DIN DKE

DEUTSCHE NORMUNGSROADMAP
KÜNSTLICHE INTELLIGENZ

AUSGABE 2

Gefördert durch:

Bundesministerium
für Wirtschaft
und Klimaschutz

aufgrund eines Beschlusses
des Deutschen Bundestages

HERAUSGEBER
Wolfgang Wahlster
Christoph Winterhalter

DIN e. V.

Am DIN-Platz
Burggrafenstr. 6
10787 Berlin
Tel.: +49 30 2601-0
E-Mail: presse@din.de
Internet: www.din.de

DKE Deutsche Kommission Elektrotechnik
Elektronik Informationstechnik in DIN und VDE
Merianstraße 28
63069 Offenbach am Main
Tel.: +49 69 6308-0
Fax: +49 69 6308-9863
E-Mail: dke@vde.com
Internet: www.dke.de

Referenzierung der Deutschen Normungsroadmap Künstliche Intelligenz:
DIN, DKE (2022): Deutsche Normungsroadmap Künstliche Intelligenz (Ausgabe 2);
www.din.de/go/normungsroadmapki

Bildnachweise:
Titelbild: pinkeyes – stock.adobe.com
Kapiteleingangsgrafiken:
kras99 (S. 11, 55), assistant (S. 33), Thitichaya (S. 41), Maxim (S. 57), peshkov (S. 105), gunayaliyeva (S. 127), pickup (S. 153, 247, 273, 297, 307), LuckyStep (S. 177), kaptn (S. 199), ryzhi (S. 225), Alex (S. 285) – stock.adobe.com

Stand: Dezember 2022

VORWORT

Prof. Dr. rer. nat. Dr. h.c. mult. Wolfgang Wahlster
CEA des Deutschen Forschungszentrums für Künstliche Intelligenz (DFKI)

Christoph Winterhalter
Vorsitzender des Vorstands, DIN

Sehr geehrte Leserinnen und Leser,

Mit der zweiten Ausgabe der Deutschen Normungsroadmap Künstliche Intelligenz können wir heute eine erweiterte und aktualisierte Analyse des Bestands und des Bedarfs an internationalen Normen und Standards für diese Schlüsseltechnologie vorlegen. Damit wollen wir an den großen Erfolg der im November 2020 publizierten Roadmap anschließen, deren Ergebnisse nach der ersten Vorstellung auf dem Digital-Gipfel der Bundesregierung ein großes internationales Echo nicht nur in Fachkreisen, sondern auch in politischen Gremien und der Presse erfahren haben.

Auch von der neuen Bundesregierung wird Normung und Standardisierung als Teil der KI-Strategie und als ein Thema mit sehr hoher Bedeutung angesehen. Nach einer Auftaktveranstaltung von DIN und DKE am 20.01.2022 wurde die nun vorliegende neue Ausgabe der Normungsroadmap KI unter aktiver Mitwirkung von mehr als 570 Fachleuten aus Wirtschaft, Wissenschaft, Zivilgesellschaft und Politik in neun Arbeitsgruppen erarbeitet. Diese wurden von einer hochrangige Koordinierungsgruppe zur KI-Normung und -Konformität mit Mandat der Bundesregierung begleitet. Der Fokus der Arbeiten lag dabei auf neun Kernthemen (Grundlagen, Sicherheit, Prüfung/Zertifizierung, Soziotechnische Systeme, Medizin, Industrielle Automation, Mobilität, Energie/Umwelt, Finanzdienstleistungen), wobei KI-Themen im Bereich von Soziotechnischen Systemen, Energie und Umwelt sowie Finanzdienstleistungen neu hinzugekommen sind. Dabei wurde auch der umfangreiche Abschlussbericht der Enquete-Kommission KI des Bundestages vom Oktober 2020 berücksichtigt.

Im Rahmen der KI-Strategie der Bundesregierung wurden inzwischen sechs KI-Kompetenzzentren als Kern der deutschen KI-Forschungslandschaft etabliert. Neben der Weiterentwicklung des DFKI wurden fünf weitere an Hochschulen angesiedelte KI-Kompetenzzentren auf- und ausgebaut, die ab Mitte dieses Jahres in eine dauerhafte institutionelle Förderung übergingen. Zusammen mit 100 neu besetzten KI-Professuren ergibt sich hiermit ein großer Schub für die KI-Forschung und den Transfer in die Wirtschaft, zu dem die Umsetzung der vorliegenden Normungsroadmap KI einen wesentlichen Beitrag leisten kann. Für die dringend benötigten Fachkräfte wurden u. a. Initiativen wie der KI-Campus zur Online-Schulung und das AI Grid für die Vernetzung und Betreuung forschender KI-Nachwuchstalente in internationalen Mikrofachgemeinschaften gestartet. Die Plattform Lernende Systeme wurde zur zentralen KI-Dialogplattform zwischen Wissenschaft, Wirtschaft, Gesellschaft und Politik weiterentwickelt und ist jetzt auch für das regelmäßige Monitoring der Umsetzung der KI-Strategie anhand von aktuellen Kennzahlen zuständig.

Mit der KI-Strategie der Bundesregierung ist eine „Blütezeit“ für die deutschen Forschenden angebrochen, um die uns viele Kolleg*innen in anderen Ländern beneiden. Selbst in den USA und China stellt der Staat in Relation zur Bevölkerungszahl kaum so viele langfristige Fördermöglichkeiten für die KI bereit.

Jetzt gilt es aber, durch verstärkten Transfer in die wirtschaftliche Wertschöpfung auch in kleinen und mittleren Unternehmen (KMUs) und durch Start-ups die Früchte dieser staatlichen Investitionen zu ernten. Dabei ist die nun startende Umsetzungsphase der Normungsroadmap KI in den Jahren 2023/2024 von größter Bedeutung, weil Normen und die darauf basierende Zertifizierung von KI-Lösungen den Unternehmen Investitionssicherheit, Rechtssicherheit und die Interoperabilität zwischen Plattformen und Wertschöpfungsnetzwerken ermöglichen und Marktanteile sichern.

Steigende Energiekosten und drohende Versorgungsengpässe, hohe Inflationsraten sowie unterbrochene Lieferketten als Folge von Coronapandemie und dem russischen Krieg gegen die Ukraine setzen der gesamten Wirtschaft zu. Künstliche Intelligenz kann als Zukunftstechnologie in dieser schwierigen geopolitischen Lage rasch mit Teillösungen zur Überwindung dieser schweren wirtschaftlichen und gesellschaftlichen Belastungen beitragen. Dazu muss aber der Zugang der für den Einsatz der KI-Technologien notwendigen Daten erleichtert und nicht innovationshemmend blockiert werden. Durch Normung und Standardisierung kann auch die oftmals kritische öffentliche Debatte zu KI-Risiken versachlicht werden, sodass die notwendigen Investitionen in diese Zukunftstechnologie von Entscheider*innen in den Unternehmen vermehrt getätigt werden.

Für Deutschland, dem innovativsten Fabrikausstatter der Welt, spielt die industrielle KI eine besondere Rolle. Industrie 4.0 ist ein Exportschlager, wie die Hannover Messe auch in diesem Jahr wieder gezeigt hat. Daran trägt die industrielle KI einen wesentlichen Anteil. Sie bildet die Basis zur Umsetzung der vierten industriellen Revolution in wandlungsfähigen, cyber-physischen Fabriken für kleine Losgrößen, in denen kollaborative und kognitive Roboter Hand in Hand mit Fachleuten in einer Kl-basierten Null-Fehler-Produktion hochqualitative Hightechprodukte klimafreundlich produzieren. Digitale Zwillinge messen dabei den $C0_2$-Abdruck während der Produktion und helfen mit KI-Algorithmen, den Energieverbrauch zu reduzieren.

Deutschland hat heute die höchste Roboterdichte in Europa und die ersten kollaborativen Roboter wurden in Deutschland bis zur kommerziellen Reife entwickelt. Aktuell gibt es hier mehr Hersteller oder deren Forschungslabore für kollaborative kognitive Roboter als in anderen Teilen der Welt. Auf dem Gebiet der industriellen KI ist Deutschland noch klar führend, wie uns auch die Kolleg*innen aus den USA und China bestätigen. Aber nun gilt es, diesen Vorsprung auch durch Normen und Standards langfristig in die industrielle Praxis zu überführen und abzusichern.

Natürlich gibt es etliche KI-Anwendungsgebiete, in denen andere Nationen derzeit führend sind – meist aber aus sehr gutem Grund: So wird in Deutschland nicht an der Überwachung der Zivilbevölkerung durch KI gearbeitet. Auch die Forschung an KI zur personalisierten und allgegenwärtigen Werbung im Internet oder zur Herstellung vollautonomer Waffensysteme ist in Deutschland nicht erwünscht und wird staatlich auch nicht gefördert.

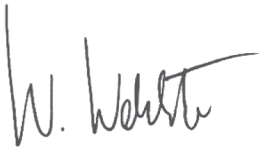
Ohne den unermüdlichen Einsatz unserer ehrenamtlich arbeitenden Fachleute in den Arbeitsgruppen wäre die rechtzeitige Erstellung dieser zweiten Ausgabe der KI-Normungsroadmap nicht möglich gewesen. Die mit hochrangigen Persönlichkeiten besetzte Koordinierungsgruppe zur KI-Normung und -Konformität hat regelmäßige Sitzungen durchgeführt und zusätzlich in einer Klausursitzung zusammen mit den Leitenden der neun Arbeitsgruppen die abschließenden Handlungsempfehlungen konsensual verabschiedet.

Auch im Namen der Koordinierungsgruppe möchten wir uns an dieser Stelle bei allen aktiven Mitgliedern der Arbeitsgruppen für das große Engagement bedanken, wobei wir Frau Filiz Elmas als exzellente Leiterin der Geschäftsstelle des Gesamtprojekts und ihrem Team unser besonderes Lob aussprechen möchten.

Die Normung ist in Deutschland eine Gemeinschaftsaufgabe, die auf der breiten Beteiligung und Mitarbeit fachkundiger Expert*innen aus Wirtschaft, Wissenschaft, Staat und Gesellschaft basiert. Nur ein frühzeitiges Engagement von Fachleuten mit breiten Erfahrungswerten wird es ermöglichen, markt- und bedarfsgerechte Normen und Standards für KI zu erarbeiten und deren Akzeptanz zu gewährleisten. Wenn Deutschland sicherstellen möchte, dass seine Interessen angemessen in internationalen KI-Standards berücksichtigt werden, gilt es, KI-Fachleute in unsere Normungsgremien zu integrieren und die Mitwirkung deutscher Expert*innen in internationalen KI-Normungsgremien zu stärken.

Wir wünschen allen Leserinnen und Lesern eine spannende Lektüre und bitten Sie um aktive Unterstützung bei der Umsetzung dieser Normungsroadmap in den kommenden Jahren.

Gemeinsam können wir dann etliche Herausforderungen der „Zeitenwende“ durch einen gezielten Einsatz von zertifizierten KI-Technologien gemäß unseres europäischen Wertesystems meistern.



Prof. Dr. rer. nat. Dr. h.c. mult. Wolfgang Wahlster,
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

Christoph Winterhalter
Vorsitzender des Vorstands, DIN

GRUSSWORT

Dr. Robert Habeck
Bundesminister für Wirtschaft
und Klimaschutz

Sehr geehrte Leserinnen und Leser,

mit dem Begriff „Künstliche Intelligenz“ (KI) verbinden wir große Hoffnungen an eine Zukunftstechnologie, die Produktions-, Arbeits- und Verwaltungsprozesse erleichtern kann. Gleichzeitig gibt es auch ernstzunehmende Vorbehalte, zum Beispiel was die Kontrolle und den Schutz von Bürgerrechten angeht. Die Aufgabe besteht nun darin, die Risiken zu minimieren, um die vielfältigen Chancen, die sich uns durch KI eröffnen, bestmöglich zu nutzen.

Dafür brauchen wir einen rechtlichen Rahmen, der einerseits die verantwortungsvolle, an den Menschen und dem Gemeinwohl orientierte Entwicklung von KI-Technologien und deren Einsatz fördert und andererseits Vertrauen bei Anwenderinnen und Anwendern schafft.

Einen solchen Rechtsrahmen wollen wir mit dem Artificial Intelligence Act schaffen, der derzeit auf EU-Ebene verhandelt wird. Ziel ist es, dass KI-Systeme, die auf dem europäischen Binnenmarkt in Verkehr gebracht werden – abhängig von ihrem Risikolevel – bestimmte Anforderungen zum Beispiel an Transparenz, Genauigkeit, IT-Sicherheit und Aufsicht durch Menschen erfüllen. Die konkrete technische Umsetzung wird dann durch europäische Normen spezifiziert werden.

Parallel zu den europäischen Prozessen werden auch auf internationaler Ebene KI-Standards entwickelt, die das künftige globale Marktumfeld der Technologie abstecken. Wer im globalen Wettbewerb um die besten KI-Lösungen an der Spitze mit dabei sein möchte, muss auch dieses Marktumfeld mitgestalten. Im Koalitionsvertrag haben wir daher festgelegt, dass wir unseren Einsatz in internationalen Normungs- und Standardisierungsprozessen stärken wollen. Durch das aktive Mitwirken bei der Entwicklung internationaler KI-Standards und Normen können wir unseren Innovationen den Weg auf globale Märkte ebnen und sicherstellen, dass europäische Werte bei der Entwicklung und Vermarktung neuer Technologien berücksichtigt werden.

In der zweiten Ausgabe der Deutschen Normungsroadmap Künstliche Intelligenz finden Sie einen Überblick über die bestehende Normungslandschaft, konkrete Standardisierungsbedarfe, die wir aus Deutschland und Europa in die internationale Normung einbringen können und Handlungsempfehlungen, welche Themenbereiche prioritär bearbeitet werden müssen. Die Deutsche Normungsroadmap KI ist eine wichtige Grundlage, um „Artificial Intelligence (AI) made in Germany“ als weltweit anerkanntes Gütesiegel für eine vertrauenswürdige Technologie zu schaffen. Ich danke allen Beteiligten bei DIN und DKE für die hervorragende Arbeit.

Im nächsten Schritt gilt es nun, die identifizierten Handlungsempfehlungen umzusetzen. Hier sind alle an der Normung beteiligten Stakeholder gefragt. Bringen Sie sich ein und ergreifen Sie die Chance, die Spielregeln für Künstliche Intelligenz mitzugestalten.

Ihr
Dr. Robert Habeck
Bundesminister für Wirtschaft und Klimaschutz

Zusammenfassung

Im Auftrag des Bundesministeriums für Wirtschaft und Klimaschutz haben DIN und DKE im Januar 2022 die Arbeiten an der zweiten Ausgabe der Deutschen Normungsroadmap Künstliche Intelligenz gestartet. In einem breiten Beteiligungsprozess und unter Mitwirkung von mehr als 570 Fachleuten aus Wirtschaft, Wissenschaft, öffentlicher Hand und Zivilgesellschaft wurde damit der strategische Fahrplan für die KI-Normung weiterentwickelt. Koordiniert und begleitet wurden diese Arbeiten von einer hochrangigen Koordinierungsgruppe für KI-Normung und -Konformität.

Mit der Normungsroadmap wird eine Maßnahme der KI-Strategie der Bundesregierung umgesetzt und damit ein wesentlicher Beitrag zur „KI – Made in Germany“ geleistet.

Die Normung ist Teil der KI-Strategie und ein strategisches Instrument zur Stärkung der Innovations- und Wettbewerbsfähigkeit der deutschen und europäischen Wirtschaft. Nicht zuletzt deshalb spielt sie im geplanten europäischen Rechtsrahmen für KI, dem Artificial Intelligence Act, eine besondere Rolle.

Die vorliegende Normungsroadmap KI zeigt die Erfordernisse in der Normung auf, formuliert konkrete Empfehlungen und schafft so die Basis, um frühzeitig Normungsarbeiten auf nationaler, insbesondere aber auch auf europäischer und internationaler Ebene, anzustoßen. Damit zahlt sie maßgeblich auf den Artificial Intelligence Act der Europäischen Kommission ein und unterstützt dessen Umsetzung.

Kapitel 1 der Roadmap führt in das Thema ein und stellt die wirtschaftspolitische Bedeutung der Normung sowie Ziele und Vorgehen der Roadmap dar.

Das aktuelle Akteurs- und Normungsumfeld für KI ist in Kapitel 3 beschrieben. Dort wird eine Übersicht über relevante innovationspolitische Initiativen, Forschungsprojekte sowie Normungs- und Standardisierungsaktivitäten gegeben.

Der Fokus der Normungsroadmap KI liegt auf neun Schwerpunktthemen, die in Kapitel 4 behandelt werden:

- → Den Ausgangspunkt bilden die **Grundlagen** wie beispielsweise Terminologien und Begriffsbestimmungen, Klassifizierungen und ethische Fragestellungen. Sie sind die Basis für Diskussionen rund um KI und damit zentraler Kern der Roadmap.
- → Für eine breite Nutzung von KI-Lösungen spielt die **Sicherheit** von KI-Systemen eine entscheidende Rolle. Nur eine tiefergehende Betrachtung von Anforderungen beispielsweise an die Betriebs- und Informationssicherheit kann einen umfassenden Einsatz von KI-Systemen in Wirtschaft und Gesellschaft ermöglichen.
- → Ein weiteres Schwerpunktthema und Grundlage für einen breiten Markterfolg von KI sind die **Prüfung und Zertifizierung**. Hierfür braucht es verlässliche Qualitätskriterien und reproduzierbare Prüfverfahren, mit denen sich die Eigenschaften von KI-Systemen überprüfen lassen. Sie sind eine Schlüsselvoraussetzung für die Bewertung der Qualität von KI-basierten Anwendungen und tragen maßgeblich zur Erklärbarkeit und Nachvollziehbarkeit bei – zwei Faktoren, die Vertrauen und Akzeptanz schaffen.
- → Eine weitere Herausforderung beim Einsatz von KI, insbesondere für kleinere und mittlere Unternehmen, stellt die Integration der KI-Technologien in Organisationen dar. Im Mittelpunkt stehen **soziotechnische Aspekt**e wie die Mensch-Technik-Interaktion, die humane Arbeitsgestaltung sowie Anforderungen an Unternehmensstrukturen und -prozesse, die in der Roadmap untersucht werden.
- → Die Anwendungsgebiete von KI sind äußerst vielfältig. In nahezu allen Wirtschafts- und Anwendungsbereichen kommen KI-Technologien zum Einsatz und bieten großes Potenzial. Um ein breites Spektrum an Anwendungen abzudecken, werden in der Roadmap neben den oben genannten querschnittlichen Themen insbesondere auch branchenspezifische Herausforderungen für die folgenden fünf Sektoren betrachtet: **Industrielle Automation, Mobilität, Medizin, Finanzdienstleistungen** sowie **Energie / Umwelt**.

Die vorliegende Roadmap skizziert für alle neun Schwerpunktthemen die Arbeits- und Diskussionsergebnisse und gibt einen umfassenden Überblick über Status quo, Anforderungen sowie Handlungsbedarfe.

Mit mehr als 116 identifizierten Normungs- und Standardisierungsbedarfen zeigt die Roadmap konkrete Potenziale in allen Kernthemen auf und formuliert in Kapitel 2 sechs zentrale Handlungsempfehlungen:

- → Entwicklung, Validierung und Standardisierung eines horizontalen Konformitätsbewertungs- und Zertifizierungsprogramms für vertrauenswürdige KI-Systeme
- → Aufbau von Dateninfrastrukturen und Erarbeitung von Datenqualitätsstandards zur Entwicklung und Validierung von KI-Systemen
- → Betrachtung des Menschen als Teil des Systems in allen Phasen des KI-Lebenszyklus
- → Entwicklung von Vorgaben für die Konformitätsbewertung von kontinuierlich oder stufenweise lernenden Systemen im Bereich der Medizin
- → Entwicklung und Einsatz sicherer und vertrauenswürdiger KI-Anwendungen in der Mobilität durch Best Practices und Absicherung
- → Entwicklung übergreifender Datenstandards und dynamischer Modellierungsverfahren zur effizienten und nachhaltigen Gestaltung von KI-Systemen

Die hohe Dynamik in der KI-Technologieentwicklung und der schnelle Anstieg der industriellen Anwendungen von KI-Systemen stellen auch neuartige Anforderungen an die Normungsprozesse und an die Bereitstellung und Weiterverwertung von Normeninhalten. Um diesen Herausforderungen zu begegnen, erarbeiten die Normungsorganisationen neue Ansätze, die in Kapitel 5 der Roadmap aufgeführt sind. Im Fokus stehen dabei die Überprüfung und Anpassung des Normenbestands, die Analyse von Standardisierungsbedarfen sowie die agile Entwicklung und bedarfsgerechte Bereitstellung von Normen und Standards.

Die vorliegende Normungsroadmap gibt den Weg für die zukünftige Normung und Standardisierung im Bereich der Künstlichen Intelligenz vor. Bereits in Ausgabe 1 der Roadmap wurden erste Handlungsbedarfe identifiziert, von denen ein Großteil als Normungs- und Forschungsprojekte angestoßen oder umgesetzt werden konnte. Den aktuellen Stand der Umsetzungsaktivitäten der ersten Ausgabe beschreibt Kapitel 6.

Die Veröffentlichung der zweiten Ausgabe der Normungsroadmap stellt den Startpunkt für die Umsetzung der Ergebnisse dar. Auch hier gilt es, Normungs- und Standardisierungsaktivitäten entlang der Handlungsempfehlungen auf den Weg zu bringen und mithilfe der entstehenden Normen und Standards die identifizierten Potenziale zu heben. Normen und Standards werden die deutsche Wirtschaft und Wissenschaft dabei unterstützen, innovationsfreundliche Bedingungen für die Technologie der Zukunft zu schaffen. Insbesondere zur gesellschaftspolitischen Debatte über die Rolle und den Einsatz von KI können die Ergebnisse der Roadmap einen wichtigen Beitrag leisten.

Die Normung in Deutschland basiert auf der Mitarbeit fachkundiger Expert*innen aus Wirtschaft, Wissenschaft, öffentlicher Hand und Zivilgesellschaft. Nur ein frühzeitiges Engagement von KI-Fachleuten in den Normungsgremien wird es ermöglichen, deutsche Interessen in internationalen Standards einzubringen und damit einerseits marktgerechte Normen und Standards für KI zu erarbeiten und andererseits die Position Deutschlands als Wirtschaftsnation und Exportland zu stärken.

1 Einleitung

Von Robotern, die in der Industrie 4.0 zusammenarbeiten, über intelligente Sprachassistenten bis zu autonom fahrenden Autos – Künstliche Intelligenz verändert unsere Wirtschaft und Gesellschaft nachhaltig. Die selbstlernenden und sich fortlaufend verbessernden KI-Systeme ermöglichen effizientere Abläufe in Produktion und anderen Bereichen. Vollkommen neue Geschäftsmodelle können durch sie entstehen. Die Möglichkeiten sind grenzenlos – und doch sollte sich eine so einflussreiche Technologie innerhalb bestimmter Grenzen bewegen, damit sie uns tatsächlich hilft. Eine zuverlässige, funktionale und vor allem sichere KI braucht gewisse Regeln: zunächst ein gemeinsames Verständnis und eine einheitliche Sprache, sodass alle vom Gleichen reden. Außerdem sind offene Schnittstellen nötig, damit die Systeme ihr volles Potenzial ausschöpfen und effizient zusammenarbeiten. Nur so können verschiedene KI-gesteuerte Maschinen miteinander kommunizieren, werden Produkte entlang der gesamten Wertschöpfungskette sichtbar. Gleichzeitig spielen ethische Fragen eine zentrale Rolle beim Einsatz Künstlicher Intelligenz. Verzerrung, Diskriminierung und Manipulation sollten von vornherein verhindert werden, wenn KI dem Menschen nutzen soll.

Bei all diesen Aspekten leisten Normen und Standards einen zentralen Beitrag: Sie definieren Anforderungen an Künstliche Intelligenz und strukturieren die Technologielandschaft. Damit sind sie ein strategisch wichtiges Instrument zur Stärkung der Innovations- und Wettbewerbsfähigkeit der deutschen Wirtschaft. Der geschätzte wirtschaftliche Nutzen von Normen beträgt rund 17 Milliarden Euro im Jahr [1]. Nicht zuletzt deshalb fiel jetzt der Startschuss für die Arbeiten an der zweiten Ausgabe der Normungsroadmap Künstliche Intelligenz. Die Aufgabe der vorliegenden Roadmap ist es, einen strategischen Fahrplan für die KI-Normung zu formulieren.

1.1 Rolle der Normung und Standardisierung bei KI

Die Entwicklung von KI-Anwendungen ist in den letzten Jahren rapide vorangeschritten, daher ist die Schaffung eines zukunftsfähigen Handlungsrahmens für KI essenziell.

Normen und Standards spielen dabei eine wichtige Rolle. Sie ermöglichen eine zuverlässige und sichere Anwendung von KI-Technologien und tragen zur Erklärbarkeit und Nachvollziehbarkeit bei. Das wiederum macht sie zu Schlüsselfaktoren für die Akzeptanz von KI-Anwendungen und schafft Vertrauen am Markt und bei Verbraucher*innen.

Für den breiten Markterfolg von KI sind Normen und Standards ein entscheidender Faktor: Sie helfen, Innovationen schneller zu etablieren, indem sie den schnellen Transfer von Technologien aus der Forschung in die Anwendung fördern und somit deutschen Unternehmen den Eintritt in europäische und internationale Märkte erleichtern. Gerade kleine und mittelständische Unternehmen profitieren davon, denn offene Schnittstellen und einheitliche Anforderungen erleichtern ihnen den Zugang zu internationalen Märkten.

Wer sich bei der Erarbeitung von Normen und Standards einbringt, kann die globalen technischen Regeln für KI aktiv mitgestalten und sich so einen Vorsprung verschaffen. Ein frühzeitiges Engagement deutscher Stakeholder in der nationalen, europäischen und internationalen Normung ist daher unabdingbar, um Deutschland als Weltwirtschaftsnation und Exportland zu stärken. Internationale Wettbewerber haben diesen Vorteil erkannt, vor allem China und die USA sind große Treiber der internationalen KI-Standardisierung. Wenn Deutschland und seine europäischen Partner sicherstellen wollen, dass europäische Wertmaßstäbe und ethische Richtlinien angemessen in internationalen KI-Standards berücksichtigt werden, ist die Mitarbeit in der Normung und eine verstärkte Präsenz in internationalen KI-Normungsgremien dringend angeraten.

Auch die Politik hat die Normung als strategisches Instrument für die internationale Wettbewerbsfähigkeit erkannt. Nicht zuletzt deshalb hat die Bundesregierung Normung und Standardisierung als zentrales Element in ihrer KI-Strategie (siehe Kapitel 1.3) festgelegt. Die Europäische Kommission veröffentlichte im Frühjahr 2021 einen Vorschlag für einen Rechtsrahmen für KI: den Artificial Intelligence Act (AI Act)

(siehe Kapitel 1.4). Mit diesem weltweit ersten Rechtsrahmen für KI will die EU die Sicherheit und Grundrechte von Menschen und Unternehmen beim Einsatz von KI sicherstellen und gleichzeitig Investitionen sowie Innovationen stärken. Der geplante AI Act weist dabei Normung und Standardisierung eine zentrale Rolle zu: Insbesondere im Bereich der Hochrisiko-KI-Anwendungen sollen harmonisierte Europäische Normen künftig Anforderungen an Transparenz, Robustheit und Genauigkeit technisch konkretisieren.

1.2 Ziele und Inhalte der Normungsroadmap KI

Normen und Standards leisten einen zentralen Beitrag, wenn es darum geht, Anforderungen an Künstliche Intelligenz zu definieren und die Technologielandschaft zu strukturieren.

Die frühzeitige Entwicklung eines strategischen Fahrplans, der die Erfordernisse im Bereich der Normung und Standardisierung identifiziert und Empfehlungen ausspricht, ist essenziell. Einen solchen strategischen Fahrplan stellt die Normungsroadmap KI dar. Sie beschreibt einen Handlungsrahmen für die Normung und Standardisierung im Bereich KI, der auf Basis eines breiten Abstimmungsprozesses entsteht und damit die wesentliche Grundlage ist, um entsprechende Arbeiten in der Normung und Standardisierung auf nationaler, aber vor allem auf europäischer und internationaler Ebene anzustoßen.

Die Normungsroadmap KI setzt damit eine wesentliche **Maßnahme der KI-Strategie der Bundesregierung** [2] um und trägt maßgeblich dazu bei, die nationale Position frühzeitig auf europäischer und internationaler Ebene einzubringen und damit die Rolle Deutschlands als Wirtschaftsnation und Exportland entscheidend zu stärken. Ziel der Normungsroadmap KI ist es, innovationsfreundliche Rahmenbedingungen für die Technologie der Zukunft zu schaffen und die deutsche Wirtschaft und Wissenschaft im internationalen Wettbewerb um die besten Lösungen und Produkte im Bereich KI zu unterstützen.

Die Normungsroadmap ist als „lebendes Dokument“ zu verstehen, das die aktuellen Arbeits- und Diskussionsergebnisse vorstellt und als zentrales Kommunikationsmedium zum Austausch zwischen Normungsgremien, Industrie, Verbänden, Forschungseinrichtungen, Zivilgesellschaft und Politik dient. Sie wird in einem offenen, transparenten und breit angelegten Beteiligungsprozess durch Vertreter*innen aus Wirtschaft, Wissenschaft, öffentlicher Hand und Zivilgesellschaft erarbeitet und regelmäßig fortgeschrieben.

1.2.1 Ziele der Normungsroadmap KI

Die Roadmap verfolgt zwei **wesentliche Ziele**: Erstens beschreibt sie das Umfeld, in dem sich die KI-Standardisierung bewegt, und gibt einen Überblick über bereits bestehende Normen und Standards zu Aspekten der KI. Zweitens zeigt sie Normungs- und Standardisierungsbedarfe auf und formuliert konkrete Handlungsempfehlungen. Die Roadmap gibt damit den Weg für die zukünftige Normung und Standardisierung im Bereich KI vor und leistet einen wesentlichen Beitrag, um „KI – Made in Germany“ als starke Marke zu etablieren und neue Geschäftsmodelle, disruptive Innovationen und skalierbare Anwendungen zu entwickeln. Damit ist sie der Wegweiser für die KI-Standardisierung und bietet gleichzeitig großes Potenzial, um europäische Wertmaßstäbe auf die internationale Ebene zu heben.

Zielgruppe der Normungsroadmap ist die breite KI-Fachöffentlichkeit. Ihre Empfehlungen richten sich dabei in erster Linie an die Wirtschaft, aber auch an Vertreter*innen der Qualitätsinfrastruktur, Politik, Forschung und Zivilgesellschaft.

Erste Ausgabe der Normungsroadmap KI

DIN und Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE (DKE) haben im Auftrag des Bundesministeriums für Wirtschaft und Klimaschutz (BMWK) die Arbeiten an der ersten Ausgabe der Deutschen Normungsroadmap „Künstliche Intelligenz“ bereits Ende 2019 angestoßen. Unter Beteiligung von über 300 Expert*innen aus verschiedenen Bereichen wurde die erste Normungsroadmap KI erarbeitet und im November 2020 auf dem Digital-Gipfel der Bundesregierung erstmals der Fachöffentlichkeit vorgestellt und veröffentlicht. Die Ergebnisse der Normungsroadmap KI stellen eine Bestandsaufnahme dar und dienen als strategischer Fahrplan für die Normung und Standardisierung im Bereich KI. Seit ihrer Veröffentlichung wird mit Hochdruck an der Umsetzung der insgesamt 78 identifizierten Handlungsempfehlungen und Bedarfe aus den sieben Themenschwerpunkten (Grundlagen, Ethik, Qualität/Zertifizierung/Konformitätsbewertung, IT-Sicherheit, Indust-

rielle Automation, Mobilität/Logistik und Medizin) gearbeitet. Eine Vielzahl an Normungs- und Standardisierungsprojekten konnte erfolgreich initiiert und auch europäisch bzw. international eingebracht werden (siehe Kapitel 6).

Zur Umsetzung der übergreifenden Handlungsempfehlungen der Normungsroadmap KI (Ausgabe 1) wurden darüber hinaus sogenannte „Leuchtturmprojekte der Deutschen Normungsroadmap KI“ ins Leben gerufen. Mithilfe dieser Umsetzungsprojekte werden im jeweiligen Anwendungskontext praktische Erfahrungen gesammelt, konkrete Normungs- und Standardisierungsbedarfe abgeleitet und Erkenntnisse zur Qualitäts- und Konformitätsprüfung gewonnen. Den Leuchtturmprojekten kommt damit eine besondere Bedeutung bei der Umsetzung der Normungsroadmap KI zu, weshalb sie eine erhöhte Aufmerksamkeit bei den Normungsakteur*innen genießen und in Wirtschaft, Forschung und Politik weithin sichtbar sind. Eine Übersicht zu den initiierten Normungsprojekten und zum Stand der Umsetzungsaktivitäten der Ausgabe 1 der Roadmap ist in Kapitel 6 dargestellt.

Zweite Ausgabe der Normungsroadmap KI

Die hohe Dynamik in der KI-Forschung und der industriellen Entwicklung und Anwendung einerseits und sich abzeichnende Veränderungen auf regulatorischer Ebene andererseits erfordern eine kontinuierliche Weiterentwicklung des strategischen Handlungsrahmens und der Empfehlungen der Normungsroadmap KI. Deshalb fiel im Januar 2022 im Auftrag des BMWK der Startschuss für die Arbeiten an der zweiten Ausgabe der Normungsroadmap Künstliche Intelligenz. Ziel ist die Fortschreibung und Weiterentwicklung der bisherigen Ergebnisse. Damit baut die vorliegende Roadmap auf den Ergebnissen der ersten Ausgabe auf und wird als alleinstehendes Dokument betrachtet.

Die Normungsroadmap KI fokussiert dabei auf ausgewählte Schwerpunktthemen: Es werden sowohl horizontale und querschnittliche Aspekte als auch sektorspezifische Herausforderungen beleuchtet. Neben den bisherigen Themen wie Grundlagen, Sicherheit, Prüfung und Zertifizierung, Industrielle Automation, Mobilität sowie Medizin wird das Augenmerk zusätzlich auf die neuen Schwerpunkte Soziotechnische Systeme, Finanzdienstleistungen und Energie/Umwelt gerichtet. Aufgabe der Roadmap ist es, einen umfassenden Überblick über Status quo, Anforderungen und Herausforderungen sowie Normungs- und Standardisierungsbedarfe zu den oben genannten neun Schwerpunktthemen zu geben. Im Rahmen der zweiten Ausgabe der Normungsroadmap KI werden verstärkt branchenrelevante Anwendungsfälle betrachtet und daraus konkrete Normungs- und Standardisierungsbedarfe abgeleitet.

Zudem zahlt die vorliegende Normungsroadmap KI maßgeblich auf den von der Europäischen Kommission veröffentlichten Entwurf zum Artificial Intelligence Act (AI Act) ein. Dieser geplante, weltweit erste Rechtsrahmen für KI weist der Normung eine zentrale Rolle zu. Insbesondere im Bereich der Hochrisiko-KI-Anwendungen sollen Anforderungen an KI-Systeme künftig durch harmonisierte Europäische Normen technisch konkretisiert werden (siehe Kapitel 1.4). Eine Aufgabe der zweiten Ausgabe der Normungsroadmap KI ist es deshalb auch, Bedarfe für Normen und Standards zur Umsetzung des AI Act zu identifizieren und diese bei der Ausgestaltung des Weiteren Fahrplans für die Normung und Standardisierung zu berücksichtigen.

An die Veröffentlichung der Normungsroadmap KI schließt sich die Phase der Umsetzung und Verstetigung der Ergebnisse an: Im Rahmen von Umsetzungs- bzw. Leuchtturmprojekten sollen für anwendungstypische und branchenrelevante KI-Anwendungen praktische Erfahrungen gesammelt, Anforderungen aufgezeigt sowie konkrete Normungs- und Standardisierungsbedarfe abgeleitet und umgesetzt werden. Zentrales Ziel der Verstetigung ist es, die identifizierten Themen und Bedarfe (siehe Kapitel 4) in den relevanten Normungsgremien einzugliedern, konkrete Normungs- und Standardisierungsaktivitäten anzustoßen und schließlich Normen und Standards zu entwickeln. Der Fokus für die Jahre 2023 und 2024 wird daher auf der Verstetigung der Ergebnisse der Normungsroadmap liegen.

Dabei ist zu beachten, dass die Normung stets eine Gemeinschaftsaufgabe darstellt, die auf der breiten Beteiligung und Mitarbeit fachkundiger Expert*innen aus Wirtschaft, Wissenschaft, Staat und Gesellschaft basiert. Nur ein frühzeitiges Engagement von Fachleuten mit breiten Erfahrungswerten und Einblicken aus der Praxis wird es ermöglichen, markt- und bedarfsgerechte Normen und Standards für KI zu erarbeiten und deren Akzeptanz in Wirtschaft, Wissenschaft und Gesellschaft zu gewährleisten. Wenn Deutschland sicherstellen möchte, dass seine Interessen angemessen in internationalen KI-Standards Berücksichtigung finden, sind die aktive Mitarbeit in der Normung und verstärkte Präsenz in internationalen KI-Normungsgremien dringend angeraten.

1.2.2 Koordinierungsgruppe KI-Normung und -Konformität

Die Arbeiten an der Normungsroadmap KI und deren Umsetzung werden durch die hochrangig besetzte Koordinierungsgruppe „KI-Normung und -Konformität“ [1] (kurz: Koordinierungsgruppe) gesteuert und begleitet. Diese wurde im Mai 2021 mit einem Mandat der Bundesregierung, vertreten durch das Bundesministerium für Wirtschaft und Klimaschutz (BMWK), das Bundesministerium für Bildung und Forschung (BMBF) und das Bundesministerium für Arbeit und Soziales (BMAS), gegründet und setzt sich aus führenden Persönlichkeiten aller relevanten Bereiche für KI zusammen. Die 17 Mitglieder aus Wirtschaft, Normung, Politik, Wissenschaft und Zivilgesellschaft repräsentieren wichtige Themen, Disziplinen, Branchen und Unternehmen unterschiedlicher Größe und verstehen sich als Botschafter*innen für die KI-Normung (siehe Abbildung 1). Damit löst die Koordinierungsgruppe die „Steuerungsgruppe Normungsroadmap KI“ ab, die die bisherigen Aktivitäten zur Ausgabe 1 der Normungsroadmap begleitet und gelenkt hat.

Die Koordinierungsgruppe verantwortet die inhaltliche und strategische Ausrichtung der Roadmap, gibt Impulse zu wichtigen innovations- und gesellschaftspolitischen Entwicklungen und setzt sich für die nationale und internationale Zusammenarbeit im Bereich KI ein. Darüber hinaus treibt sie die praktische Umsetzung der Empfehlungen der Roadmap gezielt voran und koordiniert sämtliche sich daraus erwachsende Aktivitäten. Gleichzeitig dient sie als allgemeine Anlaufstelle für Normung und Standardisierung zum Thema Künstliche Intelligenz und als Ort, an dem sich die gesamte deutsche KI-Landschaft koordinieren, austauschen und beteiligen kann. Fachlich wird die Koordinierungsgruppe durch den Expert*innenkreis unterstützt, dessen 24 Mitglieder [2] jeweils zur Unterstützung der Koordinierungsgruppenmitglieder und/oder zur Verzahnung mit relevanten Initiativen oder Akteur*innen berufen wurden.

1.2.3 Methodisches Vorgehen

Den Auftakt der Arbeiten für die zweite Ausgabe der Normungsroadmap KI gab eine virtuelle Veranstaltung [3] am 20. Januar 2022 unter Teilnahme von mehr als 600 Teilnehmer*innen (siehe Abbildung 2). Redner*innen aus Politik, Wirtschaft, Normung, Wissenschaft sowie zivilgesellschaftlichen Organisationen erläuterten Ziele und Vorgehen der Roadmap und gaben thematische Einblicke in die Schwerpunktthemen. Darüber hinaus wurden praktische Umsetzungsprojekte beleuchtet, die aus den Handlungsempfehlungen der ersten Ausgabe der KI-Roadmap hervorgegangen sind.

Wie auch schon in der ersten Ausgabe der Roadmap stellt die Mitwirkung von Expert*innen aller relevanten Kreise die wesentliche Grundlage bei der Erarbeitung der Normungsroadmap dar.

Interessierte Vertreter*innen aus Wirtschaft, Wissenschaft, öffentlicher Hand und Zivilgesellschaft sowie Repräsentant*innen bereits konstituierter und mit dem Thema KI befasster Kreise waren eingeladen, sich bei der Erarbeitung der Normungsroadmap mit ihrer Expertise aktiv einzubringen. Hierbei ist die Berücksichtigung verschiedener Sichtweisen und damit verbundener Anforderungen von hoher Bedeutung, sodass sowohl technische als auch nicht-technische Aspekte gleichermaßen Eingang in den Entstehungsprozess der Normungsroadmap KI fanden.

Mehr als 570 Fachleute aus verschiedenen Branchen und mit unterschiedlichen Erfahrungshintergründen konnten für die Mitarbeit gewonnen werden und brachten ihr Fachwissen ein. Die Erarbeitung der Roadmap erfolgte in neun Arbeitsgruppen zu verschiedenen Schwerpunktthemen (siehe Kapitel 4) und wurde komplett virtuell auf der Kollaborationsplattform DIN.ONE [4] organisiert. Die Zusammensetzung der Arbeitsgruppen zeigt Abbildung 3.

1 www.din.de/go/koordinierungsgruppe-ki

2 Mitglieder des Expert*innenkreises sind: Dir. u. Prof. Dr. Lars Adolph (BAuA), Nikolas Becker (GI), Dr. Tarek R. Besold (DEKRA DIGITAL), Jens Brinckmann (BMWK), Egbert Fritzsche (VDA), Dr. Patrick Gilroy (TÜV-Verband), Dr. Sebastian Hallensleben (VDE), Taras Holoyad (Bundesnetzagentur), Dr. Maximilian Hösl (acatech), Dr. Jürgen Klippert (IG Metall), Alena Kühlein (DIHK), Daniel Loevenich (BSI), Dr. Christoph March (BMBF), Manfred Meiss (BMWK), Dr. Maximilian Poretschkin (Fraunhofer IAIS), Prof. Dr. Georg Rehm (DFKI), Guido Reimann (VDMA), Jochen Reinschmidt (ZVEI), Dr. Kinga Schumacher (DFKI), Rosmarie Steininger (Chemistree), Dr. Christina Strobel (KI Bundesverband), Hauke Timmermann (eco – Verband der Internetwirtschaft), Merle Uhl (Bitkom e. V.), PD Dr. Marc Wittlich (IFA der DGUV)

3 Mitschnitt der Veranstaltung: https://youtu.be/jt_xen012xU; weitere Informationen zur Veranstaltung: www.din.de/go/auftakt-ki)

4 www.din.one/site/ki

MITGLIEDER DER KOORDINIERUNGSGRUPPE:

Dr. Daniela Brönstrup
Bundesministerium für Wirtschaft und Klimaschutz (BMWK)

Dr. Joachim Bühler
TÜV-Verband

Dr. Detlef Gerst
IG Metall

Dr. Tobias Heimann
ZVEI - Zentralverband Elektrotechnik- und Elektronikindustrie und Siemens Healthineers

Dr. Wolfgang Hildesheim
Bitkom und IBM Deutschland

Dr. Vanessa Just
KI Bundesverband

Julia Kloiber
Superrr Lab

Prof. Antonio Krüger
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

Dr. Christoph Peylo
Verband Deutscher Maschinen- und Anlagenbau / Verband der Automobilindustrie und Robert Bosch GmbH

Alexander Rabe
eco Verband der Internetwirtschaft

Prof. Ina Schieferdecker
Bundesministerium für Bildung und Forschung (BMBF)

Dr. Volker Treier
Deutscher Industrie- und Handelskammertag (DIHK)

Prof. Wolfgang Wahlster
Plattform Lernende Systeme und Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

Prof. Dieter Wegener
Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE (DKE)

Christoph Winterhalter
DIN Deutsches Institut für Normung

Prof. Stefan Wrobel
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS)

STÄNDIGE GÄSTE DER KOORDINIERUNGSGRUPPE:

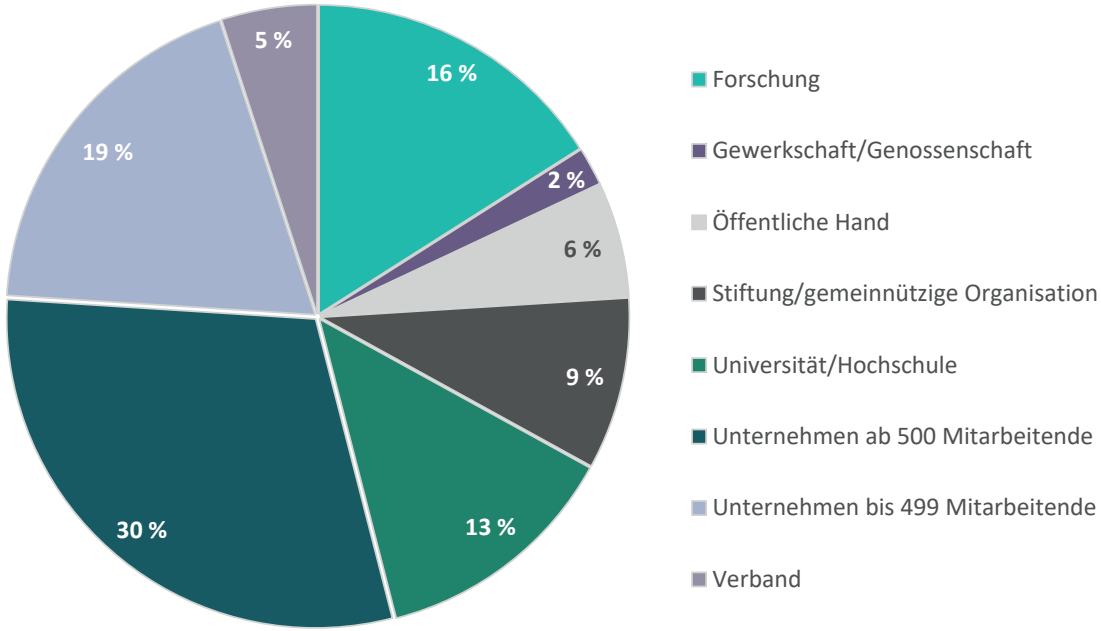
Dr. Johannes Winter[5]
Plattform Lernende Systeme

Abbildung 1: Mitglieder der Koordinierungsgruppe KI-Normung und -Konformität (Quelle: DIN)

5 bis 08/2022 Plattform Lernende Systeme, ab 09/2022 L3S AI Research Center

Abbildung 2: Impressionen von der Auftaktveranstaltung (Quelle: DIN)

Abbildung 3: Zusammensetzung der neun Arbeitsgruppen der Normungsroadmap KI (Quelle: DIN)



Für die Leitung der Arbeitsgruppen konnten erfahrene Expert*innen gewonnen werden (siehe Abbildung 4), die die inhaltlichen Arbeiten leiteten und regelmäßig an die Koordinierungsgruppe und den Expert*innenkreis berichteten.

1. Grundlagen (Leitung: Dr. Peter Deussen, Microsoft Deutschland GmbH, und Annegrit Seyerlein-Klug, neurocat GmbH)
2. Sicherheit (Leitung: Dr.-Ing. Rasmus Adler, Fraunhofer-Institut für Experimentelles Software Engineering (IESE), und Annegrit Seyerlein-Klug, neurocat GmbH)
3. Prüfung und Zertifizierung (Leitung: Dr. Maximilian Poretschkin, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS), und Daniel Loevenich, Bundesamt für Sicherheit in der Informationstechnik (BSI))
4. Soziotechnische Systeme (Leitung: Rosmarie Steininger, CHEMISTREE GmbH, Dr.-Ing. Patricia Stock, REFA-Institut e. V., und Lajla Fetic, Bertelsmann Stiftung)
5. Industrielle Automation (Leitung: Dr.-Ing. Christoph Legat, HEKUMA GmbH)
6. Mobilität (Leitung: Prof. Dr. Simon Burton, Fraunhofer-Institut für Kognitive Systeme (IKS), und Dr. Christian Müller, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI))
7. Medizin (Leitung: Dr. Jackie Ma, Fraunhofer-Institut für Nachrichtentechnik, Heinrich-Hertz-Institut (HHI), und Dr. Dirk Schlesinger, TÜV AI Lab)
8. Finanzdienstleistungen (Leitung: Dr. Oliver Maspfuhl, Deutsche Bank AG)
9. Energie/Umwelt (Leitung: Dr.-Ing. Mathias Uslar, OFFIS – Institut für Informatik, und Maximilian Schildt, Lehrstuhl für Energieeffizientes Bauen (E3D) RWTH Aachen University)

Abbildung 5 zeigt die Gesamtstruktur des Projekts der Normungsroadmap KI.

Die vorliegende Normungsroadmap KI wurde Ende 2022 veröffentlicht und der Bundesregierung übergeben. Sie steht in deutscher und englischer Sprachfassung unter http://www.din.de/go/normungsroadmapki kostenlos zum Download bereit.

Mit der Veröffentlichung der Normungsroadmap KI beginnt unmittelbar die Umsetzung und Verstetigung der Ergebnisse. Dann gilt es, möglichst viele der Handlungsempfehlungen mit Unterstützung aller Bundesministerien und der Mitwirkung von Expert*innen aus Wirtschaft, Forschung und Zivilgesellschaft in Form von konkreten Umsetzungsprojekten und Normungs- und Standardisierungsaktivitäten rasch umzusetzen.

LEITENDE DER ARBEITSGRUPPEN:

Dr. Peter Deussen
AG Grundlagen
Microsoft Deutschland GmbH

Annegrit Seyerlein-Klug
AG Grundlagen
AG Sicherheit
neurocat GmbH

Dr.-Ing. Rasmus Adler
AG Sicherheit
Fraunhofer-Institut für Experimentelles Software Engineering (IESE)

Daniel Loevenich
AG Prüfung und Zertifizierung
Bundesamt für Sicherheit in der Informationstechnik (BSI)

Dr. Maximilian Poretschkin
AG Prüfung und Zertifizierung
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS)

Lajla Fetic
AG Soziotechnische Systeme
Bertelsmann Stiftung

Rosmarie Steininger
AG Soziotechnische Systeme
CHEMISTREE GmbH

Dr.-Ing. Patricia Stock
AG Soziotechnische Systeme
REFA-Institut e. V.

Dr.-Ing. Christoph Legat
AG Industrielle Automation
HEKUMA GmbH

Prof. Dr. Simon Burton
AG Mobilität
Fraunhofer-Institut für Kognitive Systeme (IKS)

Dr. Christian Müller
AG Mobilität
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

Dr. Jackie Ma
AG Medizin
Fraunhofer-Institut für Nachrichtentechnik, Heinrich-Hertz-Institut (HHI)

Dr. Dirk Schlesinger
AG Medizin
TÜV AI Lab

Dr. Oliver Maspfuhl
AG Finanzdienstleistungen
Deutsche Bank AG

Dr.-Ing. Mathias Uslar
AG Energie/Umwelt
OFFIS – Institut für Informatik

Maximilian Schildt
AG Energie/Umwelt
Lehrstuhl für Energieeffizientes Bauen (E3D)
RWTH Aachen University

Abbildung 4: Leitende der Arbeitsgruppen (Quelle: DIN)

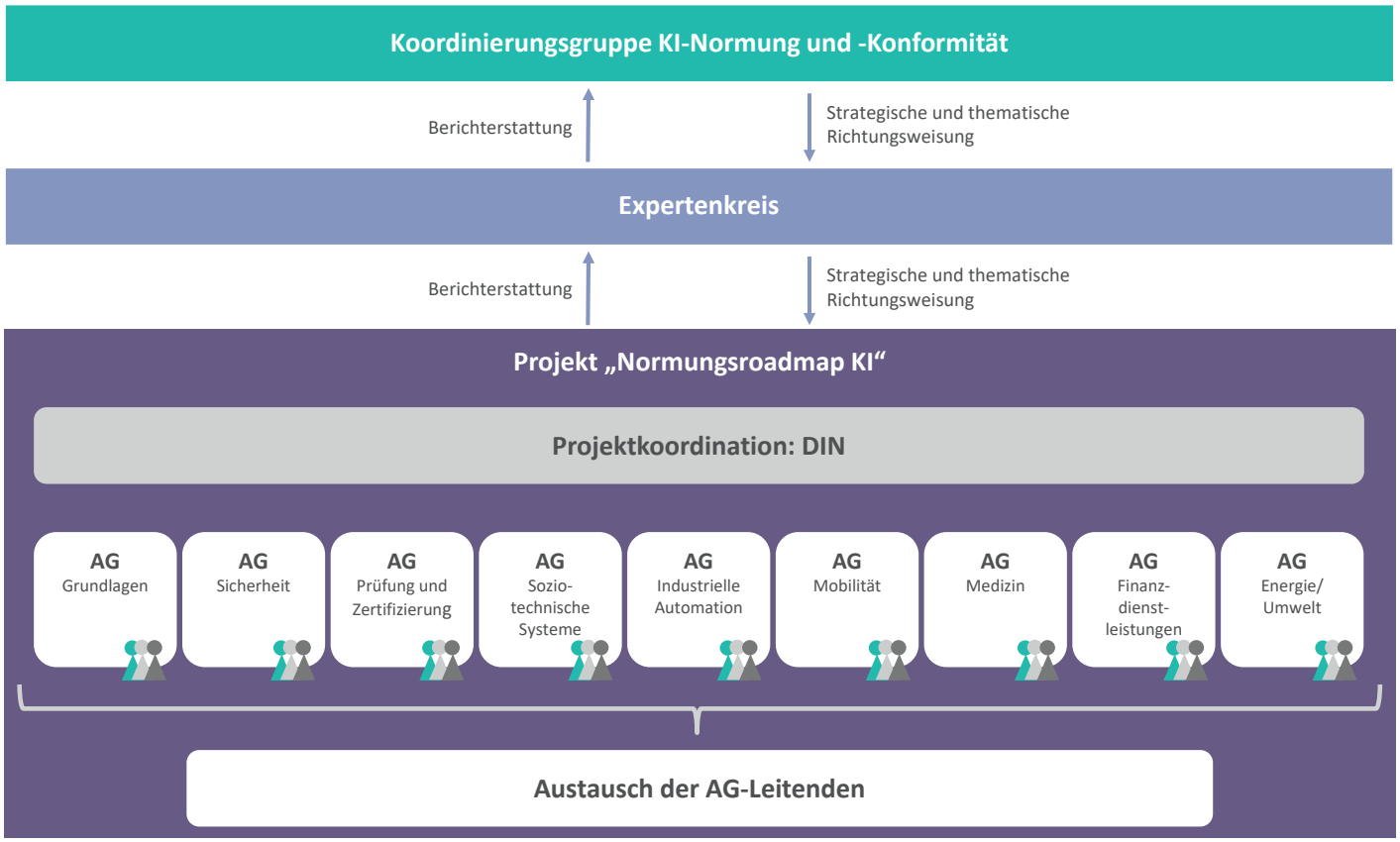


Abbildung 5: Projektstruktur der Normungsroadmap KI (Quelle: DIN)

1.3 KI-Strategie der Bundesregierung

Die Bundesregierung hat im November 2018 die nationale Strategie „Künstliche Intelligenz“ [2] verabschiedet und will mit ihr Deutschland zu einem führenden Standort für KI weiterentwickeln und die Wettbewerbsfähigkeit der deutschen und europäischen Wirtschaft, vor allem gegenüber den USA und China, stärken. Das Potenzial einer menschenzentrierten KI soll dabei entsprechend der europäischen Wirtschafts-, Werte- und Sozialstruktur genutzt werden, um die Anwendung von KI in der Breite zu fördern.

Mit der Fortschreibung der nationalen KI-Strategie Ende 2020 reagiert die Bundesregierung auf neue Entwicklungen und Bedarfe, die sich seit Veröffentlichung der Erstausgabe ergeben haben [3]. Im Fokus der Fortschreibung stehen Entwicklungen infolge der Covid-19-Pandemie, Nachhaltigkeitsthemen, insbesondere Umwelt- und Klimaschutz, sowie die europäische und internationale Vernetzung. Die finanziellen Mittel zur Umsetzung der Strategie bis 2025 wurden von bisher drei Milliarden Euro auf fünf Milliarden Euro erhöht.

Konkret will die Bundesregierung mit der Fortschreibung der KI-Strategie

- → mehr KI-Fachkräfte ausbilden, anwerben und in Deutschland halten,
- → leistungsstarke und international sichtbare Forschungsstrukturen etablieren und insbesondere modernste KI- und Rechnerinfrastrukturen auf international konkurrenzfähigem Niveau bereitstellen,
- → ausgehend von exzellenten Forschungs- und Transferstrukturen KI-Ökosysteme von internationaler Strahlkraft etablieren, um die Anwendung von Forschungsergebnissen in der betrieblichen Praxis, insbesondere im Mittelstand, zu forcieren und die Gründungsdynamik anzukurbeln,
- → die Rahmenbedingungen für innovative und menschenzentrierte KI-Anwendungen in Deutschland und Europa durch den Auf- und Ausbau der Qualitätsinfrastruktur auf der Basis eines angemessenen Ordnungsrahmens zu einem System für sichere und vertrauenswürdige KI stärken und
- → die zivilgesellschaftliche Vernetzung und die Einbeziehung in die Entwicklung und Nutzung von gemeinwohlorientierter KI unterstützen [3].

Schon in der Erstausgabe der KI-Strategie weist die Bundesregierung Normen und Standards eine zentrale Rolle zu und stellt die Normung und Standardisierung als einen zentralen Baustein der Strategie dar. Darin heißt es: „Die Bundesregierung wird (u. a.) in einem gemeinsamen Projekt mit DIN eine Roadmap zu Normen und Standards im Bereich KI entwickeln." Ferner wird die Überprüfung bestehender Normen und Standards auf „KI-Tauglichkeit" sowie die Entwicklung maschinenlesbarer und von Maschinen interpretierbarer Normen und Standards (Smart Standards) für KI-Anwendungen angeregt.

Auch in der Fortschreibung [3] wird die Bedeutung von Normen und Standards im Bereich der KI deutlich betont und herausgearbeitet. So heißt es dort:

→ „Durch das Setzen klarer Regeln sowie Standards und Normen können die Grundrechte von Bürgerinnen und Bürgern geschützt, Vertrauen in die KI gestärkt, ein nachhaltiger Einsatz sowie Innovation und Wettbewerb gefördert werden." (S. 6)

→ „[Die] Normungsroadmap für KI [...] bildet die Basis für ein nachfolgendes Umsetzungsprogramm, das auf Grundlage der Roadmap konkrete Normungsvorhaben einleiten, Zertifizierungsfragen lernender Systeme bearbeiten und die schnelle Übertragbarkeit gewonnener Erkenntnisse in internationale Standards und Prüfkriterien einleiten soll. Zentrale Themen hierbei sind u. a. Sicherheit, Robustheit, Transparenz und Nicht-Diskriminierung bei KI-Systemen." (S. 21)

→ „Zusammen mit Metrologie, Akkreditierung, Konformitätsbewertung, Marktüberwachung und Umweltprüfungen bilden Regeln, Normen und Standards die Qualitätsinfrastruktur – das Rückgrat der Marke ‚Made in Germany '." (S. 21)

→ „Umsetzung der in der Normungsroadmap KI definierten Roadmap: Entwicklung von Prüfkriterien auf der Basis etablierter und zu entwickelnder Prüftechnologien zur Prüfung der Robustheit, Sicherheit, Verlässlichkeit, Integrität, Transparenz, Erklärbarkeit, Interpretierbarkeit und Nichtdiskriminierung von (hybriden) KI-Systemen." (S. 33)

Indem die Entwicklung technischer Normen und Standards im Rahmen der nationalen KI-Strategie hervorgehoben und gefördert wird, werden Wirtschaftsprozesse erleichtert, der Technologietransfer begünstigt und über die nationale Qualitätsinfrastruktur das Vertrauen in KI-Produkte und KI-Dienstleistungen gestärkt.

1.4 KI-Regulierung auf europäischer Ebene

Die Europäische Kommission hat im April 2021 einen wegweisenden Entwurf zur Regulierung der Anwendung von Künstlicher Intelligenz veröffentlicht – den Artificial Intelligence Act (AI Act) [4]. Der Verordnungsentwurf stellt den weltweit ersten Rechtsrahmen für KI dar und basiert auf dem „Koordinierten Plan zur Künstlichen Intelligenz" [5], den „Politik- und Investitionsempfehlungen für vertrauenswürdige Künstliche Intelligenz" [6] sowie dem „Weißbuch zur Künstlichen Intelligenz" [7]. Erklärtes Ziel des geplanten AI Act ist es, beim Einsatz von KI die Grundrechte und die Sicherheit in der Europäischen Union sicherzustellen und gleichzeitig Investitionen und Innovationen in den EU-Mitgliedsstaaten zu fördern.

In dem Entwurf geht die Kommission hierbei grundsätzlich von einem sehr weiten Begriff von Künstlicher Intelligenz aus und verfolgt einen technologieneutralen und risikobasierten Ansatz. Auf Basis der Empfehlungen der von der Kommission eingesetzten unabhängigen und hochrangigen Expert*innengruppe (High Level Expert Group, HLEG) wurden drei essenzielle Komponenten für vertrauenswürdige Künstliche Intelligenz definiert: Rechtmäßigkeit, Ethik und Robustheit. Als Fundamente einer vertrauenswürdigen KI wurden vier ethische Grundsätze identifiziert: Achtung der menschlichen Autonomie, Schadensverhütung, Fairness und Erklärbarkeit. Die Verwirklichung dieser Grundsätze wurde in sieben Kernanforderungen beschrieben (siehe auch Kapitel 4.1.2.1):

→ Vorrang menschlichen Handelns und menschliche Aufsicht
→ Technische Robustheit und Sicherheit
→ Datenschutz und Datenqualitätsmanagement
→ Transparenz
→ Vielfalt, Nichtdiskriminierung und Fairness
→ Gesellschaftliches und ökologisches Wohlergehen
→ Rechenschaftspflicht [8]

Der geplante AI Act unterstützt die Empfehlungen der HLEG. Weltweit soll durch das Gesetz eine europäische Führungsrolle bei der Entwicklung von sicherer, vertrauenswürdiger und ethisch vertretbarer KI eingenommen werden. Im Rahmen der technischen Ausgestaltung der rechtlichen Anforderungen des Rechtsaktes werden Europäische Normen eine wichtige Rolle spielen.

1.4.1 Geltungsbereich

Der Gesetzesvorschlag wird aktuell im Europäischen Rat und im Europäischen Parlament beraten. Der Trilog zwischen den drei gesetzgebenden Institutionen (Europäischer Kommission, Europäischem Rat und Europäischem Parlament) wird für die Verabschiedung vorbereitet. Das Inkrafttreten des AI Act wird bei erwarteter Verabschiedung 2022/2023 voraussichtlich 24 Monate später, also 2024/2025, erfolgen [9].

Das „Gesetz über Künstliche Intelligenz" soll als Verordnung erlassen werden. Eine Verordnung ist ein verbindlicher Rechtsakt der Europäischen Union mit allgemeiner Gültigkeit und unmittelbarer Wirksamkeit in allen Mitgliedsstaaten; eine Umsetzung in nationales Recht ist nicht erforderlich. Zivilrechtliche Fragen beim Einsatz von KI (z. B. Haftung, Zurechnung von Willenserklärungen, Schaffung von geistigem Eigentum etc.) werden durch den Verordnungsentwurf nicht geregelt. Es handelt sich primär um ein Gesetz, das den Einsatz von KI-Systemen in bestimmten Anwendungsszenarien verbietet oder von technisch-organisatorischen Voraussetzungen abhängig macht. Die Konkretisierung der technischen Anforderungen an erlaubte Hochrisiko-KI-Systeme soll dabei durch harmonisierte Europäische Normen erfolgen. Art. 3 Abs. 1 des Verordnungsentwurfs definiert KI-Systeme hierbei als „Software, die mit einer oder mehreren der in Anhang I aufgeführten Techniken und Konzepte entwickelt worden ist und im Hinblick auf eine Reihe von Zielen, die vom Menschen festgelegt werden, Ergebnisse wie Inhalte, Vorhersagen, Empfehlungen oder Entscheidungen hervorbringen kann, die das Umfeld beeinflussen, mit dem sie interagieren". Die Definition kann im Trilog mit Europäischem Parlament und Europäischem Rat noch verändert werden.

Tabelle 1: Anwendungsbereich des und Strafzahlungen nach dem geplanten AI Act (Stand: Kommissionsentwurf [4])

Anwendungsbereich	a) Anbieter, die KI-Systeme in der Union in Verkehr bringen oder in Betrieb nehmen, unabhängig davon, ob diese Anbieter in der Union oder in einem Drittland niedergelassen sind b) Nutzer*innen von KI-Systemen, die sich in der Union befinden c) Anbieter*innen und Nutzer*innen von KI-Systemen, die in einem Drittland niedergelassen oder ansässig sind, wenn das vom System hervorgebrachte Ergebnis in der Union verwendet wird
Regulatorischer Fokus	Hochrisiko-KI-Systeme
Normungsbezug	Der Verordnungsentwurf sieht vor, zur technischen Ausgestaltung der grundlegenden Anforderungen an Hochrisiko-KI-Systeme auf harmonisierte Europäische Normen zu verweisen, die von den Europäischen Normungsorganisationen auf Basis eines Normungsauftrags der Europäischen Kommission erarbeitet werden (Art. 40 des Kommissionsentwurfs).
Strafzahlungen	a) 6 % des gesamten weltweiten Jahresumsatzes des vorangegangenen Geschäftsjahres oder 30 Millionen Euro – je nachdem, welcher Betrag höher ist (für Nichtkonformität mit Art. 5 und 10) b) 4 % des gesamten weltweiten Jahresumsatzes des vorangegangenen Geschäftsjahres oder 20 Millionen Euro – je nachdem, welcher Betrag höher ist (für Nichtkonformität)
Betroffene KI-Systeme	Alle KI-Systeme (spezifiziert in Anhang 1)
Zeitleiste	→ Erwartete abschließende Beratung im Europäischen Parlament: frühestens Q4 2022 → Erwartete abschließende Beratung im Europäischen Rat: frühestens Q4 2022 → Erwarteter Trilog: 2023 → Inkrafttreten (nach aktuellem Entwurf): 20 Tage nach Verabschiedung Anwendung (nach aktuellem Entwurf): 24 Monate nach Inkrafttreten

In Anhang I werden diese in der vorgeschlagenen Definition benannten Techniken und Konzepte wie folgt näher bestimmt[6]:

- → Konzepte des Maschinellen Lernens, mit beaufsichtigtem, unbeaufsichtigtem und bestärkendem Lernen unter Verwendung einer breiten Palette von Methoden, einschließlich des tiefen Lernens (Deep Learning)
- → Logik- und wissensgestützte Konzepte, einschließlich Wissensrepräsentation, induktiver (logischer) Programmierung, Wissensgrundlagen, Inferenz- und Deduktionsmaschinen, (symbolischer) Schlussfolgerungs- und Expertensysteme
- → Statistische Ansätze, Bayes'sche Schätz-, Such- und Optimierungsmethoden

Gemäß Art. 4 und 73 des Verordnungsentwurfs soll der Europäischen Kommission die Befugnis übertragen werden, Anhang I jederzeit an Marktentwicklungen und technische Entwicklungen anzupassen.

Der geplante AI Act ist folglich eine vorausschauende Reaktion der Europäischen Union auf die zunehmende Anzahl von Produktion und Dienstleistungen, welche mit Technologien der Künstlichen Intelligenz arbeiten und zukünftig auf dem europäischen Markt in Verkehr gebracht werden. Die Autor*innen der Normungsroadmap begrüßen dieses Vorhaben.

1.4.2 Gesetzgeberisches Umfeld

Der geplante AI Act ist eingebunden in eine Reihe weiterer Gesetze auf EU-Ebene, die ebenfalls bei der Entwicklung von KI-Systemen zu beachten sind. Diese adressieren Fragestellungen aus den Bereichen Zugang zu bzw. Verwendung von Daten und zugehörige Service-Strukturen (z. B. der geplante Data Act, Data Governance Act, Digital Services Act, Digital Markets Act), Fragenstellungen der Datensicherheit (z. B. Datenschutz-Grundverordnung, Cybersecurity Act oder der geplante Cyber Resilience Act) sowie allgemeinere Vorschriften wie z. B. bezüglich der Produkthaftung oder Arbeits- und Produktsicherheit (z. B. Produkthaftungsrichtlinie, Rahmenrichtlinie – Sicherheit und Gesundheitsschutz bei der Arbeit oder Maschinenrichtlinie). Dazu kommen sektorale Komponenten der Gesetzgebung, wie z. B. die Medizinprodukteverordnung in Bezug auf die Sicherheit von Produkten oder den European Health Data Space in Bezug auf den Zugang zu Daten in dem jeweiligen Anwendungsbereich. Übergeordnet ist zudem die EU-Grundrechtecharta als zentraler Baustein für die europäische Rechtsordnung zu nennen. Ein Großteil der genannten Gesetze ist bereits in Kraft, ein anderer Teil befindet sich aktuell im Gesetzgebungsprozess.

Abbildung 6 gibt einen Überblick über die Zusammenhänge zwischen dem geplanten AI Act und EU-Gesetzen, die mit ihm in Verbindung stehen. Im Fokus der Darstellung steht der geplante AI Act mit den Bereichen „Allgemeines & Produktsicherheit", „Datenschutz (Security und Privacy)" und „Daten und Services". Die verwendete Nummerierung verweist auf die detailliertere Beschreibung in Tabelle 16 (siehe Kapitel 13.1). Über die EU-Gesetze hinausgehende Vorschriften und Normen sind in der Übersicht angedeutet. Besondere Bedeutung kommt dabei den harmonisierten Europäischen Normen zu, die von den Europäischen Normungsorganisation auf Basis eines Auftrags der EU-Kommission zu Konkretisierung technischer Anforderungen des Rechtsaktes erarbeitet werden und eine zentrale Referenz für die Umsetzung der Anforderungen des geplanten AI Act darstellen.

Bereits bestehende, nicht im rechtlichen Sinne „harmonisierte" Europäische Normen leiten sich oftmals aus internationalen Normen ab, z. B. IEEE P7000™-Serie ([10], [11], [12], [13]) oder Normen des Subkommitees ISO/IEC JTC1/SC 42 [14], wie der Technische Bericht ISO/IEC TR 24368:2022 [15], der einen Überblick gibt über ethische oder soziale Bedenken bei der Verwendung von KI-Komponenten.

Bei den anderen Rechtsakten ergeben sich vielfältige Anforderungen an die Umsetzung und den Betrieb KI-basierter Systeme, die in Tabelle 16 (siehe Kapitel 13.1) näher erläutert sind. Beispielsweise enthält die Datenschutz-Grundverordnung Anforderungen bezüglich der Erhebung und Verarbeitung personenbezogener Daten und der damit verbundenen Einschränkungen oder des Rechts, nicht durch eine ausschließlich auf einer automatisierten Verarbeitung beruhenden Entscheidung beeinträchtigt zu werden (Art. 22 DSGVO) und somit die menschliche Autonomie zu erhalten. Es ergeben sich auch potenzielle Konfliktsituationen, da der geplante AI Act unter gewissen Umständen uneingeschränkten Zugang zu Trainings-, Validierungs- und Testdatensätzen erfordert (Art. 64 des AI Act). Insbesondere wenn hier noch personenbezogene Inhalte vorhanden sein sollten, könnten Sicherheitslücken entstehen.

6 Siehe Entwurf zum AI Act, Anhang I [4]

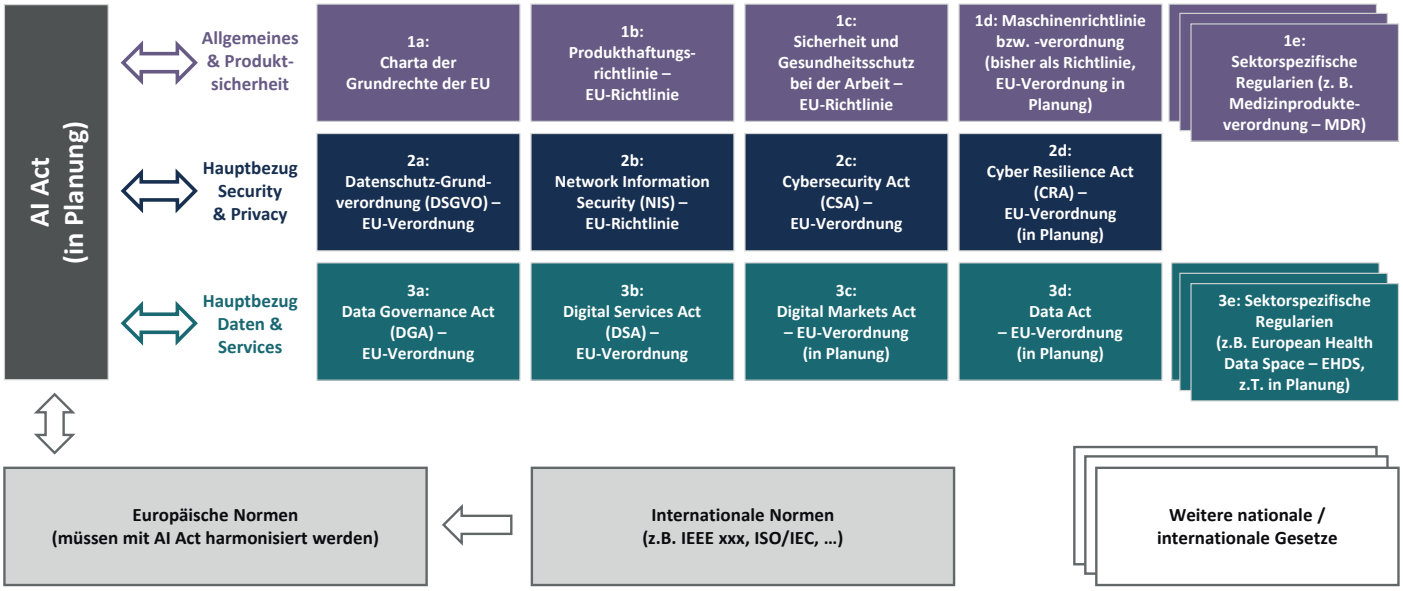


Abbildung 6: Überblick über EU-Gesetze mit verstärktem Bezug zum geplanten AI Act (Quelle: Martin Haimerl)[7]

Die aufgelisteten Gesetze enthalten aber auch Unterstützungsmaßnahmen, wie z. B. im Data Act bezüglich einer besseren Verwertung von Daten allgemein oder im European Health Data Space in Bezug auf eine passende Infrastruktur für den Zugang zu medizinischen Daten. Diese Ansätze stehen im Einklang mit den Bestrebungen der EU, durch den AI Act die Umsetzung von Innovationen im Bereich KI zu fördern. Insgesamt betrachtet wird auf EU-Ebene versucht, einen umfassenden Rahmen für die Harmonisierung und damit auch für Rechtssicherheit im Umgang mit KI-basierten Systemen zu schaffen.

1.4.3 Zusammenfassung: Ziele des geplanten AI Act

Zusammenfassend lassen sich folgende zentrale Aufgaben des AI Act identifizieren:

- → Verankerung europäischer Werte in KI-Systemen
- → Gewährleistung der EU-Grundrechte
- → Etablierung von nationalen und supranationalen Kontrollinstanzen
- → Definition von ethischer Anwendung der KI
- → Definition von Künstlicher Intelligenz und KI-Systemen
- → Einführung eines einheitlichen Frameworks, um Fragmentation zu verhindern
- → Konformitätsstandards durch eine verpflichtende CE-Kennzeichnung
- → Rechtssicherheit zur Förderung von Innovation und Investition in KI
- → Etablierung sicherer KI-Lebenszyklen
- → Regulierung von Hochrisiko-KI
- → Aufbau einer zentralen Datenbank für Hochrisiko-KI
- → Stärkung von Innovation im Bereich Künstliche Intelligenz

Der Verordnungsentwurf enthält im Weiteren noch Maßnahmen zur Innovationsförderung, wie etwa die Einrichtung von Reallaboren und „Sandboxes“ (Titel V), Vorgaben für die Leitungsstrukturen auf Unions- und nationaler Ebene, z. B. die Einrichtung eines Europäischen Ausschusses für Künstliche Intelligenz, die Einrichtung einer unionsweiten Datenbank für eigenständige Hochrisiko-KI-Systeme und die Einführung bestimmter Beobachtungs- und Meldepflichten für die Anbieter von KI-Systemen (Titel VI, VII und VIII). Titel IX enthält die Grundlagen zur Schaffung von Verhaltenskodizes, die Anbietern von KI-Systemen, die kein hohes Risiko darstellen, Anreize geben sollen, die zwingend vorgeschriebenen Anforderungen an Hochrisiko-KI-Systeme freiwillig anzuwenden.

1.4.4 Bedeutung harmonisierter Europäischer Normen für die Umsetzung des AI Act

Normen spielen in dem geplanten AI Act eine wichtige Rolle. Sie dienen der verlässlichen Umsetzung der Anforderungen des AI Act und helfen, die Entwicklung von KI-Systemen effizienter und verlässlicher zu machen. Eine besondere

7 Die Darstellung erhebt keinen Anspruch auf Vollständigkeit.

Bedeutung kommt dabei den harmonisierten Europäischen Normen (hEN) zu, insbesondere denjenigen, die zum geplanten AI Act im Amtsblatt der Europäischen Union gelistet sind. Wenn Inverkehrbringer von Hochrisiko-KI-Systemen diese hEN einhalten, wird davon ausgegangen, dass sie damit auch die korrespondierenden Anforderungen des Rechtsaktes, die von der Norm abgedeckt werden, einhalten. Diese sogenannte „Vermutungswirkung" erleichtert das Inverkehrbringen auf dem europäischen Binnenmarkt. Die Normanwendung bleibt freiwillig, das Inverkehrbringen von Hochrisiko-KI-Systemen ohne die Anwendung der hEN ist allerdings voraussichtlich mit einem erhöhten technischem Dokumentationsaufwand verbunden.

Eine harmonisierte Europäische Norm ist definiert als eine Norm, die von den Europäischen Normungsorganisationen CEN, CENELEC und/oder ETSI „auf Grundlage eines Auftrags der Kommission zur Durchführung von Harmonisierungsrechtsvorschriften der Union angenommen wurde".[8] Der Prozess der Erarbeitung einer hEN ist in Abbildung 7 dargestellt.

In ihrem jährlichen Arbeitsprogramm für die Europäische Normung kündigt die Europäische Kommission an, in welchen Bereichen sie im jeweiligen Jahr Normungsaufträge an die Europäischen Normungsorganisationen erteilen möchte.

Auf dieser Basis erstellt die Kommission einen Entwurf für einen Normungsauftrag, der mit Normungsorganisationen, Stakeholdern und Sektorexpert*innen sowie im Rahmen der dienststellenübergreifenden Konsultation beraten und ggf. angepasst wird. Der Normungsauftrag wird dann im Ausschuss für Normung (Committee on Standards), einem Mitgliedsstaatengremium[9], beraten und abgestimmt. Anschließend wird der Normungsauftrag an die Europäischen Normungsorganisationen übergeben, die nach Annahme des Auftrags die Erarbeitung der Norm in ihren Gremien umsetzen. Der Erarbeitungsprozess wird von einem durch die EU-Kommission finanzierten Harmonised Standards (HAS) Consultant begleitet, der/die der Kommission eine Bewertung vorlegt, ob die in der Norm enthaltenen Inhalte dem Normungsauftrag der Kommission entsprechen und die grundlegenden Anforderungen aus dem Harmonisierungsrechtsakt technisch abbilden. Die EU-Kommission entscheidet nach der Fertigstellung der Norm über ihre Listung im Amtsblatt der Europäischen Union. Erst mit der Listung der Norm im Amtsblatt wird diese zu einer harmonisierten Europäischen Norm und entfaltet die Vermutungswirkung. Ist ein Mitgliedsstaat oder das Europäische Parlament der Auffassung, dass eine gelistete hEN den Anforderungen, die sie abdecken soll, nicht voll entspricht, können diese einen formellen Einwand gegen diese hEN nach Art. 11 der Normungsverordnung erheben.

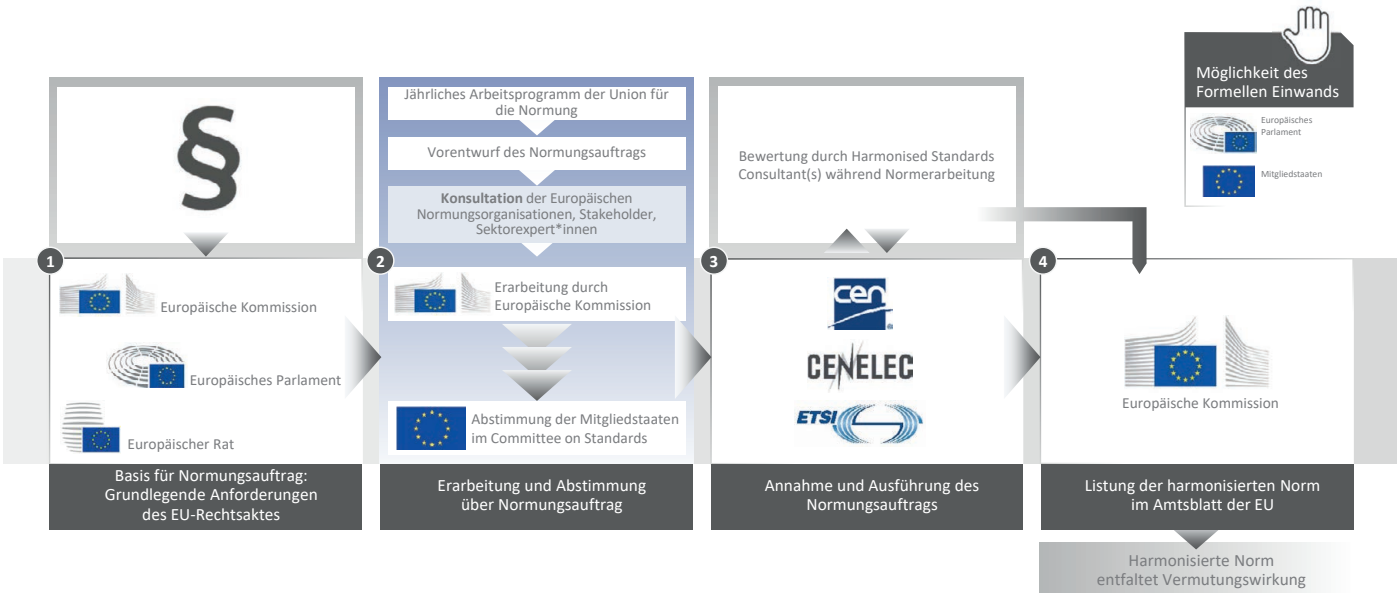


Abbildung 7: Prozess der Erstellung harmonisierter Europäischer Normen (Quelle: DIN)

8 Siehe EU-Normungsverordnung (1025/2012) Art. 2 Abs. 1c.

9 Nach EU-Normungsverordnung (1025/2012) Art. 22

Im Mai 2022 hat die EU-Kommission einen Entwurf für einen Normungsauftrag veröffentlicht, mit dem sie die Europäischen Normungsorganisationen damit beauftragen möchte, Normen zur technischen Ausgestaltung der grundlegenden Anforderungen aus Kapitel 2 des Kommissionsentwurfs für einen AI Act zu erarbeiten. Der Entwurf konzentriert sich insbesondere auf die folgenden Themenbereiche als Normungsanforderungen:

1. Risikomanagement
2. Daten und Daten-Governance
3. Management von Aufzeichnungen und eingebaute Logging-Mechanismen
4. Transparenz und Informationen für Benutzer*innen
5. Menschliche Aufsicht
6. Genauigkeitsspezifikationen für KI-Systeme
7. Robustheitsspezifikationen für KI-Systeme
8. Spezifikationen bezüglich der Cybersecurity für KI-Systeme
9. Qualitätsmanagementsysteme für Anbieter von KI-Systemen inklusive Monitoring-Prozesse nach der Inverkehrbringung
10. Konformitätsbewertung von KI-Systemen

Die geplante Verabschiedung des AI Act erfordert eine zeitnahe Erarbeitung von harmonisierten Europäischen Normen. Da die Konsensbildung im Normungsprozess Zeit in Anspruch nimmt, ist es wichtig, dass die Übergangsfristen im AI Act so großzügig gestaltet werden, dass zum Zeitpunkt der verpflichtenden Anwendung des AI Act alle relevanten Normen vorliegen. Mit Art. 41 (Common Specifications; gemeinsame Spezifikationen) des Kommissionsentwurfs, mit dem die Europäische Kommission ermächtig werden soll, mittels Durchführungsrechtsakten selbst technische Konkretisierungen für die grundlegenden Anforderungen festzulegen, zeigt der geplante AI Act eine zweite Möglichkeit auf. Diese ist allerdings kritisch zu sehen: Sie birgt die Gefahr, ein Parallelsystem zur Europäischen Normung mit inhaltlich konkurrierenden technischen Anforderungen zu schaffen, und zeugt nicht in vergleichbarer Weise wie das Europäische Normungssystem von Inklusivität und Transparenz. Common Specifications nach Art. 41 sollten daher nur die letztmögliche Fallback-Option darstellen.

Insofern sollte auch in Zukunft die Entwicklung harmonisierter Normen der bevorzugte Weg zur technischen Ausgestaltung grundlegender Anforderungen des geplanten AI Act sein. Die bereits bestehenden bzw. in Erarbeitung befindlichen internationalen Normen (IEEE 7000™-Serie ([10], [11], [12], [13]) und ISO/IEC im Rahmen des JTC 1/SC 42 [14]) können dafür ein guter Ausgangspunkt sein. Weitere Handlungsbedarfe müssen gezielt geklärt werden, insbesondere mit Blick auf die im geplanten AI Act gelisteten Anforderungen und speziell in den im Request-Entwurf für den Normungsauftrag angestrebten zehn Themenbereichen. Die vorliegende Normungsroadmap leistet dazu einen zentralen Beitrag.

1.4.5 Risikoklassifikation und Struktur des AI Act

Der geplante AI Act sieht eine Kategorisierung von KI-Systemen in vier Risikoklassen vor (siehe Abbildung 8):

- → KI-Systeme, die nicht in Verkehr gebracht werden dürfen: Art. 5 Abs. 1 lit. a)–d)
- → High-Risk-KI-Systeme: Art. 6 Abs. 1 i. V. m. Annex II, Abschnitt A, Nr. 1–12, Art. 6 Abs. 1 i. V. m. Annex II, Abschnitt B, Nr. 1–7 und Art. 6 Abs. 2 i. V. m. Annex III, Nr. 1–8
- → KI-Systeme mit besonderen Transparenzanforderungen: Art. 52 Abs. 1–3
- → Low-Risk-KI-Systeme: Alle KI-Systeme, die in keine der genannten Gruppen fallen

Primär erfolgt die Einteilung im AI Act nach Branchen bzw. Anwendungsbereichen (siehe Art. 5–7 sowie zugehörige Anhänge). Bei der Einordnung gemäß Art. 6 wird die Klassifizierung an die Anforderung gekoppelt, dass in den zu der jeweiligen Branche gehörigen und in Anhang II gelisteten sektoralen Harmonisierungsvorschriften (z. B. Maschinenrichtlinie, Medizinprodukteverordnung) die Durchführung eines Konformitätsverfahrens durch Dritte erforderlich ist. Insofern wird in diesem Fall die Klassifizierung indirekt durch die sektorale Harmonisierungsvorschrift mitbestimmt. In Art. 7 bzw. Anhang III ist zudem eine Reihe von Anwendungsfällen gelistet, die grundsätzlich in die Kategorie Hochrisiko eingeordnet werden.

Die Struktur des geplanten AI Act orientiert sich zentral an diesen Risikoklassen. Abbildung 9 gibt einen Überblick über die Struktur des geplanten AI Act mit den enthaltenen Titeln (I–XII) und den jeweils dazugehörigen Artikeln (1–85). Das Diagramm ordnet die einzelnen Titel dabei den jeweiligen Risikoklassen zu. Es zeigt an, welche Anforderungen für welche Risikoklasse zu beachten sind. Die Anforderungen der niedrigeren Risikoklassen übertragen sich dabei auf die höheren Klassen, d. h. sie sind auch dort anzuwenden.

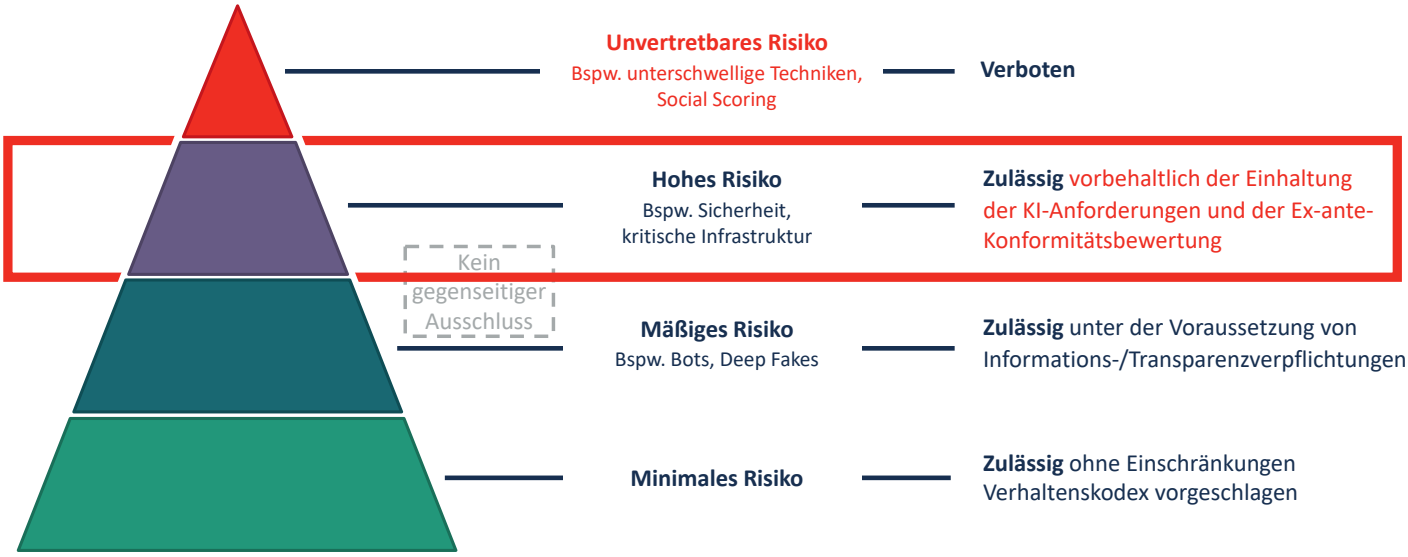


Abbildung 8: Risikoklassen des geplanten AI Act (Quelle: in Anlehnung an [4])

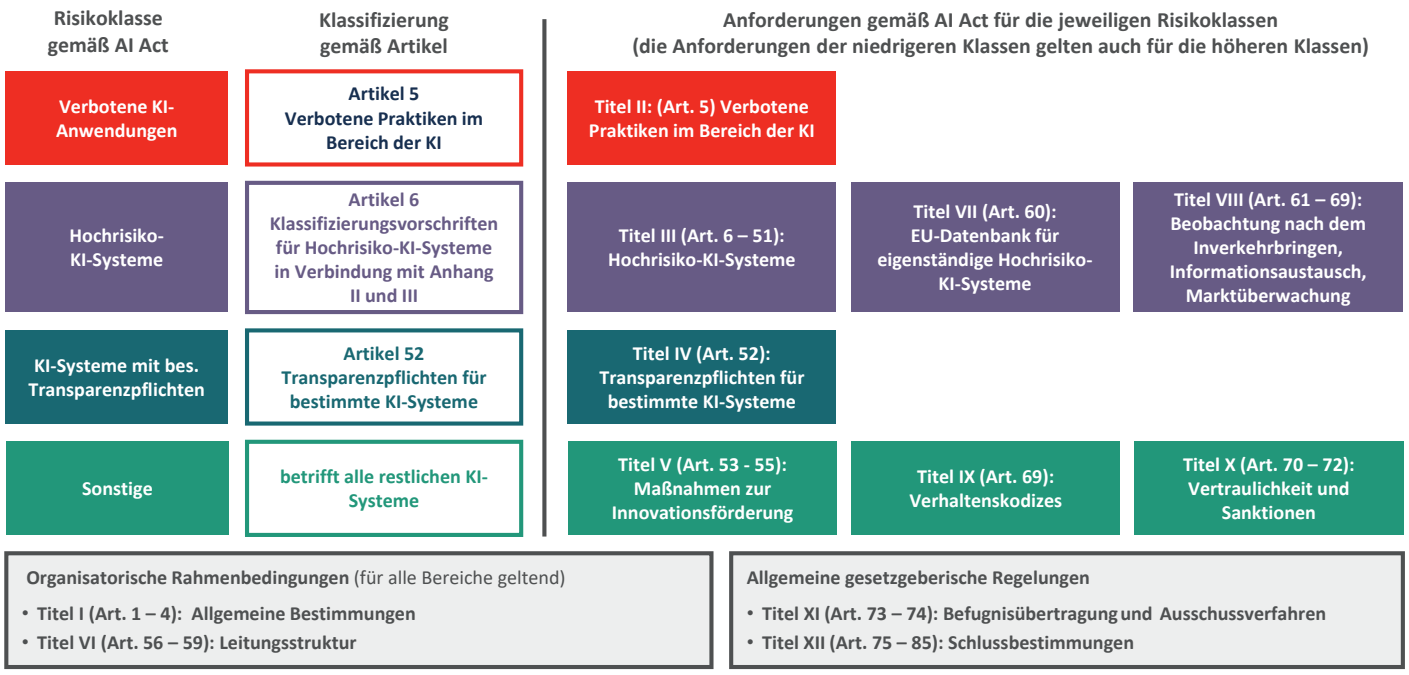


Abbildung 9: Überblick über die Inhalte des geplanten AI Act (Quelle: Martin Haimerl)

1.4.6 Konformitätsbewertung von KI-Systemen und -Produkten

Die Regulierung von KI nach dem geplanten AI Act teilt sich in die beiden Phasen vor sowie nach Markteinführung. Vor Ersterer wird der Zugang zum europäischen Binnenmarkt für Waren und Dienstleistungen anhand der Erfüllung der im Act gelisteten grundlegenden Anforderungen im Rahmen eines Konformitätsbewertungsverfahrens ermöglicht. Das betrifft insbesondere solche KI-Systeme, die in die Kategorie „Hochrisiko“ fallen. Dabei sind ggf. zusätzliche sektorale Harmonisierungsvorschriften der EU (z. B. Maschinenrichtlinie, Medizinprodukteverordnung) zu beachten, die im Konformitätsbewertungsverfahren berücksichtigt werden müssen. Die Einordnung in die Kategorie „Hochrisiko“ und die damit verbundene Anforderung, ein Konformitätsbewertungsverfahren durchzuführen (Art. 16e), gilt dabei gemäß Art. 6 sowohl für Sicherheitskomponenten von Produkten als auch für KI-Systeme, die eigenständige Produkte darstellen.

Im Rahmen des Konformitätsbewertungsverfahrens sollen Herstellende von Hochrisiko-KI-Anwendungen (gemäß der Definition des AI Act) sicherstellen bzw. nachweisen, dass ihre Produkte insbesondere die in Titel III, Kapitel 2 gelisteten Anforderungen an KI-Systeme erfüllen. Darunter fallen u. a. Anforderungen im Bereich Risikomanagement (Art. 9), Daten und Daten-Governance (Art. 10), Technische Dokumentation (Art. 11), Aufzeichnungspflichten (Art. 12), Transparenz (Art. 13), Menschliche Aufsicht (Art. 14) sowie Genauigkeit, Robustheit und Cybersicherheit (Art. 15). Zudem müssen die Anbieter von Hochrisikosystemen weitere Pflichten wie die Einrichtung eines Qualitätsmanagementsystems (Art. 17), Registrierungspflichten (Art. 51), Marktüberwachungsmaßnahmen (Art. 61) und die Meldung schwerwiegender Vorfälle und von Fehlfunktionen (Art. 62) umsetzen.

Darüber hinaus müssen Hochrisiko-KI-Systeme aus Anhang 3 für den Marktzugang in einer öffentlich zugänglichen europäischen Datenbank registriert werden (Art. 16 (f), Art. 51). Die Marktbeobachtung muss so ausgelegt sein, dass Daten zur Leistung der Hochrisiko-KI-Systeme über deren gesamte Lebensdauer hinweg aktiv und systematisch erfasst, dokumentiert und analysiert werden können. Das ist so auszulegen, dass die Marktbeobachtung im Verhältnis zur Art der KI-Technik und zu den Risiken des Hochrisiko-KI-Systems steht (Art. 61). Insgesamt müssen das Risikomanagementsystem und auch die menschliche Aufsicht so umgesetzt werden, dass sie das KI-System während des gesamten Lebenszyklus begleiten (Art. 9 und 14). Des Weiteren muss eine digitale Nachverfolgung der Funktionalität erfolgen (Art. 12).

Aus Sicht der technischen Regulierung wird die Funktionalität von KI-Systemen im Rahmen der Marktüberwachung durch nationale Aufsichtsbehörden der einzelnen europäischen Mitgliedsländer beaufsichtigt (Art. 59, 63). Für einzelne Branchen im Hochrisikobereich übernehmen die nach Rechtsakten zuständigen Behörden die Marktüberwachung. Für Finanzgeschäfte wäre beispielsweise die Finanzaufsicht zuständig. Im Rahmen der Marktüberwachung würden von den zuständigen Behörden der Zugang zu Daten von datengetriebenen Modellen und produkt-/systembeschreibenden Unterlagen sichergestellt (Art. 64) sowie die Verpflichtung zur Kontrolle von KI-Systemen auferlegt werden (Art. 63-67).

Für die Durchführung des Konformitätsbewertungsverfahrens gibt es gemäß Art. 43 zwei unterschiedliche Wege, die in Abbildung 10 dargestellt sind. Falls entsprechend harmonisierte Normen vorliegen, die die Erfüllung der grundlegenden Anforderungen des geplanten AI Act und auch der sektoralen Harmonisierungsvorschriften abdecken, kann der Herstellende diese heranziehen. Sollte das der Fall sein und der Herstellende die harmonisierten Normen vollständig angewendet haben, so kann er das Verfahren auf Basis einer internen Kontrolle unter Verwendung des in Anhang VI beschriebenen verkürzten Vorgehens durchführen. Nach Erfüllung der grundlegenden Anforderungen erhält das Hochrisiko-KI-System eine CE-Kennzeichnung (Art. 16 (i), Art. 19, Art. 49). Diese Optionen gelten für Hochrisikosysteme gemäß Art. 6, Abs. 2 bzw. Anhang III, d. h. für Systeme, bei denen die Klassifizierung als Hochrisikoprodukt nicht durch andere sektorale Harmonisierungsvorschriften (gemäß Art. 6, Abs. 2 bzw. Anhang II) gefordert ist. In dem letztgenannten Fall erfolgt die Umsetzung des Konformitätsverfahrens in Verbindung mit den jeweiligen sektoralen Vorschriften.

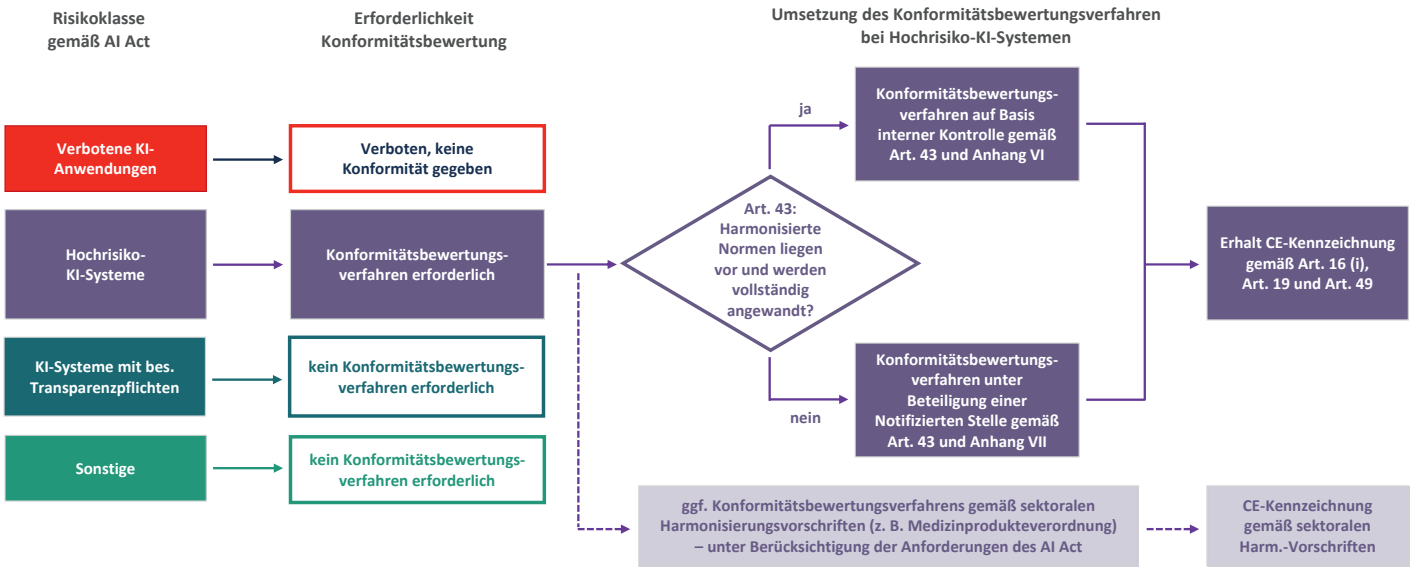


Abbildung 10: Varianten der Konformitätsbewertung gemäß Entwurf zum AI Act (Quelle: Martin Haimerl)

Sollten entweder keine harmonisierten Normen vorliegen oder diese nicht oder nur teilweise angewandt werden, so ist das Konformitätsbewertungsverfahren auf der Grundlage der Bewertung des Qualitätsmanagementsystems und der Bewertung der technischen Dokumentation unter Beteiligung einer Benannten Stelle durchzuführen. Dafür sind in Anhang VII die zugehörigen Anforderungen gelistet. In diesem Fall wird das Qualitätsmanagementsystem durch die Benannte Stelle geprüft (Anhang VII, Pos. 3.2) und bei positivem Bescheid auch im weiteren Verlauf kontrolliert (Anhang VII, Pos. 3.3/3.4). Zudem wird die technische Dokumentation des jeweiligen KI-Systems von der Benannten Stelle geprüft. Diese muss dabei uneingeschränkten Zugang zu den Test- und Trainingsdaten über eine geeignete API (Anhang VII, Pos. 4.3) und ggf. auch Zugang zum Quellcode erhalten (Anhang VII, Pos. 4.5). Zudem kann sie weitere Tests verlangen, um eine ordnungsgemäße Bewertung der Konformität durchführen zu können (Anhang VII, Pos. 4.4). Bei positiv verlaufener Prüfung stellt die Benannte Stelle eine EU-Bescheinigung über die Bewertung der technischen Dokumentation aus (Anhang VII, Pos. 4.6). Zudem bedarf jede Änderung des KI-Systems, die sich auf dessen Konformität mit den Anforderungen oder auf seine Zweckbestimmung auswirken könnte, der Genehmigung der Benannten Stelle, die die genannte Bescheinigung für das KI-System ausgestellt hat (Anhang VII, Pos. 4.7).

1.4.7 Zusammenfassung und Diskussion

Zusammenfassend weist der derzeit diskutierte Vorschlag für eine europäische KI-Regulierung aus Sicht der an der Normungsroadmap Beteiligten trotz seines Umfangs und der textlichen Komplexität einige zentrale Stärken auf: Durch den geplanten AI Act soll eine einheitliche, gemeinsame Position in Europa im Bereich KI entstehen, die Transparenz und Rechtssicherheit schafft. Dadurch können Markteinführung und -kontrolle im Sinne einer wertbasierten Regelung verbessert und im Idealfall auch beschleunigt werden. Der auf eine Risikobetrachtung ausgerichtete Ansatz zielt darauf, Gefahren zu minimieren und gleichzeitig Innovationen und Marktverbreitung zu fördern, insbesondere bei KI-Anwendungen mit niedrigem Risiko. Das Vorgehen bei der Zulassung bzw. Inverkehrbringung kann damit an das Risikopotenzial angepasst werden, wobei perspektivisch der Bedarf gesehen wird, noch stärker das Risiko konkreter Produkte zu betrachten. Der Fokus des geplanten AI Act auf die ethischen Werte und rechtlichen Grundlagen der Europäischen Union, die durch KI-Anwendungen auf Basis des Rechtsakts ebenfalls umgesetzt werden sollen, ist ein besonderes Alleinstellungsmerkmal von „KI made in Europe“ und kann idealerweise Vertrauen in die neuen KI-Technologien sicherstellen sowie deren Verbreitung fördern.

Offene Fragen ergeben sich zum Zeitpunkt des Redaktionsschlusses für diese Normungsroadmap u. a. noch aus der weit gefassten Definition von KI im Kommissionsentwurf. Es bedarf einer einheitlichen Beschreibung des Begriffsverständnisses im Rahmen des Rechtsaktes, damit sich aktuell bestehende Ungenauigkeiten im Umgang mit dem KI-Begriff nicht auf die Umsetzung von KI-basierten Produkten (z. B. in der geforderten technischen Dokumentation und auf die geplante Europäische KI-Datenbank) auswirken. Insgesamt bedürfen die Anforderungen zum Management von Hochrisiko-KI-Systemen noch einiger Klärungen und Überprüfungen. Zum Beispiel ergeben sich bei Produkten, die sowohl als KI-basierte Systeme den AI Act zu erfüllen haben, als auch sektorspezifische Harmonisierungsvorschriften, z. B. bei Medizinprodukten der Medical Device Regulation (MDR), genügen müssen, erhöhte Aufwendungen. Das wäre insbesondere der Fall, wenn Inkonsistenzen zwischen den Harmonisierungsvorschriften verblieben oder unterschiedliche Benannte Stellen herangezogen werden müssten, z. B. weil die bisher vom Unternehmen beauftragte Benannte Stelle nicht beide Harmonisierungsvorschriften abdecken kann (siehe Anhang 13.1, Abschnitt „Exemplarische Darstellung am Beispiel Medizinprodukte“ für eine ausführliche Darstellung der Problematik). Das würde insbesondere dann verstärkt wirken, wenn aufgrund zu kurzer Übergangsfristen zum Geltungsbeginn des AI Act noch nicht ausreichende Kapazitäten bei den Benannten Stellen aufgebaut werden konnten.

Ähnliches gilt für die Verfügbarkeit harmonisierter Normen. Durch sie können die Zulassungsprozesse schneller, zuverlässiger und einheitlicher umgesetzt werden. Dafür ist es zentral, dass die noch vorhandenen Unklarheiten im geplanten AI Act beseitigt und die für die Normung relevanten Prozesse rechtzeitig angestoßen werden. Den dort involvierten Expert*innen muss ausreichend Zeit zur Verfügung stehen, um die komplexen Anforderungen in praktikable Normen zu übersetzen. Die Möglichkeit, dass im Falle des Mangels verfügbarer harmonisierter Normen gemeinsame Spezifikationen (Common Specifications) an deren Stelle treten, wird im Vergleich zum Prinzip der Normung als nachteilig gesehen. Dass auf fachliche Expertise von Vertreterinnen und Vertretern aus der Industrie zurückgegriffen wird, trägt maßgeblich zur erfolgreichen Implementierung regulatorischer Vorgaben bei. Die Bestrebungen in dieser Normungsroadmap zeigen auf, welche Bedarfe umzusetzen sind, um die Grundlagen dafür zu schaffen.

1.5 Begriffsbestimmung KI

Im emergenten allgemeinen Fachbereich der Künstlichen Intelligenz ist es aufgrund einer Vielzahl unterschiedlicher Perspektiven und Akteurshintergründe schwierig, eine präzise Begriffsdefinition zu gewährleisten. Folgende Kernfragen kommen in Diskussionen über den Kern von „Künstlicher Intelligenz“ immer wieder auf:

a) Soll sich der Begriff auf einen wissenschaftlichen oder einen technischen Hintergrund beziehen?
b) Soll sich der Begriff auf eine Systemeigenschaft oder eine Systemfähigkeit beziehen?
c) Soll sich der Begriff auf eine Umschreibung der Funktion von KI-Systemen beschränken oder auf ihre Implementierung referenzieren?
d) Sollen Begriffe, die gewöhnlich mit menschlicher Intelligenz assoziiert werden (wie „Wissen“ oder „Fertigkeiten“), verwendet werden, um KI zu erklären?

Zur begrifflichen Präzisierung wird daher häufig zwischen „KI-Systemen“ und „KI“ unterschieden. Beinahe jede Organisation, welche sich mit Künstlicher Intelligenz befasst, definiert diese, mehr oder weniger stark, unterschiedlich.

Aus der Menge der verschiedenen Begriffsbestimmungen zur KI-Thematik sollen hier zwei zentrale Definitionen von „KI-System“ und „KI“ hervorhoben werden.

Auf politischer Ebene gesellschaftlicher Regelsetzung sei hier der Entwurf des AI Act der Europäischen Kommission [4] genannt.

„KI-System“ bezeichnet hier Software, welche unter Verwendung bestimmter Techniken oder Herangehensweisen Ergebnisse (beispielsweise Inhalte, Vorhersagen oder Entscheidungen) nach menschlichen Zielsetzungen generiert, welche den Kontext der KI-Systeme selbst wiederum beeinflussen.

Die spezifischen Techniken und Herangehensweisen sind folgende:

a) Ansätze zum Maschinellen Lernen inklusive überwachtes Maschinelles Lernen, unüberwachtes Maschinelles Lernen, bestärkendes Lernen und eine Vielzahl an Methoden inklusive Deep Learning
b) Logik- und wissensbasierte Ansätze einschließlich Knowledge Representation, Inductive (Logik) Programming, Knowledge Bases, Inference Engines und Deductive Engines, (Symbolic) Reasoning und Expert Systems
c) Statistische Modelle, Bayes'sche Schätzung sowie Such- und Optimisierungsmethoden

Als „KI" wird im geplanten AI Act allgemein das Feld der sich schnell entwickelnden Technologien der KI-Systeme bezeichnet.

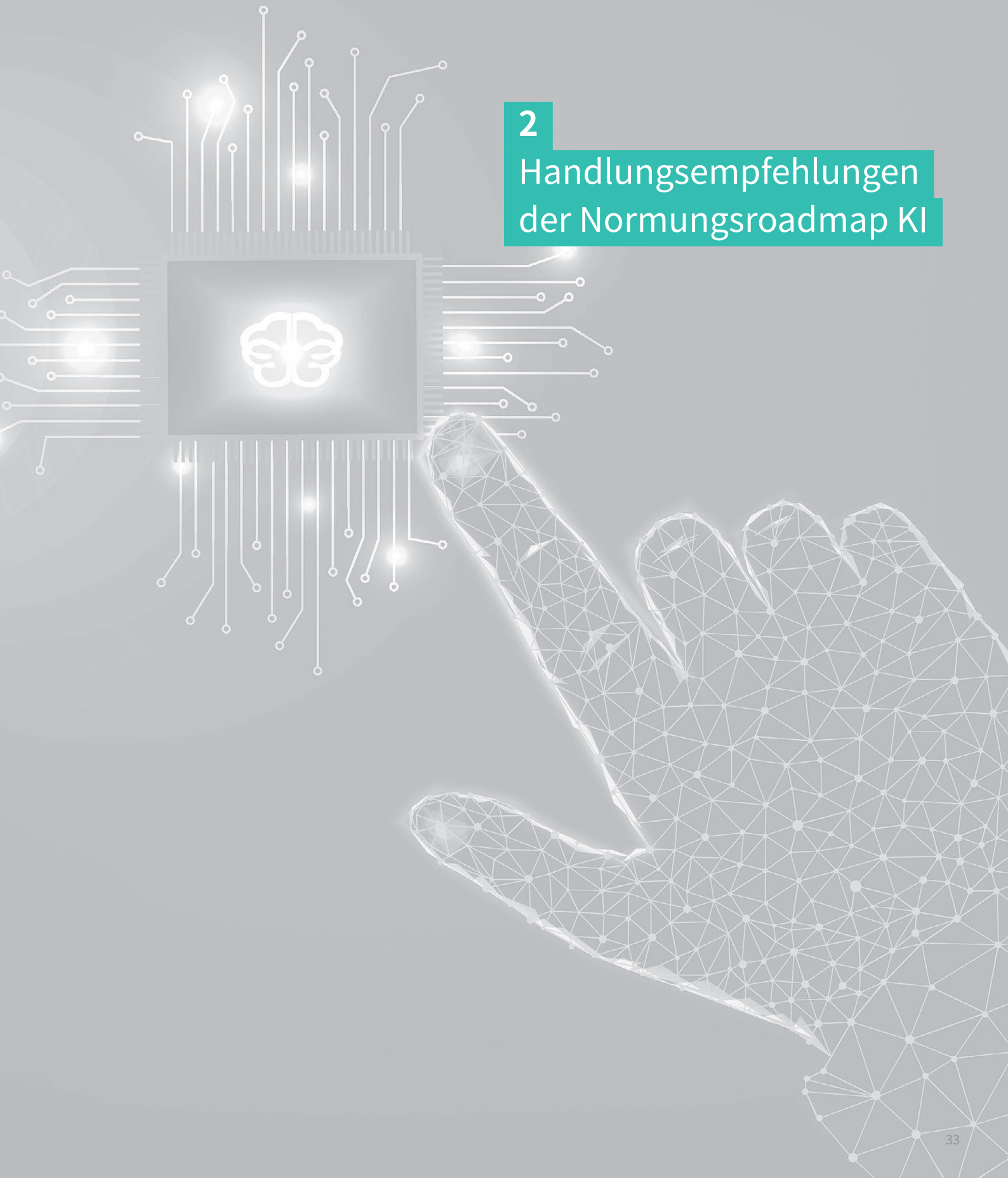
Auf der internationalen Ebene technischer Regelsetzung existiert währenddessen der internationale Standard zu Konzepten und Terminologien in KI (ISO/IEC 22989:2022 [16]).

„KI-Systeme" bezeichnet hier ein konstruiertes System, welches Ergebnisse (beispielsweise Inhalte, Vorhersagen, Empfehlungen und Entscheidungen) nach menschlichen Zielsetzungen generiert. Es werden vier Kerneigenschaften von KI-Systemen festgestellt:

a) Interaktivität: Informationsregistrierung durch Sensoren oder menschliche Eingabe
b) Kontextsensibilität: Manche KI-Systeme reagieren auf mehrere Informationsquellen
c) (menschliche) Überwachung: KI-Systeme können unter variierender Intensität menschlicher Überwachung handeln
d) Anpassungsfähigkeit: Manche KI-Systeme sind so angelegt, dass sie dynamisch auf (Echtzeit-)Daten reagieren und ihre Handlung auf Basis dieser neuen Informationen neu interpretieren und anpassen

„KI" wird hier als die Disziplin der Erforschung und Entwicklung von Mechanismen und Anwendungen von KI-Systemen bezeichnet.

Die vorliegende Normungsroadmap KI bezieht sich für ihre Begriffsbestimmung von KI und KI-Systemen auf die internationale Norm ISO/IEC 22989:2022 [16].



2 Handlungsempfehlungen der Normungsroadmap KI

Ziel der Normungsroadmap KI ist es, einen Handlungsrahmen zu beschreiben, der die deutsche Wirtschaft und Wissenschaft im internationalen Wettbewerb um die besten Lösungen und Produkte im Bereich der Künstlichen Intelligenz stärkt und innovationsfreundliche Rahmenbedingungen schafft. Damit leistet sie einen wesentlichen Beitrag, um „KI – Made in Germany“ als starke Marke zu etablieren und neue Geschäftsmodelle, disruptive Innovationen und skalierbare Anwendungen zu entwickeln. Insbesondere der deutsche Mittelstand und die wachsende Start-up-Szene in Deutschland können davon profitieren. Normen und Standards bilden die Grundlage für technische Souveränität und schaffen einen Rahmen, der Transparenz fördert und Orientierung bietet. Somit sorgen sie für Sicherheit, Qualität und Zuverlässigkeit und tragen maßgeblich zur Erklärbarkeit von KI-Lösungen bei – eine wesentliche Grundlage, wenn es um die Akzeptanz von KI-Anwendungen geht. Die Normungsroadmap KI bietet großes Potenzial, um sowohl die Wettbewerbsfähigkeit Deutschlands zu sichern als auch europäische Wertmaßstäbe auf die internationale Ebene zu heben. Nicht zuletzt deshalb sollte ein besonderes Augenmerk auf die Umsetzung der Normungsroadmap KI und ihre Handlungsempfehlungen gelegt werden.

Empfehlung 1: Entwicklung, Validierung und Standardisierung eines horizontalen Konformitätsbewertungs- und Zertifizierungsprogramms für vertrauenswürdige KI-Systeme

Der aktuelle Vorschlag der EU-Kommission für einen europäischen Rechtsrahmen (AI Act) erfordert ein anwendungsagnostisches, marktfähiges Konformitätsbewertungs- und Zertifizierungsprogramm, das die Anforderungen der Wirtschaft, der Behörden und der Zivilgesellschaft an KI-Systeme objektiv überprüfbar macht.

Das Fehlen eines solchen Konformitätsbewertungs- und Zertifizierungsprogramms gefährdet das wirtschaftliche Wachstum und die Wettbewerbsfähigkeit der Zukunftstechnologie KI. So sind Aussagen über die Vertrauenswürdigkeit von KI-Systemen ohne hochwertige Prüfmethoden nicht belastbar, wodurch die Akzeptanz von KI-Systemen in Wirtschaft und Gesellschaft unklar bleibt. Eine erfolgreiche Nutzung von KI-Systemen, die den Anforderungen des europäischen Rechtsaktes und damit den europäischen Wertevorstellungen genügt, erfordert Transparenz in der gesamten Lieferkette in verteilten und hybriden KI-Systemen durch sachkundige, verlässliche und reproduzierbare Prüfungen der KI-Technologien.

Die Normungsroadmap KI empfiehlt daher die Entwicklung, Validierung und Standardisierung eines KI-Konformitäts- und Zertifizierungsprogramms mit höchster Priorität. Im Rahmen der Fortschreibung der Normungsroadmap KI sind bereits erhebliche Vorarbeiten geleistet worden, sodass mit der Umsetzung dieser Handlungsempfehlung sofort begonnen werden kann.

Die Grundlagen und die Architektur des Zertifizierungsprogramms sollten von den Standardisierungsgremien definiert und harmonisiert werden. Die Harmonisierung betrifft die Geltungsbereiche, die bedarfsgerechten Prüfkriterien, die Anforderungen und Nachweise und die Prüfverfahren für die Zertifizierung von KI-Produkten, KI-Systemen und von KI-Managementsystemen. Damit wäre Deutschland in der Lage, einen federführenden Beitrag zur Entwicklung und Standardisierung eines international anerkannten KI-Zertifizierungsverfahrens zu leisten.

Zur Umsetzung der Empfehlung ist die Förderung und Bereitstellung eines Budgets durch die Bundesregierung für folgende Vorhaben essenziell:

→ Entwicklung und Fortschreibung eines internationalen akkreditierungsfähigen KI-Zertifizierungsverfahrens, das sich in die bestehende Zertifizierungsinfrastruktur für Produkte, Dienste, Prozesse und Organisationen einpasst.[10] Dabei stehen in der horizontalen, anwendungsagnostischen Standardisierung bereits initiierte deutsche Projekte zur KI-Zertifizierung von Produkten, hybriden Systemen, Services und ganzen Supply Chains und zusätzlich Managementsysteme für Organisationen im Fokus.

→ Initiierung und Durchführung von Forschungsprojekten insbesondere im Bereich von Prüfungen mit hoher Prüfqualität und zur Reduzierung von unvertretbar großem Prüfaufwand. Dazu gehören u. a. die folgenden Forschungsfelder:
  - die Unsicherheit bei neuronalen Netzen,
  - die Erklärbarkeit und Transparenz,
  - die Entwicklung und Zertifizierung von Prüfwerkzeugen für alle Prüfdimensionen,
  - die Komposition von Prüfergebnissen.

10 DIN EN ISO/IEC 17065:2013 [17] (i. V. m. DIN EN ISO/IEC 17067:2013 [18] und den entsprechenden Spezifikationen ISO/IEC TR 17026:2015 [19] (reine Produkte); ISO/IEC TR 17028:2017 [20] (services); ISO/IEC TR 17032:2019 [21] (Processes)), DIN EN ISO/IEC 17021-1:2015 [22] (Organisationen/Managementsysteme)

→ Überführung der Ergebnisse in die Normung und Erarbeitung von Normen und Standards: Um die Ergebnisse aus den o. g. Projekten national zu koordinieren und zeitnah auf der europäischen und internationalen Normungsebene einzubringen, ist die Finanzierung entsprechender Normungsprojekte zwingend erforderlich. Insbesondere sind hierfür Fachleute zu gewinnen und Ressourcen für deren Mitwirkung in den Normungsgremien bereitzustellen.
→ Im Sinne der Entwicklung einer KI-Zertifizierung und einer entsprechenden Infrastruktur, die sich in die Prüfung und Qualitätssicherung informationstechnischer Systeme insgesamt einfügt, wird vorgeschlagen, die Leitung des o. g. Programms gemeinsam den nationalen Normungsorganisationen sowie dem Bundesamt für Sicherheit in der Informationstechnik zu übertragen.

Empfehlung 2: Aufbau von Dateninfrastrukturen und Erarbeitung von Datenqualitätsstandards zur Entwicklung und Validierung von KI-Systemen

Daten spielen für die Realisierung vieler KI-Systeme eine zentrale Rolle und die Qualität der KI-Systeme hängt oftmals entscheidend von der Datenqualität ab. Hierbei werden große Datenmengen sowohl zum Training dieser Systeme als auch zur Validierung (systematisches Testen) benötigt. Ein prominentes Beispiel ist die Entwicklung großer Sprachmodelle wie etwa Open-GPT-3 oder DALL-E 2, welche Datensätze mit einigen 100 Millionen Trainingsdaten benötigen. Neben dem Training von entsprechenden Systemen werden Daten auch für das systematische Testen von KI-Systemen benötigt. Insbesondere für die Validierung von KI-Systemen, welche in einem Open-World-Kontext operieren, werden dabei viele Testszenarien benötigt. Die Verfügbarkeit entsprechender Datensätze ist somit auch ein strategischer Erfolgs- und Wettbewerbsfaktor für die deutsche KI-Industrie und insbesondere Start-ups. Hierfür werden geeignete Dateninfrastrukturen benötigt, welche geeignete Datensätze sammeln, kuratieren, durch geeignete Metadaten beschreiben und zur Verfügung stellen. Bei der Generierung solcher Datensätze kommt dabei vor allem auch synthetischen Daten eine entscheidende Rolle zu, da für manche KI-Anwendungen nicht genügend reale Daten zur Verfügung stehen bzw. manche Testszenarien zu selten auftreten, als dass die Verfügbarkeit realer Daten für eine angemessene Validierung ausreichen würde. Je nach Einsatzzweck sind hierbei sowohl die Bereitstellung von Open-Source-Datensätzen denkbar als auch Marktplätze, welche den Handel mit entsprechenden Daten ermöglichen. Für die Zulassung von KI-Systemen mit kritischem Einsatzkontext – etwa im Medizinsektor – können zudem behördliche Dateninfrastrukturen benötigt werden, welche Datensätze zur Zulassung dieser KI-Anwendungen zur Verfügung stellen.

Die Realisierung solcher Dateninfrastrukturen sollte auf aktuelle Datenarchitekturen wie etwa Data Meshs setzen, Techniken der Datenvirtualisierung nutzen und wo möglich auf bestehende Strukturen wie etwa Gaia-X oder den European Health Data Space aufbauen. Gleichzeitig müssen entsprechende Werkzeuge entwickelt werden, welche Datensätze auf ihre Qualität hin überprüfen und Teilmengen der Daten identifizieren, auf denen die entsprechenden KI-Systeme weniger performant sind und mit denen gezielt hochwertige synthetische Daten erzeugt werden können.

Normen und Standards kommt bei der Bereitstellung solcher Datensätze und ihrer Dateninfrastrukturen eine besondere Bedeutung zu, um die Interoperabilität sicherzustellen und gleichzeitig Qualitätsstandards zu definieren. Durch entsprechende Datenqualitätsstandards kann sichergestellt werden, dass Datensätze etwa repräsentativ, vollständig, fehlerfrei und ausgewogen sind.

Die Normungsroadmap KI empfiehlt daher die Förderung solcher Dateninfrastrukturen durch die öffentliche Hand und gleichzeitig die Unterstützung der Normungsorganisationen bei der Entwicklung entsprechender Datenqualitätsstandards. Da die erfolgreiche Entwicklung von Normen und Standards maßgeblich von der Beteiligung relevanter Fachleute abhängt, ist die Bereitstellung notwendiger Ressourcen zur Mitwirkung in den Normungsgremien durch die Bundesregierung sicherzustellen.

Der Anteil der Aktivitäten, welcher auf die Validierung von KI-Systemen abhebt, sollte eng mit dem Programm zur Zertifizierung und Konformitätsbewertung (vgl. Handlungsempfehlung 1) verzahnt werden.

Empfehlung 3: Den Menschen als Teil des Systems begreifen, und zwar in allen Phasen des KI-Lebenszyklus

Im aktuellen Entwurf der Europäischen Verordnung zu Künstlicher Intelligenz (AI Act) werden insbesondere an Hochrisikosysteme umfangreiche Forderungen zur Einbindung von Menschen gestellt, z. B. Transparenz für Betroffene und Beteiligte, menschliche Aufsicht in unterschiedlichen Rollen und Eingriffsmöglichkeiten bis hin zu einer „Stopptaste“, die von Menschen ausgelöst wird. Welche Transparenz in welchem Kontext für welche Zielgruppe ausreichend ist, wie die menschliche Aufsicht umgesetzt werden sollte und welche Basisinformationen als Grundlage für menschliche Eingriffe

ins System vorhanden sein müssen – all das sind Fragestellungen, die vom Menschen aus zu denken und wonach die technischen und sozialen Komponenten zu entwickeln und auszurichten sind.

Um diese soziotechnischen Aspekte in KI-Systemen umzusetzen, sind folgende Herausforderungen zu bewältigen:

→ Angemessenheit: Technische Komponenten sind auf Grundlage der soziotechnischen Anforderungen auszuwählen.
→ Partizipation: Die Definition und Auswahl von relevanten Akteur*innen, die beteiligt werden sollten, sind zu operationalisieren.
→ Ethik: Gesellschaftliche und ethische Fragestellungen mithilfe etablierter Modelle sind zu operationalisieren, messbar bereits bei der Entwicklung der Technologie zu verankern und dabei auf dem Stand der Forschung zu Diskriminierungssensibilität aufzubauen.
→ Kultur: Eine adäquate Organisationskultur ist zu etablieren (im Arbeitskontext z. B. die Unternehmenskultur), denn auch diese muss beim KI-Einsatz mitentwickelt werden. Dafür sind relevante Akteur*innen zu sensibilisieren, zu qualifizieren und in einem geeigneten Change Management im Prozess mitzunehmen.
→ Tools: Über den Lebenszyklus eines KI-Systems hinweg ist der Mensch mit Prozessen, Methoden und Tools zu unterstützen – von der Zielsetzung über die Entwicklung bis zum Betrieb mit Iterationen und Re-Validierung.

Zu einigen der genannten Herausforderungen liegen im Rahmen von Forschungsvorhaben bereits gewonnene Erkenntnisse vor. Diese gilt es nun, den sich möglicherweise noch ändernden Anforderungen des geplanten AI Act anzupassen und weiter zu schärfen.

Daraus ergeben sich folgende konkrete Handlungsempfehlungen:

→ An Förderträger: In Leuchtturmprojekten sollte konkret erprobt werden, wie eine Einbindung von betroffenen und beteiligten Menschen in allen Phasen des KI-Lebenszyklus in unterschiedlichen Kontexten gelingen kann.
→ An die Normungsgremien: Begleitend dazu müssen zeitnah die notwendigen Normen für die soziotechnischen Aspekte des geplanten AI Act erarbeitet werden, insbesondere zur menschlichen Aufsicht und zu den notwendigen Transparenzanforderungen.
→ An die Normungsorganisationen sowie die Politik: Um der weitreichenden gesellschaftlichen Verantwortung gerecht zu werden, ist es erforderlich, dabei in den Normungsorganisationen besonders auf eine ausgewogene Beteiligung aller relevanten Zielgruppen zu achten und diese aktiv zu forcieren (z. B. Wissenschaft oder Zivilgesellschaft).
→ An die Politik: Für die in der Normung bislang unterrepräsentierte soziotechnische Perspektive ist es zudem zwingend erforderlich, Fachleute zu gewinnen und deren Kapazitäten in Normungsgremien zur Verfügung zu stellen, um die Erkenntnisse auf nationaler Ebene produktiv zu bündeln und auf der europäischen und internationalen Ebene einzubringen.

Empfehlung 4: Entwicklung von Vorgaben für die Konformitätsbewertung von kontinuierlich oder stufenweise lernenden Systemen im Bereich der Medizin

KI-Systeme können mit mehr Daten und Informationen kontinuierlich verbessert werden. Daraus ergibt sich für KI-Systeme, die bereits in der Praxis eingesetzt werden, ein erhebliches Verbesserungspotenzial, da z. B. neue Trainingsdaten sowie Informationen über Fehlverhalten und Korrekturen gewonnen werden können. Auf der anderen Seite muss die Integration der neuen Daten auf einem hohen Qualitätsniveau umgesetzt und durch entsprechende Prüfprozesse untermauert werden, um den hohen Sicherheitsansprüchen beispielsweise im Medizinbereich zu genügen.

In diesem Kontext gibt es verschiedene Herausforderungen mit zwei grundlegend unterschiedlichen Ansätzen:

Zum einen können im Sinne eines kontinuierlich lernenden Systems Daten vor Ort gesammelt und in das Modell/System eingepflegt werden, um es zu verbessern und/oder an lokale Regionen, individuelle Krankenhäuser oder einzelne Patienten(-gruppen) gezielt anzupassen. Zum anderen können die Updates im Rahmen eines vereinfachten Rezertifizierungsprozesses bzw. Konformitätsbewertungsverfahrens stufenweise erfolgen.

Im Einzelnen sind dabei Vorgaben aus der Politik und der Normung für die folgenden Aspekte zu entwickeln.

→ Es lässt sich im Vorfeld nicht bestimmen, welche und entsprechend wie viele Daten hierfür notwendig sind. Es können theoretisch alle Daten von Bedeutung sein, die in das Online-Learning geflossen sind. Diese Daten können in der Regel nicht ohne Weiteres zur Verfügung gestellt werden. Grundsätzlich ist daher zu klären, wie die Verwaltung und Weitergabe der Daten sowie die Anwendung des KI-Systems gestaltet werden müssen, um das sich verändernde KI-System nachvollziehbar und damit auch auditierbar zu machen.

→ Es wird ein Validierungsprozess benötigt, der anhand von konkreten, medizinspezifischen Schutzzielen zu definieren ist und ein Modell-Update umfassend und zuverlässig überprüfen kann. Das beinhaltet die Definitionen und die Prüfung von Anforderungen sowohl an den Validierungsprozess als auch an das KI-System (insbesondere des aktualisierten Systems). Das kann zur Folge haben, dass der Validierungsprozess selbst einer Validierung unterzogen werden muss.
→ Es wird ein „agiler Freigabe-/Konformitätsbewertungsprozess“ gebraucht, der die gemäß der Medical Device Regulation (MDR) notwendige klinische Validierung von in Teilen durch online-akquirierte Daten verbesserte KI-Systeme so umsetzt, ohne dass jedes Mal eine Rezertifizierung bzw. ein erneutes Konformitätsbewertungsverfahren des Gesamtsystems vorgenommen werden muss.

Für die Freigabe bzw. (Re-)Konformitätsbewertung solcher Systeme fehlen insbesondere in Europa anerkannte Prüfanforderungen und Prüfverfahren. Diese Prüfanforderungen und Prüfverfahren können letztlich nur erfolgreich auf den Markt gebracht werden, wenn die Akteur*innen aus Politik, Normung, Forschung und Industrie das unterstützen. Um eine Zertifizierung bzw. einen Marktzugang erreichen zu können, müssen Randbedingungen vorab spezifiziert werden, die eine automatisierte Freigabe von kontinuierlich oder stufenweise lernenden Systemen zulassen.

Daraus ergeben sich folgende konkrete Handlungsempfehlungen:
→ Es ist ein Leuchtturmprojekt zu initiieren und durchzuführen, das durch die öffentliche Hand gefördert wird (z. B. BMBF, BMWK, BMG und weitere) und verschiedene Domänen bzw. Aspekte betrachtet:
  - Analysen in der medizinischen Bildgebung,
  - Onkologie/Krebserkennung,
  - automatisierte intensivmedizinische Versorgung,
  - Identifikation und Therapie von Sepsis.
→ Jedes dieser domänenspezifischen Teilprojekte erfordert eine Zusammenarbeit zwischen universitären und außeruniversitären Forschungseinrichtungen, Krankenhäusern im Realbetrieb, Herstellenden von medizintechnischen Geräten, Tech-/IT-Firmen, TIC-Unternehmen (TIC: Testing, Inspection, Certification) sowie Normungsorganisationen und ist mit einem entsprechenden Budget und einer projektübergreifenden Governance-Struktur auszustatten, welche für die Koordination der Inhalte Sorge trägt.
→ Zusätzlich zu einem wissenschaftlichen Beirat wird die Einrichtung einer Geschäftsstelle empfohlen, welche die Übersetzung der Projektergebnisse in Normen, Standards und allgemein praktizierte Prüfverfahren und deren industrieübergreifende Verwertung und internationale Platzierung übernimmt. Ein solches Projekt sollte schnellstmöglich auf den Weg gebracht werden.

Empfehlung 5: Entwicklung und Einsatz sicherer und vertrauenswürdiger KI-Anwendungen in der Mobilität durch Best Practices und Absicherung

Der Einsatz von KI-Technologien im Kontext der Mobilität ist gekennzeichnet durch komplexe Randbedingungen. Diese sind ausgezeichnet durch komplexe Entscheidungs- und Kontrollsysteme, welche in einer sensomotorischen Schleife in einer sich stetig ändernden Umwelt sowohl mit dieser selbst als auch mit einer Fülle weiterer Akteur*innen interagieren – in Kombination mit den hohen Risiken von Fehlfunktionen für Mensch und Umwelt. Normen und Standards für die dynamische Typzulassung von Mobilitätssystemen, deren Funktionalität zumindest teilweise auf dem Einsatz von KI-Technologie basiert, sind daher dringend erforderlich, um einerseits unter diesen komplexen Randbedingungen dauerhaft eine hinreichende Performanz und andererseits die erforderliche Vertrauenswürdigkeit und Sicherheit zu ermöglichen bzw. zu garantieren. Während die verschiedenen Aspekte von Vertrauenswürdigkeit (Trustworthiness) bereits durch den Entwurf des AI Act weitgehend vorgegeben sind, bedarf es zur Operationalisierung deren Konkretisierung über den gesamten Lebenszyklus eines KI-Systems hinweg. Insbesondere sollten diese Normen und Standards dafür sorgen, dass …
→ die effiziente (Weiter-)Entwicklung, Validierung, sukzessive Einführung und kontinuierliche Absicherung im Betrieb durch einen Best-Practice-Katalog unterstützt werden, der die Leistungsfähigkeit und Vertrauenswürdigkeit der Systeme gewährleistet. Diese Maßnahmen sollten u. a. qualifizierte Verfahren und Werkzeuge für die Entwicklung und den Test sowie erklärbare KI-Verfahren umfassen, deren relevante Eigenschaften zum Nachweis der Sicherheit und Trustworthiness analysiert und getestet werden können.
→ die Nachweise und die Validierungen von Vertrauenswürdigkeit und Sicherheit gegenüber einem unabhängigen Dritten ermöglicht werden. Hierbei sind Normen und Standards für Mindestanforderungen (insbesondere nicht tolerierbare Restrisiken an die Safety) sowie die weiteren wesentlichen Vertrauensaspekte (u. a. IT-Sicherheit, Robustheit, Transparenz, Nachvollziehbarkeit, Daten-

schutz und Nicht-Diskriminierung) über den gesamten Lebenszyklus eines KI-Systems zu definieren und zu berücksichtigen, wobei die erforderliche Absicherung einerseits abhängig von der Risikoklasse der konkreten Mobilitätsanwendung und andererseits vom jeweiligen „gesellschaftlich akzeptierten Restrisiko“ zu gestalten ist.

Um die sukzessive Einführung von sicheren und vertrauenswürdigen KI-basierten Mobilitätsanwendungen trotz der verbleibenden und aufkommenden Unsicherheiten zu ermöglichen, benötigt es agile Ansätze zur Regulierung (vgl. [23]) und Standardisierung, die eine kontinuierliche Überwachung und Anpassung der Wirksamkeit der regulatorischen Stellhebel ermöglichen. Dies erfordert eine Überwachung operationeller Risikofaktoren und setzt bestimmte gesellschaftliche Erwartungshaltungen voraus sowie eine enge Verzahnung der Standardisierung und Regulierung.

Empfehlung 6: Entwicklung übergreifender Datenstandards und dynamischer Modellierungsverfahren zur effizienten und nachhaltigen Gestaltung von KI-Systemen

KI-Systeme finden verstärkt Anwendung in der Bearbeitung gegenwartsrelevanter Fragestellungen. Dies betrifft sektorübergreifend die intelligente Steuerung von Systemen und die Formulierung von Handlungsempfehlungen. Hierbei existiert ein hoher Grad der Interdisziplinarität, indem bislang trennscharfe Datendomänen zusammengeführt und statische Modellierungsstandards flexibilisiert werden. Die zielorientierte Gestaltung neuer Datenstandards und Modellierungsverfahren stützt sich sowohl auf Normen zur Interpretation und Aggregation von Daten als auch auf Verfahrensnormen. Normative Definitionen stellen branchenübergreifende Leitplanken für Wirtschaft und Wissenschaft dar, die Antworten auf folgende Fragestellungen geben können:

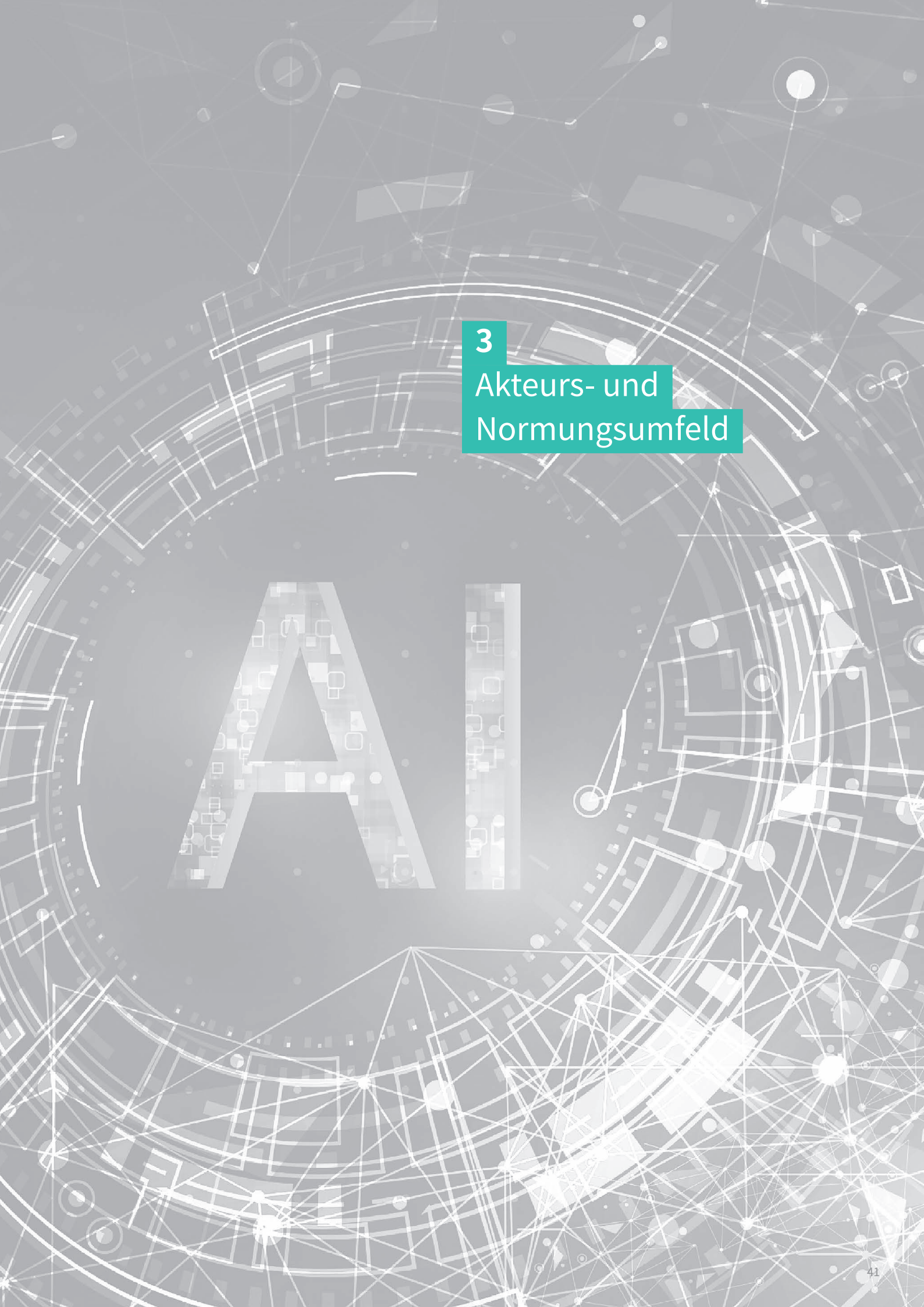
→ Wie sind Datensyntax und -semantik und darauf aufbauende KI-Systeme zu gestalten, damit (teil-)autonom agierende Systeme effizient und resilient betrieben werden können?
→ Wie können gemeinsame Ordnungsrahmen bzw. Frameworks für Datenquellen aus verschiedenen Sektoren gestaltet werden, um die kontinuierliche Datenakquisition und -kommunikation für und in KI sowie ML zu vereinfachen?
→ Wie ist die Effizienz- und Nachhaltigkeitsgüte von KI zu operationalisieren und zu evaluieren?
→ Inwiefern kann der fortlaufenden, agilen Entwicklung neuer domänenspezifischer Datenstandards in der Normung übergreifender Frameworks Rechnung getragen werden?
→ Wie kann in diesem Kontext sichergestellt werden, dass Daten mit höherer zeitlicher und geografischer Auflösung hinreichende Berücksichtigung finden?

Mithin besteht ein umfassender Bedarf an Normung und Standardisierung für die Überwindung von Datensystemgrenzen und die Entwicklung von Referenzverfahren. Diese Normungsbedarfe können durch das gemeinsame Agieren von Akteur*innen aus Normung, Wirtschaft und Wissenschaft bedient werden. Dafür werden Pilotprojekte im öffentlichen Förderkontext zu den folgenden Aspekten empfohlen:

→ Aufbau einer gemeinsamen Terminologie, Semantik, Taxonomie sowie darauf gestützter Datenmappings und -schemata in den Domänen Materialwissenschaften und Bauwesen zur Ermittlung von Energieeffizienz und Umweltwirkungen u. a. für den Aufbau von ESG-Datensätzen und deren Nutzung in KI-gestützten Planungswerkzeugen für den zukünftigen Ressourcenverbrauch. Ein solches Normungsprojekt kann Stakeholder aus der kommunalen Bauplanung, der Materialwirtschaft, der Finanzwirtschaft und der Forschung zur Energieeffizienz im Material- und Gebäudesektor involvieren, in Förderprogrammen des BMWK oder BMUV verortet sein und sollte im ersten Halbjahr 2023 initiiert werden.
→ Entwicklung eines branchenunabhängigen Kommunikationsformats für die Bestimmung des Energie- und Ressourcenverbrauchs von Gütern und Dienstleistungen. Ein derartiges Normungs- bzw. Pilotprojekt hat Bezug zu Akteur*innen aller Branchen bzw. Industriebereiche mit Privatkundenbezug und zur sozioökologischen Forschung, kann im Förderkontext des BMUV platziert werden und sollte angesichts der Bandbreite an Stakeholdern und entsprechendem Koordinationsbedarf zu Beginn des Jahres 2023 initiiert werden.
→ Entwicklung einer Methodik zur Beurteilung der Laufzeit, Akkuranz und Nachhaltigkeitsgüte von KI- und ML-Systemen. Ein entsprechendes Normungsprojekt sollte Vertreter*innen aller Branchen mit derzeit und in naher Zukunft prognostizierter intensiver KI- und ML-Nutzung, KI-Zertifizierer*innen sowie Forscher*innen mit Expertise in Algorithmik und Maschinellem Lernen involvieren, im öffentlichen Förderkontext platziert werden und angesichts der Bandbreite an möglichen Teilnehmer*innen frühzeitig (Anfang 2023) in der Teilnehmer*innenakquisition initiiert werden.
→ Aufbau eines synergetischen, dynamischen Modellansatzes für Referenzarchitekturmodelle in Smart Manufacturing und Smart Grid für die Abbildbarkeit dynamischen Variablenverhaltens und die Identifikation kritischer

Systembereiche. Auf Basis des RAMI 4.0 (Referenzarchitekturmodell Industrie 4.0) und des SGAM (Smart Grid Architecture Model) Reference Designation Modells, die die Entwicklung von Musterlösungen als „Systems of Systems“-Ansatz unterstützen und vereinfachen.

→ Entwicklung eines dynamischen Berechnungsverfahrens für die CO_2-Emissionen aus dem Strommix zur Berücksichtigung der geografisch-temporalen Volatilität nachhaltiger Stromerzeugung. Ein solches Pilot- bzw. Normungsprojekt sollte Stakeholder der Elektrizitätswirtschaft und der Geodäsie bzw. Kartografie involvieren, im Förderkontext des BMWK oder BMUV verortet sein und im ersten Halbjahr 2023 initiiert werden.



3 Akteurs- und Normungsumfeld

Sowohl auf nationaler als auch auf europäischer und internationaler Ebene gibt es zahlreiche Akteur*innen, Initiativen sowie Gremien und Normungs- bzw. Standardisierungsaktivitäten, die sich intensiv mit dem Thema KI auseinandersetzen. Das folgende Kapitel stellt eine Auswahl der wesentlichen Akteur*innen und Initiativen im KI-Umfeld dar.[11]

3.1 Innovationspolitische Initiativen

Plattform Lernende Systeme (PLS)

Die Plattform Lernende Systeme (PLS)[12] wurde 2017 durch das Bundesministerium für Bildung und Forschung mit dem Ziel initiiert, KI zum Wohl der Gesellschaft zu gestalten sowie Handlungsempfehlungen für einen verantwortlichen Einsatz von KI herauszuarbeiten. In sieben thematischen Arbeitsgruppen bündelt die Initiative das Wissen von rund 200 Expert*innen aus Wissenschaft und Wirtschaft sowie von Entscheidungsträger*innen im Innovationsökosystem und in Politik u. a. zu rechtlichen und gesellschaftlichen Rahmenbedingungen für die Anwendung von KI.

Die Arbeitsgruppen sind:

- → Technologische Wegbereiter und Data Science
- → Arbeit/Qualifikation, Mensch-Maschine-Interaktion
- → IT-Sicherheit, Privacy, Recht und Ethik
- → Geschäftsmodellinnovationen
- → Mobilität und intelligente Verkehrssysteme
- → Gesundheit, Medizintechnik, Pflege
- → Lernfähige Robotiksysteme

Die Plattform Lernende Systeme bietet zudem einen Überblick über den KI-Standort Deutschland: So werden in der KI-Landkarte[13] KI-Anwendungen, -Forschungsinstitutionen, -Transferzentren und -Studiengänge im gesamten Bundesgebiet dargestellt. Das KI-Monitoring[14] zeigt anhand verschiedener Indikatoren den Status quo sowie Entwicklungspotenziale bei Forschung und Transfer auf.

KI-Kompetenzzentren

Ein zentraler Bestandteil des deutschen KI-Kosmos und der KI-Strategie der Bundesregierung sind die Nationalen Kompetenzzentren für KI-Forschung. Seit Juli 2022 werden sechs Kompetenzzentren vom Bundesministerium für Bildung und Forschung (BMBF) dauerhaft gefördert und sollen national und international vernetzte Spitzenforschung betreiben sowie KI-Kompetenzen in Deutschland aufbauen und erweitern. Übergeordnetes Ziel ist die Sicherung der technologischen Souveränität Deutschlands bei Künstlicher Intelligenz. Mit den Kompetenzzentren sollen wissenschaftliche Durchbrüche ermöglicht, neue Start-ups und Geschäftsmodelle hervorgebracht, der Forschungstransfer beschleunigt, KI-Fachkräfte ausgebildet und neue Arbeitsplätze geschaffen werden.

Zu den sechs KI-Kompetenzzentren zählen:

- → **Berlin Institute for the Foundation of Learning and Data (BIFOLD)**
  Gefördert durch das BMBF sowie die Berliner Senatskanzlei für Wissenschaft und Forschung ist das BIFOLD[15] ein Zusammenschluss von Forschungseinrichtungen, welche sich thematisch auf Big-Data-Management und Maschinelles Lernen (ML) fokussiert haben. Konkret trägt die Forschungsinitiative zur Entwicklung von Werkzeugen und Infrastrukturen für KI-Anwendungen bei. Eine Vielzahl der entwickelten Werkzeuge wird von der Initiative als Open Source[16] angeboten.
- → **Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)**
  Als weltweit größtes unabhängiges Forschungszentrum für KI forscht das Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI)[17] an mehreren Standorten in Deutschland an Lösungen für einen menschenzentrierten Einsatz von Künstlicher Intelligenz. Es fokussiert sich dabei insbesondere auf gesamtgesellschaftliche Herausforderungen wie beispielsweise den menschengemachten Klimawandel, soziale Ungerechtigkeiten, den Kampf gegen Krankheiten und initiiert, realisiert und unterstützt zahlreiche Aktivitäten, um verlässliche und vertrauenswürdige KI aus Deutschland und Europa im internationalen Wettbewerb ganz vorne zu platzieren.
- → **Münchner Kompetenzzentrum für Maschinelles Lernen (MCML)**
  Als Zusammenschluss der Ludwig-Maximilians-Universität München (LMU) und der Technische Universität München (TUM) wird das MCML[18] von der deutschen

11 Die Darstellung erhebt keinen Anspruch auf Vollständigkeit.
12 https://www.plattform-lernende-systeme.de/startseite.html
13 https://www.plattform-lernende-systeme.de/ki-in-deutschland.html
14 http://www.kimonitoring.de/
15 https://bifold.berlin/
16 https://www.bifold.berlin/impact-transfer/open-source-systems-tools-data
17 https://www.dfki.de/web
18 https://mcml.ai/

und bayerischen KI-Strategie gefördert. Der Zusammenschluss besteht aus 50 Forschungsgruppen, welche sich auf Grundlagenforschung sowie Forschung von anwendungsbezogenem ML konzentrieren. Um den Transfer der Erkenntnisse in die Wirtschaft zu gewährleisten, unterstützt die Initiative die Ausbildung von Studierenden und bietet Schulungen in KI-Anwendungen für Industrieunternehmen an.

→ **Lamarr-Institut für Maschinelles Lernen und Künstliche Intelligenz**
Im Rahmen der deutschen KI-Strategie wird das Lamarr-Institut[19] durch das BMBF und dem Land Nordrhein-Westfalen gefördert. Das Institut geht aus einer Initiative der Technischen Universität Dortmund, dem Fraunhofer IML, dem Fraunhofer IAIS und der Universität Bonn hervor und löst das Kompetenzzentrum Maschinelles Lernen Rhein-Ruhr (ML2R)[20] ab. Neben der Forschung an ML-Technologien bietet die Initiative auch Bildungsangebote für Schüler*innen, Studierende und Nachwuchswissenschaftler*innen. Besonderes Augenmerk legt die Initiative dabei auf nachhaltige Innovationen und soziale Gerechtigkeit.

→ **Zentrum für skalierbare Datenanalyse und Künstliche Intelligenz (ScaDS.AI)**
Das Zentrum für skalierbare Datenanalyse und Künstliche Intelligenz (ScaDS.AI)[21] setzt in ihrer Forschungsarbeit auf angewandte KI und KI-Methoden im Bereich Big Data sowie Datenanalyse. Geforscht wird dabei in den drei Kernbereichen „Applied AI & Big Data", „AI Algorithms & Methods" sowie „Big Data Analytics & Engineering", wobei bei allen Bereichen auch die ethischen und sozialen Dimensionen sowie die Sicherheit und Skalierbarkeit betrachtet werden. Als Zusammenschluss von 13 Forschungseinrichtungen ist die Initiative gefördert durch das BMBF und das Bundesland Sachsen und hat sich zum Ziel gesetzt, den Transfer der wissenschaftlichen Ergebnisse durch Kooperationsprojekte mit Industriepartner*innen zu gewährleisten.

→ **Tübingen AI Center**
Das Max-Planck-Institut für intelligente Systeme und die Eberhard Karls Universität Tübingen haben sich im Tübingen AI Center[22] zusammengeschlossen, um Lernsysteme zu entwickeln, welche einen positiven Einfluss auf die Gesellschaft haben. Gefördert wird die Initiative vom BMBF und dem Ministerium für Wissenschaft, Forschung und Kunst des Landes Baden-Württemberg (MWK BW). Durch die räumliche Nähe zur Initiative Cyber Valley hat sich zwischen den beiden Initiativen eine Partnerschaft gebildet.

Mittelstand-Digital
Die Initiative Mittelstand-Digital[23] ist eine Initiative des BMWK, bei der sich Firmen aus dem gesamten Bundesgebiet zu Themen der Digitalisierung informieren und weiterbilden können. Konkret bietet die Initiative dabei digitale Lernangebote als auch Praxisbeispiele und Demonstrationsorte, um digitale Technologien in Anwendung zu sehen. Die Initiative spannt zudem ein deutschlandweites Netzwerk aus Zentren, um Firmen eine lokal verfügbare Informationsquelle bereitzustellen. Mit im Angebot der „Mittelstand-Digital"-Strategie sind auch KI-Trainer, welche mit Workshops, Vorträgen und Roadshows zum Thema aufklären und beraten.

Mittelstand-Digital-Zentren und Mittelstand-4.0-Kompetenzzentren
Die vom BMWK geförderten (Kompetenz-)Zentren[24] dienen als erste Anlaufstelle für Unternehmen, die sich zum Thema Digitalisierung informieren möchten. Insgesamt 66 Kompetenzzentren werden dafür derzeit gefördert. Zum Leistungsspektrum gehören die Klärung von Fragen und die Schulung im sicheren Umgang mit neuen Technologien, die Unterstützung beim Testen von entwickelten Anwendungen als auch Ratschläge zu IT-Recht und der Entwicklung digitaler Geschäftsmodelle. Um sich über die Standorte der Zentren zu informieren, unterhält das BMWK auf ihrer Webseite eine interaktive Karte.[25]

Kompetenzplattform KI.NRW
Das Land Nordrhein-Westfalen (NRW) hat mit der Kompetenzplattform KI.NRW[26] eine zentrale Anlaufstelle für Künstliche Intelligenz in NRW geschaffen, die den Transfer von KI aus der Spitzenforschung in die Wirtschaft beschleunigen soll. Neben dem Wissenstransfer fördert die Kompetenzplattform KI-Projekte zur Etablierung von KI-Technologien in der breiten

19 https://lamarr-institute.org/
20 https://www.ml2r.de/
21 https://scads.ai/
22 https://tuebingen.ai/
23 https://www.mittelstand-digital.de/MD/Navigation/DE/Home/home.html)
24 https://www.mittelstand-digital.de/MD/Redaktion/DE/Artikel/Mittelstand-4-0/mittelstand-40-kompetenzzentren.html
25 https://www.mittelstand-digital.de/MD/Navigation/Karte/SiteGlobals/Forms/Formulare/karte-formular.html
26 https://www.ki.nrw/

Industrie und zur beruflichen Qualifizierung. Ziel ist es, einen effizienten Technologietransfer und die enge Zusammenarbeit von Mittelstand, Start-ups, Universitäten, Hochschulen sowie Forschungseinrichtungen in NRW sicherzustellen und zu unterstützen. Zusätzlich dazu stehen auch gesellschaftliche Aspekte und ethische Grundsätze zur Gestaltung von Künstlicher Intelligenz im Fokus der Plattform.

Regionale Kompetenzzentren der Arbeitsforschung
Innerhalb des Förderschwerpunkts „Zukunft der Arbeit: Regionale Kompetenzzentren der Arbeitsforschung" werden weitere Kompetenzzentren vom BMBF gefördert mit dem Ziel, die Arbeitsforschung enger mit der Arbeitsgestaltung in der betrieblichen Praxis sowie der Hochschulausbildung zu verzahnen und den Transfer neuer Erkenntnisse in die Gesellschaft zu stärken. Zu diesen Zentren zählen:

→ **Künstlich und Menschlich Intelligent (K-M-I)**
Das Kompetenzzentrum erforscht den Einsatz von Künstlicher Intelligenz im Bereich der Arbeitsgestaltung, etwa durch die Unterstützung von intelligenten Assistenzsystemen bei der Produktionsplanung und -steuerung oder bei der Wartung und Instandhaltung komplexer Anlagen. Forscher*innen sollen das Potenzial intelligenter technischer Systeme bei der Kollaboration zwischen Mensch und Maschine arbeitswissenschaftlich untersuchen und in Betrieben erproben.

→ **WIRKsam**
Das geförderte Kompetenzzentrum WIRKsam[27] ist eine Initiative, welche an KI-Innovationen in Arbeits- und Prozessabläufen forscht. Neben der Entwicklung neuer Konzepte wird auch die betriebliche Umsetzung in der Initiative mitbetrachtet. Unternehmen aus der Kohle- und Textilregion des Rheinlands bilden dabei die Hauptzielgruppe. Das Kompetenzzentrum ist bis Oktober 2026 durch das BMBF gefördert und wird vom Projektträger Karlsruhe betreut.

Kompetenzzentrum für KI-Engineering (CC-KING)
Das Kompetenzzentrum für KI-Engineering (CC-KING)[28] wurde von drei Forschungseinrichtungen aus Karlsruhe (Fraunhofer IOSB, Forschungszentrum Informatik (FZI) und Karlsruher Institut für Technologie (KIT)) ins Leben gerufen. Gefördert vom Ministerium für Wirtschaft, Arbeit und Tourismus Baden-Württemberg (WM BW) hat sich das Kompetenzzentrum zum Ziel gesetzt, den Einsatz von Methoden der Künstlichen Intelligenz (KI) und des Maschinellen Lernens (ML) in der aus der Praxissicht der Ingenieure zu erleichtern. Mit dem Schwerpunkt auf industrielle, nachhaltige Produktion und bedarfsgerechte Mobilität werden hier Grundlagen erforscht und Methoden entwickelt, die die betriebliche Arbeit verbessern sollen.

KI-Observatorium in Arbeit und Gesellschaft
Durch die Denkfabrik des Bundesministeriums für Arbeit und Soziales (BMAS) ins Leben gerufen, beschäftigt sich das KI-Observatorium[29] mit den fünf Handlungsfeldern Technologie-Foresight und Technikfolgenabschätzung, KI in der Arbeits- und Sozialverwaltung, Ordnungsrahmen für KI und Soziale Technikgestaltung, Aufbau internationaler Strukturen und europäischer Vernetzung sowie Gesellschaftlicher Dialog und Vernetzung. Das Ziel des Observatoriums ist es, die Entwicklungen im Bereich KI zu überblicken und ihre Folgen auf die Gesellschaft abzuschätzen und positiv zu beeinflussen.

Cyber Valley
Gegründet durch das Land Baden-Württemberg sowie durch Forschungseinrichtungen und Wirtschaftsunternehmen ist das Cyber Valley[30] eine Initiative, die die Forschung im Bereich des ML, der Robotik und der Computer Vision verstärken soll. Neben dem Ausbau der Grundlagenforschung bietet das Cyber Valley auch die Möglichkeit der Förderung für Start-ups, die die gewonnenen wissenschaftlichen Erkenntnisse in die wirtschaftliche Anwendung bringen. Das Cyber Valley konzentriert seine Förderung vor allem auf den Raum Stuttgart-Tübingen.

AI Quality & Testing Hub
Das Konzept eines AI Quality & Testing Hubs[31] wird seit 2020 von VDE und TÜV im Austausch mit mehreren Bundesländern vorangetrieben. Gemeint ist eine Institution mit europäischer Reichweite, die alle Puzzleteile zusammenbringt, die für das Bewerten und Management von KI-Qualität notwendig sind, z. B. einen Überblick von Forschungsständen, Zugang und Aufbau von Trainings-Datensätzen, Simulationsumgebungen mit standardisierten Interfaces, Trainings und Kompetenzerwerb sowie maßgeschneiderter Qualitätsverbesserung für Herstellende und Anwender*innen/Betreiber*innen von KI-Produkten.

27 https://www.arbeitswissenschaft.net/wirksamweb/
28 https://www.ki-engineering.eu/
29 https://www.ki-observatorium.de/
30 https://cyber-valley.de/de
31 https://www.vde.com/de/presse/pressemitteilungen/ai-quality-testing-hub

AI4Germany
AI4Germany[32] ist eine Dachinitiative zur umsetzungsnahen Förderung und Implementierung von KI. Vom Münchener Start-up Accelerator Unternehmer TUM gegründet versteht sich die Initiative als anwendungsnahe Ergänzung zwischen PLS und AI4Europe. Festgeschriebenes Ziel des Zusammenschlusses ist die Stärkung Deutschlands als Entwicklungsstandort für Hightech-KI-Anwendungen.

Initiative for Applied Artificial Intelligence
59 Partner*innen aus Industrie, öffentlicher Hand und Forschung haben sich bei der Initiative for Applied Artificial Intelligence (appliedAI)[33] zusammengefunden. Mit dem Ziel, eine kollaborative Plattform zu schaffen, die Personen schult und Innovationen vorantreibt, ist appliedAI Europas größte Initiative. Das Angebot der Plattform reicht von Beratung und Schulung über Austausch und Vorträge bis hin zu Zugängen von KI-Werkzeugen/-Ökosystemen und Start-ups. AppliedAI ist auch Mitglied der AI4Germany.

Plattform Industrie 4.0
Die Plattform Industrie 4.0[34] ist ein Netzwerk zur Förderung der Digitalisierung der Industrie. Die Leitung teilen sich das Bundesministerium für Wirtschaft und Klimaschutz (BMWK) und das BMBF gemeinsam mit Technologieunternehmen, Verbänden und Forschungsorganisationen. Neben Themen wie geopolitischen Krisen und Lieferkettenresilienz befasst sich das Netzwerk regelmäßig mit KI, so z. B. als Fokusthema in der Arbeitsgruppe „Technologie- und Anwendungsszenarien".

KI-Transfer-Hub SH
Das Land Schleswig-Holstein startete die eigene Initiative KI-Transfer-Hub SH[35]. Mit dieser Initiative möchte das Land vor allem KMUs und Start-ups ermöglichen, KI-Technologien in ihr Geschäftsmodell mit einzubringen. Dabei unterstützen Partner*innen aus Wissenschaft und Wirtschaft aus Norddeutschland. Die Europäische Union fördert diese Initiative mit Mitteln aus dem Europäischen Fonds für regionale Entwicklung.

Neben den genannten Initiativen gibt es über die deutschen Landesgrenzen hinaus noch weitere Zusammenschlüsse, welche KI auf **europäischer und internationaler Ebene** voranbringen möchten, dazu gehören:

CLAIRE
Das Confederation of Laboratories for Artificial Intelligence Research in Europe (CLAIRE)[36] ist ein paneuropäisches Bündnis, dem sich 445 Forschungseinrichtungen angeschlossen haben und das sich zum Ziel gesetzt hat, die Forschung und Innovation im Bereich KI zu stärken. Als Partnerin von HumaneAI trägt CLAIRE zur Entwicklung von vertrauenswürdiger KI bei. Seit der Gründung im Jahre 2018 in Den Haag sind weitere Niederlassungen in Deutschland, Norwegen, Tschechien, Italien, der Schweiz und Belgien hinzugekommen. Zum Arbeitsspektrum der Initiative gehören u. a. Maschinelles Lernen, Wissensrepräsentation und Argumentation, Verarbeitung natürlicher Sprachen, aber auch Themen wie Robotik, Computer Vision sowie ethische und soziale Aspekte.

ELLIS
Das European Laboratory for Learning and Intelligent Systems (ELLIS)[37] ist ein 2018 gegründetes europäisches Netzwerk. ELLIS versteht sich als Motor, um Europas wirtschaftliche Position in der KI-Entwicklung zu stärken. Daher treibt es neben der Grundlagenforschung auch die Gründung neuer KI-Start-ups voran. Zwischen ELLIS und CLAIRE besteht eine enge Beziehung, da sich beide Initiativen in ihren Bemühungen komplementieren. ELLIS ist mittlerweile an 35 Standorten in Europa vertreten, acht davon befinden sich in Deutschland.

AI4EUROPE
Als Nachfolger zu AI4EU ist AI4EUROPE[38] im Jahre 2022 ins Leben gerufen worden und stellt eine Plattform für Forschungsgruppen dar, um wissenschaftliche Erkenntnisse zu teilen und dadurch weitere Innovationen voranzutreiben. Gestartet wurde die Plattform an der University College Cork in Irland. Neben der Forschung und Weiterbildung bietet die Plattform auch für die Wirtschaft eine Möglichkeit, sich über ihre Anwendungsfälle auszutauschen.

32 https://www.ai4germany.de/
33 https://www.appliedai.de/de/
34 https://www.plattform-i40.de/IP/Navigation/DE/Home/home.html
35 https://kuenstliche-intelligenz.sh/de/startseite
36 https://claire-ai.org/?lang=de
37 https://ellis.eu/
38 https://www.ai4europe.eu/

I-DAIR

Das International Digital Health & AI Research Collaborative (I-DAIR)[39] ist eine Initiative, welche die Forschung im Bereich der digitalen Gesundheit und der KI im Gesundheitswesen verbessern will. So soll die digitale Transformation genutzt werden, um allen Ländern und Gemeinschaften zu einer verbesserten Lebensqualität zu verhelfen. Der Zusammenschluss von internationalen Forschungsinstituten vernetzt mittlerweile über 40 Partner*innen. Zwei herausstechende Projekte der Plattform ist die Global Research Map of Digital Health and AI[40], sowie das Real Time Epidemiology & Dashboard[41].

3.2 Normungs- und Standardisierungsumfeld

Inhalte gültiger Normen und Standards repräsentieren den aktuellen Stand von Wissenschaft und Technik. Jedes Normungs- bzw. Standardisierungsdokument bildet die essenziellen Eigenschaften (beispielsweise eines Produkts), Anforderungen (beispielsweise an eine Dienstleistung) oder Verfahren (beispielsweise von Prozessen) ab, die zumeist konsensbasiert von Teilnehmer*innen aus den interessierten Kreisen (Wirtschaft, Wissenschaft, Forschung, Anwender*innen, Verbraucherschutz, Arbeitsschutz, Gewerkschaften, öffentliche Hand und Umweltschutz) entwickelt werden.

Durch die Festlegung von Technik- und Kompatibilitätsanforderungen an Produkte, Dienstleistungen oder Prozesse, aber auch die Definition von Begriffen oder Schnittstellen wird Interoperabilität gewährleistet und der Schutz von Mensch, Umwelt und Sache sichergestellt. Auf diese Weise schaffen Normen und Standards Transparenz und Vertrauen in neue Anwendungen und Technologien.

Der Bedarf für eine neue Norm oder einen neuen Standard wird oft innerhalb dieser interessierten Kreise erkannt. Grundsätzlich kann jedoch jeder die Erstellung einer Norm beim zuständigen nationalen Normungsinstitut (DIN für Deutschland) beantragen.

Je nach Art der Inhalte, Zielgruppe und weltwirtschaftlicher Relevanz erfolgt die Normungsarbeit auf nationaler, europäischer oder internationaler Ebene (siehe Abbildung 11). Wenngleich es Unterschiede zwischen diesen drei Ebenen gibt, so haben alle eines gemeinsam; die Normungsarbeit wird von Expert*innen durchgeführt, die von ihrem nationalen Normungsinstitut zur Mitarbeit auf der europäischen Ebene (bei CEN/CENELEC) oder internationalen Ebene (bei ISO/IEC) entsandt werden. Die nationalen Normungsinstitute der teilnehmenden Länder stellen somit das Bindeglied zwischen den Know-how-Trägern aus den interessierten Kreisen und der aktiven Erarbeitung von Normen dar.

Bereits seit 1975 vertritt DIN[42] als zuständige Normungsorganisation der Bundesrepublik Deutschland die deutschen Interessen in der europäischen Normung (bei CEN[43]) sowie in der internationalen Normung (bei ISO[44]). Die Normung in den Bereichen Elektrotechnik, Elektronik und Informationstechnik wird national und international durch die DKE[45] wahrgenommen, welche die deutschen Interessen sowohl bei CENELEC[46] als auch bei IEC[47] vertritt. Die Normungsarbeit konzentriert sich heute auf die europäische und internationale Ebene, wobei die Prozesshoheit innerhalb Deutschlands bei DIN und DKE liegt, welche die nationalen Arbeiten koordinieren und die durch Delegierte und Expert*innen die deutsche Stimme auf europäischer und internationaler Ebene einbringen.

Die Erarbeitung von Normen erfolgt nach festgelegten Grundsätzen auf nationaler, europäischer und internationaler Ebene und unter Berücksichtigung von Verfahrens- und Gestaltungsregeln. Im Rahmen der Ausschussarbeit wird mit Vertretenden aller interessierten Kreise (beispielsweise Herstellende, Verbraucher*innen, Handel, Hochschulen, Forschungsinstitute, Behörden, Prüfinstitute etc.) der aktuelle Stand der Technik erfasst. Normen entstehen im Konsens aller Beteiligten.

39 https://www.i-dair.org/

40 https://grm.i-dair.org/

41 https://www.i-dair.org/pathfinder/rted

42 Deutsches Institut für Normung e. V., www.din.de

43 Comité Européen de Normalisation, Europäische Organisation für Normung, https://www.cencenelec.eu/

44 International Organization for Standardization, Internationale Organisation für Normung, www.iso.org

45 DKE Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE, www.dke.de

46 Comité Européen de Normalisation Électrotechnique, Europäisches Komitee für elektrotechnische Normung, www.cenelec.eu

47 International Electrotechnical Commission, Internationale Elektrotechnische Kommission, www.iec.ch

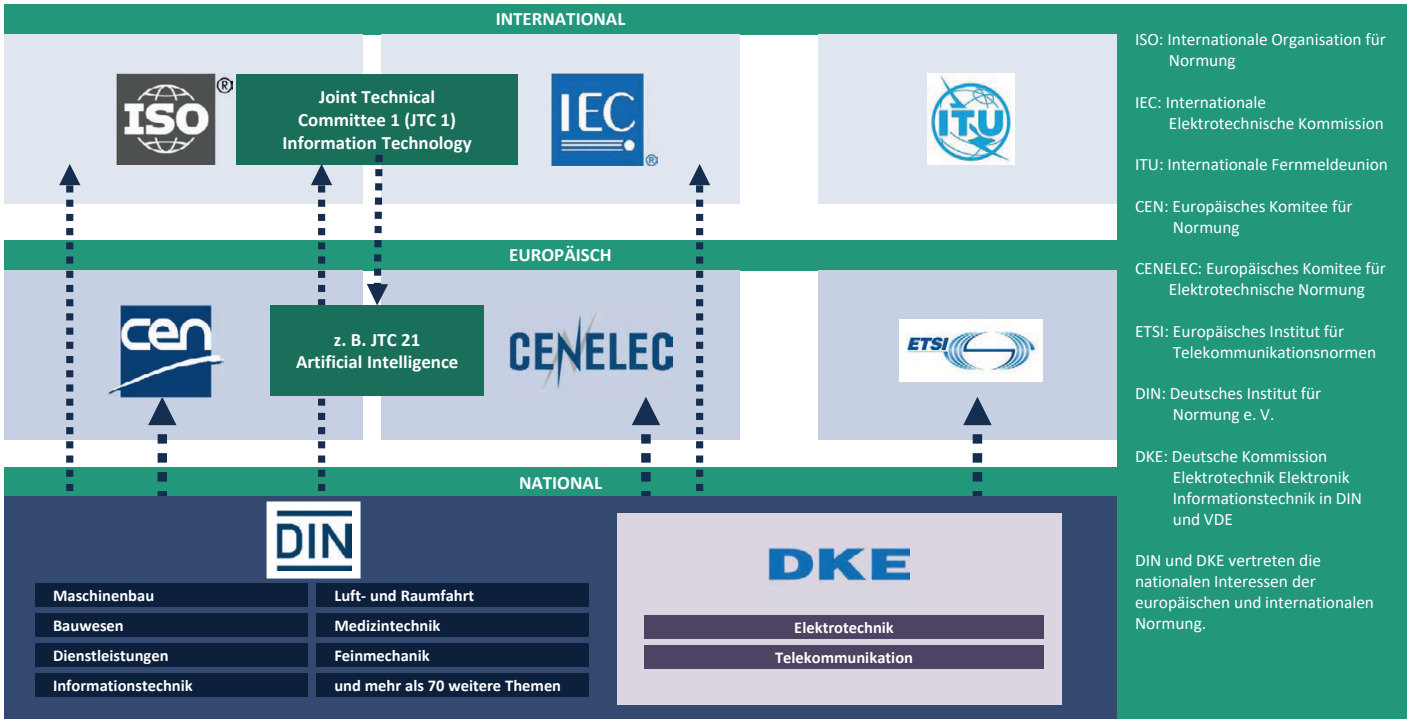


Abbildung 11: Ebenen der Normungsarbeit (Quelle: DIN)

Als „Standards“ werden Dokumente wie Technische Reports (TR), Fachberichte, Vornormen, Spezifikationen (TS, DIN SPEC), Konsortialstandards, Anwendungsregeln (AR), Richtlinien, Expert*innenempfehlungen etc. bezeichnet. Diese werden häufig für Themen mit einem geringen Reifegrad, die ggf. noch nicht vollständig am Markt etabliert sind, erstellt. Die Erarbeitung und Herausgabe erfolgt durch die Normungsinstitute und andere Organisationen und technische Regelsetzer. Bei der Entwicklung von Standards sind Konsens und Einbeziehung aller interessierten Kreise nicht zwingend erforderlich.

3.2.1 KI-Normung auf nationaler Ebene

Die Normungsarbeit zur Künstlichen Intelligenz findet derzeit auf allen drei Ebenen (national, europäisch und international) statt. Auf der nationalen Ebene ist besonders der **DIN/DKE Gemeinschaftsausschuss „Künstliche Intelligenz“ NA 043-01-42 GA**[48] hervorzuheben, der zunächst 2017 von DIN ins Leben gerufen und Ende 2021 zum DIN/DKE-Gemeinschaftsausschuss weiterentwickelt wurde. Mehr als 80 Expert*innen aus Wirtschaft, Wissenschaft, Politik und Zivilgesellschaft engagieren sich in dem Arbeitsausschuss und entwickeln Normen zu Werkzeugen, Prozessen und Anwendungsfeldern der Künstlicher Intelligenz, stets unter Berücksichtigung gesellschaftlicher Chancen und Risiken.

Der Gemeinschaftsausschuss ist als nationales Spiegelgremium für die Konsolidierung der deutschen Meinung zuständig und entsendet die deutsche Delegation sowohl in das europäische Normungsgremium (CEN/CENELEC/JTC21) als auch in das internationale Normungsgremium (ISO/IEC/JTC1).

Er zählt zu den wichtigsten KI-relevanten Gremien zur Umsetzung europäischer Vorgaben (aus Verordnungen, AI Act etc.) und spielt eine wichtige Rolle bei der Erarbeitung entsprechender Normen.

Die Abbildung 12 zeigt die Struktur des nationalen Gemeinschaftsausschusses zu KI.

48 Siehe https://www.din.de/de/interdisziplinaerer-arbeitsausschuss-zu-kuenstlicher-intelligenz-826618

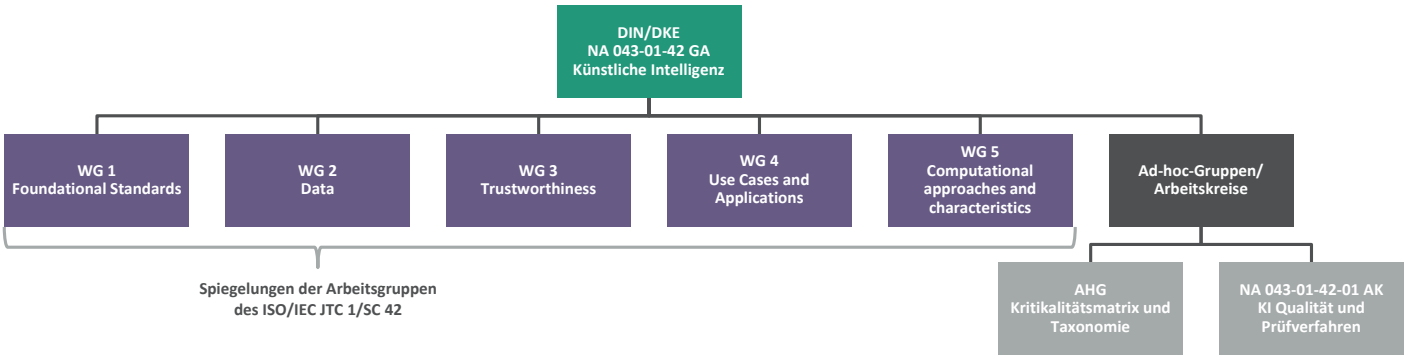


Abbildung 12: Struktur des nationalen Gemeinschaftsausschusses zu KI (Quelle: DIN)

Darüber hinaus ist noch der Arbeitsausschuss „Informationssicherheit, Cybersicherheit und Datenschutz“ NA 043-04-27 AA[49] zu nennen, dessen Themen eine besondere Relevanz für KI darstellen.

3.2.2 KI-Normung auf europäischer Ebene

Seit 2019 erlebt das Thema KI großes Interesse durch die europäische Politik. Mit dem New Legislative Framework verfügt Europa seit Langem über einen einzigartigen und bewährten Mechanismus für das Zusammenspiel von Normung und Regulierung, der auf Basis des Entwurfs des AI Act nunmehr für das Thema KI Anwendung finden soll (siehe Kapitel 1.4). Auf europäischer Ebene besteht die zentrale Aufgabe der Normung darin, europaspezifische Aspekte zu bearbeiten und die europäische Regulierung von KI (v. a. den Vorschlag zum AI Act) mit harmonisierten europäischen Normen zu unterfüttern.

In dem geplanten AI Act spielen Normen eine wichtige Rolle. Sie dienen der verlässlichen Umsetzung der Anforderungen des AI Act und helfen, die Entwicklung von KI-Systemen effizienter und verlässlicher zu machen. Bis Herbst 2024 sollen so europäische Normen beispielsweise zu Transparenz, Logging, Fairness, Risikobeurteilung oder Privatsphärenschutz entwickelt werden. Die Qualität und Geschwindigkeit der europäischen Normungsarbeit hängt von der inhaltlichen Grundlagenarbeit ab, die auf nationaler Ebene geleistet wird.

Das zentrale Gremium für die europäische KI-Normung ist das Gemeinschaftsgremium **CEN/CENELEC JTC 21 „Künstliche Intelligenz“ (CEN/CLC JTC 21)**[50], das von CEN und CENELEC auf Basis der Empfehlungen des „Weißbuches zur Künstlichen Intelligenz“ [7] und der „Deutschen Normungsroadmap Künstliche Intelligenz Ausgabe 1“ [63] im Frühjahr 2021 gegründet wurde.

Das Gemeinschaftsgremium CEN/CLC JTC 21 steht unter deutscher Leitung und wird in der Sekretariatsführung von Dänemark unterstützt. Es ist für die Entwicklung europäischer Normen zur Künstlichen Intelligenz sowie die Beratung anderer Technischer Komitees verantwortlich. Aktuell befasst sich das Gremium u. a. mit folgenden Themen: Green and sustainable AI, Data Governance and Quality for AI, AI Systems risk catalogue and risk management, Overarching unified approaches on trustworthiness-characteristics.

Die vorliegende Normungsroadmap zeigt konkrete Normungsbedarfe zu KI auf und unterstützt damit maßgeblich sowohl den nationalen Gemeinschaftsausschuss zu KI (NA 043-01-42 GA) als auch die europäische KI-Normung.

Abbildung 13 zeigt die Struktur des europäischen Gemeinschaftsausschusses zu KI (CEN/CLC JTC 21).

49 Siehe https://www.din.de/de/mitwirken/normenausschuesse/nia/nationale-gremien/wdc-grem:din21:54770248

50 Siehe https://www.cencenelec.eu/areas-of-work/cen-cenelec-topics/artificial-intelligence/

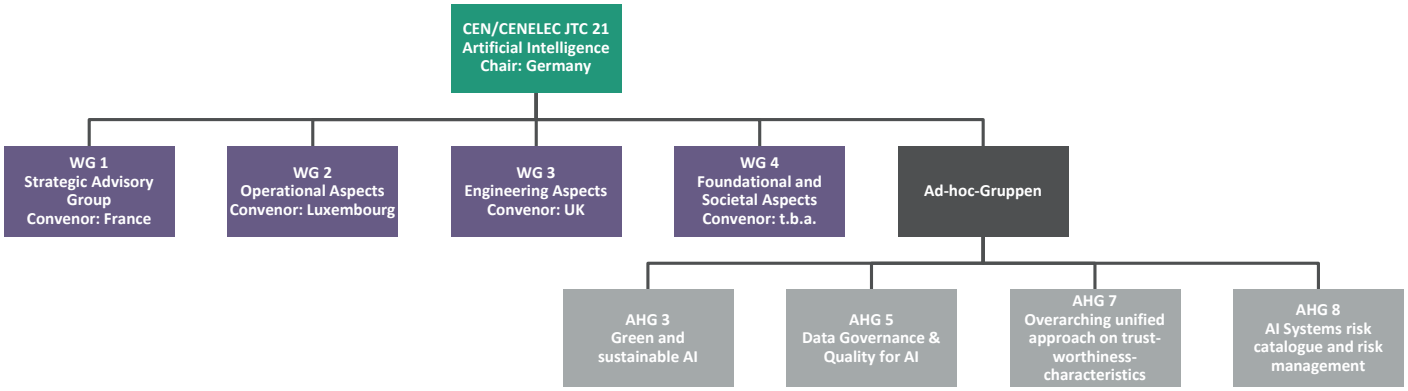


Abbildung 13: Struktur des europäischen Gemeinschaftsausschusses zu KI (Quelle: DIN)

3.2.3 KI-Normung auf internationaler Ebene

Auf internationaler Ebene wurde im Jahr 2017 unter US-amerikanischer Federführung der Gemeinschaftsausschuss **ISO/IEC JTC 1/SC 42 „Artificial Intelligence“** [51] gegründet.

Dieses von ISO und IEC eingesetzte Normungsgremium stellt die zentrale Anlaufstelle für die KI-Normung auf internationaler Ebene dar. Die aktuell 35 Mitgliedsländer des ISO/IEC JTC 1/SC 42 erstrecken sich über alle Kontinente und werden durch 15 „Observing Members“ ergänzt. Diese globale Zusammensetzung sorgt für ein ebenso global abgestimmtes Arbeitsprogramm, welches derzeit die Normung von KI-Grundlagen, Datenstandards im Zusammenhang mit KI, Big Data und Analytik, Vertrauenswürdigkeit, Auswirkungen von KI auf die Politik sowie ethische und gesellschaftliche Belange umfasst. Damit befasst sich das Gremium mit dem gesamten KI-Ökosystem und berät ISO- und IEC-Ausschüsse zu Künstlicher Intelligenz.

Seit seiner Gründung hat das ISO/IEC JTC 1/SC 42 bereits 14 internationale Normen entwickelt und veröffentlicht. Hierzu zählen Normen zu Big Data [52] und der Big Data Referenzarchitektur [53], Standards zur Bewertung der Robustheit von neuronalen Netzen [54], zur Beschreibung von Use Cases [55] sowie zu ethischen und gesellschaftlichen Belangen [56].

Die 25 derzeit laufenden Normungsprojekte thematisieren u. a. die Datenqualität für Analytik und Maschinelles Lernen, die funktionale Sicherheit sowie Qualitätsbewertungsrichtlinien und Folgenabschätzungen für KI-Systeme.

China, Irland, Japan und Deutschland haben derzeit die Sekretariatsführung der „Working Groups“ im ISO/IEC JTC 1/SC 42 inne und können somit die inhaltliche Arbeit zur KI-Normung auch auf internationaler Ebene aktiv mitgestalten.

Abbildung 14 zeigt die Struktur des ISO/IEC JTC 1/SC 42.

Weitere relevante internationale Normungsgremien sind:
Das ISO/IEC JTC 1/SC 27 [57] „Information security, cybersecurity and privacy protection“ steht unter deutscher Federführung und ist für die Entwicklung von Normen und Standards zum Schutz von Informationen und Informations- und Kommunikationstechniken zuständig. Dies umfasst u. a. Sicherheits- und Datenschutzaspekte sowie kryptografische und andere Sicherheitsmechanismen.

Das ISO/IEC JTC 1/SC 41 [58] „Internet of things and digital twin“ erarbeitet internationale Normen und Standards zu Themen wie Internet der Dinge, Digitaler Zwilling sowie verwandte Technologien.

51 Siehe https://www.iso.org/committee/6794475.html
52 Siehe ISO/IEC 20546 [443]
53 Siehe ISO/IEC TR 20547 (Reihe) [438], [439], [440], [441], [442]
54 Siehe ISO/IEC TR 24029-1 [91]
55 Siehe ISO/IEC TR 24030 [293]
56 Siehe ISO/IEC TR 24368:2022 [15]
57 Siehe www.iso.org/committee/45306.html
58 Siehe www.iso.org/committee/6483279.html

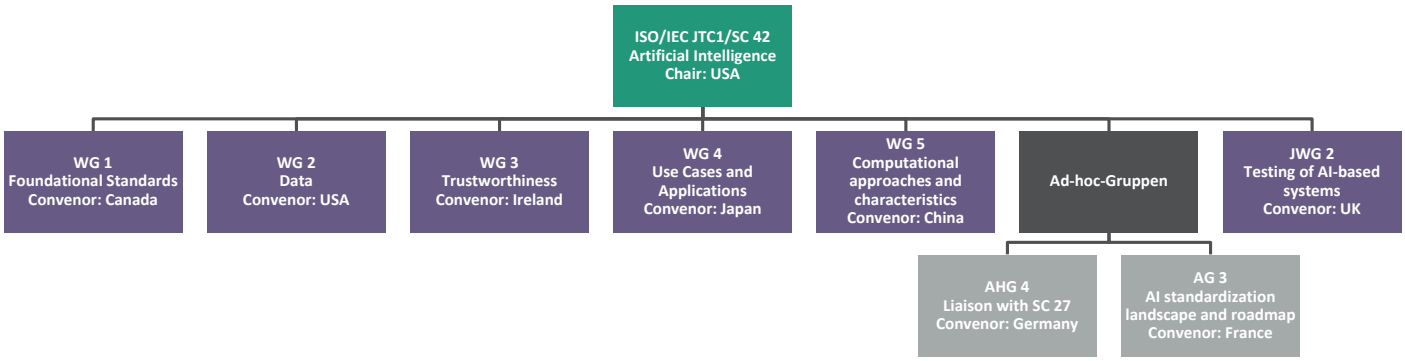


Abbildung 14: Struktur des internationalen Gemeinschaftsausschusses zu KI (Quelle: DIN)

Neben der klassischen, vollkonsensbasierten Normung werden Festlegungen und Empfehlungen zu KI auch von einigen Fachverbänden und Konsortien herausgebracht. Umfangreiche Konsortialarbeiten zur KI-Standardisierung gehen aus diversen Foren und Konsortien wie beispielsweise IETF[59], IEEE[60], CSA Group[61], OGC[62], OMG[63] oder W3C[64] hervor und ergänzen die Normung in teilweise sehr speziellen Themenfeldern.

3.3 Forschungs- und Umsetzungsprojekte zu KI

Die deutsche Forschung im Bereich der Künstlichen Intelligenz gehört zu den weltweit führenden. Um die Potenziale der Künstlichen Intelligenz nachhaltig zu heben und wirtschaftlich zu verwerten, müssen die innovativen Forschungsergebnisse aber auch in die Praxis überführt werden. Als ein anerkanntes und vertrauenswürdiges strategisches Mittel können Normen und Standards helfen, wissenschaftlichen Ergebnissen einen schnellen Zugang zum Markt zu ermöglichen.

Auf nationaler wie auch auf europäischer Ebene sind die Normungsinstitute in verschiedenen Rollen in KI-Forschungsprojekte eingebunden und unterstützen so die Identifikation wesentlicher Standardisierungspotenziale, die Entwicklung von Standardisierungsstrategien und die Initiierung von Standardisierungsaktivitäten.

59 Internet Engineering Task Force, siehe: www.ietf.org/

60 Institute of Electrical and Electronics Engineers, siehe: www.ieee.org/

61 Siehe: www.csagroup.org/

62 Open Geospatial Consortium, siehe: www.ogc.org/

63 Object Management Group, siehe: www.omg.org/

64 World Wide Web Consortium, siehe: www.w3.org/

Im Folgenden wird eine Auswahl an KI-Forschungsprojekten vorgestellt, bei denen die Normung und Standardisierung ein Kernelement darstellen.

3.3.1 KI-Forschungsprojekte[65]

VIKING

Das Projekt VIKING[66] (Vertrauenswürdige künstliche Intelligenz für polizeiliche Anwendungen) ist im Januar 2022 gestartet und wird vom Bundesministerium für Bildung und Forschung (BMBF) gefördert. Es wird das Ziel verfolgt, einen Katalog für die Einhaltung akzeptabler ethischer und hoher rechtlicher Anforderungen für KI-Verfahren im polizeilichen Alltag sowie die Operationalisierung der zugrunde liegenden Prinzipien vertrauenswürdiger KI im Licht des neuen EU-Rechts und weiterer rechtlicher und ethischer Anforderungen zu entwickeln. Mit der Verzahnung unterschiedlicher Aktivitäten im Projekt entstehen die Grundlagen für vertrauenswürdige KI. Die Anwender*innen begleiten die Arbeiten und evaluieren die Demonstratoren und Anforderungskataloge hinsichtlich Funktionalität und Praktikabilität für den polizeilichen Alltag, was eine enge Bindung der Forschungsarbeiten an die tatsächliche Relevanz für die Bedarfsträger*innen sicherstellt. Im Erfolgsfall können die Ergebnisse aus VIKING zukünftig zum Best Practice für den polizeilichen Einsatz von KI-Verfahren avancieren, Recht und Sicherheit in Europa stärken und maßgeblich die rasant wachsenden nationalen und internationalen Märkte dieses Segments durch „Technik made in Germany“ prägen.

65 Die Darstellung erhebt keinen Anspruch auf Vollständigkeit.

66 https://www.din.de/de/forschung-und-innovation/partner-in-forschungsprojekten/ki/viking-872288

STAFFEL

Das im Dezember 2021 gestartete Projekt STAFFEL[67] wurde vom Bundesministerium für Verkehr und digitale Infrastruktur (BMVI) ins Leben gerufen und widmet sich der Bereitstellung einer KI-gestützten Internetplattform, um einen sicheren, datenbasierten und speditionsübergreifenden „Staffelverkehr" zu ermöglichen. Zur Zielerreichung werden nach einer detaillierten Anforderungsanalyse die Plattform und das Diebstahlsicherungssystem prototypisch entwickelt und in zwei Feldversuchen validiert. Zunächst werden regionale Transportunternehmen über einen Lenkzeitenmarktplatz miteinander vernetzt. Danach werden Wechselstationen entlang einer Hauptverkehrsroute etabliert und der Staffelverkehr wird praktisch erprobt. Ziel ist es, Effekte, Potenziale und Herausforderungen für den Lkw-Güterverkehr zu identifizieren und eine europaweite Umsetzung vorzubereiten. Dabei wird auch die Standardisierung eine wichtige Rolle einnehmen.

BIG BICTURE

Das Projekt Big Picture[68] wird von der Europäischen Union gefördert und verfolgt seit 2021 das Ziel, eine rasche Entwicklung von KI in der Pathologie zu ermöglichen, indem die erste europäische ethische und qualitätskontrollierte Plattform unter Einhaltung der DSGVO (Datenschutz-Grundverordnung) geschaffen wird, in der sowohl groß angelegte Daten als auch KI-Algorithmen gleichzeitig vorhanden sind. Die BIGPICTURE-Plattform wird auf nachhaltige und integrative Weise entwickelt, indem sie Gemeinschaften von Patholog*innen, Forscher*innen, KI-Entwickler*innen, Patient*innen und aus der Industrie miteinander verbindet. Über die Schaffung einer gemeinsamen Infrastruktur (Hardware und Software) sollen Millionen von Bildern gespeichert, geteilt und verarbeitet werden. Dabei werden rechtliche und ethische Rahmenbedingungen und Funktionalitäten geschaffen, um eine angemessene Nutzung sowie Verarbeitung der Daten für Diagnose- und Forschungszwecke zu gewährleisten und gleichzeitig die Privatsphäre der Patient*innen und die Vertraulichkeit der Daten vollständig zu respektieren.

IMPULSE

Das Forschungsprojekt IMPULSE[69] (Identity Management in Public Services) wird von der Europäischen Union im Rahmen des Horizon-2020-Programms gefördert und konzentriert sich seit 2021 insbesondere auf zwei der vielversprechendsten und disruptivsten Technologien unserer Zeit: Künstliche Intelligenz (KI) und Blockchain, sowie deren Beiträge und Auswirkungen auf elektronisches Identitätsmanagement (eID) in öffentlichen Diensten. Es sollen zwei Hauptergebnisse produziert werden: eine ganzheitliche KI- und Blockchain-Technologie, die DSGVO-konforme eID unterstützt und handlungsfähige Roadmaps sowie Empfehlungen für die Übernahme, Ausweitung und Nachhaltigkeit solcher fortschrittlichen eID-Technologien durch öffentliche Dienste und für politische Entscheidungsträger erarbeitet.

KIOptiPack

Das Vorhaben KIOptiPack[70] (ganzheitliche KI-basierte Optimierung von Kunststoffverpackungen mit Rezyklatanteil) beabsichtigt die Entwicklung und Validierung praxisreifer KI-gestützter Werkzeuge für die erfolgreiche und qualitätsgerechte Produktion von Kunststoffverpackungen mit hohem Rezyklatanteil. Die KI- und Dateninfrastruktur wird auf den in der Gaia-X-Initiative entwickelten Konzepten und Systemen aufbauen und die verteilte KI-Anwendung und souveränen Datenaustausch ermöglichen. Mit KI-gestützten agilen Analyse-Tools soll die Materialqualifizierung unterstützt werden und eine Steigerung der Qualität, Robustheit und Produktivität in der Herstellung von rezyklathaltigen Verpackungsmaterialien erfolgen. Die Nachhaltigkeitsbewertung sowie die Weiterentwicklung von Geschäftsmodellen der Kreislaufwirtschaft werden als integraler Bestandteil verfolgt. Darüber hinaus soll ein Innovationslabor unter Einbeziehung aller relevanten Stakeholdergruppen zur kollaborativen Entwicklung innovativer Lösungen auf Basis realer Verbraucherbedürfnisse einerseits und notwendiger Vorgaben und Interessen der Akteur*innen entlang der Wertschöpfungskette andererseits aufgebaut werden.

67 https://www.din.de/de/forschung-und-innovation/partner-in-forschungsprojekten/ki/staffel-860360

68 https://www.din.de/de/service-fuer-anwender/normungsportale/gesundheit/forschung-innovation-standards/aktuelle-forschungsprojekte/bigpicture-791128

69 https://www.din.de/de/forschung-und-innovation/partner-in-forschungsprojekten/ki/impulse-799412

70 https://www.fona.de/de/massnahmen/foerdermassnahmen/ki-hub-kunststoffverpackungen.php

ZVKI

Das Projekt ZVKI[71] (Zentrum für vertrauenswürdige Künstliche Intelligenz) wurde 2021 vom Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz (BMUV) gestartet und hat sich die Förderung von Transparenz und Vertrauen in KI-Anwendungen bei Verbraucher*innen durch das Zusammenspiel von Politik, Wirtschaft, Wissenschaft und Gesellschaft zum Ziel gesetzt. Die Aktivitäten des ZVKI fokussieren sich rund um die Aufklärung und die Informationsweitergabe für Verbraucher*innen sowie die wissenschaftliche Begleitung von KI-Anwendungen hinsichtlich deren negativer Auswirkungen zum Schutz von Menschen. Darüber hinaus werden Instrumente zur Bewertung von KI-Systemen und Anforderungen für deren Zertifizierung erarbeitet, um die Grundlage für das Vertrauen der Menschen gegenüber den Entwicklungen und dem Einsatz von Künstlicher Intelligenz zu schaffen. Dabei sollen möglichst viele Stakeholder und Ideen zusammengebracht werden, um gemeinsam eine vertrauenswürdige KI zu gestalten.

KIDD

Der Fokus des vom BMAS geförderten Projekts KIDD[72] (KI im Dienste der Diversität) liegt seit 2020 darin, Betriebe zu befähigen, menschenzentrierte digitale Anwendungen in Unternehmen einzuführen. Hierbei wird ein innovativer, auf andere Unternehmen und Organisationen übertragbarer Prozess (KIDD-Prozess), für die transparente, partizipative und inklusive Einführung der KI in Unternehmen entwickelt. Ferner wird das Wie der Digitalisierung auf gerechte, transparente und verständliche Weise diskutiert, erprobt und die konkreten Ergebnisse werden nach Abschluss einer breiten Öffentlichkeit zur Verfügung gestellt, um die Digitalisierung in Unternehmen gerecht und transparent zu gestalten.

KIMEDS

Das vom BMBF geförderte und 2022 gestartete Projekt KIMEDS[73](KI-assistierte Zertifizierung medizinischer Software) verfolgt das Ziel, Zertifizierungsverfahren bei softwarebasierter Medizintechnik zu verbessern. Im Rahmen des Projekts wird an KI-Systemen geforscht, die helfen sollen, den Zulassungsprozess zu beschleunigen. Dabei soll insbesondere die Überwachung von Produktsicherheitsrisiken, von der Entwicklung der medizinischen Software bis hin zur Zertifizierung und Überwachung im Betrieb durch KI-Systeme gestützt werden. Diese Zertifizierungsprozesse stellen gerade in der Medizin häufig eine große Herausforderung dar, wenn es darum geht, bestehenden Regularien zu entsprechen. Mithilfe des Projekts soll die Frage beantwortet werden, wie ein KI-System diesen Prozess adäquat unterstützen kann.

QI-Digital

Das Ziel des Projekts QI-Digital[74] ist es, eine verlässliche Qualitätsinfrastruktur (QI) zu gestalten. Als System in verschiedenen Institutionen und Prozessen trägt es wesentlich zur Sicherheit von Produkten und Anwendungen, zum Schutz von Gesundheit und Umwelt und zum Funktionieren des Handels mit Waren und Leistungen bei. Die im Jahr 2021 gestartete Initiative QI-Digital, bestehend aus den Partnern BAM, DAkkS, DIN, DKE und PTB, entwickelt gemeinsam mit Netzwerkpartnern aus Wissenschaft und Wirtschaft sowie anderen QI-Akteur*innen ein Set an Handlungsfeldern und erarbeitet praktische Lösungen für konkrete Fallbeispiele wirtschaftspolitisch bedeutender Technologien und Innovationen. Dazu wird ein umfassendes QI-Digital-Innovationsökosystem geschaffen, das die Grundlage und den Rahmen bildet für die Entwicklung und Etablierung praxisnaher Lösungen. Beispielhaft für das angestrebte QI-Digital-Innovationsökosystem arbeitet die Initiative QI-Digital an ganz konkreten Projekten. KI in der Medizintechnik, Additive Fertigung und moderne Wasserstoffanwendungen sind drei Innovationsfelder, in denen mit Testfeldumgebungen begonnen wurde. Für diese Zukunftstechnologien sind Qualität und Sicherheit – und das daraus entstehende Vertrauen aller Stakeholder – erfolgsentscheidend.

3.3.2 Umsetzungsprojekte der Normungsroadmap KI

Neben den klassischen Forschungsprojekten widmen sich die Normungsinstitute mit verschiedenen Partner*innen Projekten zur Umsetzung der Handlungsempfehlungen aus der Normungsroadmap Künstliche Intelligenz. Die Umsetzungsprojekte betrachten anwendungstypische und branchenrelevante Use Cases, die für KI-spezifische Anwendungen Anforderungen an die Normung und Standardisierung

71 https://www.zvki.de/

72 https://kidd-prozess.de/

73 https://tu-dresden.de/tu-dresden/newsportal/news/zertifizierung-medizinischer-software-mit-ki-grundlegend-verbessern

74 https://www.din.de/de/din-und-seine-partner/presse/mitteilungen/qi-digital-792188

aufzeigen. Mithilfe dieser Vorhaben sollen im jeweiligen Anwendungskontext praktische Erfahrungen gesammelt, konkrete Normungs- und Standardisierungsbedarfe abgeleitet und Erkenntnisse zur Qualitäts- und Konformitätsprüfung gewonnen werden. Unter den Umsetzungsprojekten benennt die Koordinierungsgruppe „KI-Normung und Konformität“ anhand festgelegter Kriterien sogenannte „Leuchtturmprojekte“. Ihnen kommt eine besondere Bedeutung bei der Umsetzung der Normungsroadmap KI zu, weshalb sie eine erhöhte Aufmerksamkeit bei den Normungsakteur*innen genießen und in Wirtschaft, Forschung und Politik weithin sichtbar sind.

ZERTIFIZIERTE KI

Das Projekt ZERTIFIZIERTE KI[75] stellt ein Umsetzungsprojekt der 1. Ausgabe der Normungsroadmap KI dar. Ziel des im Jahr 2021 gestarteten Projekts ist es, Prüfkriterien, -methoden und -werkzeuge für KI-Systeme zu entwickeln und zu standardisieren und so eine vergleichbare Bewertung von KI-Systemen zu ermöglichen. Durch eine Überprüfbarkeit von technisch zugesicherten Eigenschaften soll das Vertrauen von Anwender*innen und Verbraucher*innen in KI-Technologien gesteigert werden. In branchen- und technologiebezogenen Anwenderkreisen werden die Beteiligten aus Wirtschaft und Industrie sowie Wissenschaft konkrete Bedarfe definieren, Kriterien und Maßstäbe für eine Prüfung in der Praxis festlegen und anhand von Use Cases verifizieren. Mit einem breiten Beteiligungsprozess soll sichergestellt werden, dass sich die Verfahren zu allgemein akzeptierten Standards für KI-Systeme und deren Überprüfung entwickeln und gleichzeitig rechtliche, ethische und philosophische Betrachtungen Berücksichtigung finden.

safetr.AIn

Das Projekt safe.trAIn[76] (Sichere KI am Beispiel fahrerloser Regionalzug) ist das erste offizielle Leuchtturmprojekt[77] der Normungsroadmap KI. Es wird vom BMWK gefördert und verfolgt seit 2022 das Ziel, KI-Verfahren mit den Anforderungen und Zulassungsprozessen im Bahnumfeld praktikabel zu verknüpfen. Der Fokus des Konsortiums, bestehend aus Schienenindustrie, Technologiezulieferern, Forschungseinrichtungen sowie Normungs- und Prüforganisationen, liegt dabei auf der Entwicklung standardisierter Prüfmethoden und -werkzeuge, um die zulassungsrelevante Produktsicherheit für einen breiten Einsatz vollautonomer Züge zu gewährleisten. Außerdem wird die Sicherheitsarchitektur am Beispiel des fahrerlosen Regionalzugs konkretisiert und ein vollautomatisiertes GoA4-System für diesen Anwendungsfall in einem virtuellen Testfeld konzeptionell entwickelt und validiert. Normen und Standards spielen eine entscheidende Rolle für eine beschleunigte Markteinführung und die sichere, robuste und vertrauenswürdige Anwendung KI-basierter Methoden für den führerlosen Zugverkehr.

KI-Tauglichkeit von Normen

Das Projekt KI-Tauglichkeit von Normen[78] verfolgt seit 2022 unter der Leitung von DIN insbesondere das Ziel, den inhaltlichen Bezug relevanter Normen zu Künstlicher Intelligenz zu identifizieren und zu beschreiben. KI-Technologien kommen schon heute in nahezu allen Fachbereichen zum Einsatz – auch in solchen, in denen Normen Anwendung finden, ohne dass sie dafür konzipiert wurden. Damit der Fortschritt von KI-Technologien in allen Fachbereichen ermöglicht wird, ist eine Analyse des gesamten Normenwerks und ggf. eine Anpassung relevanter Normen notwendig. Im Projekt wird eine skalierbare Methodik zur Analyse des Normenwerks hinsichtlich etwaiger Berührungspunkte zu KI-Technologien in der Praxis erarbeitet. Ergänzend dazu wird ein softwaregestütztes KI-Tool entwickelt, das bei dieser Analyse zukünftig unterstützen soll, relevante Normen zu identifizieren. Darüber hinaus wird die Entwicklung der Maschinenausführbarkeit von Normen (SMART Standards, siehe Kapitel 5.3) unterstützt, indem Anforderungen von KI-Systemen an die Struktur von Normungsdokumenten erarbeitet werden. Nähere Informationen zum Projekt sind in Kapitel 5.1 aufgeführt.

75 https://www.din.de/de/vertrauen-in-ki-staerken-mit-qualitaetskriterien-und-pruefverfahren--791046

76 https://www.din.de/de/forschung-und-innovation/partner-in-forschungsprojekten/ki/safe-train-860442

77 Siehe Kapitel 6.6

78 https://www.din.de/de/forschung-und-innovation/themen/kuenstliche-intelligenz/projekte-zu-ki-und-normung/ki-tauglichkeit-von-normen/ki-tauglichkeit-von-normen-872324

4 Schwerpunktthemen

Künstliche Intelligenz ist eine Querschnittstechnologie, die bereits heute in diversen Bereichen zum Einsatz kommt und damit nahezu die gesamte Wirtschaft und Gesellschaft beeinflusst. Umfang und Komplexität dieses Themas lassen es nicht zu, alle Bereiche im Rahmen der vorliegenden Normungsroadmap zu betrachten. Es wird daher gezielt eine Fokussierung und Strukturierung nach horizontalen Themen sowie betroffenen Wirtschafts- und Anwendungsbereichen vorgenommen. Abbildung 15 zeigt die Schwerpunktthemen der Normungsroadmap und den Aufbau des vorliegenden Kapitels.

Durch neue technische Entwicklungen werden insbesondere in der Anwendung von KI Fragestellungen zu übergreifenden Themen aufgeworfen.

Den Ausgangspunkt bilden hierbei die **Grundlagenthemen** wie beispielsweise Terminologien (Begriffsbestimmungen), KI-Klassifizierungen und Ethik. Sie sind die Basis für sämtliche Diskussionen zu KI und stellen damit eines der horizontalen Schwerpunktthemen dar (siehe Kapitel 4.1).

Der Aspekt der **Sicherheit** gewinnt im Kontext von KI zunehmend an Bedeutung – sowohl im Sinne des Schutzes vor äußeren Angriffen (Security) als auch der Fehlerfreiheit bzw. Betriebssicherheit (Safety). Nur eine tiefergehende Betrachtung der Sicherheit von KI-basierten Technologien und Anwendungen kann ihren umfassenden Einsatz in Wirtschaft und Gesellschaft ermöglichen (siehe Kapitel 4.2).

Eine weitere Schlüsselvoraussetzung für den breiten Einsatz von KI-Systemen stellen die **Prüfung und Zertifizierung dar.** Sie können maßgeblich dazu beitragen, das Vertrauen in KI-Systeme zu stärken und Akzeptanz zu schaffen. Kapitel 4.3 gibt Einblicke in den aktuellen Diskussionsstand zur Beurteilung der Qualität von KI-Anwendungen.

Als letztes horizontales Thema werden die **Soziotechnischen Systeme** betrachtet. Dabei steht insbesondere die Mensch-Technik-Schnittstelle im Fokus. Wichtige Fragestellungen sind die Integration der KI-Technologie in gesellschaftliche Subsysteme, die Mensch-Technik-Interaktion sowie die Organisationsentwicklung (siehe Kapitel 4.4).

Die Wirtschafts- und Anwendungsbereiche von KI sind äußerst vielfältig. In allen Bereichen bieten KI-Technologien großes Potenzial.

Neben den übergreifenden Themen wird in der vorliegenden Ausgabe der Normungsroadmap KI der Fokus auf die fünf Anwendungsfelder **Industrielle Automation, Mobilität, Medizin, Finanzdienstleistungen sowie Energie/Umwelt** gelegt, die ein möglichst breites und diverses Spektrum an Anwendungen abdecken (siehe Kapitel 4.5 bis Kapitel 4.9).

Im Nachfolgenden werden zu den neun Schwerpunktthemen der Roadmap die Ausgangssituation, Anforderungen und Herausforderungen sowie konkrete Normungs- und Standardisierungsbedarfe herausgearbeitet.

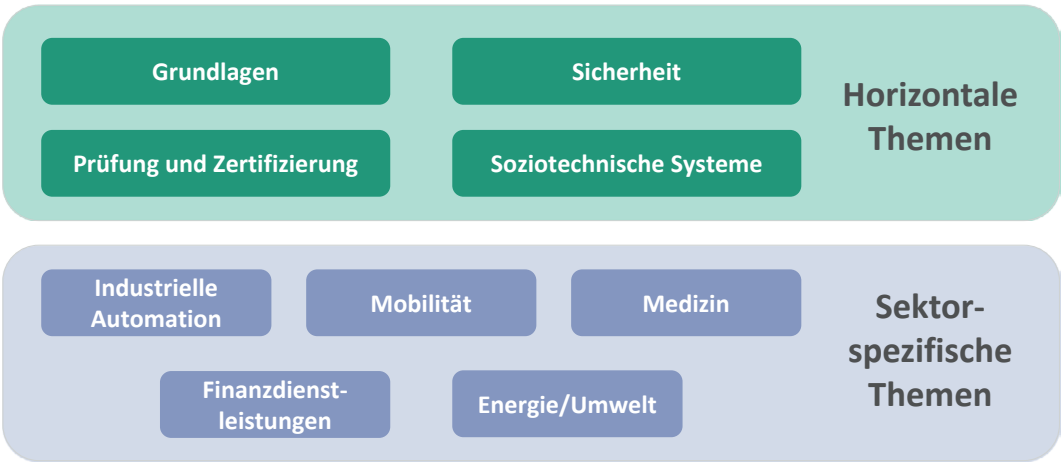
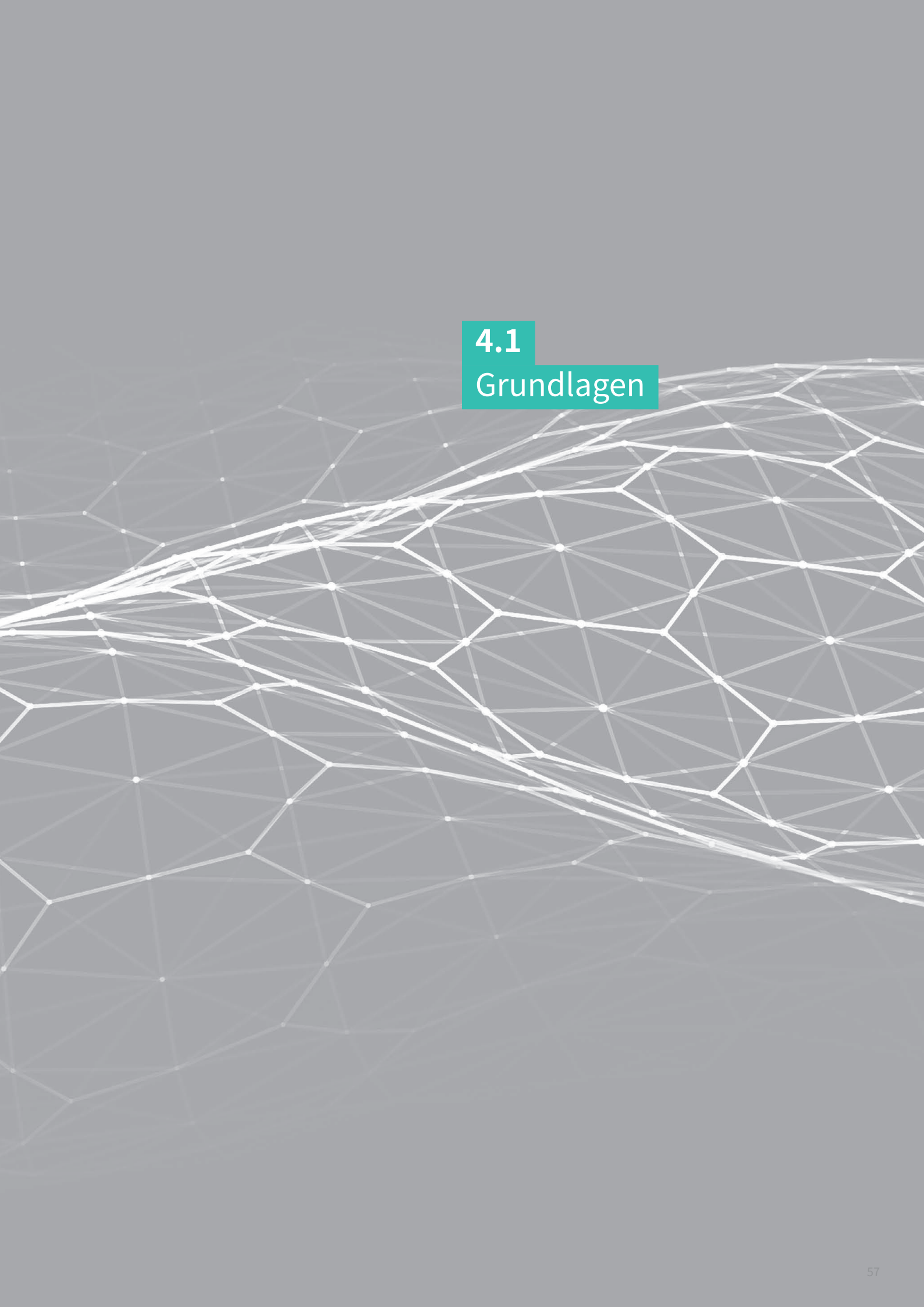


Abbildung 15: Übersichtsgrafik zu den Schwerpunktthemen (Quelle: DIN)



4.1 Grundlagen

KI ist ein Querschnittsthema, das viele Disziplinen tangiert, von denen einige in den Kapiteln 4.2 bis 4.9 betrachtet werden. Für ein grundlegendes Verständnis wurde in Kapitel 1.5 bereits die Begriffsbestimmung für das vorliegende Dokument vorgenommen, ergänzend hierzu werden übergeordnete Themen im nachfolgenden Kapitel behandelt.

4.1.1 Status quo

Im Bereich der KI-Grundlagen gibt es bereits zahlreiche Aktivitäten im Normungsumfeld. Die wichtigsten Gremien hierzu wurden bereits in Kapitel 3.2 vorgestellt, besonders hervorzuheben sind hierbei die Arbeiten des ISO/IEC JTC 1/SC 42 [14], das durch den DIN/DKE Gemeinschaftsausschuss NA 043-01-42 GA Künstliche Intelligenz gespiegelt wird. Eine Auswahl der bedeutendsten Projekte ist weiter unten aufgeführt. Weiterführend bietet Kapitel 4.1.1.1 einen aktuellen Stand zur Klassifizierung von KI.

Titel	Inhalt	Status
ISO/IEC 22989:2022, Artificial intelligence – Concepts and terminology [16]	Konzepte und Terminologie der Künstlichen Intelligenz	Veröffentlicht
ISO/IEC 23053:2022, Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) [24]	Begriffliches Rahmenwerk für Maschinelles Lernen	Veröffentlicht

Zum Thema Management von KI-Systemen sind die folgenden Arbeiten zu nennen:

Titel	Inhalt	Status
ISO/IEC 23894:2022, Information Technology – Artificial Intelligence – Guidance on risk management [25]	Richtlinien für das Risikomanagement zur Entwicklung und Nutzung von KI-Systemen. Auch diese Norm wird unter Leitung eines deutschen Editors entwickelt.	In Entwicklung, Veröffentlichung Ende 2022
ISO/IEC 38507:2022, Information technology – Governance of IT – Governance implications of the use of artificial intelligence by organizations [26]	Organisatorische Governance im Zusammenhang mit KI	Veröffentlicht
ISO/IEC 42001, Information Technology – Artificial Intelligence – Management System [27]	Zertifizierbarer Managementstandard für KI, der Anforderungen und Organisationen zur verantwortlichen Entwicklung und Nutzung von KI-Systemen enthält.	In Entwicklung, Veröffentlichung Mitte 2023
ISO/IEC 42005, Information Technology – Artificial Intelligence – AI System impact assessment [432]	Folgeabschätzung für den Einsatz von KI-Systemen	Initiiert, Veröffentlichung voraussichtlich 2025
Information technology – Artificial intelligence – Requirements for bodies providing audit and certification of artificial intelligence management systems	Anforderungen and Zertifizierungsstellen	Initiiert, Veröffentlichung voraussichtlich 2024

Das Thema Ethik (soweit nicht bereits durch die o. g. Dokumente behandelt) wird durch die folgenden Arbeiten adressiert:

Titel	Inhalt	Status
ISO/IEC TR 24028:2020, Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence [28]	Überblick zum Thema Vertrauenswürdigkeit von KI-Systemen	Veröffentlicht
ISO/IEC TR:24368:2022, Information technology – Artificial intelligence – Overview of ethical and societal concerns [15]	Überblick zu ethischen Themen mit Bezug auf das Arbeitsprogramm des SC 42	In Entwicklung, Veröffentlichung voraussichtlich Mitte 2023

Die Entwicklung von KI-Systemen und systemspezifische Aspekte ihrer Nutzung und Evaluierung sind Gegenstand der folgenden Normungsprojekte:

Titel	Inhalt	Status
ISO/IEC TS 4213 **Information technology – Artificial intelligence – Assessment of machine learning classification performance** [29]	Metriken zur Performanz Maschinellen Lernens	TBD
ISO/IEC 5338 **Information technology – Artificial intelligence – AI system life cycle processes** [30]	Lebenszyklusprozesse, basierend auf dem Lebenszyklusmodell der ISO/IEC 22989:2022 [16]	TBD
ISO/IEC 5339 **Information Technology – Artificial Intelligence – Guidelines for AI applications** [31]	Empfehlung zur Anwendung von KI-Systemen (adressiert teilweise auch ethische Aspekte)	TBD
ISO/IEC 5392 **Information technology – Artificial intelligence – Reference architecture of knowledge engineering** [32]	Referenzarchitektur für symbolische KI-Systeme	TBD
ISO/IEC TR 5469 **Artificial intelligence – Functional safety and AI systems** [33]	Überblick zur funktionalen Sicherheit von KI-Systemen	TBD
ISO/IEC TS 5471 **Artificial intelligence – Quality evaluation guidelines for AI systems** [34]	Empfehlungen zur Qualitätsevaluierung von KI-Systemen, basierend auf dem SQuaRE-Modell	TBD
ISO/IEC 25059:2022 [35] **Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality model for AI systems**	Anforderungen zum Qualitätsevaluierung von KI-Systemen, basierend auf dem SQuaRE-Modell	TBD
ISO/IEC TS 6254 **Information technology – Artificial intelligence – Objectives and approaches for explainability of ML models and AI systems** [36]	Überblick und Empfehlungen zum Umgang mit der Erklärbarkeit von KI-Systemen	TBD

Titel	Inhalt	Status
ISO/IEC TS 8200 Information technology – Artificial intelligence – Controllability of automated artificial intelligence systems [37]	Überblick und Empfehlungen zum Umgang mit Kontrollierbarkeit von KI-Systemen	TBD
ISO/IEC TS 12791 Information technology – Artificial intelligence – Treatment of unwanted bias in classification and regression machine learning tasks [38]	Empfehlungen zur Vermeidung von unerwünschter Vorurteilsbehaftung für Klassifikation und Regression	TBD
ISO/IEC 12792 [238] Information technology – Artificial intelligence – Transparency taxonomy of AI systems	Empfehlungen zur Dokumentation von Transparenzanforderungen an KI-Systeme	TBD

Datenqualitätsmanagement wird in den folgenden Normen behandelt:

Titel	Inhalt	Status
ISO/IEC 5259-1 Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 1: Overview, terminology, and examples [40]	Teil der der Normenreihe zum Datenqualitätsmanagement: Teil 1 beschreibt Terminologie und Konzepte, die in den weiteren Teilen verwendet werden.	TBD
ISO/IEC 5259-2 Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 2: Data quality measures [41]	Teil der Normenreihe zum Datenqualitätsmanagement: Teil 2 behandelt Qualitätsmaße.	TBD
ISO/IEC 5259-3 Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 3: Data quality management requirements and guidelines [42]	Teil der Normenreihe zum Datenqualitätsmanagement: Teil 3 adressiert Anforderungen.	TBD
ISO/IEC 5259-4 Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 4: Data quality process framework [43]	Teil der Normenreihe zum Datenqualitätsmanagement: Teil 4 beschreibt Prozesse, die zur Erfüllung der Anforderungen in Teil 3 implementiert werden können.	TBD
ISO/IEC 5259-5 Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 5: Data quality governance [44]	Teil der Normenreihe zum Datenqualitätsmanagement: Teil 5 behandelt das Thema Governance von Daten.	TBD
ISO/IEC 8183 Information technology – Artificial intelligence – Data life cycle framework [45]	Teil der Normenreihe zum Datenqualitätsmanagement (zu beachten ist die abweichende Nummerierung): Dieser Teil behandelt den Datenlebenszyklus als Ergänzung zum Teil 1 der ISO/IEC-5259Reihe [39].	TBD

4.1.1.1 KI-Klassifizierung

In Anlehnung an das Positionspapier „A definition of AI: Main capabilities and scientific disciplines“ der AI HLEG [46] wird zwischen Methoden und Fähigkeiten der KI unterschieden. In beiden Fällen basieren die folgenden Klassifizierungen auf einer aktuellen Übersichtsarbeit [47], die im Umfeld der ersten Version der KI-Normungsroadmap entstanden ist und den derzeitigen Stand der Technik widerspiegelt. Der Abschnitt „Klassifikation von KI-Methoden“ beschreibt, welche KI-Methoden zur Realisierung bestimmter KI-Fähigkeiten eingesetzt werden. Der Abschnitt „Klassifikation von KI-Fähigkeiten“ beschreibt grundsätzliche Fähigkeiten von KI-Systemen. In Kombination mit einer Kritikalitäts- oder Risikobewertung wird durch dieses Klassifikationsschema eine ganzheitliche Charakterisierung eines KI-Systems (Abbildung 16) möglich. Um auch den tatsächlichen Stand der aktuellen industriellen KI-Märkte adäquat abzubilden, wird zusätzlich eine Klassifizierung von KI-Anwendungen vorgenommen, die sich aus KI-Methoden und KI-Fähigkeiten ergeben. Weiterführende Informationen und Beispiele finden sich in dem kürzlich veröffentlichten Beuth Pocket [48].

KLASSIFIKATION VON KI-METHODEN

Die heutige KI basiert auf einer Vielzahl unterschiedlicher Methoden. Aufgrund der historischen Entwicklungen wird oft grob zwischen symbolischen und subsymbolischen KI-Methoden unterschieden. Beide Paradigmen bilden in der Praxis die Grundlage einer großen Anzahl von KI-Anwendungen ([49], 79–88). Auf der Seite der symbolischen Methoden stehen Techniken der Wissensrepräsentation und des logischen Schließens im Vordergrund, während die Seite der subsymbolischen Methoden vor allem durch Techniken des Maschinellen Lernens und neuronale Netze vertreten werden. Diese traditionelle Unterscheidung ist jedoch nicht umfassend. Sie vernachlässigt klassische Methoden der Künstlichen Intelligenz wie Problemlösung, Optimierung, Planung und Entscheidungsfindung. Darüber hinaus haben die Entwicklungen der letzten Jahrzehnte die traditionellen Grenzen weiter verwischt und es treten immer mehr kombinierte oder hybride Ansätze in den Vordergrund, z. B. das gesamte Feld des hybriden Lernens ([50], S. 77; [49]).

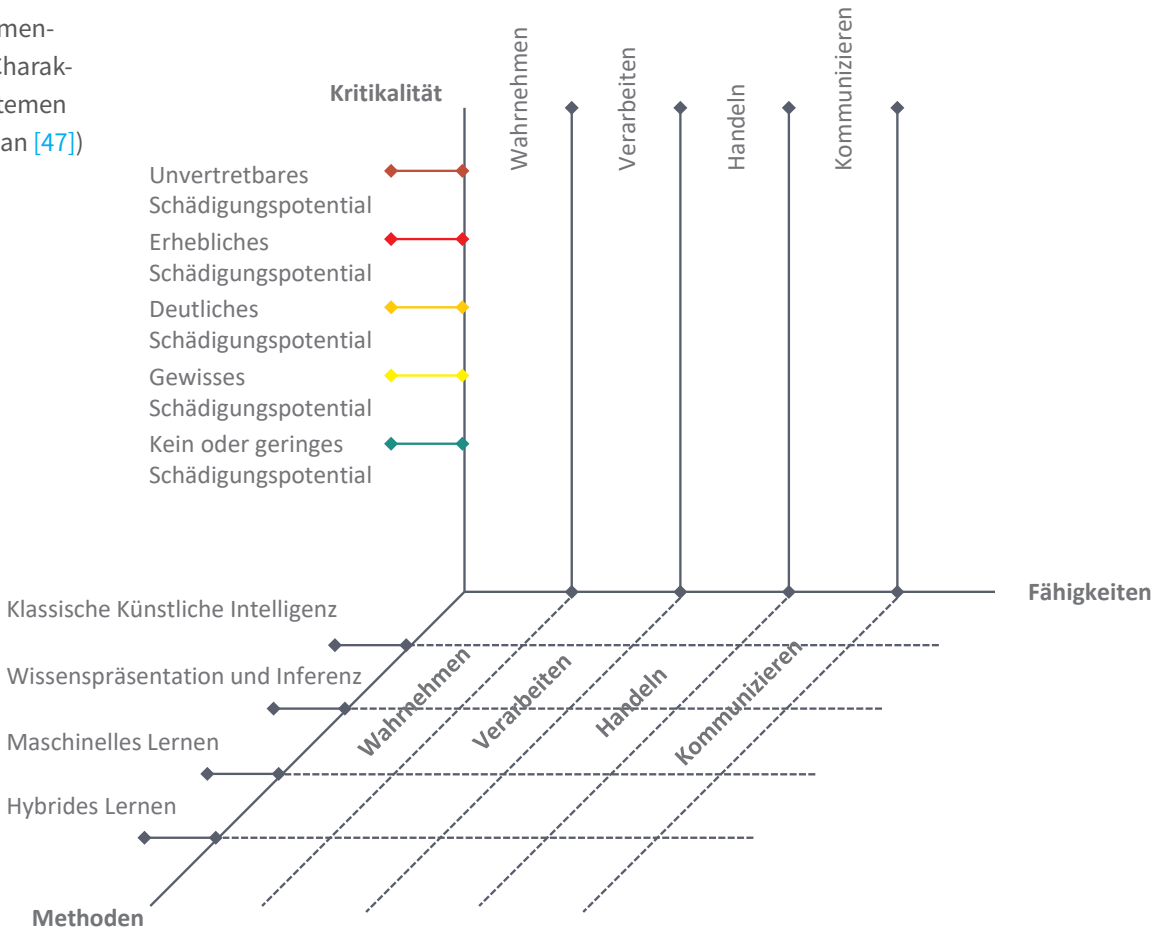


Abbildung 16: Dreidimensionales Schema zur Charakterisierung von KI-Systemen (Quelle: in Anlehnung an [47])

Tabelle 2 und Tabelle 3 geben einen Überblick über diese Methoden auf drei Granularitätsebenen (Felder der KI, Disziplinen der KI sowie Teildisziplinen) und nennen bekannte Vertreter der jeweiligen Technik. Bei diesem Schema handelt es sich um eine vorläufige Momentaufnahme, die durch neu aufkommende KI-Methoden in der Zukunft ergänzt werden kann. Auch ist es oft nicht möglich, eine unumstößliche Trennung zwischen den Kategorien zu ziehen, da manche Methoden mehreren Kategorien angehören können. In solchen Fällen wurden die Methoden nach Russell und Norvig [51] oder nach der Kategorie zugeordnet, für die sie ursprünglich vorgeschlagen wurden.

(A) Klassische Künstliche Intelligenz
Historisch gesehen gehören Ansätze zur Problemlösung, Optimierung, Planung und Entscheidungsfindung zu den frühesten KI-Methoden, die entwickelt wurden. Das Problemlösen beschreibt zielorientierte Suchstrategien und intelligente Agenten, die Probleme lösen, indem sie ein Ziel formulieren und mit einem definierten Problem als Input nach der richtigen Abfolge von Aktionen suchen, um die Lösung auszuführen und das Ziel zu erreichen. In konkurrierenden Multi-Agenten-Umgebungen, in denen die Ziele miteinander in Konflikt stehen, werden zur Lösung komplexer Probleme die kontradiktorische und die beschränkungsbasierte Suche eingesetzt.

Im Gegensatz zu Problemlösungsmethoden, die Suchräume systematisch erkunden, kümmern sich Optimierungsalgorithmen nicht um den Weg zum Ziel, sondern konzentrieren sich auf die optimale Lösung. Sie lassen sich in deterministische Ansätze wie Simplex-Verfahren, Netzalgorithmen, Entscheidungsbäume (einschließlich Branch-and-Bound-Verfahren) und klassische Gradientenabstiegsverfahren und in nicht-deterministische Ansätze unterteilen. Beispiele für nicht-deterministische Optimierungsmethoden sind genetische Algorithmen, Schwarmintelligenz und die sogenannte simulierte Abkühlung.

Bei den Planungsmethoden kann es sich um autonome oder halbautonome Techniken handeln, wie z. B. Steady-State-Search, Planungsgraphen, hierarchische Planung, nicht-deterministische Planung, Zeit- und Ressourcenplanung und Plangenerierung. Im Gegensatz zur Planung müssen Planerkennungsmodelle oder -methoden wie die deduktive und synthetische Planerkennung, die bibliotheksbasierte Planerkennung und die Planung durch abduktives Schlussfolgern tatsächliche Ereignisse oder Handlungen darstellen, die stattgefunden haben, und hypothetische Erklärungen vorschlagen. Planungsmethoden spielen in der Robotik, in Dialogsystemen und in der Mensch-Maschine-Interaktion eine Rolle.

Die Entscheidungsfindung oder Entscheidungsanalyse ist eine technische Disziplin, die sich mit der pragmatischen Anwendung der Entscheidungstheorie auf bestimmte Probleme befasst ([52], 247–302). Es gibt verschiedene Ansätze für die Entscheidungsfindung wie Prozessmodelle, Informationswert, Entscheidungsnetzwerke, Expertensysteme, sequenzielle Entscheidungsfindung und Iterationsmodelle.

(B) Symbolische Künstliche Intelligenz
Symbolische KI-Methoden zeichnen sich durch einen deduktiven Ansatz aus, d. h. durch die algorithmische Anwendung von logischen Regeln oder Beziehungen auf einzelne Fälle. Kernkonzepte der symbolischen KI sind zum einen Techniken zur Repräsentation von Wissen und zum anderen Methoden zur Anwendung dieses Wissens auf eine gegebene Eingabe. Wissen kann entweder als sicheres oder als unsicheres Wissen dargestellt werden. Mithilfe von Argumentationsketten können aus diesem Wissen Schlussfolgerungen gezogen werden.

Die formale Wissensrepräsentation umfasst Konzepte wie Ontologien, semantische Netze, Wissensgraphen und Wissenskarten, die Informationen in Strukturen, Syntax, Semantik und Semiotik zusammenfassen und systematisieren. Der Fokus von standardisierten Beschreibungssprachen wie dem Resource Description Framework (RDF) und der W3C Web Ontology Language (OWL) liegt auf der Erstellung von eindeutigen Spezifikationen für Objekte, Merkmale und allgemeine Konzepte durch logische Beziehungen. Mithilfe dieser und weiterer semantischer Webstandards und -technologien können logisch verwandte Daten über Domänen verschiedener Anwendungen hinweg gemeinsam genutzt werden, was die semantische Interoperabilität erleichtert. Im Allgemeinen basieren die grundlegenden Konzepte des Ontology Engineering auf Taxonomie, Kalkül, Deduktion, Abduktion und Verarbeitung und Modellierung von Ontologien. Darüber hinaus können logische Beziehungen und Abstraktionen von Domänen durch Wissensgraphen, semantische Netze und Wissensabbildung hergestellt werden. Im Falle einer graphenbasierten Abstraktion von Wissen bieten Algorithmen zur Durchquerung von Graphen gemeinsame Lösungen für Probleme bei der Suche, Überprüfung und Aktualisierung von Eckpunkten. Darüber hinaus können logisch verwandte Daten durch Aussagen- oder Prädikatenlogik, Logiken hoher Ordnung, nicht-monotone, temporale und modale Logiken modelliert werden.

Bei bestimmten Kenntnissen wird die Anwendung von formalem Wissen häufig mithilfe der klassischen Methoden des logischen Schließens operationalisiert. Insbesondere Erfüllbarkeit und andere Techniken der formalen Verifikation können dabei angewendet werden. Für das Schließen auf der Basis von unsicherem Wissen sind probabilistische Ansätze weitverbreitet, aber auch nicht-probabilistische Ansätze wurden vorgeschlagen. Beim probabilistischen Schlussfolgern können Informationen durch Abtasten einer Wissensbasis abgeleitet und durch relationale probabilistische Modelle oder das Konzept der Bayes'schen Inferenz verarbeitet werden. In Bezug auf unsicheres Wissen dominiert die Bayes'sche Regel den KI-Bereich bei der Quantifizierung von Unsicherheit. Nicht-probabilistisches Schließen kann für mehrdeutige Informationen im Falle von Vagheit unter Berücksichtigung von Beweisen angewendet werden. In diesen Situationen können ein Wahrheitsmanagementsystem und Schlussfolgerungen mit Standardinformationen für qualitative Ansätze verwendet werden. Darüber hinaus können Methoden des nicht-probabilistischen Schlussfolgerns mithilfe von regelbasierten Ansätzen oder Fuzzy Sets implementiert werden. Einen weiteren gängigen Ansatz für nicht-probabilistisches Schlussfolgern stellt das sogenannte Schlussfolgern mit Glaubensfunktion dar, bei dem alle verfügbaren Beweise für die Berechnung eines Glaubensgrades kombiniert werden. Andere Ansätze für unsicheres Schließen umfassen räumliches Schließen, fallbasiertes Schließen, qualitatives Schließen und psychologisches Schließen.

(C) Maschinelles Lernen

Im Gegensatz zur symbolischen KI zeichnen sich subsymbolische KI-Methoden durch ein induktives Vorgehen aus, d. h. durch die algorithmische Ableitung von allgemeinen Regeln oder Beziehungen aus Einzelfällen. Zu diesem Zweck werden typischerweise zwei große Ansätze des Maschinellen Lernens unterschieden: das überwachte Lernen mit vorgegebenen Zielparametern und das unüberwachte Lernen, bei dem diese nicht vorgegeben sind. Um diese beiden Hauptansätze herum haben sich auch alternative Lernparadigmen wie teilüberwachtes, bestärkendes oder gegenläufiges Lernen etabliert.

Überwachte Lerntechniken werden in der Regel zur Durchführung von Regressions- oder Klassifizierungsaufgaben verwendet. Praktische Anwendungen des überwachten Lernens werden seit Langem von diskriminativen Verfahren wie logistischer Regression, Entscheidungsbäumen oder neuronalen Netzen dominiert. Neuronale Netze gelten dabei als besonders flexibel, da sie theoretisch jede beliebige mathematische Funktion ohne jegliches Vorwissen erlernen können. Auch Support-Vector-Maschinen werden trotz der notwendigen, aber nicht leicht zu bestimmenden Kernel-Funktion in vielen Anwendungen erfolgreich eingesetzt.

Einige überwachte Lernalgorithmen wie Naive Bayes oder Hidden Markov Models erzeugen eine geschätzte Wahrscheinlichkeitsverteilung für Eingabe- und Ausgabevariablen. Auch wenn diese weithin zu den überwachten Lerntechniken gezählt werden, kann man hier genauer von den generativen Lernverfahren sprechen. Neuere Beispiele für generative Verfahren sind Techniken wie Generative Adversarial Networks, die jedoch einem gegenläufigen Lernparadigma folgen. Dieses Paradigma stammt ursprünglich aus dem Anwendungsbereich der Bildverarbeitung und stellt eine Weiterentwicklung des klassischen überwachten Lernens dar, das darauf beruht, dass zwei maschinelle Lernverfahren gegeneinander operieren, um die Eigenschaften eines gegebenen Datensatzes nachzubilden.

Unüberwachte Lernmethoden hingegen werden in der Regel für Clustering- oder Dimensionsreduktionsaufgaben eingesetzt. Einer der ältesten und bekanntesten Algorithmen ist dabei das Clustering-Verfahren k-means. Neben anderen statistisch motivierten Methoden wie dem hierarchischen Clustering wurden auch biologisch inspirierte Algorithmen wie die selbstorganisierende Karte von Kohonen oder die adaptive Resonanztheorie von Grossberg vorgeschlagen. Auch für Regressionsaufgaben gibt es unüberwachte Verfahren, die hauptsächlich zur Dimensionsreduktion eingesetzt werden.

Nicht alle Lernalgorithmen lassen sich eindeutig als überwacht oder unüberwacht klassifizieren. So kann z. B. ein mehrschichtiges Perzeptron, d. h. ein überwachtes Verfahren, verwendet werden, um einen gegebenen Datensatz auf sich selbst abzubilden. Entfernt man anschließend die Ausgabeschicht eines solchen Autoencoders, bleibt ein Teilnetz übrig, das eine Dimensionsreduktion des Datensatzes entsprechend der Anzahl der versteckten Neuronen durchführt. Ein solcher Einsatz überwachter Lernverfahren zur gezielten Erzeugung spezifischer Repräsentationen in einzelnen Schichten von Neuronen ist z. B. eine wichtige Grundlage des sogenannten Deep Learning. Ein weiteres Beispiel für Zwischenformen des Maschinellen Lernens sind Algorithmen des sogenannten teilüberwachten Lernens, bei denen überwachtes und unüberwachtes Lernen kombiniert werden und somit nur für einen Teil der verwendeten Daten ein vordefinierter Zielwert erforderlich ist. Dies ermöglicht nicht nur die Analyse unvollständiger Datensätze, sondern erzielt in manchen Fällen sogar bessere Ergebnisse als klassische überwachte Lernver-

fahren. Für teilüberwachte Lernalgorithmen müssen jedoch im Vorfeld Annahmen über Verteilungsdichten getroffen werden. Bei ungünstigen Annahmen können die Ergebnisse deutlich schlechter ausfallen als bei einem überwachten Lernverfahren.

Ein mit dem überwachten Lernen verwandtes, aber alternativ verfahrendes Lernparadigma ist das sogenannte bestärkende Lernen. Dieses benötigt für das Erlernen von Zusammenhängen eine Rückmeldung für getroffene Vorhersagen, nicht jedoch den genauen Zielwert. Analog zum Lernen in Form von klassischer Konditionierung (d. h. über Belohnung oder Strafe) wird bei dieser Methode nur berücksichtigt, ob das angestrebte Lernergebnis erreicht wurde oder nicht. Ein solches Lernen mit Rückmeldung, aber ohne festen Zielwert, hat sich vor allem in den Anwendungsbereichen Robotik und adaptive Steuerung als sehr nützlich erwiesen.

(D) Hybrides Lernen

Methoden des hybriden Lernens zeichnen sich durch die Kombination von Konzepten aus den zuvor vorgestellten Methoden aus, z. B. beim Training neuronaler Netze durch Anwendung genetischer Algorithmen zur Anpassung der Netzgewichte und ggf. der Netzarchitektur. Dieser kombinierte Ansatz wurde für so unterschiedliche Anwendungsbereiche vorgeschlagen wie Finanzanwendungen, ozeanografische Vorhersagen, Vorhersagen von Ambulanzaufenthalten oder die Klassifizierung von Teepflanzen.

Aufgrund der wissenschaftlichen Kreativität in diesem Bereich ist es schwierig bis unmöglich, einen umfassenden Überblick über alle Methoden hybriden Lernens zu geben. Ein großer Teil dieses Bereichs konzentriert sich jedoch auf die Kombination von symbolischer und subsymbolischer KI, um sowohl induktiv als auch deduktiv arbeitende Systeme zu schaffen. Jüngste Forschungsaktivitäten befassen sich z. B. mit der Kombination von Maschinellem Lernen und Knowledge Engineering [53]. Ein prominentes Teilgebiet sind hybride neuronale Systeme, die sich weiter in einheitliche neuronale Architekturen, Transformationsarchitekturen und hybride modulare Architekturen unterteilen lassen ([54], 62–93). Im Gegensatz zu klassischen subsymbolischen Verfahren erlauben solche Verfahren entweder die Extraktion von Regeln oder verwenden eine zusätzliche Form der Wissensrepräsentation. Im Gegensatz zu klassischen symbolischen Methoden werden solche Wissensrepräsentationen jedoch häufig auf der Grundlage gegebener Daten algorithmisch erstellt oder modifiziert.

In einem weiteren Sinne kann man hybrides Lernen auch als Lernen mit Wissen bezeichnen. Zu diesem Zweck wurden in den letzten Jahrzehnten weitere Ansätze vorgeschlagen, z. B. Lernen durch Logik und Deduktion, induktive logische Programmierung, erklärbare KI und relevanzbasiertes Lernen. Neuere Ansätze der hybriden KI sind das sogenannte konversationelle Lernen oder aktive Dialoglernen, welches darauf abzielt, die Leistung des Maschinellen Lernens durch die Einbeziehung von im Dialog gesammeltem menschlichem Wissen zu verbessern.

Tabelle 2: Klassifikation von KI-Methoden

Feld	Disziplin	Teildisziplin	Beispiele
KLASSISCHE KÜNSTLICHE INTELLIGENZ	Problemlösen	Agenten & Suchstrategien	Uninformierte & informierte Suchstrategien
			Gegenläufige Suche (Spieltheorie)
			Suche unter Randbedingungen
	Optimierung	Deterministisch	Simplex-Methoden
			Netzwerkalgorithmen
			Entscheidungsbäume (z. B. Branch & Bound)
			Gradientenabstiegsverfahren
		Nicht-deterministisch	Evolutionäre Algorithmen
			Genetische Algorithmen/Programmierung
			Schwarmintelligenz
			Simulierte Abkühlung
	Planen & Planerkennung	Autonomes & teilautonomes Planen	Steady State Search
			Planungsgraphen
			Hierarchisches Planen
			Nicht-deterministisches Planen
			Zeit- & Ressourcenplanung
			Plangenerierung
		Planerkennung	Abduktive Planerkennung
			Deduktive Planerkennung
			Bibliothekbasierte Planerkennung
			Synthese-Planerkennung
	Entscheidungs-findung	Ansätze zur Entschei-dungsfindung	Prozessmodelle
			Informationswert
			Entscheidungsnetzwerke
			Expertensysteme
			Sequenzielle Entscheidungsfindung
			Iteration Models

Feld	Disziplin	Teildisziplin	Beispiele
SYMBOLISCHE KÜNSTLICHE INTELLIGENZ	Wissens-repräsentation	Ontologien	RDF, RDFS und OWL
			Taxonomien
			Interpretation
			Calculus
			Deduktion
			Abduktion
			Ontologie-Mapping
		Wissensgraphen & semantische Netzwerke	Wissensgraphen & -netzwerke
			Existenzgraphen
			Graph-Traversal-Algorithmen
			Mapping
			Semantic Web
		Modellierung mittels formaler Logik	Propositionale Logik & Prädikatenlogik
			Logiken höherer Ordnungen
			Nicht-monotone Logiken
			Temporal- & Modal-Logiken
		Quantifizierung von Unsicherheit & Repräsentation unsicheren Wissens	Bayes'sche Regel
			Bayes'sche Netzwerke
	Logisches Schließen	Formale Verifikation	Resolution- & Konnektivitätsverifikation
			SAT & SMT (Satisfiability modulo theories) Solver
			Model Checking
		Interaktive Verifikation	Tactical Theorem Verification

Feld	Disziplin	Teildisziplin	Beispiele
	Probabilistisches Schließen	Bayes'sches Schließen	Präzise Inferenz
			Näherungsweise Inferenz
			Markov-Ketten
		Relationale probabilistische Modelle	Relationale probabilistische Modelle in geschlossenen & offenen Universen
		Probabilistisches Schließen mit Zeit & Unsicherheit	Hidden-Markov-Modelle
			Kalman-Filter
			Dynamische Bayes'sche Netzwerke
	Nicht-probabilistisches Schließen	Qualitative Ansätze	Schließen mit Standardinformation
			Wahrheitsmanagementsysteme
		Regelbasierte Ansätze	Regelbasiertes Schließen mit Sicherheit
		Schließen mit Unsicherheit	Fuzzy-Mengen & -Logik
		Schließen mit Glaubensfunktion	Dempster-Shafer-Theorie
	Weitere Ansätze für unsicheres Schließen		Räumliches Schließen
			Fallbasiertes Schließen
			Qualitative Physik
			Psychologisches Schließen
MASCHINELLES LERNEN	Überwachtes Lernen	Neuronale Netze	Multi-Layer-Perzeptron
			Learning Vector Quantization (LVQ)
			Radial Basis Networks (RBF)
			Adaptive Resonanztheorie (ART)
			Faltende neuronale Netze (CNN)
			Rekurrente neuronale Netze (RNN)
			Time-Delay-Netze (TDNN)
			Long-Short Term Memory (LSTM)
			Hopfield-Netzwerke
			Boltzmann-Maschine

Feld	Disziplin	Teildisziplin	Beispiele
		Statistisches Lernen	Entscheidungsbäume
			Random Forests
			Stützvektormaschine (SVM)
		Probabilistische Methoden	Naive Bayes
			Fuzzy-Klassifizierer
	Unüberwachtes Lernen	Clustering	k-Means
			Hierarchisches Clustering
			DBSCAN
			Fuzzy Clustering
			Selbstorganisierende Karte
		Dimensionsreduktion	Autoencoder
			Hauptkomponentenanalyse
		Probabilistische Methoden	Fuzzy c-Means
	Teilüberwachtes Lernen	Statistische Methoden	Expectation-Maximization mit generativen Modellen
			Transduktive Stützvektormaschinen
		Modifizierte Lernkonzepte	Selbstlernen
			Gemeinsames Lernen
		Graphenbasiertes Lernen	Graph-basierte Methoden
	Bestärkendes Lernen	Temporal-Differenz-Lernen	Q-Learning
			State–action–reward–state–action (SARSA)
		Monte-Carlo-Methoden	Markov-Ketten Monte Carlo
		Adaptive Dynamische Programmierung	Aktive & passive adaptive dynamische Programmierung

Feld	Disziplin	Teildisziplin	Beispiele
	Gegenläufiges Lernen	Generative Methoden	Generative Adversarial Networks (GAN)
			Bayes'sche Adversarielle Netze
			Gegenläufige Autoencoder
		One-Shot Learning	Siamesische neuronale Netze
HYBRIDES LERNEN	Hybride neuronale Systeme	Neuronale Einheits-architekturen	Konstruktivistisches Maschinelles Lernen
		Transformations-architekturen	Regelextraktion für neuronale Netze
		Hybride module Architekturen	Neuro-Fuzzy Expertensysteme
	Lernen mit Wissen	Lernen mittels Logik & Schließen	Current-Best-Learning
		Induktive logische Programmierung	Sequential-Covering-Algorithmus
			Konstruktive Induktionsalgorithmen
		Erklärbare künstliche Intelligenz	Local Interpretable Model-agnostic Explanations (LIME)
		Relevanzbasiertes Lernen	
	Konversationelles Lernen	Lernen mittels aktiver Dialog	Überwachtes konversationelles Lernen
			Bestärkendes konversationelles Lernen

KLASSIFIKATION VON KI-FÄHIGKEITEN

Wesentliche Inspiration für die Etablierung der wissenschaftlichen Disziplin Künstliche Intelligenz sind die kognitiven Fähigkeiten des Menschen [51]. Die Psychologie und Kognitionswissenschaften fokussieren sich dabei in der Regel auf Teilaspekte und weniger auf eine gesamtheitliche Betrachtung dieser Fähigkeiten. In Bildungskontexten werden menschliche Fähigkeiten dagegen bereits seit Mitte des vergangenen Jahrhunderts auf der Grundlage sogenannter Lernziele definiert und bewertet. Lernzieltaxonomien wie die von Benjamin S. Bloom klassifizieren menschliche Fähigkeiten und bilden die Grundlage der europäischen Bildungssysteme [55]. Bloom unterscheidet zunächst kognitive, affektive und psychomotorische Fähigkeiten [56], die wiederum in spezifischere Fähigkeiten unterteilt werden können. Auf der Grundlage von Blooms Ideen wurden separate Taxonomien für affektive Fähigkeiten, d. h. Verhalten, das überwiegend von kurzen, impulsartigen Gefühlsregungen und nicht von kognitiven Prozessen bestimmt ist, für psychomotorische Fähigkeiten, also die Kontrolle und Koordination der Muskeln, sowie für kognitive Fähigkeiten postuliert. Diese Dreiteilung wird im Folgenden von den drei allgemeinen Fähigkeiten Wahrnehmen, Verarbeiten und Handeln widergespiegelt (vgl. Tabelle 2).

Gemessen an derartigen Taxonomien realisiert die KI-Technologie der Gegenwart nur eine Teilmenge der kognitiven Fähigkeiten des Menschen. Gleichzeitig ermöglichen viele bestehende KI-Systeme und -Anwendungen zusätzliche Funktionalitäten, wie eine über die menschlichen Sinne

hinausgehende Sensorik oder Interaktion mit der Umwelt (z. B. nichtmenschliche Kommunikation) zu ermöglichen. Auf der Grundlage dieser Beobachtungen lassen sich sowohl die bestehenden als auch die potenziell erreichbaren KI-Fähigkeiten grob in die Bereiche Wahrnehmen, Verarbeiten, Handeln und Kommunizieren unterteilen. Diese Vierteilung von Fähigkeiten berücksichtigt zwar gängige Erkenntnisse aus der psychologischen und pädagogischen Forschung, soll aber primär die Strukturierung von Fähigkeiten ermöglichen, die heute von KI-Systemen umgesetzt werden können. Tabelle 3 gibt einen Überblick über die vorgeschlagene Klassifizierung von KI-Fähigkeiten.

(A) Wahrnehmen

Der Begriff „Wahrnehmung“ bezeichnet klassischerweise Fähigkeiten, die durch die menschlichen Sinnesorgane ermöglicht werden. Die griechischen Philosophen der Antike unterschieden z. B. die fünf Sinne Sehen, Hören, Riechen, Schmecken und Tasten. Die moderne Wissenschaft unterscheidet dagegen zwischen Sinnesorganen, die Sinnesreize weiterleiten und als eine Art Wahrnehmungsvorstufe fungieren, und Sinnesmodalitäten, die im Wesentlichen den Output der Sinnesorgane für die nachfolgende kognitive Verarbeitung beschreiben. Mit Blick auf KI-Systeme ist zu beachten, dass die Vielfalt physikalischer Größen, die von spezialisierten technischen Sensoren wahrgenommen werden können, die Zahl der vom Menschen direkt wahrnehmbaren Reize übersteigt. Technische Sensoren existieren heute für eine breite Palette akustischer, biologischer, chemischer, elektrischer, magnetischer, optischer, mechanischer, strahlender und thermischer Reize.

Um die menschliche Wahrnehmung zu beschreiben, konzentrieren sich viele Wissenschaftler heute auf die sensorischen Modalitäten. Der Begriff Modalität wird häufig verwendet, um die Kodierung oder „Darstellungsweise“ zu beschreiben, die sich aus der Transduktion eines sensorischen Inputs ergibt. Mit zunehmender Popularität einer modalitätsbasierten Perspektive auf Wahrnehmung wurde das klassische Fünf-Sinne-Schema infrage gestellt, und es wurden in den vergangenen Jahrzehnten mehrere alternative Schemata vorgeschlagen. Während die Anzahl der Sinne in solchen alternativen Klassifizierungsschemata variiert (zwischen 8 und 17), scheint ein Konsens zu bestehen, dass zusätzlich zu den klassischen fünf Sinnen, die direkt auf externe Eindrücke ausgerichtet sind, auch interne Sinne wie Körperwahrnehmung und Gleichgewicht zur menschlichen Wahrnehmung beitragen, indem sie z. B. Störungen und Anomalien erkennen ([57], S. 353–370). Auch der Tastsinn wird oft als facettenreicher Sinn betrachtet, der die Wahrnehmung von Temperatur, Druck und Schmerz umfasst [58].

Entsprechend wird vorgeschlagen, die KI-Fähigkeiten im Bereich der Wahrnehmung auf die menschlichen Sinnesmodalitäten abzustimmen. So lässt sich beispielsweise berücksichtigen, dass KI-Forschung in letzter Zeit erhebliche Fortschritte bei der Fähigkeit gemacht hat, Bilder, auditive und haptische Signale in verarbeitbare Informationen umzuwandeln. KI-Anwendungen für die Geruchs- und Geschmackswahrnehmung hingegen wurden zwar untersucht, sind aber in der Praxis noch relativ selten.

(B) Verarbeiten

Die Fähigkeit, Informationen zu verarbeiten, ist eine wesentliche Voraussetzung für intelligentes Verhalten. Um diese Fähigkeit detaillierter zu beschreiben, bietet es sich an, eine bestehende Taxonomie kognitiver Lernziele wie die sogenannte überarbeitete Bloom’sche Taxonomie ([59], S. 212–218) als Klassifizierungsschema zu verwenden. Diese gegenüber der ursprünglichen Taxonomie von Bloom aktualisierte Lernzieltaxonomie unterscheidet die menschlichen Fähigkeiten in einer primären Dimension nach den vier Bereichen der faktischen, konzeptuellen, prozeduralen und metakognitiven Kognition.

Die faktische Kognition als der am wenigsten komplexe Bereich umfasst in diesem Schema Fähigkeiten zur Verarbeitung und zum Verständnis von Terminologien sowie Wissen über spezifische Details und Elemente. Konzeptuelle Kognition umfasst die Fähigkeit, Wissen über Klassifikationen und Kategorien, Prinzipien und Verallgemeinerungen sowie Wissen über Theorien, Modelle und Strukturen zu verarbeiten und zu verstehen. Prozedurale Kognition bezieht sich auf die Verarbeitung von Wissen über fachspezifische Fertigkeiten und Algorithmen, über fachspezifische Techniken und Methoden sowie über Kriterien zur Bestimmung des richtigen Zeitpunkts für die Anwendung geeigneter Verfahren. Metakognitive Kognition umfasst Fähigkeiten zur Verarbeitung und zum Verständnis von strategischem Wissen, Wissen über kognitive Aufgaben (einschließlich kontextbezogenem und konditionalem Wissen) und Selbsterkenntnis.

Auf einer zweiten Ebene werden diese Bereiche durch sechs kognitive Stufen weiter ausdifferenziert. Bei konzeptuellem Wissen werden z. B. die grundlegenden Fähigkeiten zum Erkennen oder Klassifizieren von den mittelkomplexen Fähigkeiten zum Bereitstellen oder Unterscheiden von Informationen und den fortgeschrittenen Fähigkeiten zum Bestimmen

oder Zusammensetzen von Informationen getrennt. Unter Verwendung beider kognitiver Dimensionen erlaubt Blooms überarbeitete Taxonomie die Unterscheidung von bis zu 24 menschlichen kognitiven Fähigkeiten. In Bezug auf die derzeit realisierbaren KI-Fähigkeiten ermöglicht dies insbesondere, die Fähigkeit zur Wiedergabe von Wissen, Entscheidungen zu treffen oder Vorhersagen zu machen abzubilden. Diese Fähigkeiten bilden den Kern vieler fortgeschrittener KI-Systeme und werden oft in Kombination mit Fähigkeiten zur Wahrnehmung, zum Handeln oder zur Kommunikation eingesetzt.

(C) Handeln

Für den Menschen ist die Fähigkeit zu handeln eine grundlegende Fähigkeit. In einem allgemeineren Sinne kann eine Handlung aber sowohl auf einen menschlichen als auch auf einen nicht-menschlichen Akteur*innen bezogen werden. Sie kann als etwas beschrieben werden, das ein/e Akteur*in tut und das „unter einer gewissen Beschreibung intentional“ war [60]. Außerdem kann zwischen physischem und nicht-physischem Handeln unterschieden werden. Häufig wird dies durch die Kombination von mechatronischen und Softwarekomponenten in Robotern oder Softwarerobotern realisiert. Der Bereich der Robotik beschreibt insbesondere mechanisch oder physikalisch ausgeführte Tätigkeiten wie Roboterwahrnehmung, Bewegungsplanung, Sensorik und Manipulatoren, Kinematik und Dynamik sowie den Bereich der Mensch-Roboter-Interaktion, da sich diese Form der Interaktion auf die physische Mensch-Maschine-Interaktion konzentriert. Diese Fähigkeiten orientieren sich grob an den menschlichen Fähigkeiten der Steuerung und Koordination von Muskeln. Die Methoden, die für Softwareagenten verwendet werden, hängen von dem jeweiligen Ziel oder der Aufgabe des Agenten selbst ab. Solche autonomen Agenten sind z. B. im Bereich der Prozessautomatisierung unerlässlich.

(D) Kommunizieren

Obwohl Kommunikation eine allgegenwärtige und gut erforschte menschliche Fähigkeit ist, fällt es Kommunikationsforschern traditionell schwer, sich auf eine gemeinsame Definition oder Taxonomie zu einigen. Eine der einfachsten technologisch motivierten Definitionen versteht Kommunikation als die Übertragung von Informationen zwischen bestimmten Subjekten ([61], S. 379–423). In einem komplementären Ansatz wird Kommunikation dagegen über die Fähigkeit zu kommunizieren definiert: Diese wird dann als die Fähigkeit beschrieben, eine Äußerung zu verarbeiten, zu verstehen und zwischen Äußerung und Information zu unterscheiden und damit folglich zu unterscheiden zwischen „dem Informationswert ihres Inhalts“ und „den Gründen, aus denen der Inhalt geäußert wurde“ ([62], S. 251–259). In diesem Sinne wird die Fähigkeit, zu kommunizieren – ähnlich wie die Fähigkeit, zu handeln – als eine übergeordnete Fähigkeit betrachtet, die nicht nur die Fähigkeit zum Wahrnehmen oder Empfinden, sondern auch zum Verarbeiten und Verstehen voraussetzt. Während in der Populärliteratur häufig nach dem Medium (z. B. mündlich, schriftlich etc.) unterschieden wird, ist in der Kommunikationsforschung die Unterscheidung nach der Anzahl der Beteiligten (intrapersonal, interpersonal, transpersonal) ein weithin anerkanntes Kriterium. Auch der Einfluss des Feedbacks auf den Kommunikationsprozess (einseitig, bidirektional, omnidirektional) wird häufig als charakteristisches Merkmal von Kommunikation angesehen. Hinsichtlich der Fähigkeiten von KI-Systemen steht derzeit die Ermöglichung von Mensch-Maschine-Interaktion bzw. -Kommunikation im Vordergrund. Maschine-zu-Maschine-Interaktionen sind derzeit in der Regel „von Menschenhand gemacht“. In Zukunft wäre jedoch denkbar, dass Maschine-zu-Maschine-Kommunikation durch Algorithmen verbessert wird, um eine spontane, aufgabenorientierte und flexible Maschine-zu-Maschine-Interaktion zu erreichen.

Tabelle 3: Klassifikation von KI-Fähigkeiten

Feld	Domäne	Fähigkeit	Beispiele
WAHRNEHMEN	Extern	Sehen	Optische Texterkennung (OCR)
			Objekterkennung
			Gestenerkennung
			Infrarotsicht
		Hören	Spracherkennung
			Audioerkennung
			Erkennung von Radarsignalen
		Riechen	Duftstoffdetektion
			Säuredetektion
			Branddetektion
			Caprylsäure-Detektion
		Schmecken	Zucker-Detektion
			Säure-Detektion
			Salz-Detektion
			Bitterstoff-Detektion
			Umami-Detektion
		Tasten	Temperaturerkennung
			Druckerkennung
			Schmerzerkennung
			Elektromagnetismuserkennung
	Intern	Eigenwahrnehmung	Erkennung eigener Bewegung
			Erkennung von Körperpositionen
		Gleichgewicht	Balance-Erkennung

Feld	Domäne	Fähigkeit	Beispiele
VERARBEITEN	Faktisch	Auflisten	Kontextspezifische Terminologie
		Zusammenfassen	Automatisierte Bericht-Generierung
		Antworten	Datenassoziation
		Auswählen	Semantische Suche
		Überprüfen	Parsen (syntaktische Analyse)
		Generieren	Deduktive Wissensextraktion
	Konzeptuell	Erkennen	Erkennung benannter Entitäten
		Klassifizieren	Erkennung semantischer Domänen
		Erstellen	Erklärung
		Differenzieren	Begriffsklärung
		Bestimmen	Semantische Interpretation
		Aufbauen	Sprachübersetzung
	Prozedural	Erinnern	Prozesserinnern
		Präzisieren	Modellierung der Umwelt
		Ausführen	Diskursmodellierung
		Integrieren	Fusion von Sensordaten
		Bewerten	Nutzermodellierung
		Entwerfen	Modellierung menschlicher Verarbeitung
	Metakognitiv	Identifizieren	Strategieauswahl
		Vorhersagen	Zustandsbestimmung
		Anwenden	Transferansätze
		Dekonstruieren	Kodierungswechselstrategien
		Reflektieren	Selbstoptimierungsmethoden
		Schaffen	Narrativgeneration

Feld	Domäne	Fähigkeit	Beispiele
HANDELN	Physisch	Bewegungsplanung	Bewegungsplanung mit Unsicherheit
		Sensoren & Manipulatoren	Passive & aktive Sensoren
		Kinematik & Dynamik	Dynamische Bewegung
		Mensch-Roboter-Interaktion	Multimodale Teleoperation
	Nicht-physisch	Softwaren-Agenten	Prozessautomation
			Transaktionssysteme
			Chatbots & Kundenservice
KOMMUNIZIEREN	Sprachverarbeitung	Textgenerierung	Paraphrasierungstools
		Maschinelle Übersetzung	Sprachübersetzung
		Textanalyse	Parsen (syntaktische Analyse)
		Informations- & Wissensextraktion	Erkennung benannter Entitäten
		Information Retrieval	Semantische Suche
		Dokumentenanalyse	Erkennung semantischer Domänen
		Sprachdialogsysteme	Klärungsdialoge
			Narrativgeneration
	Mensch-Maschine-Interaktion	Kognitive Systeme	Nutzermodellierung
		Interaktionsparadigmen & Modalitäten	Multimodale Interaktion

KLASSIFIKATION VON KI-ANWENDUNGEN

Die Klassifikation von KI-Anwendungen orientiert sich häufig an den oben beschriebenen KI-Methoden und -Fähigkeiten. Ziel der KI-Anwendung ist es, die mathematischen Methoden und abstrakten Fähigkeiten mittels Software konkret zu implementieren. Auf diese Weise sind spezialisierte Softwaremärkte entstanden, die diese typischen KI-Produkte vermarkten. Diese können von Unternehmen und Anwender*innen gekauft oder gemietet werden, um die Produktivität der Geschäftsprozesse zu steigern oder Innovationen der Geschäftsmodelle zu ermöglichen. Auch sind die typischen Softwaremärkte (siehe Tabelle 4) weltweit einheitlich bezeichnet und werden von unabhängigen Marktanalysten (z. B. IDC, Gartner, Forrester etc.) regelmäßig beobachtet, sodass potenzielle Anwender*innen, Projekte und Investierende sich gut über den Stand der Fähigkeiten informieren können.

Die Softwaremärkte können grob in die Bereiche Business Intelligence & Decision Support, AI based Customer Interaction, AI based Services und AI Development Environment & Tools eingeteilt werden.

Bei Business Intelligence & Decision Support steht das zeit- und themengerechte Erstellen von Reporten im Mittelpunkt. Diese haben das Ziel, einen quantitativen und qualitativen Überblick über das Geschäft zu gewährleisten, und sind schon seit vielen Jahren in allen Bereichen – z. B. Finanz, Human Resources (HR), Entwicklung, Marketing und Vertrieb – kommerziell verfügbar. Auf diese Weise werden Entscheidungen unterstützt und komplette Planungsprozesse in komplexen Umgebungen ermöglicht. Diese Fähigkeiten beinhalten auch Analytics, da sie typischerweise die Analyse vieldimensionaler Datenräume bedingen. Wesentliche Produkte in diesem Bereich sind Softwareumgebungen zur mathematischen und KI-basierten Optimierung sowie Berechnung von Vorhersagen. Ein weiterer Bereich ist die Verarbeitung von Sprache typischerweise zur Suche, Navigation und Exploration in großen Textkörpern. Setzt man mehrere dieser Funktionen zusammen, können ganze Geschäftsprozesse automatisiert werden, was häufig als Robotic Process Automation bezeichnet wird.

Seit 2012 hat sich der KI-Trend deutlich beschleunigt, da die verfügbaren CPUs und GPUs (Central und Graphics processing units) immer leistungsfähiger werden und KI-Methoden auf der Basis künstlicher neuronaler Netze schneller und kostengünstiger realisiert werden können. Dies erlaubt neue Möglichkeiten für die Mensch-Maschine-Schnittstelle, basierend auf KI-Anwendungen, welche SMS, Chats, Sprache und physische Bewegungen simulieren und entsprechende Prozesse, z. B. einfache Dialoge in Call- und Servicecentern, automatisieren.

Um die Nutzung von KI-Anwendungen zu vereinfachen, werden typische KI-Anwendungen aus Public- oder Private-Cloud-Umgebungen angeboten. Dies erlaubt es den Anwender*innen, sofort mit der Anpassung der Anwendung an die eigenen Bedürfnisse anzufangen und nicht erst hohe Aufwände für den Aufbau von Hard- und Software zu haben. Typische KI-Services, welche out of the box angeboten werden, sind: Bilderkennung, Videoanalyse, Sprache-in-Text-Umwandlung, Text-in-Sprache-Umwandlung, Übersetzung, Textanalyse, intelligente Suche und Maschinelles Lernen. In allen wird die eigentliche Nutzung des künstlichen neuronalen Netzes gekapselt und durch eine einfache grafische Nutzeroberfläche oder durch simple Funktionsaufrufe aus Standardsprachen (z. B. Java, C, Python, etc.) erleichtert.

Für die Entwicklung von KI-Anwendungen braucht man entsprechende KI-Entwicklungsumgebungen und -werkzeuge. Diese tragen den typischen Phasen eines KI-Projekts Rechnung: Build, Train und Run. In allen Phasen kommen häufig Open-Source-Technologien und Softwarebibliotheken zum Einsatz, welche zum einen die KI-Methoden anbieten und zum anderen professionelle Softwareentwicklung, z. B. methodengestützt und in verteilten Teams.

Mittels Regulierung von Systemen auf KI-Basis können mögliche Unzulänglichkeiten von KI-Anwendungen sowie wettbewerbsverzerrende Konstellationen vermieden werden. In Anlehnung an das Weißbuch der Europäischen Kommission „Zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen“ sind mit Blick auf Regulierung folgende Aspekte von Bedeutung: Haftung, Transparenz und Zuständigkeiten sowie Trainingsdaten, Aufbewahrung von Daten und Aufzeichnungen, vorzulegende Informationen, Robustheit, Genauigkeit, menschliche Aufsicht und besondere Anforderungen an bestimmte KI-Anwendungen, z. B. Anwendungen für die biometrische Fernidentifikation.

Die ethischen Aspekte der Entwicklung, des Nutzens und der Normung von KI werden aktuell besonders diskutiert. Folgende Eigenschaften spielen hier eine wichtige Rolle, welche methodisch und technisch für jede KI-Anwendung durchdacht und sichergestellt werden sollten: Autonomie & Kontrolle, Transparenz, Stabilität gegenüber Störungen, Sicherheit und alle Fragen des Datenschutzes.

Tabelle 4: Übersicht über Softwaremärkte und typische Produkte

Software markets & typical AI-Applications	
Software market	Typical software products
Business Intelligence & Decision Support Systems	Business Intelligence
	Decision Support
	Work-Flow-Systems
	Planning Analytics
	Constraint Based Optimization
	Prediction Capability
	Text Processing Platforms & Search Engines
	Robotic Process Automation (Rule-Based)
	Cognitive Automation (Training-Based)
	Real-Time Processing
AI based Customer Interaction	Chatbots
	Voicebots
	Avatars
	Virtual & Augmented Reality
AI based Services consumed from Public- or Private-Cloud	Image Recognition
	Video Analytics
	Speech To Text
	Text To Speech
	Translation
	Deep Learning as a Service
	Knowledge Navigation
	Knowledge Exploration
	Intelligent Search
	Natural Language Processing
	Automatical Annotation
AI development environment & tools	Build & Develop AI
	Train & Optimize AI
	Run & Manage AI
	Ethic Support Tools

KLASSIFIKATION VON KI-AUTONOMIE

KI-Anwendungen und die Computersysteme, die sie implementieren, können einen unterschiedlichen Grad an Entscheidungsautonomie haben. So unterscheidet beispielsweise die Datenethikkommission der deutschen Bundesregierung drei Klassen von Autonomie:

→ Algorithmisch basierte KI-Anwendungen arbeiten als reine Assistenzsysteme ohne autonome Entscheidungskompetenz. Die von ihnen berechneten (Teil-)Ergebnisse und (Teil-)Informationen bilden jedoch die Grundlage für menschliche Entscheidungen.

→ Algorithmusgesteuerte KI-Anwendungen nehmen dem Menschen Teilentscheidungen ab oder prägen durch die von ihnen berechneten Ergebnisse menschliche Entscheidungen. Dadurch schrumpfen der tatsächliche Entscheidungsspielraum des Menschen und damit auch seine Möglichkeiten zur Selbstbestimmung.

→ Algorithmisch determinierte KI-Anwendungen treffen eigenständig Entscheidungen und weisen damit einen hohen Grad an Autonomie auf. Durch den hohen Automatisierungsgrad gibt es im Einzelfall keine menschliche Entscheidung mehr, insbesondere keine menschliche Überprüfung von automatisierten Entscheidungen.

4.1.2 Anforderungen und Herausforderungen

4.1.2.1 Ethik

Ethische Prinzipien im Kontext von KI und Normung

Eine wesentliche Aufgabe der Ethik besteht in der Aufstellung und Begründung allgemein zustimmungswürdiger Maßstäbe, orientiert an Werten und Prinzipien (z. B. Menschenwürde, Gerechtigkeit, Freiheit), aus denen sich Handlungs- und Verhaltensanleitungen für menschliches (Zusammen-)Leben mit einem berechtigten (rational nachvollziehbaren) Anspruch auf Allgemeingültigkeit ableiten lassen. Auf ihrer Grundlage werden auch etablierte gesellschaftliche Moralvorstellungen noch einmal kritisch hinterfragt (vgl. „Normungsroadmap Künstliche Intelligenz Ausgabe 1“, Kapitel 11.2 „Philosophische Grundlagen zur Ethik“, [63]).

Spätestens seit Beginn des 20. Jahrhunderts entstehen ethisch relevante Fragen und Probleme nicht mehr allein im Kontext zwischenmenschlicher Interaktion, sondern auch durch die Auswirkungen neuer Technologien auf menschliches (Zusammen-)Leben bzw. in der Interaktion zwischen Mensch und Technik/Technologie. Vor diesem Hintergrund hat sich ab der Mitte des 20. Jahrhunderts als eine weitere Teildisziplin die Angewandte Ethik entwickelt. Sie befasst sich mit entsprechenden spezifischen Aspekten, die über die klassischen Fragen der Ethik hinausgehen (z. B. Medizin-, Technik- und Wirtschaftsethik). Ihr Ziel ist es, ethische Maßstäbe und die daraus abgeleiteten Normen und Prinzipien im Sinne von allgemeinen Spielregeln in entsprechenden anwendungsspezifischen Kontexten zur Geltung zu bringen. In diesem Kontext ist auch eine KI-Ethik zu verstehen.

Moralische Prinzipien liegen nicht zwingend explizit oder gar formalisiert vor, sie können auch implizite Konventionen von Individuen und Gruppen sein, die deren Handeln beeinflussen. Diese allgemeinen Prinzipien müssen von Akteur*innen auf ihre konkrete Situation übertragen und in Bezug auf den Kontext der jeweiligen Situation operationalisiert werden. Teilbereiche etablierter Regeln werden in Rechtsstaaten ggf. als Recht formalisiert, wobei dieses nicht stets deckungsgleich mit den auch möglicherweise zeitveränderlichen allgemeinen moralischen Prinzipien und Werten einer Gesellschaft ist (siehe Abbildung 17, rechts). Beides zusammengenommen stellt den Rahmen gesellschaftlich vertretener Ziele und Erwartungen dar, demzufolge sich KI sowohl an der Moral in ihrem jeweiligen Kontext als auch an geltendem Recht orientieren muss.

Ein KI-System entfaltet Wirkungen, die innerhalb und außerhalb eines gesellschaftlich ausgehandelten ethischen Rahmens wirken (Abbildung 17, Mitte). Dabei ist zu differenzieren zwischen den Zielen, die explizit durch das KI-System verfolgt werden, und den Modalitäten der Zielerreichung oder Umsetzungsaspekten (jeweils Abbildung 17, links oben). So kann ein System naheliegenderweise deshalb gegen ethische Werte verstoßen, weil deren Erfüllung nicht hinreichend in den Systemzielen berücksichtigt wurde. Aber auch ein System, das auf die Einhaltung aller relevanten moralischen Prinzipien zielt, lässt sich aus KI-ethischer Perspektive kritisieren, wenn die Modalitäten der Zielerreichung ungenügend sind – beispielsweise, wenn für Außenstehende nicht beurteilbar ist, ob das System diese Werte tatsächlich einhält, oder wenn das System z. B. aus mangelnder Robustheit an der Erreichung dieser Ziele scheitert.

Es lässt sich demnach zwischen jenen Erwartungen, die an die expliziten Systemziele gestellt werden, und jenen Erwartungen, die an die Art der Zielerreichung gestellt werden, differenzieren. Nur die Berücksichtigung beider Aspekte führt zu einer Gesamtwirkung des Systems, das sich in der ethischen Reflexion bewährt.

Die Gewährleistung dessen ist ein Prozess, der während des gesamten Lebenszyklus eines KI-Systems von Bedeutung ist (Abbildung 17, unten links) und durch unterschiedliche verantwortliche Akteur*innen für das KI-System (kurz KI-Systemverantwortliche, beispielsweise Auftraggebende, Entwickelnde, Prüforganisationen, öffentliche Stellen oder Betreibende) sichergestellt werden muss. Entlang der sieben Phasen dieses Lebenszyklus sowie im Kontext der übergreifenden Governance ergeben sich unterschiedliche Bedarfe für Normung und Standardisierung, die im Kapitel 4.1.3 Abschnitt Ethik vorgestellt und entsprechend eingeordnet werden.

Ethisch betrachtet werden Entwicklung und Betrieb von KI-Systemen mit Blick darauf, wie sie Werte konkret operationalisieren, das heißt: umsetzen. Hier lässt sich prüfen, wie ein KI-System beispielsweise das Prinzip sicherstellt: „KI-Systeme müssen menschliche Selbstbestimmung respektieren“. Es geht also um die konkrete Anwendung von KI, die ethisch eingeordnet und bewertet wird. Die genannten KI-Systemverantwortlichen haben regelmäßig sicherzustellen und nachvollziehbar zu erläutern, dass das von ihnen verantwortete KI-System fortlaufend ethischen Prinzipien entspricht. Bei Entwurfsentscheidungen werden ebenfalls solche bevorzugt, die eine Einhaltung der ethischen Prinzipien fördern. Wo immer die Einhaltung der ethischen Prinzipien gefährdet wird, muss von den verantwortlichen Akteur*innen gründlich (im Sinne von überzeugend und rational nachvollziehbar) dargelegt werden, wieso dies der Fall ist, und ggf. Konsequenzen (z. B. Auflagen für den Betrieb, Außerbetriebsetzung bzw. keine Betriebsfreigabe etc.) gezogen werden.

Die Akteur*innen der Operationalisierung ethischer Prinzipien im Umfeld eines spezifischen KI-Systems können vielfältig sein. Es ist jedoch stets davon auszugehen, dass die benannten KI-Systemverantwortlichen maßgeblich an den ethischen Abwägungen in Bezug auf ihr KI-System beteiligt sind. Denn ihnen obliegt die Pflicht, das von ihnen verantwortete KI-System in seinem konkreten Wirkungszusammenhang in Einklang mit den rechtlichen und ermittelten ethischen Prinzipien zu bringen.

Dass sie KI-Systeme wertebasiert entwickeln und betreiben, können KI-Systemverantwortliche durch ethische Reflexion fördern und begründen (vgl. hierzu auch [64]). Diese ersichtliche Wertebasis ist ein wesentlicher Faktor dafür, dass KI-Systeme gesellschaftliche Akzeptanz finden. Sensibilität für und Kritikfähigkeit bei der technischen Umsetzung einer Wertebasis sind beispielsweise zwei Aspekte, die das Pflichtbewusstsein der KI-Systemverantwortlichen prägen. Dieses Pflichtbewusstsein ist vergleichbar mit einem Berufsethos von KI-Systemverantwortlichen (analog zum hippokratischen Eid in der Medizin). Es liegt auf der Hand, dass die genannte ethische Reflexion bei Entwicklung und Betrieb von KI-Systemen dieses Ethos stärkt und aktualisiert.

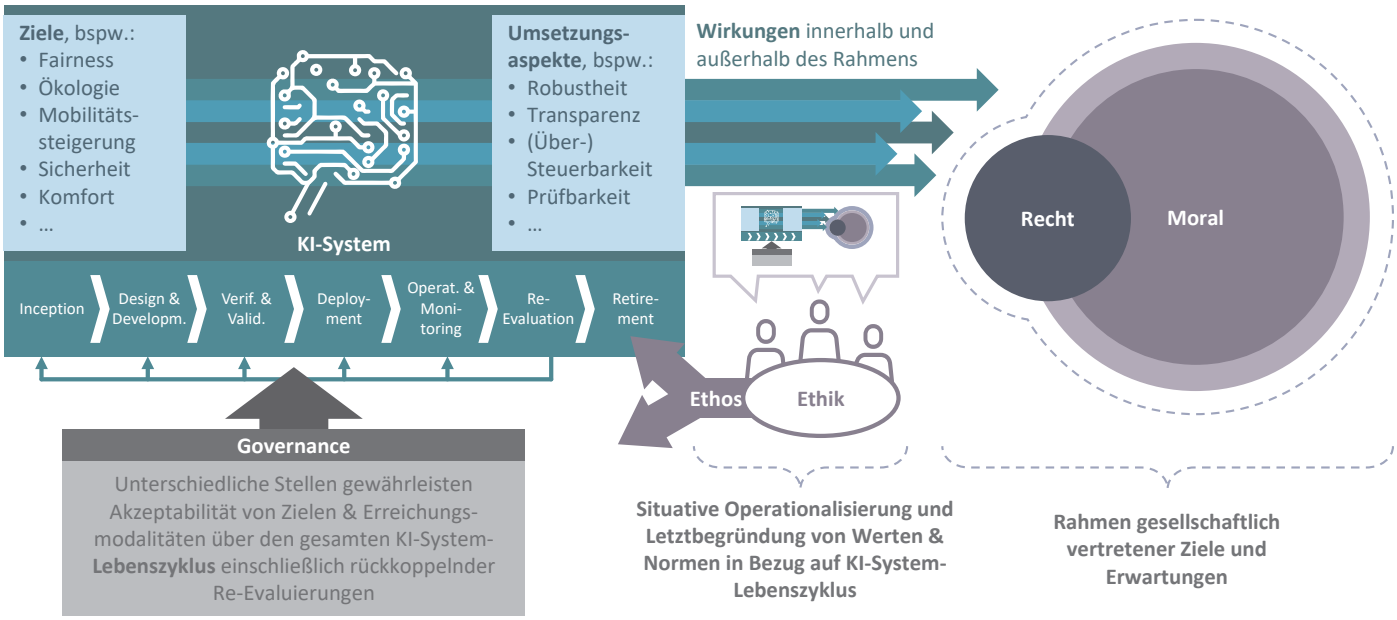


Abbildung 17: Ethik zwischen KI-System-Life-Cycle (Quelle: in Anlehnung an [16]), Arbeitsgruppe Grundlagen)

Die ethische Reflexion bei Entwicklung und Betrieb von KI-Systemen kann folgende grundlegende Schritte umfassen, die hier anhand eines durchgängigen Beispiels aus dem Bereich Medizin veranschaulicht sein sollen. Das Beispiel stellt ein KI-gestütztes Diagnosesystem zur Hautkrebserkennung als Smartphone-Anwendung dar, eine sogenannte Teledermatologie-App (im Folgenden: Derma-App). Benutzer*innen fotografieren das entsprechende Hautareal. Die Teledermatologie-App analysiert die Aufnahme und spricht eine Empfehlung aus. Im Falle eines möglichen Verdachts auf Hautkrebs rät sie, eine/n Facharzt/ärztin aufzusuchen.

Um KI-Entwicklung und -Betrieb nach ethischen Prinzipien und Werten auszurichten, bedarf es folgender Schritte:

Die KI-Systemverantwortlichen	Beispiel Derma-App: KI-gestütztes Diagnosesystem zur Hautkrebserkennung als Smartphone-Anwendung
→ entwickeln einen KI-Entwurf auf Wertebasis: Sie definieren das Wertverständnis (beispielsweise auf Basis ihres Code of Conduct) und priorisieren Werte für ihren Use Case, also den Anwendungsfall inklusive des jeweiligen Kontexts. Diesen Prozess gestalten sie transparent und nachvollziehbar.	Für die Derma-App berücksichtigen die KI-Systemverantwortlichen verschiedene Werte und ethische Prinzipien, beispielsweise: → **Gleichbehandlung** und **Erklärbarkeit**, beispielsweise von bzw. für Menschen mit unterschiedlichem Bildungshintergrund: durch übersichtliche Modelldarstellungen oder Erklärungen in leicht verständlicher Sprache muss angemessen nachvollziehbar sein, wie die App funktioniert und auf welcher Basis sie Empfehlungen ausspricht. Nutzer*innen müssen die Empfehlung als Orientierung einordnen können und nicht als Ersatz für die Behandlung durch medizinisches Fachpersonal. Es muss klar sein, dass zu einer ganzheitlichen Diagnose neben der optischen Betrachtung auch andere Untersuchungen (z. B. ein Tastergebnis) zählen, ebenso wie eine Verlaufs- und Vergleichsbeobachtung. Dies kann die App nicht leisten. Gleichbehandlung schließt zudem ein, dass die Bedienung der App (z. B. in Bezug auf die Aufnahme der Bilder oder Eingabe weiterer Daten) für unterschiedliche Benutzergruppen zuverlässig umsetzbar ist, ohne wesentliche Einschränkungen bei der Verlässlichkeit der Ergebnisse zu bewirken (vgl. u. g. Diversität). → **Verlässlichkeit**, hier in Bezug auf die Klarheit der Empfehlungsbasis: Damit Patient*innen und Ärzt*innen einschätzen können, auf welcher Basis die Derma-App Empfehlungen ausspricht und mit welcher Wahrscheinlichkeit diese zutreffen, benötigen sie konkrete Einsicht. Beispielsweise ließen sich die für die Diagnose ausschlaggebenden Teile der Haut-Fotografie markieren, sodass Nutzer*innen direkt einsehen, an welchen Merkmalen die App ihre Empfehlung festmacht. Auch ist im Rahmen der Qualitätssicherung eine stete Verbesserung der Quote zutreffender Empfehlungen erforderlich. Dies lässt sich auch durch Einbezug diverser Trainingsdaten erreichen (vgl. unten). Um Fehldiagnosen zu reduzieren, können KI-Systemverantwortliche ein Ausbalancieren von Artefaktquellen vorsehen (sodass beispielsweise Belichtungsfehler der Haut-Fotografie nicht fälschlich als pathogene Anomalie interpretiert werden). → **Diversität**, beispielsweise durch angemessen diverse Trainingsdaten: Die App muss für alle Menschen gleichermaßen eingesetzt werden können (unabhängig von Alter, Geschlecht oder Hautfarbe). Um Überanpassung an bestimmte Muster zu vermeiden, können KI-Systemverantwortliche Daten aus verschiedenen Quellen einsetzen, beispielsweise unterschiedlichen Laboren. → **Selbstbestimmung** z. B. hinsichtlich geeigneter Benutzerschnittstelle mit entsprechenden Eingriffsmöglichkeiten: Dermatolog*innen und Patient*innen müssen KI-Systemverantwortlichen Feedback geben können. Den KI-Systemverantwortlichen muss es möglich sein, darauf zu reagieren, um Fehlerquellen zu eliminieren. Sie müssen falsche oder fehlerhafte Daten löschen oder das KI-System zurücksetzen können.

Die KI-Systemverantwortlichen	Beispiel Derma-App: KI-gestütztes Diagnosesystem zur Hautkrebserkennung als Smartphone-Anwendung
	Um Werte in erforderlichem Maß einzubeziehen und zu priorisieren, gehen die KI-Systemverantwortlichen in den Dialog mit Betroffenen. Sie prüfen gemeinsam mit Repräsentant*innen beispielsweise aus den Gruppen „medizinisches Fachpersonal“ und „Patient*innen“, unter Berücksichtigung von Diversitätsaspekten (wie Alter, Bildungshintergrund, Geschlecht etc.) die gelisteten Werte und deren Priorisierung. Sie dokumentieren dieses Ergebnis und stellen es in seinen wichtigsten Aspekten und in übersichtlicher Form begleitend zu Informationen der Derma-App öffentlich zur Verfügung.
→ formulieren Anforderungen an das KI-System auf Grundlage ihrer Erkenntnisse zu den relevanten Werten: Ausgehend von dem jeweiligen Zielwert und dem Anwendungskontext des Systems ermitteln sie zentrale Anforderungen, wie dessen Funktionen unter Beachtung der erstellten Wertelistung und -priorisierung umzusetzen sind. Sie gehen systematisch vor, um die Bewertung einzelner Teilanforderungen an ihr KI-System abzustimmen und ein „Ethics by Design“ zu erreichen.	Zwei Werte, die für obigen Use Case beispielhaft als gegenseitig bestärkend genannt sind, sind **Gleichbehandlung** und **Erklärbarkeit**. Es lässt sich dafür argumentieren, dass beide auf den Zielwert **Selbstbestimmung** einzahlen, denn für fundierte Kritik- und Feedbackfähigkeit seitens Betroffenengruppen ist ein grundsätzliches Verständnis der Funktionen und Prozesse der Derma-App erforderlich. Um das ethische Prinzip „KI-Systeme müssen menschliche Selbstbestimmung respektieren“ als übergeordnetes Ziel umzusetzen, müssen KI-Systemverantwortliche sicherstellen, dass die Benutzer*innen jederzeit Hoheit über ihre Entscheidungen behalten. Aber auch das Verständnis einer Empfehlung des Systems muss bei unterschiedlichen Benutzergruppen in ausreichender Weise gegeben sein. Falls ein solches Verständnis für bestimmte Benutzergruppen, z. B. Menschen mit unzureichender Technologie- bzw. Medienkompetenz, nur bedingt zu ermöglichen ist, müssen KI-Systemverantwortliche Maßnahmen spezifizieren, die eine angemessene Einschätzung sicherstellen können. Das ließe sich z. B. durch die verpflichtende Einbindung weiterer Personen wie medizinisches Fachpersonal erreichen: Es kann ein Dialogbereich innerhalb der Derma-App eingerichtet werden, in dem Patient*innen mit ihren Rückfragen an Ärzt*innen herantreten können. Ebenso kann es einen Bereich geben, in dem weitere Kontaktmöglichkeiten zu Praxen und medizinischen Beratungsstellen zur Verfügung stehen, um eine unmittelbare Vernetzung und Hilfestellung zu erzielen.
→ müssen Zielkonflikte hinsichtlich ihrer Werte beschreiben und lösen.	In Bezug auf den oben genannten Wert **Verlässlichkeit** kann es unterschiedliche Abwägungen geben. In den Blick genommen sei an dieser Stelle der Schwellenwert, ab dem die Derma-App den Besuch eines Arztes oder einer Ärztin empfiehlt: → Ist der Schwellenwert sehr hoch, reagiert die Derma-App auf Hautveränderungen sehr sensibel. Ein Arztbesuch wird tendenziell häufig empfohlen (Risiko von False Positives). → Ist der Schwellenwert sehr niedrig, reagiert die Derma-App auf Hautveränderungen vergleichsweise unsensibel. Ein Arztbesuch wird tendenziell seltener empfohlen (Risiko von False Negatives). Im Rahmen der Verlässlichkeit gilt es nun abzuwägen: Nehmen KI-Systemverantwortliche mehr False Positives in Kauf, um die Gefahr übersehener Alarme zu reduzieren und eine in diesem Sinn verlässliche Früherkennung zu ermöglichen? Oder nehmen sie mehr False Negatives in Kauf, um das Gesundheitssystem durch unnötige Untersuchungen und Behandlungen sowie Belastungen für Benutzer*innen nicht überzustrapazieren? Diese beiden Aspekte gilt es, ins Gleichgewicht zu bringen.

Die KI-Systemverantwortlichen	Beispiel Derma-App: KI-gestütztes Diagnosesystem zur Hautkrebserkennung als Smartphone-Anwendung
	Neben der Zielabwägung innerhalb eines Wertes stellt sich auch die Frage der Abwägung von zwei Werten in Gegenüberstellung. An dieser Stelle sei die **Gleichbehandlung** (hier: bezüglich Zugang zum System) mit der **Verlässlichkeit** (hier: durch Qualitätssicherung) abgewogen. Wenn für alle Benutzergruppen der gleiche Zugang gewährt wird, kann es sein, dass bei technisch weniger versierten Menschen die Verlässlichkeit der App reduziert wird, wenn die erforderlichen Aufnahmen nicht richtig erstellt oder die Ergebnisse nicht richtig interpretiert werden können. Das betrifft z. B. unterschiedliche Altersgruppen und unterschiedliches technisches, aber auch sprachliches Verständnis. Insofern konkurriert der Wert Gleichbehandlung mit dem Wert Verlässlichkeit und es muss auch in diesen Aspekten eine Ausgewogenheit erreicht werden, indem z. B. die Benutzerschnittstelle so gestaltet ist, dass sie für unterschiedliche Benutzergruppen in geeigneter Weise zugänglich oder durch entsprechende Maßnahmen (z. B. Einbindung von Fachpersonal) abgesichert wird.
→ weisen nach, ob das KI-System letztlich nach den ermittelten Anforderungen funktioniert, und stellen eine fortlaufende Qualitätssicherung sicher.	Wie die obigen Ausführungen zeigen, müssen KI-Systemverantwortliche für die Derma-App festlegen, wie einzelne Werte nachvollziehbar umgesetzt werden. Sie legen fest, durch welche Maßnahmen → Nutzer*innen gleichberechtigten Zugang zur App erhalten, → Empfehlungen angemessen und userfreundlich eingeordnet sind, → Empfehlungen in angemessener Verlässlichkeit gewährleistet werden, → Nutzer*innen die App selbstbestimmt einsetzen können und wie das überprüft werden kann. Dabei ist zu beachten, dass es unterschiedliche Ebenen der Überprüfung bzw. Validierung gibt. Letztendlich müssen KI-Systemverantwortliche nicht nur die einzelnen Anforderungen, sondern das System als Ganzes prüfen. Nur so können sie gerade in komplexen Systemen die Wechselwirkungen einzelner Komponenten bzw. Entscheidungen und die daraus potenziell resultierenden Zielkonflikte einschätzen. Dies betrifft eine Gesamtbewertung, die Kriterien der klinischen Wirksamkeit und ethische Aspekte integriert. Im Kern muss validiert werden, ob die zu Beginn der Entwicklung vorgegebenen Werte in dem vorliegenden Anwendungsfall in ausreichender Weise umgesetzt werden konnten. Das beinhaltet eine repräsentative Abdeckung der im Anwendungsfall vorhandenen Benutzergruppen und Anwendungskontexte. Da oftmals während des Entwicklungsprozesses nicht sämtliche Situationen abgedeckt und/oder vorausgesehen werden können, ist es zudem erforderlich, Daten aus dem Betrieb des Systems im Sinne einer Quality-Backward-Chain systematisch zu erfassen und in regelmäßigen Abständen von einem geeigneten Gremium reevaluieren zu lassen. In dem vorliegenden Beispiel würde die Quality-Backward-Chain eine systematische Überprüfung beinhalten, → inwiefern die einzelnen Benutzergruppen die richtigen Entscheidungen für ihren persönlichen Fall erhalten haben bzw. → ob die Mechanismen zur menschlichen Aufsicht (Einbezug weiterer Personen) so wirksam waren, dass dieser individuelle Fall passend behandelt werden konnte.

Die KI-Systemverantwortlichen	Beispiel Derma-App: KI-gestütztes Diagnosesystem zur Hautkrebserkennung als Smartphone-Anwendung
	Bei der Reevaluation geht es um eine Überprüfung, ob über das gesamte Spektrum an Anwendungsfällen die anvisierten Ziele auch in ihrer ethischen Dimension in geeigneter Weise umgesetzt werden konnten bzw. wo es Handlungsbedarf in Hinblick auf eine Verbesserung des Systems bzw. der zugehörigen Unternehmensprozesse gibt. Ins Gewicht fallen hier fortlaufend Aspekte wie: → potenziell systematische Ungleichbehandlung von bestimmten Benutzergruppen zu vermeiden und Gleichbehandlung zu fördern oder → die Verlässlichkeit und Erklärbarkeit, also die Genauigkeit, mit der die App Empfehlungen ausspricht und für die Benutzer*innen nachvollziehbar darstellt. Dies ist Grundlage für das Vertrauen in ihre Funktionalität seitens aller Zielgruppen.

Normung kann den komplexen Prozess der Umsetzung von Werten bei Entwicklung und Betrieb von KI unterstützen. Sie …

→ gibt Impulse für Ziele, die geeignet sind, die ethische Vertretbarkeit eines KI-Systems zu begründen. Dabei greift sie die zentralen ethisch relevanten Fragen und Probleme dieses Spezialgebietes auf, die gesellschaftspolitisch identifiziert werden,
→ liefert die Basis für Argumente, die von ethisch agierenden Personen im Rahmen ihres Diskurses verwendet werden können,
→ schafft ein intersubjektives Sprachverständnis, dadurch dass sie Begriffe prägt und definiert (die Verwendung einer gemeinsamen Sprache ermöglicht überhaupt erst die Kommunikation und den Austausch von Argumenten),
→ entwickelt Schemata, um KI-Systeme einheitlich zu klassifizieren,
→ entwickelt Verfahren, die ethische Prozesse standardisieren und wertbasierte Systemanforderungen messbar machen.

Normung unterstützt die Umsetzung von Werten – mitunter das Denken, Kommunizieren und Argumentieren mit Blick auf ethisch relevante Fragen – im Kontext KI effizienter zu machen. Ein wesentliches Ziel ist dabei, die Grundlage dafür zu schaffen, KI systematisch und kontextbezogen vertrauenswürdig zu entwickeln und zu betreiben – das bedeutet: in Bezug auf den Wert Vertrauenswürdigkeit. Um diesen Aspekt und seine Voraussetzungen geht es im folgenden Kapitel.

Wertesysteme für vertrauenswürdige KI

Der Begriff „Vertrauenswürdigkeit“ kann sich grundsätzlich sowohl auf Organisationen als auch auf technische Systeme beziehen. Demgegenüber ist zu spezifizieren: Ethik referiert [65] nur auf „vernunftbegabte Wesen”, die sich zwar als Akteur*innen (z. B. KI-Systemverantwortliche) in Organisationen, jedoch nicht in technischen bzw. algorithmischen Systemen wiederfinden. Konkretere Ausführungen zu Vertrauenswürdigkeit in Bezug auf Organisationen bzw. technische Systeme finden sich exemplarisch ergänzend in Kapitel 4.1.2.2.

Werte und Anforderungen an vertrauenswürdige KI im Allgemeinen

Die „Hochrangige Expertengruppe für Künstliche Intelligenz der Europäischen Kommission“ (HLEG-KI) [8] wie auch die „Enquete-Kommission KI“ [66] haben eine Reihe von Anforderungen an KI-Systeme im Hinblick auf ihre Vertrauenswürdigkeit beschrieben. Diese als Leitlinien bezeichneten Werte bzw. Anforderungen an vertrauenswürdige KI-Systeme umfassen die folgenden Punkte (vgl. Kapitel 1.4):

1. Vorrang menschlicher Aufsicht von KI-Systemen sowie die Einhaltung und Sicherstellung von Grundrechten: Es wird gefordert, dass im Zusammenhang mit KI-Systemen Auskunfts-, Aufsichts- und Kontrollmechanismen zur Verfügung stehen sollen, um negative Auswirkungen z. B. auf Grundrechte, aber auch den Missbrauch von KI-Systemen zu vermeiden.
2. Technische Robustheit und Sicherheit, z. B. die Widerstandsfähigkeit gegen Angriffe und Sicherheitsverletzungen, Auffangplan und allgemeine Sicherheit, Präzision, Zuverlässigkeit und Reproduzierbarkeit.

3. Schutz der Privatsphäre und Datenqualitätsmanagement, z. B. die Achtung der Privatsphäre, Qualität und Integrität der Daten sowie Datenzugriff. Fragestellungen, die Standardisierungsaktivitäten betreffen, sind Datenschutzmanagement im Zusammenhang mit KI, aber auch, wie Datenqualität insgesamt sichergestellt werden kann.
4. Transparenz, Nachvollziehbarkeit und Erklärbarkeit. In der Praxis werden diese Begriffe oft synonym verwendet. Sie beziehen sich aber auf verschiedene Aspekte der Offenlegung, wie im Weiteren definiert.
   - Transparenz bezieht sich auf die Frage nach dem „Was“. Sie hat zum Ziel, den Einsatz von KI-Komponenten in einem System erkennbar zu machen und seine relevanten Eigenschaften zu beschreiben. Diese Kenntnis ist notwendig, um eine bewusste Entscheidung über die Nutzung des KI-Systems zu ermöglichen.
   - Nachvollziehbarkeit bezieht sich in diesem Zusammenhang auf die Möglichkeit, die transparent gemachten Eigenschaften eigenständig und unabhängig überprüfen zu können.
   - Erklärbarkeit bezieht sich auf die Frage nach dem „Warum“. Durch sie kann das Verhalten der KI-Komponenten und ihr Zusammenspiel in einer konkreten Situation verstanden werden. Diese Kenntnis ermöglicht es, Entscheidungen des KI-Systems auf ihre Einflussfaktoren zurückzuführen und so die Ursache einzelner Entscheidungen nachzuvollziehen. Datensätze und Prozesse, die zu der Entscheidung des KI-Systems geführt haben, sollen dokumentiert werden.
5. Fairness, Nichtdiskriminierung und Vielfalt, z. B. Vermeidung unfairer Verzerrungen, Zugänglichkeit und universeller Entwurf sowie Beteiligung der Interessenträger, Förderung von Diversität.
6. Gesellschaftliches und ökologisches Wohlergehen, z. B. Nachhaltigkeit und Umweltschutz, soziale Auswirkungen, Gesellschaft und Demokratie.
7. Rechenschaftspflicht, z. B. Nachprüfbarkeit, Minimierung und Meldung negativer Auswirkungen, Kompromisse und Rechtsbehelfe.

Vergleichend hierzu lässt sich auch die Landscape of AI ethics guidelines nennen, in deren Rahmen fünf Werte bzw. ethische Prinzipien mit grundlegender Bedeutung für KI-Systeme ermittelt wurden: Transparenz, Gerechtigkeit, Fairness, Nichtschadensprinzip, Verantwortung und Privatheit [67].
Ein weiterer Ansatz für wertorientierte Entwicklung und den Einsatz von KI-Systemen findet sich beispielsweise auch im Whitepaper Ethik-Briefing [68].

Der Wert Fairness im Besonderen

Fairness hat aus verschiedenen Gründen eine Sonderstellung als Anforderung (siehe o. g. Punkt 5) an vertrauenswürdige KI-Systeme. Zum einen fordert die Gesellschaft zu Recht ganz allgemein und grundsätzlich Fairness besonders bei exponentiellen Technologien wie der Anwendung Künstlicher Intelligenz ein, zum anderen hat sich Fairness als Operationalisierung von Nichtdiskriminierung (im Sinne von ungerechtfertigter Benachteiligung, Verzerrung oder Ungleichbehandlung) in den letzten zehn Jahren bereits in der Informatikwissenschaft und deren praktischer Anwendung etabliert.

Wenn die allgemeine Definition von Fairness nach Duden mit „anständiges Verhalten; gerechte, ehrliche Haltung andern gegenüber“ oder „den [Spiel]regeln entsprechendes, anständiges und kameradschaftliches Verhalten beim Spiel, Wettkampf o. Ä.“ breite Anerkennung findet, würde eine gemeinsame spezifischere Definition z. B. aus den beiden Blickwinkeln der Disziplinen Philosophie und Technik schon weitaus schwieriger werden. Selbst die Begrenzung des Blickwinkels auf nur eine Disziplin wie die Informatik ist bei der spezifischen Definition von Fairness immer noch eine Herausforderung.

Da jedoch beim Einsatz algorithmischer und soziotechnischer Systeme im weiteren und maschinell lernender Systeme im engeren Sinn immer häufiger Fairness gefordert wird, ist Handeln geboten. Dabei ist die Bedeutung des Begriffs auch in diesem Kontext höchst umstritten. Im Groben lassen sich zwei Hauptströme unterscheiden: Fairness als ethisches Prinzip (basierend auf Werten wie Gerechtigkeit) und Fairness als Operationalisierung von Nichtdiskriminierung. Oft ist nicht klar, wonach sich der Ruf nach Fairness im konkreten Fall richtet. Im Sinne von Operationalisierung von Nichtdiskriminierung gibt es jedoch nicht nur konkrete Umsetzungsstrategien, sondern bereits auch konkrete Vorschläge zur Messung und Beurteilung, die in der Praxis eingesetzt werden.

In den letzten zehn Jahren hat sich eine eingeschränkte Auffassung von Fairness in der Informatik parallel zu einem ethischen Verständnis aus der Angewandten Philosophie entwickelt. In der Informatik besteht die Bestrebung, „nur“ das Ausmaß von Diskriminierung durch ein algorithmisches System durch sogenannte Fairnessmaße invers (also die „Nichtdiskriminierung“) zu messen. Damit sind nicht alle Aspekte von Fairness abgedeckt.

Die Vielzahl an Ansätzen, Fairness im Sinne von Nichtdiskriminierung zu messen, vertreten verschiedene Perspektiven und Strategien und lassen sich grob in individuelle und Gruppen-Fairnessmaße einteilen. In jedem Fall setzt ein allgemeingültiges Fairnessmaß ein gemeinsames Diskriminierungsverständnis voraus. Dieses ist jedoch durch verschiedene Moralvorstellungen, Normensysteme, Prinzipien, Werte oder Dispositionen, die alle für sich den Anspruch erheben, die Grundlage richtigen Handelns zu sein (siehe Glossar Ethik), nicht vorhanden bzw. gegeben. Für eine ethische Reflexion ist es überdies unumgänglich, auftretende Diskriminierung in KI-gestützten Anwendungen (vgl. [69], 3) als mögliche Ausweitung sozialer Ungleichheiten zu prüfen, die durch Menschen als soziale Akteur*innen verkörpert werden (vgl. [69], 6). Daher sollte auch der durch mögliche hierarchische Machtasymmetrien strukturierte gesellschaftliche Hintergrund, aus dem algorithmische Systeme hervorgehen können, in den Blick genommen werden (vgl. [70], 2). Damit kann es nicht „das eine" Fairnessmaß geben, sondern es sollte eine bewusste Auswahl an Fairnessmaßen getroffen werden, um die beabsichtigten Fairnessziele messbar und nachweisbar zu fördern. Insofern erscheint es essenziell – um dem Wert Fairness in seiner je kontextbezogenen Umsetzung gerecht werden zu können –, mit einschlägigen Betroffenengruppen in Dialog zu treten (wie auch im Beispiel unter Punkt 1 dieses Unterabschnitts exemplarisch angeführt), um die Fairnesschancen und -herausforderungen eines KI-Systems in direktem Bezug auf die jeweiligen Stakeholder ermitteln und berücksichtigen zu können. Neben dem Value-based Engineering im Kontext des oben genannten IEEE 7000:2021 [64] beziehen diesen Aspekt auch Ansätze wie das Participatory Design [71] oder das Value Sensitive Design [72] ein. Uneinig ist man sich bei den Gruppenfairnessmaßen, dass es um die (bedingte) Gleichbehandlung von Gruppen gehen muss. Bei individueller Fairness hingegen herrscht die Auffassung vor, dass ähnliche Personen ähnlich behandelt werden sollen, basierend auf einer (beliebigen) Funktion, die die Ähnlichkeit bestimmt. An dieser Stelle sei darauf hingewiesen, dass eine Ungleichbehandlung auch gerechtfertigt sein kann (z. B. bei der Vergabe einer Arbeitsstelle, die eine hohe körperliche Kraft voraussetzt, oder der Priorisierung von vulnerablen Gruppen bei der Impfstoffvergabe).

Verschiedene Fairnessmaße repräsentieren verschiedene Vorstellungen von Fairness, viele davon können nicht gleichzeitig optimiert werden, da sie zu einem gewissen Grad im Widerspruch zueinander stehen. Wird gezielt auf ein bestimmtes Fairnessmaß optimiert, werden damit die Ergebnisse anderer Fairnessmaße mitunter zwangsläufig reduziert. Dadurch kann Diskriminierung nach dem Verständnis der reduzierten Maße sogar erhöht werden (vgl. Kapitel 4.8.2.3).

Da es grundsätzlich kaum möglich ist, moralisch gebotenes Handeln in festen Algorithmen oder starren Regelwerken abzubilden, zeichnet sich eine vertrauenswürdige Organisation bestehend aus „vernunftbegabten Wesen" bzw. Mitarbeitenden (frei nach Kant) dadurch aus, sich besonders in Konfliktsituationen ethisch reflektiert zu verhalten, auch wenn bestehende Gesetze oder Firmenvorschriften damit in Konflikt stehen könnten (siehe Ethische Leitlinien der Gesellschaft für Informatik e. V. [73]). Moderne Governance- und Managementsysteme (siehe Kapitel 4.1.2.2) beinhalten genau für solche Konfliktsituationen zum Schutz von Mitarbeitenden klare und wirksame Compliance-Meldewege.

Fallbeispiel Governance

Die Umsetzungsmöglichkeiten von Werten wie Vertrauenswürdigkeit sollen im Folgenden anhand eines Beispiels beschrieben werden, das auf dem konkreten Fall eines großen Softwarehauses beruht und seit mehreren Jahren im Einsatz praktisch gelebt wird.

Im Beispiel geht die Operationalisierung in Form von „Grundsätzen" zu einem ethischen Umgang mit KI auf eine Initiative der Mitarbeitenden zurück. Diese holen die Unterstützung des Topmanagements ein und führen internationale Workshops unter weltweiter Beteiligung aller von KI bzw. ML betroffenen Unternehmensbereiche durch. Die hierin erarbeiteten Grundsätze beinhalten drei Blickwinkel bzw. Rollen: Mitarbeitende/Arbeitgebende, Lösungsanbieter*innen und Gesellschaftsmitglieder. Die Grundsätze beschreiben deren Zusammenwirken gemäß dem Prinzip der Nachhaltigkeit, im Sinne des bewussten Umgangs mit materiellen und immateriellen Ressourcen in einer Weise, dass deren heutige Erstellung, Verwendung und Weiterentwicklung die Bedürfnisse künftiger Generationen nicht beeinträchtigt.

Diese abstrakten Grundsätze werden anschließend zu „Leitsätzen" konkretisiert und darüber weiter zu Handlungsanweisungen und Regeln detailliert, beispielsweise folgendermaßen:

Grundsatz	Wir entwickeln für Menschen. (Dies geht auf die Kant'sche Selbstzweckformel zurück und impliziert u. a.: Die Technologie ist stets für den Menschen da – nie umgekehrt.)

Leitsatz	Klarstellung für Mitarbeitende, wie ethische Grundsätze im Arbeitsalltag einzubringen sind
Konkrete Handlungs-anweisungen	→ keine Grey, Dark oder Black Patterns (beispielsweise gezielt irreführende Benutzerinteraktion z. B. bei Cookie-Auswahloptionen durch entsprechende Hervorhebung oder Abdunkeln der Schaltflächen) → Lieferkettencheck bei Fremddienst-leistenden → keine De-Anonymisierung → ...

Konkrete Werkzeuge wie eine Kritikalitätspyramide oder Risikomatrix (vgl. NRM KI Ausgabe A1 [63]) zur Einordnung der unternehmensinternen algorithmischen Systeme unterstützen eine nachvollziehbare und niederschwellige Umsetzung.

Darüber hinaus wird eine AI-Ethics-Governance-Struktur aus externen und internen Expert*innen aufgebaut, beispielsweise in Form eines „AI-Ethics-Steering Committee", „AI-Ethics-Office" oder „External Advisory Panel on AI". Diese Struktur ist zuständig für die dauerhafte Ausgestaltung und Weiterentwicklung der Grundsätze, Leitsätze und Handlungsempfehlungen, bildet inhärent die Unternehmenswerte ab und hält diese aktuell.

Das Beispiel zeigt praxistaugliche Schritte auf, die unternehmensintern zur Operationalisierung von Ethik möglich sind. Es gibt jedoch bereits Bestrebungen von Organisationen und Wissenschaft, unternehmensübergreifende Prozessstrukturen und Konzepte in diesem Bereich anzubieten (z. B. IEEE 7000:2021 [64] und KIDD-Prozess [74]). Normung kann hier unterstützen, eine Referenz bereitzustellen und die Vergleichbarkeit von Maßnahmen zu gewährleisten.

4.1.2.2 Umsetzung bei KI-Entwicklung und -Betrieb: Blick auf Produkte und Dienste sowie Organisationsstrukturen

Wie in Kapitel 4.1.2.1 dargestellt, kanns sich der Begriff „Vertrauenswürdigkeit" sowohl auf Organisationen wie auch auf technische Systeme beziehen. Einem technischen System (d. h. einem Produkt oder einer elektronisch bereitgestellten Dienstleistung) kann bezüglich gewisser Eigenschaften wie Sicherheit oder Zuverlässigkeit vertraut werden, wenn ein Beleg (z. B. in Form eines Prüfberichts oder eines Zertifikats) dafür vorliegt, dass das System solche Eigenschaften erfüllt. Die Vertrauenswürdigkeit einer Organisation ist weiter gefasst: Sie bezieht sich darauf, dass einer Organisation zugetraut wird, geeignete Maßnahmen durchzuführen und Managementstrukturen – ein sogenanntes Managementsystem – zu unterhalten, um die Erwartungen ihrer Stakeholder und anderer interessierter Parteien zu erfüllen. Neben einem entsprechenden Prüfbericht kann auch die Reputation einer Organisation oder ihre Akzeptanz am Markt zu ihrer Vertrauenswürdigkeit beitragen.

Vertrauen in Produkte und Dienste

Die sogenannten Common Criteria (CC) beschreiben eine Methodik zur Prüfung von Produkten und Diensten mit Fokus auf deren Sicherheit, die als Begriffsgerüst für entsprechende Prüfungen von KI-Systemen verwendet werden können. Die CC liegen ebenfalls als Internationaler Standard DIN EN ISO/IEC 15408-1:2020 [445] vor. Unterstützend wird eine abgestimmte Methodik für die Evaluierung auf Grundlage der CC im internationalen Standard DIN EN ISO/IEC 18045: 2021 [75] beschrieben. Diese Dokumente stellen die technische Basis des Common Criteria Recognition Arrangement (CCRA) [76] dar, das von einer Vielzahl von Staaten, so auch von Deutschland, unterzeichnet wurde. Weitere Informationen zu den CC finden sich u. a. auf der Website des BSI [77].

Anforderungen an eine Prüfung nach den CC werden in sogenannten Evaluation Assurance Levels (EAL) zusammengefasst:

EAL1	funktionell getestet
EAL2	strukturell getestet
EAL3	methodisch getestet und überprüft
EAL4	methodisch entwickelt, getestet und durchgesehen
EAL5	semiformal entworfen und getestet
EAL6	semiformal verifizierter Entwurf und getestet
EAL7	formal verifizierter Entwurf und getestet

Vertrauen in Organisationen

Zur weiteren Untersuchung der Anforderungen der HLEG-KI zur Vertrauenswürdigkeit von KI soll ein begrifflicher Exkurs zur Unterscheidung der Begriffe „Governance" und „Manage-

ment“ unternommen werden, wie sie im ISO/IEC zurzeit etwa in ISO/IEC 38500:2015 [78] vorgenommen wird. Es ist hierbei zu beachten, dass sich der Begriff „Managementsystem“ auf alle drei im Folgenden diskutierten Ebenen, nämlich das Leitungsgremium, das Management und konkrete technisch-organisatorische Maßnahmen, bezieht, wie in Abbildung 18 dargestellt.

Governance

Governance bezieht sich auf die allgemeinen Aufgaben und Zielsetzungen einer Organisation, ihres Selbstverständnisses und auf die sich daraus ergebenden Werte und die Kultur der Organisation, die ihr Handeln bestimmt. Dies umfasst insbesondere auch das ethische Wertesystem, zu dem sich die Organisation bekennt (s. 4.1.2.1 Ethik). Ein zentraler Begriff ist der der Risikobereitschaft. Nach dem Begriffsgerüst der ISO/IEC 38500:2015 [78] ist das Leitungsgremium (governance body) einer Organisation verantwortlich für die Umsetzung ihrer Rechenschafts- und Sorgfaltspflichten. Gerade auch Fragen der Haftung bekommen in Verbindung mit KI besondere Relevanz, da durch den möglichen Automatisierungsgrad der KI die Frage, wer bei Fehlern und Schäden haftet, wichtig ist. Dies sollte die Governance berücksichtigen, da sich der rechtliche Rahmen in diesem Feld dynamisch entwickelt.

Das Leitungsgremium leistet hierzu Vorgaben und erstellt Richtlinien, die innerhalb der Organisation umgesetzt werden müssen.

Weiterhin ist das Leitungsgremium für die Etablierung von Managementstrukturen (Prozesse, Rollen, Verantwortlichkeiten) und die Bereitstellung adäquater Ressourcen verantwortlich.

Management

Das Management einer Organisation setzt die Vorgaben und Richtlinien des Leitungsgremiums in konkrete Prozesse, Rollen und Verantwortlichkeiten um. Beispiele für Managementaufgaben sind u. a.:

- → die Identifikation und Analyse potenzieller Risiken und die Etablierung von Handlungsoptionen, basierend auf der Risikobereitschaft der Organisation,
- → die Etablierung eines Datenschutzmanagements sowie von Prozessen zur Sicherstellung ausreichender Datenqualität,
- → die Einführung eines Sicherheitsmanagements für KI-basierte IT-Systeme,
- → effektives Management der Entwicklung und des Betriebs von KI-Systemen.

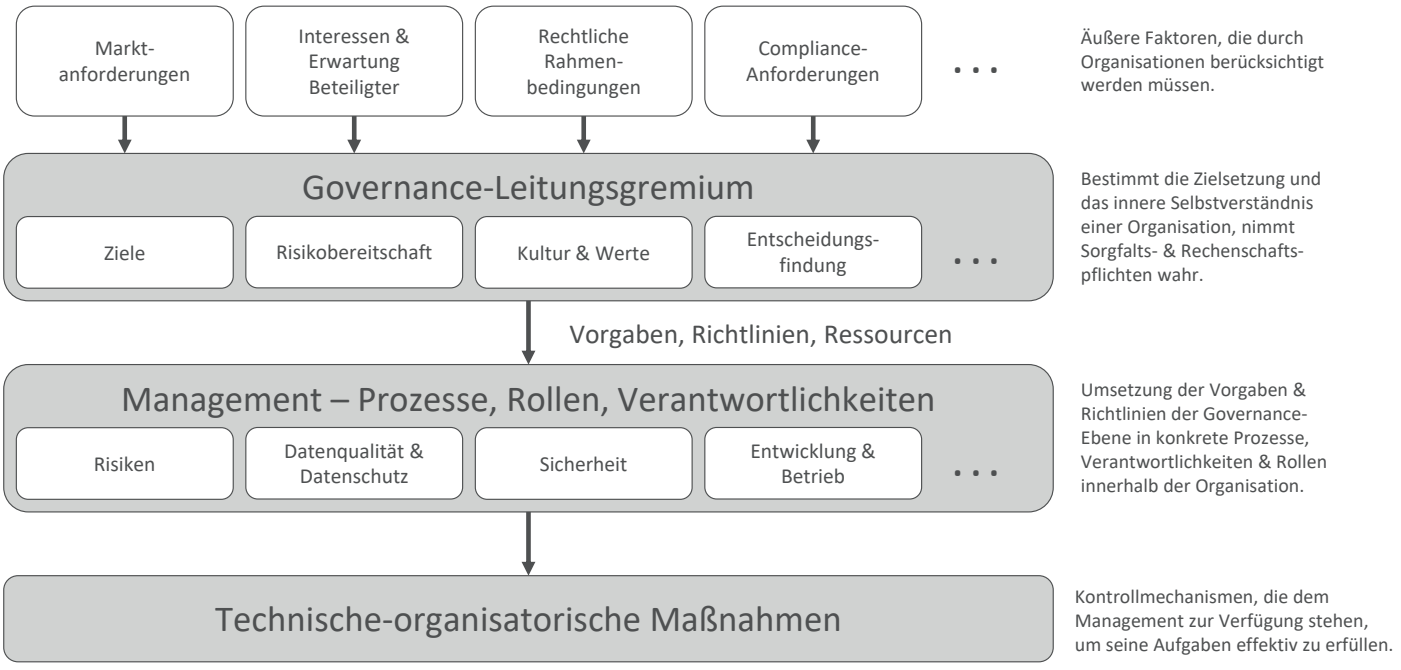


Abbildung 18: Managementsystem: Governance, Management und technisch-organisatorische Maßnahmen (Quelle: Peter Deussen)

Technisch-organisatorische Maßnahmen
Dieser Begriff umfasst alle technischen und organisatorischen Hilfsmittel, die dem Management zur Verfügung stehen, um seine Aufgaben effektiv und nachprüfbar zu erfüllen. Technisch-organisatorische Maßnahmen reichen von der Verfügbarkeit von Verschlüsselungsfunktionen zur Erhöhung der Datensicherheit über die Anwendung statistischer Methoden zur Identifikation unfairer Verzerrungen bzw. von Kontaminierung in Datensätzen bis hin zur Verfügbarkeit von Test- und Validationswerkzeugen.

Anforderungen an das Managementsystem
Im Kontext der internationalen Standardisierung spielt der Begriff des Managementsystemstandards (MSS) eine zentrale Rolle. Ein MSS definiert Anforderungen an Organisationen zur Durchführung eines effektiven und verantwortungsvollen Managements. Zum Teil werden auch Anforderungen an das Leitungsgremium einer Organisation gestellt, und viele MSS enthalten weiterhin konkrete Kontrollen im Sinne von technisch-organisatorischen Maßnahmen. Der Begriff „Managementsystem“ bezieht sich damit auf das Gesamtbild aus Abbildung 18. Mindestanforderungen an das Managementsystem sind in der sogenannte High Level Structure (HLS) [263] beschrieben:

1. **Kontext der Organisation**, hierzu zählen u. a. rechtliche Rahmenbedingungen, gesellschaftliche Erwartungen, Bedürfnisse und Erwartungen interessierter Parteien, Ziele und Werte der Organisation sowie der eigentliche Geltungsbereich des Managementsystems.
2. **Leitung**, das Leitungsgremium muss verbindliche Bereitschaften der Organisation definieren und in Form von Vorgaben niederlegen. Vorgaben zum ethischen Wertesystem sind Teil dieser Vorgaben. Weiterhin muss es Prozesse, Rollen, Verantwortlichkeiten für ein effektives Management bestimmen.
3. **Planung** umfasst Aktivitäten, um mit Risiken und Chancen umzugehen.
4. **Unterstützung** umfasst die Bereitstellung von Ressourcen, die Bestimmung notwendiger Kompetenzen, die Sicherstellung von notwendiger Achtsamkeit, die Kommunikation und Dokumentation.
5. **Betrieb** umfasst die operative Umsetzung von Managementanforderungen.
6. **Leistungsevaluation** umfasst das Monitoring, die Analyse und Evaluation, die interne Auditierung und Begutachtung durch das Management.
7. **Verbesserung** befasst sich mit der Identifikation von Nonkonformität bezüglich der Anforderungen des MSS, korrektiven Maßnahmen und der kontinuierlichen Verbesserung des Managementsystems.

Organisationen können Konformität mit MSS nachweisen (z. B. durch eine Selbstbewertung oder Zertifizierung durch eine unabhängige dritte Partei) und damit die Vertrauenswürdigkeit der Organisation bezüglich der spezifischen Aspekte des MSS erhöhen. Betrachtet man den Einsatz einer Klasse von Technologien wie die der KI, muss das Managementsystem einer Organisation deshalb auf die besonderen Charakteristiken und Wirkungsreichweiten der KI Bezug nehmen. Dies kann geschehen, indem existierende MSS um KI-spezifische Anforderungen erweitert werden. Da jedoch die verschiedenen MSS durch unterschiedliche Gremien im ISO und IEC publiziert und gewartet werden, die weder über ein gemeinsames Begriffsgerüst noch über eine synchronisierte Arbeitsweise verfügen, und es darüber hinaus nicht klar ist, ob existierende MSS überhaupt ausreichend sind, um alle Aspekte der KI zu berücksichtigen, ist es erfolgversprechender, einen neuen MSS zu entwerfen, der sich auf KI-spezifische Anforderungen konzentriert.

Unterstützende Standards
MSS umfassen lediglich Anforderungen an ein Managementsystem, beschreiben jedoch nicht seine Implementierung. Dies erlaubt es Organisationen, ihre eigenen Managementstrukturen in der für sie angepassten Weise zu definieren, solange ein Nachweis erfolgen kann, dass die Anforderungen des MSS erfüllt sind. Solche Strukturen, aber auch unterliegende technische und organisatorische Maßnahmen werden in der Regel in ergänzenden Standards beschrieben, die nun keine Anforderungen, sondern lediglich Richtlinien enthalten.

4.1.2.3 Entwicklung von KI-Systemen

Mit Software erhalten Maschinen einen immer größer werdenden Funktionsumfang. Hardware und Software bilden dabei eine Symbiose. Für Software mit einem vorbestimmten Funktionsablauf gibt es allgemein akzeptierte Entwicklungs- und Qualitätssicherungsverfahren, wie z. B. Code Reading, Modul- und Applikationtests auf verschiedenen Integrationsstufen, Verifikation und Validierung. Diese Methoden und Verfahren wirken auch bei der Software mit regelbasierten KI-Systemen. Neben der Qualität des Softwarecodes und der verwendeten Compiler kommt bei der Entwicklung von KI-Systemen der Softwarearchitektur, der Qualität der verwendeten Daten und der Lernphase eine besondere Bedeutung zu.

Lernende KI-Systeme erhalten wesentliche Funktionalitäten durch die Lernphase. Diese Lernphase kann statisch oder

dynamisch erfolgen, überwacht (supervised) oder unüberwacht (unsupervised). Wie beim Menschen auch stellt die Prüfung dessen, was gelernt wurde, eine große und für die Softwareentwicklung neue Herausforderung dar. Dieses ist insbesondere dadurch kritisch, dass KI-Systeme besonders dort ihre Stärke zeigen, wo Entscheidungen oder Entscheidungsempfehlungen auf Basis vieler Daten sehr zeitnah getroffen werden sollen.

Werden KI-Systeme zur automatisierten oder autonomen Entscheidungsfindung im sicherheitskritischen Bereich eingesetzt, so werden darauf bezogene Verfahren zur Nachweisführung und Konformitätsbewertung auch durch Dritte erforderlich. Dies gilt insbesondere für Nachweise im Rahmen der Nachweisführung zur funktionalen Sicherheit bei der Produkthaftung.

Ein zweckmäßiger Ansatz basiert auf der Betrachtung des gesamten Lebenszyklus eines KI-Systems in dessen Anwendungsumfeld sowie der Sicherstellung der Datenqualität in der Lern- und Anwendungsphase.

Der Lebenszyklus eines KI-Systems

Der International Standard ISO/IEC 22989:2022 [16] beschreibt ein generisches Lebenszyklusmodell für KI-Systeme, das die folgenden Phasen umfasst (vgl. Abbildung 19):

- → Initiierung (inception): Anfangsphase des Entwicklungsprozesses eines KI-Systems, in der die wesentlichen Anforderungen und Designparameter für das Projekt festgelegt werden.
- → Design und Entwicklung (design and development): Konstruktionsphase des KI-Systems, in der eine funktionsfähige Version für die folgende Phase der Verifikation und Validierung zur Verfügung gestellt wird.
- → Verifikation und Validierung (verification and validation): Prüfung des KI-Systems bezüglich Anforderungen und der Erfüllung von Projektzielen.
- → Überführung in die Einsatzumgebung (deployment): Das KI-System wird in seine Einsatzumgebung installiert. Diese Phase umfasst weitere Prüfungen, um sicherzustellen, dass das System in dieser Umgebung zufriedenstellend arbeitet.
- → Betrieb und Überwachung (operation and monitoring): Das System ist in Betrieb genommen und wird im laufenden Betrieb überwacht.

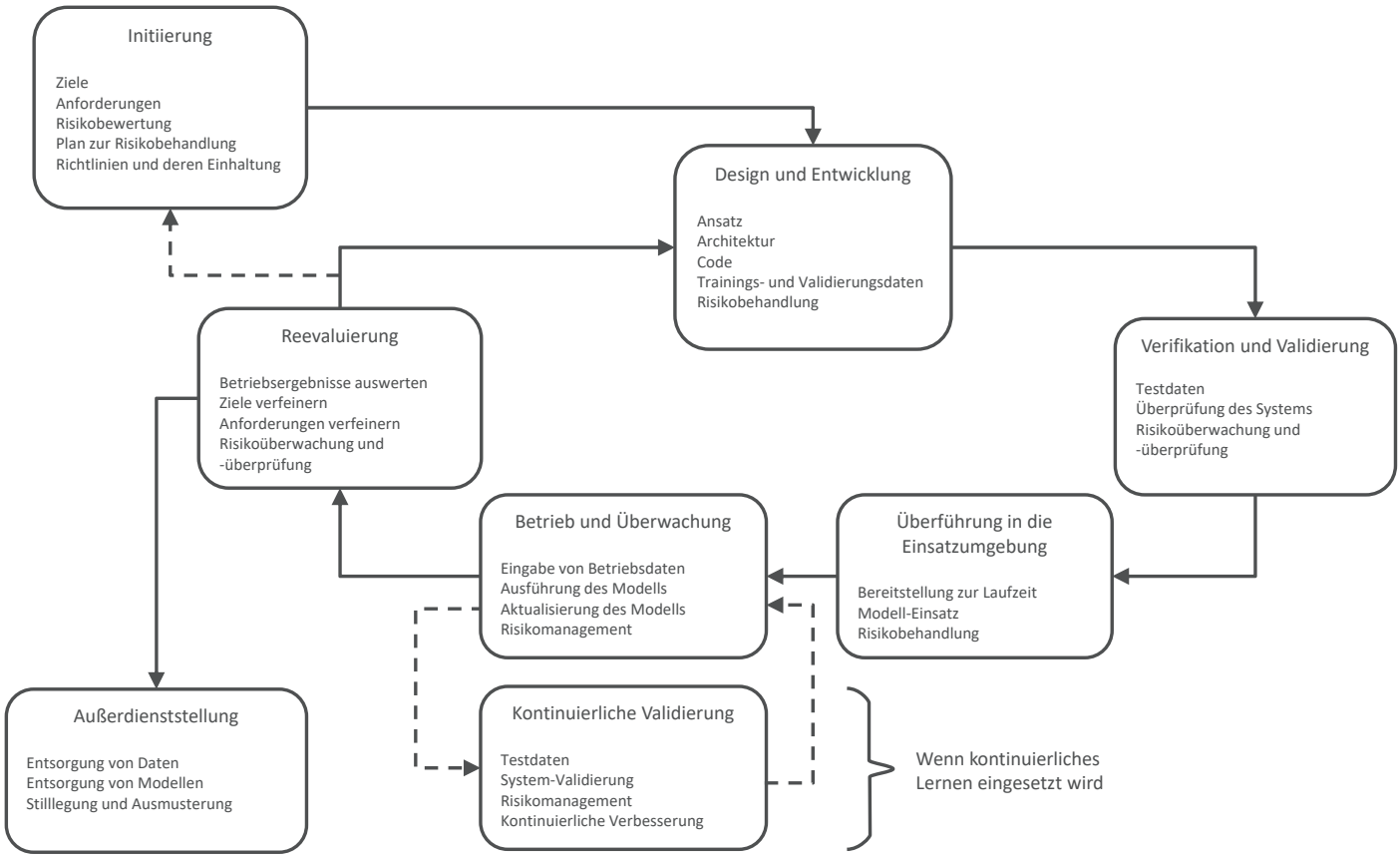


Abbildung 19: Lebenszyklus für KI-Systeme (Quelle: in Anlehnung an [16])

→ Kontinuierliche Validierung (continuous validation): KI-Systeme, die sich – z. B. durch kontinuierliches Lernen – fortlaufend an veränderte Umstände ihrer Betriebsumgebung anpassen, müssen entsprechend ihrer fortgesetzten Funktion entweder kontinuierlich oder in gesetzten Intervallen geprüft werden.
→ Reevaluierung (re-evaluation): In längeren Phasen wird eine Reevaluierung des KI-Systems bezüglich geänderter Ziele oder Anforderungen vorgenommen.
→ Außerdienststellung (retirement): Das KI-System wird stillgelegt.

Diese Phasen sind nicht im Sinne eines linearen Ablaufs voneinander abhängig, sondern müssen verzahnt durchlaufen werden. Der Wiedereintritt in eine bereits abgeschlossene Phase ist möglich. Präsentation

Datenqualität

Analog zur Verwendung eines System-Lebenszyklus, der die Organisation von Qualitäts- und Risikomanagement-Aktivitäten entlang der Entwicklungs- und Betriebsphasen eines KI-Systems erlaubt, kann ein Datenlebenszyklusmodell verwendet werden, um das Management der Datenqualität zu beschreiben. Die Normenserie ISO/IEC 5259 [39], die sich zurzeit in Entwicklung im ISO/IEC JTC 1/SC 42 befindet, adressiert Datenqualitätsmanagement.

Abbildung 20 setzt den Datenlebenszyklus mit einer spezifischen Verfeinerung für das Datenqualitätsmanagement in Beziehung. Die Phasen des Datenqualitätsmanagement-Lebenszyklus umfassen:

1. **Datenmotivation und Konzeptualisierung**. Basierend auf der intendierten Nutzung von Daten werden Konzepte zum Datenmanagement abgeleitet, die die Relevanz der Daten, Compliance-Anforderungen und ggf. ethische Anforderungen berücksichtigen.
2. **Datenspezifikation** umfasst die Beschreibung erforderlicher Daten, verwendbarer Datenformate, die Identifikation von fehlerhaften oder widersprüchlichen Anforderungen der Spezifikation.
3. **Datenplanung** umfasst die Planung der Implementierung der Datenspezifikation einschließlich der Planung spezifischer Tasks zur Datenbeschaffung und -verarbeitung und die Bereitstellung der hierfür notwendigen Ressourcen.
4. **Datenbeschaffung** umfasst die Sammlung von Daten, ggf. im Fall synthetischer Daten ihre Erzeugung, und die Kombination mit existierenden Daten.
5. **Datenvorbereitung** umfasst Aktivitäten wie die Reinigung und Filterung der Rohdaten oder die Reduktion des Datenumfangs.
6. **Datenanreicherung** umfasst die Ergänzung von Daten mit Metadaten, die Kategorisierung von Daten (Labeling) usw.

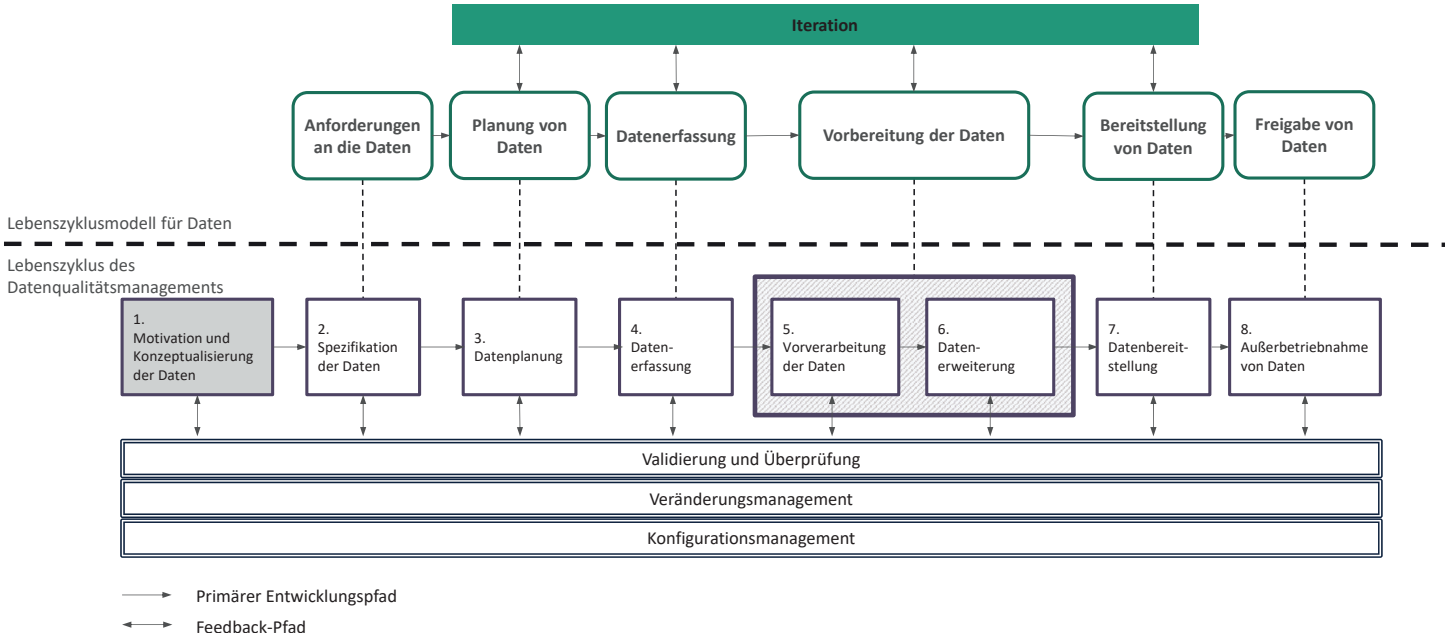


Abbildung 20: Datenlebenszyklus und Datenqualitätsmanagement-Lebenszyklus (Quelle: in Anlehnung an [39])

7. **Datenbereitstellung** umfasst die Verwendung von Daten für den vorgesehenen Zweck, z. B. für das Anlernen eines neuronalen Netzes.
8. **Datendekommissionierung** umfasst die Löschung von Daten bzw. den Transfer der projektbezogenen Daten in eine allgemeine Datenbasis oder ein neues Projekt.

Qualitätskriterien für Daten

Qualitätskriterien für Daten werden im Internationalen Standard ISO/IEC 5259-2 [41] diskutiert. Dieser Internationale Standard beschreibt insgesamt 19 Qualitätsmerkmale für Daten:

1. **Portierbarkeit**: Übertragbarkeit von Daten von einem System auf ein anderes.
2. **Verständlichkeit**: Grad der Verständlichkeit von Daten für den Nutzenden.
3. **Auditierbarkeit**: Verfügbarkeit von Daten für interne oder externe Audits.
4. **Identifizierbarkeit**: Grad der Identifizierbarkeit von Personen, mit denen Daten assoziiert werden können.
5. **Aktualität**: Grad der zeitlichen Angemessenheit von Daten.
6. **Glaubhaftigkeit**: Grad des Vertrauens, das ein Nutzer oder eine Nutzerin in den Wahrheitsgehalt von Daten setzen kann.
7. **Vollständigkeit**: Grad der Abdeckung der erwarteten Informationen durch einen Datensatz.
8. **Skalierbarkeit**: Grad, in dem die Datenqualität bei einer Erhöhung der Datenmenge oder Dateneingangsgeschwindigkeit erhalten bleibt.
9. **Generalisierbarkeit**: Grad, in dem Daten in einem Kontext verwendet werden können, für den sie ursprünglich nicht gesammelt wurden.
10. **Wirksamkeit**: Grad, in dem Daten bestimmte Anforderungen erfüllen.
11. **Akkuratheit**: Grad, in dem Daten einen bestimmten Sachverhalt korrekt wiedergeben.
12. **Präzision**: Grad der Genauigkeit, in der Daten einen bestimmten Sachverhalt von anderen Sachverhalten unterscheidbar machen.
13. **Konsistenz**: Grad der Widerspruchsfreiheit von Daten.
14. **Relevanz**: Grad der Angemessenheit von Daten für einen bestimmten Zweck.
15. **Rechtzeitigkeit**: Grad der Verzögerung der Datenverfügbarkeit in Bezug auf den Zeitpunkt ihrer Erhebung.
16. **Repräsentativität**: Grad, in dem Daten alle relevanten Aspekte eines gegebenen Sachverhalts beschreiben.
17. **Ausgewogenheit**: Grad, in dem alle relevanten Aspekte eines gegebenen Sachverhalts durch ausreichende Datenmengen beschrieben werden.
18. **Ähnlichkeit**: Grad, in dem relevante Sachverhalte durch ähnlich strukturierte Daten beschrieben werden.
19. **Diversität**: Grad der Vielfältigkeit von Daten.

Handlungsempfehlungen

Durch die Initiierung der Normenreihe der ISO/IEC 5259 [39] sind die Themen Datenqualität und Datenmanagement in der internationalen Standardisierung zumindest allgemein adressiert. Dennoch ist zu erwarten, dass für spezifische Sektoren und Anwendungen verschärfte und ggf. von den oben genannten abweichende Qualitätskriterien relevant werden. Auch Qualitätsmanagementprozesse müssen sektorspezifisch implementiert und ggf. angereichert werden. Somit wird empfohlen, in der vertikalen Standardisierung zum Datenqualitätsmanagement zu prüfen, inwieweit die ISO/IEC-5259-Reihe [39] als allgemeine Referenz herangezogen werden kann und inwieweit sektorspezifische Adaptionen notwendig werden.

4.1.2.4 Quanten-KI

Moderne Verfahren des Maschinellen Lernens (ML) sind einerseits, insbesondere während ihrer Entwicklung, oftmals extrem ressourcenintensiv und können andererseits bestimmte anspruchsvolle Problemstellungen nach wie vor nicht oder zumindest nicht effizient lösen. Quantencomputer zeigen hier das Potenzial, die in dieser Hinsicht bestehenden Limitationen zu überwinden.

Das Gebiet des „Quantum Machine Learning“ (QML) hat sich dabei als eigenständige Disziplin etabliert, die Ansätze des Maschinellen Lernens und der Quanteninformationsverarbeitung verbindet (siehe z. B. [79], [80]). Aus der Perspektive von ML-Entwickelnden und -Forschenden ist vor allem die Verwendung von Quanten-Algorithmen als Teil des klassischen ML-Lebenszyklus, vor allem während der Trainingsphase, ein relevanter Ansatzpunkt. Die grundlegende Idee, mit der die eingangs erwähnten Beschränkungen klassischer ML-Verfahren möglicherweise aufgelöst werden können, besteht in der Auslagerung bestimmter Teilprozesse und -berechnungen auf die Quanten-Hardware.

QML ist aktuell ein sehr dynamisches Forschungsgebiet, bei dem noch viele Fragen, insbesondere hinsichtlich der Praxistauglichkeit der diskutierten Verfahren, offen sind. Dennoch

sind hier, insbesondere bedingt durch die rasanten Entwicklungen im Bereich des Quantencomputings, in den nächsten Jahren deutliche Fortschritte zu erwarten. Ein Szenario, in dem die ML-Praxis durch den Einsatz von Quantencomputern nachhaltig verändert wird, sollte daher bereits jetzt diskutiert werden. Die Chancen, aber eben auch die Risiken, die QML dabei möglicherweise mit sich bringt, sind dazu eingehend zu betrachten.

Insbesondere im Bereich der IT-Sicherheit ergeben sich viele Fragestellungen, die vorausschauend zu behandeln sind [81]. Zwei Aspekte sind in dieser Diskussion wesentlich: Zum einen ist bereits bekannt, dass Angriffe auf klassische ML-Systeme [83] (Stichwort: Adversarial Machine Learning) prinzipiell auch auf QML-Systeme übertragbar sind. Inwiefern QML hier eine höhere Anfälligkeit aufweist oder aber eine verbesserte Resilienz bieten kann, ist jedoch noch unklar. Zum anderen profitieren Anwendungen in der IT-Sicherheit, die derzeit herkömmliche ML-Methoden nutzen, womöglich von etwaigen Effizienzsteigerungen durch den Einsatz von Quantencomputern. Dies betrifft grundsätzlich sowohl die Angreifer- als auch die Verteidigerperspektive.

Um QML einerseits von Beginn an als sichere Technologie zu etablieren und andererseits die Auswirkungen auf die IT-Sicherheit selbst abzuschätzen und zu adressieren, sind demnach noch erhebliche Forschungsbemühungen notwendig.

4.1.2.5 Sprachtechnologien

Die Sprachtechnologie ist ein interdisziplinäres Gebiet, das sich in erster Linie aus Informatik, Künstlichen Intelligenzforschung und Computerlinguistik speist und Anwendungen, Methoden und Lösungen für die Analyse oder Generierung geschriebener oder gesprochener Sprache entwickelt, wobei aktuell auch Multimodalität eine wichtige Rolle spielt, etwa die simultane Verarbeitung von Sprach- und visuellen Daten.

Ein zentrales Merkmal der Sprachtechnologie betrifft den Umstand, dass es eine große Bandbreite von Applikationen umfasst: vom klassischen Anwendungsfall der maschinellen Übersetzung (geschriebener oder gesprochener Sprache) reicht das Spektrum über die Synthese gesprochener (z. B. natürlich klingende Ansagen auf Bahnsteigen) oder Generierung geschriebener Sprache (z. B. automatische Erstellung von Produktbeschreibungen), die Erkennung gesprochener Sprache (z. B. Erkennung und Transkription von Textnachrichten auf dem Telefon) bis hin zur Analyse geschriebener Sprache (z. B. Textklassifikation, Informationsextraktion, Erzeugung von Wissensgraphen, Textzusammenfassung, Erkennung von Entitäten, syntaktisches oder semantisches Parsing etc.). Sprache wird zudem auch immer häufiger als Kanal für die Mensch-Maschine-Interaktion eingesetzt, z. B. für Frage-Antwort-Systeme, Information Retrieval und Suchmaschinen, für Chatbots sowie für smarte persönliche Assistenten, wie sie seit einigen Jahren in allen modernen Betriebssystemen, Telefonen sowie diversen Haushaltsgeräten und auch Autos verfügbar sind.

Der historische Kern der Sprachtechnologie ist die Computerlinguistik sowie die sprachverarbeitende, wissensbasierte KI, die seit den 1970er-Jahren insbesondere manuell entwickelte Regelsysteme und symbolverarbeitende Methoden nutzte (semantische Netze, Taxonomien, Ontologien, Wissensgraphen).

Nach erneuter wissenschaftlicher Hinwendung zu statistischen KI-Verfahren in den späten 1990er-Jahren dominieren seit etwa zehn Jahren korrelative (nichtdeterministische) neuronale Netze, wobei diese Entwicklung auch gefördert wurde durch die günstige Verfügbarkeit leistungsfähiger GPUs. Maschinelle Lernverfahren, die dem Bereich des Deep Learning zugeordnet werden, dominieren Wissenschaft und Technologie in zahlreichen Teilbereichen der Sprachtechnologie, wobei u. a. große Sprachmodelle eingesetzt werden, die u. a. auf der Transformer-Architektur basieren und auf der Grundlage sehr großer Mengen von Sprachdaten hochdimensionale Repräsentationen lernen, deren Performanz in zahlreichen konkreten sprachtechnologischen Aufgaben weit über die Leistungsfähigkeit rein statistischer Verfahren hinausgehen. Entsprechend nutzen nahezu alle modernen sprachtechnologischen Systeme neuronale Verfahren oder große Sprachmodelle in unterschiedlichen Ausprägungen, oftmals auch in Verbindung mit neuen symbolischen, funktionalen Methoden (z. B. Wissensgraphen mit aktiven Ontologien), wenn Wissen deterministisch repräsentiert werden soll, um z. B. auch Ergebnisse von neuronalen Systemen zu plausibilisieren (hybride KI).

Derzeit wird daher u. a. an der Verbindung symbolischer und subsymbolischer Methoden geforscht, um die jeweiligen Vorteile zu kombinieren und Nachteile zu kompensieren, etwa durch die Integration komplexer Ontologien oder einfacher Wissensgraphen in große Sprachmodelle, sodass das explizit kodierte symbolische Wissen von dem Sprachmodell gelernt werden kann. Entsprechende Prototypen und Sprachressourcen sowie kommerzielle Lösungen und Technologien

werden von etwa 800 universitären Forschungsgruppen und unabhängigen Einrichtungen sowie etwa 800 bis 1000 Unternehmen in Europa entwickelt. Speziell die kommerziellen Produkte werden entweder in bestehende Systeme integriert oder über Remote APIs zur Verfügung gestellt, sodass sie prinzipiell in beliebiger Hardware eingesetzt werden können. Neben den extrem großen Mengen an Sprachdaten, die für das Training von Sprachmodellen notwendig sind, werden für diesen Zweck sehr leistungsstarke Rechensysteme benötigt – beide Aspekte stellen für Akteur*innen, die auf diese Ressourcen keinen Zugriff haben, Flaschenhälse dar.

Die im Folgenden dargestellten Empfehlungen stammen u. a. aus der akademischen und industriellen Praxis, die die Teilnehmer*innen der Gruppe „Sprachtechnologie“ aus dem Arbeitsalltag kennen. Zudem wurden weitere Bedarfe und Ideen für Normung und Standardisierung in verschiedenen Fokusgruppen gesammelt, die u. a. mit den Konsortien verschiedener BMWK- und Projekte der Europäischen Union (EU) organisiert wurden.

Neben den hier betrachteten Sprachtechnologien für natürliche Sprache haben sich auch Kommunikationstechnologien entwickelt wie z. B. machine-to-machine communication, die an dieser Stelle nicht berücksichtigt werden.

Stand von Wissenschaft und Technik

Generalisierende Künstliche Intelligenz steht für eine neue Generation von KI, die Aufgaben lösen kann, für die sie nicht spezifisch trainiert wurde. Das Ziel der Entwicklung im Bereich generalisierender KI ist es, menschliches Denken in seiner Dynamik und Vielfalt nachzubilden, wobei dies aktuell nicht absehbar scheint. Fortgeschrittene Beispiele aus dem europäischen Raum mit Schwerpunkt auf neuronalen Netzen sind z. B. große KI-Sprachmodelle wie Generative Pretrained Transformer (GPT-3) (OpenaAI) oder Luminous (Aleph Alpha). Hybride Ansätze der generalisierenden Künstlichen Intelligenz, die einen Schwerpunkt auf symbolischer, funktionaler Wissensverarbeitung auf Basis aktiver Ontologien haben, repräsentiert z. B. OntoBroker (semafora systems). Neuronale Sprachmodelle werden einmalig mit riesigen Datenmengen trainiert, wodurch diese kontextuelles Weltwissen anstreben. Sie sind in der Lage, mit geringem menschlichem Input ein breites Spektrum von Texten zu verstehen und zu produzieren um somit verschiedenste informationsbasierte Arbeitsschritte zu unterstützen. Symbolische, funktionale Sprachmodelle extrahieren Sprache deterministisch z. B. aus für die Anwendung relevanten Texten, um das darin enthaltene Wissen ohne Verlust strukturiert zugreifbar zu machen.

In der akademischen sowie der industriellen Forschung und Entwicklung (F&E) fokussiert man sich neben anwendungsspezifischen Fragestellungen im Wesentlichen auf die Themen Erklärbarkeit/Transparenz, Skalierbarkeit und Metriken zur Bewertung der teilweise umfangreich benötigten Datenmengen. Der Bedarf, Methoden der Sprachverarbeitung in ihrer Qualität zu bewerten, ist nicht nur aus wissenschaftlicher und industrieller sowie Anwendersicht sehr hoch, sondern auch aus gesellschaftlichem Interesse. Die derzeit in der Forschung entwickelten Metriken und Verfahren unterliegen aber auch selbst einem Wandel, da einerseits die Anzahl der einzelnen Schritte einer prüfbaren Sprachverarbeitungskette zunehmen, andererseits die Aussagekraft der verwendeten Metriken wesentlich von den festgelegten Testdaten (Benchmarks) abhängt. Letztere sind zwar für einige Anwendungsbereiche vorhanden, jedoch von sehr unterschiedlicher Qualität.

Die zuvor aufgeführten Herausforderungen betreffen die verfügbaren Methoden jedoch durchaus unterschiedlich, so sind z. B. neuronale Verfahren stark bei der Skalierbarkeit und Flexibilität und symbolische/semantische Verfahren bei der Transparenz. Forschung und Industrie arbeiten daher daran, die Vorteile beider methodischen Richtungen in Hybridsystemen zu kombinieren.

Stand der Standardisierung

Im Bereich der Sprachtechnologien gibt es eine Reihe von Standardisierungsaktivitäten, die in Anhang 13.2 dargestellt sind. Für Künstliche Intelligenz arbeitet auf europäischer Ebene im Bereich CEN/CENELEC JTC 21 eine Working Group and Natural Language Processing für KI, die als Gremium für die erarbeiteten Handlungsempfehlungen empfohlen wird.

Stand der Regulierung

Der in Arbeit befindliche EU Artificial Intelligence Act (AI Act) adressiert alle Anwendungen, die mit Künstlicher Intelligenz arbeiten und einem Risiko, insbesondere einem hohen Risiko, unterliegen. Zu den Anwendungen, die als Hochrisiko angesehen werden, siehe Entwurf AI Act; Anhang III, gehören beispielsweise Anwendungen mit „biometrischer Fernwirkung“. KI mit Sprache, insbesondere wenn Sprache biometrisch verwendet wird, kann dann unter die Regulation des AI Act fallen und erfordert dann eine passende Zertifizierung. Der AI Act und seine standardisierungsrelevanten Anforderungen werden in Kapitel 1.4 dargestellt. Der Entwurf der EU-Kommission zu Standardisierungsthemen liegt ebenfalls vor und sollte bezüglich Anforderungen der Sprachtechnologien geprüft werden.

4.1.2.6 Bildgebende Sensorik

Unter dem Thema „bildgebende Sensorik“ werden in diesem Kapitel alle KI-Anwendungen gebündelt, die sich mit Daten ortsauflösender Sensorik befassen. Damit sollen neben Kameraeinzelbildern und Bildfolgen im sichtbaren Spektrum weitere Spektralbereiche (beispielsweise nahes und fernes Infrarot) oder andere Sensorprinzipien abgedeckt werden, die im engeren oder weiteren Sinne bildgebend sind und bei denen verwandte KI-Verfahren Anwendung finden – etwa Laserscanner, Radarsignale oder medizinische Tomografie. Das Kapitel spricht vereinfachend einheitlich von „Bilddaten“, „Einzelbildern“ und „Bildfolgen“.

Die zugeordneten KI-Verfahren wiederum sind grob unterscheidbar in drei Kategorien:

→ Verfahren, die Einzelbilder oder Bildfolgen zu abstrakteren Informationen verarbeiten (beispielsweise durch Objektdetektion oder Segmentierung),
→ Verfahren, die aus abstrakten, parametrischen Eingabedaten realitätsnahe Bilddaten synthetisieren (d. h. künstlich erzeugen),
→ Verfahren, die Bilddaten in andere Bilddaten vergleichbarer Abstraktion umwandeln, beispielsweise durch sogenannte „Style Transfers“ von Sommer- zu Winterfotografien.

Damit hat bildgebende Sensorik große Relevanz für heterogene Anwendungsfelder der Hochrisiko-KI (insbesondere im Sinne des EU-AI Act [4], vgl. dazu auch Kapitel 4.3), beispielsweise in den Bereichen automatisiertes Fahren, Medizintechnik oder zivile Sicherheit, wie etwa der Personen- und Gesichtserkennung. Der geplante EU AI Act adressiert insbesondere Bildverarbeitung im Sinne von biometrischer Identifikation sowie Hochrisikoanwendungen in Anhang 3. Diese sind die Grundlage dafür, dass der AI Act direkt zur Geltung kommt bzw. über die dafür geplanten harmonisierten Standards inklusive Konformitätsbewertung wirkt und eine entsprechende Zertifizierung ansetzt. Dementsprechend finden sich vertiefte Darstellungen und Handlungsbedarfe zu diesem Themenfeld insbesondere in den Kapiteln 4.6 und 4.7 sowie in Kapitel 1.4. Das folgende Kapitel fasst lediglich übergreifende Herausforderungen in diesen Bereichen grundlegend zusammen.

In den genannten KI-Bereichen werden mit modernen Verfahren des Deep Learning erstmals Ergebnisse der komplexen Bildinterpretation erreicht, die selbst das menschliche Leistungsvermögen übertreffen [84]. Dieser Umstand hat Implikationen sowohl für die Potenziale der Anwendung als auch für Risiken in der menschlichen Beurteilung entsprechender Systeme.

Status quo
Bestehende grundlegende Standards zum Umgang mit entsprechenden Sensordaten betreffen einerseits das Rohdatenformat (beispielsweise JPEG-Bildkompression, das Digital-Imaging-and-Communications-in-Medicine (DICOM)-Format für Robotiksensordaten als De-facto-Standard) sowie das Format von Annotationen für maschinelle Lernverfahren, beispielsweise der ASAM-Standard OpenLABEL.

Insbesondere die Anwendung maschineller Lernverfahren, spezifisch im Rahmen des Deep Learning, bringt jedoch komplexe Herausforderungen mit sich. Diese liegen einerseits in Eigenschaften der KI-Systeme und ihrer Entwicklung, andererseits jedoch, durch den unmittelbaren und umfassenden Datenbezug entsprechender ML-Verfahren, auch in den benötigten Datenumfängen und damit einhergehend dringenden datenschutztechnischen Fragestellungen. Ursächlich für diese Herausforderungen ist primär, dass im gegebenen Anwendungsfeld immense Datenmengen zur Parametrierung und Evaluierung von KI-Methoden benötigt werden. Dabei gilt etwa, dass zu einer hohen Anzahl an benötigten Trainingsbeispielen (wie bei ML-Verfahren üblich) auch ein erheblicher Datenumfang eines einzelnen Beispielbilds hinzukommt, bestehend aus meist Tausenden bis mehreren Millionen Pixeln und mehreren Farb- oder Informationskanälen. Je nach Anwendung ergeben sogar erst Bildfolgen mehrerer Bilder ein einzelnes Trainingsbeispiel. Damit sind in bildgebender Sensorik immense Mengen an Rohdaten erforderlich, zu denen – anwendungsabhängig – auch vergleichbar umfangreiche Annotationen (Labels) für Training und Test gehören. So gibt es zwar Anwendungen, die beispielsweise ein gesamtes Einzelbild oder eine Bildfolge nur in eine einzige Kategorie oder Klasse zuordnen; jedoch auch Anwendungen, die eine pixelfeine Annotation von Objektklassen erfordern (z. B. Segmentierungsaufgaben).

Wesentliche Herausforderungsfelder, die dementsprechend mit der Anwendung einhergehen, sollen im Folgenden kurz motiviert werden.

Herausforderungsfeld „Zugang zu Perzeptionsdaten mit ausreichender Datenqualität“
Die Bereitstellung der erforderlichen Umfänge an Trainings-, Test- und Validierungsdaten stößt einerseits auf unterschiedliche technische Herausforderungen. Insbesondere für die

Verarbeitung hochauflösender Daten sind oft erhebliche Mengen an entsprechend hochauflösenden Annotationen als Trainingsgrundlage erforderlich. Damit gehen nicht nur erhebliche finanzielle Aufwände einher, sondern auch die Herausforderung, die Qualität eines entsprechenden Datensatzes für spezifische Anwendungen (Klassifikation, Segmentierung etc.) zu gewährleisten. Für Hochrisiko-KI-Anwendungen fordert beispielsweise der geplante AI Act [4] „hinreichende Relevanz sowie Repräsentativität, Fehlerfreiheit und Vollständigkeit in Hinblick auf die beabsichtigte Anwendung". Diese Zielgrößen für einen Datensatz, bestehend etwa aus Tausenden oder Hunderttausenden Einzelbildern, zu bewerten, stellt Entwickler und Prüfeinrichtungen vor erhebliche Herausforderungen. Es gibt bereits erste Verzeichnisse von Sonderfällen für die Anwendung der Bildverarbeitung wie z. B. CV-HAZOP [85]. Menschliche Annotationsfehler können hier weitreichende Folgen haben, die nur schwer zu erkennen sind. Doch auch die alternative Praxis, Annotationen für Trainings- und Testzwecke nicht mehr menschlich anfertigen zu lassen, sondern mithilfe von KI-basierten Werkzeugen, wirft Fragen der Nachweisführung auf. Ebenso komplex ist die Nachweisführung, dass ein Rohdatensatz beispielsweise von Straßenszenen „repräsentativ" oder „vollständig" ist, insbesondere frei von unzulässigem Bias.

Diese Herausforderungen multiplizieren sich in KI-Anwendungen, die Sensordaten in einer „offenen Welt" erheben, etwa im automatisierten Fahren (vgl. auch Kapitel 4.6), und nicht in einer kontrollierten Umgebung wie beispielsweise oft im Umfeld von Medizin oder Bauteilprüfung. Hier weisen Sensordaten eine erhebliche Variabilität auf, deren korrekte Abbildung in Trainings- und Testdatensätzen sicherheitskritisch sein kann. Beispielsweise enthalten die prominenten Automotive-Datensätze KITTI [86] und Cityscapes [87] noch keine E-Scooter. Ferner sind Perzeptionsdatensätze je nach Anwendungsfall in unterschiedlicher Menge vorhanden. Beispielsweise konnten für den Automotivbereich 60 und nur zwei Perzeptionsdatensätze für den Eisenbahnbereich gefunden werden [88].

Im Umfeld des automatisierten Fahrens lässt sich spezifisch diese Herausforderung auch in der Normung ablesen, im Übergang der ISO-26262-Reihe [455] (Funktionale Sicherheit von Straßenfahrzeugen) und DIN EN 50657:2017 [89] (Software für Schienenfahrzeuge) hin zur ISO 21448:2022 [90] (Sicherheit der Sollfunktion). Dieser entspricht einem wesentlichen Perspektivwechsel weg von der Betrachtung primär stochastischer „Ausfälle" einer Komponente, hin zur Betrachtung der Robustheit eines Gesamtsystems in seiner Umwelt gegenüber potenziell auch unerkannten Herausforderungen (sogenannten unbekannten unsicheren Zuständen und mithin „unbekannten Unbekannten"), die wesentlich mit der offenen Welt einhergehen. Damit einhergehend ist der ASAM-Standard OpenODD (Operational Design Domain) zu nennen, dessen Ziel es ist, zulässige Anwendungsfelder für eine Fahrfunktion möglichst exakt spezifizierbar zu machen.

Entsprechend fortgeschrittene Betrachtungen im Umfeld der Normung fehlen bislang in anderen Branchen, beispielsweise für KI-basierte Baumaschinen, der Mensch-Roboter-Kollaboration oder der zivilen Sicherheit. So legen existierende Normungsaktivitäten zu KI-Methoden wie ISO/IEC TR 24029-1:2021 [91] und ISO/IEC 24029-2 [92] (die auch einen Überblick über Methoden beinhalten) die Verwendung formaler Verifikationsansätze nahe, die angesichts von Umfang und Komplexität dieser Daten nur schwer praktisch auf das Feld bildgebender Sensorik angewendet werden können. Häufig müssen daher empirische Testverfahren die Basis für Robustheitsanalysen bilden (z. B. Common Corruptions & Adversarial Attacks). Hier kann Standardisierung einen wesentlichen Beitrag leisten, indem sie für industrielle Anwendungen Leitlinien aufstellt, die u. a. folgende Fragen im Bereich der Robustheitsanalyse mit empirischen Testverfahren adressieren:

- → Wie sieht die „optimale" Teststrategie / der „optimale" Testprozess mit empirischen Testverfahren aus?
- → Wie kann man verschiedene Testverfahren aggregieren?
- → Wie kann man eine Risikoeinschätzung aus empirischen Testergebnissen extrahieren?
- → Wie wird ein „diverses Set" von Testverfahren entwickelt?
- → Wie definiert man „Erfolg" bei einem adversarialen Angriff auf ein KI-Modell?

Herausforderungsfeld Synthetisierung

Synthetisierung von Bilddaten, also deren künstliche Erzeugung, kann u. a. einen wesentlichen Beitrag dazu leisten, bestehende Datenbedarfe zu decken. Gleichzeitig wirft sie aber auch eigene Herausforderungen und Fragen auf in Abhängigkeit von Zweck und Technik der Synthetisierung.

Einerseits kann Synthesierung genutzt werden, um Test-, Trainings- und Validierungsdaten für KI-Verfahren bereitzustellen. Hier können etwa klassische Verfahren der Computergrafik zum Einsatz kommen, um realitätsnahe Bilddaten zu erzeugen. Ein wesentlicher Faktor dabei ist, dass bei entsprechend synthetisierten Daten die Annotation (also beispielsweise die im Bild enthaltenen Objekte und deren Positionen) in der Regel ebenfalls direkt vorliegen und somit der Annotati-

onsaufwand entfällt. Auch können Rohdaten seltener oder risikoreicher Ereignisse simulativ erzeugt werden. Aber auch KI-Verfahren, konkret Verfahren des Maschinellen Lernens, können zu Synthetisierungszwecken genutzt werden. Ein erheblicher Durchbruch liegt in der Entwicklung von GANs als maschinelles Lernprinzip, das beispielsweise fotorealistische menschliche Gesichter erzeugen kann, die selbst für Menschen nicht ohne Weiteres von realen Fotos unterscheidbar sind. Werden jedoch synthetische Daten einer beliebigen Erzeugungsmethode für Training und Test von KI-Verfahren genutzt, stellen sich auch hier wesentliche Fragen an Repräsentativität und Korrektheit, insbesondere jedoch auch an den Realismusgrad der synthetisierten Daten, der insbesondere für Hochrisikoanwendungen nachzuweisen ist. Auch hier besteht ein erheblicher Bedarf, analog zum Bedarf der Qualitätsbewertung von Trainingsdatensätzen, spezifische Kriterien der Qualitätsbewertung synthetisch erzeugter Daten oder Synthetisierungsmethoden mit den spezifischen Herausforderungen zu standardisieren.

Darüber hinaus können synthetische Daten auch für Nicht-KI-Anwendungen genutzt werden, beispielsweise für Kunst oder Unterhaltung. Die dazu verwendeten Verfahren, insbesondere auch hier GANs, können allerdings vielfach mit geringem Aufwand zu missbräuchlichen Zwecken genutzt werden, beispielsweise im Rahmen sogenannter „Deep Fakes“, bei denen Fotos von Personen täuschend echt in Videoaufnahmen anderer Personen eingefügt werden können, und dabei sogar deren Mimik und die Beleuchtung der Szene realistisch abbilden. Diese Herausforderung betrifft nur zu einem vergleichsweise geringen Teil die Normung und Standardisierung, sondern erfordert primär die Steigerung gesellschaftlicher Kompetenz im Umgang mit Bilddaten. Jedoch ist eine relevante Perspektive, dass auch Authentifizierungsverfahren beispielsweise in Unternehmensprozessen, die bisher beispielsweise auf Foto- oder Videodaten beruhten, dieser neuen Entwicklung Rechnung tragen sollten – beispielsweise durch standardisierte Richtlinien, mit welchen Prüfungs- oder Mehrfaktor-Authentifizierungsschritten entsprechende Deep Fakes ausgeschlossen werden können.

4.1.3 Normungs- und Standardisierungsbedarfe

4.1.3.1 Allgemein

Bedarf 01-01: Sektorübergreifende Normung von Begriffen

Gerade durch die querschnittliche Bedeutung von „KI“ als Technologie führen die benannten Bedeutungsunterschiede in interdisziplinären Diskussionen oft zu erheblichen Missverständnissen. Das erzeugt Reibungsverluste auch ohne inhaltlichen Dissens und entsprechend ohne inhaltliche Fortschritte. Da die Operationalisierung von KI und KI-Diskussionen vermehrt sektor- und domänenübergreifende Maßnahmen erfordert, wird erwartet, dass gemeinsame Begrifflichkeiten für diese ein notwendiges Fundament darstellen.

Wie im Glossar deutlich wird, gibt es bei verbreiteten Begriffen (beispielsweise „bias“, „safety“) mitunter erhebliche Abweichungen in Standards und Konventionen unterschiedlicher Domänen oder Sektoren. Es wird vorgeschlagen, sektorübergreifend vereinheitlichte Definitionen zu schaffen, um eine übergreifende Terminologie gerade in KI-Debatten sicherzustellen.

Bedarf 01-02: Verwendbarkeit der Normenreihe ISO/IEC 5259 [39] für sektorspezifisches Datenqualitätsmanagement

Die Verwendung der ISO/IEC-5259-Reihe [39] als gemeinsamer Ausgangspunkt für vertikale Standardisierungsaktivitäten im Bereich Datenqualität erlaubt es, auf ein gemeinsames Gerüst zurückzugreifen und Terminologie, Konzepte und Prozesse für Datenqualitätsmanagement sektorübergreifend zu beschreiben.

Durch die Initiierung der Normenreihe der ISO/IEC-5259-Reihe [39] sind die Themen Datenqualität und Datenmanagement in der internationalen Standardisierung zumindest allgemein adressiert. Dennoch ist zu erwarten, dass für spezifische Sektoren und Anwendungen verschärfte und ggf. andere als die oben genannten Qualitätskriterien relevant werden. Auch Qualitätsmanagementprozesse müssen sektorspezifisch implementiert und ggf. angereichert werden. Somit wird empfohlen, in der vertikalen Standardisierung zum Datenqualitätsmanagement zu prüfen, inwieweit die ISO/IEC-5259-Reihe [39] als allgemeine Referenz herangezogen werden kann und inwieweit sektorspezifische Adaptionen notwendig werden.

Bedarf 01-03: Erstellung einer Technologie-Roadmap für KI
Eine technologische Roadmap für KI-Entwicklungen kann eine wertvolle Grundlage darstellen, um Normungsbedarfen eine konkretisierte Zeitschiene aus technologischen Entwicklungen und Bedarfen gegenüberzustellen und damit den Fokus der Normungsroadmap in dieser Hinsicht zu schärfen. Während KI-Entwicklungen im Allgemeinen sehr dynamisch sind, sind Trends in Verfahren gerade dort früh absehbar, wo sie in kritische Produktbereiche, insbesondere auch mit Blick auf ethische Abwägungen beim KI-Einsatz, spielen (Beispiel: der Einsatz neuronaler Netze für Perzeption im automatisierten Fahren). Entsprechende Entwicklungen können mit vertretbarer Robustheit und zunächst unabhängig von konkreten Normungsbedarfen abgeschätzt werden, gleichzeitig kann eine solche Darstellung dazu beitragen, Normungsbedarfe schärfer und entlang Markt und Technologie zu erkennen.

Ergänzend zu der in Kapitel 4.1.1.1 skizzierten Klassifikationsmethodik der KI wird empfohlen, Arbeiten zur Erstellung einer Technologie-Roadmap zu fördern, die augenblickliche Technologietrends in der KI zusammenfasst und Empfehlungen für eine perspektivische Weiterentwicklung des Standorts Deutschlands gibt.

Bedarf 01-04: Prüfstandard für KI-Systeme in Anlehnung an die CC
Da die CC ein weltweit akzeptierter Ansatz zur Sicherheitsevaluation von IT-Systemen darstellt, der von Prüflaboren und Zertifizierungsstellen angewendet wird, wird so Mehraufwand bei der Produktzertifizierung von KI-Systemen vermieden bzw. minimiert, da auf bewährte Verfahrensweisen zurückgegriffen werden kann.

Zur Prüfung und Evaluation von KI-Systemen soll ein horizontaler Prüfstandard entwickelt werden, der sich in Terminologie, Methodik und Strukturvorgaben an die Dokumente zu den Common Criteria anlehnt.

Bedarf 01-05: Anforderungen an zertifizierende Stellen
Erforderliche Kompetenzen von Auditoren bzw. der Zeitaufwand für ein Audit gemäß ISO/IEC 42001 [27] unterscheiden sich ggf. von Audit-Anforderungen in anderen Bereichen.

Formulierung von Anforderungen an die Zertifizierung gemäß ISO/IEC 42001 [27], die durch zertifizierende Stellen erfüllt werden müssen. Ein Projektvorschlag von deutscher Seite zu diesem Thema ist in Vorbereitung; die Projektdurchführung muss jedoch von deutscher Seite maßgeblich unterstützt werden.

Bedarf 01-06: Einheitliche Form der Beschreibung von KI-Lösungen
Ergänzend zum Entwurf des EU AI Act sollte eine einheitliche Form der Beschreibung von KI-Lösungen verfügbar sein. Als Grundlage hierfür kann Kapitel 4.1.1.1 der hier vorliegenden Deutschen Normungsroadmap KI verwendet werden. Auf der Basis einer solchen wissenschaftlich fundierten Beschreibung der benutzten KI-Technologien können auch die entsprechenden Testverfahren passgenau detailliert werden. Aufwände der Regulierung und Zertifizierung sinken daher, während die Qualität steigt. Gleiches gilt für die Beschreibung ganzer KI-Anwendungen, in denen z. B. mehrere KI-Technologien zum Einsatz kommen:

- → Damit könnte auch die geforderte Technische Dokumentation (Entwurf AI Act Art. 11) deutlich verbessert werden, womit Transparenz (Entwurf AI Act Art. 13) und Vertrauenswürdigkeit der KI steigen.
- → Die European AI Database zum Management der in der EU gelisteten „Hochrisiko“-Anwendungen würde ebenfalls von einer einheitlichen Taxonomie zur Beschreibung von KI profitieren.
- → Weiterhin ist es denkbar, dass unter Nutzung der vorgeschlagenen KI-Klassifikation in Zukunft einheitliche „harmonisierte europäische Label“ entstehen, die die Verbreitung von Transparenz und Qualität von KI weiter fördern und beschleunigen.
- → Ein weiterer wichtiger Punkt ist, dass Conformity Assessments durch einheitliche Klassen von KI-Anwendungen einfacher und leichter standardisierbar werden. Gleiches gilt für die Marktüberwachung.

Es wird deshalb empfohlen, ein Standardisierungsprojekt zur Klassifizierung von KI-Systemen auf europäischer Ebene zu initiieren.

4.1.3.2 Ethik

Bedarf 01-07: Schnittstellen des Entwicklungsprozesses von KI gestalten
Standardisierte Schnittstellen und ein modulares Modell typischer KI-Bausteine kann die austauschbare Entwicklung und Einzelbewertung nach standardisierten Kriterien ermöglichen und damit zur übergreifenden Nutzbarkeit, zur Übertragbarkeit von Zulassungen und zur Transparenz beitragen. Entsprechende Methoden zur Einsichtnahme in Modelle und

Datensätze fordert auch der Entwurf zum AI Act [4]. Darauf aufbauend können standardisierte Vorgehensmodelle geschaffen werden (vgl. beispielsweise [93]), die die Bereitstellung entsprechender Schnittstellen als reguläres Artefakt der Entwicklung integrieren und Zusatzaufwände minimieren. Die dadurch entstehende Vergleichbarkeit des Schnittstellenmanagements von unterschiedlichen Institutionen schafft Orientierung und zahlt so auf den Wert Selbstbestimmung i. S. v. selbstbestimmter Nutzung ein.

Standardisierte Schnittstellen in KI-Systemen sollen bereits in der Entwicklungsphase externen Prüfern Einblick etwa in Trainingsdatensätze und Modelle geben und KI-Subsysteme, wo möglich, auf gängige einheitliche Funktionsbeschreibungen zusammenführen, um Entwicklung, Prüfung und Einsatz zu vereinfachen, insbesondere im Hinblick auf Ziele der Ethik und Vertrauenswürdigkeit (beispielsweise hinsichtlich Nachvollziehbarkeit, Authentizität der Daten, Transparenz). Es sollten standardisierte Rollenbeschreibungen von KI-Komponenten und von Akteur*innen definiert werden. Ferner soll eine standardisierte Beschreibung des Zusammenspiels der einzelnen Komponenten untereinander sowie im Gesamtkontext (inklusive Nicht-KI-Systemteile und Systemumgebung) geschaffen werden. Es ist zu definieren, welcher Abstraktionsgrad dabei praktisch ratsam ist – beispielsweise, um mit Rücksicht auf Datenschutz, Datensparsamkeit und Datenumfang nicht alle Bestandteile eines Datensatzes offenlegen zu müssen, sondern lediglich abstrahierte Merkmale.

Bedarf 01-08: Gestaltung der Inhalte einer Quality Backward Chain
Um Systeme künstlicher Intelligenz auch in ihrer ethischen Dimension während ihres Einsatzes evaluieren und ggf. Entscheidungsgrundlagen modellieren zu können, ist der Einsatz einer Quality Backward Chain zu empfehlen. Diese gewinnt im Rahmen des Einsatzes Felddaten, welche ein Urteil über ethische Entscheidungen des Systems ermöglichen. Grundlegende Korrekturen des Systems sind hierbei nicht vorgesehen, vielmehr soll verhindert werden, dass auf Schäden durch den Einsatz nicht (angemessen) reagiert werden kann. Die Quality Backward Chain liefert Daten für die nachträgliche Beurteilung möglicher Fehlentscheide und hilft dabei sowohl dem Anbieter als auch dem Anwendenden.

Verpflichtende Inhalte im Rahmen der Felddatengewinnung im Sinne einer Quality Backward Chain, die neben technischen auch ethische Aspekte systematisch abdecken muss, bedürfen einer Normung sowie einheitlicher Datenformate, um künftige Meldepflichten zu sichern. Damit soll gewährleistet werden, dass die Option, Meldungen zu machen, möglichst niederschwellig und für alle Benutzergruppen möglichst gut erreichbar ist. Damit soll hinsichtlich der Wertebene eine demokratische Nutzung sichergestellt sein. Ebenso ist dies hinsichtlich Interoperabilität erforderlich, um eine freie Nutzung von Produkten, Dienstleistungen und Systemen abseits von Monopolen zu ermöglichen und User*innen auch in dieser Hinsicht in ihrer souveränen Entscheidung zu unterstützen.

Bedarf 01-09: Möglichkeiten zur Reevaluierung vorsehen
Die ethische Reevaluierung von KI-Systemen findet anhand ihrer Kernwerte statt. Diese Kernwerte gilt es vorher im Entwicklungsprozess durch das Unternehmen im Rahmen eines Stakeholderprozesses zu identifizieren. Anhand der erfolgten Abwägung von Werten stuft das Unternehmen intern Ergebnisse bzw. Entscheidungen des KI-Systems in seiner ethischen Dimension im Betrieb ein, aber auch schon im Rahmen des Entwicklungsprozesses. Felddaten aus einer Quality Backward Chain können diese Bewertung unterstützen. In die Prüfung sind die relevanten Stakeholder einzubinden. Sie kann durch ein Expert*innengremium, z. B. ein Expert Review Board, oder anderes geschultes Personal vollzogen werden. Die Prüfung schließt mit ein, dass auch die Unternehmensprozesse in Hinblick auf die Gewährleistung ethischer Prinzipien betrachtet und ggf. korrigiert werden. Sollte ein Verstoß gegen o. g. Kernwerte entdeckt werden, so ist eine größere Prüfung der Prozesse und Datengrundlage nötig. Ebenso wäre eine Meldepflicht analog zu Datenschutzverstößen denkbar. Die Reevaluierung findet bedarfsgebunden oder in festen Abständen statt, beispielsweise alle drei Jahre. Kernelemente dieses Prozesses sind bereits in der ISO/IEC 38507:2022 [26] adressiert, wobei zum Großteil die Kernziele des Unternehmens in den Vordergrund gestellt werden und ethische Aspekte eher als Nebenanforderung auftreten. Dabei ist zudem nicht herausgearbeitet, welche konkreten Inhalte in Bezug auf die ethische Bewertung berücksichtigt und in welchem Umfang diese betrachtet werden sollen.

Dokumentationspflichten und Zeitabstände für verpflichtende Reevaluierungen sind zu normen.

Bedarf 01-10: Normung eines Konzepts für Privacy Ethical Design
Privacy Ethical Design unterlegt alle Systeme mit dem Grundsatz der Privatsphäre des Einzelnen. Dabei geht es über das Konzept der Privatsphäre an sich hinaus und weist ihr eine klare ethische Dimension zu, bei der nicht nur direkte

Einflüsse, sondern auch indirekte Einflüsse auf die Bedarfe des Anwendenden berücksichtigt werden. Damit wird ein Grundvertrauen in neue Technologien gefördert und dadurch die Marktakzeptanz erhöht. Auch Interoperabilität zwischen verschiedenen Anbietern, wie beispielsweise SSO, kann durch Privacy Ethical Design mehr Anwender*innen ansprechen. Dies kann unter Berücksichtigung des aktuell im ISO/IEC JTC 1/SC 42 initiierten Projekts zu einen MSS für KI (siehe Kapitel 4.1.3, Bedarf 1 „Unterstützung der internationalen Standardisierungsarbeiten zu einem MSS für KI“) erfolgen, indem die Erklärbarkeit von KI-Systemen in den Anforderungskatalog des entstehenden Dokuments aufgenommen wird, sowie durch eine Ausweitung des Risikobegriffs auf ethische Risiken, wie sie bereits im Projekt ISO/IEC 23894:2022 [25] Risk Management vorgenommen wurde.

Um effektives Privacy Ethical Design zu fördern, gilt es, ethische Risiken gezielt und systematisch zu beleuchten. Im Rahmen eines Risikomanagementprozesses sollen sie identifiziert und analysiert werden, um sie durch gezielte Maßnahmen zu mitigieren. Dies kann beispielsweise in Form und Umfang einer möglichen Dokumentationspflicht gestaltet werden – zur Förderung von Transparenz und Verhinderung reiner Scheinmaßnahmen. Ein solches Vorgehen zahlt u. a. auf den Wert der Nachvollziehbarkeit ein. Ein weiteres Beispiel wäre die Verbesserung der Benutzerschnittstelle im Hinblick auf Privacy-Einstellungen, um für die Beteiligten möglichst gute Möglichkeiten zu schaffen, Privacy effektiv und intuitiv umzusetzen.

Bedarf 01-11: Zweckbindung von Daten gestalten
Um ein für alle Parteien transparentes Agieren im Interesse vertrauenswürdiger KI-Entwicklung zu ermöglichen, gilt es, die Zweckbindung von Daten weiter auszugestalten. Nach Art. 5 Datenschutz-Grundverordnung (DSGVO) dürfen personenbezogene Daten nur für „festgelegte, eindeutige und legitime Zwecke erhoben werden“ sowie „nicht in einer mit diesen Zwecken nicht zu vereinbarenden Weise weiterverarbeitet werden“. Ausnahmen gelten hierbei nach Art. 89 DSGVO für „im öffentlichen Interesse liegende Archivzwecke, für wissenschaftliche oder historische Forschungszwecke oder für statistische Zwecke“. Hier kann Normung ansetzen und im Rahmen der gesetzlichen Leitplanken der DSGVO eine innovative Datennutzung fördern, durch die Unternehmen in der Lage sind, neue Produkte auf Basis ihrer Stammdaten zu entwickeln, ohne die Rechte der Verbraucher*innen zu verletzen. Eine gute Option bietet sich, zu diesem Punkt den Dialog zu einschlägigen Gesetzesvorschlägen der Europäischen Kommission (DSA, DGA) zu pflegen, um die Regulierungsabsichten in dieser Hinsicht stimmig fortzuführen. Hierbei soll über gemeinsamen Austausch bestenfalls auch die Expertise der Aufsichtsbehörden einbezogen werden. Gleichzeitig sollen die Verbraucher*innen jederzeit in der Lage sein, eine angemessene Übersicht zu erhalten, zu welchen Zwecken ihre Daten verwendet werden. Normung kann hierbei Unternehmen und Institutionen unterstützen, ein erforderliches Consent Management zu entwickeln und zu integrieren.

Für eine sichere und innovative Zweckbindung von Daten kann Normung einheitliche Dokumentationen und Einverständniserklärungen fördern, welche Anwender*innen und Anbieter*innen schnell und unkompliziert Einsicht in die möglichen Verwendungszwecke bieten.

Bedarf 01-12: Gestaltung des Wertesystems
Der Bedarf geht davon aus, dass Maschinen derzeit kein ethisches Verhalten zugeordnet werden kann; wohl aber die technologische Implementierung implizit oder explizit Rückschlüsse auf ethische Annahmen im Entwicklungsprozess zulassen. Dazu können beispielsweise die Mechanismen zählen, mit denen z. B. Fairness oder Safety (in Bezug auf Risikoabwägungen) umgesetzt werden. In welchem Grad definierte Werte über KI zur Umsetzung kommen, soll use-case-abhängig abstufbar sein, verbunden mit einer begründenden Dokumentation zu dieser Entscheidung. Hier wird empfohlen, auf eine Vereinheitlichung der Darstellungen hinzuarbeiten, die Entwicklungsrisiken mindert und gesellschaftliche Transparenz im Betrieb schafft. Orientiert sein soll diese Vereinheitlichung an der einschlägigen Normung sowie dem gesellschaftlichen Diskurs zum betroffenen Phänomen. Zudem ließe sich dies in den Leitplanken von Unternehmen, beispielsweise dem Code of Conduct, verbindlich für alle Mitarbeitenden integrieren (vgl. ISO/IEC 38507:2022 [26], ISO/IEC 42001 [27], Letztere derzeit in Ausarbeitung). Es gibt zahlreiche Phänomene, bei denen die Frage nach einem Wertesystem von KI eine Rolle spielt. Beispielhaft genannt sei die Frage, inwieweit sich KI in Recruiting-Prozessen zur Systematisierung und Kategorisierung von Bewerbungsdaten einsetzen lässt. Berührt sind an dieser Stelle u. a. die Aspekte Diversität und Fairness. Als weiteres veranschaulichendes Beispiel lässt sich die Frage nennen, ob Risikoabwägungen im Betrieb (wie beispielsweise im Fall von risikobasierten Planungsalgorithmen im automatisierten Fahren bzw. Dynamic Risk Management [94], [95]) gegenüber der Vorgabe der Ethikkommission automatisiertes und vernetztes Fahren des BMVI [96]) zulässig sein können, und wenn ja, welche konkreten Anforderungen sich daraus an Technik ableiten lassen. Es wird eingeschätzt, dass diese Abbildung ethischer Konzepte

in maschinenlesbare Form noch nicht unmittelbar in der Normung umgesetzt werden kann, sondern eine fokussierte interdisziplinäre Vorarbeit durch die Forschung voraussetzen (siehe auch Forschungsfelder Artificial Ethics bzw. Artificial Moral).

Wo ethische Konzepte die Entscheidungen von KI-Systemen zur Laufzeit beeinflussen sollen, ist deren Formalisierung und Abbildung in maschinennutzbarer Form erforderlich, beispielsweise in Form von Ontologien oder in Form von Rechenprinzipien für zulässige Risikoabwägungen.

Bedarf 01-13: Verbesserter und niederschwelligerer Überblick über das Zusammenspiel zwischen Kritikalitätsstufen und zugehörigen Anforderungen (speziell bei KI-Systemen mit geringem Risiko)
Um KI-Systeme bezüglich ihrer Kritikalität schnell einordnen zu können und die damit verbundenen Anforderungen gut erfassen zu können, wären für Herstellende klar strukturierte Vorgaben hilfreich. Das gilt insbesondere für die Frage, welche Anforderungen KI-Anwendungen mit niedrigem Risiko erfüllen sollten, um die gesetzlichen Vorgaben zu erfüllen, aber auch um ein hohes Maß an Vertrauenswürdigkeit zu erreichen. Der geplante AI Act gibt zwar für den Bereich der EU eine Einordnung in bestimmte Klassen, indem er z. B. verbotene Bereiche oder auch Hochrisikosysteme definiert, wobei die Einordnung primär gemäß dem Anwendungsgebiet und weniger nach dem für das jeweilige konkrete Produkt entstehende Risiko erfolgt. Gerade für den Bereich der weniger kritischen Systeme verbleiben jedoch wenig konkrete Anforderungen, sodass die Herstellenden in diesem Fall kein klares Bild bekommen, welche Anforderungen umzusetzen sind. Dieser Effekt wird dadurch verstärkt, dass es inzwischen vielfältige andere Gesetze auf EU-Ebene gibt, wie u. a. die Datenschutz-Grundverordnung, der Digital Service Act, der geplante Data Act oder auch die Grundrechtecharta der EU, die weitere wichtige Anforderungen liefern, die bei der Entwicklung von KI-basierten Systemen eine zentrale Rolle spielen. Auch auf Seite der Benutzer*innen wird es damit unübersichtlich, wie sie die Systeme einzuordnen haben, was ein vertrauenswürdiges System ausmacht und welche Anforderungen diese in welcher Weise erfüllen.

Eine bessere Transparenz und Übersichtlichkeit in Bezug auf die unterschiedlichen Stufen der Kritikalität (auch jenseits der Einordnung im geplanten AI Act [4] und der damit verbundenen Anforderungen) soll geschaffen und in entsprechenden Normen verankert werden. Es soll auf niederschwellige Weise vermittelt werden, was vertrauenswürdige KI ausmacht, wie die Systeme einzuordnen sind und welche Anforderungen aus welchen Gesetzen umzusetzen sind.

Konkret beinhaltet das die folgenden Punkte:
- → Niederschwellige und für Hersteller*innen und Benutzer*innen transparente Zuordnung von KI-Anwendungen in Hinblick auf ihre Kritikalität
- → Für Hersteller*innen: gezielte Klärung, welche Anforderungen aus welcher Gesetzgebung für welche Anwendungen bzw. Kritikalitätsstufen umzusetzen sind, um gesetzeskonforme und vertrauenswürdige KI-Systeme entwickeln zu können. Durch geeignete Normen/Werkzeuge soll ein gut erfassbarer Überblick geschaffen werden, der die Zusammenhänge zwischen den Anforderungen, den zugehörigen Gesetzen sowie den für den jeweiligen Use Case erforderlichen Schritten aufschlüsselt.
- → Für Benutzer*innen: schneller und niederschwelliger Einblick in die unterschiedlichen Kritikalitätsstufen und deren Anforderungen auf einem verständlichen Niveau, um die Vertrauenswürdigkeit von KI-Systemen in geeigneter Weise erfassbar zu machen.

4.1.3.3 Quanten-KI

Bedarf 01-14: Künstliche Intelligenz (insbesondere Maschinelles Lernen) und Quantencomputing im Kontext der IT-Sicherheit
Der Einsatz von Quantencomputern hat das Potenzial, die Praxis in der Künstlichen Intelligenz, insbesondere des Maschinellen Lernens, nachhaltig zu beeinflussen. Obwohl der Entwicklungsstand aktueller Quanten-Hardware der tatsächlichen Anwendung von Quanten-KI zur Bearbeitung praktischer Problemstellungen noch starke Limitationen setzt, sind hier in den nächsten Jahren deutliche Fortschritte zu erwarten. Auch hinsichtlich des entsprechenden Quanten-Softwarestacks werden zahlreiche nationale und internationale Förderprojekte engagiert vorangetrieben. Methoden der Quanten-KI spielen an dieser Stelle als Komponenten der ersten wesentlichen Anwendungen für Quantencomputer und insbesondere auch bereits in der NISQ-Ära eine ganz entscheidende Rolle.

Die Entwicklungen im Bereich der Quanten-KI und insbesondere auf dem Gebiet des Quantum Machine Learning (QML) müssen in den nächsten Jahren fortlaufend beobachtet und eingehend bewertet werden. Hier ergibt sich die große Chance, frühzeitig auf die sichere Ausgestaltung der neuen Technologie hinzuwirken und sowohl Potenziale als auch

Risiken bei deren Verwendung zu erkennen und zu behandeln. Mit Blick auf den aktuellen Stand der Technik erfordert dies zielgerichtete Forschungsbemühungen und angrenzende Aktivitäten, die einerseits die Sicherheitseigenschaften von QKI-Systemen untersuchen und stärken, und andererseits die Verwendung von QKI innerhalb der IT-Sicherheit betrachten. Neben der bloßen Fokussierung der QKI-Modelle sind auch die Gegebenheiten und Herausforderungen der entsprechenden Quanteninfrastruktur zu bedenken. Hierzu zählen u. a. die Schnittstellen zwischen klassischer IT und Quantencomputern in Form von hybriden Systemen sowie der Umstand, dass die Verbreitung und Verteilung von Quantencomputern aufgrund ihrer technischen Beschaffenheiten aller Wahrscheinlichkeit nach nicht gleichartig zu derjenigen von klassischer IT erfolgen wird. Insgesamt ist ein konsequentes Verfolgen der eben genannten Aspekte wesentlich, um perspektivisch die Entwicklung geeigneter Sicherheitsstandards für QML-Systeme zu ermöglichen. Synergieeffekte zwischen der Vielzahl an national und international geförderten Projekten zur Entwicklung von Quanten-Hardware und -Software sowie der jeweiligen Sicherheitsanforderungen können nur dann genutzt werden, wenn hierzu frühe Abstimmungs- und Austauschprozesse stattfinden.

4.1.3.4 Sprachtechnologien

Bedarf 01-15: Standardisierung von Language Technology und Natural Language Processing APIs und Datenstrukturen

Die APIs von sprachtechnologischen Cloud-Services sind nicht standardisiert und somit jeweils unterschiedlich, was Vergleich, Testen, Benchmarken und Austausch unterschiedlicher APIs erschwert bzw. unmöglich macht, d. h. aktuell ist keine Interoperabilität gegeben. Zur im besten Fall automatisierten Nutzbarmachung von Datensammlungen ist es notwendig, Metadatenbeschreibungen so zu standardisieren, dass alle wesentlichen Eigenschaften einer Datensammlung in maschinenlesbarer, semantisch annotierter Form vorliegen. Zahlreiche Initiativen arbeiten an dieser Thematik, insbesondere Nationale Forschungsdateninfrastruktur (NFDI), European Open Science Cloud (EOSC) und Gaia-X.

Für Automatic Speech Recognition (ASR)-Verfahren existieren außerdem bisher keinerlei Vorgaben oder Richtlinien, auf welche Weise z. B. Interpunktion oder Zahlen behandelt, d. h. transkribiert werden. Für den besseren Vergleich, für das Benchmarking und auch für den Austausch entsprechender Services ist eine Standardisierung notwendig.

DFKI hat in diesem Bereich bereits erste Erfahrungen im Rahmen des EU-Projekts European Language Grid gemacht sowie unter Mitwirkung der University of Sheffield erste Vorschläge vorgelegt. Dieser Aspekt betrifft auch eine Reihe beigelagerter Themen, z. B. Annotationsformate, Workflows, Benchmarks, Transferlearning bei Sprachmodellen. Das Problem: Alle Anbieter verfolgen jeweils ihre eigene Philosophie, d. h. sie bieten unterschiedliche, proprietäre APIs an. Hilfreich wäre es, die Technologien eines Anbieters mit Standarddatensätzen (oder eigenen Daten) und Standardmetriken zu evaluieren und somit vergleichen zu können (z. B. WER für ASR). Dieses Thema betrifft auch große Sprachmodelle, d. h. insbesondere, wie Sprachmodelle angesprochen werden, um Transferlearning durchzuführen. Zur Relevanz für die Industrie: Kein Unternehmen kann allein ein großes Sprachmodell entwickeln, weshalb Finetuning und Transfer auf Basis standardisierter Methoden und Schnittstellen missionskritisch sind, um das Sprachmodell an den jeweiligen Use Case anzupassen.

Mindestens europaweite Standardisierung von **Language Technology und Natural Language Processing APIs** bezüglich Funktionsumfang und Parametrisierung sollte erfolgen, sodass mehr Interoperabilität und auch bessere Vergleichbarkeit zwischen den Cloud-Services einzelner Anbieter entsteht. In diesem Zusammenhang können auch Datenformate, z. B. bezüglich Datenaustausch, und semantische Annotationsformate betrachtet werden. Dazu gehören die Standardisierung von **Metadaten**, Datensammlungen, Data-Sheets, Model-Cards, Sprachmodelle, Zugänglichkeit, Nutzung von Daten und Datensammlungen für Forschungszwecke und kommerzielle Anwendungen (kann ggf. in NFDI, EOSC, Gaia-X etc. eingebettet werden). Des Weiteren ist die Standardisierung von **Richtlinien für Transkriptionsverfahren** hilfreich, die oft ASR beinhalten oder auf ASR-Ausgaben aufsetzen, z. B. Zahl als Zahl, Zahl als Wort etc., Interpunktion, Groß- und Kleinschreibung etc.

Dieser Punkt schließt auch die **Orchestrierung** von Services in Form von **Workflows** oder **Pipelines** ein. Der Aspekt betrifft zudem die Standardisierung von **Benchmarks** zum Vergleich diverser Lösungen, z. B. ASR oder NaturalLanguage Understanding (NLU). Im Rahmen von Anwendungen im Bereich des Dialogmanagements betrifft dieser Aspekt auch die Standardisierung von **Ressourcen für die Modellierung von Dialogen**.

Bedarf 01-16: Standardisierung der Messung von Performanz, Korrektheit, Präzision und Plausibilität großer Sprachmodelle sowie der Datenqualität
Sprachmodelle stellen derzeit für viele sprachtechnologische Anwendungen den Stand der Forschung und Technik dar, allerdings existieren noch keine Standards bzw. Messung grundsätzlicher Eigenschaften wie z. B. Korrektheit, Präzision, Faktizität, Selbstkonsistenz etc. – u. a., um ein Sprachmodell einschätzen und unterschiedliche Sprachmodelle vergleichen zu können. Die Selbstkonsistenz eines Modells kann z. B. beinhalten, ob sich ein Modell bei bestimmten verwandten Fragen widerspricht. (Anm.: Regelbasierte/symbolische Modelle sind allerdings heute schon Teil von Hybridsystemen bzw. Pipelines). Beispielsweise kann die Messung des Wahrheitsgrads des Outputs von sprachmodellbasierten Anwendungen (bzw. die Selbstkonsistenz des Modells) – falls technisch möglich (und wenn auch nur in einigen klar definierten Bereichen) und belastbar realisierbar – die Qualität des Sprachmodells signalisieren. Zu beachten ist dabei, dass vermehrt auch multimodale Modelle, Bildverstehen, Kombination von Sprache und Bild, Zeichensprache (Erkennung und Generierung) auf Basis großer Sprachmodelle (Stanford nennt diese auch foundation models) durchgeführt werden.

Für das Training von Sprachmodellen und anderen maschinellen Lernverfahren werden u. a. Text-, Audio- und Videodaten eingesetzt. Derzeit existieren noch keine Standards zur Messung der Qualität derartiger Daten und Datensammlungen, u. a. um zu entscheiden, ob sie im Rahmen eines Trainingsdatensatzes nutzbar gemacht werden sollten. Standardisierte Verfahren zur Messung von Datenqualität besitzen ebenfalls eine große Relevanz für den Aspekt von Datenbias.

Standardisierung der **Messung** der **Performanz**, **Korrektheit**, **Precision**, **Plausibilität** im jeweiligen Anwendungskontext großer Sprachmodelle ist wünschenswert. In diesem Zusammenhang ist auch die Messung der Qualität des Outputs von generierenden Sprachmodellen relevant, z. B. bezüglich Sinnhaftigkeit, Grammatikalität, Semantik. Hier existiert ein Bedarf für standardisierte Metriken. Ferner müsste der Begriff „Sprachmodell" definiert werden, und zwar bezüglich Abgrenzung zu textverarbeitenden, evtl. auch regelbasierten Modellen.

Die Standardisierung von Ansätzen zur **Messung von Datenqualität für Sprachmodelle**, d. h. insbesondere Textqualität, aber auch Audioqualität und Videoqualität sind relevant für die Zusammenstellung von Datensets, die z. B. für das Training von Sprachmodellen benutzt werden, sowie für die Messung von Bias. Dies betrifft u. a. die Auswahl der Daten, die für das Training von Sprachmodellen eingesetzt werden, um z. B. Bias und Hatespeech zu bewerten/zu vermeiden etc. Auch für die Beschreibung und Messung von Bias selbst (inklusive einer Spezifizierung der unterschiedlichen Dimensionen von Bias, z. B. political bias, gender bias etc.) müssen Ansätze beschrieben und standardisiert werden.

Bedarf 01-17: Wissensgraphen und Ontologien in große Sprachmodelle
Während Sprachmodelle den Stand der Wissenschaft und Technik für eine Vielzahl sprachtechnologischer Aufgaben darstellen, existieren zahlreiche Wissensbasen, Wissensgraphen und Ontologien, die symbolisches Wissen bzw. semantisches Wissen in symbolischer Repräsentation enthalten. Derzeit existieren noch keine Standards, wie derartige Wissensbasen und Ontologien in Sprachmodelle integriert und der jeweiligen Anforderung entsprechend **sicher** (Bewertung der „Kritikalität") nutzbar gemacht werden können. Dieser Aspekt betrifft auch die Zusammenführung und Integrierung unterschiedlicher Wissensbasen und Wissenspakete.

Die Standardisierung von Ansätzen, wie **Wissensgraphen und Ontologien in große Sprachmodelle**, die integrierbar und nutzbar gemacht werden können, dient der Nutzung existierender symbolischer Wissensbestände im Rahmen der Stand der Forschung und Technik von Sprachtechnologien, die typischerweise auf großen Sprachmodellen basieren. Hierbei sollte auch die Zusammenführung, Integration und Verwaltung von Ontologien und Ontologiemodulen bzw. Ontologiepaketen aus unterschiedlichen Quellen betrachtet werden. Dabei können auch Ansätze betrachtet werden, wie (eher ontologiebasiertes) Weltwissen in (eher dokumentbasierte) Wissensgraphen integriert werden kann. Diese Aspekte sind wichtig und relevant für die Nutzung symbolischer Wissensbasen (d. h. Ontologien) im Rahmen von Knowledge-Graph-basierten Anwendungen.

Bedarf 01-18: Test- und Auditing-Prozesse für KI-Sprachanwendungen
Im Kontext von vertrauenswürdiger KI wird die Standardisierung von **Test- und Auditing-Prozessen** auch für (lernende und kontinuierlich lernende) NLP-Systeme an Bedeutung gewinnen.

Insbesondere, wenn NLP-System wie Suchmaschinen, Empfehlungssysteme oder Chatbots als Entscheidungsunterstützungssysteme in kritischen Anwendungen dienen, wird es nötig sein, Test- und Auditingprozesse zu definieren. Dazu gehören neben den direkten Variablen (Art und Erzeugung

der Testitems, Metriken zur Auswertung der Ergebnisse) auch die Frage der Prozessbeteiligten. Beispielsweise kann es im Gesundheitsbereich geboten sein, Patient*innenvertreter in einem partizipativen Prozess in die Gestaltung und Ausführung der Tests einzubeziehen. Kontinuierlich lernende Systeme werden in bestimmten Zyklen erneut getestet und auditiert werden müssen. Hier muss festgelegt werden, nach welchen Kriterien die Zyklen bestimmt werden.

Bedarf 01-19: Unterstützung Digitale Sprachgerechtigkeit
Von den zahlreichen europäischen Sprachen werden nur einige gut oder sehr gut von Technologien unterstützt. Neben dem Englischen zählen hierzu das Französische, Spanische und Deutsche. Zur Messung und Einschätzung der Unterstützung einer Sprache durch Sprachtechnologien liegen aktuelle Ergebnisse aus dem EU-Projekt European Language Equality vor: die Digital Language Equality Metric. Eine derartige Metrik könnte europaweit standardisiert werden, sodass sich die jeweiligen Sprachgemeinschaften sprachspezifische Ziele und Key Performance Indicators (KPI) im Kontext aller Sprachen Europas setzen können, die zudem gemeinsam beobachtet werden können.

Digitale Sprachgerechtigkeit – Sicherstellung, dass alle Sprachen einer wie auch immer dimensionierten Sprachgemeinschaft (Stadt, Region, Organisation etc.) in ähnlicher, balancierter, ausgeglichener Weise von Sprachtechnologien unterstützt werden – wichtig für Internationalisierung von Inhalten und Technologien sowie für die Skalierbarkeit von Technologien.

Bedarf 01-20: Standardisierungsanforderungen aus dem geplanten AI Act für Sprachanwendungen überprüfen und ggf. ergänzen
Sprache kann u. U. als Hochrisikosystem unter die Regulierung des AI Act und dessen Anforderungen wie beispielsweise der Risikobewertung, des Qualitätsmanagements oder des Nachweises von Robustheit, Transparenz fallen. Für die Standardisierungsanforderungen aus dem geplanten AI Act sind für Sprachtechnologien Untersuchungen vorhandener Standards erforderlich, um zu klären, inwieweit die Anforderungen bereits abgedeckt sind und was ggf. noch mit Standards ergänzt werden muss. Möglicherweise sind Forschungs- und Entwicklungsaktivitäten erforderlich, um die gewünschten Methoden bereitzustellen, z. B. für „record keeping through built-in logging capabilities“ oder für die „robustness specifications“.

Die genannten Standardisierungsanforderungen aus dem geplanten AI Act erfordern einerseits angepasste Standardisierung z. B. als biometrische Systeme einer KI, andererseits sind die Erfordernisse ggf. nicht vollständig technisch erforscht und entwickelt. Dafür wird weitergehende Standardisierung und Forschung empfohlen.

4.1.3.5 Bildgebende Sensorik

Bedarf 01-21: Bewertungsmetriken und Methoden für Bilddatensätze und Erhebungs-/Synthetisierungsverfahren und bildauswertende ML-Verfahren entwickeln
Datensätze übernehmen insbesondere bei modernen ML-Verfahren zunehmend die Rolle von Parametern. Entsprechend werden, beispielsweise im Entwurf AI Act, Anforderungen an KI-Systeme auch mittels Anforderungen an Datensätze formuliert. Jedoch fehlen derzeit standardisierte Verfahren, anhand derer Qualitätseigenschaften von Datensätzen übergreifend beschrieben werden könnten. Einzelfallspezifische Verfahren erreichen jedoch keinerlei Vergleichbarkeit und begrenzen damit die Einschätzbarkeit unterschiedlicher KI-Verfahren. Eine Standardisierung entsprechender Verfahren zur Güteabschätzung sowie die gezielte Entwicklung von standardisierungsfähigen, anwendungsübergreifenden Verfahren kann hier wesentlich zu einem besseren, übergreifenden Verständnis beitragen – auch wenn die standardisierten Metriken nicht den Anspruch eines unumstrittenen, absoluten Gütekriteriums erfüllen, sondern lediglich eine transparente, übergreifende Indikation ermöglichen.

Es sollten standardisierte Bewertungsmetriken erarbeitet werden, die es erlauben, entweder Datensätze (aus echten oder aus synthetisch erzeugten Bilddaten) oder Verfahren, die diese Datensätze erzeugen, nach gängigen Gütekriterien zu bewerten. Diese Metriken sollten gängige Zielvorgaben, beispielsweise gemäß Entwurf EU-AI Act, aufgreifen (vgl. [4]), „Relevanz, Repräsentativität, Fehlerfreiheit und Vollständigkeit in Hinblick auf die beabsichtigte Anwendung“) und geeignete Messprinzipien dieser Zielvorgaben spezifizieren. Diese Metriken sollten weitgehend unabhängig von KI-Methoden oder Anwendungen sein, einschränkende Annahmen/Anwendbarkeiten, wo erforderlich, jedoch klar benennen. Bestehende Ansätze (beispielsweise [97]) sollen auf Eignung untersucht werden. Wo keine geeigneten Verfahren bestehen, die eine Abschätzung leisten können, sollen im Rahmen von F&E neue Ansätze erarbeitet werden.

Bedarf 01-22: Metriken zum Test bildverarbeitender KI-Systeme standardisieren
Analog zum Bedarf „Bewertungsmetriken und Methoden für Bilddatensätze und Erhebungs-/Synthetisierungsverfahren entwickeln“ besteht ein Bedarf zur Standardisierung von Metriken, die die Bewertung bildverarbeitender KI-Systeme ermöglichen und gleichzeitig deren Anwendungsgebiet definieren. Beispielsweise hat sich in der wissenschaftlichen Gemeinschaft die Metrik der „mean Intersection over Union“ (mIoU) für die Bewertung von ML-Verfahren zur Bildsegmentierung etabliert. Entsprechende Metriken sollten auch für andere Aufgabenstellungen wie z. B. Objektdetektion, Klassifizierungen oder Bildumwandlung bereitgestellt werden. Analoge Metriken für gängige KI-Verfahren gemeinsam zu standardisieren kann zur Vergleichbarkeit heterogener Ansätze beitragen.

Dabei ist zu berücksichtigen, dass die Metriken ggf. risikoabhängige Komponenten enthalten können (z. B. risikoabhängige Bewertungen von Segmentierungsfehlern z. B. in kritischen Regionen bei medizinischen Bilddaten). Diese Mechanismen sollten dabei so generisch/modellagnostisch gestaltet werden, dass sie leicht auf unterschiedliche Szenarien übertragen werden können.

Bedarf 01-23: Verfahren zur cybersicheren Authentifizierung auf Basis von Bilddaten
Es sind Verfahren zu entwickeln, die beurteilen, inwieweit gegebene Bildmerkmale nach dem Stand der Technik noch vertrauenswürdig sind (und mithin zur Authentifizierung genutzt werden können) und ab wann entsprechende Merkmale beispielsweise durch „Deep Fakes“ manipuliert sein können. Vorgehensweisen zur Sicherstellung der Authentizität von Identitäten und Informationen sind zu spezifizieren, anhand derer für unterschiedliche Anwendungen ein entsprechend benötigter Grad an Vertrauen hergestellt werden kann.

Bedarf 01-24: Metriken für die Bewertung von Datenschutzrisiken durch Reverse Engineering von ML-Modellen entwickeln
ML-Modelle können, beispielsweise im Rahmen von „Overfitting“, personenbezogene Informationen aus den Trainingsdatensätzen speichern, sodass dieser Umstand den Entwickler*innen unbekannt ist, die Informationen jedoch von sachkundigen Angreifern missbräuchlich rekonstruiert werden können. Gezielte Forschung zur Bewertung des Risikos soll darauf hinarbeiten, perspektivisch Standards für Entwicklung oder Bewertung von ML-Modellen in der Anwendung zu erarbeiten, die diese Risiken beherrschbar machen. Es sollen Metriken erforscht werden, anhand derer beurteilt werden kann, welche Art von personenbezogenen Informationen möglicherweise latent in einem gegebenen ML-Modell enthalten sein könnte und wie entsprechende Missbrauchspotenziale eingegrenzt werden können.

Bedarf 01-25: Forschungsarbeiten zum datenschutzsicheren Entwickeln von KI/ML stärker betreiben und fördern
Trotz der oft vertretenen Annahme, dass die Datenbedarfe in ML-Anwendungen mit Datensparsamkeit oder Anonymisierung nicht zu vereinbaren sind, ist festzuhalten, dass unklar ist, ob nicht deutliche Potenziale bestehen, Datenschutz und KI/ML-Performanz miteinander zu vereinbaren, ohne erhebliche Defizite in Kauf nehmen zu müssen – vorausgesetzt, dass entsprechende Verfahren bereitgestellt werden. Gerade für Hochrisikoanwendungen ist jedoch der Nachweis erforderlich, dass die Performanz beispielsweise durch Anonymisierung tatsächlich nicht unzulässig beeinträchtigt wird. Um hier eine fundierte Einschätzung zu erreichen und entweder geeignete Verfahren standardmäßig anzuwenden oder im Bedarfsfall begründet eine Abwägung zwischen beispielsweise Safety und Privacy zu treffen, ist eine quantitative Bewertung von technischen Potenzialen und Risiken auf der Grundlage wissenschaftlicher Erkenntnisse erforderlich.

Es sollen gezielt Forschungsarbeiten betrieben und gefördert werden, die sich zum Ziel setzen, die Entwicklung hochperformanter KI/ML-Verfahren auf Bilddaten unter Einhaltung datenschutztechnischer Auflagen zu ermöglichen und zu quantifizieren, inwieweit dies möglich ist. Das meint beispielsweise die Erarbeitung von Metriken zur Abschätzung von Performanzverlusten durch Anonymisierung oder Datensparsamkeit sowie etwa die Erarbeitung geeigneter Anonymisierungsverfahren, die Entwicklung von KI/ML-Verfahren, die robust gegenüber Anonymisierung sind, oder die Entwicklung von Methoden, die Datenerhebungen gezielt auf relevante Fälle eingrenzen und damit Datenvolumina von Erhebung, Speicherung und Annotation reduzieren. Geeignete Verfahren sollen perspektivisch in die Standardisierung überführt werden.

Bedarf 01-26: Umwandlung von DIN SPEC 13266:2020 [98] in eine Norm
Es scheint keine Norm für Deep-Learning-Systeme zu geben.

DIN SPEC 13266:2020 [98] ist eine Spezifikation für Deep-Learning-Systeme und beschreibt den aktuellen Stand der Technik sehr gut. Daraus soll eine Norm werden.

Bedarf 01-27: Erweiterung von ISO 21448:2022 [90] auf andere Anwendungsfälle
Der Hauptinhalt von ISO 21448:2022 [90] passt für die meisten Mobilitätsanwendungen. Ein Teil des Inhalts kann auch für Fälle jenseits der Mobilität verwendet werden.

Die Norm ISO 21448:2022 [90] hat nur Straßenfahrzeuge im Titel. Sie soll auch für andere Anwendungsfälle jenseits der Mobilität erweitert werden.

Die Arbeitsgruppe Grundlagen hat die identifizierten Bedarfe nach der Dringlichkeit ihrer Umsetzung bewertet. Abbildung 21 zeigt die Dringlichkeit der Umsetzung, kategorisiert nach den Zielgruppen Normung, Forschung und Politik.

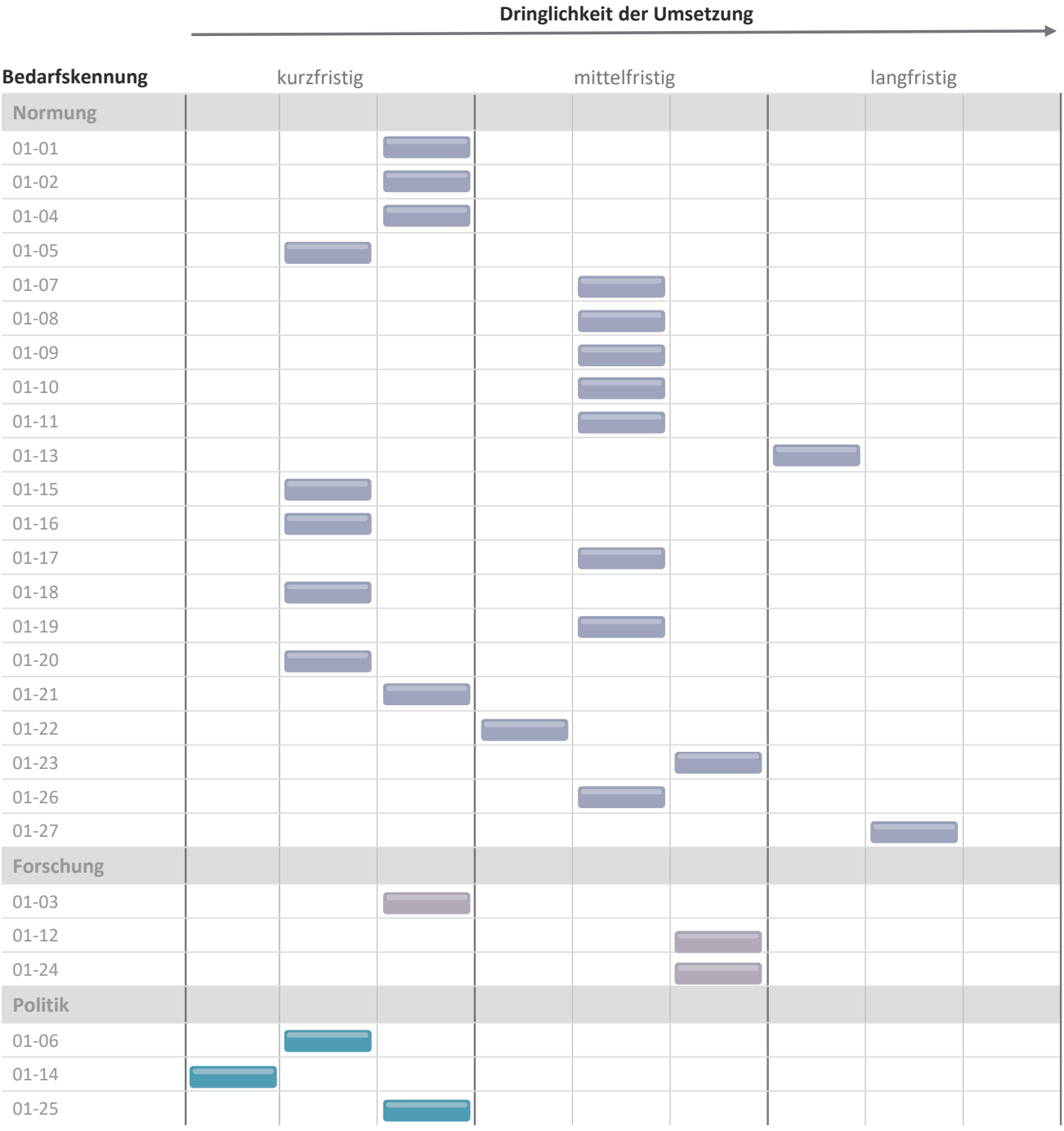


Abbildung 21: Priorisierung der Bedarfe aus Schwerpunkt Grundlagen (Quelle: Arbeitsgruppe Grundlagen)

4.2 Sicherheit

Die grundsätzliche Notwendigkeit für eine Prüfung und Zertifizierung der Sicherheits- (Safety und Security) und Privacy-Eigenschaften eines KI-Systems ergibt sich schon fast allein aus dem Kontext der Nutzung von KI-Systemen in existierenden Prozessen und Produkten und den existierenden Vorgaben an eine Risikominimierung und an einen sicheren Betrieb.

Aufgabe des folgenden Kapitels ist es, Handlungsempfehlungen zu erarbeiten, die es ermöglichen, möglichst sinnvoll existierende Prüf- und Zertifizierungsmodelle aus der Produktsicherheit (Safety) und der IT-Sicherheit (Security) für KI-Systeme nutzbar zu machen. Auch für KI-Systeme soll eine Möglichkeit geschaffen werden, ihre Sicherheit mittels geeigneter Verfahren (Controls) zu erhöhen und ein entsprechendes Sicherheitsniveau nachweisen zu können. Wie schon in der ersten Ausgabe der Normungsroadmap dargelegt, vertraut der Mensch auf sicherheitsgeprüfte Kaffeemaschinen oder sichere Komponenten in Atomkraftwerken sowie geschultes Personal, und für alle diese Bereiche gibt es entsprechenden Normen und Verfahren, die den Umsetzungsgrad einer Norm bewertbar und damit den Sicherheitsgewinn zertifizierbar machen. Letztendlich soll genau dieser zertifizierte Nachweis der Einhaltung der Grundprinzipien in allen Bereichen der Sicherheit (Safety und Security) eines KI-Systems dazu beitragen, Vertrauen zu schaffen.

4.2.1 Safety

4.2.1.1 Status quo

Nach Auffassung der Autoren kommt dem Thema Safety eine besondere Bedeutung zu und es wurde in der ersten Ausgabe der Normungsroadmap (NRM) intensiv in den sektorspezifischen Kapiteln behandelt, wobei es viele sektorübergreifende Aspekte gab. Deswegen wurde das Thema Safety (Produktsicherheit) als horizontales Thema in der NRM aufgenommen. Weiterführende sektorspezifische Safety-Aspekte werden in den Kapiteln 4.6 und 4.7 beschrieben.

Der Begriff „Safety“ steht in engem Zusammenhang mit dem Begriff „Risiko“, wobei der Begriff „Risiko“ unterschiedlich aufgefasst werden kann. Weiterhin kann mit dem deutschen Begriff „Sicherheit“ sowohl „Safety“ als auch „Security“ gemeint sein. Deswegen erfolgt zunächst eine Begriffsklärung. Anschließend werden die Themen „KI“ und „Safety“ in Beziehung gesetzt und zwei Typen von Beziehungen abgeleitet: „direkter Safety-Bezug“ und „indirekter Safety-Bezug“. Danach werden die beiden Beziehungstypen näher beleuchtet. Abschließend erfolgt ein Fazit mit wesentlichen Handlungsempfehlungen.

Im Folgenden wird zunächst erklärt, wie der Begriff „Safety“ mit dem Begriff „Risiko“ zusammenhängt und wie er sich von anderen Qualitätscharakteristiken abgrenzt. Anschließend wird auf das Verständnis des Begriffs „Risiko“ im geplanten AI Act eingegangen. Abschließend wird festgelegt, wie die Begriffe im restlichen Kapitel verwendet werden.

„Safety und Risiko“ nach ISO/IEC Guide 51:2014 [99]
Orientierung bei der Arbeit an Normen und Richtlinien bezüglich des Einbezugs der „Safety“ gibt der ISO/IEC Guide 51:2014 [99]. „Work on standards deals with safety aspects in many different forms across a wide range of technologies […].“ In diesem Guide steht der Begriff Safety für „freedom from risk which is not tolerable”, Risiko steht für „combination of the probability of occurrence of harm and the severity of that harm“ und Schaden (Harm) ist definiert als „injury or damage to the health of people, or damage to property or the environment“.

Die existierende Konsequenzkette im Themengebiet Safety (funktionale Sicherheit) wird häufig in Bereiche erweitert, die nicht absolut unabhängig davon sind, aber eben nicht ursächlich zum Thema Safety gehören. Ein Beispiel wäre der Ausfall eines Kraftwerks, was für die Verfügbarkeit der Stromversorgung problematisch ist, aber eigentlich kein originäres Thema der Safety darstellt. Trotzdem kann ein Stromausfall zu konkretem Schaden auch für Menschen in der Konsequenz führen.

Neben der Konsequenz gibt es aber auch Erweiterungen der Kausalzusammenhänge in die Richtung der Ursachen, z. B. dem Themengebiet Security (Informationssicherheit). Hier ist anzumerken, dass die Ursache einer manipulativen Absicht des Angreifers eine Dimension der Betrachtung erfordert, die nicht originär bei Safety-Risikoanalysen verortet ist. Die Debatten um die gedankliche Zuordnung der Themen Safety und Security wurden und werden in der technischen Regulierung und Normung debattiert, wobei Konsens darin besteht, dass beide Themen zu betrachten sind und Security als eine Grundvoraussetzung für einen sicheren Betrieb im Sinne der Safety gilt. Zum Querschnittsthema Safety und Security sei dabei auf Dokumente wie den Technischen Bericht DIN CLC IEC/TR 63069:2021 „Industrielle Prozess-Leittechnik, Steuerungs- und Automatisierungstechnik – Rahmenbedingungen für Funktionale Sicherheit und IT-Sicherheit“ [100]

sowie auf die Arbeitsergebnisse des Maintenance-Team zur DIN EN 61508-1:2011 [101], DIN EN 615082:2011 [102] und DIN EN 615083:2011 [103] verwiesen, welche sich über längere Zeit mit diesen Fragen auseinandergesetzt haben, um festzustellen, dass es eine wichtige Voraussetzung ist, dass die Safety-Betrachtung einen wirksamen Schutz durch Security-Maßnahmen voraussetzt und nur unter Annahme dieser Voraussetzung Gültigkeit besitzen kann. Im Rahmen dieses Kapitels wird dieser Ansatz aufgegriffen und bei ursächlichen Security-Gefährdungen auf den Bereich Security verwiesen.

„Risiko“ gemäß dem EU AI -Act der Europäischen Union (EU)

Im Bezug auf KI verfolgt die EU einen „risk-based approach of AI regulation“. Dabei ist der Begriff Risiko weiter gefasst als im Kontext „Safety“ bzw. ISO/IEC Guide 51:2014 [99]. Er bezieht sich darüber hinaus auch auf Risiken bezüglich der Grundrechte. Dies beinhaltet Themen wie Datenschutz, Diskriminierungsfreiheit, Schutz der Privatsphäre und Schutz vor unterschwelliger Manipulation von Personen, siehe hierzu Kapitel 1.4.4.

Begriffsdiskussion und -festlegung

Im Sinne des EU AI Act fällt jedes Safety-relevante KI-System in die Klasse „hohes Risiko“. Im Sinne des ISO/IEC Guide 51:2014 [99] hat hohes Risiko eine andere Bedeutung, da es bei Risiken grundsätzlich um Safety-Risiken geht. Diese Safety-Risiken können marginal und somit akzeptabel bis sehr hoch und somit inakzeptabel sein. Im Folgenden werden die Begriffe „Safety“, und „Risiko“ im Sinne des ISO/IEC Guide 51:2014 [99] verwendet.

Bezüglich des Begriffs „KI-System“ wird die Definition des geplanten AI Act verwendet. Diese Definition legt ganz klar fest, dass ein „KI-System“ eine bestimmte Art von Software ist. Sie ist allerdings sehr breit und unscharf bezüglich der Festlegung der Art von Software. Sie definiert keine klare Grenze zwischen einer konventionellen Software und KI-Software. Im Folgenden werden wir die Rolle von KI-Software im Kontext Safety genauer herausarbeiten.

Bezug zwischen KI und Safety

Safety wird durch einen iterativen Prozess aus Risikoidentifikation, -bewertung und -reduktion erreicht. Ein KI-System kann bei diesem Risikomanagementprozess auf unterschiedliche Art und Weise eine Rolle spielen. Im Folgenden wird zunächst der Risikomanagementprozess erklärt. Danach wird darauf eingegangen, wie Software generell und speziell KI in diesem Risikomanagementprozess eine Rolle spielen kann. Anschließend wird auf den Unterschied eingegangen, dass KI verwendet werden kann, um das „normale“ Verhalten eines Systems zu realisieren oder um die notwendige Risikoreduktion zu erreichen. Darauf basierend werden Besonderheiten bezüglich des Verhaltens von autonomen Systemen in komplexen Umgebungen diskutiert. Abschließend werden zwei Klassen von Safety-Bezügen vorgestellt: „KI mit direktem Safety-Bezug“ und „KI mit indirektem Safety-Bezug“.

Iterativer Risikomanagementprozess nach ISO/IEC Guide 51:2014 [99]

„The increasing complexity of products and systems entering the market makes it necessary to place a high priority on consideration of safety aspects“, heißt es in der Einleitung des ISO/IEC Guide 51:2014 [99]. Im Weiteren wird der Ansatz der Risikoidentifikation und -bewertung sowie entsprechender Reduktion von Risiken über den gesamten Lebenszyklus des Produkts beschrieben.

Die Risikobetrachtung ist der erste und wichtigste Schritt, um geeignete Maßnahmen zur Safety zu planen und zu bewerten. Der ISO/IEC Guide 51:2014 [99] beschreibt die elementarsten Überlegungen dazu und die Schritte, die zu einem akzeptablen Risiko führen sollen (siehe Diagramm in Abbildung 22).

Ein Einsatz von KI im Safety-Kontext muss seinen Platz in dieser iterativen Vorgehensweise der Risikobewertung finden und Systemverantwortliche haben zu entscheiden, inwiefern die KI-Software in der angedachten Applikation wirken soll.

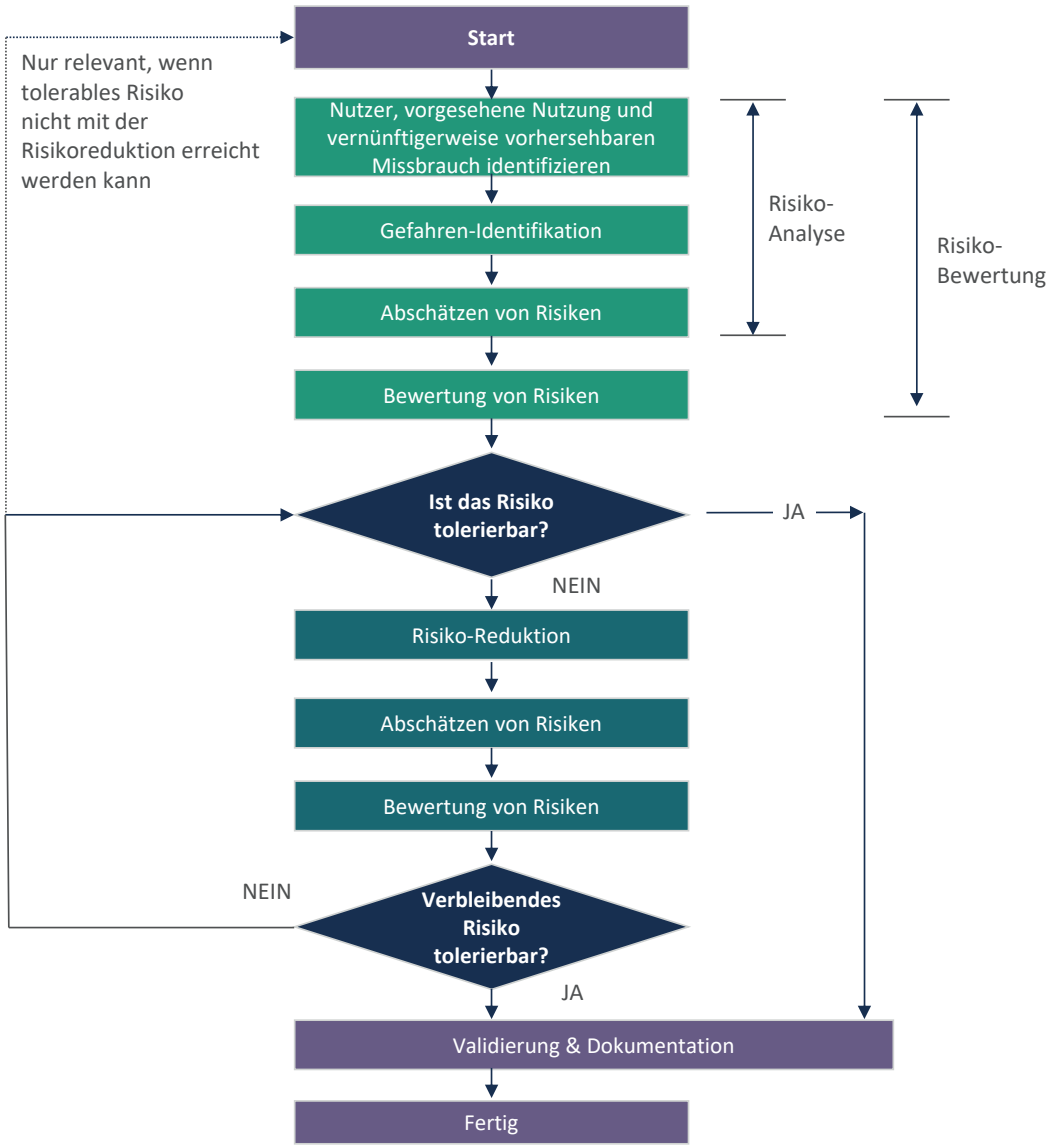


Abbildung 22: Iterativer Prozess von Risikoassessment und Risikoreduktion (Quelle: in Anlehnung an [99])

4.2.1.2 Anforderungen und Herausforderungen

Software (KI oder konventionelle Software) im Risikomanagementprozess

KI als Software gedacht kann auf unterschiedliche Art und Weise eine Rolle für die verschiedenen Schritte im Prozess spielen. Beispielsweise kann Software Betrachtungsgegenstand in einem Prozessschritt sein oder dazu genutzt werden, einen Prozessschritt durchzuführen, wobei sie dadurch in den Scope der Risikobetrachtungen rückt.

Software selbst stellt keine Gefährdung im Sinne der Safety dar, ist aber maßgeblich und zunehmend für das Verhalten technischer Systeme verantwortlich. Über das Systemverhalten kann Software zur Entstehung von Gefährdungssituationen beitragen. Der Bezug von Software zu Safety geht also immer über das Verhalten eines technischen Systems, und das gilt auch für KI.

Das Verhalten eines technischen Systems hängt aber nie nur von Software ab. Es ergibt sich typischerweise aus einem Zusammenspiel von Software, Hardware und anderen Elementen in einer u. U. komplexen Umwelt mit Menschen und weiteren technischen Systemen. Software muss immer auf einer Hardware ablaufen. Entsprechend sind Safety-Betrachtungen auf Software und Hardware in Kombination anzuwenden. Dieser Aspekt kommt bei dem geplanten AI Act zu kurz.

Hardwarefehler und Ausfälle sind zwar bei der Anforderung an die Robustheit des KI-Systems mitzubetrachten, da der

AI Act sich aber bei der KI-System-Definition spezifisch auf Software bezieht, kommt der Hardware-Bezug nicht explizit zum Ausdruck.

Nominalverhalten vs. Verhalten zur Risikoreduktion
KI kann Einfluss auf das Nominalverhalten des Systems im Sinne des „intended use“ haben, z. B. im Fahrerassistenzsystem eines Autos (siehe Abbildung 22).

Das Nominalverhalten kann dazu beitragen, dass aus Gefährdungen wie mechanischen Gefährdungen, elektrischen Gefährdungen, thermischen Gefährdungen etc. (siehe z. B. [517]) Gefährdungssituationen entstehen. Um die Risiken zu bewerten, die von der Gefährdungssituation ausgehen, bedarf es keiner neuen Risikoansätze. Die Bewertung, ob ein bestimmtes Systemverhalten akzeptabel oder zu riskant ist, hängt nicht davon ab, ob KI oder eine andere Technologie für die Realisierung des spezifizierten Verhaltens verwendet wird. KI wird allerdings häufig für die Realisierung von Funktionen eingesetzt, die schwer vollständig zu spezifizieren sind. Das gilt insbesondere für Systeme, die komplexe Aufgaben in einer komplexen und sich ändernden Umgebung automatisch, d. h. ohne Eingriff durch einen Benutzer oder seine Benutzerin, erfüllen sollen. Für solche komplexen Szenarien besteht die Herausforderung darin, sicherzustellen, dass alle relevanten Situationen betrachtet und in der Spezifikation des Systemverhaltens berücksichtigt werden. Je nach Art der KI kann es auch sein, dass sich das Nominalverhalten ändert. Diese Aspekte können zu einer Unschärfe bei der Risikoabschätzung führen. Die Unschärfe des Risikos eines betrachteten Systems (SoC, System under consideration) ist in Abbildung 23 mit einem blauen Kreis um einen roten Punkt dargestellt. Der rote Punkt bezieht sich auf das Risiko des SoC und der blaue Kreis beschreibt, wie dieses Risiko abweichen kann aufgrund der Dynamik des SoC oder seiner Einsatzumgebung.

Abbildung 23 beschreibt auch den Fall, dass KI eingesetzt wird, um ein identifiziertes Risiko zu reduzieren. KI könnte beispielsweise in Schutzmechanismen eingesetzt werden.

Mit Schutzmechanismen kann eine signifikante Risikoreduktion erreicht werden, aber entsprechend stringent sind die Anforderungen an die Korrektheit der Umsetzung. Inwieweit KI diese Anforderungen an Korrektheit erfüllen kann, ist Gegenstand aktueller Forschung. Je nach Art von KI, Komplexität des Schutzmechanismus und der Einsatzumgebung kann der aktuelle Stand in Bezug auf Wissen und Technik ausreichen oder nicht.

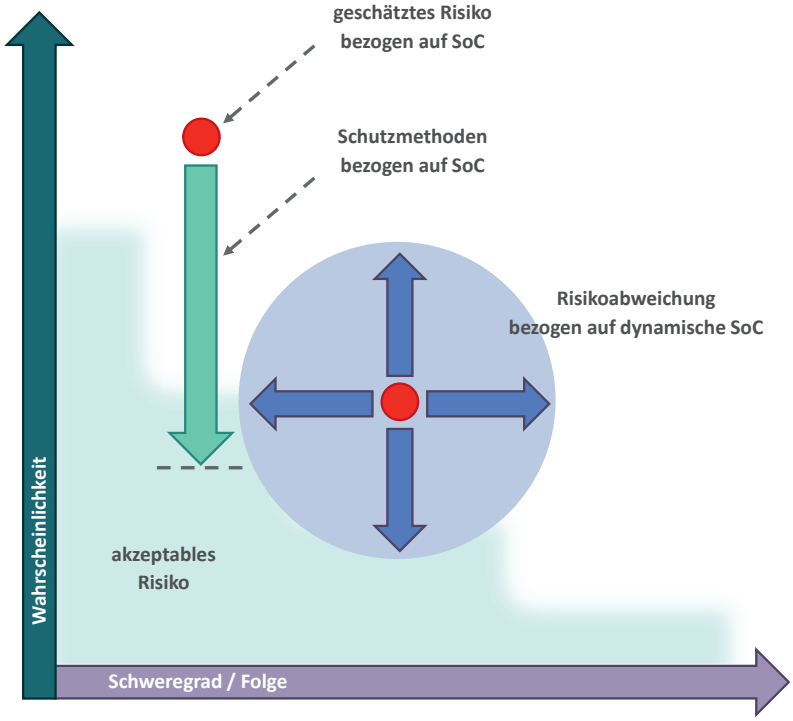


Abbildung 23: Risikodiagramm (Wahrscheinlichkeits-Folgeabschätzung) (Quelle: Holger Laible)

Eine weitere Möglichkeit, Risiken zu reduzieren, besteht darin, Warnhinweise zu geben. Der Einsatz von KI für die Realisierung von Warnfunktionen geht typischerweise mit einer geringeren Risikoreduktion einher und somit auch mit weniger stringenten Anforderungen an die Korrektheit der Umsetzung. Dennoch kann der aktuelle Stand der Wissenschaft und Technik im Bereich KI nicht ausreichend sein, um den Anforderungen zu genügen, und es besteht auch in diesem Bereich Forschungsbedarf.

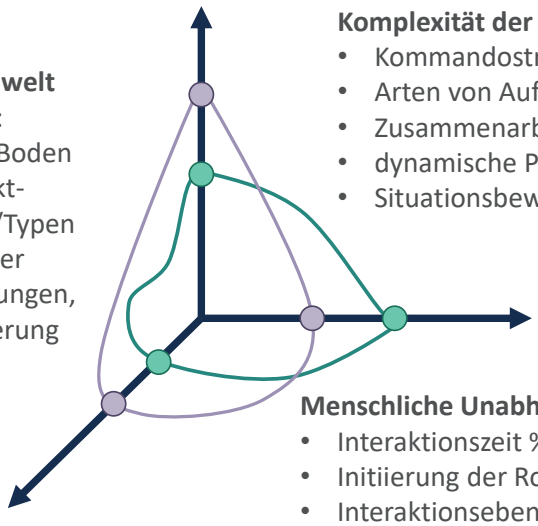
Autonome Systeme in komplexen Umgebungen

Bei steigendem Grad an Automatisierung, komplexen Aufgabenstellungen und unstrukturierten Umgebungen wird es schwieriger, eine Spezifikation zu erstellen, die ausreichend vollständig ist, widerspruchsfrei und ein Verhalten beschreibt, das akzeptabel ist. Entsprechend gibt es Diskussionen bezüglich der Grundannahmen zur Sicherheit des Nominalverhaltens. Ist es beispielsweise sicher, wenn ein fahrerloses Fahrzeug mit einer bestimmten Geschwindigkeit an einem parkenden Auto vorbeifährt (ein Standardfall im normalen Fahrbetrieb), wobei theoretisch ein kleines Kind plötzlich hervorspringen könnte? Dies ist eine Situation, die bei menschlicher Fahrzeugführung riskant ist, aber vorkommt. Wie kann nun ein akzeptables Betriebsverhalten in solchem Kontext aussehen und gibt es überhaupt die Möglichkeit, sämtliche Eventualitäten abzudecken? Man kann zur Designzeit nicht wissen, was während des Betriebs passieren wird. Solange bekannt ist, was man nicht weiß, kann man damit umgehen, aber in komplexen Umgebungen gibt es auch unbekannte Wissenslücken („Unknown Unknowns“).

Die Herausforderungen bezüglich steigender Autonomie und Komplexität der Mission und Aufgabe wurde im Papier „Autonomy Levels for Unmanned Systems (ALFUS) Framework, Volume I – Terminology“ [104] schon 2007 thematisiert (siehe Abbildung 24).

Die drei Dimensionen „Missionskomplexität“, „Umgebungskomplexität“ und „Autonomie (Human Independence)“ haben zunächst nichts mit KI zu tun, sondern mit den Charakteristiken des Gesamtsystems. Allerdings ist KI häufig ein essenzielles Mittel, womit man versucht, die Komplexität zu beherrschen und einen hohen Automatisierungsgrad zu erreichen. Im Hinblick auf Safety stehen entsprechende Aufgabenstellungen und Erwartungshaltungen aber im Spannungsfeld mit dem KISS (Keep It Simple, Stupid)-Ansatz für Safety-Lösungen.

Die Sicherheit des Nominalverhaltens ist insbesondere beim automatisierten Fahren relevant und wird dort teilweise unter dem Titel „Safety of the intended functionality“ (SOTIF) diskutiert. Allerdings geht es bei SOTIF in erster Linie darum, Fehler in der Verhaltensspezifikation zu finden, und nicht um die Frage, ob ein spezifiziertes Verhalten zu riskant oder noch akzeptabel ist.



Komplexität der Umwelt
Lösungsverhältnisse:
- statisch: Terrain, Boden
- dynamisch: Objekt-Frequenz/Dichte/Typen
- Stadt, Land, Wetter
- operativ: Bedrohungen, Täuschung, Kartierung

Komplexität der Mission
- Kommandostruktur
- Arten von Aufgaben, erforderliches Wissen
- Zusammenarbeit
- dynamische Planung, Analyse
- Situationsbewusstsein

Menschliche Unabhängigkeit
- Interaktionszeit %, Planungszeit %,
- Initiierung der Roboterkommunikation
- Interaktionsebenen
- Arbeitsbelastung/Fähigkeitsniveau

Abbildung 24: Drei Dimensionen von Komplexitäten (Quelle: in Anlehnung an [104])

Ein Forschungsgebiet, das sich damit beschäftigt, autonome Systeme sicher zu machen, ist das dynamische Risikomanagement. Dabei geht es darum, autonomen Systemen die Fähigkeit zu verleihen, Risiken selbst abzuschätzen und zu kontrollieren. Dies bedeutet nicht, dass autonome Systeme irgendein echtes Verständnis von Risiko erlangen. Es geht um die Entwicklung eines Schutzmechanismus, basierend auf Risikometriken oder komplexeren Funktionen, um Risiken von Verhaltensoptionen in der aktuellen Situation abzuschätzen. Die Entwicklung von Risikometriken und Funktionen, um Risiken zur Laufzeit zu bestimmen und zu kontrollieren, ist zwar primär ein Forschungsthema, findet aber bereits Einzug in die Normung wie bei der Anwendungsregel für autonom kognitive Systeme VDE-AR-E 2842-61-2:2021 [105] oder ISO 21815 [106], [107], [108] für Kollisionsvermeidung von Erdbaumaschinen.

Darstellung von direktem und indirektem Bezug zwischen KI und Safety

Es wurden bereits einige Fälle erwähnt, wie KI in Beziehung zu Safety stehen kann, beispielsweise KI im Nominalverhalten oder KI zur Risikoreduktion oder KI als Werkzeug bei der Durchführung des Risikomanagementprozesses. Ein wesentliches Unterscheidungsmerkmal zwischen diesen und anderen Fällen besteht darin, dass KI mehr oder weniger direkt mit Safety in Beziehung steht. Dabei kann man grob zwischen einem direkten und einem indirekten Safety-Bezug unterscheiden.

Ein System hat einen direkten Safety-Bezug, wenn ein Versagen oder Fehler des Systems direkt zu einem gefährlichen Zustand für Mensch und Umwelt führt. Dies gilt auch für ein Softwaresystem, das als KI klassifiziert wird. Dedizierte Safety-Systeme für die Umsetzung von Safety-Funktionen zur Risikoreduktion haben typischerweise einen direkten Safety-Bezug. Ein direkter Safety-Bezug bedeutet nicht, dass es zwangsläufig zu einem Unfall kommt, wenn das System versagt, da ein gefährlicher Zustand nicht zwangsläufig eine Unfallsituation bewirkt. In der Regel gibt es noch die normalen Betriebsfunktionen und es muss auch die Anforderungssituation an die Safety-Funktion eintreten.

Ein System, bei dem ein Fehler oder Versagen nur indirekt zu einem gefährlichen Zustand für Mensch und Umwelt führt, hat einen indirekten Safety-Bezug. Indirekte Safety-Bezüge können sehr komplex sein und Analysen der Kausalzusammenhänge insbesondere bei interdisziplinären Vorgängen sehr aufwendig werden. Die Analyse indirekter Kausalitätsketten ergibt oft, dass die Wahrscheinlichkeit für den Eintritt eines gefährlichen Ereignisses sehr gering ist. Je nach Komplexität der Analyse ist es aber schwer zu garantieren, dass die Analyseergebnisse verlässlich sind. In der Praxis sind daher direkte Safety-Bezüge durch in der Architektur angelegte systematische Aspekte entscheidender für die (Analyse/Betrachtung der) Systemsicherheit als solche indirekten Zusammenhänge.

KI MIT DIREKTEM SAFETY-BEZUG

Bei KI mit direktem Safety-Bezug ist es unabdingbar, sehr genau zu untersuchen, ob entweder KI als funktionale Komponente möglicherweise das Risikoniveau erhöht oder ob KI als Absicherungskomponente eine notwendige Risikoreduktion wirklich erreicht. Im Folgenden werden zunächst Aspekte vorgestellt, die berücksichtigt werden sollten, wenn es um die Frage geht, ob der Einsatz von KI als funktionale Komponente das Risikoniveau möglicherweise erhöht. Anschließend wird auf die Nutzung von KI zur notwendigen Risikoreduktion eingegangen und auf die damit verbundenen Qualitätsansprüche. Abschließend wird erklärt, wie strukturierte Sicherheitsnachweise (Safety Assurance Cases) genutzt werden können, um die erreichte Qualität transparent zu machen und Schwächen in Argumentationen bezüglich des Einsatzes von KI im Safety-Kontext aufzudecken.

KI erhöht möglicherweise das Risikoniveau durch:

a) **mangelhaftes Systemverständnis**

KI bietet Lösungsmöglichkeiten für Fälle, wo es schwierig ist, explizit zu spezifizieren und einzuprogrammieren, welche Ausgaben aus Eingaben generiert werden sollen. Das Maschinelle Lernen ermöglicht es dem Entwickler, Konzepte aus Daten zu lernen und indirekt in einem Algorithmus abzubilden. Das Fehlen einer expliziten Spezifikation und Programmierung des Zusammenhangs zwischen Eingabe- und Ausgabedaten und der oftmals riesige Parameterraum insbesondere von konnektionistischen KI-Systemen begrenzt allerdings das Systemverständnis. Da Sicherheitsnachweise auf dem Systemverständnis beruhen, limitiert das mangelnde Systemverständnis den Einsatz im Safety-Kontext.

b) **die Veränderung von Randbedingungen**

Beim z. T. schon heute stattgefundenen und dem noch zu erwartenden stärkeren Einzug von KI-Software in Betriebsfunktionen bzw. dem Nominalverhalten, um verschiedene Use Cases zu realisieren, ist es durchaus denkbar, dass bisherige Risikobetrachtungen überdacht werden müssen.

Einflüsse auf das Ausgangsrisiko sind zu untersuchen, sollten aber bei direktem Safety-Bezug und Einbeziehung von Worst-Case-Überlegungen im Bereich Safety weniger kritisch sein.

Stärker jedoch könnte KI-Software bestehende Schutzmaßnahmen, welche auf Annahmen von Gefährdungssituationswahrscheinlichkeiten dimensioniert sind, beeinflussen, indem diese Randbedingungen verändert werden. So ist es denkbar, dass KI-Software z. B. für die Regelung der Maschinenentwärmung eingesetzt wird, aber durch Unzulänglichkeiten die Temperaturen für sicherheitsrelevante Bauteile unzulässig hoch werden und damit systematisch aus der Spezifikation laufen, was zu nicht abschätzbaren Fehlern dieser Bauteile führt. Denkbar sind auch Szenarien, welche Anforderungsraten an Safety-Systeme (d. h. wie oft das Safety System eingreifen muss) und damit die Randbedingungen ihrer Auslegung verändern.

An dieser Stelle sei angemerkt, dass auch normale Software zu veränderten Randbedingungen führen kann. Allerdings ist das Risiko bzw. die Unsicherheit bezüglich veränderter Randbedingungen als Folge des Einsatzes von KI-Systemen i. d. R. höher: durch ggf. unvollständige Spezifikationen, Intransparenz der Algorithmen und, abhängig vom konkreten Modell, eine große Sensitivität im Hinblick auf Parameteränderungen in Kombination mit Anpassungen der Parameter im Kontext von Updates und kontinuierlichen Lernansätzen („Online-Learning“).

c) direkten Einfluss auf Safety-Funktionen

KI-Software, die im Zusammenhang mit Safety-Funktionen zum Einsatz kommt bis hin zum Einsatz als Schutzmaßnahme, beschäftigt die Expert*innenkreise und benötigt weiterhin starke Aufmerksamkeit. In diesem Zusammenhang sei auch an die speziellen Use Cases aus der Industrie verwiesen, die in diesem Papier beschrieben sind.

d) Verhaltensänderungen des Menschen im Zusammenhang mit Automatisierung

Der Umgang des Menschen mit dem technischen System ist im Wandel durch die zunehmende Automatisierung.

Bei technischen Assistenzsystemen, die Menschen eine umfangreiche Unterstützung bieten, stellt sich die Frage nach Änderungen im menschlichen Verhalten. Bei der Risikoabschätzung für Safety wird traditionell auch auf diesen Aspekt geachtet. Neben dem Verlust physischer Fähigkeiten kann es auch dazu kommen, dass Zusammenhänge kognitiv nicht mehr bewertet werden können (unbewusste Inkompetenz). Auch gibt es den Effekt, dass die Assistenzsysteme den Eindruck erwecken könnten, sehr zuverlässig zu sein, was z. B. durch Aufmerksamkeitsverlust des Menschen eine Verhaltensänderung bewirken kann.

Bei höheren Automatisierungsgraden fungiert der Mensch oft noch als Überwacher, wobei das System dann wiederum überwacht, ob der Mensch seiner Funktion als Überwacher noch gerecht wird. KI bietet an dieser Stelle besondere Möglichkeiten.

QUANTIFIZIERBARE RISIKOREDUKTION DURCH KI

Dies ist der häufig gewünschte und intendierte Einsatz von KI-Software, jedoch ist hier weiterhin noch Forschung und Entwicklung (F&E) zur Zuverlässigkeit von KI-Technologien notwendig und es sind bis zum jetzigen Zeitpunkt, abhängig von der Art der KI Technologie, noch Herausforderungen zu bewältigen.

Es wurde bereits auf die Unterschiede der Schutzmaßnahmen eingegangen und im Bereich der Funktionalen Sicherheit ist es üblich, dass die Wertigkeit der Maßnahme quantifiziert wird. Aber auch hier setzen sich Safety-Level (SIL nach DIN EN 61508-1:2011 [101] oder PL nach DIN EN ISO 13849-1:2016 [109]) aus quantifizierbaren (statistische Fehler) und nicht quantifizierbaren (systematischen) Aspekten zusammen. In anderen Bereichen, z. B. der elektrischen Sicherheit, sind Maßnahmen in Normen definiert, aber nicht mit Wahrscheinlichkeiten hinterlegt. Daher gibt es eine Debatte, wie mit KI-Software umzugehen ist, und diese Debatte ist weiter fortzuführen. Unabhängig vom Ergebnis der Debatte bleibt festzuhalten, dass die Erreichung bestimmter Qualitäten bei KI-Applikationen eine Grundvoraussetzung für deren Einsatz als bedeutende Schutzmaßnahme wäre.

Wegen der mangelnden Interpretierbarkeit und dem mangelnden Systemverständnis (s. a) mangelhaftes Systemverständnis) von vielen heutigen KI-Systemen fehlt es oft an Nachweisverfahren, die garantieren (oder zumindest plausibilisieren), dass die KI ihre Aufgabe mit der erforderlichen Zuverlässigkeit erfüllt. Es gibt heute etwa für die Objekterkennung (z. B. Fußgängererkennung) kein Verfahren, das zuverlässige Aussagen liefern kann, ob ein Fußgänger in jedem Fall erkannt wird. Es gibt zwar Nachweisverfahren, aber die damit erhaltenen Aussagen zur Risikoreduktion erfüllen nicht die z. B. für die Perzeption beim hochautomatisierten Fahren erforderlichen Anforderungen. Ebenso gibt es Forschungsarbeiten zur Verbesserung der situativen Unsicherheitsabschätzung und -behandlung sowie der Erklärbarkeit von KI, die

solche Nachweise erlauben würden. Auch diese sind bisher aber noch nicht ausreichend, um die erforderlichen Nachweise zu erbringen.

Es gibt durchaus Applikationen und Technologien, welche aufgrund des heutigen Stands der Technik einsetzbar sind, doch werden diese Bereiche von KI-Expert*innen oft nicht als „echte KI“ gesehen (z. B. Entscheidungsbäume), fallen aber trotzdem unter die recht breite Definition von KI nach ISO/IEC 22989:2022 [16] und des EU AI Act. Der ISO/IEC TR 5469 [33] hat deshalb begonnen, Überlegungen zur Einordnung der KI-Safety-Applikationen anzustellen, um davon ausgehend geeignete Anforderungen und Maßnahmen ableiten zu können. Diese Arbeiten sollten weiter Unterstützung finden, um zu einem umfassenden Bewertungskonzept zur Safety von KI-Software zu kommen, denn insbesondere bei „echten KI-Technologien“ sind die Ansätze noch unvollständig.

Als Rahmenwerk für so ein umfassendes Konzept bieten sich Assurance Cases an, wie auch in der VDEARE 2842-61-2:2021 [105] und der in der Entwicklung befindlichen ISO PAS 8800 [110] bereits berücksichtigt.

QUALITATIVE RISIKOREDUKTION DURCH KI (ASSURANCE CASES)

Für den Einsatz „normaler“ Software im Kontext Safety gibt es umfangreiche Erfahrungen und einen Konsens, sodass mithilfe von Safety-Leveln (z. B. SIL nach DIN EN 61508 [101], [102], [103], [433], ASIL nach ISO-26262-Reihe [455], PL nach DIN EN ISO 13849-1:2021 [111], AgPLr nach DIN EN ISO 251191:2021 [112]) Maßnahmen empfohlen werden können. Das Level wird typischerweise mit einem Risikographen bestimmt. Dort werden Risikoparameter wie das Schadensausmaß bestimmt und jede Parameterkombination einem Safety-Level zugeordnet. Das Safety-Level legt dann fest, welche Maßnahmen wie Testmethoden, Code-Review usw. durchgeführt werden sollten. Die Grundlage für diese Festlegung ist die Erfahrung. Entsprechend sollten Safety-Level auch nur genutzt werden, wenn es ausreichend Erfahrung gibt. In der ISO/IEC/IEEE 15026-3:2022 [113] heißt es: „Integrity levels shall be defined for an area only if a substantial body of relevant experience exists for the area that is well understood by those performing the definition.“ Da diese Erfahrung beim Einsatz von KI im safety-kritischen Kontext noch fehlt, kann das Konzept der Integritätslevel noch nicht sinnvoll für KI angewandt werden.

Es kann aber Dokumente geben, die einen hilfreichen Maßnahmenkatalog zusammenstellen. Solch ein Katalog kann verwendet werden, um Maßnahmen auszuwählen. Allerdings stellt sich dann die Frage, ob diese Maßnahmen für den anvisierten Anwendungsfall ausreichen. Assurance Cases sind ein geeignetes Mittel, um eine Antwort auf diese Frage zu finden und zu evaluieren, ob sich eine stichhaltige Argumentation für Safety finden lässt. Neben der Forschung zu KI-Absicherungsmaßnahmen und deren Sammlung ist somit die Forschung zu Assurance Cases für KI essenziell, um das aktuelle Potenzial von KI auch im safety-kritischen Kontext optimal auszunutzen und inakzeptable Risiken zu vermeiden.

Assurance Cases werden im System- und Software Engineering eingesetzt, um trotz zunehmender Komplexität der Aufgabenstellung (z. B. entlang der drei Dimensionen Missionskomplexität, Umgebungskomplexität und Autonomie) und/oder dem Einsatz neuer Technologien Safety zuzusichern. KI ist eine neue Technologie im Hinblick auf die Nutzung im safety-kritischen Kontext. Es gibt kaum Erfahrung in diesem Nutzungskontext und keinen genormten Konsens darüber, was wann ausreicht. Entsprechend sollten sie auch genutzt werden, um die Safety beim Einsatz von KI zu garantieren.

Assurance Cases bestehen aus einer strukturierten Argumentation, basierend auf Evidenzen, um eine Behauptung (Claim) zuzusichern. Durch die Strukturierung kann jedes einzelne Argument dediziert auf seine Validität geprüft werden. Die Nutzung von Assurance Cases hat sich bei anderen Technologien bewährt und es gibt einen breiten Konsens, dass sie auch bei KI hilfreich ist, um zu verhindern, dass unsichere Systeme auf den Markt gebracht werden. Weiterhin begünstigen sie den Aufbau der notwendigen Erfahrung, da Felddaten genutzt werden können, um die Argumente zu stärken und invalide Argumente frühzeitig zu erkennen.

Es sollte aber auch erwähnt werden, dass konzeptuell zwar Assurance Cases für Systeme mit KI-Komponenten aufgestellt werden können, aber für viele Komponenten der Nachweis (oder zumindest die Plausibilisierung) der für den konkreten Assurance Case notwendigen Evidenzen aktuell schwer möglich ist. Aktuelle Forschungsansätze zur Strukturierung und Prüfung von Assurance Cases sollten im Hinblick auf die intersubjektive Bewertbarkeit und andere Kriterien für Safety-Normung weiterentwickelt und evaluiert werden.

ZUSÄTZLICHE RISIKOREDUKTION DURCH KI (ÜBER DEN PUNKT DES AKZEPTIERTEN RISIKOS HINAUS)

Dabei kann Risikominimierung auch den Punkt des akzeptierten Risikos überschreiten. Der Stand der Technik ist für die Minimierung des Risikos anzuwenden und damit können

auch Anwendungen eingesetzt werden, welche für sich selbst betrachtet keine definierte Schutzqualität beweisen können, also nicht quantifizierbar sind.

Ein Beispiel für eine solche Anwendung wäre eine Maßnahme zur Kontrolle der Einhaltung der vorgesehenen Verwendung oder des korrekten Betriebs („intended use“), um eine mögliche Fehlverwendung („foreseeable misuse“) zu erkennen und/oder zu unterbinden, sofern dies konstruktiv nicht möglich ist.

Solche Anwendungen erfordern allerdings eine genaue Betrachtung, wie sie langfristig wirken können. Folgende Fallunterscheidungen bezüglich Verhaltensänderungen können hier sinnvollerweise gemacht werden:

→ Risikoreduktion wirkt verhaltensändernd
Bei Assistenzsystemen, die im täglichen Gebrauch gut funktionieren, wird der Mensch dazu tendieren, sich auf deren Wirkung zu verlassen. Dies kann bei Fahrzeug-Assistenzsystemen sehr gut empirisch beobachtet werden. Diese Verhaltensänderung führt dazu, dass das technische System eine höhere Wertigkeit bekommt, als ursprünglich angedacht war. Damit entsteht eine Rückwirkung auf die ursprüngliche Risikoabschätzung und die Wirksamkeit der Schutzmaßnahmen.

→ Risikoreduktion wirkt nicht verhaltensändernd
Bei Funktionen, welche nicht in die tägliche Verwendung Einzug halten, wie z. B. KI-unterstützte Not-Halt-Applikationen, ist die Adaption des Menschen an das System nicht gegeben, da die Unannehmlichkeiten für solche Safety-Auslösungen zu hoch für den Normalbetrieb sind. Allerdings können durch Falschauslösungen von Safety-Systemen wiederum neue Gefährdungen entstehen.

Auch dieses Thema steht nur indirekt im Zusammenhang mit KI-Software. KI-Software ist eher ein technologisches Mittel, um Assistenzsysteme und autonome Systeme zu bauen. Die Verhaltensänderung ist aber nicht davon abhängig, welche Mittel verwendet wurden, um ein System zu bauen, sondern davon, was gebaut wurde. Die Verhaltensänderungen des Menschen würden auch auftreten, wenn das gleiche Systemverhalten nicht mit KI-Software, sondern mit einer anderen Art von Software implementiert worden wäre.

KI mit indirektem Safety-Bezug

Im Folgenden werden einige Anwendungen von KI mit indirektem Safety-Bezug diskutiert. Zunächst wird die Anwendung von KI im Risikomanagement diskutiert und dann KI in der Arbeitssicherheit.

KI IM RISIKOMANAGEMENT/ASSESSMENT

Eine denkbare Anwendung von KI im Risikomanagement unterscheidet sich von einer klassischen Tool-Unterstützung, wie sie heutzutage zur Risikodokumentation eingesetzt wird. Smarte Features, die ggf. Vorbewertungen zum Risiko durchführen, können wiederum dazu führen, dass die beteiligten Expert*innen mehr auf die Vorschläge oder Einschätzung vertrauen, als dies angezeigt wäre.

Streng genommen fallen aber solche Systeme nicht unter eine sicherheitsrelevante Tool-Klasse (und stehen deshalb unter indirektem Safety-Bezug), aber es greifen Aspekte der Verhaltensänderung und Fragen zum menschlichen Verständnis des bewerteten Risikos von komplexen Systemen und Zusammenhängen, wie sie zuvor bereits angesprochen wurden (siehe 4.2.1.2).

Zu untersuchen wäre, ob solche Tools oder Systeme zur Risikoanalyse in der Lage sein könnten, ausreichendes Vertrauens- und Zuverlässigkeitspotenzial zu haben, um Expert*inneneinschätzungen in einer Risikositzung sinnvoll zu ergänzen, und wie sie umgesetzt und eingesetzt werden sollten, um eine riskante Verhaltensänderung in der Anwendung zu vermeiden bzw. zu adressieren.

KOMPLEXITÄT DES SYSTEMS, EINSATZ VON KI FÜR RISIKOBEWERTUNGEN

Die Zunahme der Systemkomplexität stellt eine große Herausforderung dar, in den Risikobewertungen die entscheidenden Punkte zu analysieren. Bei einem eventuellen Einsatz von KI für die Durchführung von Risikobewertungen solcher komplexen Systeme überkreuzen sich verschiedenste Fragestellungen.

VOLLSTÄNDIGKEIT DER ERFASSUNG DER NOTWENDIGEN RISIKOKRITERIEN

Tiefere Analysen von Risiken durch KI-Software bedingen die umfassende und weitreichende Festlegung von Beschreibungen (Semantik) über Zusammenhänge, die interdisziplinär über verschiedenste Domänen zu vereinbaren wären.

Folgende komplexe Punkte sind nur einige der Herausforderungen:

a) Kann die reale Welt ausreichend in eine digitale Beschreibung überführt werden oder ist mit dieser Überführung eine unterkomplexe Darstellung verbunden?
b) Besteht Einigkeit über die wissenschaftlichen und technischen Zusammenhänge?

c) Gibt es zu viele unterschiedliche Interessen, die zu vereinbaren wären?
d) Wie kann die Richtigkeit der Kriterien und auch der gemachten Ableitungen bewertet werden?

MENSCH IN DER VERANTWORTUNG UND KONTROLLE HALTEN (KEIN SAFETY OHNE DIE LETZTLICHE MENSCHLICHE ENTSCHEIDUNG)

Alle Safety-Aspekte, wie Vollständigkeit, Korrektheit und Haftung, sind an den Menschen gebunden. Werden Safety-Aspekte auf technische Systeme heruntergebrochen, gibt es hohe Anforderungen an deren Wirksamkeit und Korrektheit. Bei Unfällen kann beobachtet werden, dass häufig menschliches Versagen oder technisch gesprochen systematische Fehler die letztliche Ursache sind. Streng genommen gilt dies für alle Vorfälle, weil ein technisches System vom Menschen entwickelt ist. Die Verkettung unglücklicher Ereignisse oder aber eine unerwartete Überschreitung zuvor festgelegter Randbedingungen werden beobachtet, wenn Unfälle untersucht werden.

Die Versuchung, daraus zu folgern, sobald „Verantwortung" der Maschine übertragen wird, würden die Risiken sinken, verkennt die ungleich höhere Wahrscheinlichkeit, dass die Entwicklung der Maschine mit bereits potenziell unzulänglichen Parametern und Vorgaben seitens weniger Expert*innen zuvor erfolgt ist.

Dieses Spannungsfeld sollte unvoreingenommen beleuchtet werden. Auch stellt sich die ethische Frage, inwieweit eine Maschine überhaupt „eigenständig" Gefahrensituationen für den Menschen einschätzen sollte. Heutige Sicherheitsfunktionen wie die Kollisionsvermeidung bei fahrerlosen Transportsystemen detektieren Gefahrensituationen und reagieren, um Kollisionen zu vermeiden. Das Systemverhalten ist im gewissen Sinne „eigenständig", da es keine Benutzervorgaben für das Verhalten gibt. Allerdings ist das Systemverhalten für Entwickler nachvollziehbar und kein Ende-zu-Ende gelerntes Verhalten, das den Eindruck eines selbstbestimmten „eigenständigen" Handelns vermittelt. Selbst wenn diese Art von Ende-zu-Ende-Lernen im safety-kritischen Kontext irgendwann Einzug fände, wäre immer noch der Mensch für die Folgen verantwortlich. Bei explizit einprogrammiertem Verhalten zum Umgang mit Gefahrensituationen stellt sich die ethische Frage, welche Regeln zulässig sind und welche moralischen Werten gelten. Inwieweit ist es beispielsweise akzeptabel, Kollisionsrisiken zu detektieren, algorithmisch zu bewerten und trade-off-Entscheidungen zu treffen? Diese Frage gehört zu dem neuen Forschungsfeld „dynamisches Risikomanagement", welches sich damit beschäftigt, autonome Systeme in die Lage zu versetzen, Risiken zu detektieren und zu managen.

DYNAMIK DER ÄNDERUNGEN IN KOMPLEXEN SYSTEMEN

Die beständige Veränderung an Systemen wird durch neue Technologien zunehmen und es besteht die Herausforderung für die Domäne Safety, die Bewertung neuer Konfigurationen und Funktionalitäten zeitnah leisten zu können. Die oben beschriebenen Problematiken werden durch die Dimension der dynamischen Veränderung erneut verkompliziert. Dabei ist der Stand der Dinge heute, dass die Risikoanalysen der einzelnen Maschinen- und Anlagenteile verschiedenster Herstellender nicht grundsätzlich einsehbar und leicht verfügbar sind. Grund hierfür ist nicht zuletzt das Wissen (intellectual property), das in diesen Informationen steckt. Die relevanten Ableitungen aus der Risikoanalyse, welche für den bestimmungsgemäßen sicheren Betrieb getroffen wurden, werden in den sicherheitsrelevanten Begleitunterlagen weitergegeben, welche von hoher rechtlicher Bedeutung sind. Obwohl es in den Begleitunterlagen Schlüsselwörter für Sicherheitshinweise gibt, ist es doch im konkreten Schadensfall das gesamte Dokument, ja, sogar Material aus dem Bereich des Marketings, welches für ein juristisches Verfahren herangezogen wird. Lediglich der definierte Austausch dieser öffentlich gemachten Teilinformationen aus der Risikobetrachtung zwischen den Beteiligten ist heute noch nicht standardisiert.

Bei Änderungen der darin angesetzten bestimmungsgemäßen Betriebsvorgaben ist häufig eine erneute Risikobewertung, zumindest teilweise, erforderlich. Die häufig beworbenen dynamisch adaptierenden Anlagen sind nach Stand der Technik Anlagen, deren Sicherheit (Safety) in allen Ausprägungen und Optionen zuvor umfassend von Expert*innen bewertet wurden.

KI IN DER ARBEITSSICHERHEIT

Alle Aspekte der Arbeitssicherheit können potenziell von KI betroffen sein und erfordern eine Risikobetrachtung (die Safety-Risikoanalyse ist damit nicht identisch). Arbeitssicherheit umfasst eine breite Aufgabenstellung, die sich z. B. von Maschinensicherheit über chemische Gefahren, Fragen der Arbeitsplatzergonomie bis hin zu Fragen eines sicheren Arbeitsweges beschäftigt. Grundsätzlich ist es denkbar, dass sämtliche Aufgaben der Arbeitssicherheit mit KI-Softwaretools unterstützt werden und die Überlegungen sind ähnlich zu den zuvor angestellten Betrachtungen KI beim Risikomanagement/Assessment einzusetzen.

Schlussbemerkung
Dem Wunsch und der Forderung zum Einsatz von KI-Software, insbesondere bei Safety-Anwendungen, ist rational zu begegnen. Eine möglichst vorbehaltlose Analyse ist für die jeweilige Applikation notwendig und die Risikobewertung ist ein zentraler Punkt der ergebnisoffenen Abwägung. Eine Risikobewertung dient nicht dazu, den Einsatz von KI in jedem Falle zu rechtfertigen, sondern sich für die vernünftigste Safety-Lösung zu entscheiden. Insbesondere der Einsatz bestimmter KI-Softwaretechnologien in direkt sicherheitskritischen Applikationen verlangt noch einiges an Forschungs- und Entwicklungsarbeit. Diese Ergebnisse spielen eine wichtige Rolle, denn für High-Risk-Applikationen (entsprechend dem EU-Vorschlag zur KI-Regulierung), die auch nicht originär Safety-Applikationen sind, werden Bewertungskriterien aus dem Bereich Safety verlangt und eine Nachvollziehbarkeit korrekter Arbeitsweise gefordert. Damit können die Arbeiten an einer Safety-Argumentation für KI-Software breite Bedeutung auch für andere Anwendungsbereiche erhalten.

Zu empfehlen ist, dass die ersten Schritte zu einer Safety-KI-Software zunächst über einfache Use Cases erfolgen sollten, um damit die Aufgabenstellung zu vereinfachen und die Methoden besser bewerten zu können. Bedauerlicherweise erhalten diese einfachen Aufgabenstellungen meist nicht die erforderliche Aufmerksamkeit in F&E.

4.2.1.3 Normungs- und Standardisierungsbedarfe für Safety

Bedarf 02-01: Geeignete Definitionen und regulative Kriterien als Grundlage
Die KI-Definition von Regulierung schärfen im Hinblick auf Safety-Handlungsbedarfe

High-Risk-KI-Systeme (im Sinne des EU AI Act [4]) können auch Systeme sein, welche nicht als Safety-Systeme gelten. Allerdings gelten ähnliche Anforderungen. Ist es vom Gesetzgeber gewünscht, künftig alle High-Risk-KI-Systeme als Safety-Systeme (im Sinne von Fail-Safe, funktionale Sicherheit) auszuführen?

Eine Regulierung, deren zu regulierender Kernaspekt nicht definiert ist, kann nicht angewandt werden.

Die aktuelle Regulierung und Normung bezüglich Safety berücksichtigt Software. Aus Safety-Sicht sollte die Definition von KI-System nur Arten von Software betreffen, die von aktueller Regulierung und Normung noch nicht ausreichend adressiert sind.

Bedarf 02-02: Forschung zu Safety-Konzepten und Standards evaluieren
F&E zur Zuverlässigkeit von KI-Technologien ist notwendig. Es werden Methoden und Verfahren benötigt, diese Technik vertrauenswürdig einzusetzen. Der aktuelle Stand ist nicht ausreichend, um belastbare risikoreduzierende Maßnahmen mit KI umzusetzen. Erste Schritte zu einer Safety-KI-Software sollten zunächst über einfache Use Cases erfolgen, um damit die Aufgabenstellung zu vereinfachen und die Methoden besser bewerten zu können. Diese einfachen Aufgabenstellungen erhalten heute meist nicht die erforderliche Aufmerksamkeit. Es gibt hier Interesse seitens Forschung und Industrie, diese Use Cases umzusetzen, aber es mangelt an Förderungen.

Bedarf 02-03: Forschung zu Safety Assurance Cases fördern und Standards evaluieren
Neben der Forschung zu KI-Absicherungsmaßnahmen und deren Sammlung ist somit die Forschung zu Assurance Cases für KI essenziell, um das aktuelle Potenzial von KI auch im safety-kritischen Kontext optimal auszunutzen und inakzeptable Risiken zu vermeiden. Verfahren sind heute nicht ausreichend, um die erforderlichen Nachweise zu erbringen. Die Tragfähigkeit von Safety-Konzepten für KI muss argumentativ basierend auf Fakten belegt werden können (Assurance-Case-Ansatz). Insbesondere der Einsatz bestimmter KI-Softwaretechnologien in direkt sicherheitskritischen Applikationen verlangt noch einiges an Forschungs- und Entwicklungsarbeit. Verfahren sind heute nicht ausreichend, um die erforderlichen Nachweise zu erbringen.

Bedarf 02-04: Safety bei autonomen Systemen
Bei explizit einprogrammiertem Verhalten zum Umgang mit Gefahrensituationen stellt sich die ethische (juristische) Frage, welche Regeln zulässig sind und welche moralischen Werte gelten. Inwieweit ist es beispielsweise akzeptabel, Kollisionsrisiken zu detektieren, algorithmisch zu bewerten und Trade-off-Entscheidungen zu treffen? Ethische Fragestellungen sind eine kulturelle und gesellschaftliche Frage und können nicht im Rahmen von Normung vereinheitlicht werden. Der Umgang mit Dilemmasituationen ist in der Praxis weniger relevant. Relevant sind algorithmische Entscheidungen bezüglich des Umgangs mit Risiken und Unsicherheiten (z. B. bei Perzeption).

4.2.2 Security

4.2.2.1 Status quo

Grundsätzlich können bei KI-Systemen, wie bei anderen IT-Systemen auch, alle üblichen Sicherheitsschutzziele wie Vertraulichkeit, Verfügbarkeit, Integrität oder auch Belastbarkeit gefährdet sein. Man spricht dann von Risiken der Verletzung eines Schutzzieles für einen Schutzgegenstand. Für Information Security steht mit DIN EN ISO/IEC 27701:2021 [128] und weiteren Unternormen bereits ein standardisiertes Informationssicherheitsmanagement inklusive diverser Maßnahmen und einer Risikobewertung zur Verfügung; sowie auch eine Prüf- und Zertifizierungsmöglichkeit. In anderen Bereichen oder Branchen gibt es ähnliche Standards, beispielsweise TISAX für die Automobilindustrie.

Ein Beispiel für zusätzliche Risiken bei Künstlicher Intelligenz ist das Risiko einer unautorisierten, unerkannten und gezielten Manipulation von Trainingsdaten. Beim sogenannten Data-Poisoning-Angriff werden die Trainingsdaten mit dem Ziel manipuliert, das gesamte KI-System zu beeinflussen, indem das KI-Modell auf Basis der manipulierten Trainingsdaten falsch trainiert wird. Die Standards z. B. der DIN-EN-ISO/IEC-27000-Reihe [131] sind eine Grundlage für Information Security und bedürfen einer Prüfung auf Ergänzungen für Systeme mit Künstlicher Intelligenz. Schwachstellen im KI-System mit Auswirkungen auf Safety, Security und Privacy können ohne Transparenz, Nachvollziehbarkeit und Erklärbarkeit sehr viel schwerer identifiziert und behoben werden. Ohne weitere Informationen über die inneren Abläufe des KI-Systems ist eine Schwachstellenanalyse vergleichbar der eines Softwaresystems mittels closed-box testing (ISO/IEC/IEEE 29119-1:2022 [464]), häufig besser bekannt als Blackbox Testing; die Schwachstellenanalyse ist dann vergleichsweise schwieriger als bei einem Glassbox-Test, da dem Tester hier lediglich eine Spezifikation oder nur noch ein Zugriff auf externe Eingaben und die Antworten des KI-Systems gewährt werden, aber keine Interna wie das verwendete KI-Modell. Es gibt hinsichtlich IT-Sicherheit und Softwaresicherheit also bereits viele Vorarbeiten. Dennoch stellt der benötigte Brückenschlag zwischen den existierenden IT-Sicherheitsstandards und der KI eine ganz wesentliche Herausforderung dar.

Ebenso wie IT-Sicherheit muss KI-Sicherheit auch entsprechend zeitlich (= über den gesamten Lebenszyklus) und umfänglich (= für alle Komponenten des KI-Systems) betrachtet werden. Um beim Beispiel zu bleiben, ein notwendiger Schutz vor Manipulation des KI-Systems beinhaltet den Schutz vor einer Manipulation des trainierten Modells und einen Schutz vor Manipulation der Trainingsdaten, und dieser Schutz muss über den gesamten Lebenszyklus erfolgen: Beginnend bei der Erstellung der Daten und Modelle; weiter über deren Verwendung und auch während des Betriebs muss der Schutz und dessen Einhaltung durch ein entsprechendes Monitoring des KI-Systems begleitet werden.

Status der Handlungsempfehlungen für IT-Sicherheit aus der Normungsroadmap Künstliche Intelligenz Ausgabe 1 (NRM KI Ausgabe 1)

In Tabelle 5 sind die Handlungsempfehlungen aus der ersten Version der NRM wieder aufgenommen, um kurz darzulegen, wo es bereits Überlegungen, bzw. Arbeiten in der Standardisierung gibt, und wo weiterhin ein Handlungsbedarf fortbesteht.

Tabelle 5: Übersicht über die Handlungsempfehlungen der Normungsroadmap 1 und deren Status quo

	Bedarf aus NRM KI Ausgabe 1	Beschreibung des aktuellen Status	in der Normung Arbeit
1	Recherche/Prüfung/Bewertung bestehender Normen, Konformitäts- und Zertifizierungsverfahren sowie vorhandener Gesetze	Eine vollständige Recherche und Übersicht ist bisher nicht erstellt worden. Normungsgremien erstellen eher keine Übersichten, sondern erarbeiten konkrete Kriterien, die noch nicht vorliegen. Für eine Studie ist ein Budget z. B. aus der Politik erforderlich. Die Normungs- und Regulierungslandschaft zu KI ist zudem noch sehr stark in Bewegung. Einen Überblick liefert das EU Observatory for ICT Standardisation [115] mit dem Report of TWG AI: Landscape of AI Standards [116].	ja teilweise

	Bedarf aus NRM KI Ausgabe 1	Beschreibung des aktuellen Status	in der Normung Arbeit
2	Empfehlungen für Akteur*innen/Marktteilnehmende	Diese wurden bisher nicht durch ein Normungsgremium aufgenommen.	nein
3	Erarbeitung von Ergänzungen/Anpassungen im Risikomanagement	Bei ISO/IEC SC 42 AI ist ISO/IEC 23894:2022 [25] zu AI-Risikomanagement kurz vor der Veröffentlichung. Aus der Perspektive Security/Privacy und hinsichtlich Safety-Themen wäre dieser Standard nochmals zu überprüfen.	ja ISO SC 42/CEN JTC21
4	Kritikalitätsstufen und IT-Sicherheit verbinden	Bei CEN/CENELEC gibt es einen Standardisierungsantrag aus Deutschland zur Klassifizierung von KI. Es ist geplant, aus KI-Sicht Risiko und Kritikalität zu behandeln. Das Ergebnis ist noch offen.	ja CEN JTC21
5	IT-Sicherheitskriterien für Trainingsmethoden definieren	Dieser Punkt wird in Handlungsempfehlung 7 mitverhandelt.	ja teilweise
6	Explainable AI schaffen	Zu diesem Thema wird bei ISO/SC 42 an einer Technical Specifications gearbeitet: ISO/IEC TS 6254 [36] Information technology – Artificial intelligence – Objectives and approaches for explainability of ML models and AI systems. Bei DIN wurde aktuell eine DIN SPEC 92001-3 [117] zum Thema Explainability gestartet. Forschungsbedarf: Offen ist weiterhin die Initiierung von Grundlagenforschung, die zusätzlich erforderlich ist, da Methoden noch nicht vollständig und breit erforscht und anwendbar sind.	ja teilweise ISO SC 42; DIN SPEC 92001-3 [117]
7	Controls für IT-Sicherheit für KI definieren	Bisher gibt es Studien der European Union Agency for Cybersecurity (ENISA) [118], [119], dem BSI [81] und der Fraunhofer-Gesellschaft zusammen mit dem BSI [120]. Der Bedarf wurde bisher nicht durch ein deutsches Normungsgremium aufgenommen und deshalb erneut in die NRM Ausgabe 2 aufgenommen im Rahmen von Prüfung und Zertifizierung. Hinzu kommt die Standardisierungsanforderung aus dem Entwurf des AI Act zu Cybersecurity. Auf ISO/IEC Ebene gibt es in ISO/IEC JTC 1/SC 27 eine erste Aktivität, Angriffen wie Data Poisoning Maßnahmen gegenüberzustellen, das Projekt ISO/IEC 27090 [121] ist aber noch ganz am Anfang (WD-stage). Hierzu sollten Aktivitäten verstärkt werden, da Listen mit für KI geeigneten Maßnahmen (Controls) relevant für eine Zertifizierung sind.	ja, teilweise ISO Liason SC 27/SC42

	Bedarf aus NRM KI Ausgabe 1	Beschreibung des aktuellen Status	in der Normung Arbeit
8		Security by Design und Default wird im Cybersecurity Act gefordert. Das Kriterium ist Bestandteil einer Standardserie für sichere Softwareentwicklung, z. B. ISO/IEC 27034 Informationstechnik – IT Sicherheitsverfahren – Sicherheit von Anwendungen [122], [123], [124], [125], [126], [127]. In einem möglichen ergänzenden Security-Standard für KI müsste dieser Punkt übernommen werden, ebenso wie Prüf- und Zertifizierungsmechanismen.	nein
9	KI-Security-by-Design und KI-Security-by-Default	Für diese Anforderung ist bei ISO/IEC JTC 1/SC 42 die KI-Standardserie ISO/IEC 5259 Artificial intelligence – Data quality for analytics and machine learning (ML) [39] in Erarbeitung.	ja ISO SC 42
10	Verifikation der Herkunft und Schutz der Daten	Alle Datensituationen einer KI erfordern IT-Sicherheitsmaßnahmen. Für Aktivitäten siehe Handlungsempfehlung 7 der NRM KI Ausgabe 1.	ja, teilweise ISO SC 42
11	IT-Sicherheit der Trainings-Daten	Siehe Handlungsempfehlung 7 der NRM KI Ausgabe 1.	nein
12	IT-Sicherheitskriterien für lernende Systeme definieren	KI-Algorithmen müssten mit einer verifizierbaren Identität versehen und ihre Funktion und Arbeitsweise in Dokumentationen festgehalten werden. Ergebnisse sollten möglichst nicht nur als Wahrscheinlichkeitswert einer Ergebnisklasse als Grundlage einer Entscheidung, sondern als Konfidenzintervall ausgewiesen werden. Für diese Empfehlung ist aktuell kein Gremium tätig.	nein
13	Verifizierbare Identität für KI-Algorithmen	Forschungsbedarf	nein
14	Auswirkungen von Verfügbarkeit von Ressourcen	Forschungsbedarf	nein

Regulierung, Entwurf AI Act und Standardisierungsanforderungen

Auf europäischer regulatorischer Ebene wurde dem Thema Security in den letzten Jahren mehr Aufmerksamkeit gewidmet mit speziellen Verordnungen zu Cybersecurity. Die EU-Richtlinie für Netzwerk- und Informationssicherheit (NIS), die aktuell überarbeitet wird, ist in Deutschland als IT-Sicherheitsgesetz 2.0 umgesetzt. Hinzu kommt der europäische Cybersecurity Act (CSA), der weitere Security-Anforderungen, u. a. Security/Privacy by Design and Default, beschreibt und der ENISA mehr Kompetenzen und Aufgaben verleiht. Geplant ist weiterhin der EU Cyber Resilience Act, der die Anforderungen weiter erhöht (siehe Kapitel 1.4).

Für KI befindet sich der Artificial Intelligence Act (AI ACT) in Vorbereitung (siehe Kapitel 1.4). Dieser beinhaltet Handlungsanweisungen in Abhängigkeit von einer Risikobewertung zum Einsatz der Künstlichen Intelligenz insbesondere für Hochrisiko KI-Anwendungen. Enthalten ist neben anderen die Anforderung an Cybersecurity, welche immer in Zusammenhang mit weiteren Anforderungen wie Risiko- und Qualitätsmanagement, Logging and Monitoring, Transparenz und Informationen, menschliche Aufsicht, Genauigkeit und Robustheit steht. In Summe wird eine Konformitätsbewertung erwartet, die durch Standards unterlegt werden soll. Deren Entwicklung und Bereitstellung soll möglichst durch die europäischen Standardisierungsorganisationen erfolgen.

Der zugehörige Entwurf der Standardisierungsanforderungen enthält in Kapitel 2.8 die Anforderung an „Cybersecurity specifications for AI systems“ [4]: Der Entwurf fordert europäische Standards, die geeignete organisatorische und technische Lösungen bereitstellen, um sicherzustellen, dass KI-Systeme resistent gegen Veränderungen ihrer Nutzung, ihres Verhaltens, ihrer Performanz sind und Schutz bieten gegen eine Verletzung der Sicherheit (Security). Die organisatorischen und technischen Lösungen sollen soweit möglich Cyberangriffe auf spezielle KI-Subkomponenten wie beispielsweise Trainingsdaten (vgl. Data-Poisoning-Angriff) verhindern sowie die Sicherheit und die Funktion der zugrunde liegenden Informations- und Kommunikationsinfrastruktur gewährleisten. Die technischen Lösungen sollen dabei immer entsprechend der relevanten Umstände und der Risiken gewählt werden.

Status quo

Prüfung und Zertifizierung Privacy/Datenschutz beim Einsatz künstlicher Intelligenz (Level 5)

Der Schutz personenbezogener Daten gemäß der Datenschutz-Grundverordnung (DSGVO) gilt auch bei Künstlicher Intelligenz. In Art. 4 Nr. 2 DSGVO heißt es: Verarbeitung ist jeder mit oder ohne Hilfe automatisierter Verfahren ausgeführte Vorgang oder jede solche Vorgangsreihe im Zusammenhang mit personenbezogenen Daten. Dazu gehört das Erheben, das Erfassen, die Organisation, das Ordnen, die Speicherung, die Anpassung oder Veränderung, das Auslesen, das Abfragen, die Verwendung, die Offenlegung durch Übermittlung, Verbreitung oder eine andere Form der Bereitstellung, der Abgleich oder die Verknüpfung, die Einschränkung, das Löschen oder die Vernichtung von Daten. Wenn diese Vorgänge mittels automatisierter Verfahren stattfinden, spricht man von „automatisierter Verarbeitung“. Hinzu kommt die Sicherstellung der Datensicherheit gemäß Art. 32, der eine Risikobetrachtung und IT-Security-Maßnahmen beinhaltet.

Damit muss bei jeder Risikobetrachtung und zugehörigen Maßnahmen die DSGVO berücksichtigt werden. In Deutschland wurde das Datenschutz-Anpassungs- und -Umsetzungsgesetz EU (DSAnpUG-EU) [129] ergänzt. Je nach Anwendung und auf Basis der Zustimmung zur Verarbeitung spielen insbesondere die Methoden der Pseudonymisierung und/oder Anonymisierung eine wichtige Rolle, die leider in Zusammenhang mit KI-Systemen oftmals nicht ausreichend sind, da z. T. Rückidentifizierungen von Personen möglich sein können, z. B. durch die MRT-Aufnahme (Magnetresonanztomografie) des Kopfes bei einer seltenen Erkrankung, die in einem bestimmten Zeitraum in einem bestimmten Krankenhaus behandelt wurde. Je nach Eingruppierung der Daten (personenbezogen oder nicht und anonymisiert oder nicht) muss die Privatsphäre der Daten und der zugrunde liegenden Personen entsprechend anders geschützt werden. Zu berücksichtigen sind die spezifischen datenschutzrechtlichen Vorgaben gemäß DSGVO sowie die Anforderungen aus der DIN EN ISO/IEC 17065:2013 [17] für die Konformität. Weiterhin stellt die Standardisierung eine Normenreihe als Privacy Framework mit DIN EN ISO/IEC 29100:2020 [133] und DIN EN ISO/IEC 29134:2020 [134] Impact Assessment und DIN EN ISO/IEC 29151:2022 [135] Leitfaden zur Verfügung sowie DIN EN ISO/IEC 27701:2021 [128] (Datenschutz entlang DSGVO).

Eine offizielle Prüfung und Zertifizierung entsprechend des Verfahrens aus DIN EN ISO/IEC 17065:2013 [17] nach DSGVO ist noch nicht veröffentlicht, aber in Kürze geplant.

Eine Prüfung und Zertifizierung nach DIN EN ISO/IEC 27701:2021 [128] (bei Verwendung von DIN EN ISO/IEC 27001:2017 [480]) ist möglich, wobei diese keine Prüfung und Zertifizierung nach DSGVO umfasst, sondern diese unterstützt.

Weitere Informationen bietet der Bitkom-Leitfaden „Machine Learning und die Transparenzanforderungen der DSGVO“ [136].

Für Beispiele zu existierenden Prüfungen und Zertifizierungen in den Bereichen Safety, Security und Privacy siehe Anhang 13.3.

4.2.2.2 Anforderungen, Herausforderungen und Normungs- und Standardisierungsbedarfe für Security

1. **Herausforderung: Definition von Schutzzielen auf der Ebene von Prozessen und Daten innerhalb der KI-Komponente**

Wie ausgeführt beziehen sich IT-Sicherheitsziele wie beispielsweise Integrität immer auf einen Gegenstand, für welchen dieses Schutzziel erreicht werden soll. Gegenstand im Sinne einer Prüfung können Prozesse, Daten oder physikalische Komponenten darstellen, beispielsweise KI-Trainingsdaten. Zur gezielten Beschreibung und Prüfung ist es daher notwendig und wünschenswert, die KI-Komponente, also die Systemkomponente, die Künstliche Intelligenz bereitstellt,

weiter zu unterteilen. Damit lassen sich gezielter entsprechende Schutzziele und damit auch einzelne Maßnahmen zur Erhöhung des Schutzes (Controls) auf Basis kleiner und abgegrenzter Bereiche betrachten. Um bei dem konkreten Beispiel zu bleiben: Das IT-Sicherheitsschutzziel Integrität soll für die Trainingsdaten gelten, um dem Angreifenden weniger Möglichkeiten für einen Data-Poisoning-Angriff zu bieten.

Die Herausforderung besteht darin, die Zerlegung der KI-Komponente selbst in verschiedene Daten oder verschiedene Teilprozesse so vorzunehmen, dass existierende Angriffe möglichst als eine Verletzung von Schutzzielen einzelner Subkomponenten beschrieben werden können. Im Detail kann dies zwar für verschiedene KI-Verfahren anders sein, aber eine möglichst abstrakte Zerlegung einer KI-Komponente hilft, Angriffe und Gegenmaßnahmen für ganze oder gar mehrere Klassen von KI-Verfahren zu beschreiben. Dies ermöglicht dann auch, KI-verfahrensübergreifend für eine erfolgreiche Zertifizierung eines KI-Systems den Einsatz dieser Maßnahmen (Controls) vorzuschreiben. Zugleich hilft eine Aufteilung der KI-Komponente, die Komplexität besser zu verstehen und Probleme, die im Zusammenspiel der KI-Komponente mit dem Gesamtsystem entstehen können, zu erkennen. Innerhalb des Lebenszyklus eines KI-Systems haben einzelne Subkomponenten unterschiedliche Wirkungen, auch dieses soll ein Modell abbilden. Letztlich ist diese Zerlegung auch ökonomisch geboten, da sie hilft, den Wirkungsbereich für entsprechende Maßnahmen sinnvoll zu begrenzen und so Ressourcen für die Durchsetzung der Sicherheits-, Safety- und Privacy-Schutzziele gezielt einzusetzen.

Zur Zerlegung in KI-Subkomponenten macht ISO/IEC 22989:2022 [16] bereits einen guten Anfang. Diese Zerlegung, zu der es in der Forschung bereits weitere Ansätze gibt (z. B. ISTQB-Lehrplan zum Certifed Tester AI Testing [137] [79] (Bild 1, Seite 30)), sollte in der Standardisierung ebenfalls weiterverfolgt und für einzelne KI-Verfahren oder Verfahrensklassen aufgegriffen und ausgebaut werden. Im Folgenden ein Diskussionsvorschlag für ein abstraktes Komponentendiagramm, welches über die Grenzen von KI-Verfahren hinweg generisch genug ist, um als Basis für verfeinerte Komponentendiagramme zu dienen. Es sollte aber untersucht werden, inwieweit es nötig ist, dies für verschiedene KI-Verfahren zu adaptieren. Für einzelne KI-Verfahren (Kapitel 4.1.1.1) können manche KI-Subkomponenten oder Prozessschritte aus dem generischen Komponentendiagramm weggelassen werden.

In Abbildung 25 ist die KI-Komponente zerlegt in verschiedene Prozesse und Daten dargestellt, die als Subkomponenten die KI-Funktionalität beeinflussen und somit mit Maßnahmen zur Erreichung der jeweils relevanten Schutzziele geschützt werden. Nicht alle Subkomponenten sind in allen KI-Verfahren vorzufinden. Die Begriffe der Abbildung 25 erweitern die in ISO/IEC 22989:2022 [16] beschriebenen Subkomponenten und sind außerdem farblich den dort vorgeschlagenen Lebenszyklusphasen zugeordnet. Die Begriffe sind in Tabelle 6 näher erläutert.

79 https://www.istqb.org/certifications/artificial-inteligence-tester

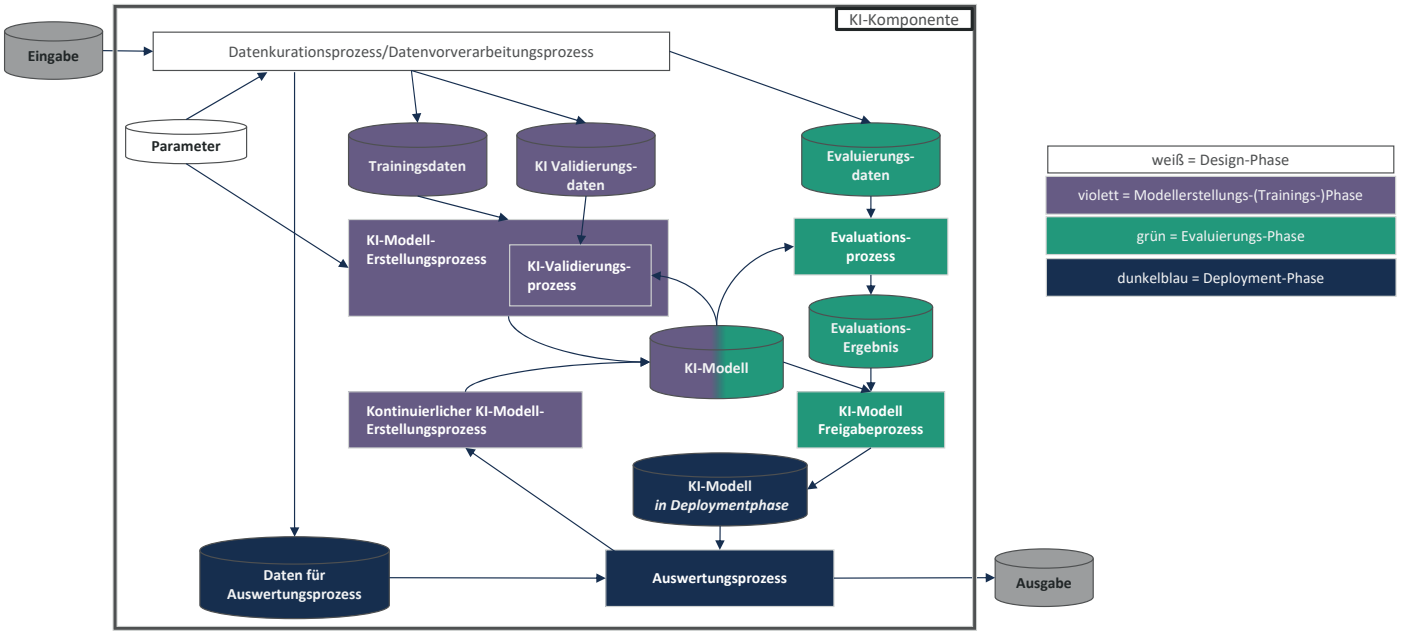


Abbildung 25: Komponentendiagramm (Quelle: Dr. Henrich Pöhls)

Tabelle 6: Beschreibung der KI-Subkomponenten (Prozesse, Daten), die sich abstrakt in einer KI-Komponente identifizieren lassen. Nicht alle Subkomponenten sind in allen KI-Verfahren vorzufinden.

Übergeordnete Begriffe	Beschreibung
KI-Komponente	Systemkomponente, die Künstliche Intelligenz bereitstellt; bestehend aus mehreren Subkomponenten
KI-Algorithmus	Gesamtheit aller Prozesse und aller Daten und Parameter, welche ein KI-Verfahren ausmachen
KI-Subkomponente (Daten, Prozesse, Modell)	**Beschreibung**
KI-Modell (in mehreren Phasen)	Daten, die das Wissen bereitstellen, welches mittels des KI-Algorithmus KI-Modellerstellungsprozess unter Zuhilfenahme weiterer Informationen wie der Trainingsdaten erstellt wurde. Das KI-Modell wird während der Trainingsphase erstellt, während der Evaluierungsphase bewertet und erst das KI-Modell in der Deployment-Phase wird vom KI-Auswertungsprozess zur Erzeugung der Ausgabe der KI-Komponente benutzt.
Auswertungsprozess	Prozess, der mittels der Daten für Auswertungsprozess und des KI-Modells eine Ausgabe erzeugt
Trainingsprozess/ KI-Modellerstellungsprozess	Prozess, der mittels der Trainingsdaten und weiterer Eingaben wie KI-Modellevaluierungsergebnis, Hyperparameter oder interner Modellparameter ein neues KI-Modell erzeugt oder ein existierendes KI-Modell erweitert. Wenn es sich um einen kontinuierlichen KI-Modellerstellungsprozess handelt, dann fließen außerdem Informationen aus dem Auswertungsprozess (Ergebnisse, interne Werte, aber auch Daten für den Auswertungsprozess) mit ein, beispielsweise im Rahmen eines KI-Verfahrens, das weiterlernt (continuous learning).
KI-Validierungsprozess	Prozess, der mittels Validierungsdaten und anderer Parameter ein bestehendes KI-Modell bewertet und den KI-Modellerstellungsprozess steuert

Übergeordnete Begriffe	Beschreibung
KI-Evaluationsprozess (im Rahmen des Testprozesses)	Prozess, der mittels der Evaluierungsdaten ein existierendes KI-Modell evaluiert und ein Modellevaluierungsergebnis erzeugt
Datenkurationsprozess/ Datenvorverarbeitungs-prozess	Prozess zur Überführung der Eingabe (raw data) in eine Darstellung, welche sich für die Anwendung der Prozesse (KI-Trainingsprozess, KI-Auswertungsprozess, KI-Validierungs-prozess) des KI-Algorithmus eignet
Trainingsdaten	Daten für den Trainingsprozess, erzeugt durch einen Datenkurationsprozess/Datenvorver-arbeitungsprozess
KI-Validierungsdaten	Daten für die interne Bewertung des Trainingsprozess erzeugt durch einen Datenkurations-prozess / Datenvorverarbeitungsprozess
Testdaten	Daten für den KI-Evaluationsprozess, erzeugt durch einen Datenkurationsprozess/Datenvor-verarbeitungsprozess
Eingabe (raw data)	Unaufbereitete Daten, die der KI-Komponente als Eingabe aus dem KI-System übergeben werden, beispielsweise Bildinhalte bei einer KI-Komponente zur Bilderkennung
Daten für Auswertungs-prozess	Durch den Datenkurationsprozess/Datenvorverarbeitungsprozess aufbereitete Daten, die dem KI-Auswertungsprozess zugeführt werden, um eine Ausgabe zu erzeugen
Ausgabe	Ergebnis der KI-Komponenten, welche durch Anwendung eines KI-Algorithmus und eines KI-Modells aus der Eingabe ermittelt wurden

Die unten stehende Tabelle 7 ordnet Prozesse und Daten der abstrakten Subkomponenten den Phasen des Lebenszyklus zu.

Tabelle 7: Lifecycle stages in Anlehnung an ISO/IEC 22989:2022 [16].

Lebenszyklus	Beschreibung
Design-Phase (siehe Abbildung 25 weiße Markierung)	Entwicklung der Parameter und Auswahl des KI-Verfahrens; die zur Generierung der Parame-ter nötigen Prozesse fallen in die Inception-Stage (ISO/IEC 22989:2022 [16]), sowie teilweise in die Phase Design and Development.
Modellerstellungs-(Trainings-)Phase (siehe Abbildung 25 violette Markierung)	Erstellung eines KI-Modells; umfasst teilweise die Phase Design and Development nach ISO/IEC 22989:2022 [16].
Evaluierungs-Phase (siehe Abbildung 25 grüne Markierung)	Umfasst u. a. den Evaluierungsprozess und den KI-Modell-Freigabeprozess; diese Prozesse sind ggf. in den Phasen Verification and validation, Continuous validation, Re-evaluate nach ISO/IEC 22989:2022 [16] anzusiedeln. (Hinweis: ISO/IEC 22989:2022 [16] verwendet hier den Begriff „Validation“ – dieser hat mehrere Bedeutungen in unterschiedlichen aber hier relevanten Kontexten, Näheres hierzu in Kapitel 9 und Kapitel 4.4.2.3.)
Deployment-Phase (siehe Abbildung 25 dunkel-blaue Markierung)	Nutzung des KI-Modells zur Erzeugung von Ausgaben auf Basis von Eingaben; entspricht der Deployment Phase nach ISO/IEC 22989:2022 [16].

ANGRIFFE AUF DIE KI-SUBKOMPONENTE ALS VERLETZUNG VON SCHUTZZIELEN

KI-Angriffe sollten nach Möglichkeit als Verletzung von Schutzzielen für KI-Subkomponenten aufgefasst werden. Neue spezielle Angriffsmöglichkeiten auf die Daten wie ein Data-Poisoning-Angriff oder Angriffe zur Extraktion von Trainingsdaten aus trainierten Modellen (Privacy Attacks) können dann auch durch IT-Sicherheitsmaßnahmen verringert werden.

Viele Angriffe von KI-Systemen hängen dabei stark von den zugrunde liegenden Daten ab. Gerade bei personenbezogenen Daten zielen die Angriffe entsprechend auf die Privatsphäre ab. Bei nicht personenbezogenen Daten geht es primär um den wirtschaftlichen Schaden z. B. durch die Manipulation der Systeme oder durch die Preisgabe von geheimen Daten, die beispielsweise nur für das Training genutzt wurden, sich aber später aus dem trainierten Modell gewinnen lassen. Daher sind die verschiedenen Daten und die mit ihnen verbundenen Prozesse ein primärer Schutzgegenstand. Die Risikoanalyse und das Datenmanagement sollten natürlich immer über den gesamten Lifecycle erfolgen.

Es bietet sich zuerst eine Analyse vorhandener Standards zu Security-Management (DIN EN ISO/IEC 27000er Folge [131]), Lifecycle, Funktionsdarstellung, Modularisierung, sicherem Softwareentwurf (ISO/IEC/IEEE 29119-2:2021 [465], DIN EN ISO/IEC 27037:2016 [130]) und KI-Ecosystem in Hinblick auf Security und Privacy in KI an; zusätzlich die Analyse der in Arbeit befindlichen KI-Standards der Gremien CEN CENELEC JTC 21 und ISO/IEC SC 42, ISO/IEC 27090 [121], ISO/IEC TR 27563 [138]; desweiteren die Analyse der geltenden Regulierungen im Zusammenhang mit KI wie der EU Cyber Security Act (CSA), die ENISA-Studie „Securing Machine Learning Algorithms“ [119], die Network Information Security Directive (NIS-Richtlinie) und der dortigen geplante Artificial Intelligence Act (AI Act) inklusive der Standardisierungsanforderungen. Auch die Studien des deutschen BSI [81] und der Fraunhofer-Gesellschaft [120] mit untersuchten Kriterien und Verfahren unterstützen die Entwicklung eines Standards mit handhabbaren Prüfungs- und Zertifizierungskriterien sowie -verfahren.

Bedarf 02-05: Abstrakte Zerlegung der KI-Komponente in Daten und Prozesse

Aktuelle Komponenten eines KI-Systems, aufbauend auf dem aktuellen Stand aus ISO/IEC 22989:2022 [16], weiter verfeinern (der aktuellen Forschung bzw. dem Diskussionsvorschlag entsprechend) und zerlegen zur genauen Beschreibung der Angriffe und Verwundbarkeiten. Das Ziel ist ein abstraktes Komponentenmodell zur weiteren Verwendung für die Beschreibung von Risiken und Maßnahmen für verschiedene KI-Verfahren und zur KI-Zertifizierung.

Bedarf 02-06: Existierende KI-Angriffe und Risiken mit existierenden zertifizierbaren IT-Sicherheitszielen abgleichen

Schafft man eine Abbildung von Angriffen auf KI-Komponenten (z. B. Data Poisoning) auf IT-Sicherheits-Schutzziele entsprechend einer Beschreibung der schutzwürdigen Gegenstände der KI-Komponenten, so ermöglicht dies, existierende Bausteine aus der Prüfung und Zertifizierung von IT-Systemen auch möglichst schnell für KI-Systeme wiederzuverwenden. Als Basis für ein solches Mapping sollten die bestehenden Dokumente der ENISA [119] oder des BSI (referenz cloud AI-Katalog) [81] (Letzeres ggf. mit ISO SC 38) weiterverfolgt und der Normung möglichst widerspruchsfrei zwischen ISO/IEC SC 27 (IT-Sicherheit) und ISO/IEC SC 42 (KI) zugeführt werden. Es existieren bereits Prüfprozesse und entsprechende Zertifizierungen für IT-Sicherheit. Diese sollten, wo möglich, Anwendung auch für die Prüfung und Zertifizierung der IT-Sicherheit des KI-Systems bzw. der einzelnen KI-Komponenten im Einsatz für das gesamte System finden. Um nicht unnötig neue Prozesse und Controls für KI-Systeme und die dort verwendete(n) KI-Komponente(n) zu beschreiben, gilt es, existierende Bedrohungen für KI-Komponenten hinsichtlich des Schutzgegenstands (ggf. auch nur für Subkomponenten der KI-Komponente wie Daten, Modell, Prozess etc.) und des IT-Sicherheitsschutzzieles (beispielsweise Integrität) zu beschreiben. Dies würde dann ermöglichen, bestimmte Controls wiederzuverwenden, beispielsweise führt Data Governance zu einem Überblick, woher Daten stammen, erschwert damit Angriffe auf die Integrität von Trainingsdaten und vermindert damit das Risiko für einen sogenannte Data-Poisoning-Angriff. Dies ermöglicht einen ersten Maßnahmenkatalog (wie im Anhang DIN EN ISO/IEC 27001 [480] oder in DIN EN ISO/IEC 27002 [481]) für KI-Sicherheit und KI-Privacy, basierend auf existierenden Maßnahmen, aufzustellen. Dies zeigt auch mögliche Lücken, also Schutzbedarfe, für die es KI-spezifischer Maßnahmen bedarf. Wo die Angriffsvektoren sehr speziell sind und sich nicht (oder nicht einfach) auf eine Menge existierender IT-Sicherheitsschutzziele abbilden lassen, sind dann spezielle Kriterien zu erarbeiten.

Bedarf 02-07: Standardisierung von KI-Produkt- und Prozessprüfverfahren für Security und Privacy

IT-Security und Privacy für KI ist sowohl ein Thema eines KI-Security-Managementsystems in der Organisation, über den Lebenszyklus und die Lieferkette, als auch aus einer funk-

tionalen Produktsicht einer singulären Softwarekomponente oder aus der Perspektive des umfangreichen KI-Systemkomplexes inklusive der möglichen Wechselwirkungen. Für alle Bereiche sollte Security- und Privacy-Standardisierung mit passenden Kontrollkriterien, Prüfwerkzeugen und Prüfverfahren sowie Managementsystemanforderungen für Prüfung und Zertifizierung erarbeitet werden, insbesondere für Machine-Learning-Methoden und in kritischen Umgebungen/Infrastrukturen. Für die Prüfung der IT-Sicherheit von Produkten, Systemen und Prozessen gibt es verschiedene etablierte Prüfverfahren und Zertifizierungsschemata. Es befinden sich neue Ansätze in Entwicklung, um sich den sich ändernden Herausforderungen in der IT-Sicherheit anzupassen. Prüfverfahren und Akkreditierungsverfahren sind essenziell, um die Qualität der Prüfung durch unabhängige Dritte sicherzustellen sowie die Nachvollziehbarkeit und Vergleichbarkeit von Ergebnissen zu verbessern.

Wie im ENISA-Report [119] angeregt, sollte weitere Forschung angepasste Security Controls für Machine Learning untersuchen, validieren, Benchmarks für ihre Wirksamkeit erstellen und hinsichtlich ihrer Implementierung standardisieren.

2. Herausforderung: Ausarbeitung eines horizontalen Querschnittsstandards und vertikale Ausprägungen

Im Laufe der Jahre haben sich in den verschiedensten Sektoren und Handlungsfeldern Security-Standards für Prüfung und Zertifizierung und in letzter Zeit auch Ansätze von KI-Standards entwickelt. Allerdings ist es aus Unternehmenssicht sinnvoll, mit möglichst wenigen und umfassenden und anerkannten Standards zu arbeiten. Eine allgemeingültige horizontale Standardisierung auch von Cybersecurity und Privacy für KI wäre aus Sicht der Wirtschaft sehr hilfreich, ebenso wie die Möglichkeit, bei Sektorenbesonderheiten ergänzende Standards zur Verfügung zu haben.

Zum Beispiel hat für Medical Devices die Food and Drug Administration (FDA) mit Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) [139] ein Framework vorgeschlagen, wie kontinuierlich lernende Systeme geprüft und zugelassen werden könnten. Das Framework fordert SaMD (Software as a Medical Device) Pre-Specifications (SPS), um die vorhergesehenen Änderungen (in Bezug auf „performance“, „inputs“ oder „intended use“) zu beschreiben, die als zulässig betrachtet werden können, dass das System bei Änderungen / kontinuierlichem Lernen keine neue Zertifizierung benötigt. Zudem fordert es ein Algorithmic Change Protocol (ACP), damit über entsprechende Tests nachgewiesen werden kann, dass die Risiken, die gemäß SPS als zulässig betrachtet werden können, ausreichend kontrolliert werden. Über das ACP erfolgt in gewissem Sinn eine automatisierte Revalidierung des Systems.

Bedarf 02-08: Ausarbeitung eines horizontalen Querschnittsstandards und vertikale Ausprägungen zu Security

Empfehlenswert wäre die Herausarbeitung von horizontalen Themen zu Cybersecurity und Privacy für KI zur Prüfung und Zertifizierung, die alle Sektoren betreffen, sowie eine Schnittstelle zu sektorspezifischen Anforderungen. Ein horizontales Thema wäre beispielsweise die Anforderung an eine geeignete Zugriffskontrolle. Als vertikale Ausprägung können wiederum spezielle Security-Anforderungen aus dem sektoralen Umfeld angesehen werden, wie u. a. für den Bereich der Medizinprodukte.

3. Herausforderung: Entwicklung von Metriken und Controls gemäß den Standardisierungsanforderungen des EU AI Act

Der geplante AI Act enthält diverse Anforderungen an Cybersecurity. In dem Entwurf des Standardisierungsrequests ist deshalb eine Standardisierung von Cybersecurity im Zusammenhang mit AI erhalten. Die erforderlichen Standards sollen zusammen mit dem Inkrafttreten des AI Act voraussichtlich ab 2024 zur Verfügung stehen.

Bedarf 02-09: Entwicklung von Metriken und Controls gemäß den Standardisierungsanforderungen des geplanten EU AI Act

Entwicklung von Standardisierung zu Cybersecurity-Anforderungen aus dem AI Act für Metriken und Controls zur Messung und Vermeidung von Cyberangriffen sowie Methoden für Prüfung, Auditierung und Zertifizierung inklusive Anforderungen an die Kriterien für die Prüfmaßnahmen und Prüfenden.

Dabei erscheint es wichtig, eine gemeinsame Arbeitsgruppe mit den Gremien der Cybersecurity und KI in den Standardisierungsorganisationen von Deutschland, der EU und eventuell auch international zu etablieren.

4. Herausforderung: Prüfkriterien für Prüfwerkzeuge zu Cybersecurity und Privacy für KI

Prüfwerkzeuge und Prüfkriterien für Prüfwerkzeuge zu Cybersicherheit und Privacy für KI gibt es aktuell nicht. Dadurch, dass KI-Systeme grundsätzlich komplexere IT-Systeme sind, gibt eine entsprechende Überlappung bei der Anwendung von Prüfwerkzeugen zu IT-Security und Privacy.

Bedarf 02-10: Prüfkriterien für Prüfwerkzeuge zu Cybersecurity und Privacy für KI
Sobald es um die Prüfung der KI-Komponente / des KI-Algorithmus geht, fehlt es noch an Prüfwerkzeugen und Methodiken. Prüfwerkzeuge für die Prüfung KI-spezifischer Kriterien sowie passende Prüfkriterien für die Prüfwerkzeuge selbst sollten entwickelt bzw. bestehende Verfahren der Prüfung von IT-Sicherheit entsprechend ergänzt werden.

5. Herausforderung: Quantifizierung von Robustheit für Modelle des Maschinellen Lernens
Eines der Zertifizierungsziele von KI-basierten Systemen ist die Quantifizierung der Robustheit. Hierbei sollen zwei Arten von Robustheit betrachtet werden: (1) Robustheit gegenüber natürlich vorkommenden Perturbationen der Eingangsdaten und (2) Robustheit gegenüber speziellen Angriffen, z. B. sogenannten Adversarial Examples. Für eine entsprechende Robustheitszertifizierung sollen Methoden und Schemata entwickelt werden, deren Ergebnisse in den Zertifizierungsprozess einfließen und die zu einer geeigneten Einschätzung der Sicherheit des Gesamtsystems beitragen.

Die Herausforderung bei der Quantifizierung von Robustheit für KI-Modelle liegt in der Wahl geeigneter Methoden. So existieren bereits aufwendige Ansätze zur empirischen Messung der Robustheit von Modellen gegenüber Angriffen, welche auf State-of-the-Art-Angriffsmethoden aufbauen [140]. Diese Ansätze sind jedoch noch nicht für alle Modelle, Architekturen und Anwendungsfälle verwendbar. Auch bieten aktuelle Methoden keine Referenzwerte zur Einordnung der Robustheitswerte.

Bedarf 02-11: Quantifizierung der Robustheit von Machine-Learning-Modellen
Aufgrund der oben genannten Herausforderungen ergibt sich die entsprechende Handlungsempfehlung und folgender Forschungsbedarf: Es sollten weitere Methoden zur Quantifizierung der Robustheit von KI-Modellen erarbeitet werden. Diese sollen in einen potenziellen standardisierten Zertifizierungsprozess aufgenommen werden. Die zukünftigen Methoden sollen es erlauben, eine Messung der Robustheit unabhängig von der Modellarchitektur und anderer Eigenschaften des Systems durchzuführen. So sollte die Anwendbarkeit auch bei großen Modellen gewährleistet bleiben. Auch sollen neue Methoden beispielsweise durch passende Referenzwerte eine relative Einstufung der Robustheit des Modells sowie des Gesamtsystems erlauben.

Die Arbeitsgruppe Sicherheit hat die identifizierten Bedarfe nach der Dringlichkeit ihrer Umsetzung bewertet. Abbildung 26 zeigt die Dringlichkeit der Umsetzung, kategorisiert nach den Zielgruppen Normung, Forschung und Politik.

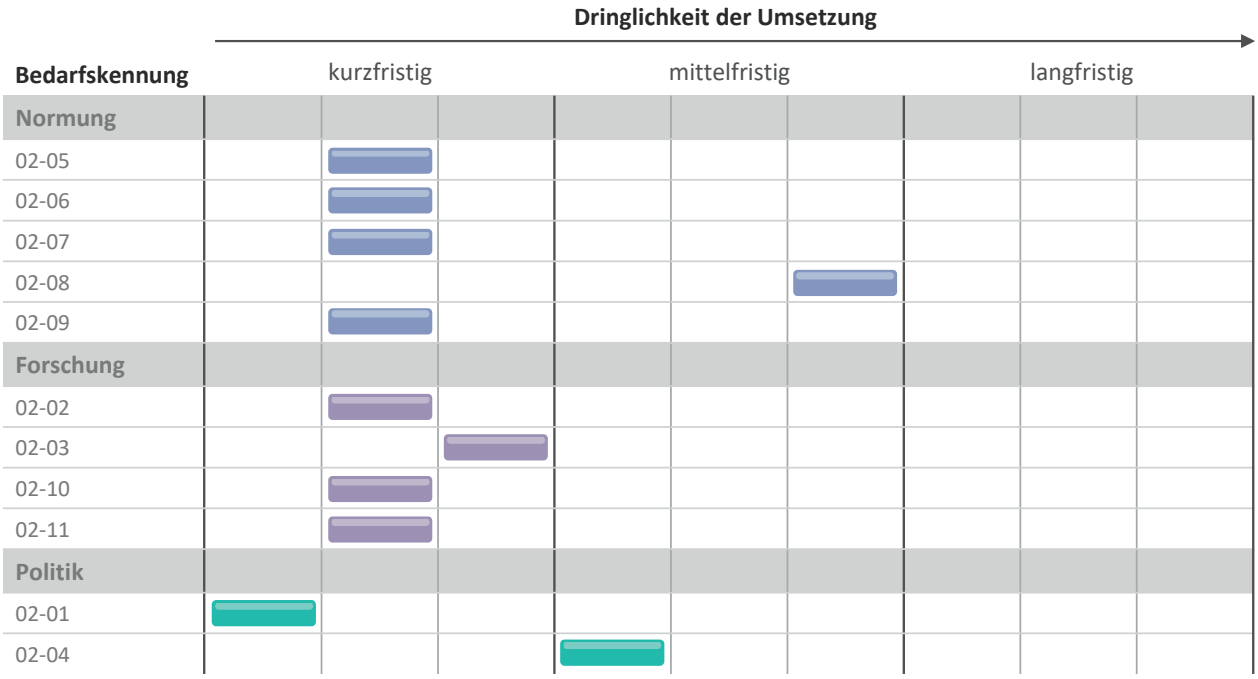


Abbildung 26: Priorisierung der Bedarfe aus Schwerpunkt Sicherheit (Quelle: Arbeitsgruppe Sicherheit)

4.3 Prüfung und Zertifizierung

Grundsätzlich lässt sich die Beurteilung der Qualität einer KI-Anwendung aus drei Perspektiven vornehmen:

1. **Gesellschaftlich-normative Bewertung**: Die erste Perspektive bezieht sich auf die Frage, ob der Einsatz einer KI-Anwendung mit gesellschaftlichen Wertevorstellungen übereinstimmt. Diese Perspektive wird vor allem im Kapitel zu den soziotechnischen Systemen behandelt (Kapitel 4.4).
2. **Sektorale, einsatzspezifische Ausprägungen**: Diese Perspektive bezieht sich auf die Frage, ob die zu beurteilende KI-Anwendung den Anforderungen ihrer Einsatzumgebung entspricht.
3. **Horizontale, anwendungsagnostische Technikbetrachtung**: Diese Perspektive bezieht sich vor allem auf die Bewertung der eingesetzten KI-Technologien und trifft beispielsweise Aussagen über die Robustheit eines ML-Modells.

Der Unterschied zwischen der dritten und der zweiten Perspektive ist, dass eine KI-Komponente (etwa ein spezifisches ML-Modell) üblicherweise in ein größeres (IT-)System (z. B. ein hochautomatisiertes Fahrzeug) eingebettet ist. Dabei beziehen sich die sektoralen Anforderungen (z. B. eine Safety- oder Security-Anforderung an das hochautomatisierte Fahrzeug) auf das Gesamtsystem, die horizontalen Anforderungen beziehen sich auf eine einzelne KI-Komponente bzw. -Technologie. Diese Anforderungen an das Gesamtsystem müssen dann in Anforderungen an die einzelnen Bestandteile, z. B. an die KI-Komponenten oder auf einzelne Datensätze, heruntergebrochen werden (vgl. auch Darstellung zu Sicherheitsargumentationen, für ein konkretes Beispiel aus dem Bereich hochautomatisiertes Fahren siehe etwa [141]).

Die normativen Anforderungen spiegeln im Kern die gesellschaftspolitische Fürsorgepflicht wider. Die Konformität mit gesellschaftlichen, ethischen und rechtlichen Rahmenbedingungen dient hauptsächlich dem Schutz von Rechtsgütern und darüber hinausgehenden Anforderungen, welche sich aus ethischen und gesellschaftlichen Debatten ergeben [68]. Als Schlüsseltechnologie der Digitalisierung durchdringt KI viele Lebens- und Arbeitsbereiche. Die KI-Konformitätsprüfungen in diesen Kategorien sollen Beeinträchtigungen von Gruppen und Einzelpersonen, Unrecht bzw. ethisch nicht gerechtfertigte Zustände der Gesellschaft verhindern und vermeiden helfen. Dies erfordert die Verankerung grundsätzlicher Handlungsprinzipien bis in die KI-Technologie hinein.

Die ethischen Anforderungen an KI

→ formulieren Grundsätze des übergeordneten Handlungsrahmens,
→ basieren auf gesellschaftlichem und politischem Konsens,
→ legen die Dimensionen der Verantwortung der Akteur*innen fest
→ und sollten europaweit in einem übergeordneten politischen Harmonisierungsprozess standardisiert werden.

Normative oder ethische Anforderungen an die KI behandeln Fragen nach dem Für und Wider des Einsatzes von KI, z. B. nach dem Einsatz von medizinischen Diagnosesystemen als Entscheidungssysteme, die Fachärzte ersetzen, oder als Systeme, die Fachärzte bei ihren Entscheidungen unterstützen (siehe u. a. [143]). Speziell für ethische Betrachtungen sind Gütesiegel vorgeschlagen worden, welche auf Wertanalyseverfahren aus einer Kombination von Zielkriterien, Indikatoren und messbaren Größen beruhen (Kapitel 4.8.3). Sämtliche Bewertungen normativer Anforderungen sollten sich auf technische Prüfungen stützen können. Normative oder ethische Anforderungen an die KI werden im Kapitel über soziotechnische Systeme näher betrachtet (Kapitel 4.4).

Die anwendungsspezifischen Anforderungen überführen die normativen Grundsätze in die konkrete Anwendung und fügen spezielle Einsatzanforderungen hinzu. Sie

→ bilden die Grundlage der Risikoeinstufung, z. B. gemäß AI Act,
→ greifen dabei relevante ethische Aspekte auf,
→ nutzen das New Legislative Framework, um komponentenweise die Konformitätsvermutung von Herstellenden zu unterstützen (Kapitel 4.3.1)
→ und formulieren im Grundsatz Anforderungen an das gesamte technische System, in dem KI eingebettet ist.

Anwendungsspezifische Anforderungen stehen im Zentrum der vertikalen Bewertung von KI, bei der geprüft wird, ob die KI für einen bestimmten Einsatzzweck geeignet ist, beispielsweise wird bewertet, ob die geprüfte Genauigkeit eines Diagnosesystems als Entscheidungsunterstützung für einen Radiologen geeignet ist oder ob ein hochautomatisiertes Fahrzeug den Anforderungen an die Safety genügt.

Die technischen Anforderungen an eine KI schließlich dienen der Überprüfung, ob die KI-Anwendung korrekt spezifiziert, entwickelt und betrieben wird. Hier wird z. B. mit bestimmten Maßnahmen festgestellt, wie genau das Diagnosesystem klassifiziert, also wie korrekt es Röntgenbilder Pathologien zuordnet.

Technische Anforderungen an eine KI
- → formulieren technisch prüfbare Anforderungen anwendungsübergreifend auf horizontaler Ebene,
- → konzentrieren sich auf KI-Technologien und ihre technisch motivierten typischen Verwendungen,
- → geben dabei ein Spektrum von Prüfmethoden unterschiedlicher Prüftiefe zur Auswahl an
- → und ermöglichen die Prüfung hybrider und eingebetteter KI in konkreten Einsatzszenarien.

Es ergibt sich von der ethischen über die normative bis zur technischen Ebene also eine dreistufige Anforderungskaskade (siehe Abbildung 27), wobei die Einstufung in Risikogruppen sinnvollerweise auf der anwendungsorientierten Ebene erfolgt, die eigentliche technische Überprüfung der Anforderungen aber auf der horizontalen Ebene geleistet wird.

Die Kaskade und die Beziehungen und Verantwortlichkeiten ist für die erfolgreiche Umsetzung gesetzlicher Handlungsrahmen in den Blick zu nehmen. Beispielsweise bestehen bei der Beantwortung der Standardisierungsanforderungen der Europäischen Kommission im Bezug auf den geplanten Artificial Intelligence Act (AI Act) in der vorliegenden Fassung viele Spielräume, um Standards auf allen genannten Ebenen und mit beliebigen Mischformen vorzuschlagen. Die Anforderungskaskade spiegelt sich auch in der Organisation der nationalen und internationalen Standardisierungs- und Normungsgremien für KI wider (Kapitel 3.2). Arbeitsgruppen und Projekte für vertikale, d. h. anwendungsbezogene Standards greifen konzeptuell auf horizontale, anwendungsübergreifende Standards zurück. Und auch die vorliegende Roadmap folgt dieser Betrachtungslogik: Die sektoralen Untersuchungen bauen auf den grundlegenden technologischen und prüfmethodischen Aspekten auf.

In Standardisierungsgremien und Expert*innengruppen herrscht folgender Konsens: Es besteht dringender Klärungs- und Handlungsbedarf, um Prüfverfahren auf den unterschiedlichen Bewertungsebenen zu etablieren und die Qualitätssicherung für vertrauenswürdige KI in Wirtschaft und Gesellschaft transparent zu gestalten. Das vorliegende Kapitel gibt Einblick in die verschiedenen Dimensionen der Umsetzung, formuliert die daraus resultierenden Fragen als Standardisierungsbedarfe und bündelt sie abschließend als dringende Empfehlung der Entwicklung und Etablierung eines horizontalen, anwendungsübergreifenden KI-Zertifizierungsprogramms.

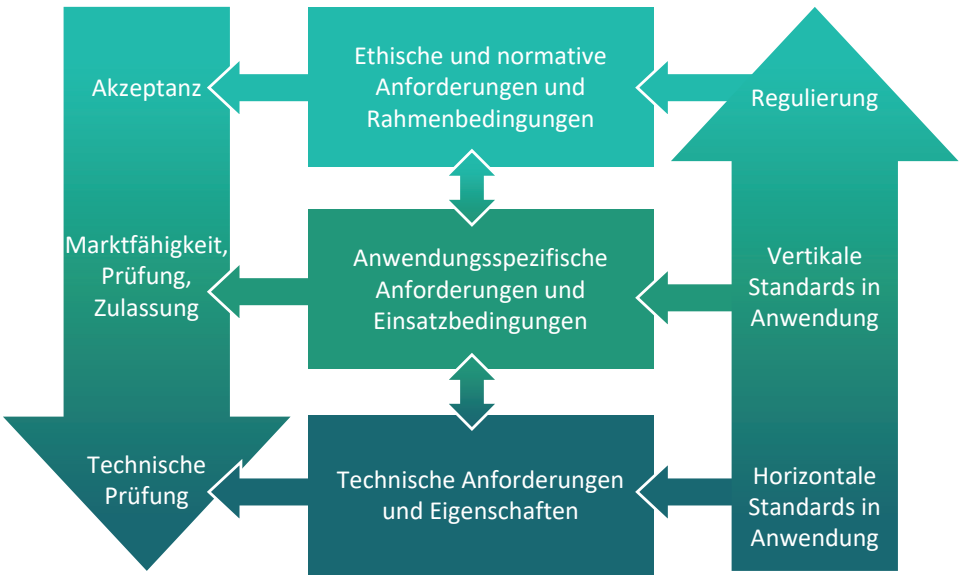


Abbildung 27: Dreistufige Anforderungskaskade (Quelle: BSI)

4.3.1 Status quo

Ein Programm zur Entwicklung einer anwendungsübergreifenden KI-Zertifizierung kann große Strahlkraft entfalten, wenn es mit bestehenden Konformitätsbewertungsverfahren und der Qualitätsinfrastruktur verträglich ist. Im Folgenden steht der Begriff „KI-Zertifizierung“ für einen Werkzeugkasten, der verschiedenen Arten der Konformitätsbewertung, die im Zusammenhang mit KI als Evaluationstätigkeiten zum Tragen kommen können, beinhaltet. Die hier naturgemäß kurze Beschreibung des Status quo konzentriert sich auf wenige zentrale Fragestellungen, wie etwa:

→ Welche Gestalt und welchen Umfang können die KI-Zertifizierungen haben?
→ Welche Qualitätsdimensionen einer KI-Zertifizierung lassen sich identifizieren? Wie lassen sie sich einordnen und ggf. auch zu Regulierungsvorgaben in Beziehung setzen?
→ Wie ist der Gegenstand der Konformitätsbewertung zu identifizieren und auszuwählen?
→ Welche Typen der Konformitätsbewertung sind relevant? Welche Prüf- und Inspektionsverfahren sowie Validierungen spielen eine Rolle?
→ Wie können KI-Zertifizierungen aussehen? Wie lassen sie sich zu bestehenden horizontalen Standards einsetzen?
→ Welche vertikalen Standards zur Umsetzung von Prüf- und Inspektionsverfahren sowie Validierungen können eingesetzt werden? Welche müssen weiterentwickelt werden und welche sind neu zu entwickeln? Siehe auch KI-Tauglichkeit in (Kapitel 3.3).
→ Welchen Nutzen haben Anwender*innen, Anbieter*innen, Hersteller*innen und Entwickler*innen von Nachweisen der KI-Vertrauenswürdigkeit? Welchen Beitrag kann ein anwendungsübergreifendes Zertifizierungsverfahren für die Akzeptanz von KI in Wirtschaft und Gesellschaft leisten?

In diesem Unterkapitel werden wichtige Konzepte und Begrifflichkeiten eingeführt, welche die Grundlage für die weitere Diskussion des Themenkomplexes „Prüfung und Zertifizierung“ von KI-Systemen bilden. Hierzu wird zu Beginn diskutiert (Kapitel 4.3.2.1), welche Entitäten (etwa Systeme, Organisationen, Personen etc.) Gegenstand einer KI-Zertifizierung sein können. Da die Diskussion so weit wie möglich auf den etablierten Konzepten von Konformitätsbewertungen aufbauen sollte, werden im Folgenden (Kapitel 4.3.2.1) wichtige Grundlagen und Begriffe der Konformitätsbewertung eingeführt. Das Unterkapitel endet mit einer Darstellung der wichtigsten Qualitätsdimensionen für vertrauenswürdige KI (Kapitel 4.3.2.1).

4.3.1.1 Regulatorische Anforderungen

Es gibt eine Reihe internationaler und nationaler Regularien, von denen für Konformitätsbewertungen von KI-Anwendungen, -Diensten und -Systemen drei für die Normungsroadmap Künstliche Intelligenz als besonders bedeutsam angesehen werden:

1. die europäische Datenschutz-Grundverordnung (DSGVO) mit den Umsetzungs- bzw. Begleitnormen für ein Datenschutzzertifikat und ein Datenschutzmanagement,
2. die Maschinenrichtlinie (wird zeitnah durch Maschinenverordnung abgelöst) und das Produktsicherheitsgesetz und ihre Umsetzung mit dem Schwerpunkt der Unfallverhütung und
3. der Verordnungsentwurf zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz der Europäischen Kommission.

Die beispielhaft genannten europäischen Regularien sind Bestandteil der Umsetzung des „New Legislative Framework“ (NLF). Es ist ein Maßnahmenpaket zur Verbesserung der Marktüberwachung (Beschluss des Parlaments der Europäischen Union (EU),768/2008/EG [144]) und für das Inverkehrbringen von industriellen Produkten (VO (EG) Nr. 1025/2012 [169]) in den Mitgliedsstaaten sowie zur Steigerung der Qualität der Konformitätsbewertung durch klare Regeln der Akkreditierung (Verordnung (EG) Nr. 765/2008 [145]). Mit dem Inkrafttreten dieser Verordnung im Rahmen dieses Maßnahmenpakets ist die Akkreditierung in der gesamten EU eine hoheitliche Aufgabe und wird in den jeweiligen Mitgliedsstaaten durch eine einzige nationale Akkreditierungsstelle wahrgenommen. In Deutschland ist die Deutsche Akkreditierungsstelle (DAkkS) die zuständige Behörde. Für unabhängige Konformitätsbewertungsstellen der ersten, zweiten oder dritten Seite mit Hauptsitz in Deutschland ist damit nur eine Akkreditierung durch die DAkkS gestattet. Details zu diesen Regularien werden in Kapitel 1.4 und Anhang 13 erläutert.

Die meisten der für KI relevanten EU Verordnungen und Richtlinien erwarten eine risikobasierte Prüfung und ggf. Zertifizierung für definierte Hochrisikoanwendungen und dafür geeignete harmonisierte Standards. Handlungsempfehlungen für diese und weitere Anforderungen zur Prüfung und Zertifizierung zu entwickeln ist Ziel dieses Kapitels.

4.3.1.2 Kompetenz von Organisationen sichern und Verbraucher*innen schützen

Eine Arbeitsgruppe der Internationalen Normungsorganisationen entwickelt z. B. einen internationalen Standard ISO/IEC 42001 [27] für KI-Managementsysteme (AI Management Systems, AIMS). Ein AIMS unterstützt Unternehmen, Organisationen und Institutionen. Dieses sollte geeignete Strategien und Prozesse für die vertrauenswürdige Entwicklung und Nutzung von KI-Systemen festlegen. Damit soll das Vertrauen und die Akzeptanz von KI als Schlüsseltechnologie der Digitalisierung erhöht werden. Die Entwicklung von KI und besonders von automatisierten Entscheidungsfindungsprozessen führt zu Herausforderungen hinsichtlich des Vertrauens und Wohlergehens der Verbraucher*innen.

Die Sicht des Verbraucherschutzes auf lernende Systeme fällt naturgemäß kritisch aus. Da lernende Algorithmen Daten mit von Menschen nicht mehr nachvollziehbarer Präzision und Geschwindigkeit verarbeiten können, verweist der Verbraucherschutz auf damit verbundene Risiken, insbesondere dann, wenn Entscheidungen getroffen werden, ohne dass die Ergebnisse von Menschen überprüft werden. Ein großes Problem ist die Verzerrung von relevanten Daten. Maschinelles Lernen beruht auf der Erkennung von Mustern innerhalb von Datensätzen. Probleme entstehen dann, wenn die Datenbasis keinen repräsentativen Querschnitt bildet und die Lernprozesse verzerrt. Dieses Problem wird besonders in der Prüfdimension „Bias, Fairness und Vermeidung von unerwünschter Diskriminierung“ behandelt (vgl. Kapitel 4.3.2.1).

Verbraucherschützer weisen auch auf die möglichen Folgen solcher Verzerrungen in algorithmischen Entscheidungssystemen (ADM-Systemen, engl.: algorithmic decision making) als spezifische KI-Systeme hin. In vielen Fällen kann eine mithilfe solcher Systeme getroffene Entscheidung signifikanten Einfluss auf natürliche Personen haben, beispielsweise in der Kreditwirtschaft, auf dem Arbeitsmarkt, im Gesundheitswesen oder bei juristischen Auseinandersetzungen. Eine Entschließung des Europäischen Parlaments fordert die Europäische Kommission zur Untersuchung der Frage auf, ob in einer zunehmend von KI und automatisierter Entscheidungsfindung beeinflussten Welt Rechtssicherheit für die Verbraucher*innen besteht.

Das Bundesministerium der Justiz und für Verbraucherschutz will die Entwicklung von KI-Systemen im Verbraucherschutz vorantreiben. Das Ministerium fördert speziell die Entwicklung von KI-Anwendungen mit dem Programm zur Innovationsförderung im Verbraucherschutz. KI-basierte Anwendungsszenarien und prototypische Lösungen, die Verbraucher*innen den Alltag und ihre Selbstbestimmung erleichtern, zielgruppengerecht konzipiert sind, die Lebensqualität erhöhen und zum Verbraucherschutz beitragen, stehen im Vordergrund.

4.3.2 Anforderungen und Herausforderungen

4.3.2.1 Grundkonzepte

Gegenstände von Konformitätsbewertungen
In diesem Kapitel wird erklärt, welche Entitäten Gegenstand einer Konformitätsbewertung sein können. Hierzu werden die entsprechenden Entitäten eingeführt, durch Beispiele untermauert und relevante Standards referenziert.

Im Allgemeinen umfassen Konformitätsbewertungen u. a. Prüfung, Inspektion, Validierung und Verifizierung.

Die Begriffe „Zertifizierung“ und „Prüfung“ werden im Nachfolgenden untechnisch als Synonyme für die verschiedenen Arten der Konformitätsbewertung sowie deren Evaluationstätigkeiten verwendet.

Wegen der Einsatzbreite und der technischen Komplexität ist eine Differenzierung der unterschiedlichen Gegenstände der Konformitätsbewertung, die auf die Erfüllung ihrer Anforderungen hin bewertet werden sollen, angemessen. Die technikspezifischen Prüfgegenstände von Konformitätsbewertungen (wie z. B. einer Prüfung) sind die KI-Systeme selbst i. S. einer Software mit oder ohne Hardwareanteilen. Es ist leicht einzusehen, dass die KI-Qualitätssicherung solche Produkte, Systeme und Lösungen als Gegenstand der Konformitätsbewertung betrachtet. Ihr Wirken und Handeln beinhaltet unmittelbare Auswirkungen auf die KI selbst oder für die Umgebung. Alle anderen Entitäten, z. B. Personen und Organisationen, (informations-)technische Systeme, Infrastrukturen usf. mögen zwar eng mit den KI-Lösungen verknüpft sein, bringen aber im Regelfall eigene, über die KI hinausgehende Wirkungspotenziale mit. Sie sind aus Sicht der KI eher übergeordnete (mittelbare) Gegenstände der Konformitätsbewertung im Bereich KI. Diese haben in der Regel auch eigene, spezifische Standards für ihre Konformitätsbewertung (für ihre „Prüfung“), wie z. B. ein KI-Managementsystem für Unternehmen und Organisationen.[80]

80 Eine Liste relevanter vertikaler Prüfstandards findet sich im folgenden Kapitel.

TECHNIKSPEZIFISCHE GEGENSTÄNDE VON KI-PRÜFUNGEN

Die Menge und Art der eingangs genannten Anforderungen richtet sich naturgemäß am Gegenstand der Konformitätsbewertung aus. Die Konformitätsbewertung von KI muss dabei der zunehmenden Digitalisierung in Wirtschaft und Gesellschaft Rechnung tragen. So findet KI zunehmend Anwendung in unterschiedlichen Produkten des Alltags, in komplexen technischen Systemen der Industrie und in speziellen informationstechnischen Lösungen. Hauptgegenstand einer KI-Konformitätsbewertung ist eine Software, welche KI-Komponenten enthält, etwa ein Modul, welches auf Maschinellem Lernen beruht. Wesentlich in der Diskussion ist, dass klassische Konzepte zur Softwarequalitätssicherung und -prüfung im Fall von KI zu kurz greifen. Ferner kann die Implementierung der Software auf einer entsprechenden Hardware eine wichtige Rolle bei der Beurteilung spielen. Prominente Beispiele sind hier Cloud-Umgebungen, Edge-Computing oder als Spezialfall neuromorphe Hardware. Die Beurteilung der Performanz einer KI-basierten Software kann im Allgemeinen stark von der Implementierung abhängen.

Mit Blick auf die eingangs beschriebenen Anforderungen lassen sich insbesondere vier technikspezifische Kategorien von KI-Prüfgegenständen unterscheiden.

1. KI-Anwendungen: Mit diesem Begriff wird eine KI-basierte Softwarelösung bezeichnet, welche Teil eines größeren IT-Systems (vgl. auch 4.) sein kann. Beispielsweise ist hier an eine Anwendung zum Kredit-Scoring oder Anomalieerkennungen zu denken.

Abbildung 28 stellt eine ML-basierte KI-Komponente und dessen Anwendung dar.

2. KI-Dienste: Hierunter wird eine KI-Anwendung verstanden, welche als Softwaredienst zur Verfügung gestellt wird. Ein typisches Beispiel sind bestimmte KI-Basisdienste (z. B. Optical-Character-Recognition-Systeme (OCR)), welche etwa von großen Clouddienstleistern zur Verfügung gestellt werden. KI-Dienste können als cloudbasierte Lösungen realisiert werden, wobei die Cloud privat oder öffentlich sein kann, oder auch über hybride Cloud-Edge-Systeme.

3. KI-Modul: Mit KI-Modul werden KI-Dienste als Bausteine in einer Kette von Lieferbeziehungen mit mehreren IT-Komponenten oder KI-Diensten bezeichnet. Technisch gesehen unterscheidet sich das KI-Modul nicht von einem KI-Dienst, die Definition hier hebt auf die Bedeutung solcher Komponenten als wichtiger Bestandteil der KI-Supply-Chain und die damit einhergehende Frage nach der Verantwortlichkeit für die Qualität des KI-Gesamtsystems ab.

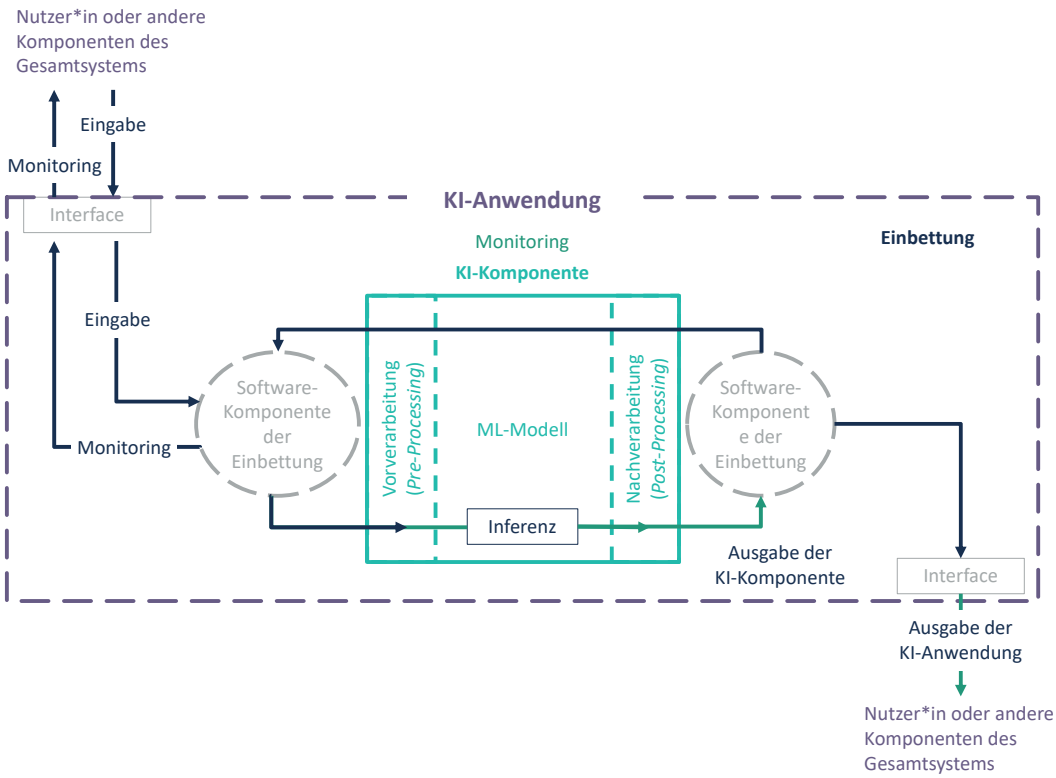


Abbildung 28: Darstellung einer ML-basierten KI-Komponente (Quelle: in Anlehnung an [120])

4. KI-Systeme: Unter einem KI-System wird schließlich ein IT-Gesamtsystem verstanden, welches eine oder mehrere KI-Anwendungen als eingebettete Bestandteile enthält. Festzuhalten ist, dass KI-Systeme üblicherweise hybrid sind, das heißt, das intelligente Verhalten wird über das Zusammenspiel mehrerer KI-Komponenten und andere klassische Softwaremodule realisiert, wobei bei den KI-Komponenten neben Maschinellem Lernen eine Vielzahl anderer KI-Methoden zum Tragen kommen kann. Die Risiken von Systemen mit KI-Komponenten müssen bei der Prüfung der KI-Komponenten berücksichtigt werden.

Lebenszyklus von KI-Systemen

Aufgrund der Einsatzbreite von KI-Systemen, der Einbettung von KI-Komponenten in komplexe technische Systeme und der Vielfalt von Technologien in teilweise hybride KI-Systeme nimmt die KI-Standardisierung den gesamten Lebenszyklus eines KI-Systems in den Blick [16][81].

Dabei werden die Phasen
- → Initiierung,
- → Design und Entwicklung,
- → Verifikation und Validierung,
- → Überführung in die Einsatzumgebung,
- → Betrieb und Überwachung,
- → kontinuierliche Validierung
- → Reevaluierung und
- → Außerdienststellung

explizit sowohl auf Dimensionen der KI-Vertrauenswürdigkeit (s. u.) als auch auf KI-Bewertungsprozesse, z. B. den kontinuierlichen Risikomanagementprozess, bezogen.

Für die Entwicklung und den Betrieb von KI-Systemen bestehen Standards und Normen aus anderen Bereichen mit Relevanz für KI-Qualität und Konformitätsbewertung. So können bestimmte KI-Prozesse in bereits bestehende, für die Softwareentwicklung erstellte Normen integriert werden, beispielsweise in die ISO/IEC/IEEE 12207:2017 [148], ISO/IEC-27034-Reihe [122], [123], [124], [125], [126], [127], sowie die ISO/IEC 25010:2011 [152] sowie ISO/IEC 25059:2022 [35].

81 Eine ausführliche Betrachtung des KI-Lebenszyklus anhand unterschiedlicher Modelle findet sich im Kapitel „Grundlagen“ (Kapitel 4.1.2.3).

Datengetriebene Systeme können sich beispielsweise in den Betriebsphasen durch einen hohen Grad an Anpassung an ihre Umwelt auszeichnen. Das Lebenszyklusmodell sieht im Betrieb als spezielle Überwachungsprozesse ein fortlaufendes Monitoring und eine kontinuierliche Validierung des KI-Systems vor. Für die Überwachung speziell von lernenden Systemen können verschiedene Verfahren in Betracht gezogen werden. Zu erwähnen ist insbesondere der MLOps-Prozess (Machine Learning Operations Prozess). MLOps setzt Machine-Learning-Modelle wiederholt ein und überwacht diese kontinuierlich. Dadurch werden die Modelle im Produktiveinsatz optimiert, wenn sich die Datenbasis ändert. Für eine Einführung in MLOps-Prozesse siehe etwa Beck et al. [153]. Im Rahmen der Konformitätsbewertung kann für solche Prozesse ggf. zukünftig eine kontinuierliche Inspektion (embedded audit agent) im Sinne der DIN EN ISO/IEC 17020:2012 [157] umgesetzt werden.

Für die technikspezifischen Prüfgegenstände – KI-Anwendungen, KI-Dienste, KI-Komponenten und KI-Systeme und für die korrespondierenden Entwicklungsprozesse und die Beobachtung und laufende Bewertung im Lebenszyklus – fordern der AI Act und die Standardization Requests der Europäischen Kommission horizontale, anwendungsübergreifende Verfahren zur Konformitätsbewertung (wie z. B. KI-Zertifizierungen). Wichtig ist es daher im Rahmen der entsprechenden Evaluationstätigkeiten der Konformitätsbewertung wie Inspektion, Prüfung oder Validierung und Verifizierung, entsprechende Prüfstandards und Prüfgrundlagen auf Level 4 zu entwickeln. Die Entwicklung und Standardisierung ist kurzfristig zu konzipieren, mittelfristig zu etablieren und mittel- und langfristig anhand des technologischen Fortschritts und der wachsenden Einsatzbreite kontinuierlich anzupassen.

ÜBERGEORDNETE (MITTELBAR) RELEVANTE PRÜFGEGENSTÄNDE

Aus der Phasenbetrachtung ergeben sich nach ISO/IEC/IEEE 12207:2017 [148], ISO/IEC 27034-1:2011 [122] und ISO/IEC 25010:2011 [152] gemeinsam mit ISO/IEC DIS 22989:2022 [16] abgeleitete Fragestellungen für die Überprüfung von Qualitätseigenschaften.

1. Welche Mindeststandards für Organisationen, Institutionen und – im Fall von verteilten KI-Systemen – Infrastrukturen sollten bei der Entwicklung und im laufenden Betrieb von KI-Anwendungen überprüft werden?
2. Welche Rollen von Personen bei der Entwicklung und im laufenden Betrieb von KI-Anwendungen sind aus den o. g. Standards identifizierbar und wie sind die entsprechenden Qualifikationsprofile z. B. für KI-Quali-

tätsbeauftragte zu prüfen? Lassen sich aus dem Qualitätssicherungsprozess Forderungen an die Personenzertifizierung z. B. für KI-Entwickler oder KI-Qualitätsauditoren herleiten?
3. Qualitätsinfrastruktur. Welche Anforderungen müssen mit der Konformitätsbewertung (z. B. Prüfung oder Zertifizierung) von KI-Systemen befasste Prüflaboratorien und Zertifizierungsstellen erfüllen?

Der Beantwortung der erstgenannten Frage nimmt sich der KI-Managementstandard an (ISO/IEC 42001 Information technology – Artificial intelligence – Management system) [27], [142]. Um die dort aufgeführten Anforderungen als erfüllt nachzuweisen, muss eine Managementsystemzertifizierung nach DIN EN ISO/IEC 170211:2015 [22] durch eine unabhängige Zertifizierungsstelle erfolgen. Im Rahmen dieser Zertifizierung wird das Managementsystem mit seinen definierten Prozessen und Rollenverteilungen sowie die Kompetenz der Organisation, dieses System normkonform zu verwalten und zu betreiben, auf wissenschaftlich reproduzierbare Weise analysiert, evaluiert und bewertet. Die Zertifizierung eines Managementsystems wird im Regelfall auf Antrag des Kunden durchgeführt. Die Zertifizierungsstelle hat dabei sicherzustellen, dass alle für die Zertifizierung notwendigen Standards Anwendung finden.

Zum zweiten Fragenkomplex sei beispielhaft auf die Personenzertifizierung des Bundesamtes für Sicherheit in der Informationstechnik (BSI) im Rahmen der Umsetzung von Zertifizierungsprogrammen zur IT-Sicherheit verwiesen. Das BSI führt Zertifizierungen von Personen nach § 9 BSI Gesetz durch und akkreditierte Konformitätsbewertungsstellen können Personen nach DIN EN ISO/IEC 17024:2012 [155] zertifizieren, weil zur Durchführung von Evaluierungen und Prüfungen zum Zwecke der Zertifizierung von Produkten und Managementsystemen sowie zur Unterstützung des BSI im Bereich IT-Sicherheitsdienstleistungen qualifizierte Personen benötigt werden. Ebenfalls können in spezifischen internationalen Standards Qualifikationsniveaus festgeschrieben sein. Zum Beispiel können in einer untergeordneten Norm zu DIN EN ISO/IEC 17021-1:2015 [22] spezifische Qualifikationsanforderungen für Zertifizierungsstellen und Auditoren, die KI-Managementsysteme bei Organisationen auditieren und zertifizieren, festgeschrieben werden. Ziel eines solchen Verfahrens für KI ist es, kompetente Personen in den Geltungsbereichen bereitzustellen sowie die Qualität und Vergleichbarkeit der Evaluierungen, Audits und Dienstleistungen sicherzustellen. Die Qualifikationsanforderungen sowie Ausbildungsprogramme für KI-Entwickler*innen, Auditor*innen, Qualitätsbeauftragte und spezielle Anwender*innen sind kurzfristig gemeinsam mit den Prüfstandards auf Level 4 zu entwickeln.

Die dritte Fragestellung führt unmittelbar zu den Rahmenbedingungen und den Verfahren der Qualitätsinfrastruktur. Sie werden im Folgenden kurz erläutert.

Typen von KI-Konformitätsbewertungen

KI-Prüfungen können grundsätzlich als Konformitätsbewertungen auf der Grundlage eines oder mehrerer Konformitätsbewertungsstandards verstanden werden, in welchem die Geltungsbereiche, die bedarfsgerechten Prüfkriterien, die Anforderungen und Nachweise, das Verfahren und das Management zur Durchführung der Bewertung beschrieben sind. Eine wichtige Kategorisierung in der Konformitätsbewertung ist, in welcher Beziehung derjenige, der die Bewertung durchführt, zum Gegenstand der Konformitätsbewertung steht. Hierbei gibt es drei Typen von Konformitätsbewertungen, die in DIN EN ISO/IEC 17000:2020 [147] definiert sind.

1. **Konformitätsbewertungstätigkeit durch eine erste Seite:** Die Konformitätsbewertungstätigkeit wird von der Person oder von der Organisation, die Gegenstand der Konformitätsbewertung ist oder die diesen anbietet, durchgeführt.
2. **Konformitätsbewertungstätigkeit durch eine zweite Seite:** Die Konformitätsbewertungstätigkeit wird von einer Person oder einer Organisation durchgeführt, die an dem Gegenstand der Konformitätsbewertung ein Interesse als Anwendender hat.
3. **Konformitätsbewertungstätigkeit durch eine dritte Seite:** Die Konformitätsbewertungstätigkeit wird von einer Person oder einer Organisation durchgeführt, die unabhängig vom Anbietenden des Gegenstands der Konformitätsbewertungstätigkeit ist und kein Interesse als Anwendender hat.

Insofern Anforderungen mit oder auch mit dem Ziel einer Darlegung ihrer Erfüllung festgelegt werden, müssen diese auch für eine solche Konformitätsbewertung geeignet sein. Ferner müssen nach ISO/IEC Direktiven die Anforderungen für den jeweils zu bewertenden Gegenstand gelten. Details der Durchführung werden separat festgelegt, z. B. Evaluationsverfahren (z. B. Prüfverfahren), Kompetenzkriterien und andere Anforderungen an den Konformitätsbewerter.

Es gilt das „Neutralitätsprinzip“, wonach die Anforderungen so formuliert bzw. separiert sein müssen, dass es keine Rolle spielt, wer ihre Erfüllung ermittelt und bewertet. Dies können interne Stellen (Konformitätsbewertung durch eine erste Seite), potenzielle Käufer*innen/ Anwender*innen (Konformitätsbewertung durch eine zweite Seite) oder Unabhängige (Konformitätsbewertung durch eine dritte Seite) sein. Ebenfalls nach ISO/IEC Direktiven dürfen von sektoralen TC keine neuen Konformitätsbewertungsstellen implementiert werden. Die Anforderungen müssen sich demnach von einem Prüflaboratorium (nach DIN EN ISO/IEC 17025:2018 [156]), einer Inspektionsstelle (nach DIN EN ISO/IEC 17020:2012 [157]), einer Validierungs-/Verifizierungsstelle (nach DIN EN ISO/IEC 17029:2020 [158]) oder einer der Zertifizierungsstellen (nach DIN EN ISO/IEC 17021-1:2015 [22], DIN EN ISO/IEC 17024:2012 [155], DIN EN ISO/IEC 17065:2013 [17]) bewerten bzw. anwenden lassen.

Dabei ist allein die Zertifizierung per Definition eine Tätigkeit „von dritter Seite“. Die übrigen Stellen (Prüflaboratorium, Inspektions-, Validierungs-, Verifizierungsstelle) können durchaus als interne oder nicht völlig organisatorisch unabhängige, gleichwohl kompetente und unparteiliche Konformitätsbewertungsstellen anerkannt oder akkreditiert werden.

Eine Akkreditierung eines Prüflaboratoriums, einer Inspektionsstelle, Validierungs- oder Verifizierungsstelle ist empfehlenswert, wenn die Konformitätsbewertungstätigkeiten für ein spezifisches KI-System z. B. unabhängig von einer späteren KI-Zertifizierung unter Anwendung von DIN EN ISO/IEC 17065:2013 [17] angeboten werden sollen. Prüfungen, Inspektionen und Validierungen sichern die Qualität eines KI-Systems als Produkt gemäß Art. 6 Abs. 1 i. V. m. Anhang II EU AI Act.

Für jeweils spezifische KI-Systeme bzw. für solche Produkte (Art 6 Abs. 1 i. V. m. Anhang II EU AI Act [4]) können Prüfergebnisse von akkreditierten Prüflaboratorien gemäß DIN EN ISO/IEC 17025:2018 [156], Inspektionsergebnisse von akkreditierten Inspektionsstellen gemäß DIN EN ISO/IEC 17020:2012 [157] oder auch Validierungen gemäß DIN EN ISO/IEC 17029:2020 [158] potenziell unabhängig von KI-Zertifizierung erzeugt werden. Gleichwohl bedarf es Prüf-, Inspektions- oder Validierungsergebnisse für jeweils spezifizierte KI-Systeme, um eine KI-Produktzertifizierung zu ermöglichen. Hierbei kann eine Zertifizierungsstelle gemäß DIN EN ISO/IEC 17065:2013 [17] diese selbst ermitteln oder erwägen, vorangegangene Prüf-, Inspektions- oder Validierungsergebnisse zu übernehmen.

Die Akkreditierung ist dabei ein wichtiges Werkzeug, welches das Vertrauen in die Vergleichbarkeit der Arbeit der Konformitätsbewertungsstellen absichert und somit aktiv zum Abbau technischer Handelshemmnisse zwischen Staaten beiträgt. Grundlage der praktischen Arbeitsweise für Akkreditierungsstellen ist die internationale Norm DIN EN ISO/IEC 17011:2018 [159]. Sie legt die Anforderungen an die Kompetenz, die einheitliche Arbeitsweise und die Unparteilichkeit von Akkreditierungsstellen fest, die Konformitätsbewertungsstellen begutachten und akkreditieren (siehe Abbildung 29).

Die gesetzlichen Anforderungen bezüglich Konformitätsbewertung und Akkreditierung sind – wie in der Produktwelt – in technischen Normen konkretisiert. Im Bereich der Konformitätsbewertung und Akkreditierung gibt es klare Unterscheidungsmerkmale an die Ebene, die bewertet wird. Daraus ergibt sich ein Levelsystem, welches in den Dokumenten EA-1/06 A-AB:2022 [170] und IAF PR4: 2015 [171] zu finden ist.

Das System der Akkreditierung und Konformitätsbewertung dient dabei zur Absicherung der Qualitätssicherungs-Prozesskette. Von Level 5 ausgehend wird immer der Gegenstand der Konformitätsbewertung betrachtet, der durch ein Unternehmen produziert oder erstellt wurde. Es handelt sich hier um Produkte, Prozesse, Dienstleistungen oder um Personen, an die gewisse Anforderungen spezifischer Qualifikationen gestellt werden. Die gesetzlichen Anforderungen, die in Normen spezifiziert wurden, sind dementsprechend von den Unternehmen einzuhalten. Sie müssen die Konformität mit diesen Anforderungen erklären und – auch für KI-Systeme als Produkte gemäß Art. 6 Abs. 1 i. V. m. Anhang II EU AI Act – nachweisen. Bei einer Konformitätsbewertung durch eine Konformitätsbewertungsstelle sind von dieser Stelle die normativen Anforderungen von Level 3 einzuhalten, um auf wissenschaftlicher Basis eine Evaluation des Gegenstands der Bewertung (Level 5) vorzunehmen, vergleichbare Ergebnisse zu erzielen und damit die erklärte Konformität des Herstellenden oder Inverkehrbringers (Kapitel 4.3.1.1) bestätigen zu können.

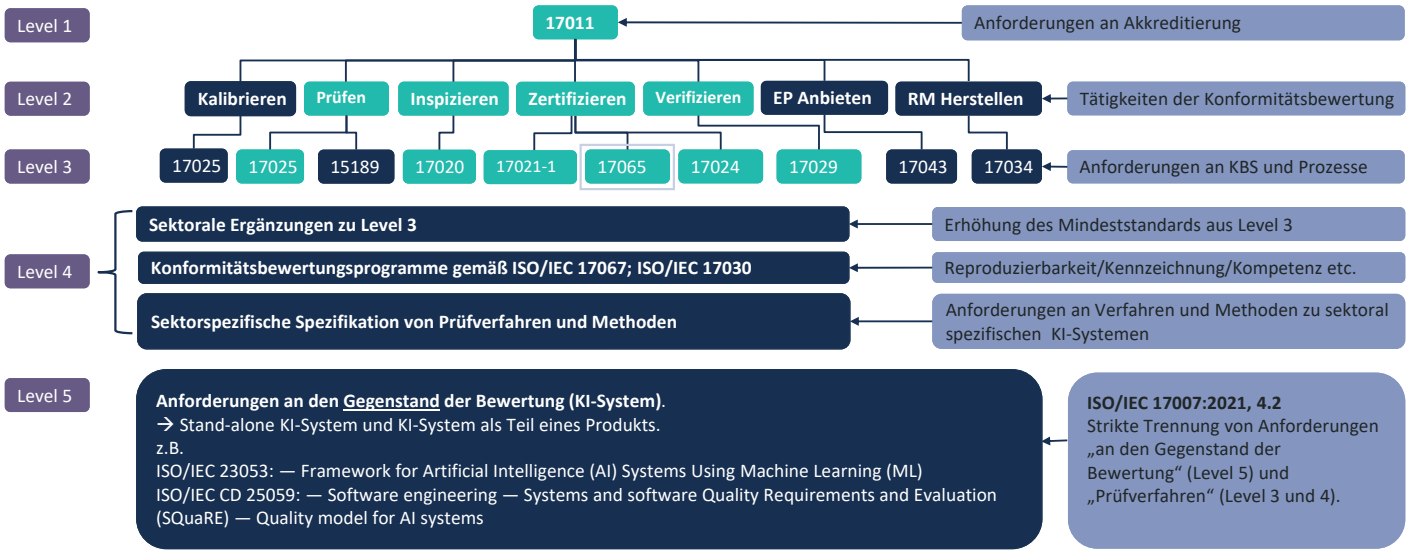


Abbildung 29: Einordnung von Konformitätsbewertungsverfahren in internationale Levelstruktur (Quelle: DAkkS)

Dimensionen der KI-Vertrauenswürdigkeit (Prüfdimensionen)

DATENQUALITÄT UND DATENMANAGEMENT

Die Qualität und Vertrauenswürdigkeit einer KI-Anwendung ist eng mit der Datenqualität verbunden. Qualitätsanforderungen an Daten umfassen beispielsweise korrekte Datenannotation oder vertrauenswürdige und relevante Datenquellen [81]. Hinreichende Datenqualität stellt für viele der weiteren Dimensionen eine wichtige Grundlage dar, wie beispielsweise als Maßnahme, um Fairness sicherzustellen, oder um ausreichende Performance eines KI-Systems zu erreichen. Eng damit verbunden sind die Anforderungen an eine sinnvolle Datenverwaltung, die beispielsweise diese Qualitätsanforderungen abbildet oder den Datenzugang regelt. Der EU-Regulierungsentwurf [4] formuliert ebenfalls Datenqualitäts- und -verwaltungsanforderungen für Hochrisikosysteme.

BIAS, FAIRNESS UND VERMEIDUNG UNERWÜNSCHTER DISKRIMINIERUNG

Eine grundlegende Anforderung für die Vertrauenswürdigkeit eines KI-Systems ist die Vermeidung unerwünschter Diskriminierung [154]. Diese Anforderung soll sicherstellen, dass eine ungerechtfertigte Ungleichbehandlung von Individuen oder Gruppen im Vergleich mit anderen Gruppen verhindert wird [63]. Ursachen für unerwünscht diskriminierendes Modellverhalten resultieren häufig aus historischen Daten, die unbalanciert sind oder ein Bias bezüglich einer bestimmten Gruppe aufweisen. Auf Grundlage von Bewertungsmaßen [siehe AG Grundlagen] lässt sich die Nichtdiskriminierung einer KI-Anwendung quantifizieren, wobei Bias und unerwünschte Diskriminierung einerseits in den Trainingsdaten gemessen werden kann, anderseits in der Ausgabe des Modells.

AUTONOMIE UND KONTROLLE

Durch die Möglichkeit, selbstständig Modelle und Trainingsparameter aus Daten zu lernen, ergibt sich für bestimmte KI-Anwendungen ein gewisser Grad an Autonomie. Je nach Kontext und Kritikalität einer Anwendung entsteht aus der Autonomie der KI-Anwendung ein Spannungsfeld zur menschlichen Autonomie der Nutzer*innen und Betroffenen. Um den Vorrang menschlichen Handelns abzusichern, muss dieses Spannungsfeld durch einen angemessenen Autonomiegrad zwischen der KI-Anwendung und Nutzerautonomie kontrolliert werden. Gleichzeitig umfasst die Dimension der Autonomie und Kontrolle aber auch, dass Nutzer*innen und Betroffene angemessen informiert und befähigt sind, um mit einer KI-Anwendung zu interagieren [120].

ERKLÄRBARKEIT, INTERPRETIERBARKEIT UND TRANSPARENZ

Die Transparenz umfasst verschiedene Aspekte wie die Interpretierbarkeit, die Erklärbarkeit, die Nachvollziehbarkeit oder die Reproduzierbarkeit der Ergebnisse und der Funktionalität einer KI-Anwendung. Die Reproduzierbarkeit von Ergebnissen eines Systems ist eine minimale Anforderung zur Nachvollziehbarkeit der Resultate. Während die Interpretierbarkeit

eines Systems impliziert, dass das System als Ganzes nachvollziehbar ist [120], bezeichnet die Erklärbarkeit lediglich, dass verständlich ist, welche Faktoren zum Zustandekommen des Ergebnisses geführt haben [16]. Die Transparenz muss dabei auf geeignete Art und in einem angemessenen Maß gewährleistet werden, sodass sie zugänglich und angepasst an den jeweiligen Nutzenden ist [4].

PERFORMANCE, LEISTUNGSFÄHIGKEIT, VERLÄSSLICHKEIT, ROBUSTHEIT, VOLLSTÄNDIGKEIT

Um eine KI vertrauenswürdig zu gestalten, müssen Nutzer*innen sich auf das System verlassen können. Aus technischer Sicht umfasst die Verlässlichkeit eines Systems verschiedene Aspekte wie die Korrektheit der Ausgaben im Regelfall, die Einschätzung der Unsicherheit der Ergebnisse oder die Robustheit gegenüber Angriffen, Fehlern und unerwarteten Situationen [120]. Performanzmetriken erlauben dabei eine messbare, qualitative und quantitative Einschätzung des Systems [16]. Wenn auch beispielsweise in Art. 15 der EU-Regulierung Anforderungen an die Verlässlichkeit von Hochrisikosystemen gestellt werden, bleibt die Übersetzung der Anforderungen in quantitative Maße und Zielwerte bislang offen und erfordert spezifisches Domänen- und Anwendungswissen.

SAFETY, SECURITY UND PRIVACY

Eine weitere Dimension für die Prüfung und Zertifizierung der Vertrauenswürdigkeit ist Sicherheit mit den Themen Safety, Information Security, Privacy, Security und Reliability. Diese werden ausführlich im Kapitel 4.2 Sicherheit vorgestellt.

4.3.2.2 Operationalisierung von KI-Prüfungen

Ziel dieses Kapitels ist die Darstellung der komplexen Zusammenhänge und Verantwortlichkeiten für KI-Systeme und der daraus sich ergebenden Konsequenzen für KI-Prüfverfahren, die über die bisherigen Betrachtungen von KI-Konformitätsbewertungen hinausgehen. Nach den bisherigen Vorüberlegungen sollten Prüfverfahren für die Vertrauenswürdigkeit von KI-Systemen drei Beobachtungen gerecht werden: komplexen KI-Lieferketten, dem hybriden Charakter vieler KI-Systeme und der Einbettung in technische Systeme.

KI-Lieferketten

KI-Anwendungen und KI-Dienste können als eigenständige Module in Liefer- und Leistungsbeziehungen zu anderen Komponenten eines (informations-)technischen Systems stehen, die für die Bewertung des Gesamtsystems relevant sind. Beispielsweise lässt sich eine KI-basierte Lösung für die Erkennung von Kreditkartenbetrug als ein zusammengesetztes KI-System aus drei Modulen aufbauen. Der Kreditkartenbetreiber liefert Transaktionsdaten, die als Rohdatensätze in ein lernendes System eines KI-Dienstleisters eingehen. Das lernende System produziert Regeln für ein Expertensystem, das bei einem Finanzinstitut angesiedelt ist und online „just in time“ Empfehlungen für die Betrugserkennung macht. In diesem Fall sind drei Akteur*innen beteiligt, die unabhängig voneinander Teile des Gesamtsystems realisieren:

- → Der Kreditkartenbetreiber ist für die Qualität der Trainings- und Testdatensätze verantwortlich (B2B-Beziehung im Gesamtsystem).
- → Der KI-Dienstleister ist für die Qualität der gelernten Regeln verantwortlich (B2B-Beziehung im Gesamtsystem).
- → Das Finanzinstitut ist dem Endkunden gegenüber für die Qualität des gesamten Betrugserkennungsvorgangs verantwortlich (B2C-Beziehung im Gesamtsystem).

Diese Liefer- und Leistungsbeziehungen mit Blick auf die KI-Anforderungen der Komponenten können im Regelfall nicht hinreichend detailliert vertraglich abgebildet werden, sondern das Gesamtsystem, d. h. jedes der enthaltenen Module, muss untersucht und die einzelnen Prüfergebnisse müssen zu einer Konformitätsbewertung des Betrugserkennungssystems zusammengefügt werden. Der Sachverhalt ist schematisch in Abbildung 30 unter Einbindung von Clouddienstleistern dargestellt.

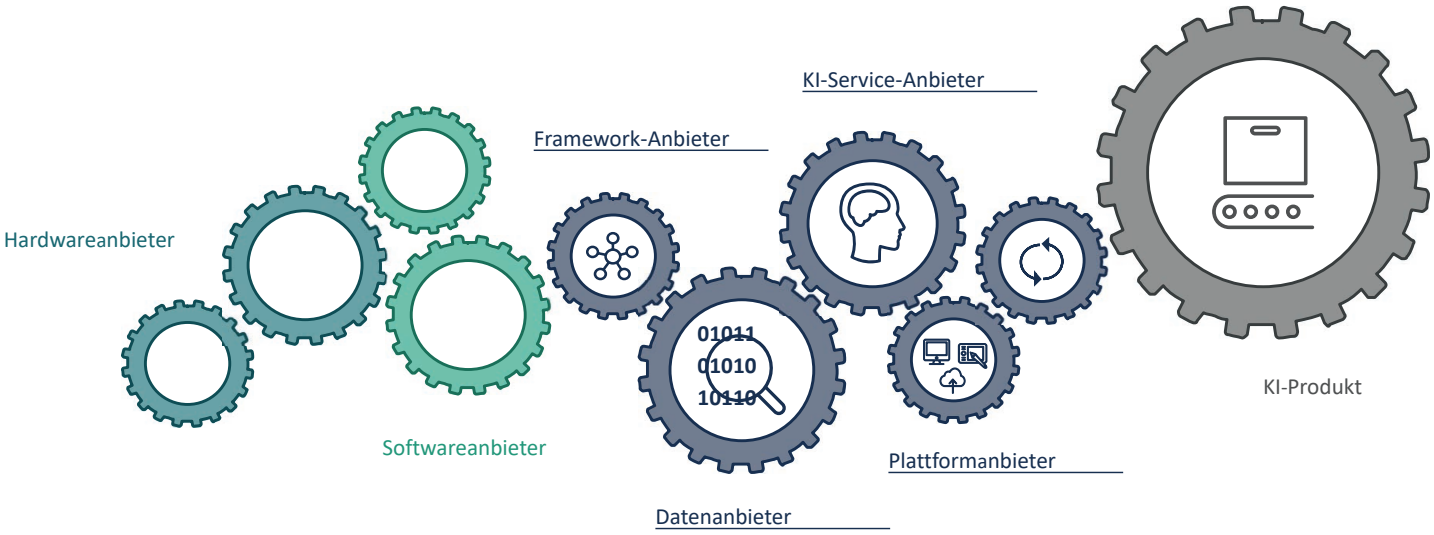


Abbildung 30: Akteur*innen in einer cloudbasierten KI-Supply-Chain (Quelle: PwC)

Hybride KI-Systeme

KI-Systeme können technologisch hybrid sein, d. h. sie bestehen aus mehreren Modulen mit unterschiedlichen KI-Technologien. Ein System zur Erkennung gesprochener Sprache besteht z. B. sinnvollerweise mindestens aus

→ einem Analog-Digital-Konverter (Mikrofon), um mittels Fourier-Transformation ein Sprachspektogramm zu erzeugen, aus dem Phoneme über die Frequenz, die Zeit und die Intensität der analogen Signale digitalisierbar sind.

→ Phoneme können je nach Sprecher, Akzent, Alter, Geschlecht oder Position im Wort variieren. Zur Erkennung von Worten und Sätzen können spezielle KI-Technologien eingesetzt werden. Als statische Modelle bieten sich spezielle Markov-Modelle an.

→ Diese sind aber insbesondere im Fehlerfall nicht flexibel genug, sodass sie durch weitere Verfahren, z. B. durch spezielle neuronale Netze, unterstützt und abgesichert werden.

Die entstehenden Produkte arbeiten sehr zufriedenstellend, was sich im technikbestimmten Alltag bestätigt. Die Technologien müssen aber auf der Basis bestehender oder zu entwickelnder KI-Prüfverfahren so überprüft werden können, dass eine qualifizierte Aussage über die Vertrauenswürdigkeit des Gesamtsystems gemacht werden kann. Interessant wird in solchen Fällen die prüfrelevante Verknüpfung mit anderen Fragestellungen, z. B. welche Technologien laufen auf dem Client, welche in der Edge? Mit anderen Worten: Worauf stützt sich die Vertrauenswürdigkeit des KI-Systems in Teilen ab?

Eingebettete KI-Systeme

KI-Anwendungen, KI-Dienste und KI-Module in komplexen Systemen sind spezielle Informationstechnologien, deren Einzelprüfung für bestehende Prüfverfahren der Gesamtsysteme, in welche die KI eingebettet sein kann, brauchbare Ergebnisse liefern muss. Die Ergebnisse von KI-Prüfungen müssen in Resultate übergeordneter Prüfungen auf der Basis bestehender Prüf- und Zulassungsverfahren einzahlen können. Für KI-Konformitätsbewertungen muss die Möglichkeit eröffnet werden, in bestehende Zertifizierungsverfahren auf Basis der DIN EN ISO/IEC 17065:2013 [17] als Teilbewertungen im Sinne einer Übernahme der Ergebnisse (nach DIN EN ISO/IEC 17065:2013 [17] T. z. 9.6) aufgenommen zu werden.

Diese drei Beobachtungen führen zu einer ganzen Reihe von Verflechtungen mit in Arbeit befindlichen und bestehenden Normen. Die für dieses Kapitel relevanten Standards sind nachfolgend in dem bestehenden Framework von Konformitätsbewertungen aufgelistet. Mit Blick auf die obigen Beobachtungen sind gleichzeitig Anpassungen dieses Frameworks an den speziellen Charakter von KI-Systemen zu prüfen.

Dokument	Titel
Level-5-Normen (Anforderungen an Gegenstand der Konformitätsbewertung)	
ISO/IEC 5259-2 [41]	Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 2: Data quality measures
ISO/IEC 5259-5 [44]	Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 5: Data quality governance
ISO/IEC TR 5469 [33]	Artificial intelligence – Functional safety and AI systems
ISO/IEC TS 5471 [34]	Artificial intelligence – Quality evaluation guidelines for AI systems
ISO/IEC 24029-2:2022 [92]	Artificial intelligence (AI) – Assessment of the robustness of neural networks – Part 2: Methodology for the use of formal methods
ISO/IEC TR 24029-1:2021 [91]	Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview
ISO/IEC 22989:2022 [16]	Artificial intelligence – Concepts and terminology
DIN SPEC 92001-2:2020 [240]	Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 2: Robustness
ISO/IEC 5259-1	Data quality for analytics and ML – Part 1: Overview, terminology, and examples
ISO/IEC 5259-3 [42]	Data quality for analytics and ML – Part 3: Data Quality Management Requirements and Guidelines
ISO/IEC 5259-4 [43]	Data quality for analytics and ML – Part 4: Data quality process framework
ISO/IEC TS 8200 [37]	Information technology – Artificial intelligence – Controllability of automated artificial intelligence systems
ISO/IEC 8183 [45]	Information technology – Artificial intelligence – Data life cycle framework
ISO/IEC 42001 [27]	Information Technology – Artificial intelligence – Management system
ISO/IEC TS 6254 [36]	Information technology – Artificial intelligence – Objectives and approaches for explainability of ML models and AI systems
ISO/IEC TR 29119-11 [132]	Information technology – Artificial intelligence – Testing for AI systems – Part 11:
ISO/IEC TS 12791 [38]	Information technology – Artificial intelligence – Treatment of unwanted bias in classification and regression machine learning tasks
ISO/IEC 24668	Information technology – Artificial intelligence – Process management framework for Big data analytics
ISO/IEC 5338 [30]	Information technology – Artificial intelligence – AI system life cycle processes
ISO/IEC TS 4213 [29]	Information technology – Artificial Intelligence – Assessment of machine learning classification performance

Dokument	Titel
ISO/IEC 5339 [31]	Information Technology – Artificial Intelligence – Guidelines for AI Applications
ISO/IEC 5394 [149]	Information Technology – Criteria for concept systems
ISO/IEC 5392 [32]	Information technology – Artificial intelligence – Reference Architecture of Knowledge Engineering
ISO/IEC 23894:2022 [25]	Information Technology – Artificial Intelligence – Risk Management
ISO/IEC TS 24462 [150]	Ontology for ICT Trustworthiness Assessment
ISO 24089 [151]	Road vehicles – Software update engineering
ISO/IEC 23053:2022 [24]	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)
ISO/IEC 27034-1:2011 [122]	Informationstechnik – IT Sicherheitsverfahren – Sicherheit von Anwendungen – Teil 1: Überblick und Konzept
ISO/IEC 27034-2:2015 [123]	Informationstechnik – IT Sicherheitsverfahren – Sicherheit von Anwendungen – Teil 2: Organisation des normativen Rahmen
ISO/IEC 27034-3:2018 [124]	Informationstechnik – Sicherheit von Anwendungen – Teil 3: Managementprozess für die Sicherheit von Anwendungen
ISO/IEC 27034-5:2017 [125]	Informationstechnik – IT Sicherheitsverfahren – Sicherheit von Anwendungen – Teil 5: Protokolle und Datenstruktur zur Kontrolle der Anwendungssicherheit
ISO/IEC 27034-6:2016 [126]	Informationstechnik – IT Sicherheitsverfahren – Sicherheit von Anwendungen – Teil 6: Fallstudien
ISO/IEC 27034-7:2018 [127]	Informationstechnik – IT Sicherheitsverfahren – Sicherheit von Anwendungen – Teil 7: Model zur Voraussage der Zusicherung von Sicherheitsanwendungen
DIN EN ISO/IEC 29101:2022 [493]	Informationstechnik – Sicherheitstechniken – Architekturrahmenwerk für Datenschutz
DIN EN ISO/IEC 29134:2020 [134]	Informationstechnik – Sicherheitsverfahren – Leitlinien für die Datenschutz-Folgen-abschätzung
DIN EN ISO/IEC 29147:2020 [494]	Informationstechnik – Sicherheitstechniken – Offenlegung von Schwachstellen
DIN EN ISO/IEC 29151:2022 [135]	Informationstechnik – Sicherheitsverfahren – Leitfaden für den Schutz personen-bezogener Daten
ETSI DGR SAI 002:2021 [497]	Securing Artificial Intelligence (SAI); Data Supply Chain Report
ETSI DGS SAI 003 [336]	Securing Artificial Intelligence (SAI); Security Testing of AI
DIN EN ISO/IEC 27001:2017	Informationstechnik – Sicherheitsverfahren – Informationssicherheitsmanagement-systeme – Anforderungen
DIN EN ISO/IEC 27002 [481]	Informationstechnik – Sicherheitsverfahren – Leitfaden für Informationssicherheits-maßnahmen

Dokument	Titel
DIN EN ISO/IEC 27701:2021 [128]	Sicherheitstechniken – Erweiterung zu ISO/IEC 27001:2021 und ISO/IEC 27002 für das Management von Informationen zum Datenschutz – Anforderungen und Leitlinien
ISO/IEC 25000:2014 [472]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE
ISO/IEC 25024:2015 [473]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of data quality
ISO/IEC 25020:2019 [474]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality measurement framework
ISO/IEC 25010:2011 [152]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models
ISO/IEC 25021:2012 [475]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality measure elements
ISO/IEC 25012:2008 [463]	Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model
DIN ISO 31000:2018 [160]	Risk management – Guidelines
ISO/SAE 21434:2021 [324]	Road vehicles – Cybersecurity engineering
ISO-26262-Reihe [455]	Road vehicles – Functional safety
ISO/IEC TR 24027:2021 [436]	Information technology – Artificial intelligence (AI) – Bias in AI systems and AI aided decision making
ISO/IEC TR 24372:2021 [437]	Information technology – Artificial intelligence (AI) – Overview of computational approaches for AI systems
ISO/IEC TR 24030:2021 [293]	Information technology – Artificial intelligence (AI) – Use cases
ISO/IEC 38507:2022 [26]	Information technology – Governance of IT – Governance implications of the use of artificial intelligence by organizations
ISO/IEC TR 24368:2022	Information technology – Artificial intelligence – Overview of ethical and societal concerns
ISO/IEC TR 24028:2020 [28]	Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence
ISO/IEC 25059:2022 [35]	System- und Software-Engineering – Qualitätskriterien und Bewertung von Systemen und Softwareprodukten (SQuaRE) – Qualitätsmodell für KI-Systeme
Level 4 Normen (Normen zu Kennzeichnungen, Spezifikation von Prüfverfahren und Methoden)	
DIN EN ISO/IEC 17050-1:2010 [489]	Konformitätsbewertung – Konformitätserklärung von Anbietern – Teil 1: Allgemeine Anforderungen

Dokument	Titel
DIN EN ISO/IEC 17050-2:2005 [490]	Konformitätsbewertung – Konformitätserklärung von Anbietern – Teil 2: Unterstützende Dokumentation
DIN EN ISO/IEC 17030:2021 [486]	Konformitätsbewertung – Allgemeine Anforderungen an Konformitätszeichen einer dritten Seite
DIN EN ISO/IEC 15408-1:2020 [445]	Informationstechnik – IT-Sicherheitsverfahren – Evaluationskriterien für IT-Sicherheit – Teil 1: Einführung und allgemeines Modell (ISO/IEC 15408-1:2009); Deutsche Fassung EN ISO/IEC 15408-1:2020
DIN EN ISO/IEC 15408-2:2020 [446]	Informationstechnik – IT-Sicherheitsverfahren – Evaluationskriterien für IT-Sicherheit – Teil 2: Sicherheitsfunktionskomponenten (ISO/IEC 154082:2008); Deutsche Fassung EN ISO/IEC 15408-2:2020, nur auf CD-ROM
DIN EN ISO/IEC 15408-3:2021 [447]	Informationstechnik – IT-Sicherheitsverfahren – Evaluationskriterien für IT-Sicherheit – Teil 3: Komponenten zur Sicherheitskontrolle (ISO/IEC 15408-3:2008, korrigierte Fassung 2011-06-01); Deutsche Fassung EN ISO/IEC 15408-3:2020, nur auf CD-ROM
ISO/IEC 15408-4:2022 [448]	Informationstechnik – IT-Sicherheitsverfahren – Evaluationskriterien für IT-Sicherheit – Teil 4: Rahmen für die Festlegung von Bewertungsmethoden und -tätigkeiten
ISO/IEC 15408-5:2022 [449]	Informationstechnik – IT-Sicherheitsverfahren – Evaluationskriterien für IT-Sicherheit – Teil 5: Vordefinierte Pakete von Sicherheitsanforderungen
DIN EN ISO/IEC 18045:2021 [75]	Information technology – Security techniques – Methodology for IT security evaluation
Level 3 Normen (Anforderungen an Konformitätsbewertungsstellen)	
DIN EN ISO/IEC 17020:2012 [157]	Konformitätsbewertung – Anforderungen an den Betrieb verschiedener Typen von Stellen, die Inspektionen durchführen
DIN EN ISO/IEC 17021-1:2015 [22]	Konformitätsbewertung – Anforderungen an Stellen, die Managementsysteme auditieren und zertifizieren – Teil 1: Anforderungen
DIN EN ISO/IEC 17021-3:2019 [485]	Konformitätsbewertung – Anforderungen an Stellen, die Managementsysteme auditieren und zertifizieren – Teil 3: Anforderungen an die Kompetenz für die Auditierung und Zertifizierung von Qualitätsmanagementsystemen
DIN EN ISO/IEC 17024:2012 [155]	Konformitätsbewertung – Allgemeine Anforderungen an Stellen, die Personen zertifizieren
DIN EN ISO/IEC 17025:2018 [156]	Allgemeine Anforderungen an die Kompetenz von Prüf- und Kalibrierlaboratorien
DIN EN ISO/IEC 17029:2020 [158]	Konformitätsbewertung – Allgemeine Grundsätze und Anforderungen an Validierungs- und Verifizierungsstellen
DIN EN ISO/IEC 17043:2022 [488]	Konformitätsbewertung – Allgemeine Anforderungen an die Kompetenz von Anbietern von Eignungsprüfungen
DIN EN ISO/IEC 17065:2013 [17]	Konformitätsbewertung – Anforderungen an Stellen, die Produkte, Prozesse und Dienstleistungen zertifizieren

Dokument	Titel
Level 1 Norm (Anforderungen an Akkreditierungsstellen)	
DIN EN ISO/IEC 17011:2018 [159]	Konformitätsbewertung – Anforderungen an Akkreditierungsstellen, die Konformitätsbewertungsstellen akkreditieren
Level 0 Norm (all. Grundlagen der Akkreditierung und Konformitätsbewertung)	
DIN EN ISO/IEC 17000:2020 [147]	Konformitätsbewertung – Begriffe und allgemeine Grundlagen (ISO/IEC 17000:2020)

In der Qualitätssicherung und -bewertung von Informationstechnologien finden sich

→ Prüfkriterien für die Definition und Beschreibung der Systemfunktionalität,
→ Kriterien, nach denen das Vertrauen in die Wirksamkeit von Systemfunktionen bewertet werden kann, und
→ Kriterien, nach denen bei Inbetriebnahme und im laufenden Betrieb die Korrektheit des Prüfgegenstands im Hinblick auf die Vorgaben der Vertrauenswürdigkeit untersucht werden kann.

Für die eingangs betrachteten hybriden, eingebetteten KI-Systeme und ihre Liefer- und Leistungsbeziehungen werden alle drei Arten von Kriterien in einem gemeinsamen Prüfverfahren erforderlich. Eine solche kriterienbasierte Prüfung und Bewertung von KI-Systemen kann im Rahmen eines Zertifizierungsprogramms anwendungsspezifisch abgedeckt werden und wird als Evaluation bezeichnet.

Abbildungen von vertikalen Risiken der Gesamtsysteme in horizontalen Prüfanforderungen für KI-Komponenten

Vom Standpunkt der Evaluation der Vertrauenswürdigkeit gegenüber Aspekten der Prüfdimensionen gibt es zwei Ausgangssituationen. Entweder wird der Evaluationsgegenstand (EVG) in einer konkreten Umgebung innerhalb eines technischen Systems beschrieben, z. B. als kamerabasiertes Objekterkennungssystem in Kraftfahrzeugen, oder – und dies ist immer häufiger anzutreffen – der EVG liegt als KI-technologischer Standard vor, der als „Rohling“ verwendet und dann gemäß konkreten Einsatzanforderungen individualisiert und angepasst wird, etwa bei einem KI-Dienst eines Finanzdienstleisters, der Transaktionsdaten als Rohdaten verarbeitet und Indikatoren für die Prognose von Geschäftsvorfällen liefert.

In beiden Fällen wird eine Prüfung vom gesamten technischen System ausgehen. In jedem Fall wird eine Risikoanalyse auf der Basis von Einsatzszenarien bzw. von Liefer- und Leistungsbeziehungen vorgenommen. Dabei können je nach Prüfdimension unterschiedliche Vorgehensweisen angewendet werden (etwa Betrachtung von Worst-Case-Szenarien gegenüber reinen Bedrohungsanalysen). Entscheidend ist, dass anhand vordefinierter Ausnahmesituationen die Gefährdungen des technischen Systems auf seine Umwelt und die Gefährdungen des Einwirkens der Umwelt auf das technische System über alle Prüfdimensionen hinweg betrachtet werden, damit eventuell auftretende (Inter-)Dependenzen zwischen einzelnen Prüfaspekten ermittelt und in den Risikoanalyseprozess eingeordnet werden. Der Begriff der Gefährdung wird im Zusammenhang mit der Risikoanalyse von KI-Systemen für Ereignisse verwendet, die zu unerwünschten Abweichungen des spezifizierten Verhaltens des technischen Gesamtsystems führen.

Als Risikoanalyse wird in diesem Zusammenhang der komplette Prozess bezeichnet, um Risiken zu beurteilen (identifizieren, einschätzen und bewerten). Risikoanalyse bezeichnet aber nach den einschlägigen ISO-Normen DIN ISO 31000:2018 [160] und ISO/IEC 27005:2018 [161] nur einen Schritt im Rahmen der Risikobeurteilung, der zur Risikobehandlung erforderlich ist. Die Risikoanalyse für KI-Systeme lehnt sich an ISO/IEC 23894:2022 [25] an und besteht aus einer Erstellung einer Gefährdungsübersicht, d. h. aus einer Liste möglicher elementarer Gefährdungen und der Ermittlung zusätzlicher Gefährdungen, die über die elementaren Gefährdungen hinausgehen und sich aus dem spezifischen Einsatzszenario ergeben und einer Risikoeinstufung, also einer Einschätzung der Risiken nach Ermittlung von Eintrittshäufigkeit und Schadenspotenzial, und der darauf basierenden Einordnung in eine Risikokategorie. Die Risikobehandlung schließt sich der Risikoanalyse an und besteht aus Vermeidungs-, Reduktions-, Transfer-, und Akzeptanzstrategien einschließlich der Definition und Prüfung von Gegenmaßnahmen. Ergriffene Maßnahmen zur Vermeidung bzw. Reduzierung der KI-basierten Risiken umfassen vertragliche Vereinbarungen mit KI-Dienstleistern, Software License Agreements und andere

Qualitätssicherungsmaßnahmen z. B. durch den Einsatz von Prüfwerkzeugen, u. v. m.

Im deutschen Sprachgebrauch hat sich der Begriff „Risikoanalyse“ für den kompletten Prozess der Risikobeurteilung und Risikobehandlung etabliert. Bei der Evaluation von KI-Systemen sind beide Schritte jedoch voneinander zu trennen.

Der KI-Risikoanalyseprozess[82] leistet den Transfer von Gefährdungen von oder an das technische System in Risiken, die in einen Anforderungskatalog an Struktur und Funktionsweise des technischen Systems münden. Die konkrete Form der Spezifikation ist je nach Prüfdimension den einschlägigen Normen und Standards (vgl. Liste oben) zu entnehmen oder kann beispielsweise nach der Systemdekomposition nach DIN SPEC 92001-1:2019 [162] erfolgen. In spezifischen Sektoren kann die Notwendigkeit bestehen, risikobasiert zusätzliche Standards, Prüfschemata und technische Kontrollwerkzeuge heranzuziehen. Nur so können im jeweiligen Kontext die dem Prüfgegenstand angemessenen Anforderungen berücksichtigt werden. Die Spezifikation enthält im Regelfall die aus den Risiken ableitbaren Mindestanforderungen, die an eine Systemkomponente gestellt werden bzw. die eine Komponente an andere Systemkomponenten stellt. Für die im technischen System enthaltenen KI-Module bzw. für die in der Supply-Chain enthaltenen KI-Komponenten ist in einem gesonderten Dokument festzuhalten, welche Anforderungen welches KI-Modul oder welche KI-Komponente bezüglich welcher Prüfdimensionen erwartet bzw. erfüllen muss.

Durch diese schrittweise Verfeinerung extrahiert der Risikoanalyseprozess schließlich auf der Ebene der KI-Module und KI-Komponenten Zielobjekte mit Mindestanforderungen, deren Einhaltung unabdingbar für die Risiken des Gesamtsystems sind. Diese Anforderungen bilden die Grundlage für die Spezifikation des EVG in der Systembeschreibung. Nach erfolgter Prüfung können die Ergebnisse analog zur schrittweisen Verfeinerung zurückverfolgt und schließlich den Risiken auf der Ebene des technischen Systems bzw. des verteilten KI-Systems zugeordnet werden. Dieser mehrschrittige Verfeinerungsprozess und seine Rückverfolgung mit den Prüfergebnissen der KI-Komponenten und KI-Module ist notwendig, um die in den internationalen Standardisierungsorganisationen geforderten Grundlagen für eine anwendungsübergreifende KI-Zertifizierung entwickeln zu können.

Konformitätsbewertungen für KI-Systeme leiten aus der KI-Risikoanalyse und den oben skizzierten Abbildungsprozessen qualitative Mindestforderungen ab (siehe Abbildung 31). Solche Mindestforderungen können sich an die Betriebsumgebung des KI-Systems richten oder sich auf den Entwicklungs- und Spezifikationsprozess des KI-Systems selbst beziehen. Beispielsweise sollte im Zusammenhang mit der Entwicklung eines verteilten KI-Systems ein Informationstransferprozess initiiert worden sein, es können Rollen und Aufgaben definiert werden und eine Strukturanalyse ermittelt die wichtigsten Informationen über das gesamte System. Daraus können sich weitere Betrachtungsschwerpunkte ergeben, z. B.:

- → bei der Betrachtung von Prozess- und Geschäftsrisiken können KI-Komponenten als Risikoursache explizit betrachtet und bewertet werden,
- → bestimmte risikorelevante Parameter können bei der Risikobewertung sofort einbezogen werden, beispielsweise ob personenbezogene Daten verwendet werden, ob externe Daten verwendet werden und ob durch die KI-Komponente ein Sach- oder Personenschaden entstehen kann,
- → spezielle Dokumente können zum Nachweis von Prüfzielen einbezogen werden, z. B. Assurance Cases als Output der Assurance-Case-Methode.

Diese Mindestforderungen und die Anleitungen für die Dokumentation der Abbildung von Risiken in Anforderungen an die KI-Module und Komponenten ist zu entwickeln.

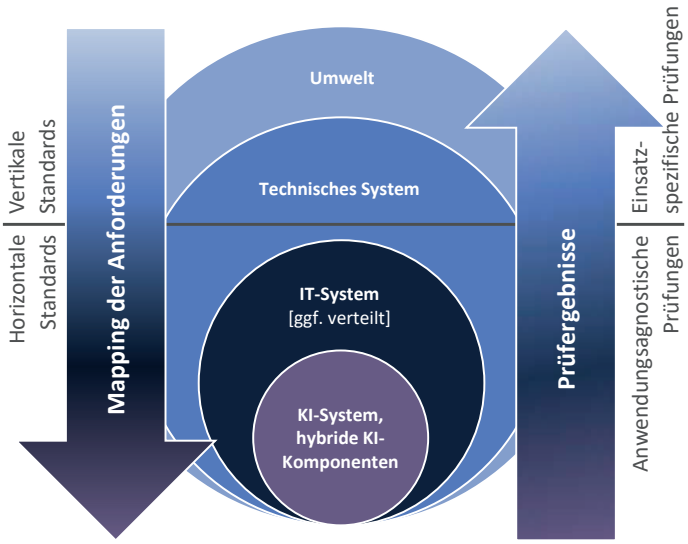


Abbildung 31: Schrittweise Verfeinerung der Prüfanforderungen und Rückbezug der Prüfergebnisse (Quelle: BSI)

82 Für einen risikobasierten Zugang zur Evaluation von KI-Systemen siehe auch [120].

Grundlagen für die KI-Prüfung

Im Folgenden werden die Grundkonzepte für KI-Prüfungen dargestellt. Hierzu muss im ersten Schritt der Prüfgegenstand beschrieben und eine Risikoanalyse durchgeführt werden. Vor dem Hintergrund der Ermittlung des Schädigungspotenzials etwa für Daten, Finanzen, Fairness und das menschliche geistige sowie physische Wohlbefinden ist die anwendungsspezifische Beschreibung von KI-Systemen unabdingbar. Als Ansätze eignen sich zur Konkretisierung von Risiken aktuelle Standardisierungsvorhaben, darunter die Beschreibungen zum Risikomanagement im Dokument ISO/IEC 42001 [27] (Allgemeine Beschreibung eines KI-Managementsystems, für eine Darstellung des KI-Managementsystems siehe auch die Studie [120]). Grundlage jeder Evaluation ist die Beschreibung des Evaluationsgegenstands (EVG), also des KI-Systems, dessen Vertrauenswürdigkeit geprüft werden soll. Ein EVG, der vertrauenswürdig sein soll, muss bestimmte Eigenschaften aufweisen. Damit ein angemessener Grad an Vertrauen in die Eigenschaften gesetzt werden kann, müssen diese selbst hinreichend genau beschrieben werden. Die Genauigkeit der Beschreibung hängt dabei davon ab, welche KI-Technologie(n) der EVG zu welchem Zweck in welcher Weise nutzt und wie tief das Vertrauen ist, das diesen Eigenschaften entgegengebracht werden soll. Diese Angaben, Darstellungen und Beschreibungen bilden einen Satz von Dokumenten, der als Prüfvorgaben bezeichnet wird. Im Regelfall benötigt jede Konformitätsprüfung eines KI-Systems i. o. S. eigene Prüfvorgaben. Die Prüfvorgaben behandeln aus Sicht des EVG die Fragestellungen:

→ Was soll überprüft werden?
→ Mit welcher Prüftiefe soll geprüft werden?

Daraus kann eine mit der Prüfung beauftragte Stelle einen konkreten Prüfplan ableiten.

VORGEHENSMODELL

Die erste Frage zielt auf den Funktionsumfang des EVG, also seine Funktionalität. Die zweite Frage zielt auf das Vertrauen, das durch eine Prüfung in diese Funktionalität entstehen kann. Die Unterscheidung zwischen der Funktionalität eines Systems und der Vertrauenswürdigkeit, die durch die Prüfqualität und die Prüftiefe gefordert wird, ist eines der grundlegenden Paradigmen für die kriterienbasierte Prüfung und Bewertung von Sicherheitseigenschaften programmierbarer IT-Systeme – und also auch für KI-Systeme. Kriterienbasierte Prüfverfahren erzeugen zunächst individuelle, auf das KI-System zugeschnittene Prüfpläne mithilfe der Prüfvorgaben. Die Funktionalitätsuntersuchung ordnet den Risiken zunächst Prüfziele zu, die dann schrittweise verfeinert werden. Den Prüfzielen werden auf der Ebene der Grobspezifikation Systemfunktionen zugeordnet. Eine Betrachtungsebene tiefer werden den Funktionen – in der Feinspezifikation – konkrete Maßnahmen zugeordnet, die die Funktionen umsetzen. Die Prüfqualität betrachtet Gesichtspunkte der Wirksamkeit der Maßnahmen und der Korrektheit der Implementierung. Grundsätzlich kann diese Vorgehensweise als rückgekoppeltes Wasserfallmodell verstanden werden. Für KI-Systeme müssen beide Aspekte – Wirksamkeit und Korrektheit – erweitert werden.

WIRKSAMKEITSANALYSE

Die im Rahmen eines Prüfschemas zu entwickelnden Wirksamkeitskriterien sollten die Lebenszyklusphasen des Systems berücksichtigen und je nach Phase unterschiedliche Prüfschwerpunkte haben, etwa …

Konstruktion:

→ Analyse der Eignung der Mechanismen,
→ Analyse des Zusammenwirkens der Mechanismen,
→ Analyse der Mechanismenstärke,
→ Analyse der Konstruktionsschwächen (bei implementierten Mechanismen).

Betrieb:

→ Analyse der Prüfprozesse im Life Cycle oder bei Wiederholungsprüfungen (bei Prüfmechanismen).

Aus den Eigenschaften der Wirksamkeitskriterien lassen sich Anforderungen an Prüfwerkzeuge ableiten. Prüfwerkzeuge sollten alle notwendigen Informationen liefern, um Ergebnisse angemessen zu interpretieren. Derartige Informationen sollten mindestens die folgenden Dimensionen abdecken:

→ Umfang und Tiefe: Welcher konkrete Teil des KI-Systems wird geprüft? Was sind Input und Output dieses Teils? Welche und wie viele Daten werden für die Prüfung des Systems verwendet?
→ Funktionszuordnung: Welche Funktionen werden mit dem Werkzeug unterstützt? Was ist ein gewünschtes Ergebnis der Prüfung? Was ist ein unerwünschtes Ergebnis der Prüfung?
→ Funktionsweise des Prüfwerkzeugs: Die zur Prüfung des KI-Systems verwendete technische Methode soll beschrieben werden. Auch Limitationen der angewandten Prüfmethode sollten explizit dargestellt werden, ebenso wie Informationen zur Stabilität und Reproduzierbarkeit der Prüfergebnisse.

KORREKTHEIT
Für Korrektheitskriterien bietet sich die Vorgehensweise der stufenweisen Definition von Prüfkriterien an, wobei jede Stufe auf der nächsttieferen Stufe aufbaut. Solche Evaluation Assurance Levels (EALs) werden in Kapitel 4.1.2.2 dargestellt. Mithilfe der EALs wird die Prüfqualität und auch die Prüftiefe schrittweise erhöht. Im Rahmen der Grundlagen für KI-Prüfungen wird in jeder Stufe die Unterscheidung in Konstruktions- und Betriebsphasen im KI-Lebenszyklus gemacht werden müssen. Für jede einzelne Evaluationsstufe werden die Evaluationskriterien in verschiedenen Phasen dann weiter zu untergliedern sein. Bisher scheinen folgende Phasen für die Korrektheit relevant zu sein:

Konstruktion, Entwicklungsprozess:
1. Anforderungen an die Prüfvorgaben
2. Architekturentwurf
3. Feinspezifikation
4. Implementierung

Konstruktion, Entwicklungsumgebung:
1. Vorgehensweise
2. Kontrollprozesse
3. Vertrauenswürdigkeit beim Entwickler

Betrieb:
1. Vorgaben für den Betrieb
2. Auslieferung und Konfiguration
3. Anlauf und Betrieb
4. Betriebsdokumentation
5. Betriebsbegleitende Prüfung
6. Nachweissicherung
7. Betriebsende

Jede Phase wird die Prüfmaßnahmen und die bereitzustellenden Dokumente bei Beginn der Prüfung definieren und die Mindestanforderungen an die Prüfergebnisse vorgeben.

Qualitätsinfrastruktur
Dieses Kapitel argumentiert auf der Grundlage existierender Normen und Standards und auf Basis aktueller internationaler KI-Standardisierungsaktivitäten für ein universelles Zertifizierungsverfahren für KI-Systeme. Es wurde gezeigt, wie ein solches Verfahren konzipiert werden kann, damit es einerseits für die vertikale KI-Standardisierung verwendet werden kann und andererseits an bestehende Prüf- und Zertifizierungsverfahren der Informationstechnik angebunden werden könnte. Es wurde dargelegt, dass ein solches Verfahren für die Umsetzung der KI-Regulierung in Europa richtungsweisende Umsetzungsimpulse geben und gleichzeitig internationale Marktdurchdringung erreichen kann. Das Plädoyer geht eindeutig in Richtung eines auf der Grundlage bestehender Normen und Standards zu entwickelnden KI-Zertifizierungsprogramms innerhalb einer Qualitätsinfrastruktur, die folgende Rahmenbedingungen erfüllt:

→ Das Zertifizierungsprogramm wird international in zwei Standards – „Trustworthy Artificial Intelligence Systems Evaluation Criteria“ und „Trustworthy Artificial Intelligence Systems Evaluation Methodology“ – verankert.
→ Das Zertifizierungsprogramm lässt sich an die bestehenden IT-Prüfinfrastrukturen anschließen.
→ KI-Prüfer*innen bei Konformitätsbewertungsstellen werden in speziellen Ausbildungs- und Fortbildungsprogrammen (Lizenzierung, Personenzertifizierung) im Rahmen der Aufgaben aus der Normungsroadmap KI auf Basis von international entwickelten Qualitätsanforderungen gefördert.
→ Man prüft in der Konformitätsbewertung gegen geltende rechtliche Anforderungen und technische Spezifikationen – normative und ethische Aspekte werden ausgeklammert.
→ Die Schnittstellen zu KI-Managementsystemen – insbesondere zum AIMS – werden klar definiert.
→ Zertifizierung und Zulassung von KI-Prüfwerkzeugen werden als Schwerpunkt fest in den o. g. Kriterienwerken verankert.

4.3.2.3 Bestehende Ansätze und Ergebnisse

In diesem Kapitel werden kurz Projekte und Initiativen dargestellt, die im Rahmen der Prüfung und Zertifizierung von KI-Systemen nationale und internationale Bedeutung haben.

ZERTIFIZIERTE KI
Im Leuchtturmprojekt „ZERTIFIZIERTE KI“ der Kompetenzplattform KI.NRW entwickelt ein Konsortium aus Fraunhofer IAIS, BSI, DIN und weiteren Forschungspartner*innen Prüfkriterien, -methoden und -werkzeuge für KI-Systeme, um die Qualität von KI-Anwendungen durch unabhängige Prüfende beurteilbar zu machen. Hierbei werden industrielle Bedarfe durch die aktive Einbindung von zahlreichen assoziierten Unternehmen und Organisationen berücksichtigt, die unterschiedliche Branchen wie etwa Telekommunikation, Banken, Versicherungen, Chemie und Handel repräsentieren. Die Ergebnisse werden in die Standardisierung überführt.

Ein erstes Projektergebnis ist der „Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz“ [120], welcher Entwickler*innen eine Richtschnur an die Hand gibt, um neue KI-Anwendungen systematisch vertrauenswürdig zu gestalten. Zum anderen leitet er Prüfer*innen dazu an, KI-Anwendungen strukturiert auf Vertrauenswürdigkeit zu untersuchen. Hier verfolgt der Leitfaden ein vierstufiges Vorgehen:

1. Eine umfassende Risikoanalyse entlang der Dimensionen Fairness, Autonomie und Kontrolle, Transparenz, Verlässlichkeit, Sicherheit und Datenschutz.
2. Die Festlegung objektiver, möglichst messbarer Zielvorgaben, um die Mitigation der unter 1 identifizierten Risiken nachweisbar zu machen.
3. Eine systematische Auflistung von Maßnahmen entlang des Lebenszyklus einer KI-Anwendung, um die in 2 gesetzten Zielvorgaben zu erreichen.
4. Die Erstellung einer stringenten Argumentation, dass die unter 2 formulierten Zielvorgaben erreicht wurden („Absicherungsargumentation für die Vertrauenswürdigkeit“), wobei auch KI-spezifische Trade-offs, z. B. Sicherheit vs. Transparenz, berücksichtigt werden.

Weitere Informationen sind über die Projekthomepage www.zertifizierte-ki.de erhältlich.

Ein Prüfstandard für cloudbasierte KI[83]

Der breite Marktzugang für geprüfte KI in Clouds kann dadurch sichergestellt werden, dass auf Ebene des KI-basierten Clouddienstes (relativ kostengünstige) Konformitätsprüfungen durchgeführt werden, die die Wirksamkeit von Maßnahmen gegen Gefährdungen oder gar Risiken an die KI innerhalb des Clouddienstes zum Gegenstand haben. Diese Prüfungen, die regelmäßig im Lebenszyklus der KI-Anwendung wiederholt werden, stützen sich auf drei Säulen ab:

1. Die grundsätzliche Prüfung des gesamten unterliegenden Cloudsystems von der Infrastruktur (IAAS) über die Plattform (PAAS) bis zu den Schnittstellen zum KI-Dienst (SAAS). Für solche Prüfungen gibt es bereits Kriterienwerke, etwa den C5-Kriterienkatalog des BSI. Dieser stützt sich wiederum dort, wo technische Prüfverfahren nicht mehr anwendbar sind, auf personelle, organisatorische, institutionelle oder räumliche Randbedingungen des Providers ab, wenn die Restrisiken über die technische Prüfung hinaus nicht tragbar sind.
2. Die tiefgehende technische Prüfung des KI-Frameworks des Providers, die ja für jeden Kunden zunächst gleich angeboten wird. Dabei wird für die einzelnen KI-Technologien und -Verfahren von Prüfungen verschiedener Art, Qualität und Tiefe bis hin zu Zertifizierungsvorgängen ausgegangen. Die Prüfschemata dafür müssen im Projekt entwickelt und evaluiert werden. Sie können über verschiedene Prüfschemata hinaus so gestaltet werden, dass die inhaltlichen Anforderungen als Kriterienkatalog aus einem erweiterbaren Satz von Bausteinen zusammengesetzt werden. Für die Prüfqualität und Prüftiefe wird eine allgemeine Prüfmethodologie entwickelt, die zusammen mit dem Kriterienkatalog in die Standardisierung einfließt. Aus der allgemeinen Prüfmethodik können dann für verschiedene Prüfschemata (Konformitätsprüfung vs. Zertifizierung) spezielle Prüfmethodiken abgeleitet werden.
3. Die genannten Elemente müssen in einen übergeordneten Standard überführt werden, aus dem für jeden der Cloudservice-Provider ein entsprechendes individualisiertes Prüfschema ableitbar ist, das im Sinne der Instanz eines risikobasierten Managementstandards alle relevanten Risiken mit personellen, organisatorischen, technischen und räumlichen Dimensionen abdeckt. Die Mindeststandards (z. B. Grundschutz) des BSI können sich über die technischen Qualitätsmerkmale hinaus bei der Instanziierung als sinnvolle Bausteine erweisen.

Es gilt also für den Gesamtprojekterfolg und seine Umsetzung, aus einem Standard für vertrauenswürdige KI in Cloudsystemen alle Einzelelemente abzuleiten und im konkreten Fall aus unterliegenden Schemata die Bausteine zusammenzufügen, die für die Vertrauenswürdigkeit im konkreten Anwendungsfall relevant sind.

Das Leuchtturmprojekt der NRM KI wird vom BSI geleitet und durchgeführt und mit internationalen Standardisierungsprojekten begleitet.

KI-Standards für medizinische Diagnosesysteme[84]

Das Projekt hat zum Ziel, Prüfkriterien und Prüfmethoden für den Einsatz von KI in medizinischen Diagnose- und Prognosesystemen zu entwickeln und sie in relevante Normen so einzubetten, dass Prüfstandards für KI in der Medizintechnik etabliert werden können.

83 Siehe Kapitel 6.6.

84 Siehe Kapitel 6.6.

Dazu sind folgende Meilensteine zu erfüllen:

→ Entwicklung von erweiterbaren Prüfkriterien für relevante KI-Technologien in der Medizintechnik auf der Basis existierender Normen und etablierter Standards,
→ Evaluation dieser Prüfgrundlagen in Pilotprojekten mit eingesetzten KI-Lösungen im Zuge eines kontinuierlichen Verbesserungsprozesses,
→ Ableitung und Entwicklung von Referenzarchitekturen und Prüfprofilen für die unten betrachteten Use Cases im Einsatzbereich und für verwendete KI-Technologien mit dem Ziel der Reduzierung von Prüfaufwänden,
→ Standardisierung und Normung der entwickelten Prüfgrundlagen und Kriterienwerke und Einordnung auf Basis bestehender Normen und schließlich
→ Etablierung der KI-Prüfstandards auf internationaler Ebene.

Das Leuchtturmprojekt der NRM KI wird vom BSI geleitet und durchgeführt.

ExamAI: Assurance Cases und Acceptance-Test-Driven Development

Im vom Bundesministerium für Arbeit und Soziales (BMAS) geförderten Projekt „ExamAI – Testing und Auditing von KI“ wurde eine Kombination aus Assurance Cases und Acceptance-Test-Driven Development (ATDD) vorgeschlagen, um das Auditieren und langfristig auch Zertifizieren von extrafunktionalen Anforderungen zu unterstützen. Bei Assurance Cases handelt es sich um eine strukturierte Argumentation, die erläutert, warum ein System als ausreichend gut in Bezug auf eine festgelegte Eigenschaft eingeschätzt wurde, um eingesetzt zu werden. Ein Assurance Case startet mit einer Qualitätsbehauptung (Claim) wie z. B., ein System sei fair. Diese Behauptung wird nun basierend auf Argumenten (reasoning) in Teilbehauptungen unterteilt. Jedes Argument kann zusätzlich durch Kontextinformationen (context) und Annahmen (assumptions) ergänzt werden. Am Ende stehen für jede Behauptung Beweise (evidences), die belegen, dass die jeweilige Behauptung zutrifft. Das Konzept stammt ursprünglich aus der Philosophie und ist aktuell ein gängiges Framework im Safety Engineering, um zu argumentieren, auf Basis welcher Argumente ein System als ausreichend sicher angesehen wird. Die Erweiterung um ATDD sieht vor, dass der Assurance Case vor Entwicklungsbeginn erstellt wird. Es handelt sich damit um ein Konzept der Test-First-Philosophie. Eine möglichst diverse Gruppe aus Stakeholdern (Projektverantwortliche, Entwickler*innen, Nutzer*innen, Betroffene, Jurist*innen, ...) trifft sich dazu, um theoretische Szenarien zu entwickeln, in denen das System entgegen der sicherzustellenden Eigenschaft handeln könnte. Darauf basierend werden mögliche Gegenmaßnahmen entwickelt, wie z. B. Tests. Am Ende wird der Assurance Case erstellt, der argumentiert, warum die Tests als ausreichend angesehen werden.

ENISA

Die Agentur der EU für Cybersicherheit, ENISA, hat die Aufgabe, zu einem hohen gemeinsamen Maß an Cybersicherheit in ganz Europa beizutragen [118]. Die für die ENISA geltende Verordnung ist die Verordnung (EU) 2019/881 [163] des Europäischen Parlaments und des Rates vom 17. April 2019 über die ENISA und über die Zertifizierung der Cybersicherheit von Informations- und Kommunikationstechnik und zur Aufhebung der Verordnung (EU) Nr. 526/2013 [164] (Rechtsakt zur Cybersicherheit). Zum Thema Künstliche Intelligenz hat die ENISA zwei Publikationen veröffentlicht:

– ENISA Report – Artificial Intelligence Cybersecurity Challenges mit drei Themenbereichen: AI LIFECYCLE; AI Assets; AI THREATS.

Inhaltlich findet sich eine Übersicht des KI-Cybersicherheitsökosystems und seiner Bedrohungslandschaft unter Berücksichtigung des AI Lifecycle. In fünf Kapiteln werden ein generisches Referenzmodell, Details zum KI-Ecosystem, eine Bedrohungstaxonomie mit Verbindung zwischen relevanten Bestandteilen und zugehörigen Bedrohungen vorgestellt sowie die Herausforderungen im Zusammenhang mit der Cybersicherheit für KI.

– ENISA Report – SECURING MACHINE LEARNING ALGORITHMS; December 2021 [119]

Dieser Report enthält eine Taxonomie von ML-Algorithmen, die Identifizierung relevanter Bedrohungen und Schwachstellen sowie eine Liste von Sicherheitskontrollen.

Aufbauend auf der KI-Bedrohungslandschaftskartierung der ENISA konzentriert sich diese Studie auf Cybersicherheitsbedrohungen, die für ML-Algorithmen spezifisch sind. Darüber hinaus werden Schwachstellen im Zusammenhang mit den oben genannten Bedrohungen und vor allem Sicherheitskontrollen und Minderungsmaßnahmen vorgeschlagen.

Die angenommene Beschreibung von KI ist eine bewusste Vereinfachung des Stands der Technik in Bezug auf diese riesige und komplexe Disziplin, mit der Absicht, sie nicht genau oder umfassend zu definieren, sondern die spezifische Technik des Maschinellen Lernens pragmatisch zu kontextualisieren.

Als Ergebnis wurde festgestellt, dass es keine eindeutige Strategie für die Anwendung eines bestimmten Satzes von Sicherheitskontrollen zum Schutz von maschinellen Lernalgorithmen gibt. Die allgemeine Cybersicherheitslage von Organisationen, die maschinelle Lernalgorithmen verwenden, kann verbessert werden, indem die für diese Algorithmen entwickelten Kontrollen sorgfältig ausgewählt werden.

Aktuelle Aktivität: KI-Nachwuchsforschergruppe BAuA
In der zwischen dem BMAS und der BAuA geschlossenen Verwaltungsvereinbarung wird eine Forschungsstrategie zum Thema „KI in einer sicheren und gesunden Arbeitswelt" beschriebenen. Zur Umsetzung der Strategie hat die BAuA eine Nachwuchsforschergruppe für die nächsten fünf Jahre eingerichtet. Ziel der Gruppe ist es, im Rahmen von Promotionsvorhaben, die in Zusammenarbeit mit einschlägigen Universitätsinstituten durchgeführt werden, Antworten auf anwendungsorientierte Fragen zu KI in der Arbeitswelt zu geben. Angelehnt an die zwei Rechtsbereiche, auf denen die Vorschriften und Regeln zur Gewährleistung der Sicherheit und Gesundheit bei der Arbeit in Deutschland beruhen, werden zwei Themengebiete unterschieden: der betriebliche Arbeitsschutz und die Produktsicherheit. Die Herausforderungen, die in dem jeweiligen Themengebiet durch die Nutzung von KI entstehen, werden von zwei Teams fachlich vertieft, einem Team in Dortmund (betriebliche Gestaltungsmaßnahmen) und einem Team in Dresden (Produktsicherheit).

Aktuelle Aktivität: KI-LOK – Ein Verbundprojekt über Prüfverfahren für KI-basierte Komponenten im Eisenbahnbetrieb
Der Entwurf und Betrieb innovativer Fahrzeuge im schienengebundenen Verkehr fordert verstärkt den Einsatz KI-basierter lernender Systeme, um die Qualität des Verkehrsangebots zu verbessern, die Ressourceneffizienz und damit die Nachhaltigkeit der Züge zu steigern sowie neue Funktionalitäten bereitzustellen. Eine der größten Herausforderungen ist dabei die Entwicklung entsprechender Verifikations- und Validierungsverfahren, die in ihrer Gesamtheit den Zielen der datenbasierten Mobilität wie auch den Qualitäts- und Sicherheitsansprüchen des Bahnverkehrs gerecht werden müssen und für den Nachweis der funktionalen Sicherheit von KI-Systemen geeignet sind. Ziel des Projekts ist es, Testverfahren und Methoden zur Absicherung und Zertifizierung von KI-gestützten Technologien für sicherheitskritische Anwendungen in der Bahntechnik zu entwickeln. Die zu entwickelnden Techniken und Werkzeuge werden anhand von praktischen Anwendungsbeispielen entwickelt, um praxisgerecht zu sein. Auf Grundlage zweier Fallstudien – „Objekterkennung im vorausliegenden Lichtraumprofil" und „Sichere Eigenlokation als Teil des Fahrzeug-Odometriesystems" – soll daher die Trainings- und Teststrategie für KI-Systeme entwickelt und für den industriellen Einsatz nutzbar gemacht werden. Der Wirkraum des Projekts KI-LOK wird durch die drei Eckpunkte Zulassungsprozesse, Risiko- und Gefährdungsanalyse sowie Analysemethoden für KI definiert. Die Ergebnisse des Projekts bilden die Grundlage für eine werkzeuggestützte Methode zur Validierung und Verifizierung von KI-basierten Komponenten im industriellen Umfeld und definieren darüber hinaus einen systematischen Rahmen zur Definition von Zulassungsprozessen für KI-basierte Anwendungen im Eisenbahnbetrieb. Das Projekt KI-LOK wird vom Bundesministerium für Wirtschaft und Klimaschutz (BMWK) im Rahmen der Förderrichtlinie „Neue Fahrzeug- und Systemtechnologien" gefördert und finanziert [165].

Aktuelle Aktivität: Industrial Grade Machine Learning for Enterprises (IML4E)
In Analogie zur klassischen Software muss KI-basierte Software entsprechend den Anforderungen des Endbenutzers implementiert und validiert werden und die etablierten Qualitätsmerkmale klassischer Software sowie eine Reihe neuer Qualitätsmerkmale (z. B. Interpretierbarkeit, intelligentes Verhalten, Diskriminierungsfreiheit usw.) erfüllen. Ihr Einsatz muss technologisch, sozial und ethisch akzeptabel und sicher sein. All dies muss sorgfältig geplant, realisiert, validiert und über den gesamten Softwarelebenszyklus hinweg gewartet werden. Vor diesem Hintergrund bringt das IML4E-Projekt Unternehmen aus den Hauptsektoren der deutschen und europäischen Softwareindustrie zusammen, um ein europäisches Rahmenwerk für die Entwicklung, den Betrieb und die Wartung von KI-basierter Software zu entwickeln und dadurch die Entwicklung von intelligenten Diensten und intelligenter Software im industriellen Maßstab zu gewährleisten. Das Projekt konzentriert sich dabei auf die Bereitstellung von industrietauglichen Techniken, Methoden und Werkzeugen, die aktuell nicht durch Open-Source-Lösungen frei zugänglich sind, und adressiert etablierte Softwareentwicklungsprinzipien wie Wiederverwendung, Automatisierung und die enge Integration von Entwicklung und Betrieb über den gesamten Softwarelebenszyklus, sodass deutsche Unternehmen in die Lage versetzt werden, KI-basierte Software in ihre Entwicklungsprozesse und Produkte zu integrieren. Das Projekt IML4E wird vom Bundesministerium für Bildung und Forschung (BMBF) im Rahmen der europäischen ITE-Initiative gefördert [168].

4.3.3 Normungs- und Standardisierungsbedarfe

Bedarf 03-01: Spezifikation von formalen Anforderungen an „explainable“ AI („XAI“)-Methoden
Formulierung konkreter operationalisierbarer/prüfbarer Anforderungen an XAI-Methoden.

Welche formalen Aussagen sollen anhand der Ergebnisse einer XAI-Methode möglich sein?
→ Die Trainingsdaten betreffend?
→ Das Testdatum betreffend?
→ Das Modell betreffend?
→ Den Zusammenhang zwischen Ein- und Ausgabedaten (Prädiktionen) betreffend?
→ Den Zusammenhang zwischen Modell, Ein- und Ausgabedaten betreffend?

Welche praktischen Konsequenzen sollen sich sicher aus diesen Aussagen ableiten lassen? Welcher Mehrwert an „Verlässlichkeit“ soll wirklich geschaffen werden, und wie kann er nachgewiesen werden?

Eine sektorübergreifende und auch im Entwurf AI Act verankerte Forderung ist die nach „Erklärbarkeit“, „Interpretierbarkeit“ etc. Es klafft aber eine große Lücke zwischen den gesetzlichen/regulatorischen Anforderungen und der konkreten Umsetzung von XAI. Die in der Literatur publizierten XAI-Methoden schließen diese Lücke noch nicht, da die Anforderungen an die Methoden von den Autor*innen meist nicht konkret genug spezifiziert werden. Dementsprechend ist die Validierung/Verifikation dieser Methoden tendenziell oft eher qualitativ, subjektiv und zirkulär.

Formale Kriterien sind notwendig, um zu spezifizieren, welche Aussagen / praktischen Konsequenzen auf Basis des Ergebnisses einer gegebenen XAI-Methode korrekt und zulässig sind. Die Einhaltung dieser Kriterien muss formal oder empirisch verifiziert werden. Nur so können Fehlinterpretationen vermieden werden.

Bedarf 03-02: Operationalisierung der „Erklärgüte“ von XAI-Methoden
Entwicklung von ground-truth-Referenzdatensätzen. Die Antworten auf Fragen, die von XAI-Methoden geliefert werden sollen, sind für diese Daten per Konstruktion bekannt und können daher mit dem Ergebnis von XAI-Methoden abgeglichen werden. Die Daten können durch mathematische Bildungsvorschriften, physikalische Simulation oder Manipulation realer Daten generiert werden.

Entwicklung geeigneter Metriken für die „Erklärgüte“ von XAI-Methode auf ground-truth-Referenzdaten (z. B. Precision/Recall, andere Metriken aus der Signaldetektionstheorie).

Ohne eine ausreichende Verifikation von XAI-Methoden selbst bleibt unklar, welchen Nutzen sie für die Qualitätssicherung von ML-Systemen haben können.

Bedarf 03-03: Entwicklung eines Standards mit Guidance-Dokumenten für die Abbildung von Risiken eines Systems in die Funktionalität von KI-Komponenten
KI-Systeme sind:
→ möglicherweise hybrid,
→ möglicherweise Komponenten eines technischen Systems,
→ möglicherweise Teil einer verteilten Architektur auf verschiedenen Plattformen und in verschiedenen Infrastrukturen.

Die Risikoanalyse für das KI-System erfolgt mit Blick auf das gesamte technische System. Daraus müssen Safety-, Security-, …-Anforderungen an die Teile und Komponenten des KI-Systems abgeleitet werden. Dies wird unter Berücksichtigung des Einsatzzweckes und vorhandener Prüfvorschriften und Rahmenbedingungen erfolgen müssen (ISO-26262-Reihe [455], Maschinenrichtlinie etc.). Man bildet also Risiken ganz oder teilweise auf Prüfanforderungen an das ganze KI-System oder auf Teile davon ab. Diese Abbildung bildet den Anker für die Einbettung der Prüfergebnisse in bestehende Prüfverfahren und ihre Bewertung.

→ Contributions CEN/CLC JTC 21 & ISO SC 42 WG 3 „TAISEC“ & „TAISEM“

Einbettung von KI-Prüfungen in die bestehende Prüfinfrastruktur.

Bedarf 03-04: Entwicklung von Funktionalitätsklassen für KI-Technologien
An jedes KI-System oder -Produkt werden eigene Anforderungen bezüglich der Einhaltung von Vertrauenswürdigkeit gestellt. Um diese Anforderungen zu erfüllen, stehen technische Funktionen zur Verfügung, die das KI-System entweder selbst enthält oder die seine Umgebung zur Verfügung stellt, beispielsweise für die Erkennung und Abwehr von Adversarials, die Protokollauswertung oder die Fehlererkennung und -überbrückung. Gefordert wird ein angemessenes Vertrauen in diese Funktionen, unabhängig davon, ob es sich um das Vertrauen in die Korrektheit der speziellen Funktionen

(sowohl vom Gesichtspunkt der Entwicklung als auch von dem des Betriebs) oder um das Vertrauen in die Wirksamkeit dieser Funktionen handelt. Um beides überprüfen zu können, muss eine Funktion immer in Zusammenhang mit der KI-Technologie gebracht werden, die das KI-System enthält. Somit ergeben sich für verschiedene KI-Technologien verschiedene Möglichkeiten z. B. der Fehlererkennung. Diese Funktionalitäten müssen klassifiziert werden, damit die Funktionen den Anforderungen leicht zugeordnet werden können. Es wird also ein Baukasten von relevanten Funktionalitätsklassen benötigt.

→ Contributions CEN/CLC JTC 21 & ISO SC 42 WG 3 „TAISEC“ & „TAISEM“

Bedarf 03-05: Entwicklung von Werkzeugkriterien für die Prüfung von KI-Systemen
Werkzeuge zur Messung von Eigenschaften eines KI-Systems, z. B. der Performance, spielen die entscheidende Rolle für die Prüfung des Systems. Die Aussagekraft der Ergebnisse solcher Messungen bestimmt die Aussagekraft des gesamten Prüfvorgangs. Für die Prüfung und Zertifizierung solcher Werkzeuge werden die entsprechenden Prüfkriterien und Prüfmethoden benötigt. Die entstehenden Prüfverfahren sind Teil des zu entwickelnden KI-Zertifizierungsprogramms in o. g. Standardisierungsbeiträgen.

→ Contributions CEN/CLC JTC 21 & ISO SC 42 WG 3 „TAISEC“ & „TAISEM“

Bedarf 03-06: Entwicklung von ineinandergreifenden Standards für KI-Systeme und notwendiger Konformitätsbewertungsverfahren
Damit Konformitätsbewertungsverfahren für KI-Systeme nutzbar sind, ist es wichtig, dass die geltenden Normen der Reihe DIN EN ISO/IEC 17000:2020 [147] (Level 3 Normen) beachtet werden. Für spezifische Anforderungen an bestimmte Evaluierungsaufgaben innerhalb der definierten Konformitätsbewertungsaktivität auf Level 3 sind nach sektoralen oder technischen Anforderungen auf Level 4 differenzierte Normen für KI-Systeme zu entwickeln.

Daneben besteht Normungsbedarf im Bereich der Grundlagen, insbesondere bezüglich Kalibrierung und Eignungsprüfungsanbieter (Ringversuche). Auch hier sind auf Level 4 Normen zu entwickeln, welche die technischen Besonderheiten und Risiken der KI-Systeme berücksichtigen.

Besonders wichtig ist es, die Normungsvorhaben an den Gegenstand der Konformitätsbewertung (KI-System / Organisation i. S. v. Herstellenden oder Inverkehrbringer) (Level 5) von den Normungsvorhaben, die sich auf Konformitätsbewertung beziehen (Level 4 und 3), zu trennen.

Nur so ist es möglich, die einzelnen Rollen und Verantwortlichkeiten im Hinblick auf Herstellende, Inverkehrbringer, Nutzer*innen und Konformitätsbewertungsstellen richtig zuzuschreiben.

Nur durch klare Vorgaben an Qualifikationen und klare Anforderungen können Prüfverfahren, die sich durch Ringversuche in ihrer Bewertungsqualität messen müssen, entwickelt werden. Zuerst sind die Anforderungen an den Gegenstand zu kennen, bevor festgelegt werden kann, wie diese überprüft werden können.

→ Contributions CEN/CENELEC JTC 21 WG 2 „Conformity Assessment“

Bedarf 03-07: Entwicklung von Qualifikationskriterien für Prüfer und Zertifizierter zu Cybersecurity und Privacy für KI
Entwicklung eines Standards mit Kriterien für die Qualifikation von Prüfern, Auditoren und Zertifizierern für Cybersecurity und Privacy bei KI unter Berücksichtigung bestehender Standards aus der DIN EN ISO/IEC 27000er-Folge [131].

Aktuell bestehen etablierte Prüf- und Zertifizierungsverfahren an die Qualifikation von Expert*innen zur Prüfung und Zertifizierung von Cybersecurity und Privacy, aber noch nicht für KI. Diese sind ergänzend notwendig.

Bedarf 03-08: Vernetzung aller Akteur*innen
Bei der Erarbeitung von Normen gilt es, alle beteiligten und interessierten Kreise einzubeziehen, insbesondere Behörden gemäß Art. 5 und Art. 7 Verordnung (EU) Nr. 1025/2012 [169], und die Vernetzung von Expert*innen aus allen benötigten Bereichen sicherzustellen.

Normungsvorhaben, die Methoden, Verfahren oder Prozesse vorsehen, die z. B. eine Konformitätsbewertung (z. B. als Prüfung) von Anforderungen vorsehen, müssen mit einem breiten Expert*innenfeld aus dem Bereich der Konformitätsbewertungsstellen und Akkreditierungsstellen besetzt werden.

Dabei gilt es, auch innerhalb der Normungsarbeit ein gemeinsames Verständnis des notwendigen Zusammenwirkens der verschiedenen Ebenen (Metrologie, Konformitätsbewertung, Akkreditierung, Herstellung, Inverkehrbringung und Anwendung) zu etablieren, um geeignete und ineinandergreifende Normen für den Gegenstand (z. B. KI-System) sowie für die Konformitätsbewertung (z. B. im Rahmen einer Prüfung) zu entwickeln.

In der Normung gibt es kein übergreifendes Verständnis, wie in der Praxis Normanforderungen an den Gegenstand und Normanforderungen an Prüfprozesse ineinandergreifen. Dies sollte zukünftig besser versucht werden, am Anfang eines Normungsvorhabens hervorzustellen, um besser abgestimmte Normungsvorhaben zu haben. Je besser das gegenseitige Verständnis ist, desto leichter wird die praktische Umsetzung.

In Kapitel 4.3 wird deutlich, dass das Verständnis von „Prüfung und Zertifizierung“ je nach Anwendungsfeld und Berufskontext unterschiedlich aufgefasst wird. Dabei existiert ein gesetzlich geregeltes System in der EU, welches die Qualität von Produkten, Prozessen, Services und Dienstleistungen absichert: die Qualitätsinfrastruktur.

Bedarf 03-09: Definition von Kontrollpunkten
Anhand des KI-Lebenszyklus sind einzelne Prüfpunkte, an denen eine Konformitätsbewertung (Level 4 und 3) stattfinden muss, mit einem Minimalset an Evaluationstätigkeiten zu definieren, um die Konformität mit den rechtlichen Anforderungen, die in Gesetzesvorhaben wie dem europäischen AI Act oder dem kanadischen Artificial Intelligence and Data Act [170] definiert werden, bewerten und bestätigen zu können.

Dabei ist eine klare Rollendefinition auf der Ebene der KI-Entwickler*innen/Herstellenden/Inverkehrbringer als auch auf der Ebene der Konformitätsbewertungsstellen und Akkreditierungsstellen notwendig.

Nach einer klareren Rollenstruktur gilt es dann, herauszuarbeiten, welche Rolle (aus Level 5 oder Level 3) an welchem Punkt im KI-Lebenszyklus in die Entwicklung, Evaluierung, den Einsatz und die Stilllegung des KI-Systems integriert werden muss, um die gesetzlichen Anforderungen zu erfüllen.

Bessere Verzahnung von Unternehmen, die KI-Systeme entwickeln und/oder in Verkehr bringen, mit den Konformitätsbewertungsstellen (erster, zweiter und dritter Seite).

Die Arbeitsgruppe Prüfung und Zertifizierung hat die identifizierten Bedarfe nach der Dringlichkeit ihrer Umsetzung bewertet. Abbildung 32 zeigt die Dringlichkeit der Umsetzung, kategorisiert nach den Zielgruppen Normung, Forschung und Politik

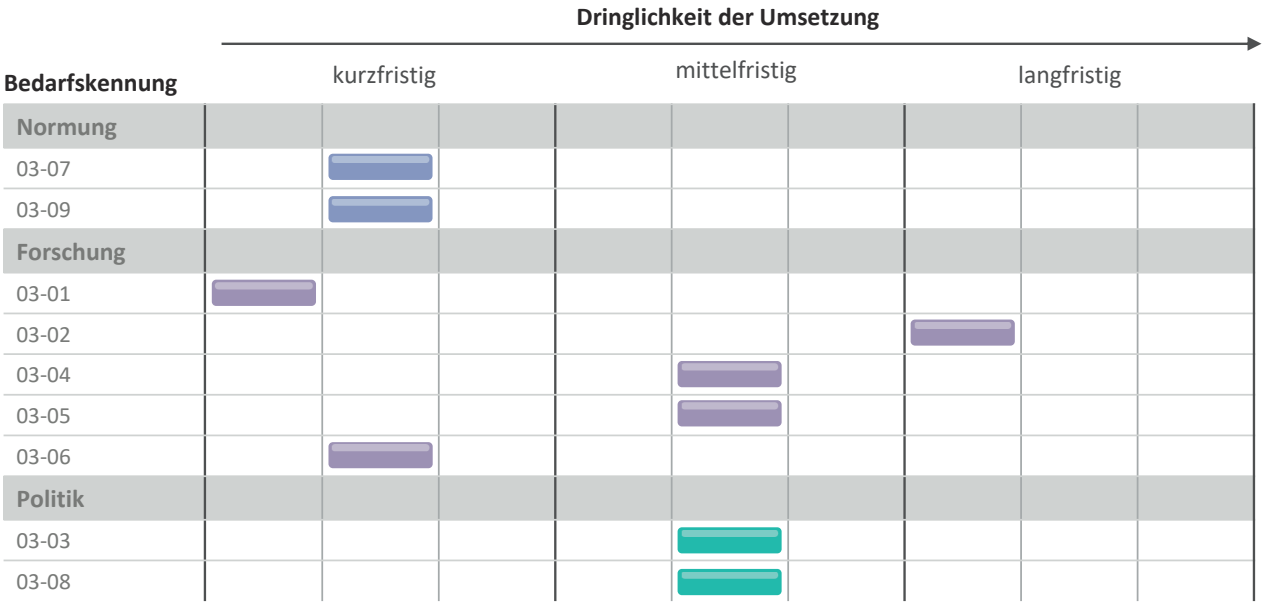
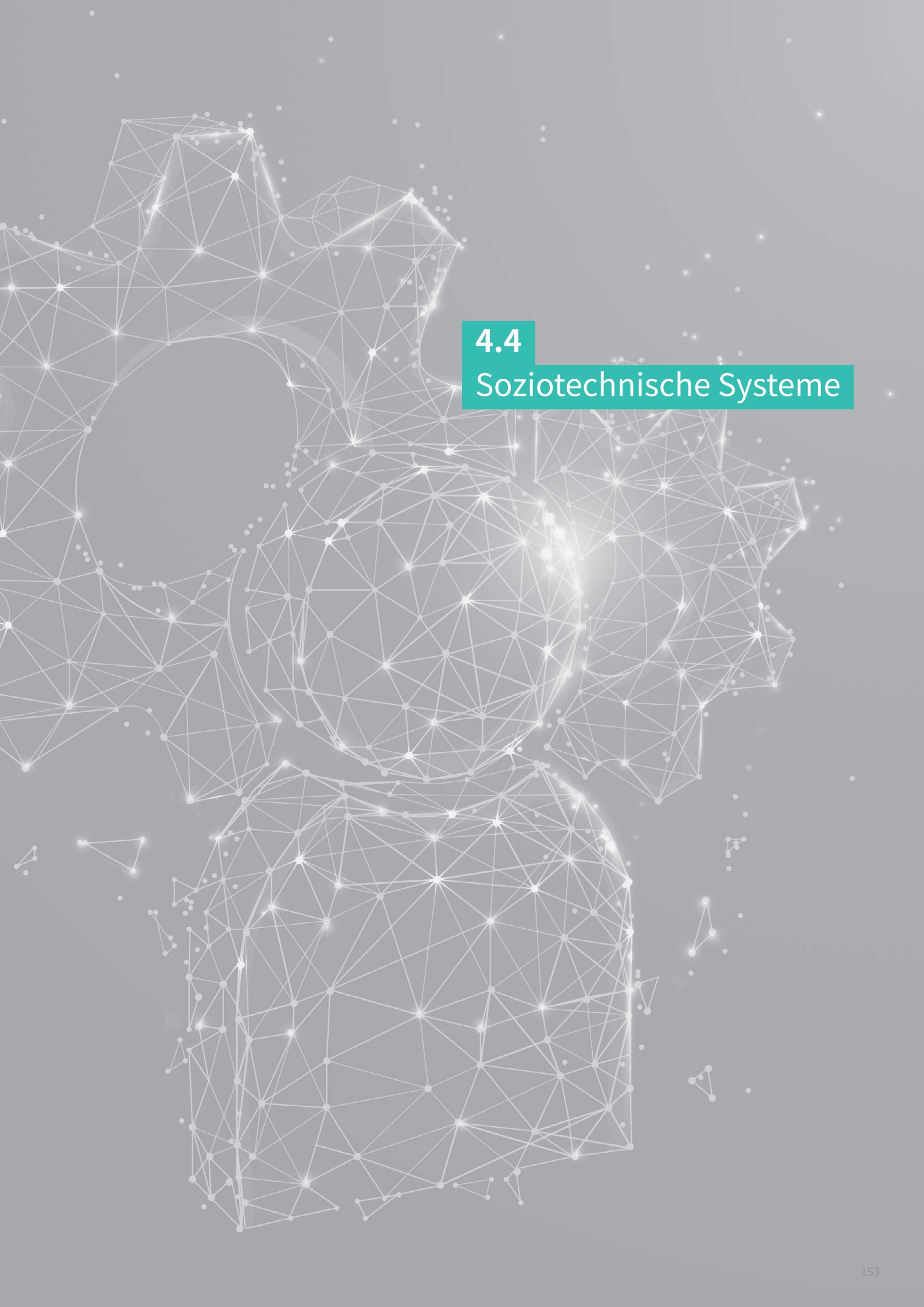


Abbildung 32: Priorisierung der Bedarfe aus Schwerpunkt Prüfung und Zertifizierung (Quelle: Arbeitsgruppe Prüfung und Zertifizierung)



4.4 Soziotechnische Systeme

In der vorliegenden zweiten Ausgabe der Normungsroadmap KI wird das Themenfeld Soziotechnische Systeme erstmals in einem eigenständigen Kapitel betrachtet. Wichtige Vorarbeiten finden sich bereits in der ersten Ausgabe im Kapitel Ethik/Responsible AI mit dem Anspruch, wertorientierte Anforderungen an IT-Systeme zu stellen und Lösungen zu entwerfen und umzusetzen, die den Menschen in den Mittelpunkt stellen. Zudem wurden und werden ethische Leitlinien für algorithmische Entscheidungssysteme in unterschiedlichen Kontexten diskutiert (vgl. [67], [173]). Wie es gelingen kann, die zugrunde liegenden ethischen Werte zu operationalisieren, um diese Anforderung konkret umzusetzen, ist der Fokus dieses Kapitels.

4.4.1 Status quo

4.4.1.1 Einordnung des soziotechnischen Systems im KI-Kontext

Soziotechnische Systeme beinhalten die Subsysteme Mensch und Technik, die miteinander verknüpft sind und in Wechselwirkung zueinander stehen oder stehen sollten (i. A. a. [174], [175], [176]). Die KI-Technologie steht dabei im Kontext zum Menschen, dem organisatorischen Umfeld und der Gesellschaft als Ganzes. Daher sind wichtige Fragestellungen die Integration der Technologie in gesellschaftliche Subsysteme, die Mensch-Technik-Interaktion [177] sowie die Organisationsentwicklung [178].

Soziotechnisches Gestalten von IT-Systemen erfordert, dass sie (Arbeits-)Aufgaben von Menschen in unterschiedlichen Rollen und im Nutzungskontext unterstützen können, d. h. für Menschen etwa über ergonomisch gestaltete Schnittstellen (z. B. Anzeigen und Stellteile) zugänglich machen (z. B. [179], Reihen DIN EN 614 [180], [181], [182] und DIN EN 894 [515]). Der nutzerzentrierte bzw. menschenzentrierte Ansatz [183] stellt den Menschen in den Mittelpunkt. Das Grundprinzip basiert darauf, die Bedürfnisse der Menschen zu erkennen, zu analysieren und daraus Produkte (KI) zu gestalten, die den Nutzenden dabei helfen, ihre Aufgabe effektiv, effizient und zufriedenstellend zu erledigen.

Technologie – Mensch – Organisation – Gesellschaft

Die Einführung von KI-Anwendungen in bestehende wie auch neue (Arbeits-)Prozesse ermöglicht die Generierung positiver Potenziale, ist jedoch auch mit Herausforderungen hinsichtlich der Governance dieser verbunden. KI sollte als „eine neue Klasse von Agenten in der Organisation" [184] betrachtet werden, was den Begriff deutlich weiter fasst als das reine Verständnis als technisches Werkzeug. Dies erfordert ein Verständnis für die Funktionsweise der KI und impliziert die organisatorische bzw. prozessuale Integration von KI-Anwendungen [185], [186], [184], [187]. Viele der Punkte und Fragestellungen, die hier für KI-Anwendungen aufgeführt werden, gelten ebenso für „klassische" Algorithmen.

In der Interaktion zwischen Mensch und KI lassen sich Autonomiegrade unterscheiden (siehe z. B. [188]). Diese hängen davon ab, wie die Interaktion gestaltet wird [189]:

- → KI kann z. B. nur dann etwas ausführen, wenn der Mensch dies vorher bestätigt.
- → KI ist autonomer, wenn sie eigenständig handelt, jedoch der Mensch ein Veto einlegen kann.
- → KI könnte zudem autonom handeln und den Menschen nur dann informieren, wenn dieser bewusst danach fragt.
- → Und schließlich könnte KI handeln, ohne den Menschen einzubeziehen.

Gestaltungskonzepte in Ergonomics/Human Factors (EHF) (u. a. zur soziotechnischen Gestaltung) bezogen sich in der Vergangenheit vorwiegend auf statische technische Systeme (z. B. Schnittstellengestaltung zu statischer und stationärer Maschine). Nicht nur, aber auch durch KI (als inhaltlich und zeitlich dynamisches System mit nicht mehr dokumentierbaren Ursache-Wirkungs-Beziehungen) muss das EHF-Gestaltungskonzept erweitert werden, damit Dynamik von Schnittstellen, Funktionsweisen und Auswirkungen auch für Menschen passend gestaltet werden.

Die Art und Weise des Arbeitens verändert sich mit der Einführung von KI-Anwendungen, die Anforderungen an Arbeitskräfte ebenso. Menschliche Attribute wie Empathie oder die der emotionalen Dimensionen werden sich in den Skill-Bedarfen hervorheben [190], [191], [192], [193]. Im Kontext von KI haben Menschen unterschiedliche Rollen: Der Mensch beauftragt, entwickelt, überarbeitet und nutzt die KI und ihre Ergebnisse für eigene Zwecke oder im Auftrag anderer (z. B. [194]). Nicht alle Menschen nutzen KI-Anwendungen in gleichem Umfang und brauchen daher entsprechende Kompetenzen in gleicher Breite und Tiefe (vgl. [190]). Die KI wiederum wirkt auf den Menschen und sein Verhalten, daher muss die Gestaltung die Leistungsvoraussetzungen des Menschen miteinbeziehen in Schnittstelle, Funktion und Wirkung [195]. Ähnlich zum Konzept der Kommunikation [196], wonach der Mensch nicht „nicht kommunizieren" kann, kann der Mensch nicht „nicht mit KI interagieren", sofern er/sie davon betroffen (z. B. Auftraggebende*r, Nutzende*r, Betroffene*r von

Auswirkungen) ist. Das macht soziotechnisches Gestalten von KI-Technik für ihre Zielsetzung, Funktionsweise und Wirkung in einem Gesamtsystem sowie auch für die Aufgaben-, Interaktions- und Informationsschnittstellen der Mensch-Technik-Interaktion erforderlich.

Das Konzept der soziotechnischen Systemgestaltung postuliert explizit die Notwendigkeit, den Technologieeinsatz und die Organisation gemeinsam zu optimieren („joint optimization“) (vgl. [197] bzw. [198], [199]). Die Organisation stellt somit den Rahmen und zudem eine zentrale (im besten Falle sozialpartnerschaftliche) Regulierungsebene des Verhältnisses zwischen Mensch und Technik dar und wird zugleich in ihrer inneren Struktur dadurch bedingt. Zentral ist die Beschreibung der Arbeitsaufgabe [199]. Maßgeblich sind dabei wiederum nach Ulich u. a. „die Unternehmensziele, die Unternehmensstrategie, die Unternehmensorganisation, die Marktposition, die Produkte und die Produktionsbedingungen, die Personalstruktur, der Technikeinsatz, das Qualitätsmanagement, das Innovationsverhalten, das Lohnsystem, die Arbeitszeitmodelle, die Art der Mitarbeitervertretung und der Aushandlungsprozesse sowie die soziotechnische Geschichte des Betriebes.“ Eine solche Organisation ist wiederum in eine Umwelt integriert (Staat und Gesellschaft, europäische und internationale Vereinbarungen, Standards und Rechtsetzungen).

Nicht nur der Organisation und dem Menschen, auch der gesellschaftlichen Perspektive kommt in soziotechnischen Systemen eine wichtige Rolle zu. Dabei ist „die Gesellschaft“ im jeweiligen Anwendungskontext von ganz unterschiedlichen Akteur*innen und Werten geprägt. Konfigurationen der Subsysteme Mensch und Technik kommen an der Schnittstelle „Gesellschaft“ zusammen. Dadurch können sich auch Macht- und Ungleichheitsverhältnisse oder diskriminierende Muster verfestigen. Technologien, die auf eine Automatisierung der Intelligenz des Menschen abzielen, sind nicht objektiv oder neutral und können zu einer Verstärkung von Rassismus und anderen Phänomenen sozialer Ungleichheit beitragen [200].

Mensch und Maschine beeinflussen und verändern sich wechselseitig im Nutzungsprozess [176]. Neue Entwicklungen wie das Maschinelle Lernen verdeutlichen dies, wenn Softwareprogramme dynamisch und „adaptiv“ auf ihre Nutzer*innen reagieren [201]. Dieses Verständnis von Mensch und Maschine hinterfragt die bisherige Konzeption von autonomen und strikt trennbaren Entitäten: Erst durch die wechselseitige Übernahme etwa sprachlicher Regeln wird den menschlichen und maschinellen Akteur*innen ihre Handlungsfähigkeit zuteil und es kommt zu einer Verständigung, die kollaborativ in der Interaktion entsteht [176].

Die Gestaltung soziotechnischer Systeme orientiert sich am MTO-Konzept, das davon ausgeht, dass die Teilsysteme Mensch (M), Technik (T) und Organisation (O) durch die Arbeitsaufgabe verknüpft sind und aufeinander einwirken [199]. Hierbei sind nicht nur die drei Teilsysteme selbst zu betrachten, vielmehr muss auch die Aufmerksamkeit auf die Schnittstellen Mensch-Technik, Mensch-Organisation und Technik-Organisation gelenkt werden. Für jede dieser Schnittstellen gibt es Zielkriterien. Darüber hinaus lassen sich für die soziotechnische Gestaltung übergeordnete Zielkonzepte formulieren. Zum Beispiel könnte das Zielkonzept „Adaptivität, Human-in-the-Loop und menschzentrierte Technik“ für die Mensch-Technik-Schnittstelle mit Zielbildern wie „ganzheitliche Aufgaben und Sinnstiftung“ (Schnittstelle Mensch-Organisation) und „Dezentralität“ (Schnittstelle Organisation-Technik) einhergehen. Vom Ausgangspunkt eines konsistenten soziotechnischen Mensch-Technik-Organisation-Modells können dann zentrale Fragen der Einführung, Nutzung und Folgeabschätzung von KI zielgerichteter bearbeitet werden [202], [203]:

- → Mit welchen Daten wird KI verbunden und zu welchen Zwecken eingesetzt?
- → Wie wirkt sich der KI-Einsatz auf menschliche Verhaltensweisen aus (z. B. Autonomie, Entscheidungsdilemma, Verhaltensanpassungen des Menschen)?
- → In welchem Verhältnis steht die KI-Anwendung zu menschlichen Bedürfnissen und Erwartungen (z. B. das Bedürfnis, sein Gegenüber einschätzen und überzeugen zu können)?
- → Welche systemischen Folgewirkungen hat der KI-Einsatz innerhalb des Systems, für seine Subsysteme, aber auch für Systemumwelt und Gesellschaft (z. B.: einfache Aufgaben werden automatisiert; schwierige Aufgaben werden schwieriger; veränderte Sicherheitsrisiken, da der Nutzende sein Verhalten an die automatisierte Technik anpasst)?

Die Besonderheit der soziotechnischen Perspektive

Als erprobtes Denkmodell verbindet die soziotechnische Perspektive in anschlussfähiger Weise frühere Industrialisierungsstufen mit ihren Mensch-Technik-Interaktionen und der digitalen Transformation. Der Einsatz von KI im Arbeitsprozess macht die Interaktion des Menschen mit KI erforderlich, die als Arbeitsaufgabe beschrieben werden kann. Bei der Arbeitsaufgabe [204], [205] wirken technische, organisatorische und qualifikatorische Elemente zusammen; ihre hier-

archische und sequenzielle Vollständigkeit (i. S. d. psychologischen Handlungsregulationstheorie, u. a. [206] kann als Beurteilungsmaßstab für die Qualität der Arbeit herangezogen werden (nach [207]).

Chancen und Risiken von KI hängen nicht allein von der Technik und deren Entwicklung ab, sondern vom Kontext der Anwendung. Die soziotechnische Perspektive stellt diesen Zusammenhang dar, erleichtert die Operationalisierung und ist zudem das geeignete multiperspektivische „Gegengewicht" zu einer rein technikzentrierten Sicht auf KI. Gleichzeitig bietet dieser Ansatz Innovationspotenzial, weil er Betroffene zu Beteiligten zu machen vermag und ein Modell der subsidiären (Fein-)Regulierung z. B. auf Unternehmensebene anbietet.

So wie Menschen systematisch Entscheidungsfehler machen [208], können auch bei der Entwicklung und dem Einsatz von KI „Bias"-Effekte oder Entscheidungsfehler bezüglich Fairness entstehen. „Bias" steht dabei für unerwünschte Verzerrungen, die teils bereits bei der Erhebung der Datensätze selbst oder durch die Selektion bzw. Art der Verarbeitung aufkommen können. Nicht zuletzt gehen Verzerrungen zurück auf Designentscheidungen (z. B. Datenbank und Logik) und die zugrunde liegenden Vorannahmen der Problemkonstruktion. Mit Einsatz einer KI entstehen so auch Herausforderungen hinsichtlich Verantwortlichkeit (Accountability) und Fairness, wenn KI in kognitiven Anwendungen genutzt wird (s. Praxisbeispiel Bewerbungen bearbeiten). Die unerwünschten Effekte rund um Bias oder Fairness [209] verweisen auf die Unsicherheit hinsichtlich der Konsequenzen der KI-Anwendung. Risiko bezeichnet für den Entscheidenden das Eintreffen eines oder mehrerer bekannter Umweltzustände mit einer empirisch ermittelten Eintrittswahrscheinlichkeit (z. B. morgen wird es regnen und die Eintrittswahrscheinlichkeit ist 70 %). Das heißt: Risiko ist quantifizierbar und somit möglicherweise steuerbar. Unsicherheit unterscheidet sich vom Risiko dahingehend, dass weder die möglichen Umweltzustände noch die mögliche Eintrittswahrscheinlichkeit bekannt sind (z. B. der Ausbruch der Covid-19-Pandemie und Folgeeffekte)[85]. Die existierenden Algorithmen für Risikosituationen und -abschätzung wägen das Risiko in Form von Eintrittswahrscheinlichkeiten und gewünschten Optimierungsleveln ab [208]. Neben Risiko und Unsicherheit gibt es weitere Faktoren, die zu berücksichtigen sind (z. B. wahrgenommene Prozesskontrolle [210] oder Entscheidungstiefe von Algorithmen [211] usw.). Insgesamt lässt sich festhalten, dass die menschliche Wahrnehmung bei der Analyse, Gestaltung und Bewertung von KI-Systemen eine entscheidende Bedeutung hat.

Aspekte sozialer Nachhaltigkeit im soziotechnischen Kontext

Was unter Nachhaltigkeit normativ verstanden wird, wird häufig in Parlamenten und in nationalen und internationalen Gremien verhandelt, entschieden und ggf. auch gesetzlich umgesetzt. Bei der Gestaltung von KI-Anwendungen ist somit sicherzustellen, dass diese Nachhaltigkeitskriterien genügen. Hieraus resultiert als Anforderung an das KI-System die Parametrisierbarkeit in Bezug auf quantitative Zielsetzungen aus Nachhaltigkeitsvorgaben.

Im soziotechnischen Kontext sind dabei insbesondere Aspekte sozialer Nachhaltigkeit zu betrachten. „Im Hinblick auf die Entwicklung, Nutzung und den Einsatz von KI-Systemen bedeutet Nachhaltigkeit vor allem, dass die Würde des Menschen respektiert wird, keine Menschen ausgeschlossen, benachteiligt oder diskriminiert werden und die menschliche Autonomie und Handlungsfreiheit durch KI-Systeme nicht eingeschränkt werden dürfen. In einer erweiterten Perspektive auf Nachhaltigkeit bedeutet soziale Nachhaltigkeit auch, dass neben körperlicher Unversehrtheit und menschenwürdigen Lebensbedingungen auch die Fähigkeit, auf menschliche Art und Weise zu denken, zu argumentieren und zu handeln, nicht eingeschränkt werden sollte. Hier zeigt sich schon, dass ein umfassendes Verständnis von sozialer Nachhaltigkeit sehr weitreichende Konsequenzen für die Gestaltung von KI-Systemen hat." [223]. Zugleich wird anhand der vielfältigen zu berücksichtigenden Ziele und Aspekte deutlich, dass Gesetze und Normen an ihre Grenzen stoßen und nicht jedes Detail regeln können. Notwendig werden subsidiäre Aushandlungssysteme, z. B. auf betrieblicher Ebene, sowie auch individuelle Entscheidungsrechte. Die Nachhaltigkeit von KI-Systemen wird letzten Endes unter Nutzung unterschiedlicher Indikatorensysteme und Regeln auf der MTO-Ebene (MTO: Mensch, Technik und Organisation) verhandelt und entschieden.

KI-Systeme können Einzelpersonen und Gruppen von Personen Schaden zufügen, die Muhammad (2022) [224] verschiedenen Typen zuordnet:

→ So gibt es „Vergabe-Fehler", indem das System Möglichkeiten, Ressourcen oder Informationen zurückhält oder unfair zur Verfügung stellt. Ein Beispiel sind hier die vielfach rezipierten Benachteiligungen bei Bewerbungsverfahren, aber auch eine Ungleichbehandlung von Menschen mit und ohne Internetzugang [225].

85 https://wirtschaftslexikon.gabler.de/definition/risiko-44896/version-268200

- → Eine weitere Kategorie sind die „Servicegüte-Fehler“, bei denen das System nicht für alle Gruppen ähnlich gut arbeitet.
- → Ein „Repräsentations-Fehler“ tritt auf, wenn die Entwicklung oder die Verwendung eines Systems einzelne Gruppen über- oder unterrepräsentiert. Hier geht es beispielsweise um das überwiegende Anzeigen männlicher Personen bei einer Bildersuche nach „CEO“ [226], [227].
- → Weiterhin wird ein möglicher „Stereotyp-Fehler“ aufgeführt, bei dem das System Stereotypen reproduziert und verstärkt, indem beispielsweise stereotypische Charakteristika unreflektiert allen Angehörigen einer Gruppe zugewiesen werden.
- → Ein „Verunglimpfungs-Fehler“ tritt auf, wenn das System aktiv abwertend oder beleidigend wird, wie beispielsweise das auf Reichweite optimierende Verhalten des Twitter-Bots Tay von Microsoft.
- → Als „Prozess-Fehler“ wird schließlich das Verhalten eines Systems bezeichnet, welches Entscheidungen aufgrund von Charakteristika trifft, die nicht für die Aufgabe relevant sein sollten. Ein Beispiel hierfür ist ein Bewerbungsprozessmanagement, welches Menschen mit zu viel Berufserfahrungen als die benötigte abwertet [228].

Der Artificial Intelligence Act (AI Act) der Europäischen Union (EU)

Der vorliegende Entwurf zur KI-Verordnung der EU adressiert die soziotechnische Perspektive: „Künstliche Intelligenz (KI) sollte ein Werkzeug für die Menschen sein und eine Kraft für das Gute in der Gesellschaft darstellen, mit dem letztendlichen Ziel, das menschliche Wohlbefinden zu steigern. Das europäische Konzept für künstliche Intelligenz setzt auf Exzellenz und Vertrauen; es zielt darauf ab, die Forschung und die industriellen Kapazitäten zu fördern und gleichzeitig die Sicherheit und die Grundrechte zu gewährleisten.“ (Europäische Kommission: Entwurf zum „Umsetzungsplan von Standardisierungsanforderungen durch die europaweiten Normungsorganisationen“). „Mit dem vorgeschlagenen KI-Gesetz werden u. a. Anforderungen an das Inverkehrbringen und die Inbetriebnahme von KI-Systemen mit hohem Risiko eingeführt. Diese Anforderungen beziehen sich auf die Bereiche Risikomanagement, Datenqualität und Governance, technische Dokumentation, Aufzeichnungen, Transparenz und Bereitstellung von Informationen für die Nutzer*innen, menschliche Aufsicht, Genauigkeit, Robustheit und Cybersicherheit.“ Dies kennzeichnet den Entwurf bezogen auf die Definition der Schutzziele. Die Regulierungen betreffen hingegen mehr das Produkt KI und weniger ihre Anwendung im Rahmen von (Arbeits-)Prozessen. Einer soziotechnischen Betrachtung wird der Entwurf deshalb nur teilweise gerecht.

Die Anforderungen, Transparenz und Informationen für Benutzende zur Verfügung zu stellen und eine menschliche Aufsicht zu gewährleisten, können nur erfüllt werden, wenn das KI-System als soziotechnisches System verstanden und der Mensch als Teil des Systems mitgedacht wird. Aus diesem Grund ist es von größter Notwendigkeit, bei der Beschäftigung mit KI klar die Systemgrenzen des soziotechnischen KI-Systems zu definieren, das Zusammenwirken von dessen Systemelementen zu betrachten und vor allem das Wechselverhältnis der technischen KI-Komponenten mit dem menschlichen Verhalten zu beurteilen und zu gestalten. Da manche Systeme im Einsatz weiterlernen, ist eine einmalige Prüfung und Optimierung zu einem bestimmten Zeitpunkt für die Lebensdauer eines Systems nicht ausreichend (siehe Kapitel 4.4.2.4).

Beispiele aus der Praxis

Konkrete Spannungsszenarien werden in der öffentlichen Debatte diskutiert: Hier gibt es z. B. die automatisierte Auswahl von Bewerbungsunterlagen: Ein bereits 2014 von Amazon entwickeltes Verfahren sorgte für Schlagzeilen, da es strukturell Frauen benachteiligte. Laut Reuters (2018) [212] war der Algorithmus mit den Datensätzen der angenommenen Bewerber trainiert worden – in den zugrunde gelegten zehn Jahren waren allerdings vor allem Männer eingestellt worden, sodass der Algorithmus zu dem Schluss kam, dass Bewerbungen von Männern zu bevorzugen seien. Doch auch bei anderen Personalauswahl-Softwarelösungen wurden ähnliche Entscheidungsmuster nachgewiesen (vgl. dazu z. B. [213]). Trainingsdaten oder Lernverfahren, die selbst schon einen Bias enthalten, können zu einem „falschen“ Ergebnis mit einem sogenannten „Bias“ führen. Ähnlich ist die Entwicklung eines reichweitenoptimierenden Twitterbots fehlgeschlagen, der innerhalb kürzester Zeit lernte, rechtsradikales Gedankengut zu verbreiten. Eine weitere aktuelle Debatte entbrennt um autonom fahrende Autos und deren „Entscheidungen“ in Dilemmasituationen ([214], [215]). Diese Probleme stellen die Konsequenzen maßgeblicher Entscheidungen im Hinblick auf die Datengrundlage (z. B. sexistischer „Bias“ schon in den Trainingsdaten), Nutzung und Beeinflussbarkeit (z. B. Lernen in Echtzeit aus ungefilterten Daten) und Interdependenzen (z. B. rechtliche Konsequenzen, Userakzeptanz etc.) bei der Entwicklung KI-basierter Algorithmen in den Vordergrund – und begründen die Notwendigkeit einer sorgfältigen Abwägung der auszuwählenden Lösung und der Auswahl der Trainingsdaten.

4.4.1.2 Schnittstellen zu nicht-normungsfähigen Bereichen

Vorrang vor der nationalen und internationalen Normungsarbeit haben europäische und nationale Gesetze, Vorschriften und Regeln. In Deutschland betrifft dies z. B. die Themen Arbeitsschutz sowie Datenschutz.

Den Sozialpartner*innen obliegt gemäß Art. 9 Abs. 3 GG die „Wahrung und Förderung der Arbeits- und Wirtschaftsbedingungen“. Die Sozialpartnerschaft erstreckt sich daher auf alle Gebiete der Wirtschafts- und Sozialpolitik. Insbesondere gehört dazu die Regelung aller Vergütungs- und sonstigen Arbeitsbedingungen durch Tarifverträge (Tarifautonomie). Sozialpartnerschaftliche Aufgaben zählen demnach zu dem nicht-normungsrelevanten Bereich. Die Normung kann hier allenfalls den rechtlichen Rahmen ergänzen oder konkretisieren.

Vorschriften und Regeln zu Sicherheit und Gesundheit
Für die Gestaltung von Arbeitssystemen ist die Entwicklung und Berücksichtigung von Normen und arbeitswissenschaftlichen Erkenntnissen nicht hinreichend. Anforderungen an Arbeitsplätze und Arbeitsmittel sind national und europäisch u. a. hinsichtlich Sicherheit und Gesundheit bei der Arbeit gesetzlich reguliert. Grundsätzlich zu unterscheiden sind hierbei gesetzliche Anforderungen, die sich einerseits auf die Gestaltung und das Inverkehrbringen von Produkten und Arbeitsmitteln (Verantwortung beim Herstellenden) und andererseits auf den betrieblichen Arbeitsschutz (Verantwortung beim Betreiber) beziehen.

Bei Produkten und Arbeitsmitteln hat die europäische Maschinenrichtlinie [216], [217] eine herausgehobene Bedeutung. Diese ist in Deutschland durch das Produktsicherheitsgesetz (ProdSG) und die darauf gestützte Maschinenverordnung (9. ProdSV) national umgesetzt. Die EU-Maschinenrichtlinie wird derzeit im Zusammenhang mit dem EU KI Act als EU-Maschinenprodukteverordnung novelliert, was umfangreiche Anpassungen in Normen zur Folge haben wird. Im Bereich der Maschinensicherheit sind durch die Europäische Kommission mandatierte, harmonisierte Normen (siehe Kapitel 1.4.4) von besonderer Relevanz. Diese lösen bei ihrer Anwendung die Vermutung aus, dass die Gestaltung einer Maschine den rechtlichen Erfordernissen entspricht. Sachverhalte, die nicht in harmonisierten Normen geregelt sind, müssen im Rahmen der immer erforderlichen Risikobeurteilung bewertet werden und ggf. sind entsprechende Maßnahmen zu treffen.

Neben der Maschinenrichtlinie existieren eine Zahl weiterer europäischer Richtlinien inklusive nationaler Umsetzungen (bzw. sind in Erarbeitung, wie z. B. der AI Act), die bei der technischen Gestaltung von KI-Systemen zu berücksichtigen sind.

Für den betrieblichen Arbeitsschutz ist in Deutschland das Arbeitsschutzgesetz maßgeblich, welches im Wesentlichen eine nationale Umsetzung des europäischen Arbeitsschutzrechts darstellt. Zentrales Instrument des Arbeitsschutzgesetzes ist die Gefährdungsbeurteilung, welche die Arbeitsbedingungen und die damit verbundenen Risiken für die Sicherheit und Gesundheit der Beschäftigten zum Gegenstand hat. Das deutsche Arbeitsschutzgesetz wird national durch Verordnungen konkretisiert, die rechtlich bindende Vorschriften sind. Zur weiteren Konkretisierung der Verordnungen (z. B. BetrSichV, ArbStättV, GefStoffV, ArbMedVV) werden staatliche Regeln in beim Bundesministerium für Arbeit und Soziales (BMAS) beratend angesiedelten Ausschüssen[86] unter Beteiligung von Ländern, Arbeitgebern, Gewerkschaften, der Deutschen gesetzlichen Unfallversicherung (DGUV), Wissenschaft und ggf. weiterer Institutionen/Verbände formuliert.

In Dualen Arbeitsschutz wird in Deutschland ein kohärentes Vorschriften- und Regelwerk in Abstimmung zwischen Staat und DGUV erstellt, sodass für Unternehmen in Deutschland auch das Vorschriften- und Regelwerk der branchenspezifisch aufgestellten Unfallversicherungsträger zu beachten ist bzw. weitere konkretisierende Regeln und Informationen unterstützen.

Für den Einsatz von und die Arbeitsgestaltung zu KI-Systemen gelten hiermit in Deutschland über das Vorschriften- und Regelwerk zum Arbeitsschutz zentrale Grundprinzipien der Prävention. Die Grundprinzipien der Prävention sind für Industrie 4.0 und ansatzweise für KI-Systeme in dem DGUV-Positionspapier 2/2017 [218] erläutert. Erste Konkretisierungen im Technischen Regelwerk des Dualen Arbeitsschutzes in Deutschland sind verfügbar und werden laufend weiterentwickelt. Vorhaben oder Realisierungen von Vorhaben zum Einsatz von KI erfordern jeweils fachkundige und spezifische Prüfungen des Regelwerks.

Aspekte des Datenschutzes
KI-Systeme nutzen typischerweise große Datenmengen. Hierdurch können Wechselwirkungen mit den Anforderungen des

86 https://www.baua.de/DE/Aufgaben/Geschaeftsfuehrung-von-Ausschuessen/Geschaeftsfuehrung-von-Ausschuessen_node.html

Datenschutzes und den Persönlichkeitsrechten von Nutzenden entstehen.

Die Grundprinzipien des Datenschutzes wie
→ Zweckfeststellung bzw. Zweckbindung von Daten,
→ Erforderlichkeit,
→ Transparenz,
→ Datenvermeidung und Datensparsamkeit

erfordern bei der Gestaltung des soziotechnischen Systems sorgfältige Berücksichtigung: Ob die Prinzipien inhaltlich erfüllt sind, kann von den verwendeten Technologien sowie dem jeweiligen Anwendungsfall abhängen. Bewerbungsunterlagen, die durch KI-System zu Rekrutierungszwecken diskriminierend vorsortiert wurden (s. o.), könnten mit anderem Einsatz von KI stattdessen dazu verwendet werden, Diskriminierung im System entgegenzuwirken.

Da das Datenschutzrecht für Daten, die im Beschäftigungsverhältnis erhoben werden, bisher kein einheitliches Regelwerk bietet, können im betrieblichen Kontext Betriebs- und Dienstvereinbarungen sinnvoll sein, die datenschutzbezogene Aspekte, technologische Herangehensweisen und wertebasierte Prinzipien (beispielsweise Ethikkodex) miteinander verbinden.

4.4.2 Anforderungen und Herausforderungen

4.4.2.1 Die soziotechnische Perspektive im KI-Lebenszyklus

Die soziotechnische Perspektive muss während des kompletten KI-Lebenszyklus betrachtet werden (vgl. Kapitel 4.1.2.3). In jeder Phase des KI-Lebenszyklus (vgl. ISO/IEC 22989:2022 [16], ISO/IEC 23053:2022 [24]) wird dabei der Fokus auf spezifische Aspekte des soziotechnischen Systems gelegt.

Hierbei gilt es zu beachten, dass die Ansätze der allgemeinen Systemtheorie (von [219] & [220]) wie auch der soziologischen Systemtheorie (z. B. [221], [222]) bei der Entwicklung von KI-Systemen nur eingeschränkt greifen. Während sich ein klassisches System während seiner Betriebsphase nicht oder nur geringfügig weiterentwickelt, haben KI-Systeme die Fähigkeit, sich in einem gesteckten Rahmen weiterzuentwickeln. Ohne eine sorgfältige Abwägung dieses Rahmens in der Designphase kann es zu unerwartetem und unerwünschtem Verhalten führen. Während der Betriebsphase können nur „Standbilder“ des Systemzustands aufgenommen werden, die das System und seine soziotechnische Interaktion zu einem speziellen Zeitpunkt abbilden. Dies erschwert Bewertung und Design, sodass ein besonderes Augenmerk auf die Wirksamkeit des Systems von einer soziotechnischen Perspektive nötig wird.

Im Folgenden wird der KI-Lebenszyklus bezüglich der jeweils relevanten soziotechnischen Fragestellungen beleuchtet. Kapitel 4.4.2.2 betrachtet die Phase „Initiierung“, Kapitel 4.4.2.3 befasst sich mit den Aktivitäten in den Phasen „Design und Entwicklung“ sowie „Verifikation und Validierung“ und Kapitel 4.4.2.4 beleuchtet die Phasen „Überführung in die Einsatzumgebung“, „Betrieb und Überwachung“, „Reevaluierung“, „Kontinuierliche Validierung“ und „Außerdienststellung“. Hierbei gilt es zu beachten, dass die Betrachtung nicht den Anspruch einer vollständigen Erörterung aller soziotechnischen Perspektiven und Fragestellungen erhebt, da dies den Rahmen der Normungsroadmap Künstliche Intelligenz gesprengt hätte.

Essenziell bei der Betrachtung der soziotechnischen Perspektive im KI-Lebenszyklus ist auch, dass die betrachteten Subsysteme Mensch und Technik sowie deren Wechselwirkung in jeder Phase beschrieben und dokumentiert werden.

4.4.2.2 Initialisierung

Diese Phase korrespondiert mit der ISO/IEC 22989:2022 [16] Phase „Initiierung“. Aus soziotechnischer Sicht werden in dieser Phase insbesondere die Ziele der Anwendung und Anforderungen definiert. Zu welchem Zweck braucht es eine KI-basierte Anwendung? Welche Anforderungen muss sie aufgrund der soziotechnischen Einbettung erfüllen? Das sind die wesentlichen Fragen, die sich die relevanten Akteur*innen zu Beginn des KI-Lebenszyklus stellen müssen und die keine ausschließlich technischen Antworten erfordern. Dennoch geben die Antworten vor, wonach in den nächsten Schritten die technischen Komponenten gewählt werden sollen. Es geht schlussendlich darum, auf Basis einer eingehenden Problemanalyse in einer gegebenen Situation von der Idee zur Entscheidung für ein KI-System zu gelangen und den Entwicklungsprozess anzustoßen. Folglich verlangt diese Phase eine intensive Analyse des Wirkfeldes des KI-Systems und seiner möglichen soziotechnischen Folgen. Mit der Definition von Zielen und Anforderungen ist der Rahmen für den weiteren Prozess gesetzt und Beteiligte können an dieser Stelle die Entwicklung des KI-Systems wesentlich ausrichten. In den nächsten Phasen wird es immer wieder Rückbezüge zu den hier festgelegten Parametern geben.

Ziel der Phase Initialisierung
Ziel dieser Phase ist aus soziotechnischer Sicht die Kontextualisierung der Anwendung in ihr Umfeld. Dafür gilt es, zunächst eine initiale Entscheidung zu treffen,
→ wieso ein KI-System entwickelt werden soll,
→ welches Problem es löst,
→ welches Bedürfnis der Zielgruppe erfüllt werden soll und
→ welches die Erfolgsparameter sind.

Soziotechnische Schritte der Phase Initialisierung

Relevante Personengruppen definieren und beteiligen
Dabei ist insbesondere zu beachten, dass nicht das technisch Mögliche entscheidend ist, sondern der reale Bedarf, der sich aus der Problemanalyse ergibt. Wissen über die Zielgruppe ist dabei essenziell und sollte mit Blick auf Diversität keine ausschließlich stereotypischen Vorstellungen von Menschen widerspiegeln. Auch Betroffene und ihre Rechte sind Bausteine für die eingehende Problemanalyse. Eine Einbindung dieser Gruppen in den Entwicklungsprozess ermöglicht direkte und ungefilterte Einblicke in die jeweiligen Bedarfe und kann die Qualität des KI-Entwicklungsprozesses wesentlich beeinflussen. Im Rahmen des BMAS-geförderten Forschungprojekts KIDD wird ein Ansatz für die Auswahl und Beteiligung von relevanten Stakeholdern in verschiedenen Experimentierräumen praktisch erprobt.

Participatory-Design-Ansatz
Die Einbindung relevanter Personengruppen kann auch über einen Participatory-Design-Ansatz erfolgen (vgl. [229]). Hierbei geht es vor allem um das gemeinsame Antizipieren von Zukunftsszenarien. Dies lässt sich in dem Terminus „reflection-in-action“ (ebd.) kondensieren. Es geht darum, den User*innen eine Stimme zu verleihen, ohne dass diese dabei selbst zu Entwickler*innen werden müssen (vgl. ebd.).

Mögliche Methoden für diese Übersetzungsleistung sind das Kreieren von Prototypen, Lehrmodellen und Simulationen (vgl. ebd.), Exkursionen von ähnlichen, bereits laufenden und funktionierenden Systemen, Szenarien, Zukunftsworkshops, Spiele oder „design fiction“ [71]. Dies kann mit dem Begriff „storytelling methods“ (ebd.) zusammengefasst werden.

Eine weitere wichtige Säule des Participatory Design ist die anhaltende Evaluation durch die User*innen. Anwendungen Künstlicher Intelligenz stellen die bestehenden Konzepte des Participatory Design gerade in Bezug auf anhaltende Evaluationen der Systeme vor neue Herausforderungen. Dies liegt an der neuartigen Beschaffenheit der KI-Komponenten, die sich etwa durch die hohe Dynamik und Verwobenheit von Algorithmen, Parametern und Daten sowie durch statistische Inferenzen und der Komplexität von Trainingsdatensets nicht unmittelbar jenen erschließen, die am Designprozess nicht beteiligt waren.

Um eine sichere Nutzung über den gesamten Lebenszyklus der Software zu gewährleisten, bedarf es gerade für die anhaltende Evaluation neue Beteiligungskonzepte. Erste Ansätze können in den Arbeiten zu XAI (Explainable AI) gefunden werden, die etwa über die Offenlegung kritischer Entscheidungspunkte oder über neue Visualisierungskonzepte einen ersten Zugang zu den zugrunde liegenden Softwarelogiken ermöglichen. Diese Ansätze bedürfen weiterer Ausarbeitung.

Soziotechnische Anforderungen definieren
In der Initialisierungsphase werden darüber hinaus insgesamt Anforderungen definiert, die sich über den gesamten Lebenszyklus hinweg spannen. Neben technischen Anforderungen gilt es, insbesondere soziotechnische Aspekte zu definieren. Die Gestaltungsanforderungen können dabei auch ethische Aspekte beinhalten. Empfohlen wird, diese Anforderung so zu operationalisieren, dass daraus diese Anreize für eine Umsetzung resultieren. Beispielhaft können gängige Anforderungskataloge an die Entwicklung von KI-Systemen aus ethischer Sicht sogenannte ethics-by-design-Kataloge sein. Hier sind beispielsweise die von der Bertelsmann Stiftung entwickelten Algo.Rules zu nennen, die anhand von neun Gestaltungsprinzipien und etwa 120 Fragen Anforderungen entlang des gesamten KI-Lebenszyklus definieren [230]. Darüber hinaus bietet das WKIO-Modell der AI Ethics Impact Group eine etablierte Methode, vorab definierte ethische Werte zu operationalisieren [231]. Werte werden dementsprechend durch Kriterien auf klar definierte Teilaspekte heruntergebrochen und diese durch Indikatoren und korrespondierende Observablen messbar gemacht. Über die ethischen Aspekte hinaus gilt es die weiteren hier im Schwerpunktkapitel definierten soziotechnischen Anforderungen zu berücksichtigen.

Risiko analysieren
Des Weiteren umfasst diese Phase eine initiale Risikoanalyse, die die soziotechnischen Folgen aus Sicht mehrerer Stakeholder identifiziert, bevor es überhaupt zur Entwicklung und Umsetzung der Anwendung kommt (siehe ISO/IEC 23894:2022 [25]). Dabei sind neben technischen und rechtlichen Folgen entsprechend insbesondere ethische und soziale Folgen aus Sicht von Mensch und Gesellschaft zu adressieren. Daraus ergibt sich eine vielschichtige Bedeutung potenzieller und zu identifizierender Risiken. Folgende Fragen

können dabei helfen: „Welche Grundrechte oder -werte könnten von dem Einsatz der Software potenziell berührt sein? Welches sind die beabsichtigten Auswirkungen der Software? Wer ist von dem Einsatz des algorithmischen Assistenzsystems betroffen? Welche potenziellen Auswirkungen hat der Einsatz der Software auf die unterschiedlichen Stakeholder? Welche potenziellen Auswirkungen hat der Einsatz der Software auf Gesellschaft, Wirtschaft oder Umwelt? Welche Risiken könnten bei möglichen Fehlern bei der Entwicklung oder dem Einsatz der Software entstehen? Welche Szenarien sind hier denkbar?“ [232].

Drei Kritikalitätsmodelle

Es gibt darüber hinaus im Wesentlichen drei Varianten der Kritikalitätseinordnung von KI-Systemen, die helfen können, die multidimensionalen Risiken zu sortieren und entsprechende Maßnahmen in einem nächsten Schritt zu empfehlen. Mokänder et al. [233] unterscheidet drei gängige Modelle:

→ den Schalter,
→ die Leiter und
→ die Matrix.

Das Schalter-Modell fungiert als binärer Klassifizierungsansatz. Die von der EU-Kommission vorgeschlagene KI-Verordnung nutzt das Schalter-Modell, indem bestimmte Bedingungen an ein System definiert werden, das später unter den Anwendungsbereich der KI-Verordnung fallen soll. Dieses Modell ist ein relativ intuitives, wenig aufwendiges Verfahren, das allerdings Gefahr läuft, zu wenige oder zu viele Systeme für eine weitere Befassung zu definieren [233].

Das Leiter-Modell stellt diesbezüglich eine höhere Komplexitätsstufe dar. Dieses Modell unterscheidet KI-Systeme anhand verschiedener Faktoren und gruppiert sie in unterschiedliche Risikoklassen. Ein bereits etabliertes Modell der Leiter wird von der AI Ethics Impact Group – unter der Leitung der Bertelsmann Stiftung und des VDE – vorgestellt. Die dort präsentierte Risikomatrix nach Krafft und Zweig (2019) unterschiedet die Intensität des möglichen Schadens durch das KI-System und die Abhängigkeit der betroffenen Person(en) von der jeweiligen Entscheidung (AI Ethics Impact Group 2020: 35). Anhand dieser Faktoren werden fünf Risikoklassen unterschieden, die jeweils verschiedene Risikomanagementschritte im Anschluss erfordern. Auch die vorgeschlagene KI-Verordnung etabliert eine ähnliche Risikomatrix. Die Leiter-Modelle eint die Erkenntnis, dass nicht die technische Komplexität, sondern die Modalitäten der sozialen Einbettung das Risiko der Systeme wesentlich definieren. Obwohl die Leiter-Modelle komplexer sind, eröffnen sie in der praktischen Anwendung ausreichend Orientierung zur Einordnung der Kritikalität.

Das dritte Modell zur Klassifizierung der KI-Systeme beschreibt [233] als Matrix-Modell und ist ein multi-dimensionaler Ansatz. Ein Beispiel hierfür ist der OECD[87]-Ansatz zur Klassifizierung von KI-Systemen anhand von fünf Dimensionen. Dieses Modell entspricht den sehr vielfältigen Anwendungsfällen von KI-Systemen und präsentiert folglich das komplexeste Modell zur Einordnung der Risiken.

Die drei verschiedenen Modelle zur Einordnung der Kritikalität haben für sich genommen jeweils Vor- und Nachteile – insbesondere in Hinblick auf Praktikabilität und Aussagekraft. In der Praxis sind Mischformen denkbar, so arbeitet die KI-Verordnung sowohl mit einem binären Ansatz als auch mit abgestuften Risikoklassen. Je nach vorab definierten Bedingungen bzw. Dimensionen kann ein KI-System als risikobehaftet definiert sein oder nicht. Die soziotechnische Natur der Systeme erfordert deshalb eine qualitative Befassung mit den Modellen und eine sensible Abwägung der möglichen Risiken für Mensch und Gesellschaft. Normungsansätze sollten die Vielschichtigkeit von Kritikalität beachten.

Risiko managen

Die vorab identifizierten Risiken sollten in einem nächsten Schritt mit einem entsprechenden Plan angegangen werden. Hierbei können bereits etablierte Risk-Managementsysteme unterstützen, die identifizierten Risiken entlang des gesamten KI-Lebenszyklus – prozessorientiert – zu verringern (vgl. [25]).

Ansprüche an Transparenz und Accountability konkretisieren

In dieser Phase konstituieren sich auch die Ansprüche an Transparency und Accountability: Welche Informationen müssen offengelegt werden? Für wen müssen diese Informationen offengelegt werden? Und mit welcher technischen Tiefe müssen Informationen angereichert werden, um gleichzeitig hilfreich und verständlich zu sein? Wer kann zur Rechenschaft gezogen werden bei eventuell auftretenden Schadensereignissen? Ohne eine Klärung dieser Aspekte kann eine weitere Entwicklung des KI-Systems zu schädlichen Folgen für Mensch und Gesellschaft führen.

87 The Organisation for Economic Co-operation and Development [OECD(2022)]

Umsetzbarkeit evaluieren
Darüber hinaus werden Kosten, Aufwand und Ressourcen in dieser Phase antizipiert und die grundsätzliche Umsetzbarkeit der Anwendung evaluiert. ISO/IEC 22989:2022 [16] definiert hier insbesondere auch kaufmännische Überlegungen. Daneben sollten insbesondere auch gesellschaftliche Abwägungen betrachtet werden.

Nach vollständiger Initialisierung können weitere Schritte innerhalb der Phase Planung und Entwicklung angestoßen werden. Neue Informationen, beispielsweise über Risiken, können eine Rückkehr zu den Schritten in der Initialisierung erforderlich machen und sollten beispielsweise in die Risikoanalyse und das Risikomanagement einfließen.

4.4.2.3 Planung & Gestaltung

Dieses Kapitel befasst sich mit den Phasen „Design und Entwicklung“ und „Verifikation und Validierung“ der ISO/IEC 22989:2022 [16].

Ziele der Phase aus soziotechnischer Sicht
In dieser Phase wird das KI-System gemäß den zuvor definierten Zielen und Anforderungen (vgl. Kapitel 4.4.2.2) konkretisiert. Hierbei werden i. d. R. zunächst mehrere Groblösungen entwickelt, die hinsichtlich der Erfüllung der Ziele und Anforderungen überprüft werden. Nicht jede erarbeitete Groblösung kann die gesetzten Ziele optimal erfüllen, sodass evtl. mehrere Planungsschleifen erforderlich sein können, bevor die gewählte Groblösung feingeplant und alle notwendigen Schritte für die Inbetriebnahme und den späteren Betrieb (vgl. Kapitel 4.4.2.4) vorbereitet werden können. Hierbei ist es wichtig, alle Beteiligten (z. B. Betreiber des KI-Systems, spätere Nutzende des KI-Systems, Interessensvertretungen von Betreiber und Nutzenden, Vertretende der Zivilgesellschaft; zur Vertiefung vgl. Kapitel 4.4.2.3) frühzeitig und partizipativ in die Planung einzubinden (vgl. z. B. [203]).

Bei der Planung und Gestaltung von KI-Systemen sind somit die Umsetzung von ergonomischen Grundsätzen und Prinzipien sowie eine gebrauchstaugliche Gestaltung von Produkten und Arbeitsmitteln erfolgskritische Ziele. Damit ist die Anwendung dieser ergonomischen Grundsätze und Prinzipien auch ein wesentliches Gütemerkmal von KI-Systemen als Arbeitsmittel bzw. Gebrauchsgegenständen. Das gilt im gesamten Produktlebenszyklus des KI-Systems (vgl. Kapitel 4.1.2.3 sowie dort Abbildung 19; nach ISO/IEC 22989:2022 [16]). Vom Produktentstehungsprozess über Inbetriebnahme und alltäglicher operativer Anwendung bis hin zur Außerbetriebsetzung sind nicht nur der Stand der technologischen Entwicklung sowie der spezifische Anwendungsfall zu berücksichtigen, sondern auch die Grundsätze und Prinzipien einer menschengerechten und partizipativen soziotechnischen Gestaltung. Dieses Erfordernis spiegelt sich bislang meist nicht in den korrespondierenden Normen wider.

Inhalt der Phase

DIMENSIONEN DER GESTALTUNG ERMITTELN
Zur systematischen und zielgerichteten Gestaltung des KI-Systems müssen die zugrunde liegenden Wirkzusammenhänge im betrachteten soziotechnischen System bekannt sein. Die Dimensionen der Gestaltung umfassen somit alle zu klärenden Fragestellungen im Planungs- und Gestaltungsprozess, grenzen die gesetzlichen Rahmenbedingungen (vgl. Kapitel 4.4.1.2) sowie gültige Normen und Standards ein und geben Hinweise darauf, wer zu beteiligen ist und welche Methoden oder Instrumente eingesetzt werden können.

Bei der Analyse, Bewertung und Gestaltung soziotechnischer Systeme ist zu beachten, dass sie stets von sachlichen (technisch-organisatorischen) und zugleich von menschlichen (persönlichen) Gegebenheiten beeinflusst werden (z. B. DIN EN ISO 6385:2016 [235]). Das MTO-Konzept geht davon aus, dass Mensch, Technik und Organisation stets in ihrer gegenseitigen Abhängigkeit und ihrem Zusammenwirken zu reflektieren sind. Der Arbeitsaufgabe kommt dabei eine zentrale Rolle zu, da diese die drei Elemente Mensch, Technik und Organisation miteinander verknüpft [236], [205]. Die Dimensionen der Gestaltung eines KI-Systems ergeben sich daher stets aus den drei Elementen Mensch, Technik und Organisation sowie aus deren Schnittstellen (also Mensch-Technik, Mensch-Organisation und Organisation-Technik) zueinander.

Zur Konkretisierung spezifischer Fragestellungen können verschiedene Handlungsrahmen herangezogen werden. Tabelle 8 skizziert exemplarisch einige einschlägige Handlungsrahmen mit den dort beschriebenen Gestaltungsdimensionen. Zur Vertiefung wird auf die jeweiligen Quellen verwiesen.

Tabelle 8: Exemplarische Handlungsrahmen zur Konkretisierung der Dimensionen der Gestaltung eines KI-Systems

Autor	Zu betrachtende Dimensionen der Gestaltung
Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (Hrsg.) [237]	→ Veränderbarkeit (System, Umfeld) → Transparenz (Expert*innen, Beteiligte) → Vernetzung (intern, nach außen) → Kontrollierbarkeit (Emergenz, Beschränkungen) → Widerstandsfähigkeit (Robustheit, Resilienz) → Involviertheit des Menschen (Handelnder, Gefährdeter) → Schadensfolgen (Personenschäden, sonstige Schäden)
Huchler et al. (2020) [202]	→ Schutz des Einzelnen (Sicherheit und Gesundheitsschutz, Datenschutz und verantwortungsbewusste Leistungserfassung, Vielfaltssensibilität und Diskriminierungsfreiheit) → Vertrauenswürdigkeit (Qualität der verfügbaren Daten, Transparenz, Erklärbarkeit und Widerspruchsfreiheit, Verantwortung, Haftung und Systemvertrauen) → Sinnvolle Arbeitsteilung (Angemessenheit, Entlastung und Unterstützung, Handlungsträgerschaft und Situationskontrolle, Adaptivität, Fehlertoleranz und Individualisierbarkeit) → Förderliche Arbeitsbedingungen (Handlungsräume und reichhaltige Arbeit, Lern- und Erfahrungsförderlichkeit, Kommunikation, Kooperation und soziale Einbindung)
IG Metall Vorstand (2019) [234]	→ Mensch-Technologie (Adaptivität, Transparenz, Komplementarität) → Mensch-Organisation (Ganzheitlichkeit, Polyvalenz, Akzeptanz und Partizipation) → Organisation-Technologie (Dezentrale Regelungskreise, Optimierung der Schnittstellen)
The AI Methods, Capabilities and Criticality Grid [47]	→ Methoden der KI-Komponenten → Fähigkeiten der KI-Komponenten → Gestufte Taxonomie einer allgemeinen Risikoeinschätzung in Bezug auf die Systeme
ISO/IEC 12792 [238]: „Transparency taxonomy of AI systems“-Projekt	→ Basisinformationen → Organisationsprozess → Anwendbarkeit der KI → Technische Informationen → Qualität- und Leistungsfähigkeit
The Fairness Handbook [224]	→ KI-Folgen- und Risikoabschätzung → Analyse der Stakeholder und betroffenen demografischen Gruppen → Fairness-Definition & -Metriken → Soziotechnische Kontextuntersuchung → Bias-Analyse

Aus den identifizierten Gestaltungsdimensionen resultieren dann verschiedene Normen und Standards, welche für die Planung und Gestaltung heranzuziehen sind.

Dimensionen der Gestaltung eines KI-Systems aus Ergonomie-/Human-Factors-Sicht beziehen sich bei soziotechnischen Systemen
→ einerseits auf Elemente des Systems (vgl. Kapitel 4.4.2.3) und
→ andererseits auf Mensch-Technik-Interaktionen (vgl. Kapitel 4.4.2.3).

Diese werden in den folgenden Kapiteln näher beleuchtet.

Gestaltungsanforderungen und -empfehlungen aus Ergonomie/Human Factors beziehen sich dabei derzeit vorwiegend auf statische und stationäre technische Systeme und Anlagen. Vorliegende Anforderungen bzw. Gestaltungsdimensionen sind einerseits für neue Technologien (z. B. KI-Systeme) nicht ausreichend beschrieben. Andererseits sind zusätzliche Anforderungen etwa durch inhaltlich und zeitlich dynamische Systeme (wie z. B. KI, aber auch bereits einfache mobile Maschinen) bisher nur unzureichend dokumentiert. Wesentliche Dimensionen sind dabei u. a.:
→ Digitalisierung (z. B. digitale Repräsentationen realer Lösungsmengen)
→ Vernetzung (z. B. Variabilität der Zugriffsbreite)
→ Dynamisierung (z. B. zeitliche und inhaltliche Veränderlichkeit)
→ Ambiguität (z. B. Unbestimmtheit des Lösungsraums)
→ Autonomiegrad des KI-Systems (vgl. Kapitel 4.1.2.2).

(Soziotechnisches) System analysieren, in dem die KI eingesetzt werden soll

Die Elemente des soziotechnischen Systems stellen bei der Nutzung oder beim Einsatz z. B. eines KI-Systems jeweils wirksame Ausführungsbedingungen für eine Aufgabenbearbeitung durch einen Menschen dar und können auch als Taxonomie verschiedener Konstellationen von z. B. KI-Systemen herangezogen werden. Daher ist bei der Planung und Gestaltung des KI-Systems zwingend eine Analyse des zugrunde liegenden soziotechnischen Systems durchzuführen.

Im Arbeitskontext handelt es sich bei dem soziotechnischen System um das „Arbeitssystem“ (vgl. [199]). Die DIN EN ISO 6385:2016 [235] definiert das Arbeitssystem als „System, welches das Zusammenwirken eines einzelnen oder mehrerer Arbeitender/Benutzer*innen mit den Arbeitsmitteln umfasst, um die Funktion des Systems, innerhalb des Arbeitsraumes und der Arbeitsumgebung unter den durch die Arbeitsaufgaben vorgegebenen Bedingungen, zu erfüllen“.

Die DIN EN ISO 6385:2016 [235] legt Grundsätze der Ergonomie in Form von grundlegenden Leitlinien zur Gestaltung von Arbeitssystemen fest und definiert die dafür relevanten grundsätzlichen Begriffe. Ergonomie ist die „wissenschaftliche Disziplin, die sich mit dem Verständnis der Wechselwirkungen zwischen menschlichen und anderen Elementen eines Systems befasst, und der Berufszweig, der Theorie, Grundsätze, Daten und Verfahren auf die Gestaltung von Arbeitssystemen anwendet mit dem Ziel, das Wohlbefinden des Menschen und die Leistung des Gesamtsystems zu optimieren“ [239]. Da die Ergonomie gleichermaßen das Wohlbefinden des Menschen als auch die Leistung des Gesamtsystems berücksichtigt, beinhaltet eine ergonomiezentrierte analytische Betrachtungsweise auch das Berücksichtigen von Sicherheitsaspekten, d. h. von 1) „safety“ (unfallrelevante Ereignisse) sowie von 2) „security“ (angriffsrelevante Ereignisse). Im Hinblick auf „security“ wird allerdings dabei i. d. R. nur eine „inside-out“-Perspektive betrachtet, d. h. Bedrohungen, die durch das soziotechnische System selbst ausgehen (z. B. durch fehlende Qualifizierung der Menschen).

Die DIN EN ISO 6385:2016 [235] sieht im ersten Schritt eine Anforderungsanalyse zur Formulierung der Ziele vor. Hierauf aufbauend können sich die folgenden Gestaltungsfelder laut DIN EN ISO 6385:2016 [235] ergeben:
→ Gestaltung der Arbeitsaufgaben und Tätigkeiten
→ Gestaltung der Arbeitsorganisation
→ Gestaltung der Arbeitsumgebung
→ Gestaltung der Arbeitsmittel und Schnittstellen
→ Gestaltung des Arbeitsraumes und des Arbeitsplatzes

Die skizzierten Gestaltungsfelder lassen sich häufig auch auf andere Anwendungskontexte übertragen. Dies ist aber für den konkreten Anwendungsfall zu prüfen und bei Bedarf sind die Systemelemente und Gestaltungsfelder entsprechend anzupassen.

Generelle Prinzipien und Konzepte der Ergonomie, welche für die Gestaltung soziotechnischer Systeme herangezogen werden können, werden in der DIN EN ISO 26800:2011 [239] spezifiziert, nämlich insbesondere:

→ Prinzipien der Ergonomie
  - Menschorientierter Ansatz: Anpassung der Komponenten eines Systems an die Merkmale des Benutzenden unter Berücksichtigung von
    - Zielpopulation
    - Aufgabenorientierung
    - Umgebungskontext
  - Kriterienbasierte Bewertung: Bewertung der Anwendung von ergonomischen Kriterien

→ Konzepte der Ergonomie:
  - Gebrauchstauglichkeit
  - Zugänglichkeit/Barrierefreiheit
  - Systemkonzept
  - Belastungs-Beanspruchungs-Konzept

→ Ergonomieorientierter Gestaltungsprozess über den gesamten Lebenszyklus

Anhand der identifizierten Gestaltungsfelder können dann die anzuwendenden Normen und Standards sowie die relevanten Methoden und Instrumente abgeleitet werden. Zudem bestimmen die Gestaltungsfelder maßgeblich, welche Daten für die Phasen Gestaltung und den späteren Betrieb benötigt werden und wie die Qualitätsansprüche an diese sind.

Darüber hinaus sind ethische Aspekte bei der Planung und Gestaltung des soziotechnischen Systems stets zu beachten und für den gesamten Lebenszyklus des KI-Systems zu gestalten. Ethische Aspekte sind z. B. Transparency, Accountability, Privacy, Justice, Reliability und Sustainability (z. B. AI Ethics Label der AI Ethics Impact Group [231]; vgl. Kapitel 4.4.2.2). Das soziotechnische System muss in dieser Phase auch hinsichtlich der ethischen Aspekte analysiert werden, um die in der Phase Initialisierung (vgl. Kapitel 4.4.2.2) ermittelten Anforderungen weiter zu konkretisieren und adäquat in die Planung und Gestaltung einfließen zu lassen.

Die einschlägigen Normen bzw. Standards zu KI (z. B. ISO/IEC 22989 [16], ISO/IEC 42001 [27], DIN SPEC 92001 Reihe [162], [240], [117], ISO IEC 25059:2022 [35]), Ergonomie & Organisation (z. B. DIN EN ISO 6385:2016 [235], DIN EN ISO 26800:2011 [239], DIN EN ISO 9241 Reihe [514], DIN EN ISO 10075 Reihe [513], DIN EN ISO 27500:2017 [271], VDI/VDE-MT 7100 [241]) und Ethik (VDE SPEC 90012 [242], IEEE 7000 Serie [10], [11], [12], [13], ISO IEC/TR 24028 [28]) berücksichtigen die resultierenden Anforderungen aus der soziotechnischen Gestaltung eines KI-Systems i. d. R. noch nicht hinreichend und lassen oft die Wechselwirkungen zwischen Mensch, Technik und Organisation außer Acht. Daher sind diese zu überprüfen und bei Bedarf zu ergänzen.

Aufgabenteilung zwischen Mensch und KI sowie Interaktionsprozess definieren

Die Rolle des Menschen im KI-System variiert abhängig von der eingesetzten KI-Technologie. Aufgaben des Menschen im KI-System sowie hieraus resultierende Anforderungen und Qualifizierungsbedarfe lassen sich z. B. aus den drei Dimensionen der KI-Klassifikation (zur KI-Klassifikation vgl. Kapitel 4.1.1.1) ableiten, also:

→ KI-Methoden (Klassische Künstliche Intelligenz / Wissensrepräsentation und Inferenz / Maschinelles Lernen / Hybrides Lernen)

→ KI-Fähigkeiten (Wahrnehmen / Verarbeiten / Handeln / Kommunizieren)

→ Kritikalität (Kein oder geringes / Gewisses / Deutliches / Erhebliches / Unvertretbares Schädigungspotenzial)

Je nach Verwendungszweck des KI-Systems kann es zudem weitere Dimensionen geben, die sich auf die Rolle des Menschen im KI-System auswirken.

Grundsätzlich sind bezogen auf Mensch-Technik-Interaktionen in soziotechnischen Systemen drei hierarchisch strukturierte Schnittstellen mit jeweils darauf bezogenen Gestaltungsprinzipien von besonderem Interesse:

→ Aufgabenschnittstelle, z. B. nach DIN EN 614-2:2008 [181], Reihe DIN EN ISO 11064:2011 [243]

→ Interaktionsschnittstelle, z. B. nach DIN EN 894-1:2009 [244], DIN EN ISO 9241-11:2018 [245], DIN EN ISO 9241-110:2020 [246], ISO/IEC 29138-1:2018 [247]

→ Informationsschnittstelle, z. B. nach VDI/VDE 3850-1:2014 [248], ISO 9241-112:2017 [249]

Entsprechend lässt sich i. A. a. Hacker ([250], [251]) eine Hierarchie der Gestaltungsebenen der Mensch-Technik-Interaktion ableiten (zitiert bei Böde et al. (2013) [252]):

1. **Mensch-Technik-Interaktion (im engeren Sinne):** Die Mensch-Technik-Interaktion im engeren Sinne fokussiert auf die Gestaltung der Aufgaben-, Interaktions- und Informationsschnittstellen. Hierbei stehen die Konzepte von Ergonomie (vgl. [239]) sowie Usability/Gebrauchstauglichkeit und User Experience (vgl. [246]) im Fokus. Die Normenreihen DIN EN 614 [180], [181], [182], DIN 894 [515] sowie DIN EN ISO 9241 Reihe [514] präzisieren die zugrunde liegenden Prinzipien und Konzepte und geben

Hinweise für die Gestaltung von Mensch-Technik-Interaktionen im Bereich von Maschinen und Anlagen sowie von Konsumgütern.
Die Prinzipien der Aufgabengestaltung leiten sich vom Primat der Aufgabe aus Ergonomie/Human Factors ab ([181], [253], [254]) und beziehen sich auf Vollständigkeit, Handlungsspielraum, Bewertbarkeit, Abwechslung, Kompetenzbezug, Ergebnisbeitrag, Entwicklungsförderlichkeit, Kooperation (vgl. [255], [248]).
Die Grundsätze der Informationsdarstellung werden in DIN EN ISO 9241-112:2017 [249] erläutert, und zwar handelt es sich um Entdeckbarkeit, Ablenkungsfreiheit, Unterscheidbarkeit, eindeutige Interpretierbarkeit, Kompaktheit und (interne und externe) Konsistenz.
Als maßgebliche Interaktionsprinzipien nennt DIN EN ISO 9241-110:2020 [246] dabei die Aufgabenangemessenheit, die Selbstbeschreibungsfähigkeit, die Erwartungskonformität, die Erlernbarkeit, die Steuerbarkeit, die Robustheit gegen Benutzungsfehler und die Benutzerbindung.
Es ist im Rahmen der Normung zu prüfen, ob die existierenden Normen die neuen Anforderungen der KI-Systeme bereits angemessen abbilden oder entsprechend angepasst werden müssen. Normungsbedarfe können sich z. B. durch die dynamische Funktionsallokation und bezüglich erforderlicher Strategien zur Abwendung von Ironien der Automation ergeben.

2. **Funktionsteilung Mensch-Technik:**
Für die Gestaltung der Funktionsteilung zwischen Mensch und KI-System gilt grundsätzlich das Primat der (Arbeits-) Aufgabe, d. h. die Gestaltung der Aufgabe steht am Anfang des Gestaltungsprozesses und ordnet ihr die Gestaltung der Ausführungsbedingungen unter ([195], [205], [181]). Das Vorgehen zur Gestaltung von Arbeitsaufgaben ist in DIN EN 614-2:2008 [181] definiert. Die gewählte Funktionsteilung repräsentiert den Autonomiegrad des KI-Systems (zu den Automatisierungsgraden z. B. [256] sowie Kapitel 4.1.2.2). Die einschlägigen Normen sind dahingehend zu prüfen, ob diese die verschiedenen Autonomiegrade angemessen berücksichtigen.
Für die Funktionsteilung werden teilweise die MABA-MABA-Liste (= „men are better at“ – „machines are better at“) herangezogen, die ursprünglich von [257] entwickelt wurde (vgl. z. B. [258], [175]). In der Ergonomie/Human-Factors-Forschung ist dieser Ansatz bereits in den frühen 1960er-Jahren kritisiert und seitdem alternativ diskutiert worden [259]. Die Manifestation einer festen Funktionszuteilung zwischen den Subsystemen Mensch und Technik greift zu kurz, da sie (1) ein mechanistisches Zusammenwirken von Faktoren bzw. Subsystemen postuliert, (2) Fertigkeiten, Fähigkeiten und Wissen des Subsystems Mensch pauschalisiert und in ihrer tatsächlichen Tiefe und Interaktionsleistung nicht berücksichtigt, (3) Dynamik und Weiterentwicklung der Subsysteme nicht berücksichtigt, (4) die Lebenszyklusperspektive für beide Subsysteme nicht berücksichtigt und (5) die Zielsetzung der Systemgestaltung verfehlt, dessen Erfolg allenfalls auf einer komplementären Ergänzung beruhen kann [259], [260], [261]. Eine wissenschaftliche Aufarbeitung einer komplementären Ergänzung der Subsysteme Mensch und KI steht noch aus.
Darüber hinaus kann sich die Funktionsteilung im Laufe der Nutzung abhängig von der Situation dynamisch anpassen (z. B. in Entscheidungs- oder Gefahrensituationen, bei denen der Mensch übernehmen muss). Diese Adaptivität ist aktuell noch nicht in der Normung abgebildet (und fehlt allgemein für automatisierte Systeme, nicht nur KI-Systeme).
Zudem können im Kontext der dynamischen Funktionsallokation die sogenannten „Ironien der Automatisierung“ zum Tragen kommen (vgl. [262]), die aufzeigen, dass mit der Automatisierung die Systemkomplexität steigt und deshalb neue Aufgaben der Überwachung, Steuerung und Korrektur entstehen, für die die menschlichen Qualifikationen oft nicht ausreichen. Dies muss bei der Gestaltung der Funktionsteilung und Automatisierung berücksichtigt werden und in die relevanten Normen und Standards einfließen.
Zu prüfen ist auch, wie mit dem Konzept der Individualisierbarkeit im Kontext der KI-Systeme (bzw. generell bei automatisierten Systemen) umzugehen ist. Dies ist momentan nicht in der Normung abgebildet.
Schließlich spielt bei der Gestaltung der Funktionsteilung auch das zugrunde liegende Leitbild der Technikgestaltung eine wesentliche Rolle: Wird der Mensch vom Entwicklenden als Fehlerquelle gesehen, so wird die Gestaltung tendenziell versuchen, den Einfluss des Menschen im KI-System weitgehend zu reduzieren. Wird das KI-System hingegen als Unterstützung für den Menschen betrachtet, so wird die Funktionsteilung eher komplementär erfolgen. Dies sollte somit zu Beginn des Gestaltungsprozesses kritisch hinterfragt werden – ein Hinweis hierzu sollte in die Normung aufgenommen werden. Im Interesse einer menschenzentrierten KI-Nutzung sollte dem Leitbild einer komplementären Funktionsteilung der Vorzug gegeben werden.

3. **Voraussetzungen und Folgen der Mensch-Technik-Interaktion:**
   Schließlich gilt es auch, die Organisation und Prozesse zu gestalten, in der das KI-System eingebettet ist. Hierbei sind verschiedenste Aspekte zu berücksichtigen, wie z. B.
   - Systemvertrauen („Trust in Automation“)
   - Belastung und Beanspruchung durch die Nutzung des KI-Systems (z. B. Technikstress), vgl. DIN-EN-ISO-10075-Reihe [513]
   - systemische Effekte (beispielsweise Aufschaukelungseffekte durch den Eingriff des Menschen in das KI-System)
   - veränderte Risikokompensation des Nutzenden sowie deren Folgen beim (unbemerkten) Ausfall des Systems
   - Veränderungen des Verhaltens des Nutzenden (z. B. bezüglich Kommunikation oder Kompetenz) und deren Folgewirkungen
   - der Kulturbegriff im zugrunde liegenden soziotechnischen System
   - Fragen der Verantwortung und Haftung

Bei vielen genannten Aspekten spielt die Qualifizierung der Nutzenden, die partizipative Gestaltung sowie ein geeignetes Change Management eine entscheidende Rolle. Es ist zu prüfen, inwieweit diese Aspekte in den relevanten Normen und Standards (z. B. DIN EN ISO 27500:2017 [271], VDI/VDE-MT 7100 [241], DIN EN ISO 9001:2015 [263]) abgebildet sind.

KI-Lösung feinplanen
Nachdem die Rahmenbedingungen für den Einsatz der KI-Lösung in den vorangegangenen Schritten geklärt wurden, müssen nun alle Details für den Einsatz der KI-Lösung feingeplant werden, wie z. B. die Auswahl der einzusetzenden Technologie und Arbeitsmittel, prospektive Gefährdungsbeurteilung der KI-Lösung, Schaffung der erforderlichen Rahmenbedingungen im Unternehmen, Qualifizierungsmaßnahmen.

Die Vorbereitung für den betrieblichen Einsatz erfordert i. d. R. ein spezifisches Training der KI-Lösung. Hierbei gilt es, besonderes Augenmerk auf die Auswahl der Trainings, Validierungs- und Testdaten zu legen, um Diskriminierung etc. zu vermeiden (vgl. Kapitel 4.4.2.2). Darüber hinaus sind die zu verwendenden Daten auf ihre Qualität hinsichtlich des geplanten Einsatzes zu überprüfen (z. B.: Genug Daten? Inkonsistente Daten? Zu alte, zu neue Daten? Falsche Daten?). Zudem sind Auswahl der verwendeten Datensätze sowie das Training, die Verifizierung, die Validierung und die Testung der KI-Lösung adäquat zu dokumentieren.

Diese Phase ist stets spezifisch für das jeweilige Projekt zu klären und kann daher verschiedenste Aspekte umfassen. In der Regel kommen hier die spezifischen Produktnormen bezogen auf die eingesetzte Technik zum Tragen.

Darüber hinaus können auch einschlägige Prozessnormen relevant sind, z. B. für die Gestaltung der Organisation [271], Qualitätsmanagementsysteme (DIN EN ISO 9000 ff. [264], [263]), Umweltmanagementsysteme [265], Energiemanagementsysteme [266], Managementsysteme für Sicherheit und Gesundheit bei der Arbeit [267]. Es ist zu prüfen, ob die relevanten Normen bereits den Einsatz von KI-Lösungen ausreichend betrachten oder diesbezüglich zu ergänzen sind.

Inbetriebnahme planen
Nach Abschluss der Feinplanung der KI-Lösung muss die Inbetriebnahme geplant werden, also das erstmalige Betreiben der KI-Lösung. Dies beinhaltet typischerweise die Terminplanung sowie die Gestaltung des Kommunikationsprozesses mit den Beteiligten, Evaluations-, Feedback- und Schlichtungsmechanismen für die Inbetriebnahme. Hierfür ist i. d. R. das Projektmanagement entscheidend (z. B. DIN ISO 21500:2016 [268], Reihe DIN 69901 [269], Reihe DIN 69909 [270]. Es ist zu prüfen, ob KI-Projekte hinsichtlich des Projektmanagements Besonderheiten aufweisen, die bei Bedarf in der Normung abzubilden sind.

Darüber hinaus spielen in diesem Kontext die Prozessnormen erneut eine große Rolle.

Regelbetrieb planen
Abschließend muss auch der Regelbetrieb geplant werden. Diese Planung basiert auf den Ergebnissen der KI-Feinplanung sowie der Planung der Inbetriebnahme und setzt somit auch die dort definierten Werkzeuge, Methoden und Prozesse ein. Weitere Planungsaspekte für den Regelbetrieb können z. B. Aspekte des operativen Personaleinsatzes, die Einrichtung eines Beteiligungsprozesses zur kontinuierlichen Verbesserung oder das Monitoring der aktuellen Entwicklung der eingesetzten KI-Technologie hinsichtlich Änderungsbedarfe sein. Die konkreten Planungsbedarfe hängen stark vom jeweiligen Anwendungsfall ab.

Für die Planung des Regelbetriebs bilden die einschlägige Prozessnormen eine wichtige Grundlage, z. B. für die Gestaltung der Organisation [271], Qualitätsmanagementsysteme [264], Umweltmanagementsysteme [265], Energiemanagementsysteme [266], Managementsysteme für Sicherheit und Gesundheit bei der Arbeit [267]. Diese Prozessnormen

berücksichtigen die besonderen Anforderungen beim Einsatz von KI-Lösungen meist noch nicht hinreichend und sind daher zu ergänzen, insbesondere im Hinblick auf die sozio-technischen Aspekte.

Methodeneinsatz bei der Gestaltung und Planung

Die Normungsroadmap Künstliche Intelligenz hat nicht den Anspruch, eine Übersicht über mögliche Methoden zur Planung und Gestaltung von KI-Lösungen zu geben, da diese stets fallspezifisch zu wählen sind. Es ist daher während der Planung und Gestaltung immer zu prüfen, was für den konkreten Planungsfall Stand der Technik ist, und dies entsprechend zu berücksichtigen. Die resultierenden Normen und Standards sind zu befolgen.

Bei der Planung und Gestaltung von KI-Lösungen sind z. B. Methoden erforderlich für:

- → Prozess zur Gestaltung von KI-Systeme
- → Technische Ausgestaltung des KI-Systems
- → Ergonomische Ausgestaltung des KI-Systems
- → Gestaltung der Schnittstellen
- → Partizipation Betroffener; Unterstützung der Deliberation der Beteiligten über den Inhalt durch Prozesse
- → Technikfolgeabschätzung, Schadenanalyse und Gefährdungsbeurteilung
- → Evaluation, Feedback und Schlichtung
- → Qualitäts- bzw. Ergebniskontrolle der KI-Lösung
- → Projektmanagement
- → Kommunikation
- → Qualifizierung, Kompetenzentwicklung, Change Management
- → Dokumentation des Planungs- und Gestaltungsprozesses

Beteiligte

Grundsätzlich wäre die idealtypische Forderung bei der Planung und Gestaltung einer KI-Lösung, eine Vertretung aller Stakeholder zu beteiligen. Die ISO/IEC 22989:2022 [16] unterteilt sogenannte Stakeholder in „AI-Provider“, „AI-Producer“, „AI-Customer“, „AI-Partner*innen“, „AI-Subject“ und „Relevant Authorities“.

Konkret können z. B. die folgenden Personen einzubinden sein:

- → Expert*innen mit Domänenwissen (KI-Expert*innen, Data Scientists, Informatiker*innen usw., Prozessgestaltende, Usability- Expert*innen, Produktgestaltende usw., Softwaretester*innen, Ergonomen, Psycholog*innen usw., Sicherheitsexpert*innen in der jeweiligen Domäne, Expert*innen für Ethik, Diversity, Fairness usw.),
- → im Unternehmen: Expert*innen aus den betroffenen Fachabteilungen,
- → Nutzende des KI-Systems,
- → Interessensvertretungen von Betreibenden und Nutzenden,
- → Entscheider über den Einsatz der KI-Lösung,
- → Vertreter*innen der Zivilgesellschaft,
- → und sonstige Perspektiven.

Art, Inhalt und Form von Kommunikation und Beteiligung sind dabei abhängig vom jeweiligen Zeitpunkt, bezogen auf den Projektlebenszyklus, insbesondere

- → während der Zielfindung,
- → während der Planung und Gestaltung,
- → während der Inbetriebnahme,
- → im laufenden Betrieb bzw. im kontinuierlichen Verbesserungsprozess.

Doch nicht nur der Zeitpunkt der Interaktion ist relevant, auch ist zu berücksichtigen, welche Stakeholder bei der Interaktion beteiligt sind. Hier ist darauf zu achten, dass die Kommunikation immer zielgruppengerecht und inklusiv erfolgt. Unterschiede ergeben sich z. B. bei der Kommunikation

- → zwischen den Expert*innen mit Domänenwissen untereinander,
- → zwischen den Expert*innen mit Domänenwissen und den Nutzenden,
- → zwischen Nutzenden und Technik,
- → zwischen Expert*innen und sonstigen Beteiligten,

Vor diesem Hintergrund müssen die Prozesse der Kommunikation und Beteiligung entsprechend geplant und methodisch durchgeführt werden. Einschlägige Normen und Standards (z. B. VDI-MT 7001:2021 [241]) können hierbei unterstützen. Es ist zu prüfen, ob die relevanten Normen bereits den Einsatz von KI-Lösungen ausreichend betrachten oder diesbezüglich zu ergänzen sind.

Zudem können Good-Practice-Beispiele oder Experimentierräume unterstützen (z. B. der KIDD-Prozess (2022) [74], moderierte Spezifikationsdialoge zwischen arbeitsweltlichem Erfahrungswissen und softwaretechnischem Fach- und Sachwissen [272].

4.4.2.4 Betrieb

Dieses Kapitel befasst sich mit den Phasen „Überführung in die Einsatzumgebung“, „Betrieb und Überwachung“, „Kontinuierliche Validierung“ und „Reevaluierung“ der ISO/IEC 22989:2022 [16].

Ziele der Phase aus soziotechnischer Sicht
Aus soziotechnischer Sicht ist das Ziel, in dieser Phase sicherzustellen, dass die in den Phasen Initialisierung sowie Planung und Gestaltung festgelegte gewünschte Funktionsweise eingehalten wird, sowie in regelmäßigen Abständen zu entscheiden, ob die gewünschte Funktionsweise geänderten Rahmenbedingungen angepasst werden muss. Dazu sollten im Betrieb Echtdaten (je nach Anwendungsfeld anonymisiert oder pseudonymisiert) gesammelt und für die relevanten Akteur*innen im soziotechnischen System verständlich und transparent aufbereitet werden. Auf dieser Grundlage kann ein kontinuierlicher Verbesserungsprozess stattfinden und Personen, die mit dem System interagieren, erhalten eine belastbare Basis für eine informierte Entscheidung über einen möglichen Eingriff oder andere notwendige Maßnahmen. In eine kontinuierliche Evaluierung und Anpassung des Systems sollten die Betroffenen (im Unternehmenskontext z. B. die Beschäftigten) eingebunden sein und ihre Erfahrung mit dem System sollte Grundlage und Ausgangspunkt für Verbesserungen sein [203].

Für dieses Monitoring sind technische Lösungen erforderlich, die im Sinne eines Transparency-by-Design- bzw. Transparency-by-Default-Ansatzes jederzeit den nötigen Überblick ermöglichen. Dies kann entweder in einem Modul innerhalb des KI-Systems erfolgen oder durch ein eigenständiges Command Tool.

Begleitend ist es wichtig, im Rahmen von Trainings die mit dem KI-System interagierenden Personen zu fachlichen und KI-spezifischen sowie interaktionsbasierten Inhalten (z. B. den Auswirkungen von Over-Reliance oder Under-Reliance) zu schulen.

Beteiligte Akteur*innen im Betrieb des soziotechnischen Systems und ihre Bedürfnisse
Menschen nehmen während des Betriebs des soziotechnischen Systems unterschiedliche Rollen ein, z. B.:

- → das Management einer Organisation, in der ein solches System im Einsatz ist;
- → Personen in beteiligten Fachstellen; Betriebsräte und Betriebsrätinnen sowie andere Vertreter*innen von Mitarbeitendenrechten;
- → Nutzende des soziotechnischen Systems;
- → an der Entwicklung und Weiterentwicklung beteiligte IT-Fachleute;
- → Interessierte aus der betroffenen oder der breiten Öffentlichkeit.

Eine besondere Bedeutung kommt den Rollen zu, die im Sinne der High Level Expert Group (HLEG) die menschliche Aufsicht gewährleisten:

- → dem oder der Human-in-the-Loop (HITL, im Entscheidungszyklus der KI involviert),
- → dem oder der Human-on-the-Loop (HOTL, beim Design der KI und im Monitoring involviert) sowie
- → dem oder der „Human in Command“ (HIC, soll die Gesamtaktivität inklusive breitere ökonomische, soziale, rechtliche und ethische Auswirkungen überblicken können).

Die Rolle des oder der HIC wird im Vorschlag der EU-Verordnung KI eingebracht. HIC soll insbesondere für Hochrisikosysteme, im Grunde aber für jede KI unabhängig von ihrer Kritikalität über geeignete Interventionsmöglichkeiten verfügen, d. h. z. B. in der Lage sein, eine „Stopptaste“ für die KI betätigen zu können [273]. Mit der Forderung nach einer „Stopptaste“ ist nicht gemeint, eine laufende KI-Prozedur bei aufkommenden Zweifeln zu unterbrechen, sondern die Möglichkeit, einer durch KI getroffenen Entscheidung nicht zu folgen oder die KI-Nutzung für einen bestimmten Zeitraum auszusetzen und stattdessen Menschen entscheiden zu lassen.

Bei der Nutzung von KI in soziotechnischen Systemen sollten also Interventionen von Menschen vorgesehen werden. Diese könnten beispielsweise darin bestehen, dass Menschen Ausnahmen von den Entscheidungen der KI treffen oder dass sie Parameter des Systems (Schwellenwerte, Eingangsgrößen) rekonfigurieren können. Beides ist denkbar als unmittelbare Intervention durch die Nutzenden oder alternativ nach Hinzuziehung von autorisierten Personen im Betrieb. Während beim Ansatz „keep the human in the loop“ einzelne Individuen im Verhältnis zur KI betrachtet werden, gibt es auch das Gestaltungsprinzip „keep the organization in the loop“. Damit ist gemeint, dass beim Einsatz von KI auch die Interaktion der relevanten Stakeholder betrachtet und laufend optimiert werden sollte [274].

Um ihre jeweiligen Rollen auszufüllen bzw. ihre Bedürfnisse in Bezug auf die Transparenz eines KI-Systems zu befriedigen, benötigen diese genannten Akteur*innen zielgruppenorientierte technische Lösungen – sozusagen Editionen eines Transparency bzw. Command Tools. Diese Editionen müssen sich in vier Punkten unterscheiden:

→ In der Informationstiefe und -darstellung: für die breite Öffentlichkeit anders als für das Management oder für HICs.
→ In den Einflussmöglichkeiten: Ein HIC oder eine HIC benötigt laut Verordnungsentwurf KI für Hochrisikosysteme die Möglichkeit, „in den Betrieb des Hochrisiko-KI-Systems einzugreifen oder den Systembetrieb mit einer „Stopptaste“ oder einem ähnlichen Verfahren zu unterbrechen.“ Personen aus dem Management können z. B. das Recht benötigen, Zielvariablen zu verändern.
→ In den Feedbackmöglichkeiten: bei einer (vermuteten) Fehlleistung des KI-Systems.
→ Und schließlich im Qualifizierungskonzept, das notwendig ist, um die Rolle im soziotechnischen System auszufüllen und eine Transparency- bzw. Command-Funktionalität zu verwenden und so ein soziotechnisches System zu steuern und zu beaufsichtigen.

Der Aspekt der Organisation im Betrieb eines soziotechnischen Systems

Die Einführung von KI-Systemen im Arbeitsprozess bedeutet für die einführende Organisation und deren Akteur*innen immer Veränderung. Durch die Einführung von KI können neue Kompetenzen erforderlich, aber auch vorhandene Kompetenzen entwertet werden. Zugleich können sich Aufgaben, Rollen und Kooperationszusammenhänge verändern. Vor diesem Hintergrund wird es erforderlich, frühzeitig Veränderungswiderstände zu erkennen und darauf einzugehen. Um den Change-Prozess erfolgreich zu gestalten, ist es wichtig, Beteiligte und ihre Interessensvertretungen zu informieren, Partizipation zu ermöglichen sowie Transparenz und Einflussmöglichkeiten zu schaffen. Eine begleitende Organisationsentwicklung ist daher bereits in der Initiierung und der Planung und Gestaltung eines soziotechnischen Systems wichtig, damit dieses im Betrieb die intendierte Wirkung entfalten kann. In den Betriebsphasen Überführung in die Einsatzumgebung, Betrieb und Überwachung, kontinuierliche Validierung und Reevaluierung kommt der begleitenden Organisationsentwicklung ebenfalls eine entscheidende Rolle zu.

Monitoring im Betrieb: Transparenz und Eingriffsmöglichkeiten für Menschen im soziotechnischen System

Die Forderung nach Transparenz für Beteiligte, die sich aus der Organisationsentwicklung und auch aus dem Entwurf der EU-Verordnung KI ergibt, umfasst eine Reihe von Bausteinen: zum einen die Transparenz über die definierten Ziele und die intendierte Funktionsweise, also das Narrativ hinter dem soziotechnischen System [64]. Zum zweiten Governance-Gesichtspunkte, also die Einordnung in die Risikomatrix sowie festgelegte Verantwortlichkeiten, definierte Kompetenzen und Rechte entlang der „Langen Kette der Verantwortlichkeiten“ [63]. Auch die Ergebnisse aus dem Planungs- und Gestaltungsprozess sowie Erklärungen zur Wahl von bestimmten Kalibrierungen der Hyperparameter und Evaluationskriterien transparent und verständlich darzustellen ist wichtig für das Monitoring im Betrieb.

Ein Beispiel dazu: Ob die Ergebnisse und die Funktionsweise eines soziotechnischen Systems im Betrieb als fair eingestuft werden können, hängt davon ab, welche Fairnessaspekte im Kontext relevant sind. Bei einem KI-System, das Bewerberinnen und Bewerber für Jobs vorschlägt, könnte z. B. die Verteilung der unterschiedlichen Geschlechter bei den Vorschlägen ein Fairness-Aspekt sein. Gemäß dem WKIO-Modell wäre also ein Wert „Fairness in Bezug auf das Geschlecht“, das dazugehörende Kriterium könnte „Prozentualer Anteil von Männern und Frauen sowie anderen Geschlechtern“ sein. Ein möglicher Indikator wäre, dass bezogen auf die Verteilung der Geschlechter unter gleich bzw. ähnlich qualifizierten Bewerbenden (es bewerben sich z. B. 30 % Frauen auf diese Stelle) ein als fair empfundener Anteil (z. B. 25 bis 35 % Frauen) auch auf der Vorschlagsliste landet.

Um nun bei der Nutzung eines KI-Systems das Monitoring durchführen zu können, sind die Entscheidungen aus den früheren Phasen wichtige Grundlagen, die es gilt, transparent zu haben:

→ Warum wurden welche Ziele gewählt (z. B. Fairness in Bezug auf Geschlecht)?
→ Warum wurden welche Zielkorridore gewählt (z. B. 25 bis 35 % Frauen in der Job-Shortlist)?
→ Welche Autonomiestufen gibt es, wie sind sie definiert und warum – also: Ab welcher Abweichung vom Zielkorridor ist mehr menschlicher Eingriff oder sogar ein Aussetzen des KI-Systems nötig?
→ Findet automatisiert einen Wechsel in eine niedrigere oder höhere Autonomiestufe statt?
→ Welche Teile sind als zum KI-System zugehörig definiert und müssen z. B. im Fall von Hochrisikosystemen ganz

ausgeschaltet werden können – und welche Teile des Systems könnten weiter in Betrieb bleiben?
→ Falls eine komplette Aussetzung des KI-Systems ermöglicht wird – also eine „Stopptaste“ integriert ist: Wie kann diese in den verschiedenen Anwendungsfällen überhaupt aussehen?
→ Kann das soziotechnische System seine Funktion weiterhin erfüllen, wenn die Stopptaste gedrückt wird? Falls ja, unter welchen Voraussetzungen und Rahmenbedingungen?

Der Normung kommt die wichtige Aufgabe zu, einen Rahmen für die Klärung dieser Fragen zu definieren.

Zusätzlich zur Transparenz über die definierten Ziele, die Gestaltungsentscheidungen und deren Hintergründe ist nun im Betrieb die Transparenz darüber nötig, wie das soziotechnische System sich tatsächlich im Betrieb verhält, also die systematische Auswertung der Leistung und Risiken des Systems im Feld, wie sie auch im EU AI Act, Art. 61 gefordert wird. Wichtige Fragestellungen sind:
→ Welche Inputdaten werden verwendet?
→ Wie verändern sich diese über die Zeit?
→ Ist die Qualität in Bezug auf die in der Planung und Gestaltung festgelegten Maßstäbe noch hoch genug?
→ Wo liegen die konkreten Messwerte in Bezug auf die Zielkorridore, also im Beispiel: Wie hoch ist der prozentuale Anteil eines bestimmten Geschlechts in der Vorschlagsliste wirklich?
→ Wie verändert sich dies über die Zeit?

Je nach Zielgruppe und Informationstiefe ist hier ein einfaches Ampelsystem hilfreich bzw. sind tiefergehende Informationen notwendig. Ein intuitives und nutzerfreundliches Oberflächendesign und eine für die jeweilige Zielgruppe verständliche Aufbereitung sind dabei erfolgskritisch.

Dem Themenbereich XAI (Explainable AI, Erklärbarkeit des KI-Systems) kommt in diesem Kontext eine wichtige Bedeutung zu. Dabei geht es um diese Fragestellungen: Aufgrund welcher Inputs werden welche Outputs generiert? Welche Aspekte können a priori festgelegt werden, welche mit Erklärbarkeitsmetriken identifiziert und welche ex post über Zielkorridore erhoben werden?

Transparente Informationen zu Zielen und tatsächlichen Messwerten sind notwendige Voraussetzungen zum Monitoring eines soziotechnischen Systems im Betrieb. Es gibt weitere Aspekte, um auch hinreichende Voraussetzungen zu erfüllen: Wie wird etwa mit dem Dilemma of Automation umgegangen? Sprich: Kann der Mensch überhaupt entscheiden, ob ausgeschaltet werden soll? Kann der Mensch die Aufgabe ohne KI noch erfüllen? Welche Qualifikation ist nötig, um ihn zu befähigen?

Das Vertrauen von Menschen in die Mensch-Maschine-Interaktion und das soziotechnische System insgesamt ist messbar. Es gibt verschiedene Möglichkeiten, solche Messungen vorzunehmen. Sie unterscheiden sich wesentlich hinsichtlich der Zeitigkeit. Ziel muss es sein, zu erkennen, wann zu wenig (under-reliance) oder übermäßiges Vertrauen (over-reliance) des Menschen in den automatisierten Prozess oder die Entscheidung vorliegt, um so anschließende Maßnahmen einleiten zu können: „Die Benutzer*innen müssen in der Lage sein, eine klare Zuordnung zwischen den über die Schnittstelle dargestellten Systemfunktionen und ihren Zielen vorzunehmen.“ [275].

Schulungen
Unterschiedliche Akteur*innen im soziotechnischen System haben spezifische Bedürfnisse in Bezug auf Schulungen: Zum einen können Schulungen in Bezug auf das technische KI-System nötig sein. Die Art und Weise, wie Menschen mit KI interagieren, kann ebenfalls ein wichtiges Thema sein: Ist im Kontext Over-Reliance, Under-Reliance oder beides zu erwarten? Welche organisatorischen und gesellschaftlichen Aspekte sollten miteinfließen?

Einen Überblick über die KI-Kompetenzen und ihre Entwicklung gibt Abbildung 33.

Die Entwicklung von Kompetenzen gilt als Schlüsselfaktor für eine erfolgreiche Implementierung eines KI-Systems. Schulungen zur Kompetenzentwicklung der Beschäftigten sollten passgenau auf den Wissensstand der Beschäftigten und die unternehmerischen Ziele ausgerichtet werden. Um dies zu erreichen, muss im ersten Schritt festgestellt werden, welche (Job-)Rollen sich im Betrieb im Kontext der KI ergeben. Diese müssen dann als Aufgaben formuliert werden, um daraus die notwendigen KI-Kompetenzen abzuleiten, siehe Abbildung 34. Diese Kompetenzprofile sollten dann den jeweiligen (Job-)Rollen zugeordnet werden. Um den Wissensstand der Beschäftigten zu berücksichtigen, sollte, basierend auf dem erforderlichen Kompetenzprofil, für ihre (Job-)Rolle eine individuelle Kompetenzbedarfsanalyse durchgeführt werden, woraus sich dann der individuelle Bedarf für Weiterbildungsmaßnahmen ableiten lässt [190].

Aufgaben	Cluster	Kompetenz
Aufgabe 1	Anwendung von Fach- und Grundwissen	• Fachkompetenz • Grundlegende digitale Kompetenzen • Grundwissen: Maschinelles Lernen
Aufgabe 2	Umgang mit KI-Systemen	• MMI-Kompetenzen • Prozess- und Systemkompetenzen • Problemlösungskompetenz, Resilienz • Reflexionskompetenz
Aufgabe 3	Gestaltung von Arbeitsprozessen	• Selbstkompetenzen • Soziale und Kommunikationskompetenz • (Personal-)Management, Führungskompetenz, Change-Management • Entscheidungskompetenz • Anpassungsfähigkeit, Transfer • Organisatorische Kompetenzen • Strategische Kompetenzen
Aufgabe 4		
...		

Prozess der Kompetenzentwicklung: Ableitung von Kompetenzen aus den (rollenspezifischen) Aufgaben

Abbildung 33: Prozess der Kompetenzentwicklung und Systematisierung von KI-Kompetenzen (Quelle: in Anlehnung an [190])

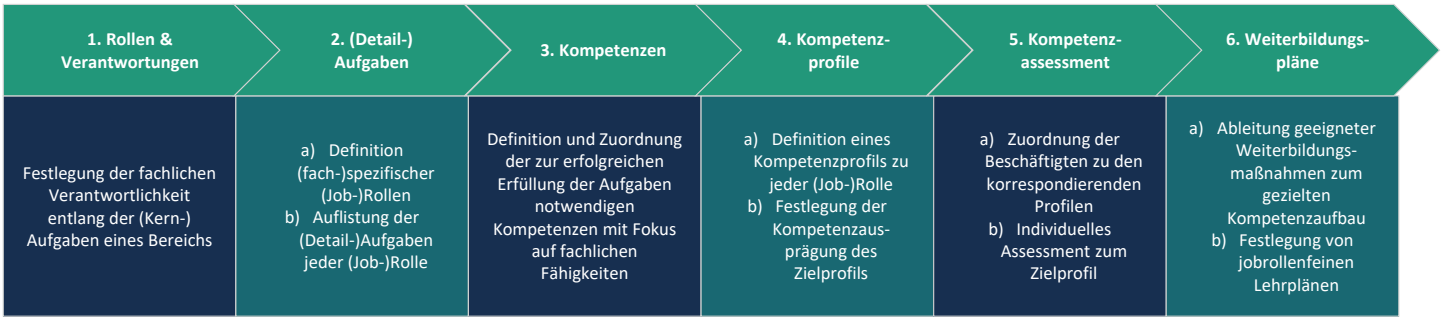


Abbildung 34: Schritte eines aufgabenorientierten Kompetenzmanagementprozesses (Quelle: in Anlehnung an [190])

Iterativer Prozess: Continuous Validation, Re-Evaluation und Continuous Improvement

Bei der Planung eines soziotechnischen Systems werden auch unter Berücksichtigung von Folgeabschätzungen Ziele und Maßnahmen definiert. Aber nicht alle Entscheidungen können a priori getroffen werden, da oft nicht alle erforderlichen Informationen vorhanden sind und Rahmenbedingungen sich über die Zeit ändern. Zudem könnten nicht intendierte Effekte eintreten oder es könnte sich herausstellen, dass geplante Maßnahmen unzureichend oder unvollständig waren. Eine Veränderung der zugrunde gelegten Datenlage im operativen Betrieb eines KI-Systems gegenüber der Datenlage zum Zeitpunkt der KI-Systemerstellung (Trainings- und Testdaten) kann durch Methoden im Themenaspekt „Drift" (Concept Drift, Data Drift ...) erkannt werden und sollte Standardfunktion sein.

Aus den bereits genannten Gründen ist eine kontinuierliche Validierung und Evaluation der Ziele und Gestaltungsentscheidungen in Bezug auf das KI-System notwendig und gemäß AI Act, Art. 61 für Hochrisikosysteme sogar verpflichtend. Bei einer Validierung sind die entscheidenden Fragen jeweils neu zu beantworten:

→ Sind weitere / andere Ziele zu berücksichtigen?
→ Ist durch die vorhandenen Ziele und Korridore eine korrekte Funktionsweise noch sichergestellt?
→ Oder kann auch bei eingehaltenen Zielen/Korridoren ein Problem mit dem KI-System vorliegen?
→ Falls ja, wie kann dieses detektiert werden?
→ Muss das der Mensch machen, kann er das überhaupt oder welche Unterstützung ist dafür notwendig?
→ Wird ein Korridor ständig in eine Richtung ausgenutzt und welche „Near Misses" („Fast"-Ausfälle) entstehen?

Diese „Near Misses" sind oft häufiger als wirkliche Ausfälle und geben wertvolle Einblicke in das System, wenn es an seinen Grenzen betrieben wurde. Die Einführung von Berichtsstrukturen über Ausfälle oder auch „Fast"-Ausfälle ermöglicht die Analyse und Verbesserung von KI-Systemen, um auch zukünftige Systemresilienz sicherstellen zu können oder Risiken zu simulieren [276]. Auch dies ist im geplanten AI Act, Art. 62 verpflichtend umzusetzen, zumindest bei Hochrisiko-KI-Systemen.

Da eine KI häufig im Zusammenspiel mit anderen Systemen oder anderen KIs eingesetzt wird, sind Integrationstests notwendig. Insbesondere vor der ersten Inbetriebnahme, aber auch bei Updates des Gesamtsystems sollte ein solcher Integrationstest durchgeführt werden. Hier ist das komplette KI-basierte System zu testen:

→ Entstehen Seiteneffekte, wenn die KI in das Gesamtsystem integriert wird (Datenformatierung, Timing)?
→ Gibt es Probleme mit der Bedienung (Usability) der KI im Gesamtsystem?
→ Hat die Integration Einfluss auf die Performanz des Gesamtsystems?
→ Hat die Integration Einfluss auf die Security des Gesamtsystems?

Eine weitere wichtige Grundlage für die iterative Überprüfung des soziotechnischen Systems sind Feedbacks. Diese können vom System aktiv angefordert („programmatisches Feedback", regelbasiert), durch einen Operator angefragt („triggered Feedback") [277] oder in Form einer Problemanzeige gemeldet werden. Der Folgeprozess (Wie und von wem wird das Feedback bewertet und welche Schritte werden daraus abgeleitet?) muss klar definiert sein.

Bei Softwareupdates sowie anderen Änderungen des KI-Systems ist ein erneuter Abgleich mit der festgelegten gewünschten Funktionsweise notwendig. Außerdem sollten Regressionstests hinsichtlich der Performanz, der Security und auch der Usability durchgeführt werden. Jede Änderung kann Seiteneffekte verursachen, die so nicht gewollt sind.

Idealerweise ist regelmäßig zu prüfen, ob es grundlegende Änderungen in der KI-Technologie gibt, die sich ggf. auf die eigene Lösung auswirken bzw. bessere Resultate erzielen können – oder ob es gar anstatt einer KI inzwischen andere Lösungen gibt, die das ursprüngliche Problem schneller oder besser lösen.

Im Anschluss an die Evaluation erfolgt eine Optimierung des soziotechnischen Systems. Diese kann einzelne Komponenten wie die eingesetzte KI, die Verbindungen von KI mit anderen IT-Systemen und Datenbeständen, Benutzeroberflächen oder Qualifizierungskonzepte umfassen. Charakteristisch für eine soziotechnische Betrachtung ist stets die ganzheitliche Sicht auf die Schnittstellen zwischen den Elementen Technik, Organisation und Mensch und Gesellschaft. Die kontinuierliche Verbesserung soziotechnischer Systeme zielt insbesondere auf diese Schnittstellen. Beispiele für soziotechnische Optimierungen wären veränderte Autonomiegrade oder angepasste menschliche Interventionsmöglichkeiten.

4.4.3 Normungs- und Standardisierungsbedarfe

Bedarf 04-01: Berücksichtigung der Dynamik von KI-Systemen bei der Gestaltung von Aufgaben-, Interaktions- und Informationsschnittstellen
Bei der Planung und Gestaltung von KI-Systemen sind die Umsetzung von ergonomischen Grundsätzen und Prinzipien sowie eine gebrauchstaugliche Gestaltung von Produkten und Arbeitsmitteln erfolgswirksame Ziele. Damit ist die Anwendung dieser ergonomischen Grundsätze und Prinzipien auch ein wesentliches Gütemerkmal von KI-Systemen als Arbeitsmittel bzw. Gebrauchsgegenständen.

Gestaltungskonzepte in Ergonomie/Human Factors (u. a. zur soziotechnischen Gestaltung) bezogen sich in der Vergangenheit vorwiegend auf statische technische Systeme (z. B. Schnittstellengestaltung zu statischer und stationärer Maschine). Nicht nur, aber auch durch KI (als inhaltlich und zeitlich dynamisches System mit nicht mehr dokumentierbaren Ursache-Wirkungs-Beziehungen) muss das EHF-Gestaltungskonzept erweitert werden, damit Dynamik von Schnittstellen, Funktionsweisen und Auswirkungen auch für Menschen passend gestaltet werden.

Die einschlägigen Normen zur Ergonomie (z. B. DIN EN ISO 6385:2016 [235], DIN EN ISO 26800:2011 [239], DIN-EN-ISO-9241-Reihe [514], DIN-EN-ISO-10075-Reihe [513], DIN-EN-614-Reihe [180], [181], [182], DIN EN 894-1:2009 [244], DIN EN ISO 11064:2011 [243]) berücksichtigen die resultierenden Anforderungen aus der soziotechnischen Gestaltung eines KI-Systems i. d. R. noch nicht hinreichend und lassen oft die Wechselwirkungen zwischen Mensch, Technik und Organisation im Zusammenspiel mit KI-Systemen außer Acht. Zudem werden Interaktionskonzepte und Anforderungen an die Informationsdarstellung derzeit nur unzureichend abgebildet für eigendynamische Systeme, für die eine kontinuierliche Aufgabenbearbeitung erforderlich ist und für die Steuerungseingriffe nicht rückgängig gemacht werden können.

Bedarf 04-02: Berücksichtigung soziotechnischer Aspekte bei der Gestaltung von KI-Systemen
Die Art und Weise des Arbeitens verändert sich mit der Einführung von KI-Anwendungen, die Anforderungen an dem Menschen ebenso. Bei der Einführung von KI-Systemen sind daher die Organisationsentwicklung, das Change Management sowie die Qualifizierung der Beteiligten wichtige Fragestellungen. Im Sinne einer soziotechnischen Systemgestaltung sind daher Technologieeinsatz und Organisation gemeinsam zu planen bzw. zu optimieren.

Einschlägige Prozessnormen, z. B. für die Gestaltung der Organisation (DIN EN ISO 27500:2017 [271]), Qualitätsmanagementsysteme (DIN EN ISO 9000:2015 [264]), Umweltmanagementsysteme (DIN EN ISO 14001:2015 [265]), Energiemanagementsysteme (DIN EN ISO 50001:2018 [266]), Managementsysteme für Sicherheit und Gesundheit bei der Arbeit (DIN ISO 45001:2018 [267]) berücksichtigen die besonderen Anforderungen beim Einsatz von KI-Lösungen meist noch nicht hinreichend und sind daher zu ergänzen, insbesondere im Hinblick auf die soziotechnischen Aspekte.

Bedarf 04-03: Erfüllung des Standardisation Requests zum EU AI Act, Aspekt „Transparenz“
Der Entwurf zur **KI-Verordnung der EU** (KI-VO) legt einen Fokus auf die soziotechnische Perspektive: Die Anforderung, Transparenz und Informationen für Benutzende zur Verfügung zu stellen, kann nur erfüllt werden, wenn das KI-System als soziotechnisches System verstanden und der Mensch als Teil des Systems mitgedacht wird.

Welche Transparenz in welchem Kontext für welche Zielgruppe ausreichend ist und welche Basisinformationen als Grundlage für menschliche Eingriffe ins System vorhanden sein müssen – das sind Fragestellungen, die nicht die KI bzw. KI-Entwickler*innen an sich betreffen, sondern vielmehr die Menschen, die mit ihr interagieren.

Zur Erarbeitung dieser Norm ist es daher entscheidend, die relevanten Akteur*innen breit zu beteiligen.

Bedarf 04-04: Erfüllung des Standardisation Requests zum EU AI Act, Aspekt „Menschliche Aufsicht“
Der vorliegende Entwurf zur **KI-Verordnung der EU** (KI-VO) legt einen Fokus auf die soziotechnische Perspektive: Die Anforderung, eine menschliche Aufsicht zu gewährleisten, kann nur erfüllt werden, wenn das KI-System als soziotechnisches System verstanden und der Mensch als Teil des Systems mitgedacht wird.

Wie menschliche Aufsicht in unterschiedlichen Rollen und mit einer Reihe von Eingriffsmöglichkeiten bis hin zu einer „Stopptaste“, die von Menschen ausgelöst wird, umgesetzt werden soll und welche Basisinformationen als Grundlage für menschliche Eingriffe ins System vorhanden sein müssen – das sind Fragestellungen, die nicht die KI bzw. KI-Entwickler*innen an sich betreffen, sondern vielmehr die Menschen, die mit ihr interagieren.

Zur Erarbeitung dieser Norm ist es daher entscheidend, die relevanten Akteur*innen breit zu beteiligen.

Die Arbeitsgruppe Soziotechnische Systeme hat die identifizierten Bedarfe nach der Dringlichkeit ihrer Umsetzung bewertet. Abbildung 35 zeigt die Dringlichkeit der Umsetzung der Zielgruppe Normung.

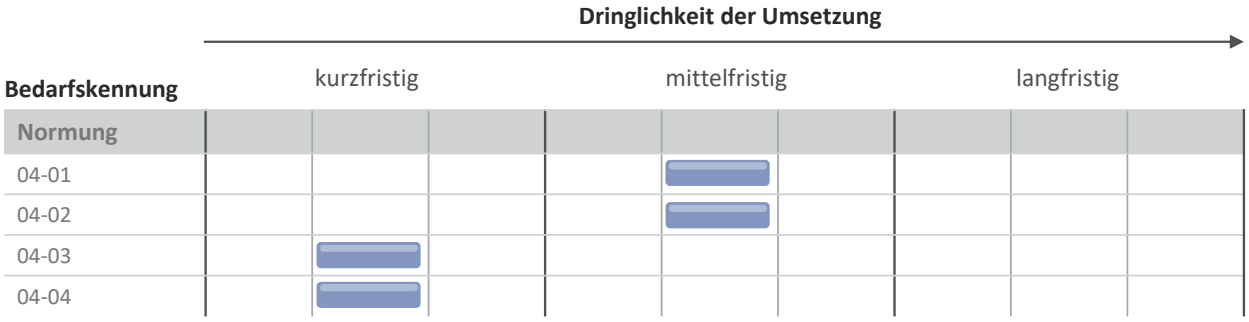


Abbildung 35: Priorisierung der Bedarfe aus Schwerpunkt Soziotechnische Systeme (Quelle: Arbeitsgruppe Soziotechnische Systeme)

4.5 Industrielle Automation

Ein Fünftel der gesamtwirtschaftlichen Bruttowertschöpfung Deutschlands wird derzeit direkt von der produzierenden Industrie bzw. dem verarbeitenden Gewerbe erbracht; eine Vielzahl weiterführender Dienstleistungen hängen zusätzlich hiervon ab [278]. Demzufolge stellt die produzierende Industrie einen wesentlichen Taktgeber der deutschen Konjunktur dar. Dies hat zur Folge, dass die digitale Transformation der produzierenden Industrie, also der steigende Einsatz von Methoden und Werkzeugen der Informationstechnologie und (Netzwerk-)Kommunikation von essenzieller Bedeutung für den Wirtschaftsstandort Deutschland ist.

Im Rahmen der Arbeiten des Zukunftsprojekts Industrie 4.0, welches durch die Deutsche Bundesregierung bereits im Jahr 2015 initiiert wurde, wurde das Themenfeld als konsequente Weiterentwicklung der Automatisierung der produzierenden Industrie (Industrie 3.0) strukturiert aufgearbeitet. Daher wird der Begriff industrielle Automation im Folgenden mit produzierender Industrie bzw. verarbeitendem Gewerbe sowie Industrie 4.0 synonym verwendet. Ausgehend von existierenden Wertschöpfungsprozessen der produzierenden Industrie [279], [280] wurden entsprechende zukünftige Anwendungsszenarien definiert [281], [282], [283]. Dabei umfassen die Anwendungsszenarien einen breiten Anwendungsbereich wie beispielsweise die auftragsgesteuerte Produktion auf Basis dynamischer Wertschöpfungs- und Liefernetzwerke, wandlungsfähige Fabriken, welche eine flexible Adaption von Fertigungsressourcen einer Fabrik ermöglichen, smarte Produktentwicklung u. v. m. Diese Anwendungsszenarien stellen dabei die Grundlage für weiterführende Verfeinerungen und Analysen zur Ableitung etwaiger Forschungs- und Normungsbedarfe dar [284], [285].

Künstliche Intelligenz (KI) stellt im Zusammenhang mit der digitalen Transformation des produzierenden Gewerbes eine wichtige und wesentliche Schlüsseltechnologie dar [285]. Insbesondere weist KI ein besonders hohes Potenzial auf, um Abläufe und Prozesse in der produzierenden Industrie [284], [287] nachhaltig zu gestalten und die Wertschöpfung durch Dynamisierung und Flexibilisierung zu steigern sowie Geschäftsmodelle in der produzierenden Industrie zu verändern. Dabei können sowohl traditionelle, aber auch neu gestaltete Produktionsabläufe und Sekundärprozesse wie beispielsweise Logistikprozesse durch KI verbessert, optimiert und flexibilisiert werden [282], [288]. Im Englischen und vermehrt auch im deutschsprachigen Raum wird hierbei auch von industrieller KI, Industrial Artificial Intelligence oder Industrial AI gesprochen, welche als Sammelbegriff für alle Anwendungsfelder von Künstlicher Intelligenz in der industriellen Anwendung dient [289], [290]. Wie in Abbildung 36 exemplarisch dargestellt, können und werden unterschiedliche Methoden und Algorithmen der Künstlichen Intelligenz zur Umsetzung verschiedener Anwendungen eingesetzt werden.

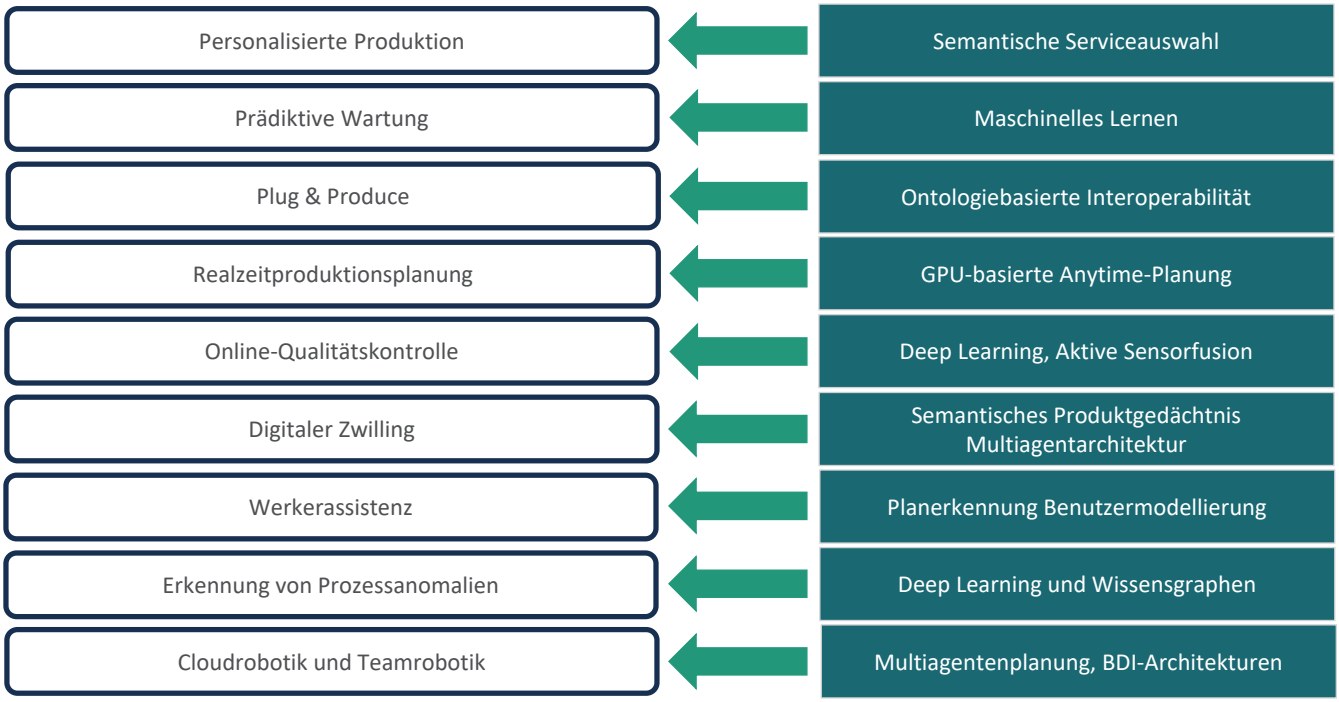


Abbildung 36: Übersicht Methoden und Algorithmen der KI und deren Anwendungen (Quelle: Prof. Wolfgang Wahlster, DFKI)

Internationale Normung und Standardisierung ist im produzierenden Gewerbe / der industriellen Automation von großer Bedeutung [285], [291]. In der industriellen Automation kommt, insbesondere bei der Entwicklung und dem Betrieb automatisierter Systeme, eine Vielzahl unterschiedlicher Herstellender zum Einsatz; große Zulieferbäume für Komponenten und Teilsysteme sind üblich. Dementsprechend kommt einer unternehmensübergreifenden Interoperabilität (mechanisch, elektrisch sowie hinsichtlich Software, Kommunikation und Daten sowie deren Beschreibung) eine große Bedeutung zu, welche durch Normen und Standards adressiert wird. Normen sind ferner für die Definition von Lösungswegen zur Einhaltung regulatorischer Rahmenbedingungen wie z. B. Maschinenrichtlinie [216] oder Einhaltung von Schutzzielen wie der sichere Betrieb (siehe Kapitel 4.2.1) von grundlegender Bedeutung. Dies gilt demzufolge auch für den Einsatz von KI und untermauert damit die durch die KI-Strategie der Bundesregierung geforderten Normungs- und Standardisierungsaktivitäten [2]. Aus diesem Grund wurde das Thema Normung und Standardisierung für die produzierende Industrie bereits seit Jahren in der DIN/DKE Normungsroadmap Industrie 4.0 detailliert untersucht und dedizierte Bedarfe abgeleitet; auch KI wurde ab Version 4 [291] explizit adressiert und kontinuierlich die Normung von KI in industriellen Anwendungen nachverfolgt [292].

Dabei besteht laut NRM KI Ausgabe 1 [63] der grundlegende Bedarf einer strukturierten Analyse von Anwendungsfällen und Ableitung normativer Anforderungen in der industriellen Automation, welcher bereits durch IEC/TC 65/WG 23 und ISO/IEC/JTC 1/SC42/WG 4 aufgegriffen wurde, wobei sich entsprechende technische Reports in Aktualisierung befinden (wie im Falle des ISO/IEC TR 24030:2021 [293]) oder kürzlich veröffentlicht wurden wie der PD IEC TR 63283-2 [294]. Demzufolge wird in der aktuellen Ausgabe das Thema Anwendungsfälle/Use Cases nicht mehr detailliert aufgeführt.

Eine wichtige Rolle in der digitalen Transformation wird der digitalen Abbildung der physischen Realität zugeschrieben: dem sogenannten digitalen Zwilling. Um die Interoperabilität innerhalb eines digitalen Ökosystems sicherzustellen, erarbeitet die Plattform Industrie 4.0 gemeinsam mit allen beteiligten Institutionen die Spezifikation der sogenannten Verwaltungsschale als digitales Abbild jedes relevanten Gegenstands (Asset) in der vernetzten Produktion [295], [296], [283]. Eine Verwaltungsschale speichert alle wesentlichen Eigenschaften eines Assets wie beispielsweise physische Eigenschaften (Gewicht, Größe), Prozesswerte, Konfigurationsparameter, Zustände und Fähigkeiten. Dabei ist die Verwaltungsschale nicht nur Informationsspeicher, sondern auch Kommunikationsschnittstelle, über die ein Asset in die vernetzt organisierte Industrie-4.0-Produktion eingebunden wird. Hierdurch ist es möglich, auf alle Informationen in einem Asset zuzugreifen und dieses zu kontrollieren. Dies stellt den Rahmen und eine wichtige Grundlage für die Anwendung von Künstlicher Intelligenz für die Industrie 4.0 dar, da hierdurch auf Daten- und Metadaten relevanter Assets einheitlich zugegriffen werden kann und diese in einem strukturierten Datenformat zur Verfügung stehen. Aktuelle Herausforderungen im Zusammenhang mit Datenmodellen und deren Semantik für den Einsatz von KI in der industriellen Automation wird in Kapitel 4.5.2 detailliert betrachtet. Bekannte Anwendungsbeispiele für KI in Industrie 4.0 sind u. a. die vorhersagende Wartung, wobei auf Basis symbolischer sowie mittels Machine-Learning-Modelle und gesammelter Betriebsdaten die Lebensdauer und der notwendige Wartungszeitpunkt von Komponenten vorhergesagt wird. Um die Verfügbarkeit notwendiger Daten (unternehmensübergreifend) sicherzustellen, bekommen Datenräume eine zunehmende Bedeutung in Industrie 4.0 [283] für die Anwendung von KI; auch hier kommen expliziten Datenmodellen und deren automatischer Verarbeitung (sogenanntes Reasoning) eine wesentliche Rolle zu.

Die Bedeutung expliziter (semantischer) Modelle in der industriellen Automation ergibt sich dabei u. a. aus deren langjähriger Anwendung bei der Entwicklung von Maschinen und Anlagen, welche häufig detailliert mechanisch, elektrisch geplant und dann softwaretechnisch automatisiert werden. Hierbei entstehen bereits eine Vielzahl von Modellen, deren Nutzung großes Potenzial durch den Einsatz von KI zeigt. Daher stellt ein nicht unwesentlicher Bestandteil aktueller Aktivitäten in Industrie 4.0 die Entwicklung technischer Systeme dar, in denen Künstliche Intelligenz zum Einsatz kommt [297]. Aus diesem Grund wird in dieser Ausgabe erstmals das Thema „KI Engineering“ detailliert betrachtet, analysiert und dessen Bezug zu Normung und Standardisierung beschrieben (siehe Kapitel 4.5.1).

Ferner werden im Kontext von Industrie 4.0 weitere Anwendungen der Künstlichen Intelligenz betrachtet. Neben der autonomen Intralogistik (siehe hierzu auch Kapitel 4.6) werden beispielsweise auch die industrielle Bildverarbeitung und Bilderkennung sowie die Verbesserung der Interaktion und Integration von Mensch und Maschine berücksichtigt. Zum einen durch den Einsatz neuer Interaktionsmechanismen wie Sprache und Geste, durch neue Darstellungsmöglichkeiten wie Augmented Reality (AR) und die Stärkung der Zusammenarbeit wie beispielsweise durch kollaborative Robotik. Hierbei finden KI-Technologien durchweg intensive Anwendung.

4.5.1 KI-Engineering

4.5.1.1 Status quo

KI-Engineering adressiert die systematische Entwicklung und den Betrieb KI-basierter Lösungen als Teil von Systemen, die komplexe Aufgaben erfüllen [vgl. Kompetenzzentrum KI-Engineering Karlsruhe CC-KING[88]]. Damit ergänzt KI-Engineering die Grundlagenforschung zu Künstlicher Intelligenz (KI) und Maschinellem Lernen (ML) und schlägt die Brücke zu den Ingenieurswissenschaften. Ziel der industriellen Automation ist es, KI- und ML-Methoden gemäß den typischen Anforderungen und Vorgehensweisen von Ingenieur*innen nutzbar zu machen, auch in sicherheitskritischen Anwendungen.

Ein wesentliches Ziel ist die Akzeptierbarkeit von KI-Methoden, insbesondere deren Einsatz in Subsystemen von kritischen Anwendungen und komplexen Systemen. Dies setzt ein hohes Maß an Verlässlichkeit, Vertrauenswürdigkeit, Sicherheit (im Sinne von Safety und Security) und Kontrollierbarkeit voraus. Dazu fehlen in der Industrie akzeptierte Vorgehensweisen und Entwicklungsmethoden. Ein erster Ansatz ist PAISE(R)[89] – Process Model for AI Systems Engineering. Nicht in allen Anwendungen wird ein gleiches Maß an diesen nicht-funktionalen Anforderungen benötigt. Deshalb muss ein Vorgehensmodell für eine gegebene Situation maßgeschneidert werden können (tailoring). Zudem stellt sich die Frage, wie Qualitätskriterien in KI-basierten Systemen beschrieben und validiert werden können.

Da in der Praxis KI-Methoden in bereits existierende Systeme (legacy systems) eingebracht werden müssen, sollte ein Vorgehensmodell auch Migrationsansätze und agile Erweiterungen bestehender Systeme unterstützen. Zudem ist im KI-Engineering der komplette Lebenszyklus von KI-basierten Systemen zu betrachten, da im Betrieb Abweichungen vom Systemkontext zum Zeitpunkt der Entwicklung entstehen können, die systematisch zu behandeln sind. Beispielsweise können die von ML-Methoden im Betrieb benutzten Sensordaten von den in der Entwicklungsumgebung benutzten Trainingsdaten in ihrer statistischen Verteilung so weit abweichen, dass die Aussagekraft des ML-Verfahrens beeinträchtigt wird (distributional shift).

KI-Engineering erfordert das koordinierte Zusammenwirken von Repräsentanten unterschiedlicher Disziplinen und Ausbildungen: Ingenieur*innen, KI-Expert*innen und Informatiker*innen. Während das Fachwissen und damit auch die fachlichen Anforderungen typischerweise von Ingenieur*innen (Maschinenbau, Chemie, Verfahrenstechnik, Elektrotechnik, ...) abgedeckt werden, ist die Kenntnis über KI-Methoden zumeist Spezialisten vorbehalten (KI-Expert*innen), die über dedizierte mathematische und statistische Methodenkompetenzen verfügen. Letztlich ist ein KI-basiertes System ein IT-System, bestehend aus Hardware- und Softwarekomponenten/Subsystemen, das gemäß den etablierten Methoden des System- und Software-Engineerings zu entwickeln und zu betreiben ist. Die dazu notwendigen Kompetenzen werden insbesondere durch IT-Expert*innen mit ausgewiesenen Informatikkompetenzen verkörpert.

Es besteht die visionäre Vorstellung, dass KI-Engineering einen Methodenbaukasten liefern kann mit klaren Aussagen, welche Fähigkeiten sowohl funktional als auch nicht-funktional/qualitativ mit welchen Methoden und unter welchen Rahmenbedingungen erreichbar sind. Dazu ist es notwendig, KI/ML-Verfahren gemäß einheitlichen (Meta-)Modellen beschreiben, bewerten und damit vergleichen zu können. Dies beinhaltet auch die folgenden Aufgaben:

- → Erarbeitung eines Ordnungsschemas für KI/ML-Methoden: überwachte vs. unüberwachte ML-Verfahren vs. Verfahren des verstärkenden Lernens (reinforcement learning); vortrainierte ML-Methoden vs. selbstlernende Systeme (reinforcement learning, Kalman-Filter u. a.)
- → Beschreibung und Validierung von Qualitätskriterien in KI-basierten Systemen
- → Erarbeitung einer fachlich orientierten Erklärbarkeit von KI-basierten Entscheidungen/Entscheidungsvorschlägen
- → Erarbeitung der Rahmenbedingungen und Metabeschreibungen in der strukturierten und systematischen Datenvorverarbeitung, sowohl für strukturierte, semistrukturierte als auch unstrukturierte Daten, sowohl für statische als auch dynamische Daten (u. a. Zeitreihen).

Zu den meisten dieser Themen und Aufgaben bestehen rudimentäre und singuläre Lösungsansätze aus der Wissenschaft und der industriellen Praxis. KI-Engineering hat das Ziel, diese Ansätze und Lösungen systematisch und interdisziplinär zusammenzuführen und dafür disziplinübergreifende und in der Praxis akzeptierte Standards zu etablieren.

88 https://www.ki-engineering.eu/de/was-ist-ki-engineering.html

89 https://www.ki-engineering.eu/content/dam/iosb/ki-engineering/downloads/PAISE(R)_Whitepaper_CC-KING.pdf

Anwendungsfälle

Die nachfolgend beschriebenen Anwendungsfälle sollen den interdisziplinären Charakter von KI-Engineering beschreiben und erläutern.

→ **Verlässliche Energieversorgung für die industrielle Produktion im Fehlerfall**

Für die industrielle Produktion ist es unerlässlich, sich auf eine zuverlässige Energieversorgung stützen zu können. Eine zuverlässige Energieversorgung bedeutet insbesondere die Vermeidung von kaskadierenden Effekten bei möglichem Versagen (failure) verteilter Energieressourcen (distributed energy resources, DER) zur Energieübertragung und -verteilung. Das Versagen bzw. der Ausfall einzelner Komponenten kann aufgrund der vorhandenen DER-Resilienzen und -Redundanzen verkraftet werden, während ein kaskadierender Effekt des Versagens gekoppelter DER-Komponenten ein hohes Risiko für einen Blackout ganzer Netzwerkteile zur Folge haben kann.
Eine Kaskade des Versagens einzelner DER-Komponenten kann auch unter normalen Umständen zu einer Kette weiterer Fehler in der Energieübertragung oder -verteilung führen. Zur Vermeidung des Kaskadierungseffekts müssen folgende drei Faktoren von den Energieversorgungsunternehmen unter Kontrolle gehalten werden [298], [299]:

- Augenmerk auf DER-Komponenten, weil Ausfälle meist in Subkomponenten geschehen.
- Das Power Management System (PMS) muss auch unter Stressbedingungen vollständig im Operationsmodus verbleiben können.
- Eine mögliche Kaskadierung ist eine Auswirkung des Verhaltens sehr großer Systeme („very large-scaled systems“ VLS, welche auch als „system-of-systems“ bezeichnet werden).

Daraus folgt, dass das PMS einer neuen angepassten „Sicherheitspolicy“ folgen muss. Um eine solche Policy durchzusetzen, braucht es ein Maß, um technischen Stress, z. B. hervorgerufen durch ein Gewitter, einordnen zu können. Diese Einordnung könnte ggf. durch ein geeignetes ML-gestütztes Kategorisierungsverfahren, das sich z. B. an der Schwere eines Gewitters orientiert, erfolgen. Weiterhin sollten regulatorische Maßnahmen zur Steigerung der Empfindlichkeit des PMS bezüglich Komponentenversagens mit hohem Risiko zur Kaskadierung vorgesehen werden. Ein praktikables Regelwerk baut u. a. auf Informationen und Wissen auf, wie das Versagen einzelner DER-Komponenten das Verhalten des gesamten Energieversorgungsnetzes (failure propagation) beeinflusst, z. B. indem Blackouts von großen Versorgungsgebieten bzw. großen industriellen Verbraucher*innen eintreten und damit großen Schaden in der Energieversorgung oder Produktion anrichten können.
Ein Blackout unterbricht die Energieversorgung, weil die unmittelbare Kompensierung des Ausfalls einer Übertragungsleitung durch andere Leitungen wegen mittelbarer Lastüberlastung oder möglicher Überhitzung nicht erfolgen kann bzw. die Leitungen zur Kompensierung ebenfalls ausfallen.
Eine formalisierte Regelung für das PMS benötigt computerisiertes Wissen über Hochspannungsübertragungsleitungen und Transformatoren, das durch Knoten (graph vertices) repräsentiert wird. Erfahrungswerte über Übertragungsleitungsausfälle, sogenannte distribution factors, kennzeichnen die möglichen Folgen eines Übertragungsleitungsausfalls auf andere Leitungen, welche als graph edges in einen Graphen eingetragen werden.
Bezug zu KI-Engineering: Wenn in einem Subsystem eines PMS KI-Methoden zum Einsatz kommen, z. B. ML-Verfahren für die Optimierung der Energienetze im Normalbetrieb, frühzeitige Erkennung von Anomalien im Energienetz zur Prognose von Energieversorgungsausfällen als auch mögliche Lösungsvorschläge zur Problembehebung oder -minderung, müssen Aussagen zur Verlässlichkeit derartiger Aussagen getroffen und auf der fachlichen Ebene begründet werden können.

→ **Zusammenspiel von KI-Methoden mit dem Industrie-4.0-Konzept der Verwaltungsschale**

Die Verwaltungsschale (VWS; engl. Asset Administration Shell AAS) ist ein wesentlicher Baustein zur Steigerung der Wertschöpfung eines Produkts. Mittels einer VWS kann der komplette Lebenszyklus eines Produkts abgebildet und optimiert werden. Je näher die Modellierung eines Produkts dem „realen“ Objekt entspricht, desto genauer kann diese das „reale“ Objekt virtuell abbilden und desto näher ist man am idealisierten Konzept eines vollumfänglichen Digitalen Zwillings. In Verbindung mit KI können somit z. B. Prozessabläufe virtuell optimiert werden, ohne dabei in reale Prozesse eingreifen zu müssen. Zum Beispiel kann eine Umsetzung in der physischen Welt erst nach einer erfolgreichen Simulation in der virtuellen Welt mittels des Digitalen Zwillings erfolgen.
Unter einem Digitalen Zwilling wird hier ein logisches Konzept verstanden, das den Zustand und das Verhalten eines realen Assets in der virtuellen Welt vollumfänglich abbildet. In der Praxis ist dies nicht wirtschaftlich umsetzbar und gemäß den Anwendungsfällen auch nicht erforderlich. Es besteht aber die Vision, dass in einem Digitalen Zwillingssystem über eine geeignete Service-In-

frastruktur eines Datenraums die jeweils notwendigen Daten für den jeweiligen Anwendungsfall über definierte Schnittstellen beschafft werden können, vgl. dazu das Reference System for Digital Twin Systems (DTS-RM).
Nachfolgend sind zwei Beispiele für die Anwendung von KI in Verbindung mit der Verwaltungsschale bzw. dem Digitalen Zwilling aufgeführt.
Beispiel 1: Einsatz von KI im Digital Twin für „Predictive Maintenance“
Durch den massiven Ausbau von Sensoren (z. B. für Temperaturen, Geräusche oder Vibrationen) innerhalb einer Anlage können Fehlerquellen schon in der Entstehungsphase erkannt und somit Ausfallzeiten minimiert werden. Für die Erkennung bietet der Digitale Zwilling – welcher den Sollzustand widerspiegelt – eine optimale Referenz für den KI-basierten Erkennungsalgorithmus (Anomaliedetektion).
Beispiel 2: KI-basierte Optimierung von 5G-Campusnetzen auf Basis eines Digitalen Zwillings
Die steigende Forderung nach hoch flexiblen Produktionssystemen (z. B. Losgröße eins) stellt neue Herausforderungen an die Planung und den Aufbau. Bewegliche Teile wie beispielsweise Automated Guided Vehicles (AGVs) müssen vollständig in die Anlage integriert werden. Eine geeignete Lösung bieten die 5G-Campusnetze.
Der Einsatz eines Digitalen Zwillings für ein 5G-Campusnetz bietet sowohl in der Planung als auch im laufenden Betrieb ein großes Potenzial für die Optimierung der gesamten Anlage. Für die Optimierungsprozesse kommen entsprechende KI-Methoden zum Einsatz. Durch die Kombination von Sensorik und entsprechender a-priori-Informationen können beispielsweise durch Bewegung entstehende Abschattungseffekte der 5G-Netzabdeckung in einer Produktionshalle prognostiziert und der Weg des AGVs entsprechend angepasst werden.
Bezug zu KI-Engineering: Um KI-Methoden auf der Basis von Daten eines Digitalen Zwillings verlässlich anwenden zu können, bedarf es wohldefinierter Herkunfts- und Qualitätsdaten des Digitalen Zwillings. Da die Daten zumeist aus unterschiedlichen Quellen stammen, ist eine standardisierte Bereitstellung der Daten und deren Verarbeitung unabdingbar.

Normungsbedarf

KI-Engineering ist auf dem Weg, sich als Teildisziplin des Systems Engineering eigenständig weiterzuentwickeln. Dazu gehören eigenständige Prozesse und Methoden, die wissenschaftlich fundiert und in der Praxis allgemein akzeptiert sind, sodass sich daraus auch Anforderungen an die geforderte Qualität und die nicht-funktionalen Eigenschaften eines Systems ableiten lassen. Diese werden in Lastenhefte einfließen und auch in Regularien verwendet werden. Ein Beispiel dafür ist die entstehende KI-Verordnung der Kommission der Europäischen Union (EU), die bezogen auf die Kritikalität eines Systems Regeln für den Einsatz von KI-Verfahren ableiten und verbindlich vorschreiben wird.

Zur effizienten und rechtssicheren Umsetzung dieser Anforderungen und Regularien bedarf es allgemein akzeptierter, abgestimmter und vorzugsweise genormter Verfahren, Modelle und Vorgehensweisen. Dies ist notwendig, da KI-Engineering nur in einem Zusammenspiel von Akteur*innen aus Ingenieursdisziplinen (Maschinenbau, Elektrotechnik, Chemie, Verfahrenstechnik, ...), der Informatik und den Datenwissenschaften (data sciences) gelingen kann [300]. Die in diesen bestehenden Disziplinen vorherrschenden Normen und Standards können nicht unverändert übernommen, sondern müssen konzeptionell zusammengeführt werden. Dies umfasst eine einheitliche Definition von Begrifflichkeiten und die Beschreibung von nicht-funktionalen Systemeigenschaften, die auch durch den Einsatz von KI-Verfahren in Teilsystemen erreichbar und ggf. zertifizierbar sind.

4.5.1.2 Anforderungen und Herausforderungen

Aus den beschriebenen Herausforderungen des Themengebiets KI-Engineering leiten sich die folgenden Themen für den Normungsbedarf ab:

- → Metadatenbeschreibungen von Eingangs-/Ausgangsdatensätzen von ML-Verfahren
- → Metabeschreibung von KI-Methoden
- → Taxonomie, textuelle und ggf. formale Beschreibung von Qualitätskriterien KI-basierter Systeme u. a. Verlässlichkeit, Zuverlässigkeit, Planbarkeit, Kontrollierbarkeit, …
- → Metriken zur Erklärbarkeit
  - Ziel: Beschreibung des Trade-off zwischen Erklärbarkeit der angewandten maschinellen Lernverfahren (Gedankenmodell der Anwender*in) und der Genauigkeit bzw. Güte
  - Nutzung semantischer Modelle in der Erklärbarkeit
- → Systematisches Vorgehen beim Einsatz von KI-Methoden in Subsystemen komplexer Systeme (KI-Engineering-Vorgehensmodell)
  - In allen Phasen des Lebenszyklus eines Systems (Entwurf → Realisierung → Betrieb und Pflege inklusive Erweiterungen/Änderungen)

→ Modellierung KI-basierter Systeme (technische und anwendungsbezogene Aspekte), z. B. Unified Modeling Language (UML) Profil, spezielle Stereotypen für KI-Aspekte

4.5.2 Datenmodellierung und Semantik

4.5.2.1 Status quo

Eine heute typische Vorgehensweise bei der KI-Datenmodellierung beruht auf der Exploration historischer Daten. Hierbei haben sich erste Methoden wie z. B. Cross Industry Standard Process for Data Mining, wie es im KI-Engineering vorgeschlagen wird, weitgehend etabliert, um die Aufbereitung im industriellen Kontext durchzuführen und eine Modellbildung mit der entsprechenden Optimierung zu erzielen. Nach der Implementierung und dem notwendigen Test eines KI-Systems erfolgt der Übergang in den operativen Betrieb, und ein nachgelagertes Überwachen stellt sicher, dass Daten und Datenmodell sowie die eingesetzten Routinen in der dann notwendigen Qualität betrieben werden. Ungelöst dabei sind beispielsweise Fragen, wie eine Qualität mit historischen Daten erreicht werden kann, die unter anderen Randbedingungen entstanden sind (Kapitel 4.5.4, Handlungsbedarf 05-09). Durch die methodische Erweiterung des Systemengineerings um die Aspekte der KI beispielsweise durch das vorgeschlagene KI-Engineering kann die Industrie diese neue Technologie gewinnbringend verwenden. Speziell bei der Modellbildung hat sich in der Industrie das Vorgehen etabliert, Elemente und Mengen in einen Zusammenhang zu setzen und in Diagrammen und Abbildungen zu erklären. Beispielsweise sind bekannte semantische Datenmodelle das Entity-Relationship-Modell oder die bei der objektorientierten Modellierung eingesetzte Unified Modeling Language. Die Abstraktion zur Modellbildung ist notwendig, um die reale Welt im digitalen Raum abzubilden (vgl. Abbildung 37).

Umgekehrt reagiert der digitale Raum auf die physische Welt, beispielweise über „Robot-Process-Automation“-Mechanismen. Die Datenmodellierung ist mit erheblichem manuellem Aufwand verbunden und unterliegt einer gewissen Willkür. Zudem besteht die Herausforderung bei der Modellierung darin, dass die Systemgrenzen der Modelle dynamisch sind und ein hoher Abstimmungsbedarf bei der Erstellung der Modelle besteht. Das zeigt sich letztlich in der Anwendungsbreite von hochspezialisierten Algorithmen bis hin zu allgemeinen Lösungen. Dabei geht es bei der Modellierung grundsätzlich immer um die Auflösung von Widersprüchen zwischen Wirklichkeit, Modellen und Artefakten. Wichtig werden die Interpretationsmechanismen (vgl. Abbildung 38), die von dem physischen Raum in den digitalen Raum und umgekehrt wirken. Ebenfalls unterschätzt wird die Qualität sowohl der Daten selbst als auch der Datenmodelle und von deren Architekturen, die aber für eine erfolgreiche Umsetzung notwendig ist. Hier kann die Normung durch Metriken und Standards zur Datenqualität gut unterstützen (Kapitel 4.5.4, Handlungsbedarf 05-09). Die Stärke des Maschinellen Lernens liegt darin, Daten zu transformieren, ohne dass a priori eine vollständige mathematische Vorschrift benötigt wird. Darin liegen auch bekannterweise die Nachteile, die speziell in der Erklärbarkeit und Vorhersagbarkeit der Ergebnisse heute diskutiert werden.

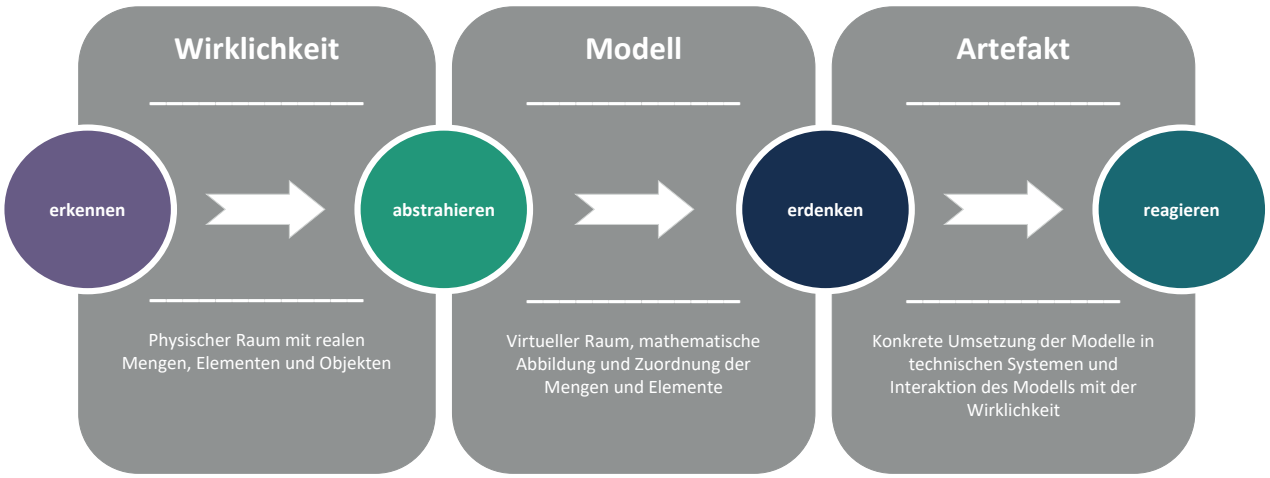


Abbildung 37: Modellbildung (Quelle: Arbeitsgruppe Industrielle Automation)

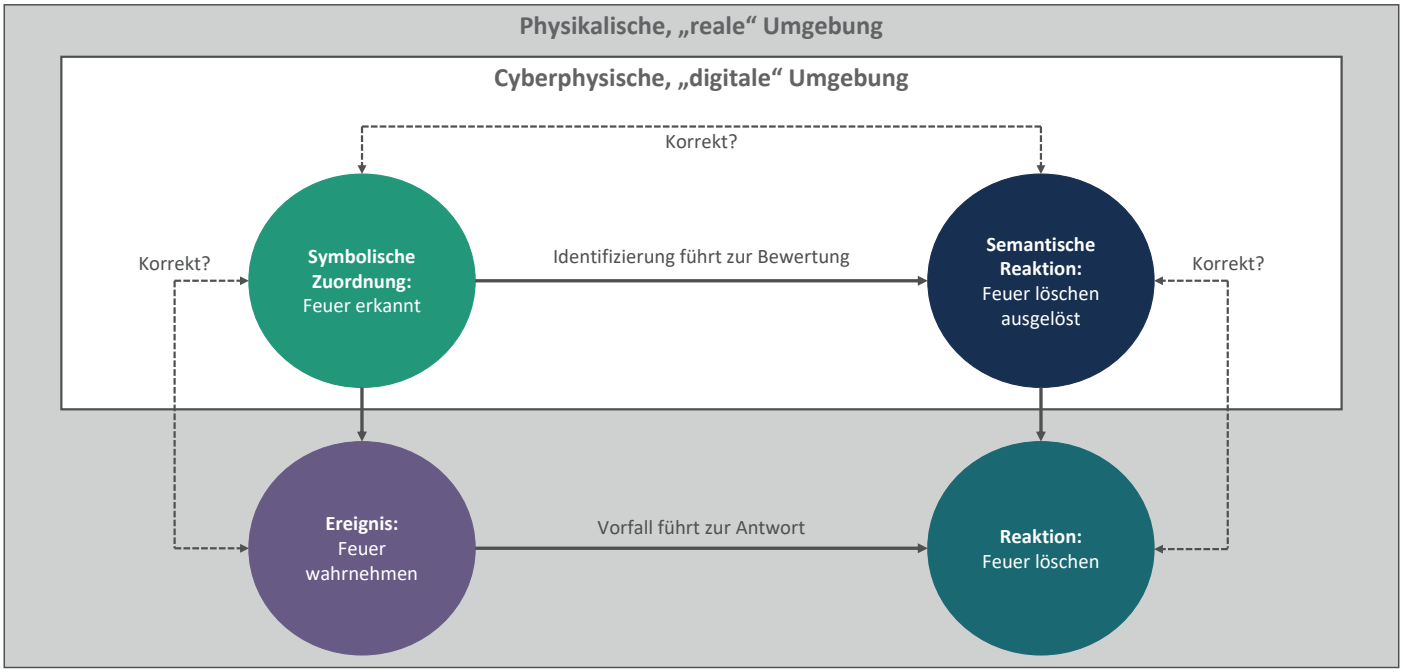


Abbildung 38: Interaktion (Quelle: Arbeitsgruppe Industrielle Automation)

Welche Bedeutung die Semantik, die Datenmodellierung, die Datenqualität und die Interaktion zwischen der realen und der cyberphysischen Welt hat, soll folgendes allgemeines Beispiel des Feuerlöschens verdeutlichen. Eine erste Frage ist, durch welche Sensorik „Feuer“ erkannt werden soll und in welchem Kontext dieser Anwendungsfall in der „realen“ Umgebung steht. „Feuer“ in der menschlichen Interpretation könnte z. B. durch einen Temperatursensor, einen Kohlenmonoxidsensor oder eine Kamera sensorisch erfasst werden. Die generierten Daten sind vollkommen unterschiedlich, denn es können Grad Celsius, Partikel oder direkt farbige oder farblose Bilder sein. Falls lediglich ein Datenmodell erzeugt werden darf, muss es mit den jeweiligen Sensordaten umgehen können. Der Anwendende steht bereits hier vor einer Entscheidung zur Festlegung der weiteren Vorgehensweise. Entweder erhält jeder Sensor ein eigenes Datenmodell oder die unterschiedlichen Sensordaten werden interpretiert und in ein Datenmodell überführt. In der Praxis werden verfügbare Sensoren über die Zeit durch neuere Sensoren ersetzt – ein nicht unübliches Szenario. Aber auch das Datenmodell selbst könnte sich auf die unterschiedlichen Sensordaten adaptieren. Unklar ist heute, wie eine Überprüfung sichergestellt wird. Um eine geeignete Antwort auf das Ereignis durch das System zu geben, ist zusätzlich der Kontext herzustellen. Dieser Kontext bildet sich über die Zuordnungen ab. Das Feuer kann das Feuer eines Feuerzeugs oder ein brennendes Fahrzeug sein – insofern wird es unterschiedliche Antworten bzw. Reaktionen darauf geben. Auch die Umsetzung der Antworten kann über unterschiedliche Aktorik erfolgen – insofern gelten ähnliche Bedingungen wie bei den Sensordaten. Im Folgenden könnte das Beispiel ein brennendes Fahrzeug sein.

In der gleichen Abbildung 38 besteht die „reale Welt“ aus einem Fahrzeug in einem bestimmten, sich dynamisch verändernden Zustand. Der digitale Raum (ggf. digitaler Zwilling mit KI-Komponente) kann in seinen Prozess- und Datenmodellen die Dynamik in der realen Welt simulieren und damit analysieren. Der Vorfall „Fahrzeug brennt“ muss im digitalen Raum schnell und korrekt erkannt werden sowie eine schnelle und adäquate Reaktion generieren. Damit muss ein Vorfall zu einer geeigneten Antwort führen, denn diese stehen im semantischen Verhältnis als Implikation zueinander. Die Ursachen-Wirkungs-Interpretation ist „es brennt, es soll gelöscht werden“. Diese Interpretation kann aber auch durch eine andere Interpretation „es brennt → Heizung in Ordnung“ und entsprechend anderen Kontext ersetzt werden. Der digitale Zwilling wird mit seinen Modellen schnell analysieren können, dass die erste Interpretation in einem Fahrzeug „korrekt“ ist und die zweite möglicherweise in einem Gebäude ihre Gültigkeit hat.

In der digitalen Welt könnten die realen Gegenstände, hier das Fahrzeug, über eine Verwaltungsschale verfügbar gemacht und die Verwendung auf der Grundlage deklarierter

Semantik, d. h. unter Beachtung axiomatischer Bedingungen, eingebunden oder überwacht werden. Dabei kann die Einhaltung der Deklarationen in der Verwaltungsschale auch über Untermodelle realisiert und administriert werden. Im Beispiel des Fahrzeugs mit KI-Komponente werden dann mindestens zwei Untermodelle benötigt: eines für die Erkennung des Vorfalls und eines für die Analyse der „korrekten" Antwort. Während für das Vorfall-Untermodell ein Datenmodell geeignet sein dürfte, wäre für die Analyse der möglichen Antwort vermutlich eher ein Prozessmodell auf graph-theoretischer Basis geeignet. Es ist bemerkenswert, festzustellen, dass bei semantischer Sicht auf die betrachteten Vorgänge die in den Untermodellen der Verwaltungsschale eingetragenen Fakten hinreichend sind. Die Prüfung auf Korrektheit bzw. die Interpretation dieser Fakten könnte von einem autorisierten Digitalen Zwilling mit Zugriff auf die Untermodelle der Verwaltungsschale von außen durchgeführt werden. Damit wird im Ansatz ersichtlich, welche Bedeutung, aber auch welche Kompliziertheit bei der Gestaltung dieser Systeme beherrscht werden muss.

Nun wurde in der ersten Version der KI-Normungsroadmap auf die Bedeutung von Deklaration und Narration in der semantischen Modellierung im Hinblick auf die Interoperabilität solcher Systeme hingewiesen, insbesondere bei deren dynamischem Zusammenspiel. Der Begriff „Deklaration" – im Sinne von deklarativer Wissensdarstellung – bezeichnet die maschinell überprüfbare, als widerspruchsfrei angestrebte Darstellung von Struktur und Verhalten interoperabler Systeme unter Verwendung von Axiomen, Fakten und Regeln bei gleichzeitigem Verzicht auf prozedurale Anteile. Der Begriff der „Narration" kann darauf aufbauend und wie ein „Programmablauf" hingegen zweierlei meinen:

→ Als Prozess stellt eine Narration eine dynamische Orchestrierung von kommunizierenden, interoperablen, modellbasiert-deklarativ beschriebenen Systemen mit ihren Variablen dar. Die jeweilige Orchestrierbarkeit der Systeme spannt einen Suchraum möglicher Graphen von Modellen auf, durch den mittels der Narration eine Trajektorie gebildet wird. Insofern hat das Konzept der Narration eine konzeptionelle Nähe zum Automatisierten Planen (siehe u. a. PDDL – Planning Domain Definition Language [301]: Plan Generation, Plan Execution).

→ Als Artefakt stellt eine Narration eine geplante oder als Ergebnis eines Narrationsprozesses tatsächliche Abfolge von Interaktionen zwischen kommunizierenden, interoperablen Systeme und ihren modellbasiert definierten Variablen oder Entitäten dar. Aufgrund von Veränderungen im Suchraum während der Narration, z. B. aufgrund von Veränderungen in den Rahmenbedingungen, insbesondere zur Laufzeit, wird die konkrete Trajektorie durch den Suchraum gebildet. Diese kann dann weiterer Verarbeitung, z. B. einer Validierung, zugeführt werden.

In diesem Kontext meint der thematisch zugehörige Begriff des „Narrativs" die Narration als Artefakt und stellt somit eine deklarativ beschriebene, wiederverwendbare, überprüfbare Trajektorie durch den Graphen-Suchraum dar. Somit kann ein Narrativ a priori vorgegeben werden, um einen Suchraum in einer bestimmten Art und Weise zu beeinflussen („Was soll getan werden?"). Ein Narrativ kann aber auch a posteriori durch Beobachtung der Veränderungen an einem Graph ermittelt werden („Was wurde getan?"). Dabei ist bei den interoperierenden (KI-)Systemen gerade mit Blick auf Kopplungen heterogener Systeme von einer unterschiedlichen und a priori einander nicht zwingend bekannten technischen Basis auszugehen, insbesondere bei dynamischen Verschaltungen. Zudem können verschiedene Parteien (Systeme, Werkzeuge, Wissensingenieure) gleiche Modelle mit nicht deckungsgleicher Interpretation ihrer Semantik nutzen. Damit ergeben sich Abweichungen und Verluste in den Verarbeitungen der Inhalte. Ursprüngliche Intentionen von Datenstrukturen und Modellen können nicht durchgängig ausgedrückt, weitergegeben und rekonstruiert werden. Damit ist eine verlustfreie Anwendung von Modellen und deren Validierung auf konsistente Interpretationen über mehrere Parteien entlang einer Verarbeitungskette („Pipeline") nicht gewährleistet (Kapitel 4.5.4, Handlungsbedarf 05-06).

Eine verlustfreie, konsistente horizontale wie vertikale Interpretierbarkeit von übermitteltem deklarativem Wissen ist jedoch eine wesentliche Anforderung für den erfolgreichen Einsatz von KI-Verfahren (Kapitel 4.5.4, Handlungsbedarf 05-07). Die Übermittlung von Wissen erfolgt in zwei Stufen, einmal auf der Ebene eines geeigneten Formats und zum zweiten auf der semantisch-interpretativen Ebene. Ein geeignetes Format von Wissen ist u. a. seine Darstellung als Ursache-Wirkungs-Implikation, was auch in Programmiersprachen dargestellt werden kann. Auf semantischer Ebene werden diese Implikationen als geordnete Paare in einem Graphen dargestellt und als Ereignis, d. h. eine aktuelle Kante, in den Graphen eingetragen. Die Graph-Semantik wird von Sender und Empfänger geteilt. Am Beispiel des W3C Semantic Web Stacks als aufeinander aufbauende Grundbausteine verdeutlicht: transportierte bzw. angelieferte Resource-Description-Framework (RDF)-serialisierte Strukturen können nicht in jedem Fall einer Verarbeitung durch RDFS- oder Web Ontology Language (OWL)-basierte KI-Mechanismen zuge-

führt werden, wenn nämlich RDF-Serialisierungsmechanismen verwendet werden, die keine standardisierte RDFS- oder OWL-konforme Interpretation erlauben. Auch wenn RDF die Grundlage für RDFS/OWL bildet und die serialisierten Inhalte grundsätzlich in ihrer Interpretation RDFS-/OWL-kompatibel und serialisierbar wären. Gleiches trifft bei der Verwendung des kommenden RDF*-Modells zu. Eine konkrete betroffene Datenstruktur ist eine RDF-Liste, die RDFS-/OWL-kompatibel ungeordnete Mengenelemente transportieren kann, für die es in den Tools aber keine standardisierten Transformationen gibt. Dies ist in der Abbildung 39 skizziert.

Zudem nehmen Werkzeuge bei Import und Export von Strukturen und Modellen individuelle Transformationen der jeweiligen Inhalte vor. Diese sind oft nicht semantikerhaltend wie auch gleichzeitig nicht überprüfbar, d. h. veränderte Interpretationen von Inhalten können nicht in jedem Fall erkannt werden. Transformationsmechanismen von Werkzeugen oder Systemen können nicht dediziert gemäß ihren Fähigkeiten angesprochen und getestet werden. Damit lässt sich nicht von extern erkennen, ob ein Werkzeug oder System angebotene Inhalte verlustfrei verarbeiten kann (Kapitel 4.5.4, Handlungsbedarf 05-08).

Der Verfügbarkeit entsprechender Transformationen, die hinsichtlich ihres Semantikerhalts klassifizier- und auch prüfbar durch Dritte sind, kommt daher eine zentrale Bedeutung zu, um verschiedene (KI)Systeme interoperabel werden zu lassen. Unterstrichen wird dies durch [302]. Die Forderungen der FAIR-Prinzipien (Findable, Accessible, Interoperable, Reusable) bezüglich Interoperabilität [303] sind hierfür zwar notwendig, aber nicht hinreichend. Zudem ist aufgrund damit einhergehender höherer Komplexität der Modellhandhabung eine verlässliche Automatisierbarkeit der Transformationsfunktionen erforderlich. Dazu können verstärkt Entwurfsmuster als Grundlage zum Einsatz kommen. Dies gewinnt noch durch das Aufkommen von Data Spaces als hochskalierte „dynamische und automatisierte Treffpunkte“ von (KI-)Systemen vor dem Hintergrund von „Trustworthy AI“ an Bedeutung. Hier treffen auch eine Vielzahl von interagierenden Digitalen Zwillingen mit individuellen Fähigkeiten und Prozessen, typischerweise bereitgestellt mittels der Verwaltungsschalen, aufeinander.

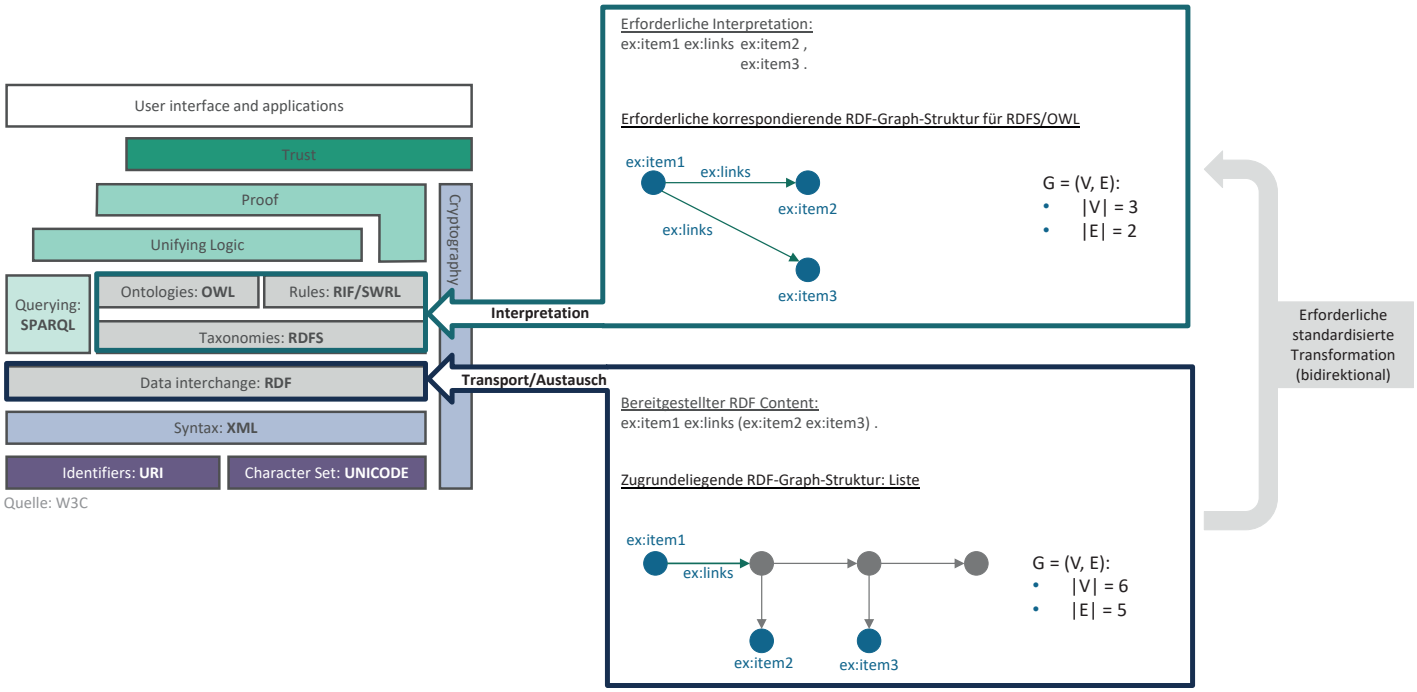


Abbildung 39: Datenmodellierung (Quelle: Arbeitsgruppe Industrielle Automation)

4.5.2.2 Anforderungen und Herausforderungen

Die übergreifende Herausforderung der „semantischen Interoperabilität“ findet sich insbesondere zwischen heterogenen technischen oder physikalischen Prozessen, die als komplexe, zeitkontinuierliche Variablen im Modell repräsentiert werden. Diese Variablen und ihre charakteristischen Eigenschaften werden etwa in den Submodellen einer Verwaltungsschale axiomatisch und regelbasiert deklariert. Bei Interoperationen zwischen administrierten Modellen findet ein Informationsfluss von Werten zwischen Instanzen, Objekten oder Prozessen statt. Dieser Wertefluss ist also das Kennzeichen der Interoperabilität zwischen den Modellen oder Systemen. Bei Störung dieser Stabilität wird die Beziehungsfolge unterbrochen, d. h. die semantischen Fakten zwischen den Modellen können nicht mehr erfüllt werden. Interoperation soll zwischen allen Modellen unterschiedlichen Typs bei gegebener Semantik möglich sein (Kapitel 4.5.4, Handlungsbedarf 05-07). Das bedeutet, eine Schnittstelle zwischen Modellen hat immer zwei Seiten, eine Sender- und eine Empfängerseite. Beide Seiten verwenden unterschiedliche Technologien, teilen aber eine Semantik, d. h. sie operieren in unterschiedlichen „Sprachen“ oder Technologien, aber mit einem gemeinsamen Verständnis z. B. des Informationstransfers.

Diese Dualität einer Schnittstelle soll anhand eines einfachen Beispiels, der Interoperabilität zwischen analogen und digitalen Geräten, gezeigt werden. Jedes Gerät für sich hat den Begriff „Information“ mit eigenen Mitteln implementiert. Während Information im analogen Gerät als Spannung, gemessen in Volt, in Abhängigkeit von der Zeit dargestellt wird, wird im digitalen Gerät Information zeitunabhängig als logisches Symbol „0“ oder „1“ repräsentiert. Bei Interoperation zwischen den Geräten unterschiedlichen Typs wird der aktuelle Spannungspegel des analogen Geräts in ein Symbol {„0“, „1“, „ungültig“} in Abhängigkeit zum analogen Spannungspegels und der definierten Spannungsbereiche für {„0“, „1“, „ungültig“} transformiert. Dieser Vorgang der Typanpassung (AD-Wandlung) wird „coercion“ genannt und an der Schnittstelle automatisch ausgeführt. Wie man leicht erkennt, gibt es für die Typanpassung in jede Richtung eine Abbildungsvorschrift mit den beiden Typparametern Spannungspegel und Spannungsbereiche. Bei Einhaltung aller deklarierter Regeln und Axiome können alle Formen kontinuierlich ineinander transformiert werden. Die Herausforderung besteht hier in der hinreichenden Deklaration aller benötigten Fakten und Transformationsregeln, die eingehalten bzw. geprüft werden müssen, um den Informationsfluss aufrechtzuerhalten. Ausgehend vom Status quo mündet dies zum einen in der Anforderung nach Standards für modellübergreifende deklarative Formate zur Herstellung semantischer Interoperabilität zwischen Modellen (Kapitel 4.5.4, Handlungsbedarf 05-06). Des Weiteren wird der Bedarf für Standards zur Repräsentation semantischer Charakteristiken technischer Prozesse insbesondere unter Einbeziehung KI-basierter Komponenten motiviert (Kapitel 4.5.4, Handlungsbedarf 05-17, 05-07, 05-08).

4.5.3 Mensch und KI

4.5.3.1 Allgemeine Betrachtungen

Das Fundament des Rechtsrahmens wird in Deutschland mit dem Grundgesetz gebildet und durch spezifischere Gesetze und Verordnungen präzisiert. Für KI-Systeme gibt es derzeit noch keine konkrete Ausgestaltung und die EU hat mit dem Entwurf Artificial Intelligence Act (AI Act) einen Vorschlag zur Ausgestaltung vorgelegt (vgl. Kapitel 1.4). Ferner gibt es schon EU-Publikationen wie die „ETHIK-LEITLINIEN FÜR EINE VERTRAUENSWÜRDIGE KI“ von 2019, welche auch auf das Spannungsfeld zwischen Grundrechten und Ethik bei der KI-Anwendung eingehen. In den Leitlinien werden als Ziel für eine vertrauenswürdige KI die drei Eigenschaften rechtmäßig, ethisch und robust angegeben.

Im Rahmen der technischen Normung können Fragen zur Rechtmäßigkeit und zu ethischen Grundsätzen nicht beantwortet werden und auch die EU-Ethik-Leitlinien erkennen an, dass es sich hier um einen politischen Prozess der Meinungsbildung handelt. In einer Normungsroadmap ist es daher notwendig, die Trennlinie zwischen technischer Normung und der politischen Meinungsbildung erkennbar darzustellen. Im Rahmen der technischen Normung ist diese Debatte weniger interessant, da der Mensch, mit Ausnahme der Produktsicherheit, nur indirekte Auswirkungen erfahren hat. Bei der Produktsicherheit war allerdings klar die Zielsetzung im Vordergrund, den Menschen nicht zu schädigen, und technische Normung konnte über den Stand der Technik in der Produktsicherheitsregulierung ihren Beitrag leisten. Normung beschreibt den Prozess, mit welchem sich Werte in KI-Entwicklung und -Betrieb einbeziehen und umsetzen lassen (vgl. Kapitel 4.1.2.2).

Wenn es also um die Robustheit und Sicherheit von KI-Systemen geht, können sich Fachleute in den Normungsgremien einbringen und Verfahren und Anforderungen erarbeiten. Für bestimmte soziotechnische Systeme und Applikationen, wel-

che in die Gesellschaft eingreifen, ergeben sich Fragestellungen jenseits der Produktsicherheit und Ökonomie. Zunächst als unproblematisch erscheinende Anwendungen können bei genauerer Betrachtung gesellschaftlich starke Auswirkungen haben. Als Beispiel sei hier die automatisierte Aufbereitung von Informationen und Nachrichten im Rahmen der gesellschaftlichen Meinungsbildung genannt. Daher erscheint es sinnvoll, dass die Trennlinie für die Normungsarbeit bestimmt wird und den Normungsgremien Anhaltspunkte an die Hand gegeben werden, in welchen Bereichen die technische Normung einen demokratischen Meinungsbildungs- und Gesetzgebungsprozess nicht ersetzen kann.

Zunächst bleibt festzuhalten, dass ein KI-System (oder Algorithmus), das lediglich technische und ökonomische Verbesserungen aufzeigt, nicht immer einen direkten Einfluss auf den Menschen haben muss. Normung trägt jedoch dazu bei, Bewusstsein hierfür zu schaffen und wachzuhalten, wenn standardisierte Prozessschritte bei Entwicklung und Betrieb von KI ethische Implikationen widerspiegeln (vgl. Kapitel 4.1.2.1). Daher richtet sich der folgende Fragenkatalog, der für eine Einschätzung als Grundlage verwendet werden kann, an der Kernfrage aus, wie stark der Einfluss und Zusammenhang mit dem Menschen jenseits der Betrachtung zu Safety ist.

Wird die KI oder der Algorithmus verwendet, um im Zusammenhang mit Menschen …
1. Informationen zu verwalten und aufzubereiten? Beispiele: Nachrichtenfilter, Zensur, Statistiken
2. Entscheidungsvorlagen zu verwalten oder aufzubereiten? Beispiele: Statistische Auswertungen, Vorschlags- oder Optimierungssysteme, Gesundheitsstatus bewerten, Scoring
3. Entscheidungen zu treffen? Beispiele: Wissensbasen, Medikamentierung festlegen, Automatisierung von Verwaltungsakten
4. Entscheidungen oder Maßnahmen zu exekutieren? Beispiele: Automatisierung von Gerichtsverfahren
5. Zwang oder Gewalt auszuüben? Beispiele: Automatisierung polizeilicher Maßnahmen; Sicherheits- und Waffensysteme
6. Krieg zu führen? Beispiele: Sicherheits- und Waffensysteme, strategische Systeme zur Kriegsführung

Der Schweregrad der Auswirkungen steigt in dem Fragenkatalog (nach unten hin) an, wobei es noch eine weitere Dimension bei der Überlegung nach den Auswirkungen gibt, die weniger offensichtlich ist, jedoch sofort leicht überzeugen kann. Das wäre die Frage nach der Möglichkeit der Erkenntnis und Einflussnahme des Menschen auf den verwendeten Algorithmus. Dabei gibt es bei soziotechnischen Systemen oder Systemen, welche die Organisation der Gesellschaft unterstützen, Unterschiede in den beteiligten Rollen, wie sie z. B. aus der Produktsicherheit bekannt sind. Zu den bekannten Rollen Betreibender, Entwickelnder, Unterweisender, Bedienender, Anwendender wird das System auf die betroffene Person oder den normalen Bürger erweitert, der u. U. auch nicht darüber informiert ist, dass ein Algorithmus aktiv war. Aus der Rollenverteilung ist auch zu erkennen, dass eine Konzentration von Verantwortung und Macht auf der Seite der Rollen zu finden ist, welche den Algorithmus kontrollieren, und ein Mangel an Transparenz und Wahlmöglichkeit auf der Seite der Betroffenen. Weshalb diese Dimension durch weitere Kriterien zu diesem Ungleichgewicht zu adressieren ist.

Die KI oder der Algorithmus arbeitet mit Daten und Metadaten von und für …
1. einzelne Menschen.
2. kleine Gruppen oder Gemeinden von Menschen.
3. große Gruppen oder Gemeinden von Menschen.
4. Landes- oder Staatsebene.
5. Länderübergreifend oder global.

Während sich die bisherigen Ausführungen auf die Umstände und Auswirkungen der Verarbeitung von Daten beziehen, die in der EU-Charta für Menschenrechte in Art. 8 berücksichtigt sind, ist dort auch das Thema des Datenschutzes erwähnt, welcher in Deutschland in der Datenschutz-Grundverordnung (DSGVO) geregelt ist. Der Datenschutz ist ein weites Feld und kann hier nicht umfassend beleuchtet werden. Klar ist aber, dass rechtskonforme Daten, welche für einen Algorithmus zur Verfügung stehen, nicht einfach so vorhanden sind und verschiedensten Kriterien, wie z. B. Einwilligung, Begrenzung, Transparenz, Vergessen und Zweckbindung, unterworfen sind.

Darüber hinaus stellt sich bei der Verarbeitung auch immer die Frage, ob die Datengrundlage und verwendeten Algorithmen frei von Diskriminierung und repräsentativ sind. Und

nicht zuletzt, ob die Anwendung der Erkenntnisse auf den Einzelnen[90] angezeigt ist oder vielleicht Themen wie Wahlfreiheit oder Unschuldsvermutung tangiert.[91]

Empfohlen wird daher, dass die oben gezeigten Aspekte als Grundlage dienen sollen, um den Arbeitsrahmen der technischen Normung konkret abzustecken bzw. die angedachten Normungsprojekte so gezielt auszuwählen, dass Konflikte mit den rechtlichen und gesellschaftlichen Aspekten vermieden werden, welche ausdrücklich nicht zum Arbeitsauftrag der technischen Normungsorganisationen gehören.

4.5.3.2 Status quo

Die Anwendung von KI im menschlichen Umfeld birgt ein großes Potenzial und wirft auch viele – derzeit noch offene – Fragen auf. Durch die Erfassung, das Speichern und das Analysieren von Daten können neue abstrahierte Kontextinformationen automatisiert gewonnen und weitergegeben werden. Basierend auf dieser Digitalisierung geht KI noch einen Schritt weiter und ermöglicht es, Prozesse und Abläufe vollständig autonom agieren zu lassen. Die Anwendungsbereiche sind sehr vielfältig und reichen von reinen Softwareapplikationen über autonomes Fahren bis hin zu Medizintechnik, Logistik oder smart manufacturing. Durch die stetig steigende Rechenleistung ist die Realisierung von KI-Systemen mit steigender Komplexität möglich.

Bereits in der ersten Ausgabe der Normungsroadmap KI wurde das Thema „Ethik/Responsible KI“ eingehend diskutiert. Ein zentraler Aspekt ist, dass Menschen Künstliche Intelligenz als Technologie so einsetzen, dass die Rechte und Freiheiten natürlicher Personen gewahrt bleiben. Auf den ersten Blick scheint dieses Problem lösbar zu sein, jedoch ergeben sich bei näherer Betrachtung erhebliche Spannungsfelder, da eine KI eigenständige Entscheidungen treffen darf. Bereits Isaac Asimov (s. a. das KI-Narrativ „vom CYBORG zum Digitalen Zwilling“) hat 1942 in seiner Kurzgeschichte „Runaround“ Grundregeln des Roboterdienstes beschrieben.

90 Die KI macht z. B. eine Ableitung aus den Daten aller Patienten für den einzelnen Patienten. Das kann (wird) für den Einzelnen maximal falsch sein, weil er ja individuell und kein statistisches Mittel ist.

91 Beispiel „Unschuldsvermutung“, wenn eine KI aufgrund von Rasterfahndungskriterien alle Bürger durchsucht und ohne konkreten Anhaltspunkt Maßnahmen veranlasst. Oder einfach nur alle Menschen am Flughafen biometrisch zu erfassen, um einen Straftäter zu suchen. Diese Erfassung ist bereits eine Maßnahme, die für den Einzelnen anlasslos geschieht.

Aktuell lassen sich folgende Probleme identifizieren:

→ Fehlende Transparenz und Erklärbarkeit autonomer Systeme
→ Fragen nach Biases (Vorurteile) und der Fairness von Algorithmen
→ Ethisches KI-Design und der Schutz der Privatsphäre
→ Prüfung der Rechtskonformität, z. B. Datenschutzregulierung mit Werkzeugen des Monitorings, Zertifizierung etc.

Um einen konstruktiven Einsatz von KI als Technologie zu gewährleisten, muss gesichert sein, dass diese dem Menschen als Werkzeug unterstützend dient. Das heißt: Der Mensch setzt KI so ein, dass Grundrechte und -freiheiten gewahrt und gestärkt werden. Das bedeutet weiterhin: Der Mensch wird nicht durch KI instrumentalisiert. Hierbei bestehen Bedenken, dass durch KI ein gläserner Mensch entsteht und viele Arbeitsplätze durch vollständig autonome (KI-basierte) Maschinen ersetzt werden könnten. Zusätzlich ergeben sich Fragen der Haftbarkeit und des von der automatisierten oder gar autonomen Maschine ausgehenden Risikos. Wird beispielsweise ein autonomes Fahrzeug betrachtet, so besteht ein Verhältnis zwischen dem Insassen, dem Herster und der KI. Das Fahrzeug ist im Auftrag des Menschen (dem Insassen) unterwegs, die KI wurde vom Hersteller entwickelt und das Verhalten des Fahrzeugs wird von der KI nach vorgegebenen Regeln bestimmt.

Aus Perspektive des Anthropozentrismus erfolgt die Beschreibung der Fähigkeiten von Künstlicher Intelligenz in der Fachliteratur einerseits auf Grundlage von menschenzentrierten Ansätzen, bei welchen u. a. menschliche Fähigkeiten und Sinne im Vordergrund stehen und nachgeahmt werden. Des Weiteren bestehen Ansätze, die von anthropozentrischen Ansätzen absehen und Fähigkeiten Künstlicher Intelligenz auf Grundlage von physikalischen Charakteristiken wie mechanischen, elektrischen und magnetischen Größen beschreiben.

Die Digitalisierung und die digitale Transformation sind u. a. Gründe für die Anwendung von KI. Im Vergleich zur Digitalisierung geht KI noch einen Schritt weiter und ermöglicht es, Prozesse und Abläufe weitgehend autonom nach den Regeln der semantischen Interoperabilität interagieren zu lassen.

„Mensch und KI“ haben bei industriellen Anwendungen sozusagen die „Pflicht zur Kooperation“, um als Enabler zu wirken. Die Pflicht zur Kooperation ergibt sich aus guten Gründen, z. B. zur Kostenvermeidung bei unkontrolliertem Verhalten (s. Beispiel „Fair Play“), bei Verletzung der Prozess- und Datensicherheit, Safety etc. Neben guten Gründen, die KI zu nutzen,

gibt es natürlich auch Schattenseiten. Ein Beispiel hierfür ist die sogenannte Deep-Fake-Technologie, um Imitationen von Personen in Bild, Video und Ton zu generieren. Das zu unterscheiden mit standardisierten Kriterien und Metriken, ist eine der Herausforderungen bei der Anwendung neuer KI-Technologien.

Ein zentraler Schwachpunkt ist die fehlende einheitliche Definition von KI und damit das Verständnis der unterschiedlichen beteiligten Menschen bei der Nutzung der KI.

Es existiert zwar eine normative Definition von Künstlicher Intelligenz (siehe Kapitel 1.5), welche jedoch sehr allgemein formuliert ist. Demzufolge bietet sie nur einen schwachen Ansatz für die Konkretisierung von Definitionen.

Im Kontext von KI finden bereits Normungsaktivitäten auf internationaler Ebene, auf europäischer Ebene sowie auf nationaler Ebene statt (vgl. Kapitel 3.2).

Grundsätze des ethischen Entwerfens von Maschinen mit KI-Komponenten

IEEE P7000(™) ethically aligned design (EAD) Grundsätze
Der ethische Entwurf einer Softwarekomponente wie z. B. eines „Digitalen Zwillings“ oder der ethische Entwurf einer Hardware-Komponente wie des sogenannten „physikalischen Zwillings“ im Rahmen der drei Achsen „Interoperationen, Anlagenhierarchien, Value Stream“ von Architekturmodellen, wie RAMI4.0, SGAM (Smart Grid Architecture Model) etc. skizziert, soll sich an den Empfehlungen der IEEE P7000™-Serie [64] ausrichten. Die P7000™-Serie enthält sogenannte EAD-Grundsätze für entwerfende und operative Ingenieurstätigkeiten wie z. B., Aufgaben zuverlässig auszuführen (functional safety), Fehleranalysen kritisch zu bewerten (functional reasoning), verhaltensvorausschauend zu berechnen (functional prediction) etc.

Eine ähnliche Rolle wie die IEEE P7000™-Serie [10], [11], [12], [13] spielt der JTC1 SC42/WG3 Standard ISO/IEC TR 24368:2022 [15] „AI Overview of ethical and societal concerns“, in welchem z. B. „Fairness“ in Bezug auf Verhalten oder die Beurteilung von Ergebnissen definiert wird.

Die acht Grundsätze eines sogenannten „ethischen Entwurfs“ (P7000(™) EAD.ed2 [64], [104]), wie sie die IEEE entwickelt hat, beschränken sich auf technische intelligente Systeme, welche fähig sind, sich autonom zu verhalten. Die Fähigkeit zur Autonomie soll den interagierenden Menschen in vielen Bereichen des Lebens stationär oder mobil in einen Zustand des „Wohlbefindens und der Sicherheit“ versetzen können. Diese Forderung steht im Einklang mit den EU-Regulierungen des AI Act, welche insbesondere vermeidbare oder unkontrollierte Verletzungen des menschlichen „Interakteurs“ ausschließen.

EU-Regulierungen AI Act, DGA, DSA etc. in Bezug zu „Mensch und KI“
Mindestens seit 2018 gibt es in der EU ein Nachdenken über eine EU-Strategie für KI [304], [305], die in einem Weißbuch, veröffentlicht Februar 2020, beschrieben steht. In dem Dokument „KI für Europa – Eine Europäische KI-Strategie“ [304] steht, dass KI uns nicht nur das Leben erleichtern kann, sondern sie kann auch dazu beitragen, Herausforderungen wie die Behandlung chronischer Krankheiten, den Kampf gegen den Klimawandel oder die Antizipation von Bedrohungen der Cybersicherheit zu meistern.

KI bezeichnet daher „Systeme mit intelligentem Verhalten“ [304] als Systeme, die ihre Umgebung analysieren (d. h. „zu verstehen“ zu versuchen), um ein bestimmtes Ziel zu erreichen. Sobald KI-Anwendungen gut funktionieren und sie ausreichend qualifizierte Daten erhoben haben, können Entscheidungen, mit abzuschätzendem Risiko zwar, automatisiert werden.

Auf der DIN/DKE-Fachkonferenz im November 2021 [305] wurden „EU AI Policy und das Weißbuch der KI“ und wie sich Unternehmen einbringen können, vorgestellt, z. B. mit der Frage, wie KI regulatorisch begriffen werden kann, indem die ethischen Bürgerrechte und das sogenannte **New Legislation Framework (NLF)** Berücksichtigung finden. Zur Vorbeugung von unterschwelliger Manipulation oder zur Vermeidung biometrischer Fernidentifizierung fordert der AI Act, eine ex-ante-Konformitätsbewertung für KI-Anwendungen zur Pflicht für alle KI-Anbietenden (supplier) zu machen.

In Ergänzung zum geplanten **AI Act** beschäftigt sich der **Data Governance Act (DGA)** [306] mit der Verfügbarkeit von Daten und wie Vertrauen in sogenannte „data intermediaries“ mitsamt ihren data-sharing-Mechanismen EU-weit gesteigert werden kann. Die data intermediaries korrelieren hierbei mit den zu gestaltenden „data spaces“.

Das Ziel des DGA ist es, die Stakeholder in allen KI-Anwendungsfeldern zu motivieren, ihre Daten gegen ein anderweitig lohnendes Entgelt der Öffentlichkeit zur Verfügung zu stellen

und gleichzeitig ihre Datenschutzrechte und Rechte Dritter zu wahren.

Im **Digital Services Act (DSA)** [307] wird ein Handlungsbedarf, der sich aus dem regulatorischen Rahmen für KI-Anwendungen ergibt, ebenso benannt. Es soll sichergestellt sein, dass KI-Systeme im Markt sicher und zuverlässig verwendet werden können und dass sie die gegebenen Grundrechte und Werte der EU respektieren, d. h. implementiert haben.

Alle KI-Produkte und -Systeme sollen so gestaltet sein, dass sie den Marktteilnehmern Gewissheit geben, ihre Investitionen zu tätigen und fortgeschrittene KI-Innovation in den Markt bringen zu können. Data-Governance-Maßnahmen und die Durchsetzung gegebener Gesetze, Grundrechte und Sicherheitsanforderungen an KI-Systeme sollen angepasst und verbessert werden. Der Europäische Binnenmarkt soll in Richtung eines gemeinsamen Marktes für gesetzestreue, sichere und vertrauenswürdige KI-Systeme vorbereitet werden, um eine Marktzersplitterung zu vermeiden.

Technologisch-ethisches Narrativ, vom CYBORG zum Digitalen Zwilling

Der hybride Organismus zwischen Mensch und Technologie, sogenannte „AI Machines (AIMs)“, wurde in den 1960er-Jahren erfunden, als der Mensch zum ersten Mal versuchte, die Erde zu verlassen, um sich entfernte Welten anzueignen. „Cyborg and Space“ war damals zwar Science-Fiction, es findet heute aber seine Renaissance in der Nutzung der KI-Technologie in vielen industriellen Bereichen, sogenannten smart spaces, wieder.

Der Digitale Zwilling, als **AI Machine** gedacht, war damals wie heute ein hybrides, nicht immer durchschaubares, d. h. nicht transparentes, u. U. nicht vertrauenswürdiges Mensch-Maschine-Produkt, das man besser einzuhegen wünschte.

Beispiel: So hat Isaac Asimov bereits 1942 in seiner Schrift „Runaround“, um befürchteten Konflikten mit den AIMs vorzubeugen, die ersten drei ethischen Sicherheitsregeln für hybride KI-Komponenten (AIM) entworfen (welche erst 1982 in deutscher Sprache erschienen sind):

- → Einer AIM ist es nicht gestattet und sie darf sich selbst nicht instruieren, einen Menschen zu verletzen;
- → eine AIM muss Befehle des Menschen immer ausführen, vorausgesetzt, Regel 1 wird nicht verletzt;
- → eine AIM muss sich schützen können, vorausgesetzt, Regel 1 und Regel 2 werden nicht verletzt.

Das AIM-Beispiel in diesem Narrativ dient der Illustrierung eines in der technologischen Entwicklung frühzeitig erkannten Handlungsbedarfs, die Kontrolle des Menschen über komplexe nicht-transparente Technologien unbedingt beizubehalten bzw. nicht zu verlieren. Heute gibt es vielversprechende Ansätze, mit den verschiedenen EU-Regulierungen, sich überlappenden Bereichen der Digitalisierung, KI, ML, Datennutzung, -verteilung und -sicherheit etc., auf diese drängenden Fragen eine Antwort zu geben.

Mit dem Einsatz KI-basierter Robotertechnologien tritt die Automatisierung aus dem Hintergrund des bloßen Engineerings heraus in das Licht der Ansprüche eines ethisch ausgerichteten Designs (EAD). Automatisierung ist längst nicht mehr nur eine Aufgabe der Regelungstechnologie für Ingenieur*innen, sondern eine gesellschaftliche Aufgabe, sobald sich die Automatisierung den Fragen der ethischen Regulierung und des Entwurfs von Robotern und mobilen Geräten öffnet. Neue Steuerungskonzepte, wie sie der Digitale Zwilling oder die sogenannte Verwaltungsschale für wertvolle Assets, z. B. industrielle Produktionsanlagen oder industrielle Produkte, darstellen, werden heute in der Normung und Wissenschaft diskutiert bzw. stellen einen Handlungsbedarf dar.

Die anderswo dargestellte Aufgabe, die Semantik bzw. die Funktionalität eines RAMI4.0/SGAM-konformen Systems während der Laufzeit zu kontrollieren und zu optimieren, kann der Digitale Zwilling [308], ggf. ausgestattet mit KI-Komponenten, übernehmen. Der Digitale Zwilling leitet Steuersignale aus dem Ist-Soll-Vergleich ab und sendet sie als Korrektursignale an die entsprechenden Elemente bzw. Artefakte in den (lifecycle) Domänen und (hierarchy level) Zonen des betrachteten Systems zurück.

4.5.3.3 Anforderungen und Herausforderungen an die Semantik KI-basierter Systeme

Das semiotische Dreieck stellt die holistische Sicht auf die semiotische Dreiecksbeziehung zwischen einem antizipierten Ding, Gerät oder Asset, seiner ontologischen prä- oder deskriptiven Beschreibung der Charakteristiken und der möglichst vollständigen Semantik bzw. Konzepte des betrachteten Dings dar. Die Beziehung wird holistisch genannt, weil sie die drei für das Verständnis mindestens notwendigen Darstellungen und ihre Beziehungen zueinander enthält. Es ist das Ding an sich, das konstruktiven Anforderungen gehorcht, um zu funktionieren; es sind die Standards und Anforderungslis-

ten, um Dinge industriell mit unterschiedlichen Maschinen an unterschiedlichen Orten fertigen zu können, und es ist das Verständnis, in Relation zur Sprache, das ausgedrückt werden muss, um die Semantik, Funktionen, Datenflüsse, Strukturen, einschließlich der Nutzung des Geräts, bauen und dimensionieren zu können.

Methodik des Semiotischen Dreiecks

Das semiotische Verfahren zur Kontrolle automatischer und autonomer Prozesse dient u. a. zur Darstellung von de- und präskriptiven Anforderungen in Ontologien und Standards, von Eigenschaften wie trustworthiness oder der Qualitätsanforderungen an die Herstellung von Produkten.

„Jedes Produkt hat drei Seiten", die sich aufeinander beziehen und daher zusammen dargestellt bzw. implementiert werden müssen. Das betrachtete Objekt oder Produkt (z. B. ein Heizgerät, das in einem linearen Arbeitsbereich von minus x bis plus y Grad funktioniert), besteht aus physischen Artefakten (Bauteilen) und folgt einer physikalischen Funktion bzw. erfüllt einen bestimmten Zweck (nämlich es erzeugt Wärme aus Elektrizität). Dieser physikalisch funktionale Zweck wird mit semantischen Artefakten (z. B. nichtlineare thermodynamische Gleichungen) eindeutig dargestellt und ggf. für einen bestimmten Arbeitsbereich gelöst. Bei dieser Lösungsaufgabe müssen die Anforderungen von gegebenen Standards und technischen Berichten in Bezug auf Konformität, Sicherheit, Zuverlässigkeit, Erklärbarkeit und Nachvollziehbarkeit etc. besonders bei nichtlinearen Vorgängen und Prozessen oder „nicht sicherem" Verhalten von KI-Komponenten, die Ingenieur*innen ggf. zur Lösung verwenden, Berücksichtigung finden.

Am Ende eines „Ethically Aligned Engineerings" für die Produktion eines Heizgeräts (um im Beispiel zu bleiben) sind alle drei semiotischen Sichten aufeinander abgestimmt: Eine eindeutige (mathematische) Semantik erklärt den Zweck und die Funktion des Geräts, die Standards liefern eine vollständige Liste der zu erfüllenden (Sicherheits-, Qualitäts-)Anforderungen an das Gerät und ggf. auch Use Cases für den Betrieb der Heizung. Das Gerät bzw. der Gerätetyp wird vom Ingenieur bzw. der Ingenieurin so entworfen, dass das Gerät seinen Zweck erfüllt und zuverlässig in seinem Arbeitsbereich und im gesamten Lebenszyklus funktioniert [309].

Architecture of Choice

„Architecture of Choice" ist der sprachliche Begriff, Dinge, Bilder, Daten, Informationen, die zur Auswahl stehen, so automatisiert anzuordnen, dass es Menschen erleichtert und u. U. auch erschwert wird, eine bestimmte Auswahl zu treffen. Zum Beispiel ist es für den betroffenen Menschen von Bedeutung, welche Informationen und Daten von einem automatisierten, teilautonomen Fahrzeug ihm zur Entscheidung vorgelegt werden. Diese Daten müssen vom Menschen auf Vertrauensbasis bewertet und entschieden werden können.

Die Intransparenz von eingebauten KI-Komponenten und -Verfahren macht es in der Regel schwer, die Ergebnisse eines autonomen Fahrzeugs oder einer automatisierten Produktionsanlage vertrauensvoll beurteilen zu können. Daher sind genormte Verfahren erforderlich, um das Risiko einer gewollten oder ungewollten Verschiebung der Zielkoordinaten oder die Veränderung des Ergebnisses eines Arbeits- oder Produktionsschrittes oder das Verhalten eines Menschen an der Mensch-Maschine-Schnittstelle (HMI) abschätzen zu können.

Genormte Verfahren und Metriken erhöhen für Nutzende die Vergleichbarkeit der Feststellung der Qualität von Produkten oder Diensten, als auch bei der Bewertung eines Herstellungsprozesses, darin Vertrauen zu haben oder nicht.

An der HMI kann der Mensch mit beabsichtigten Maßnahmen zur Beeinflussung seines Verhaltens dazu bewogen werden, Entscheidungen zu treffen, die im Interesse eines verborgenen Dritten, z. B. eines Unternehmens, das hinter der verwendeten HMI für Clouddienste steht, sind. Ein anderes Beispiel wären Versuche, gesellschaftliche Interessen des Umwelt- oder Klimaschutzes etc. durchzusetzen.

Eine der Herausforderungen einer „Architektur der Auswahl von Optionen" (architecture of choice) oder anders ausgedrückt einer Architektur zur Unterstützung der Entscheidungsfindung an einer HMI-Schnittstelle besteht in der Präsentation der Auswahloptionen an der HMI-Schnittstelle im Hinblick auf die Auswirkungen der Entscheidungen auf gegenwärtiges und künftiges Geschehen im Umkreis der Menschen, die Entscheidungen zu tätigen haben. Diese Menschen bevorzugen i. d. R. „naheliegende" Entscheidungen mit Auswirkungen auf den aktuellen Zustand, gegenüber Entscheidungen mit Auswirkungen auf mögliche künftige Zustände, z. B. die Vermeidung eines Fehlerzustands.

Menschen bzw. Personen unterscheiden sich signifikant in der Perzeption von Information und ziehen daher sehr unterschiedliche Schlüsse aus vergleichbarer Präsentation einer Auswahl von vorzunehmenden Entscheidungen.

Eine weitere Herausforderung einer „Architektur zur Unterstützung von Entscheidungsfindungen“ ist die offene Frage, wie menschliche kognitive Entscheidungsfindungsprozesse vonstattengehen und wie Architekturkonzepte die Qualität der Entscheidungsfindung unterstützen und verbessern können oder dem zuwiderlaufen.

Smarte Technologien – smarte Fähigkeiten des Menschen?
Der von der Normung empfohlene Entwicklungsprozess, der sich u. a. auf EAD (s. IEEE P7000(™) [64]) stützt, gibt Anleitung, wie die Werte über eine ethische Abwägung schon beim Entwurf neuer technischer Standards auf das Design Einfluss nehmen können.

Der Begriff der smarten Technologien wird vor allen Dingen im Bereich industrielles Internet der Dinge (s. JTC1 SC41 IIoT) verwendet, um auszudrücken, dass die „Dinge ihre Umgebung kennen“, weil sie mit Sensoren und Aktoren, z. B. das „Ding“ einer automatisierten Produktionsanlage, ausgestattet sind. Die Produktionsanlage kann bis zu einer bestimmten Qualität mit ihrer Umgebung interagieren. Das macht sie zu einer smarten Produktionsanlage, weil sie, bevor sie den nächsten Schritt ausführt, alle gegebenen Sicherheitsanforderungen, z. B. keine Personen im Sicherheitsbereich, im HMI-Bereich prüft.

Fertige und künftige Standards beeinflussen Design, Implementierung von Prozessen und Maschinen (von Dingen) besonders bei Verwendung neuer Technologien, wobei die von den EAD-Standards geforderten Maßstäbe und Transparenzanforderungen normativ verstanden werden sollten.

Die Fähigkeit zu semantischer Interoperabilität in komplexen **Systemen-von-Systemen** (very large-scaled systems) zwischen Dingen, Maschinen, Menschen, Modellen und Prozessen sollte von Normung, Wissenschaft und Politik mit Standards, Regulierungsvorschlägen und Forschungsvorhaben vorangetrieben werden.

Die Aufrechterhaltung eines kontinuierlichen Austauschs von Daten, Werten und Wissen, von Informationen über Energie oder Produkte, zwischen Kulturen und Systemen erfordert Kontrolle und Steuerung (governance) des **Wertzuwachsstroms** (flow of value) zwischen den verschiedenen Modellvorstellungen, heterogenen Systemen oder Kulturen.

Informationsmodelle sind meist **Schichtenmodelle**, wie es z. B. im SGAM oder RAMI4.0 und anderen Referenzmodellen dargestellt wird. Im Gegensatz zur semantischen Interoperabilität, wo es um die Aufrechterhaltung eines Werteflusses geht, geht es im Schichtenmodell um Formate und Protokolle in eher syntaktischen Kategorien.

Diese HMI-Faktoren sollen durch Anwendung des sogenannten Human Factor Engineerings (HFE), nach der neuen Norm DIN IEC 63351, VDE 0491-61 [310], bei heterogenen System-zu-System-Kooperationen genutzt werden können.

4.5.4 Normungs- und Standardisierungsbedarfe

Bedarf 05-01: Erstellung eines Referenzmodells für KI-Engineering
Schaffung eines gemeinsamen Grundverständnisses der Begriffe sowie der Zusammenhänge der verwendeten Konzepte als Hilfestellung für den/die Ingenieur*in in Zusammenarbeit mit Informatiker*innen und Datenwissenschaftler*innen.

Definition und Erläuterung von Begriffen und Konzepten und deren Zusammenhänge zum System Engineering unter besonderer Berücksichtigung des Einsatzes von KI-Methoden in Subsystemen; ggf. Aufbau eines formalen Modells (z. B. UML, Ontologie, ...)

Bedarf 05-02: Liste und Definition von nicht-funktionalen Merkmalen (Qualitätskriterien) für KI-basierte Systeme, bezogen auf die Entwicklung und den Betrieb
Schaffung eines einheitlichen Verständnisses für Stakeholder (z. B. Systemanforderer, Systemingenieur*in), Aufbau eines einheitlichen Rechtsrahmens, Schaffung von Rechtssicherheit für das Systemverhalten und die Zertifizierbarkeit.

Definition und Beschreibung der Bedeutung für kennzeichnende Merkmale wie Akzeptanz, Verlässlichkeit, Zuverlässigkeit, Planbarkeit, Kontrollierbarkeit, Erklärbarkeit, Cybersicherheit (Security), funktionale Sicherheit (Safety), Unsicherheit.

Bedarf 05-03: Einheitliche Vorgehensweise für die Bewertung von KI-basierten Systemen gemäß definierten Kriterien
Definition von allgemeingültigen Kriterien und Workflows zur Abnahme und zum Vergleich der Leistungsfähigkeit von KI-basierten Systemen.

Beschreibung wesentlicher Arbeitsschritte im Workflow und der Anwendung von Bewertungskriterien, insbesondere bei hochkritischen Systemen gemäß dem Entwurf zum AI Act der EU.

Bedarf 05-04: Vorgehensmodell für das Engineering und den Betrieb von KI-basierten Systemen
Entwicklung einer Hilfestellung für den/die Systemingenieur/in, wie KI-basierte Systeme grundsätzlich entwickelt, betrieben und gewartet werden sollen.

Definition einzelner Prozessschritte für Entwicklung, Test, Abnahme, Betrieb, Wartung. Beschreibung der Struktur des Systems und der Subsysteme sowie der KI-basierten Teile. Angaben zur vorteilhaften Anwendung von agilem vs. linearem Vorgehen, definierte Designartefakte, Angaben zur Dokumentation.

Bedarf 05-05: Aufbau einer standardisierten Metadatenbeschreibung von KI-Methoden
Schaffung von Möglichkeiten für den Aufbau von Lösungsräumen (u. a. Kataloge) für Anforderungsmuster.

Definition eines Ordnungsrahmens und Klassifikation von KI-Methoden, Formulierung von semistrukturierten Anwendungsfällen und Ableitung potenzieller KI-Methoden zu deren Lösung.

Bedarf 05-06: Auszeichnungen von Datenstrukturen und Modellen zu Verwendung, Erhalt und Rekonstruktion ihrer ursprünglichen Intentionen
Verschiedene Parteien (Werkzeuge, Systeme, Wissensingenieure) sollen gleiche Modelle mit deckungsgleicher Interpretation ihrer Semantik nutzen können, um Abweichungen und Verluste bei der Verarbeitung zu vermeiden. Dadurch sollen ursprüngliche Intentionen von Datenstrukturen und Modellen durchgängig ausgedrückt, weitergegeben und rekonstruiert werden. Dies ermöglicht eine verlustfreie Anwendung von Modellen und deren Validierung auf konsistente Interpretationen über mehrere Parteien entlang einer Verarbeitungskette („Pipeline“).

Es soll daher eine validierbare Semantik der Intentionen von Strukturen und Modellen über verschiedene Parteien entlang von Pipelines hinweg definiert werden. Dazu werden robuste Vorgehensweisen und Mechanismen beschrieben und definiert, anhand derer die Intentionen von Datenstrukturen, Mustern und Modellen ausgezeichnet, erhalten und validiert werden können.

Bedarf 05-07: Validierbare Transformationen von Strukturen und Modellen
Transformationsmechanismen von Werkzeugen und Systemen für den Import und Export von Strukturen und Modellen sollen transparent und überprüfbar sein, um Veränderungen der transformierten Inhalte erkennen zu können sowie Fehlinterpretationen zu vermeiden. Dadurch sollen Werkzeuge und Systeme dediziert gemäß ihren Fähigkeiten angesprochen und getestet werden können. Zusätzlich bietet ein solches Verhalten die Möglichkeit, von außen zu erkennen, ob ein Werkzeug ihm angebotene Inhalte verlustfrei verarbeiten kann.

Transformationsmechanismen sollen deshalb über entsprechende Kapselung ihre Fähigkeiten und Datenstrukturen/-formate validierbar bekannt geben, sodass entlang einer Kette von Transformationen vorab die Semantik des Ergebnisses ersichtlich ist. Dazu sind Auszeichnungsmechanismen und -strukturen auf Schnittstellenebene zu definieren, anhand derer sich Transformationsmechanismen verstehen, einordnen und testen lassen.

Bedarf 05-08: Identifikation und Behebung struktureller Probleme in den Grundbausteinen für kompatiblen Daten-/Modellaustausch und KI
Alle Verarbeitungsebenen entlang von aufeinander aufbauenden Grundbausteinen für Daten-/Modellaustausch und KI („Stacks“) erfordern eine ganzheitliche auditierbare Konformität verwendeter syntaktischer Strukturen. Aktuell werden je nach Stufe der adressierten Semantik in Stacks etwa bestimmte (Daten-)Strukturen erlaubt oder verboten (Beispiel: der W3C Semantic Web Stack erlaubt auf RDF-Ebene syntaktisch Strukturen, die auf darauf aufbauender OWL-Ebene nicht mehr zulässig sind). Dies führt dazu, dass ein „Label der Stack-Konformität“ für Werkzeuge und Pipelines nicht ausreicht. Zur Ausführung eines KI-Mechanismus erforderliche Inhalte müssen aber bei auditierbarer Konformität verwendeter Werkzeuge mit vorgegebenen Anforderungen und Stacks verlustfrei beigesteuert werden.

Für Transformationsmechanismen entlang von Stacks sollen Prüfmerkmale definiert werden, anhand derer Inhalte automatisch auf Nutz- bzw. Interpretierbarkeit mit den jeweils höheren/niedrigeren Stufen der Stacks geprüft werden können. Dazu wird vorgeschlagen, Stacks hinsichtlich der vertikal durchgängigen Interpretierbarkeit von Inhalten zu untersuchen und für die jeweilige Überbrückung von Stufen entsprechend standardisierte Transformationen zu definieren. Diese sollen sich anhand ihrer Semantik auch mit anderen Stacks

kombinieren lassen, sodass ein semantikerhaltender Transport von Inhalten hin zu KI-Mechanismen über verschiedene Stacks hinweg gewährleistet werden kann.

Bedarf 05-09: Definition von Metriken und Methoden zur Bewertung der Datenqualität u. a. in ML-Datenmodellen
Die Datenqualität ist ein entscheidender, auch wirtschaftlicher, Einflussfaktor, sobald Transaktionen über die Datenmodelle ausgeführt werden. Heute fehlen standardisierte Methoden, um dieses Merkmal zu ermitteln, als auch Metriken, um die Datenqualität bewerten zu können. Eine Aussage zur Datenqualität führt zu einer Aussage zur Modellqualität und damit zu einer erfolgreichen KI-Umsetzung.

Es wird vorgeschlagen, Methoden und Metriken zur Datenqualität einzuführen und Mechanismen zu definieren, mit denen dieses Merkmal validiert werden kann.

Bedarf 05-10: Skizzierung einer spezifischen I4.0-Methodik für den Entwurf von I4.0-Systemen mit KI-Komponenten
Bedarf für eine I4.0-Methodik ergibt sich aus Anforderungen einer einheitlichen semantischen Betrachtung von I4.0-Systemen und von Industrieanlagen samt Daten, Vorgängen und Kriterien für die Interoperabilität zwischen Mensch und Maschine und Maschine-Maschine. Dazu gehört u. a. sprachliche Ausdruckskraft zur ontologischen Charakterisierung eines Produkts oder Verfahrens.

Ziel der I4.0-Methodik ist es, ein Vokabular mit Anwendungsregeln zu haben, womit formale und vom Rechner ausführbare Ontologien erstellt und die von Mensch und Maschine jeweils auf ihre besondere Art „verstanden“ (d. h. logisch vom Menschen und operational von der Maschine) und verwendet werden können.

Bedarf 05-11: Standardisierung und Katalogisierung von allen nach dem Schema Ding-Ontologie/Symbol-Semantik kategorisierten Artefakten und ihre Sammlung in stakeholderspezifischen Katalogen für Designer, Entwickler, Operateure etc.
Die Semantik von Anwendungsszenarien soll in einer sowohl vom Menschen nachvollziehbaren Art und in einer von der Maschine ausführbaren Art dargestellt werden können. Das ist mit der Verwendung von Graph- und Datentypen der Fall. Folgen von beobachtbaren datenverarbeitenden Ereignissen werden also für die Beschreibung von I4.0-Herstellungsprozessen und Produkten verwendet. So können die Anforderungen einer bestimmten Erzählung eines Anwendungsszenariums einerseits anschaulich als Graph-Trajektorie und andererseits semantisch klar dargestellt werden. Ein Beispiel für ein I4.0-Narrativ (formal dargestellt als Graph-Trajektorie mit Zielzustand) ist der „value flow“ in den gegebenen Referenzarchitekturmodellen.

Eine I4.0-Methodik bietet standardisierte Werkzeuge und Artefakte u. a. zur Gestaltung von Anwendungsszenarien, Anwendungsbeispielen oder zum Schreiben von Narrativen an. Narrative zeichnen sich dadurch aus, dass sie ein prüfbares Ziel oder eine validierbare Absicht, z. B. eine erfolgreiche Qualitätskontrolle in der Produktion, mit der Herstellung eines Produkts verbinden. Alle Schritte, die unternommen werden, um das gesetzte Ziel zu erreichen, müssen dokumentiert werden. Dazu stehen Metadaten-Artefakte zur Verfügung. Die Vergleich- und Nutzbarkeit der katalogisierten Artefakte ergibt sich aus der Menge der angewendeten Regeln (d. h. der Semantik) zur Gestaltung oder zum Entwurf eines Dings oder Asset.

Bedarf 05-12: Formalisierung von Metriken, Evaluationen, Testing, Verifikation und Modellbildung
Da nur rudimentäre Konzepte eines gemeinsamen Verständnisses oder einer Sprache in der vertikalen Normung und in I4.0-Branchen zu beobachten sind, driften auch die Bewertungskriterien zur Prüfung von Functional-Safety- oder Security-Anforderungen auseinander. Daher gibt es großen Handlungsbedarf bei der Normung von Bewertungsschemata und -kriterien.

Eine gemeinsame Sprache erlaubt es, gemeinsame Darstellungs- und Bewertungsmaßstäbe zu etablieren. Die Common Logic/Semantics umfasst sprachliche, ontologische und logische Kategorien von Artefakten, womit z. B. die Digital-Twin-Modellbildung oder eine semantiktreue, d. h. verhaltensmäßig korrekte Implementierung (im Vergleich zum Modell) der cyberphysischen Wirklichkeit überprüft werden kann.

Bedarf 05-13: Prüf- und Evaluierungsmethoden für Assets mit eingebauten KI-Komponenten zur Abschätzung des Einflusses der KI auf die System- oder Komponentenqualität
KI oder ML werden als Werkzeuge betrachtet, die, in Assets eingebaut, die Qualität der Assets verändern können. Daraus ergibt sich Bedarf, zu prüfen, inwiefern Qualitätsveränderungen auf die Functional Safety eines Assets Einfluss haben.

Bedarf ergibt sich aus der Abschätzung der Auswirkung von neuen (KI-)Werkzeugen und Komponenten, eingebaut in Fertigungsanlagen und Produkte, bezüglich der Beziehungen zwischen Mensch und Maschine z. B. auf gemeinsam durchzuführende Aufgaben, auf die Qualität der so hergestellten Produkte etc.

Bedarf 05-14: Beschaffung von Argumenten und Metadaten, die zur Belegung der Vertrauenswürdigkeit der Maßnahmen beteiligter Stakeholder verwendet werden können
Vertrauenswürdigkeit ist nicht immer nur ein Problem der Produktqualitätsvermessung, sondern oft ein Problem der Produktnutzung, in welcher Transparenz und Selbsterklärbarkeit des Eingabe-/Ausgabeverhaltens eine Rolle spielen. Daher gibt es Bedarf an Methoden, die Wirksamkeit der Kontrolle über das Produkt oder die Produktionsstätte zu verifizieren.

Der Bedarf, die Vertrauenswürdigkeit eines Produkts oder Verfahrens zu belegen, wandelt sich u. a. mit veränderlichen Technologien und Verfahren, z. B. angewendet in einer Produktionsanlage. Es ist also ein permanenter Prozess der Erneuerung, der auch eine permanente Überprüfung der Zusicherung der Vertrauenswürdigkeit erforderlich macht.

Bedarf 05-15: Aufbau und kontinuierliche Aktualisierung einer (semantischen) Normungslandkarte mit eingebauten Hilfen zur Nutzung der Landkarte
Normen werden oft zur Gestaltung von Anlagen und Produkten isoliert betrachtet und geschrieben, ohne tiefere Kenntnisse oder Bezüge zu anderen relevanten horizontalen und vertikalen Standards, weil ein semantisches Koordinatensystem für Normen in der Normungslandschaft fehlt.

Eine standardisierte Form einer gemeinsamen Darstellung von Semantik kann hilfreich sein bei Versuchen, zusammenhängende Normungsthemen in einer heterogenen Normungslandschaft, wie sie sich z. B. aus dem RAMI4.0 ergibt, aufzufinden und ggf. zu prüfen.

Bedarf 05-16: Internationalisierung und Digitalisierung von Normen für Neue Technologien, um die automatisierte Auswertung von Systemanforderungen unterstützen zu können
Eine in „Normungsenglisch“ geschriebene Norm wird i. d. R. bis zu ihrer Erfüllung von verschiedenen Stakeholdern mehrmals „übersetzt“ oder „rückübersetzt“. Eine digitale Norm dagegen kann die zu implementierenden Anforderungen z. T. verarbeitbar in Maschinen in einer Sprache der Computational Logic bereitstellen. Diese Teile können Daten, Prozesse und „Wissens“-Datenbasen zur maschinellen Datenverarbeitung oder Entscheidungsfindung enthalten.

Bedarf 05-17: Entwicklung einer gemeinsamen (I4.0-) Sprache, die es erlaubt, ein System in unterschiedlichen Viewpoints, aber in einheitlicher regelorientierter Darstellung zu beschreiben
Üblicherweise werden beim Schreiben von Normen und Standards mehrere Logiken (sogenannte Viewpoints) gleichzeitig verwendet. Allen Logiken gemeinsam sind Gesetze, Axiome, Ableitungsregeln für ihre spezifischen Domänen, die sie abbilden. Beispiele für „Logiken“ sind Produkthaftung, Safety, Security, Privacy, Funktionalität, Interoperabilität, Qualitätsangaben, Produktionsmaße etc.

Bedarf 05-18: Katalogisierung von technischen, semantischen und juristischen Begriffen bzw. Artefakte zur konstruktiven Synthese von AAS-Teilmodellen
Es ist notwendig, eine gemeinsame Sprache (d. h. gemeinsame Darstellung der Semantik) mit Regeln, Gesetzen, Axiomen branchenspezifisch und Teile davon branchenübergreifend zu definieren, um damit u. a. Anwendersicherheit bei der Kooperation zwischen Mensch und Maschine, die von einer eingebetteten KI gesteuert wird, zu regeln und zu gewährleisten.

Bedarf 05-19: Standardisierung der Aspekte des Ökosystems „Mensch & KI“
Mensch & KI entwickeln sich zu einem Ökosystem mit Auswirkungen auf gesellschaftliches, wirtschaftliches, privates und arbeitsplatzbezogenes Handeln des Menschen und seiner Kooperation mit Maschinen, die sich auf KI stützen.

Um das Ökosystem transparent zu gestalten, ergibt sich folgender Normungsbedarf:
- → Aufzeigen der gegenseitigen Wirkungen KI vs. Mensch.
- → Beschreiben und Definieren von Verantwortlichkeiten der „KI“ und des Menschen in unterschiedlichen Rollen und Kollaborationen.
- → Beschreiben und Definieren von Szenarien im mehrdimensionalen Zusammenspiel von KIs und Menschen.

Bedarf 05-20: Realisierung und Umsetzung des Digital Service Act (DSA) im Ökosystem „Mensch & KI“ in verschiedenen vertikalen Anwendungen und Datenräumen
Der DSA enthält 35 Artikel, gruppiert in acht Kapitel, Anleitung, um Datenräume aus privatem Datenbesitz aufzubauen.

In diesen Datenräumen können KI-Komponenten zur Auswertung der verfügbaren Daten eingesetzt werden.

Normungsbedarf ergibt sich für die von der Wirtschaft benötigten Datenräume, gefüllt mit Wirtschafts-, Produktions- und Arbeitsdaten aus privatwirtschaftlicher Hand. Dieser Kooperationsprozess zwischen gebenden und nutzenden Stakeholdern muss mit Standards, Regelungen und Gesetzgebung zum Nutzen aller geformt werden.

Bedarf 05-21: Standardisierung von Methoden zur Ursprungsbewertung und Beherrschung der vertikalen Datenräume
Das Narrativ der KI zeigt, dass formalisierte Regeln die Grundlage zur Beherrschung und zum Verständnis komplexer Prozesse sind, wobei die Elemente von KI-gestützten Systemkomponenten und Datenräumen mittels Metadaten und Metaregeln sichtbar und erklärbar gemacht werden. Metadaten und -regeln stellen Wissen über die Entstehung von Dingen und Produktionsdaten bzw. über Funktionen der Dinge und Nutzung von Daten dar.

Bedarf 05-22: Neue Normungsprojekte für „formale und semiformale“ Standards zur semantischen Konkretisierung technischer Themen und dem Verhalten von Systemen, die im Rahmen der technischen Normung zu leisten sind
Mit formalen und semiformalen Standards sind Normungstexte gemeint, die teilweise oder ganz „computerisiert“, also mit einem Rechner ver- und bearbeitbar sind. Ein Beispiel dafür wäre der „Digitale Zwilling“.

Viele Themen und Aussagen in der Normung und Regulierung betreffen sich überschneidende Kompetenzbereiche. So ist z. B. der politische Meinungs- und Gesetzgebungsprozess für die regulative Ausgestaltung maßgeblich, der den konkreten Rahmen für bestimmte technische Anwendungsfälle und Zielstellungen gibt (z. B. Geräte, die auch einen militärischen Nutzen haben können).

Bedarf 05-23: Die Anwendung von ethischen Regeln soweit wie möglich und in Kooperation mit demokratischen Institutionen national und in der EU in der Normung behandeln
Das Thema Ethik und dessen Anwendung ist Bestandteil des politischen Meinungs- und Gesetzgebungsprozesses. Ethische Fragestellungen können nur teilweise oder gar nicht über Normungsverfahren erarbeitet werden. I. d. R. bilden sie einen Kernaspekt in demokratischen Prozessen und werden international unterschiedlich gehandhabt.

Daher sollten gemeinsame Normungsbemühungen angestrebt werden, um ethisch begründete (normative) Verfahren zu Regulierungen von AI biases, sludging, nudging, ethically-aligned design etc. zu finden.

Bedarf 05-24: Klare Definition von High-Risk-KI-Systemen und Abgrenzung zu Safety-Systemen
High-Risk-KI-Systeme (im Sinne des Vorschlags der EU-Kommission für einen AI Act) können auch Systeme sein, welche nicht als Safety-Systeme gelten. Allerdings gelten ähnliche Anforderungen, falls es vom Gesetzgeber gewünscht ist, künftig alle High-Risk-KI-Systeme als Safety-Systeme (im Sinne von Fail-Safe, funktionale Sicherheit) auszuführen. KI und Datenmodellregulierungen ergänzen die technischen Anforderungen, die in der Normung angegeben werden, im Sinne einer „roten Linie“, die ethisch und juristisch nicht überschritten werden soll.

Im Rahmen der laufenden Diskussion um den geplanten AI Act muss geklärt werden, ob künftig alle High-Risk-KI-Systeme als Safety-Systeme auszuführen sind, da der Gesetzesvorschlag unterschiedliche Anforderungen für High-Risk- und Safety-Systeme vorsieht.

Die Arbeitsgruppe Industrielle Automation hat die identifizierten Bedarfe nach der Dringlichkeit ihrer Umsetzung bewertet. Abbildung 40 zeigt die Dringlichkeit der Umsetzung, kategorisiert nach den Zielgruppen Normung, Forschung und Politik.

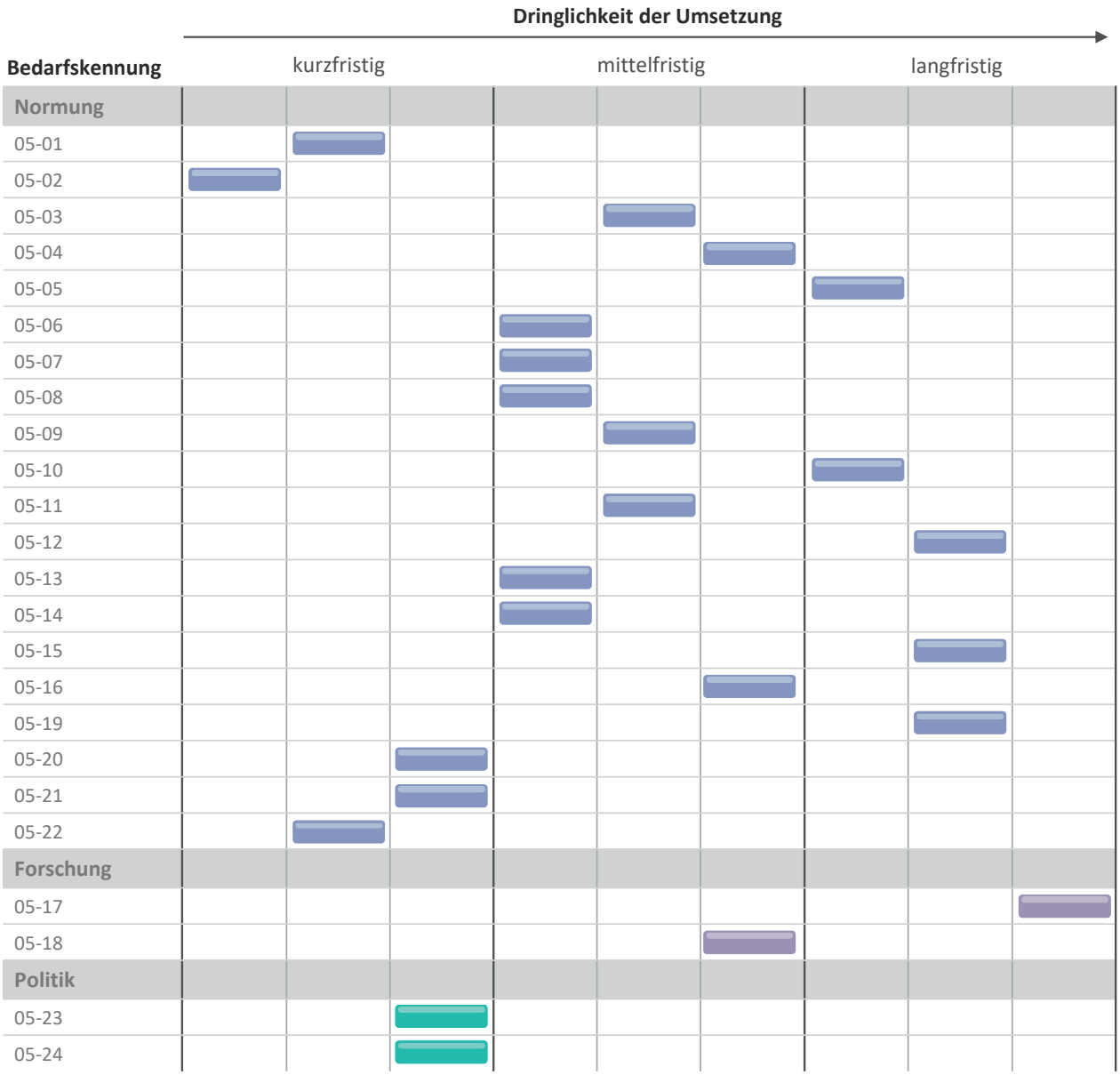
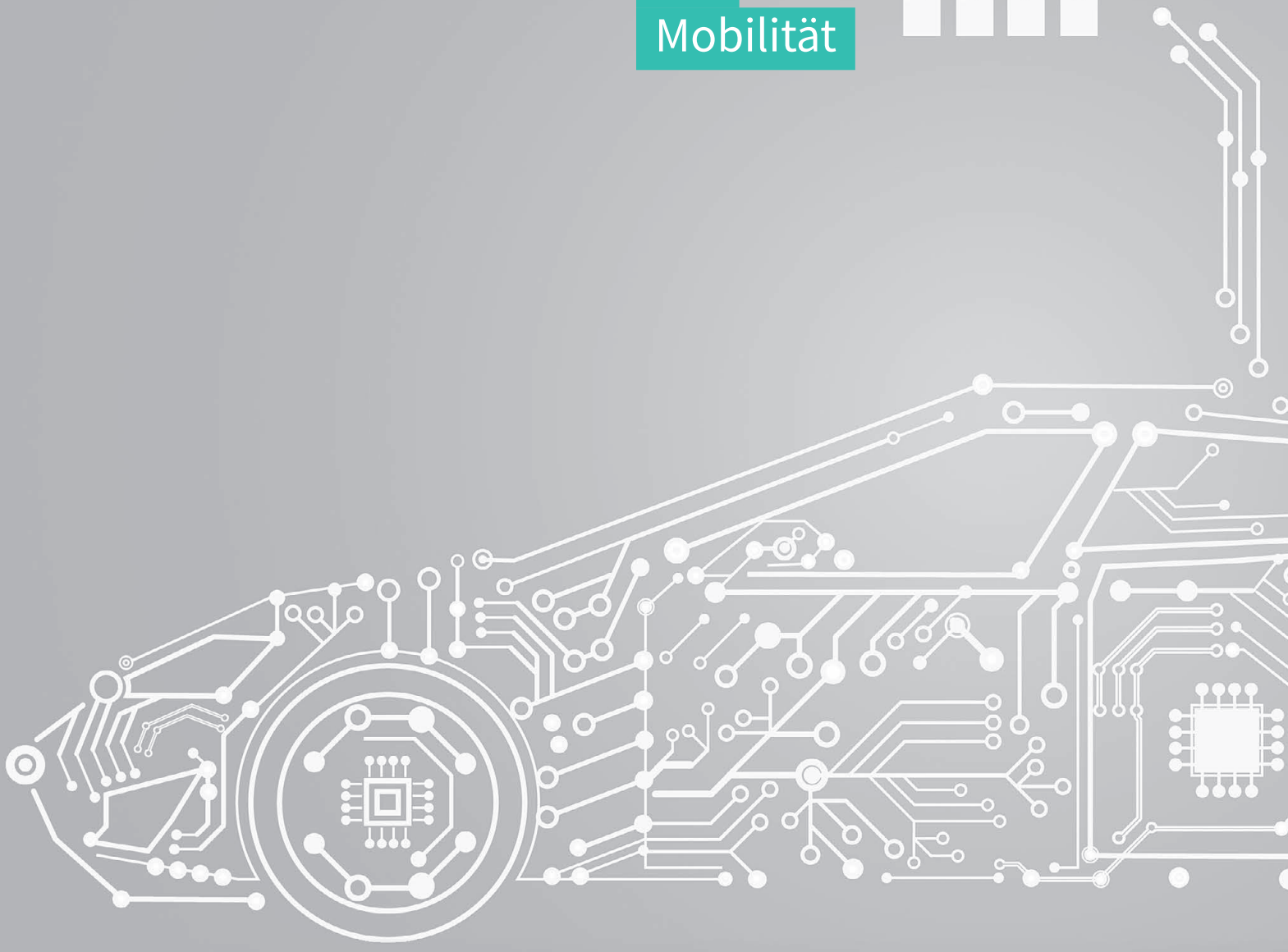


Abbildung 40: Priorisierung der Bedarfe aus Schwerpunkt Industrielle Automation
(Quelle: Arbeitsgruppe Industrielle Automation)

4.6 Mobilität



Der Sektor Mobilität spielt sowohl bezüglich seiner wirtschaftlichen als auch seiner gesamtgesellschaftlichen Bedeutung eine herausragende Rolle. Mobilität ist ein wesentlicher Faktor bei vielen wichtigen Lebensentscheidungen, ermöglicht die Teilnahme am gesellschaftlichen Leben, und der Transport von Personen und Gütern ist Grundvoraussetzung für eine funktionierende Wirtschaft. Hinzu kommt, dass der Kraftfahrzeugbau nach wie vor der umsatzstärkste Industriezweig und wichtiger Arbeitgeber in Deutschland ist.

Der Einsatz der Schlüsseltechnologie KI bietet einerseits wichtige Chancen für den Mobilitätssektor, u. a. durch die Ermöglichung komplexer automatisierter Fahrfunktionen und die Optimierung von Verkehrsströmen bzw. komplexen Mobilitätsketten, und stellt andererseits eine enorme Herausforderung dar, u. a. weil ein sicherer und vertrauenswürdiger Einsatz von KI weitreichender Anstrengungen in Forschung, Entwicklung, Standardisierung und Regulierung bedarf. Die Transformation des Mobilitätssektors durch den Einsatz von KI ist relativ weit vorangeschritten; u. a. haben bereits viele automatisierte Fahrfunktionen Einzug in Serienfahrzeugen gefunden und es werden beträchtliche Summen in entsprechende Forschung und Entwicklung (F&E) investiert.

In Anbetracht des hohen Stellenwerts von Mobilität und KI in der Mobilität liegt der Schwerpunkt im folgenden Kapitel darauf, umfassend den aktuellen Stand, die Anforderungen und Herausforderungen sowie die Normungs- und Standardisierungsbedarfe in diesem Sektor aufzuzeigen. Im Gegensatz zur ersten Ausgabe der Normungsroadmap, in welcher der rechtliche Rahmen und der Gütertransport (Logistik) Schwerpunkte bildeten, wird in der vorliegenden zweiten Ausgabe ein Fokus auf folgende Aspekte gelegt:

1. Einsatz von Trustworthy Artificial Intelligence (TAI) in der Anwendungsdomäne Mobilität und hier insbesondere im Kontext der „Cooperative, Connected und Automated Mobility“ (CCAM). CCAM umfasst dabei Vehikel verschiedener Modalitäten (Straße, Schiene, Wasser und Luft) mit automatisierten Funktionen und deren Vernetzung mit intelligenten Infrastrukturen, wie z. B. auch bei der intermodalen Mobilität.
2. Relevanz einzelner Aspekte der Trustworthiness (vgl. näher hierzu Abschnitt „Einbettung und Lebenszyklen von KI-Systemen“ weiter unten) im Rahmen der Systemeinbettung einerseits und der unterschiedlichen Lebensphasen des CCAM-Systems andererseits (vgl. näher hierzu Abschnitt „Einbettung und Lebenszyklen von KI-Systemen“ weiter unten) sowie der diesbezügliche Stand der Operationalisierung bzw. Operationalisierbarkeit. Als Aspekte von TAI werden u. a. Safety, IT-Sicherheit, Robustheit, Performanz, Erklärbarkeit, Nachvollziehbarkeit und Mensch-Maschine-Interaktion betrachtet. Hierbei kommt der funktionalen Sicherheit („Safety“) in allen Mobilitätsanwendungen eine besondere Bedeutung unter den TAI-Aspekten zu: Sie ist „nicht verhandelbar“. Die Anforderungen an die übrigen TAI-Aspekte sind anwendungsabhängig: Je nach Anwendungskontext und gültigen Regularien müssen auch hier bestimmte Mindestqualitäten realisiert sein; die konkrete Ausprägung einzelner Eigenschaften kann sich jedoch zwischen verschiedenen Anwendungen oder gleichen Anwendungen verschiedener Herstellender durchaus unterscheiden – auch gewollt zur Produktdifferenzierung – solange die hohen Mindestqualitäten für jede Eigenschaft erfüllt sind. Für Safety gilt allerdings, dass diese immer „vollumfänglich“ realisiert sein muss, d. h. in der nach dem jeweiligen Stand der Technik besten Realisierungsstufe, die nachweisbar das Restrisiko für das Auftreten eines Schadens unter ein minimales, gesellschaftlich akzeptiertes Restrisiko senkt. Da die verschiedenen TAI-Aspekte nicht unabhängig voneinander sind, gibt es kontextbedingte Abhängigkeiten. So kann in bestimmten Anwendungen z. B. die IT-Sicherheit eine unabdingbare Voraussetzung für Safety sein [311], [312] und muss daher ebenso hohe Anforderungen erfüllen.
   Vor diesem Hintergrund wird in diesem Kapitel die Trustworthy AI aus folgenden zwei Blickwinkeln betrachtet:
   a) Zunächst werden die Trustworthiness für KI-Systeme allgemein, nämlich alle TAI-Aspekte gleichwertig betrachtet (der Aspekt der Safety wird dabei nur insoweit beleuchtet, wie er für die Erläuterung des Gesamtkontextes und die Einordnung im Verhältnis zu den weiteren TAI-Aspekten notwendig ist). Hierzu werden die drei sogenannten Domänen des hochautomatisierten Fahrens, der Mobilitätsdienste bzw. -ketten und der Infrastruktur in Bezug auf bestimmte Funktionalitäten als Use Cases betrachtet bzw. miteinander verglichen.
   b) Darüber hinaus wird – vor dem oben dargestellten Hintergrund der besonderen Rolle der Safety und der beim Betrieb der CCAM begründeten Risiken für Körper und Leben verschiedener Verkehrsteilnehmender – der Aspekt der Safety einer tiefgreifenderen Betrachtung unterzogen, und zwar insbesondere mit Blick auf die Nachweisbarkeit dieser Eigenschaft, wie sie z. B. für eine Typzulassung bzw. eine Zertifizierung von Vehikeln in den verschiedenen Anwendungsdomänen notwendig ist. Hierbei wird nach den Modalitäten Automotive, Luftfahrt, Schifffahrt und Bahn differenziert.

Für gesellschaftliche oder ethische Aspekte, für die Vorgaben nicht rein aus technischer Sicht gemacht werden können, werden exemplarisch die für die Kontrolle und Durchsetzung dieser Vorgaben nötigen technischen Voraussetzungen behandelt, aber nicht die gesellschaftlichen Implikationen thematisiert.

Trustworthy-AI-relevante Aspekte von und Sichtweisen auf KI-Systeme

Die Technologie der KI, die u. a. maschinelle Lernverfahren (ML) wie z. B. Deep Learning (DL) für Deep Neural Network (DNN) umfasst, ist inzwischen eine unverzichtbare Schlüsseltechnologie für viele Anwendungsgebiete, die Menschen bei Entscheidungsprozessen unterstützt oder gar Entscheidungsprozesse ohne menschliches Zutun durchführt. Trustworthy AI, d. h. eine KI, der Menschen, Organisationen und/oder Gesellschaften vertrauen, ist nicht nur allgemein wünschenswert, sondern eine Voraussetzung für den Einsatz von KI in sicherheitskritischen Anwendungen [313], [314], [315], [316], [317]. Ob ein KI-System diese Eigenschaft „Trustworthiness“ besitzt, ist abhängig vom konkreten KI-System, der konkreten Anwendung und weiteren Rahmenbedingungen wie z. B. den rechtlichen und technischen Voraussetzungen für die Entwicklung und den Einsatz dieses Systems [311], [312], [318], [319].

Es gibt eine Vielzahl unterschiedlicher Sichtweisen auf KI-Systeme und TAI-relevanter Aspekte, aus denen eine Vielzahl maßgeblicher Kriterien abgeleitet werden können [312], [320]. Diese Aspekte und Sichtweisen können einen (primär) technischen oder gesellschaftlichen Hintergrund haben. Während die technischen Sichtweisen zu rein technischen Kriterien führen, erfordern die gesellschaftlichen Sichtweisen zwar technische Grundlagen (insbesondere geeignete Metriken wie z. B. für die Ausgewogenheit von Datensätzen), jedoch können die konkreten Anforderungen hieran nicht ausschließlich aus der technischen Perspektive festgelegt werden, sondern bedürfen einer ethischen Bewertung bzw. gesellschaftspolitischen Einordnung. Im Hinblick auf derartige Kriterien mit (primär) gesellschaftlichen Kriterien – namentlich die Akzeptanz eines KI-Systems durch einzelne Nutzer*innen oder die Gesellschaft, sogenannte Fairness oder Bias sowie Datenschutz und Privatheit – fokussiert dieses Kapitel die technische Sichtweise, nämlich auf die technischen Grundlagen für die Überprüfung und Durchsetzung dieser Kriterien (z. B.: Wie können wir die Prüfung ethischer Kriterien unterstützen? Welche Metriken sind hierfür besonders geeignet?).

Zu den in diesem Kontext besonders relevanten Aspekten der TAI gehören:

- → Performanz: Performanz des KI-Systems in Bezug auf relevante Performanzmetriken (in nachfolgenden Kapiteln wird vorrangig die englischsprachige Bezeichnung „Performance“ genutzt).
- → IT-Sicherheit: passive und aktive Robustheit des KI-Systems gegen Angriffe und insbesondere gegen KI-spezifische Angriffe (adversariale Angriffe, „Poisoning“-Angriffe und „Privacy“-Angriffe) in Bezug zu den drei Sicherheitszielen Integrität, Vertraulichkeit und Verfügbarkeit (in nachfolgenden Kapiteln wird vorrangig die englischsprachige Bezeichnung „Security“ genutzt) [83], [320].
- → Funktionale Sicherheit: Ein System ist funktional sicher („safe“), wenn von seinem Betrieb keine unakzeptablen Risiken für die Umgebung (Individuen, Umwelt, Organisationen und Gütern) ausgehen (in nachfolgenden Kapiteln wird vorrangig die englischsprachige Bezeichnung „Safety“ genutzt).
- → Robustheit und Generalisierbarkeit: Passive und aktive Robustheit gegen natürliche Variationen von Eingängen (Situationen) einschließlich derer, die vermeidbar gewesen wären, wenn sie während des Trainings adäquat berücksichtigt worden wären. Hierzu gehören die Robustheit gegenüber stochastischen Einflüssen wie z. B. Rauschen und gegenüber Störsignalen wie z. B. Interferenzen (in nachfolgenden Kapiteln wird vorrangig die englischsprachige Bezeichnung „Robustness“ genutzt).
- → Erklärbarkeit: Eigenschaften eines KI-Systems, die den Menschen in die Lage versetzen, den Entscheidungsprozess des KI-Systems zu verstehen, entweder durch inhärent interpretierbare Modelle oder durch post-hoc-Interpretation (in nachfolgenden Kapiteln wird vorrangig die englischsprachige Bezeichnung „Explainability“ genutzt).
- → Interpretierbarkeit: Eigenschaften eines KI-Systems, die es möglich machen, dass dessen Leistung im Gesamtsystem überwacht werden kann. Dazu ist es notwendig, dass Informationen zur Plausibilisierung der Ergebnisse bereitgestellt werden, die nicht notwendigerweise eine „Erklärung“ im Sinne der Erklärbarkeit darstellen (in nachfolgenden Kapiteln wird vorrangig die englischsprachige Bezeichnung „Interpretability“ genutzt).
- → Transparenz, Rechenschaft und Dokumentation: Nachvollziehbarkeit des KI-Systems über den gesamten Lebenszyklus hinweg, z. B. von Designentscheidungen, Randbedingungen, Daten, Modellen, Trainingsalgorithmen, Trainingsprozessen, Evaluationen und Betrieb u. a. durch technische Dokumentation und Logging (in nach-

folgenden Kapiteln wird vorrangig zusammenfassend die englischsprachige Bezeichnung „Tracability“ genutzt).

- → Risikomanagement: Identifikation, Analyse und Priorisierung von Risiken und koordinierter Einsatz von Ressourcen zur Minimierung von Risikowahrscheinlichkeiten oder Risikoauswirkungen (akzeptables Grenzrisiko).
- → Mensch-Maschine-Interaktion / „Human Oversight“: Implementierung von „Human-in-the-loop/ on-the-loop“-Lösungen – diese können als Maßnahmen zur Steigerung der Sicherheit oder zur Steigerung der Nutzerbeteiligung gesehen werden.
- → Akzeptanz durch einzelne Nutzer*innen und die Gesellschaft.
- → Bias, Unparteilichkeit: Maßnahmen, um den unausgewogenen Betrieb von KI-Systemen zu verhindern, z. B. durch Trainingsdatensätze, die nicht den IID-Kriterien („independent and identically distributed“, deutsch „unabhängig und identisch verteilt“) entsprechen und zu Diskriminierung führen, z. B. in Bezug auf das Geschlecht (in nachfolgenden Kapiteln werden vorrangig die englischsprachigen Bezeichnungen „Fairness“ bzw. „Impartiality“ genutzt) [317].
- → Datenschutz und Privatheit: Angemessene Behandlung sensitiver (privater und vertraulicher) Daten.
- → Redundanz: Welche Anforderungen an die redundante Erfassung und Auswertung von Daten ergeben sich, um dem Gesamtsystem vertrauen zu können (auch abhängig von der Kritikalität der Funktionen), insbesondere, wenn notwendigerweise Blackbox-KI-Ansätze (fehlende Interpretierbarkeit bzw. Erklärbarkeit) genutzt werden müssen, weil herkömmliche Algorithmen die Funktionen nicht abbilden können?

Bei genauerer Betrachtung wird deutlich, dass die oben genannten Sichtweisen nicht scharf voneinander abzugrenzen sind und es zahlreiche Abhängigkeiten gibt. So gibt es beispielsweise Überschneidungen zwischen der IT-Sicherheit (Security) und funktionalen Sicherheit (Safety), da sowohl ein erfolgreicher Security-Angriff die Funktionalität des Systems ändert und damit die funktionale Sicherheit des Systems gefährdet, als auch das Nicht-Erfülltsein einer Teileigenschaft der funktionalen Sicherheit Angriffsflächen für Security-Angriffe öffnen kann. Fehlende Robustheit durch eingebaute semantische Plausibilisierung macht z. B. eine Reihe von Angriffsmustern wahrscheinlicher oder überhaupt erst möglich. Dazu gehören u. a. „adversarial attacks“ durch Manipulation auf Pixelebene, die in TAI abgefangen werden können (vgl. [321]). Die beiden Eigenschaften sind somit gegenseitige Voraussetzungen für die Gesamtbetrachtung des Systems.

Weitere Abhängigkeiten zwischen TAI-Aspekten bestehen offensichtlich ebenso; für diesen Beitrag zentral ist die Feststellung, dass eine Vielzahl an Aspekten hohe Relevanz für TAI-Systeme haben und dementsprechend berücksichtigt werden müssen. Die Relevanz und die nötige Priorisierung der jeweiligen Aspekte ist für jede Anwendung bzw. Anwendungsklasse separat zu bewerten.

Einbettung und Lebenszyklen von KI-Systemen

Systeme, die als KI-Systeme bezeichnet werden (hier verstanden u. a. im Sinne des Entwurfs der ISO/IEC 22989:2022 [16], bestehen oftmals aus mehreren interagierenden Soft- und Hardwaremodulen und sind eingebettet in ein Gesamtsystem aus KI- und Nicht-KI-Komponenten sowie in Bezug auf einen Kontext [318]. Das Softwaresystem besteht z. B. aus einer variablen Anzahl von klassischen IT-Modulen, symbolischen KI-Modulen (z. B. logisches Schlussfolgern oder Entscheidungsbäume) und konnektionistischen KI-Modulen (z. B. neuronale Netze), die über geeignete Schnittstellen miteinander kommunizieren. Die Software läuft auf Recheneinheiten, die jeweils lokal (Edge) oder über ein Netzwerk (Cloud) angebunden sein können. Die Software interagiert über Hardwaremodule mit der Umwelt. Ein automatisiertes Fahrzeug besitzt z. B. eine Vielzahl an Sensoren und Aktuatoren, die über mechanische, elektrische und IT-Systeme im Fahrzeug-„Körper “ verbunden sind. Sensoren lassen sich in propriozeptive (interne Sensoren wie z. B. Radumdrehungssensoren), exterozeptive (externe Sensoren wie z. B. Kamerasensoren) und virtuelle Sensoren (wie z. B. Eingänge aus Kommunikationskanälen oder aus der Fusion verschiedener Sensoren) unterteilen. Aktuatoren reichen vom Antriebsstrang über das Bremssystem und das Lenksystem bis hin zum Beleuchtungssystem und zu nutzerrelevanten Informationssystemen (Display, Lautsprecher). Die Umwelt eines KI-Systems können hier u. a. verschiedene passive und aktive Verkehrsteilnehmende, Insassen des einbettenden Fahrzeugs oder Smart-City-Infrastrukturen sein. Hinter der Entwicklung eines KI-Systems steckt i. d. R. eine Organisation und eine oder mehrere Organisationen haben zudem Verantwortlichkeiten während des Betriebs des KI-Systems, bedingt durch die Bereitstellung oder Verarbeitung von Datenströmen. Daher müssen diese Organisationen auch in die Gesamtbetrachtung miteinbezogen werden. Insgesamt ergeben sich aus der anwendungsspezifischen Einbettung eines KI-Systems anwendungsspezifische Anforderungen und Risiken, die bei Entwicklung, Prüfung und Betrieb eines solchen Systems beachtet werden sollten.

Bei klassischen IT- und symbolischen KI-Systemen können Struktur und Parameter, zumindest im Prinzip, jeweils direkt

vom Entwickelnden festgelegt bzw. eingestellt und ihre Funktionsweise im operationalen Betrieb nachvollzogen werden. Bei IT- und symbolischen KI-Systemen ab einer gewissen kritischen Größe erschwert bzw. verhindert die große Anzahl an Parametern jedoch ggf. ein direktes Design und Tuning der Parameter und eine Interpretation der Funktionsweise. Bei konnektionistischen KI-Systemen wie z. B. neuronalen Netzen und Support Vector Machines trifft dieses Problem aufgrund ihrer für Menschen nicht (von vornherein) intuitiven Verarbeitungsweise wesentlich verstärkt und somit auf einen Großteil der im Einsatz befindlichen Systeme zu. Solche Systeme müssen in einem datenbasierten, iterativen Trainingszyklus mithilfe maschineller Lernverfahren entwickelt werden, in dem die Entwickelnden jeweils die Rahmenbedingungen, aber nicht mehr direkt die Parameter des operativen Systems festlegen. Hierdurch ergibt sich ein komplexer Lebenszyklus, der sich – wie sich in der Praxis herausgebildet hat – in die folgenden Phasen unterteilen lässt:

- → Planungsphase: Hier werden abhängig von den gewünschten Eigenschaften des zu entwickelnden KI-Systems u. a. geeignete KI-Modelle, Lernmethoden, benötigte Daten, Metriken und Qualitätssicherungsmaßnahmen inklusive etwaige Abhängigkeiten identifiziert und ein Entwicklungsplan festgelegt.
- → Datengewinnungs- und QS-Phase: Für das Training benötigte Daten werden in hinreichender Qualität und Quantität gewonnen. Hierbei können Daten zunächst selbst gewonnen (Datenaufnahme in der physischen Welt), aus externen Quellen bezogen oder synthetisch generiert werden. Neben einer Kombination dieser Datenquellen können Daten in vielfältiger Weise angereichert werden, z. B. um die Anzahl an Daten zu erhöhen oder gewünschte Eigenschaften in den Datensatz einfließen zu lassen. Abhängig von den konkreten Anforderungen an den Datensatz schließen sich verschiedene Qualitätssicherungsmaßnahmen an.
- → Trainingsphase: Der Entwickelnde startet iterativ einen oder mehrere Trainingsprozesse mit vorab festgelegten Modellen, Daten und Hyperparametern. Abhängig von festgelegten Abbruchkriterien, die mithilfe von geeigneten Metriken (z. B. Performanzkriterien) geprüft werden, werden die Trainingsprozesse gestoppt und mit angepassten Parametern neu gestartet, bis zumindest ein trainiertes System die vorab geforderten Anforderungen (bezüglich der festgelegten Metriken und Qualitätssicherungsmaßnahmen) erfüllt.
- → Evaluationsphase: Die Evaluation des Systems geht über die automatisierte Berechnung von Metriken im Trainingsprozess hinaus und wird vor, während und nach der Inbetriebnahme des Systems durchgeführt. Für die Evaluation können z. B. komplexe Simulationen oder Pentests[92] eingesetzt werden.
- → Deployment- und Skalierungsphase: Hier werden die KI-Systeme für den praktischen Einsatz und die Inbetriebnahme angepasst, was ggf. weitere Optimierungen einschließt, z. B. hinsichtlich einer verbesserten Skalierung oder Effizienz.
- → Operationale Phase, einschließlich Wartung. Prinzipiell wäre es denkbar, auch in der operationellen Phase weitere Trainingsphasen durchzuführen (sogenannte selbstlernende oder online-lernende Systeme). Die dadurch möglichen Änderungen des Systemverhaltens entziehen sich jedoch gegenüber dem aktuellen Stand der Technik vollständig einer Safety-Analyse und den zugehörigen Safety-Nachweisen, sodass eine Zertifizierung bzw. Typzulassung solcher Systeme aktuell nicht möglich ist. Diese Art von KI-Systemen wird daher in diesem Kapitel nicht betrachtet.
- → Außerbetriebnahme: Falls KI-Modell und/oder Trainingsdaten auch nach dem regulären Betrieb vor Privacy-Angriffen auf Modell und/oder Daten geschützt werden sollen (z. B. aus Datenschutz- oder IP-Gründen), ist eine geordnete Außerbetriebnahme, die einen öffentlichen Zugriff auf Modell und Daten dauerhaft unterbindet, erforderlich. Ansonsten hat diese Lebenszyklusphase keine KI-spezifische Relevanz.

Aufgrund sich ändernder Anforderungen, aufgrund von im Betrieb bekannt werden den Schwachstellen des Systems oder aufgrund des Ziels, ein System kontinuierlich zu verbessern, werden die oben genannten Phasen zyklisch (kontinuierlich im Sinne eines kontinuierlichen Entwicklungsprozesses) durchlaufen. Hierbei gibt es einen kontinuierlichen Übergang von seltenen, sorgfältig geplanten und durchgeführten Updates mit ggf. wesentlichen Änderungen zur Vorversion über sehr kurze Updatezyklen hin zu selbstlernenden oder online-lernenden Systemen. Während die diskreten Updates bei vielen Systemen inzwischen unabdingbar sind und regelmäßig durchgeführt werden, finden selbstlernende Systeme (d. h. Systeme, die sich im Feld aufgrund eintreffender Beobachtungen anpassen) trotz großer medialer Aufmerksamkeit in sicherheitskritischen Anwendungen wie der Mobilität bisher keine Verwendung (siehe auch oben bezüglich Zertifizierbarkeit).

92 Penetrationstests, d. h. kontrollierte Cyberangriffe mit dem Ziel der Identifikation von Schwachstellen.

4.6.1 Status quo

4.6.1.1 Grundlegende, qualitativ neuartige Eigenschaften von KI-Technologie

Einerseits eröffnet der Einsatz von KI-Technologie neue Chancen und ermöglicht Anwendungen, die mit klassischen Technologien nicht oder nur sehr eingeschränkt realisierbar sind. Andererseits führt die Komplexität der KI-Systeme und ihrer Lebenszyklen zu qualitativ neuen Problemen und Risiken [320], [83]. Wie oben beschrieben erfordert die Entwicklung von KI-Systemen i. d. R. einen datengetriebenen Ansatz und der Entwickler hat keine direkte Kontrolle über die erlernten Parameter des KI-Systems und die dadurch implizierten Ein-/Ausgabekorrelationen. Dies führt dazu, dass operative KI-Systeme Blackbox-Eigenschaften besitzen und sich ihre Funktionsweisen (und somit auch mögliche Fehler) den Entwickelnden und Nutzer*innen nicht direkt erschließen. Die Eigenschaften der mittels maschinellen Lernverfahren und Daten implizit eincodierten Funktionen hängen wesentlich von dem zugrunde liegenden Trainingsdatensatz ab. Eine hinreichende Qualitätssicherung von Trainingsdaten ist jedoch eine nicht-triviale Aufgabe, insbesondere, wenn die Daten aus externen Quellen stammen. Falls, wie häufig praktiziert, vorab trainierte Modelle verwendet werden, können schwer entdeckbare Sicherheitslücken im KI-System enthalten sein, die oftmals weitere Nachtrainingseinheiten unbeschadet überstehen. Viele KI-Systeme besitzen zudem einen riesigen Eingangs- und Parameterraum. Hier seien exemplarisch der Kameraeingang einer 4K-Kamera mit einer hohen Anzahl an Farbkanälen genannt. Als Folge dieser Komplexität sind formale Verifikationsmethoden für viele praktisch eingesetzte KI-Systeme nicht verfügbar und alternative empirische Validierungsmethoden können aus praktischen Gründen nur einen Bruchteil des Parameterraums abdecken. Somit erfüllt ein KI-System nicht unbedingt die Absicht des Programmierers und es gibt weder eine Garantie, was vom System gelernt wurde, noch eine Sicherheit hinsichtlich der eingangs aufgeführten Vertrauenswürdigkeitsaspekte (vgl. Kapitel 4.6) – so etwa, welche Performanz das System in der Praxis erzielt. Umgekehrt gibt es oftmals keine oder nur eine eingeschränkte Erklärung der Funktionsweise eines KI-Systems für Menschen. Bezüglich der verschiedenen Vertrauenswürdigkeitsaspekte von KI-Systemen ist das technische Verständnis aktuell noch unvollständig, u. a. hinsichtlich Funktionalität, Integrität, Verlässlichkeit, Safety und Generalisierbarkeit, und weitere umfassende Anstrengungen in F&E werden benötigt.

4.6.1.2 Anforderungen, Prüfung und Absicherung von KI-Systemen

Aus dem steigenden Einsatz von KI-Technologien einerseits, insbesondere auch in sicherheitskritischen Systemen, und der hohen Komplexität von KI-Systemen andererseits, die zu qualitativ neuen Risiken führt, ergeben sich steigende Bedarfe an Regulierung, Standardisierung, Absicherung und objektive Überprüfbarkeit von KI-Systemen im Hinblick auf ihre Vertrauenswürdigkeit. Erste Ansätze in diese Richtung, insbesondere auf abstrakter Ebene, existieren bereits, wie z. B. die geplante horizontale (d. h. sektorübergreifende) Regulierung von KI-Systemen in der Europäischen Union (EU) [4]. Trotz großer internationaler Anstrengungen in F&E mangelt es aktuell an hinreichend praxistauglichen und technisch fundierten Anforderungen, Prüf- und Mitigationsstrategien und an entsprechenden Werkzeugen [312]. Da formale Verifikationsmethoden aufgrund der Systemkomplexität praktisch oftmals nicht einsetzbar sind, muss auf empirische Validierungs- und Testverfahren zurückgegriffen werden. Eine hinreichende Aussagekraft erfordert hier aber eine sehr gute Abdeckung des Eingangsraums des Systems. Um eine hinreichende Testabdeckung, insbesondere inklusive relevanter Corner Cases, zu erreichen, kann es neben technischen Entwicklungen ggf. erforderlich sein, relevante Randbedingungen einzuschränken. Dies kann z. B. die Beschränkung automatisierter Fahrfunktionen auf bestimmte Verkehrssituationen und Wetterbedingungen bedeuten.

Jedenfalls sind zum gegenwärtigen Zeitpunkt die Anforderungen bezüglich der relevanten Aspekte nicht umfassend – über den gesamten Lebenszyklus eines KI-Systems hinweg – konkretisiert. Entsprechende Metriken, die – analog zum Begriff der „Key Performance Indicators“ (KPI) – als „Key Trustworthiness Indicators“ (KTI) fungieren können, sind bislang noch nicht hinreichend etabliert. Aufgrund der Komplexität des Themas bietet sich hier die zeitweise Fokussierung auf spezifische Anwendungsklassen und Anwendungen an. Dieser zum horizontalen Ansatz des europäischen Artificial Intelligence Act (AI Act) komplementäre vertikale Ansatz verfolgt das mittelfristige Ziel einer Operationalisierung des AI Act zunächst für einzelne Anwendungen und das langfristige Ziel, die sektorspezifischen Erkenntnisse zu generalisieren und das horizontale Modell iterativ zu verbessern.

Neben einer verbesserten Prüfbarkeit von KI-Systemen bezüglich ihrer TAI-Eigenschaften ist ein weiteres zentrales Ziel, KI-Systeme von Grund auf so zu entwickeln, dass sie wesentliche TAI-Eigenschaften besitzen („Trustworthy by design“).

4.6.1.3 Stand der Technik, aktuelle Anwendungsfälle

Automotive

Im Automobilbereich ist der Einsatz von KI aktuell zumeist limitiert auf nicht oder begrenzt sicherheitskritische Funktionen oder Prototypen ohne Serienzulassung. Neben Fahrassistenzsystemen, welche menschliche Fahrer*innen in bestimmten Fahrsituationen unterstützen, sind Systeme mit höherem Automationsgrad nur in sehr definierten Bereichen im Einsatz [311]. Zu solchen Systemen mit definiertem Betriebsbereich gehören automatisierte Valet-Parking-Systeme (AVP) sowie automatisierte Spurhaltesysteme (Automated Lane Keeping System, ALKS). Ein ALKS übernimmt im Wesentlichen die Längs- und Querführung eines Fahrzeugs. Dabei darf dieses System bis zu einer Geschwindigkeit von 60km/h nur in speziellen Bereichen (Operational Design Domain, ODD) eingesetzt werden, in denen eine bauliche Trennung der Fahrrichtungen vorherrscht und die für besonders schützenswerte Verkehrsteilnehmer*innen wie Fußgänger*innen und Radfahrer*innen unter normalen Umständen gesperrt sind. Dies trifft in Deutschland im Wesentlichen auf Teile der Bundesstraßen (insbesondere Kraftfahrtstraßen) und Autobahnen zu. In diesem Bereich übernimmt das System sowohl die Anpassung der Geschwindigkeit als auch der Lenkung, um innerhalb der „Freiflächen“ auf den durch das Fahrzeug genutzten Fahrstreifen der Straßenführung zu folgen. Die Sensorik des Systems übernimmt dabei die Erkennung der „Freiflächen“ in der Umgebung. Eine Überwachung der Funktion findet eingeschränkt durch das System selbst (z. B. Erkennen von Eingriffen durch oder Abwesenheit eines/einer Fahrzeugführenden sowie technische Schwierigkeiten beim Halten der Spur, der Geschwindigkeit oder bei der Bestimmung der „Freiflächen“) statt. Solange das System aktiviert ist, trägt es die Verantwortung, jedoch muss dauerhaft durch das System sichergestellt werden, dass Fahrzeugführende innerhalb eines bestimmten Zeitrahmens (z. B. zehn Sekunden) die Fahrfunktion übernehmen können. Falls eine angefragte Übergabe nicht erfolgt, muss das System ein sogenanntes Minimum Risk Manoeuvre (zum Erreichen eines Zustands, in dem das Risiko minimal ist) ausführen. Das weltweit erste zugelassene System dieser Art ist der sogenannte „Drive Pilot“ von Mercedes, welcher dem Automationsgrad SAE (Society of Automotive Engineers) Level 3 entspricht. Eine Erweiterung der Regularien, welche u. a. das Fahren bis zu einer Geschwindigkeit von 130 km/h sowie Fahrstreifenwechsel erlaubt, ist bereits vorbereitet und tritt Anfang 2023 in Kraft.

Darüber hinaus gelten nationale Regelungen wie beispielsweise das Gesetz zur Änderung des Straßenverkehrsgesetzes und des Pflichtversicherungsgesetzes – Gesetz zum Autonomen Fahren vom 12. Juli [322] und die dazu gehörende Autonome-Fahrzeuge-Genehmigungs-und-Betriebs-Verordnung [323].

Die wichtigsten, der Typzulassung zugrunde liegenden Standards und Normen sind dabei die ISO-Standards zur funktionalen Sicherheit (ISO-26262-Reihe [455]), zur Cybersecurity (ISO/SAE 21434:2021 [324]) und zur funktionalen Sicherheit der intendierten Funktion (SOTIF, Safety of the intended Function, ISO 21448:2022 [90]). Neuere, zum Teil noch in Arbeit befindliche Standards erweitern diese um Konzepte des szenarienbasierten Testens, des Tests im laufenden Betrieb sowie um die Betrachtung hochautomatisierter Fahrfunktionen, wie sie u. a. durch den Einsatz von KI-Verfahren ermöglicht werden (ISO/TR 4804:2020 [325]) und Nachfolger ISO/TS 5083 [326], ISO 22737:2021 [327], ISO PAS 8800 [110]).

Luftfahrt

Im Bereich der Luftfahrt kann zunächst zwischen verschiedenen Anwendungsdomänen unterschieden werden. Dabei grenzt sich das Feld der Urban Air Mobility (UAM) deutlich von dem Bereich der herkömmlichen Luftfahrt ab. Durch die UAM und speziell die Entwicklung von Drohnen und Flugtaxis sind viele neue Akteur*innen involviert, welche sich durch ihre kurzen Entwicklungszyklen und starke Technologieaffinität auszeichnen. Die potenziellen Anwendungsbereiche von KI sind dabei in beiden Anwendungsdomänen vielfältig, wobei aufgrund der allgemein hohen Sicherheits- und Zulassungsanforderungen sowie der allgemein in der Luftfahrt geforderten Redundanz von Systemen noch keine klaren Standards und Normen für KI-basierte Funktionen existieren. Zwar existieren bereits eine KI-Roadmap der EASA [328], siehe Abbildung 41, sowie erste konkrete Arbeiten zu spezifischen „Concepts of operations“ und Betrachtungen zur Vertrauenswürdigkeit, Erklärbarkeit und Zuverlässigkeit [329], [330] der KI, jedoch ergeben sich bei der Überführung der gesetzten Ziele in Zulassungsverfahren und Normen dieselben, wenn nicht größere Hürden, verglichen mit den anderen in diesem Kapitel adressierten Domänen. Grundsätzlich orientieren sich die Automatisierungsschritte dabei an der Übergabe der Verantwortung an KI-basierte Funktionen. So wird zunächst die Einführung von Pilotassistenzsystemen adressiert. Anschließend sollen Single-Pilot-Operations umgesetzt werden, bei der die KI den Piloten lediglich unterstützt. Wiederum

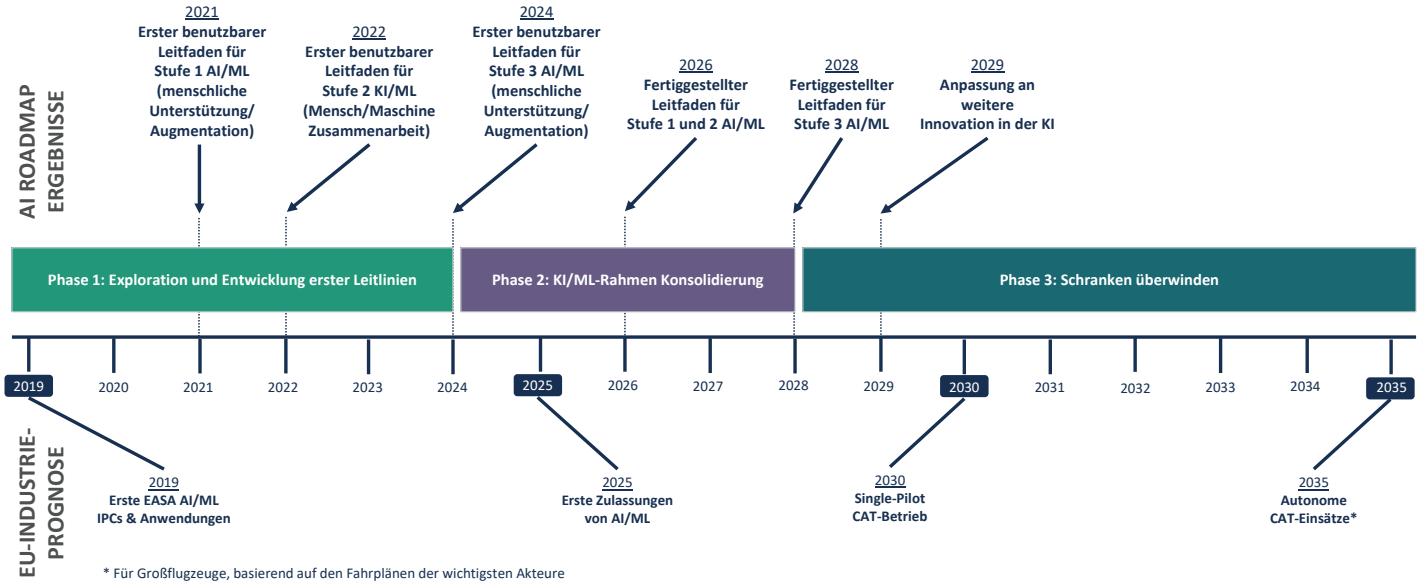


Abbildung 41: Auszug aus der EASA Artificial Intelligence Roadmap (Quelle: in Anlehnung an [328])

anschließend soll die Verantwortung schrittweise an die KI übertragen werden, sodass zunächst teilautomatisierte Funktionen und schließlich vollautomatisierte Funktionen/Aerial Vehicles umgesetzt werden.

Erste KI-Anwendungen im Bereich von Pilotassistenzsystemen (beispielsweise Deadalean und Iris Automation) und kleinen Drohnen sind aufgrund des überschaubaren Risikos (geringes Gewicht, Pilot zur Überwachung) bereits etabliert; hierbei folgen Drohnen einzelnen Personen, kartieren automatisiert Strukturen oder sind in der Lage, ihre Umgebung selbstständig dreidimensional zu erfassen (beispielsweise [331]). Es gibt jedoch auch in diesem Maßstab noch keine sinnvollen Wege, eine Zertifizierung der Verfahren zu ermöglichen, was die Notwendigkeit konkreter Handlungsempfehlungen für die Schaffung solcher Standards und Normen unterstreicht.

Denn problematisch im Bereich der Luftfahrt ist die Tatsache, dass viele der Aerial Vehicles (AV) durch ein Fehlverhalten von relevanten KI-Funktionen einen erheblichen Schaden am AV selbst, anderen AV (Air Risk) oder der Umgebung (Ground Risk) verursachen können. Insbesondere wenn Personen passiv oder aktiv involviert sind, steigt die Kritikalität zusätzlich an. So wird das Risiko, welches mit der Automatisierung der Luftfahrt einhergeht, nicht zuletzt aufgrund der kinetischen und potenziellen Energie eines Unmanned Aerial Vehicle (UAV) sowie der Beteiligung von Personen für die verschiedenen Klassen von AV und ihre jeweiligen Einsatzbereiche bewertet werden müssen. Ein weiterer entscheidender Faktor ist dabei selbstverständlich auch die Vertrauenswürdigkeit und Zuverlässigkeit der KI sowie die Kritikalität deren Einsatzgebietes für die Funktionalität des AV.

Aufgrund der hohen Risiken, die mit dem Transport von Gütern oder Personen über die dritte Dimension einhergehen, sind die Arbeiten zur Integration von KI allgemein deutlich weniger fortgeschritten als beispielsweise in der Domäne des hochautomatisierten Fahrens. Dennoch lassen sich zukünftige Anwendungsfälle insbesondere im Hinblick auf sicherheitskritische Funktionen skizzieren.

Schifffahrt

In der Schifffahrt kommen zunehmend Systeme mit KI-Komponenten bzw. eigenständige KI-Systeme zum Einsatz, um beispielsweise das nautische Personal bei der Entscheidungsfindung zu unterstützen. Hierbei handelt es sich oft um optionale Ausstattungsvarianten oder Sonderfunktionen, die neben den ausrüstungspflichtigen Anlagen nach SOLAS (International Convention for the Safety of Life at Sea) bzw. MED (Marine Equipment Directive) als zusätzliches Feature angeboten werden und keiner konkreten Zulassung unterliegen. Hierzu zählen beispielsweise 360-Grad-Perzeptionssysteme, die die Umgebung erfassen und auf einem separaten Bildschirm mit annotierten AR-Elementen anzeigen. Neben klassischen Systemen zur Kollisionsverhütung (z. B. Radar,

AIS) kommen auch datenbasierte Kollisionswarnsysteme zum Einsatz. Derartige Assistenzsysteme dienen lediglich der Information und lösen keine eigenständigen Aktionen aus. Ggf. werden automatisierte Vorschläge gemacht. Somit ist der Mensch weiterhin die Kontrollinstanz und die Schiffsführung trägt die Verantwortung für getroffene Entscheidungen. Die Schiffsbrücke muss daher dauerhaft mit qualifiziertem Personal besetzt sein. Systeme, die darüber hinaus agieren, sind aktuell Prototypen oder Teil von Forschungsvorhaben und dienen der Erprobung unter kontrollierten Randbedingungen, wobei auch hier der Mensch als Kontrollinstanz agiert, um in Gefahrensituationen einzugreifen.

Unterschieden werden vier verschiedene Automationsgrade. Hierfür hat die International Maritime Organization das Maritime Autonomous Surface Ship mit den dazugehörigen Abstufungen der Automatisierung definiert:

- → **Degree One**
  Ship with automated processes and decision support: Seafarers are on board to operate and control shipboard systems and functions. Some operations may be automated and at times be unsupervised but with seafarers on board ready to take control.
- → **Degree Two**
  Remotely controlled ship with seafarers on board: The ship is controlled and operated from another location. Seafarers are available on board to take control and to operate the shipboard systems and functions.
- → **Degree Three**
  Remotely controlled ship without seafarers on board: The ship is controlled and operated from another location. There are no seafarers on board.
- → **Degree four**
  Fully autonomous ship: The operating system of the ship is able to make decisions and determine actions by itself.

Vielen Neubauten lassen sich bereits Degree One zuordnen. Steuerungssysteme wie Autopiloten übernehmen unter menschlicher Aufsicht die Kontrolle über den Antrieb und folgen einer festgelegten Route. Solche Systeme arbeiten regelbasiert und kommen ohne KI aus.

Im Bereich der Schifffahrt gibt es derzeit keine Normen oder Standards mit Fokus auf KI-Komponenten.

Eisenbahn

Die Eisenbahn unterscheidet sich vom automatischen städtischen schienengebundenen Personennahverkehr (AUGT) nach DIN EN 62267:2010 [332], wie z. B. der Nürnberger Bahn [332], durch das Fehlen sicherer Barrieren. Der Eisenbahnverkehr findet im Freien statt und impliziert dadurch das Vorhandensein systemfremder Hindernisse. DIN EN 62267:2010 [332] listet Automatisierungsgrade von GoA0 bis GoA4 (vgl. Tabelle 9) auf. Für den Eisenbahnbereich gibt es noch keine Norm, die Automatisierungsgrade einteilt. Es wird jedoch allgemein die Einteilung ähnlich DIN EN 62267:2010 [332] benutzt. Systeminterne Hindernisse wie andere Eisenbahnfahrzeuge werden den Triebfahrzeugführenden (Tf) ab GoA1 rechtzeitig durch Signale angezeigt. GoA0 wird auch als Fahren auf Sicht bezeichnet. Eine zusätzliche visuelle Erkennung anderer Eisenbahnfahrzeuge ist daher bei GoA1 nicht nötig. Der Automatisierungsgrad GoA2 bedeutet wiederum, dass der Tf für die Erkennung systemfremder Hindernisse, das Öffnen und Schließen von Türen und Notfälle zuständig ist. GoA3 bedeutet, dass Zugpersonal nur für Notfälle zuständig ist und sich in dem Zug frei bewegen kann. GoA4 beschreibt Züge ohne Zugpersonal. Bei GoA4 kann eine von Menschen besetzte Überwachungs- und Steuerungszentrale eingesetzt werden. Rangierbahnhöfe stellen bei GoA4 eine Ausnahme dar – dort müssen auch andere Eisenbahnfahrzeuge visuell erkannt werden. GoA2 ist derzeit Stand der Technik. Grade ab GoA3 sind noch experimentell, wie z. B. im Falle von „AutoHaul“ [334].

Wie man in Tabelle 9 sieht, ist die Hindernisserkennung die zentrale Herausforderung für GoA3. Experimentelle Systeme für Hinderniserkennung im Eisenbahnbereich werden seit den 1990ern erprobt [335], wobei die Implementation oft konventionelle Bildverarbeitungstechnologien verwendet. Die Verwendung der KI-Systeme für Hinderniserkennung wird oft als eine leistungsstärkere Lösung gewertet. Zur Diagnose und Prognose von Restlebensdauer, Fehlerereignissen oder anderen Zustandseigenschaften von Elementen der Schieneninfrastruktur werden zunehmend Modelle und Verfahren der KI angewandt. Die damit zur Verfügung stehenden neuen Funktionalitäten bilden die Grundlage für die Entwicklung digitaler und datenbasierter Instandhaltungsstrategien wie zustandsbasierter und vorausschauender Instandhaltung.

Tabelle 9: Vereinfachte Übersicht der Automatisierungsgrade für die Eisenbahn

	GoA0	GoA1	GoA2	GoA3	GoA4
	Fahren auf Sicht	Fahren nach Signalen	Halbautomatischer Fahrbetrieb	Fahrerloser Betrieb	Unbegleiteter Fahrbetrieb
Weichenstellung und Fahrberechtigungen	X	S	S	S	S
Abstand zwischen Zügen	X	S	S	S	S
Bremsen	X	X und S	S	S	S
Beschleunigen	X	X	S	S	S
Hinderniserkennung	X	X	X	S	S
Überwachung des Fahrgastwechsels	X	X	X	X oder S	S
Zustandsüberwachung	X	X	X	X	S
Notfallsituationen	X	X	X	X	S und/oder Überwachungs- und Steuerungszentrale

X – Betriebsbedienstete, S – technisches System

4.6.2 Anforderungen und Herausforderungen

Wie im vorangegangenen Kapitel beschrieben, nimmt die Safety eine Sonderstellung unter den Trustworthiness-Aspekten ein. Im folgenden Kapitel 4.6.2.1 erfolgt daher zunächst eine Betrachtung der Trustworthiness-Eigenschaft als Ganzes, während in Kapitel 4.6.2.2 eine Fokussierung auf den Aspekt der Safety erfolgt und die speziell für diesen Trustworthiness-Aspekt geltenden weiteren Herausforderungen dargestellt werden.

4.6.2.1 Trustworthy AI based mobility

Der vertrauenswürdige Einsatz von KI-Technologie in der Mobilität ist komplex und bestimmt von verschiedenen Komplexitätsdimensionen (vgl. näher hierzu Kapitel 4.6.1). Der dadurch aufgespannte mehrdimensionale Raum kann hier aufgrund der begrenzten Ressourcen nur exemplarisch in Form von ausgewählten Anwendungsbeispielen erfolgen, gleichwohl mit dem Ziel, anwendungsspezifisch die Anforderungen und Herausforderungen als Begründung der in Kapitel 4.6.3 formulierten Bedarfe zu erarbeiten. Unter der Zielsetzung, weniger Anwendungsbeispiele (Use Cases) so auszuwählen, dass sie einerseits zusammen den Komplexitätsraum vertrauenswürdiger KI in der Mobilität möglichst gut abdecken und andererseits der in der AG Mobilität vorhandenen Expertise für eine hinreichend tiefe Betrachtung entsprechen, wurden folgende Anwendungsbeispiele in den drei verschiedenen Domänen zur weiteren Betrachtung ausgewählt:

1. **Ausweichmanöver** als komplexes Fahrmanöver beim automatisierten Fahren
2. **Ride Sharing** als Teil von Mobilitätsdiensten
3. Verkehrsoptimierung über eine Verbesserung der **Lichtsignalanlagensteuerung** (LSA-Steuerung) in der **Verkehrsinfrastruktur**

Für alle drei Anwendungsbeispiele wurde zunächst die Relevanz und sodann der Stand der Operationalisierung der geplanten Regulierung der KI im jeweiligen Anwendungsbei-

spiel systematisch erfasst, um diese sodann vergleichen und generalisieren zu können. Für die systematische Erfassung wurde eine abgeänderte Version der Certification Readiness Matrix aus [312] genutzt, im weiteren Trustworthiness Readiness Matrix (TRM) genannt. Die TRM bildet die folgenden zwei Dimensionen ab:

→ Einbettungsphasen (Organisation, anwendungsspezifische Anforderungen und Risiken, Verkörperung und Situiertheit des KI-Modus) und Lebenszyklusphasen (Planungsphase, Datengewinnungs- und Qualitätssicherungsphase, Trainingsphase, Evaluationsphase, Deployment- und Skalierungsphase, operationale und Wartungsphase)
→ Vertrauenswürdigkeitsaspekte (Safety, Security, Performance, Robustness, Interpretability/Explainability, Tracibility/Docu/Logs, Fairness/Impartiality, Data Privacy)

Detailliertere Ergebnisse in Form der TRMs sind im Anhang 13.4 „Trustworthiness Readiness: ausgewählte Ergebnisse“ zu finden. Im Folgenden werden die wichtigsten Ergebnisse in kompakter Form zusammengefasst.

Durch die vergleichende Gegenüberstellung von TRMs für verschiedene Anwendungsdomänen/-bereiche (beispielsweise Automotive vs. Medizin) und KI-Technologien (beispielsweise Entscheidungsbäume vs. DNNs) könnten weitere Dimensionen berücksichtigt und über die Zeit nachverfolgt werden.

Use-Case-Ausweichmanöver als komplexe Fahrmanöver beim automatisierten Fahren

Die Automatisierung von Fahrfunktionen ist wohl die bekannteste Anwendung von KI in der Mobilität und – mit Blick auf die unvorhersehbaren Umweltbedingungen (neben dem Verkehr selbst auch Wetterbedingungen, Straßenbeschaffenheit, Sensorverschmutzung, unkenntliche Straßenschilder usw.) – eine der komplexeren Anwendungen. Die hohe Komplexität dieser Anwendung, verbunden mit hohen Risiken für Körper und Leben während des Betriebs und der hohen und unmittelbaren Anwendungsrelevanz auf Basis der bereits weitreichend fortgeschrittenen Automatisierung vieler Fahrfunktionen legt eine Betrachtung dieser Anwendung nahe. Da automatisiertes Fahren seinerseits sehr vielfältige Szenarien und Funktionen umfasst, wurde eine weitere Eingrenzung des zu betrachtenden Anwendungsfalls auf die Funktionalität des Ausweichmanövers als komplexes und repräsentatives Fahrmanöver innerhalb des automatisierten Fahrens (SAE Level 2 und 3, Advanced Driver Assistance Systems) vorgenommen.

Namentlich besteht der vorliegende Anwendungsfall in der Aufgabe des KI-Systems, auf Basis der Ausgabe der Sensoren ein Ausweichmanöver zu planen und dieses durch entsprechende Befehle an das Fahrzeug zur Steuerung der Bewegungsrichtung und Geschwindigkeit durchzuführen. Die Planung beinhaltet hierbei eine Entscheidung sowohl über das Ob als auch das Wie der geplanten Aktionen, was Komplexitäten wie Sicherheitsabstand, Verzögerung im Informationsfluss (und zwar seitens der von den Sensoren eintreffenden Signale einerseits wie der Kommunikation mit der Steuerung des Fahrzeugs andererseits), Geschwindigkeitsanpassung und Bremswegberechnung einschließt. Hinsichtlich der Entscheidungsfindung ist ferner nicht nur der Aspekt der Sicherheit (d. h. Vermeidung einer Kollision) zu betrachten, sondern auch die Effizienz (d. h. die Dauer des Manövers) und der Fahrkomfort (d. h. Vermeidung plötzlicher Geschwindigkeits- und Bewegungsänderungen). Des Weiteren wurde beschlossen, einen Fokus auf die Perzeptionsebene (Objekterkennung) inklusive Sensorfunktion zu legen.

1. Eine systematische Betrachtung des Anwendungsfalls mithilfe der TRM in Bezug auf die Relevanz der Vertrauenswürdigkeitsaspekte in den jeweiligen Lebenszyklus- und Einbettungsphasen zeigt die durchgängig hohe Relevanz aller Aspekte: Eine herausgehobene Stellung nimmt hierbei der Aspekt Safety ein, der insbesondere in der operationalen Phase von höchster Bedeutung ist: Mensch und Umwelt dürfen nicht zu Schaden kommen. Vor dem gleichen Hintergrund und mit Blick darauf, dass ein Angriff auf das System durch Externe zwingend auszuschließen ist, wurde der Aspekt Security als hochrelevant bewertet. Eine ähnlich hohe Relevanz – auch als Voraussetzung bzw. Unterstützung von Safety – hat der Aspekt Performance, der ebenfalls in der operationalen Phase eine sehr hohe Relevanz hat, sowie Robustness, und zwar tendenziell in den späteren Lebenszyklusphasen. Hierbei wurde das Verständnis zugrunde gelegt, dass die Performance einen Durchschnittswert für alle auftretenden Fälle/Situationen bedeutet, während als Robustness die Verlässlichkeit des Systems auch unter vorhersehbaren „extremen“ („Edge Cases“) und unvorhersehbaren („Corner Cases“) Bedingungen verstanden wird. Auf die Berücksichtigung derartiger Corner Cases komme es gerade in Ausweichsituationen an, da diese naturgemäß ungewöhnliche Situationen im Straßenverkehr bedeuten. Corner Cases treten im vorliegenden Use Case nicht zuletzt häufig zutage aufgrund der Ungenauigkeiten/Diskrepanz zwischen simulierter Umgebung, in der das KI-System trainiert wird, und der realen Umge-

bung, in der es eingesetzt wird. Insofern sei ein Standard dafür zu entwickeln, welche Schwellenwerte bei welchen konkreten Faktoren/Funktionen erreicht werden müssen, um von Robustness sprechen zu können bzw. diese zu gewährleisten. Data Privacy wurde punktuell in der operationalen Phase als höchst relevant bewertet, insbesondere in Bezug auf die Frage des Umfangs der Speicherung von Sensordaten zur Nachvollziehbarkeit, welche regelmäßig Bilddaten von anderen Fahrzeugen und Personen umfassen. Ferner wurde eine hohe Relevanz gesehen für die Interpretability/Explainability – und zwar in der Evaluations-, Deployment- und operationalen Phase – sowie für die Tracability – und zwar neben der operationalen Phase bei der Rubrik Einbettung für die anwendungsspezifischen Anforderungen und Risiken. Hintergrund dafür ist insbesondere die Klärung der Schuldfrage bei einem Unfall (sowohl bei einem Unfall trotz Ausweichmanöver als auch bei einem Unfall aufgrund eines Ausweichmanövers). Fairness/Impartiality, welche gerade auch in der öffentlichen Diskussion bei Ausweichmanövern in Bezug auf (extreme) Konfliktsituationen eine Rolle spielt, wird als besonders relevant für die Datengewinnungs- und Evaluationsphase erachtet.

2. Während demnach die Relevanz des gesamten Trustworthiness-Komplexitätsraums für diese Anwendung insgesamt hoch bis sehr hoch ist, ist der Stand der Operationalisierung der Regulierung „durchwachsener". Inzwischen gibt es viele Initiativen zur Standardisierung (u. a. [336]; ENISA Ad-Hoc Working Group on AI Cybersecurity; NIST Trustworthy and Responsible AI; Projekt KI Absicherung; Grand Défi – Sécuriser, certifier et fiabiliser les systèmes fondés sur l'intelligence artificielle; UNECE GRVA technical workshop on Artificial Intelligence). Aspekte, die durch existierende Methoden bedient werden können, z. B. die Tracability, sind voraussichtlich mit vergleichsweise wenig Aufwand operationalisierbar. Insbesondere für die technischen Aspekte Safety, IT-Security und Robustness ist jedoch eine Operationalisierung des geplanten EU AI Act aktuell noch nicht möglich [312]. Schließlich wird die Einführung derartiger hoch- und vollautomatisierten Fahrfunktionen nur dann als gerechtfertigt erachtet, wenn eine Verbesserung der Sicherheit und Umweltverträglichkeit des Straßenverkehrs auch nachgewiesen wird. Um solche Fahrfunktionen schnell in den Verkehr zu bringen, sollte zusätzlich zur Regulierung von (KI-)Technologien eine Reform der Typenzulassung hinsichtlich deren Dynamisierung angestrebt werden.

Use Case Ridesharing als Mobilitätsdienst (Mobilitätskette)

Die flexible und zeitlich begrenzte Zuteilung von Fahrzeugen zu Kunden durch kommerzielle Anbieter im Rahmen von „Ridesharing" fällt unter die Klasse der „Mobilitätsdienste" wie „Mobilitätsketten". Der Einsatz von KI erfolgt hier einerseits im Rahmen automatisierter Fahrfunktionen, wo – anders als im regulären Individualverkehr – unterschiedlichste Nutzer*innen mit verschiedensten Fahrzeugtypen einschließlich unterschiedlicher Automatisierungsgrade unter kürzeren Eingewöhnungszeiträumen interagieren. Andererseits kommt KI zur Optimierung des Flottenmanagements inklusive Fahrzeugvorhaltung und -zuteilung sowie predictive Maintenance zum Einsatz.

Grundlage für diese Dienste sind Applikationen sowie Fahrzeuge mit Automatisierungsfunktionen, deren korrekte Bedienung ein angemessenes Vertrauensverhältnis seitens der Nutzenden erfordert. Dementsprechend sollen hier in Bezug auf das Fahrzeug nur die zusätzlichen Herausforderungen im Vergleich zum automatisierten Fahren betrachtet werden (d. h., die für das automatisierte/autonome Fahren angenommenen Anforderungen – etwa an Safety und Security – werden als Grundlage vorausgesetzt). Hierbei wird als Anwendungsfall der komplexeste Fall betrachtet, bei dem ein Mobilitätsanbieter Fahrzeuge verschiedener Herstellender mit verschiedenen Automatisierungsfunktionen, Bedienkonzepten usw. im Angebot hat.

1. Eine systematische Betrachtung des Komplexitätsraums mithilfe der TRM hinsichtlich der Relevanz in Bezug auf den Anwendungsfall zeigt, dass insbesondere folgende zusätzliche Herausforderungen im Vergleich zum automatisierten Fahren entstehen:
   a) KI-Funktionen und Implikationen der Fahrzeuge müssen für die Nutzenden (Individuen und Organisationen) in kurzer Zeit hinreichend erklärt werden, sodass eine erhöhte Relevanz in Bezug auf den Aspekt der Explainability zu legen ist.
   b) Die Funktionen sind für sehr heterogene Nutzende/Umweltprofile auszulegen, wobei Funktionen abhängig von der Nutzervorerfahrung oder von Nutzerbedürfnissen ggf. auch deaktiviert oder deaktivierbar sein sollten. Dies gilt insbesondere, soweit die Aspekte Safety und Security dies erfordern.

c) Beim „Ride Hailing“ sollte die Zuteilung von Fahrzeugen an Nutzende nicht nur nach ökonomischen Aspekten, sondern auch nach Fairnessgrundsätzen erfolgen. Dies ist insofern von besonderer Relevanz, als private Mobilitätsdienstleister grundsätzlich nach Wirtschaftlichkeits- bzw. Profitgesichtspunkten agieren werden. An dieser Stelle wird eine etwaige Incentivierung seitens der öffentlichen Hand eine wichtige, politisch zu beantwortende Frage sein. Diese Aspekte haben überragende Relevanz für die Performance solcher Mobilitätsdienste/-ketten.
d) In diesem Zusammenhang wird es auch eine Rolle spielen, in welchem Umfang / an welchen Orten die öffentliche Hand Schnittstellen zwischen den unterschiedlichen Verkehrsmitteln bereithält – so etwa für Pkw anzusteuernde sogenannte Drop-off Zones mit Zugang zum öffentlichen Verkehrsnetz (Schiene oder Wasserwege). Ähnliches gilt für die Frage, inwieweit die sogenannte Multi Layer Traffic Optimization vorangetrieben/gefördert werden soll und (private) Mobilitätsanbieter hierbei integriert werden sollen.
e) Die Wartung der Fahrzeugflotte u. a. mit Predictive-Maintenance-Methoden sollte auch bei unterschiedlichen Herstellenden und Modellen auf einheitlichen Standards basieren.
f) Da bei den vorgenannten Funktionalitäten die Daten von verschiedenen Fahrzeugen mit häufig wechselnden Nutzer*innen zu verarbeiten sind, ist der Aspekt des Datenschutzes – generell und insbesondere in der operationalen Phase – von herausragender Bedeutung. Die Notwendigkeit des Ausmaßes der Datenerhebung wird auf der einen Seite durch den Automatisierungsgrad der Fahrzeuge sowie der Flottensteuerung seitens der Mobilitätsbetreiber bedingt. Ihre Akzeptabilität liegt auf der anderen Seite aber beim Profil des Mobilitätsnutzers sowie beim Einsatzzweck, daher ist davon auszugehen, dass der Aspekt des Datenschutzes im Ridesharing nicht rein auf Fahrzeugbasis regelbar ist (vgl. [337]).
g) Als wesentlich wurde herausgearbeitet, dass die Verantwortlichkeiten zwischen Fahrzeugherstellenden und Mobilitätsdienstleistern in Bezug auf die im öffentlichen Verkehrsraum vom Gesetz- bzw. Normengeber vorgegebenen Anforderungen – so insbesondere aber nicht nur die Anforderungen an Safety und Security – verbindlich geklärt werden müssen und sodann auch transparent gegenüber dem Kunden zu kommunizieren sind. Dies gilt etwa in Bezug auf Softwareupdates, welche die Fahrzeugapplikationen einerseits und die Service-Applikation andererseits betreffen können und insofern insbesondere die Schnittstellen zwischen den Applikationen (so etwa Zugriff der Dienstleister-Applikation auf das Navigationssystem des Fahrzeugs / Original Equipment Manufacturer (OEM). Damit kommt dem Aspekt der Dokumentation eine erhöhte Relevanz zu.

2. In Bezug auf die Operationalisierung ist festzustellen, dass im Vergleich zu dem Stand / den Ausführungen beim autonomen Fahren (s. obiger Abschnitt „Use Case Ridesharing als Mobilitätsdienst (Mobilitätskette)“) die Operationalisierung der für Mobilitätsdienste/-ketten relevanten Aspekte bislang nur in wenigen Bereichen vorangetrieben wurde. Während bei dem Aspekt des Datenschutzes auf die Ansätze zum Flottenmanagement aufgesetzt werden kann, sind belastbare Ansätze weder in Bezug auf die genannten Safety-/Security-Aspekte noch die Performance und Erklärbarkeit/Fairness zu erkennen; zumal die KI-basierten Entwicklungen bislang insbesondere OEM-seitig vorangetrieben wurden und insofern das diesbezügliche Know-how bzw. die Einbettung innerhalb der Organisation bei selbigen zu beobachten ist, indes bei den Mobilitätsanbietern gegenwärtig kaum vorhanden ist.

Use Case Verkehrsoptimierung / Verbesserung der Lichtsignalanlagensteuerung (LSA) in der Verkehrsinfrastruktur

Die Optimierung des Verkehrsflusses unter Einbeziehung verschiedener Verkehrsteilnehmender hat eine hohe Bedeutung in der Mobilität. Die Steuerung von Lichtsignalanlagen an sich ist offenkundig im höchsten Maße sicherheitsrelevant. Aus diesem Grunde soll der Einsatz von KI-Technologie aktuell nur parallel zu bzw. eingebettet in klassischen, formal verifizierten Verfahren eingesetzt werden, die die Einhaltung aller sicherheitsrelevanten Aspekte garantieren. Perspektivisch sind KI-basierte Steuerungsfunktionen z. B. im Rahmen von Smart Citys jedoch nicht auszuschließen, so etwa denkbar bei der Einbeziehung von Umfelderkennung, Berechnung der optimaler Phasenfolgen/-übergänge/-dauern und insbesondere Übergangszeiten zwischen Rot-Grün-Phasen und insbesondere bei dem Zusammenspiel mit bzw. Rückgriff auf V2X-Daten (sogenannte Car-to-Infrastruktur-Daten). Deren Einfluss auf die Verkehrssicherheit soll daher in die Überlegungen miteinbezogen werden.

1. Eine systematische Betrachtung des Komplexitätsraums mithilfe der TRM hinsichtlich der **Relevanz** in Bezug auf den Anwendungsfall zeigt, dass insbesondere folgende Herausforderungen beim Einsatz von KI bei der LSA-Steuerung bestehen:
   a) Da in der Datengewinnungsphase (auch) personenbezogene Daten Verwendung finden können (wohingegen in der Deployment- bzw. Skalierungsphase und der operationalen Phase die Verwendung personenbeziehbarer oder anonymisierter Daten ausreichend ist bzw. kritische Daten on-chip ohne Zugriff auf Rohdaten bzw. personenbezogene Daten verarbeitet werden können), wurde hier eine hohe Relevanz für Data Privacy angenommen. Vor diesem Hintergrund wurde ferner der Aspekt Security – insbesondere in der Phase der Datengewinnung – als hoch relevant erachtet.
   b) Als hoch relevant wird der Aspekt der Fairness und Traceability der Entscheidungsprozesse hinsichtlich unterschiedlicher Verkehrsteilnehmer(-gruppen), Fortbewegungsmodalitäten (Pkw, motorisiertes Zweirad, Fahrrad oder Fußgänger etc.) und Routen (etwa Hauptverkehrsstrom zu Nebenströmen) eingestuft, zumal dies zugleich unmittelbaren Einfluss auf die Performance des Infrastruktursystems hat. Auch die Fairness im Hinblick auf multimodale Aspekte (beispielsweise Gewichtung verschiedener Verkehrsteilnehmertypen bei der Definition der Zielfunktion/Optimierungsgröße) spielt dabei eine besondere Rolle. Als außerordentlich relevant wird der Aspekt Traceability ferner in der Deployment-Phase insofern erachtet, als dass das „Ausrollen“ eines erfolgreich getesteten Systems auf verschiedene Kommunen einer guten Dokumentation bedarf, zumal eine Bedienung des Systems durch Verkehrsingenieure ohne KI-Expertise möglich sein muss.
   c) Daneben werden Performance und Robustness für besonders relevante Aspekte gehalten, insbesondere bei hohem Verkehrsaufkommen, schlechtem Wetter etc. Hierin kommt der Sinn und Zweck einer Verkehrsoptierung zum Ausdruck, wonach sowohl im Normal-/Durchschnittsfall als auch in Extremfällen der Verkehrsfluss besser gesteuert werden soll als heute in allgemeinen und in Spitzenzeiten bzw. Extremsituationen.
   d) Für die Interpretability wurde eine hohe Relevanz in den frühen Phasen der Datengewinnung, des Trainings und der Evaluation gesehen. Da zentrale Entscheidungen bezüglich des Designs in den Entwicklungsphasen getroffen werden, müssen diese begründbar sein. Daneben müssen die Entscheidungen der KI jedoch auch im Deployment für die Anwendenden nachvollziehbar sein, wozu wiederum der Grundstein im Systemdesgin – also in frühen Entwicklungsphasen – gelegt wird.
   e) Die Relevanzen von Safety und Security sind – jedenfalls im Vergleich zu den zuvor betrachteten Use Cases – nicht besonders hoch, da hier redundante klassische Systeme greifen und Daten vom Nutzenden nicht einfach direkt manipuliert werden können. Nichtsdestotrotz ist die Betrachtung der Safety und Security auch hier unabdingbar.

Hinsichtlich des Stands der **Operationalisierung** ist zunächst festzuhalten, dass bereits ein rechtlicher Rahmen für Verkehrssteuerungssysteme und daraus abgeleitete Anforderungen bzw. Prüfkriterien existieren. Zum einen ist jedoch festzustellen, dass auch in den Bereichen, in denen der rechtliche Rahmen bereits recht umfassend gesetzt ist – so insbesondere für den Bereich des Datenschutzes – in Bezug auf einzelne Phasen eine Operationalisierung im Sinne eines operationalisierten Anforderungskatalogs noch nicht existiert (so im besagten Bereich des Datenschutzes für die Datengewinnungs- und Deployment-Phase). Zum anderen ist der Operationalisierungsbedarf hoch in Bezug auf zukünftige multidimensionale bzw. multimodale Systeme (so insbesondere sogenannte koordinierte Ampelsteuerungen unter Beachtung von Hierarchien wie bei beispielsweise für Einsatzfahrzeuge, verschiedene Areale zu unterschiedlichen Tages- oder etwa Rushhour- bis hin zu Ferienzeiten oder Sonderereignissen) sowie Multi-Agenten-Funktionalitäten. Dies gilt umso mehr, als derartige Systeme/Funktionalitäten die oben genannten als besonders relevant eingestufte Performance und Robustness fördern. In diesem Zusammenhang ist ferner anzumerken, dass gegenwärtig für Ad-hoc-Situationserfassung – und somit insbesondere in der Phase des Deployment und der Operationalisierung – der Stand der Sensorik ungenügend ist.

Daraus resultierende Herausforderungen

Betrachtet man die Aspekte mit hoher Relevanz einerseits und/oder geringem Operationalisierungsstand andererseits, ergeben sich hohe bis sehr hohe Bedarfe fast über die ganze TRM hinweg und damit für den gesamten Komplexitätsraum. Die Bedarfe reichen von der Erfassung des aktuellen Standes über die Erarbeitung fehlender Grundlagen, der Formulierung zeitgemäßer Anforderungen, der Verfügbarkeit von Handlungsempfehlungen bis zur Bereitstellung einer geeigneten Infrastruktur von Szenarien, Daten und Simulationen.

4.6.2.2 Sichere hochautomatisierte Mobilität

Zukünftige Anwendungsfälle

Modalität Automobil

Das ALKS stellt im Automobilbereich den aktuellen Stand der Technik dar (vgl. Kapitel 4.6.1), in steigender Komplexität reihen sich der Autobahn-Chauffeur, der automatisierte Hub-to-Hub-Transport und das automatisierte Fahren im urbanen Raum mit verschiedenen Zwischenstadien ein.

Der **Autobahn-Chauffeur** stellt im Wesentlichen eine Weiterentwicklung zum ALKS dar. Zunächst zeichnet sich dies in einer Erweiterung der ODD aus, so führt der Autobahn-Chauffeur erweiterte Fahrmanöver wie das Überholen anderer Fahrzeuge oder das Wechseln der Autobahnen in Autobahnkreuzen aus und fährt im Allgemeinen auch bei höheren Geschwindigkeiten. Jedoch wird der Betrieb des Autobahn-Chauffeurs weiterhin nur in baulich abgetrennten, standardisierten Bereichen ermöglicht, die für besonders schützenswerte Verkehrsteilnehmer*innen nicht zugelassen sind. Mit der Erweiterung einher geht auch die Notwendigkeit einer komplexeren sensorischen Umfeldwahrnehmung. Dazu zählen insbesondere die Erkennung von Objekten und die Prädiktion des zukünftigen Verhaltens dieser Objekte. Jedoch scheint hier eine grobe Klassifizierung der Objekte (z. B. in statisch, dynamisch, Motorrad, Pkw, Lkw) sowie eine Schätzung des Bewegungsvektors ausreichend. Die Fahrfunktion des Systems ist in diesem Fall weiterhin redundant ausgelegt, so wird die Überwachung insbesondere der ODD-Einhaltung durch das Fahrzeug übernommen und die Kontrolle beim Verlassen der ODD aktiv an den Menschen übertragen, welcher sich nahezu zu jeder Zeit bereithalten muss. Auf der einen Seite dient der Mensch als Rückfallebene für die Leistungsgrenzen des Systems, kann aber auf der anderen Seite selbst jederzeit aktiv die Kontrolle übernehmen. Der potenzielle Einsatz von KI in diesem Anwendungsfall umfasst im System die Sensordatenaufbereitung, Sensordatenfusion, Objekterkennung und Bewegungsprädiktion für die Umfeldwahrnehmung, die taktische Planung und Trajektorienplanung inklusive einer Bewertung und Überwachung von Trajektorien sowie die Aufmerksamkeitsüberwachung der/des Fahrenden. Des Weiteren könnte KI auch im Bereich der Entwicklung und des Testens der Systeme eingesetzt werden, um intelligent Testfälle zu explorieren sowie Corner Cases zu identifizieren. Der Autobahn-Chauffeur ist wie das ALKS ein System des Automationsgrades SAE Level 3.

Der **automatisierte Hub-to-Hub-Transport** stellt im Gegensatz zum ALKS und Autobahn-Chauffeur keine Komfortfunktion für Fahrzeugführende dar. Im Gegenteil findet der Transport zumeist völlig automatisiert und ohne die Anwesenheit von Menschen im Fahrzeug statt. Die ODD ist im Vergleich zum Autobahn-Chauffeur um Baustellen und Betriebshöfe bzw. deren Ein- und Ausfahrten mit eingewiesenem Personal erweitert und somit weiterhin baulich abgetrennt. Zwar übernimmt das System innerhalb der ODD die gleichen Funktionen wie beim Autobahn-Chauffeur, jedoch entfällt das Vorhandensein eines Menschen als Rückfallebene. Aus diesem Grund müssen die Funktionen mit höherer Qualität sowohl in Bezug auf die Genauigkeit als auch die Zuverlässigkeit der Umfelderkennung ausgeführt werden und alle notwendigen Fahrmanöver (Spurhalten, Spurwechsel, Überholen, Anhalten) sowie deren Planung – von einer globalen Routenplanung und -optimierung bis zur Planung der konkreten Trajektorie für die nächsten Sekunden/Minuten – vom System ausgeführt werden; und dies möglicherweise sogar kooperativ mit anderen Verkehrsteilnehmer*innen. Die Kontrolle wird nur an den Menschen übergeben, wenn die ODD verlassen wird (z. B. beim Verlassen der Autobahn oder auf dem Betriebshof). Zusätzlich kann KI zur Prädiktion von typischem Verhalten der Verkehrsteilnehmenden und auch zur Überwachung der Einhaltung der ODD eingesetzt werden.

Das **urbane automatisierte Fahren** stellt ein vollständig automatisiertes Fahrzeug dar. Die ODD ist erweitert auf innerörtliche Bereiche sowie Städte. Zwar sind innerorts die Geschwindigkeiten niedriger, jedoch gibt es keine durchgehende bauliche Trennung zwischen Verkehrsrichtungen und besonders schützenswerten Verkehrsteilnehmer*innen, welche innerhalb des gemeinsam genutzten Verkehrsraums erlaubt sind. Die Folge ist eine dramatische Erhöhung der Diversität der wahrzunehmenden Objekte sowie durch den offenen Kontext eine zeitliche Variabilität des Kontexts (z. B. durch die Einführung neuer Verkehrsmittel). Das System benötigt damit eine fortgeschrittene Umfeldwahrnehmung bezüglich der (bekannten und unbekannten) Objekte und ihrer Intentionen (sowohl typisches als auch atypisches Verhalten). Die Planung der Fahrmanöver muss kontinuierlich unter Berücksichtigung des zukünftigen Verhaltens der anderen Verkehrsteilnehmer*innen angepasst werden. Innerhalb der ODD übernimmt das System alle Fahrmanöver, insbesondere auch bei komplexer Verkehrsführung mit kreuzenden Fahrbahnen und dedizierten Fahrwegen für andere Verkehrsteilnehmer*innen. Des Weiteren wird die Überwachung vollständig durch das System übernommen und die Kontrolle nur in Notfällen (z. B. zur Bergung) einer Remote-Steuerung

übergeben. In einem solchen System übernimmt KI keine neuen Aufgaben, allerdings sind die Anforderungen an die KI stark gestiegen. Als Besonderheit führt die zeitliche Variabilität des Kontexts zur Notwendigkeit einer kontinuierlichen Systemanpassung. Um die Sicherheit des Systems sicherzustellen, ist es empfehlenswert, den Betrieb kontinuierlich zu überwachen und Möglichkeiten zu schaffen, unbekannte Szenarien (insbesondere der Umfeldwahrnehmung) zwischen Herstellenden und Fahrzeugen auszutauschen.

Aus Sicht der Anwendung von KI-Verfahren in solchen zunehmend komplexeren und höher automatisierten Systemen wird die Minimierung möglicher gefährlicher Begegnungen z. B. durch eine innere und eine äußere Kontrollschleife erreicht. Die innere Schleife beinhaltet alle Prozesse, die automatisiert innerhalb des hochautomatisierten Fahrzeugs ablaufen können. Die äußere Schleife beinhaltet alle Prozesse, die mit der Umgebung des selbstfahrenden Fahrzeugs interagieren müssen. In nicht automatisierten Fahrzeugen bzw. zu Zeitpunkten, an denen das Fahrzeug nicht durch die Automation kontrolliert wird, ist das die Aufgabe des Fahrzeugführenden. In hochautomatisierten Fahrzeugen werden dazu intelligente KI-Komponenten und -Systeme eingesetzt, die Gegenstände, Vorgänge, Personen, andere Fahrzeuge, Muster wie hell und dunkel, Ungenauigkeiten etc. unterscheiden und zuverlässig deuten können.

Modalität Luftfahrt

Umfelderfassung: Um eine sichere hochautomatisierte Navigation und Flugführung zu ermöglichen, ist eine zuverlässige Umfelderfassung eine elementare Voraussetzung. Hierbei ist in der Regel nicht nur die dreidimensionale Geometrie der Umgebung, sondern aufgrund der hohen Sicherheitsanforderungen auch ein semantisches Verständnis des Umfeldes erforderlich. Dies gilt insbesondere bei Start- und Landevorgängen sowie im bodennahen Flug. In diesen Fällen sind je nach AV die Anforderungen an das semantische Verständnis des Umfeldes mit denen des hochautomatisierten Fahrens im Straßenverkehr vergleichbar. Jedoch ergibt sich durch die Navigation in drei Dimensionen ein größerer Lösungsraum. Speziell bei der Entwicklung von Drohnen und Hubschraubern spielt hier nicht nur das nach vorne und unten gerichtete Sichtfeld eine Rolle, sondern die Erfassung eines sphärischen 360°-Umfeldes. Dies stellt besondere Herausforderungen an die Sensorik, KI-basierte Auswertung und damit einhergehende Rechenressourcen in Bezug auf die abzudeckenden Sichtfelder. Im Kontrast dazu ergeben sich bei Reiseflugeschwindigkeiten komplementäre Anforderungen, die ebenfalls durch das Zusammenspiel von Sensorik und KI abgedeckt werden müssen. Hierbei steht das Gebiet des „Detect and Avoid“ (D&A, Erkennen und Ausweichen) im Vordergrund. Für D&A ergeben sich je nach Reiseflugeschwindigkeit deutliche Unterschiede in der notwendigen Erkennungsreichweite und bei den Verarbeitungslatenzen, wobei die Anforderungen an die Zuverlässigkeit der Erkennung und Klassifikation von Hindernissen nach wie vor sehr hoch sind.

Trajektorienplanung: Der Einsatz von KI zur Trajektorienplanung wird in verschiedenen Bereichen erforscht. Hierbei sind deterministische KI-Verfahren grundsätzlich in der Lage, Trajektorien zu finden, jedoch ergeben sich auch hier weitreichende Anforderungen an Sicherheit und Zuverlässigkeit. Zudem spielen Umgebungsbedingungen wie beispielsweise Wetter und (Auf-)Wind sowie die damit verbundenen Effekte der Flugphysik eine wichtige Rolle, die die Absicherung solcher Verfahren in der Praxis erschwert. Weiterhin arbeiten die Algorithmen speziell in der lokalen Trajektorienplanung mit Daten aus der Umfelderfassung, sodass Algorithmen mit den entsprechenden Unsicherheiten umgehen und die erfassbaren dreidimensionalen Sichtfelder berücksichtigen müssen. Beispielsweise können komplexe Anströmungsverhältnisse zu untypischen Flugsituationen oder auch hohem Side Slip (z. B. bei VTOL-G-Drohnen) führen, bei denen die Beobachtbarkeit der Umgebung nicht immer hinreichend gewährleistet werden kann.

Entscheidungsfindung/Flugplanung: Bereits vor dem Abflug müssen bestimmte Entscheidungen getroffen werden. So gilt es, beispielsweise die Flugroute festzulegen, Flugverbotszonen zu beachten sowie Wetter und andere Umwelteinflüsse einzubeziehen. Grundsätzlich sollte KI zukünftig in der Lage sein, solche Funktionen zu übernehmen oder mindestens zu unterstützen, um einen hochautomatisierten Transport von Personen und Gütern zu ermöglichen. Um in einem nächsten Schritt solche Funktionen sowie weitere missionsrelevante Entscheidungen während des Fluges unter Berücksichtigung externer Einflüsse und sonstiger Randbedingungen treffen zu können, bedarf es weiterer vertrauenswürdiger KI-Verfahren, die in der Lage sind, sinnvolle übergeordnete und nachvollziehbare Entscheidungen mit hoher Kritikalität zu treffen.

Notfalllandung: Das Verhalten des AV in Notsituationen ist entscheidend für die oben beschriebenen Sicherheitsaspekte und die Bewertung des Risikos. Da ein Ausfall, eine unkontrollierte Landung oder gar ein Absturz des AV enorme Konsequenzen mit sich bringen und nicht immer sichergestellt werden kann, dass vordefinierte abgesicherte Landepunkte erreichbar sind oder angeflogen werden können, stellt die Fä-

higkeit zur Notlandung einen elementaren Sicherheitsaspekt dar. Um eine Notfalllandung umzusetzen, müssen Funktionen aus dem Bereich der Umfelderfassung, Trajektorienplanung als auch die Entscheidungsfindung zusammengeführt werden. Dabei muss zum einen ein sicherer Landplatz in potenziell unbekanntem Terrain identifiziert werden und anschließend eine sichere Trajektorie identifiziert und angeflogen werden. Dabei spielen viele der oben angeführten KI-Anwendungen eine wichtige Rolle, wobei viele Entscheidungen auf Grundlage der Umfelderfassung getroffen werden. Daher ist die Umsetzung einer sicheren, redundanten und vertrauenswürdigen Umfelderfassung für diese Modalität von besonderem Interesse. Dabei lassen sich die Technologien auch auf andere Anwendungsfälle wie beispielsweise die automatisierte Landung von Drohnen, Helikoptern oder Flugzeugen an ihrem Bestimmungsort oder die Identifikation von Risiken am Boden (Reduzierung des Ground Risk durch die Vermeidung des Überflugs) übertragen.

Allgemeine KI-Anwendungsfelder: Weitere Bereiche, in denen KI eine kritische Funktion übernehmen kann, sind beispielsweise Predictive Maintenance und die Schätzung von Akkuzuständen und Restkapazitäten. Wie bereits oben beschrieben führt der Ausfall eines AV aufgrund von Fehlern bei diesen KI-Systemen in der Regel zu einem Absturz des AV, was einen der großen Unterschiede zur bodengebundenen Mobilität darstellt und die Zulassungs-, Sicherheits-, und Redundanzanforderungen maßgeblich begründet.

Paketdrohne: Ein bekannter Use Case ist die Auslieferung von Paketen per Drohne. Dabei kann die Paketauslieferung sowohl in Form einer Punkt-zu-Punkt-Mission (Kurierdienst) aber auch als Verteilung von Paketen im Umfeld eines Logistik-Hubs konzipiert sein. Genauso ist das Abholen von Sendungen beim Kunden denkbar. Aktuell werden diese Anwendungen intensiv erforscht, wobei erste Firmen ihre Konzepte bis zum Markteintritt gebracht haben. Speziell im Hinblick auf die Auslieferung der Pakete werden verschiedene Konzepte mit unterschiedlichen Automatisierungsanforderungen adressiert. Drohnen könnten bei der Auslieferung des Pakets sowohl einen Abwurf per Fallschirm durchführen als auch in festgelegten Bereichen landen. Aber auch hier sind weitere Automatisierungsgrade zukünftig denkbar. Äquivalent zur Identifikation sicherer Notfalllandeplätze können Drohnen in unbekanntem Terrain im Umfeld der anvisierten Adresse/Koordinate nach sicheren Landflächen suchen, um Pakete abzusetzen oder aufzunehmen. Dabei ist auch eine Interaktion des Menschen mit der Drohne denkbar, bei der der Mensch die Landefläche anzeigt oder die Drohne auf die Gestik des Menschen reagiert. Grundsätzlich müssen automatisierte Paketdrohnen in allen Szenarien in der Lage sein, Personen und kritische Umgebungen sicher zu identifizieren und eine etwaige Gefährdung zuverlässig zu vermeiden. Das gilt sowohl bei der direkten Landung als auch bei dem Abwurf von Paketen.

Automatisiertes Lufttaxi: Bei der Vision des automatisierten Lufttaxis sind von dem Punkt-zu-Punkt-Transport von Passagieren über Hubs bis hin zur flexiblen Nutzung von Lufttaxis als „Robotertaxis“ verschiedene Ausprägungen denkbar. Den einfachsten Fall stellt der geregelte Transport von einem Hub zu einem anderen dar, bei dem die Start- und Landebereiche kontrolliert sind und die Flugroute bekannt ist sowie potenziell gesichert werden kann. Sobald sich der Anwendungsbereich jedoch hiervon unterscheidet, steigen die Anforderungen an die Flexibilität der KI-Funktionen zur Umfelderfassung, Trajektorienplanung und Entscheidungsfindung maßgeblich. Mit zunehmender Frequentierung des unteren Luftraumes steigen zudem die Anforderungen an Koordination und Kommunikation. Weiterhin wird die Umsetzung von D&A-Funktionen relevanter.

Modalität Schifffahrt

In naher Zukunft werden zunehmend hochautomatisierte Schiffe auf den Weltmeeren unterwegs sein. Hochautomatisiert bedeutet aber nicht zwingend unbemannt. Es ist davon auszugehen, dass sich der Grad der Autonomie während einer Seereise ändern kann. Die Palette an möglichen Anwendungsfällen ist groß und wird je nach Einsatzzweck stark variieren. Jedoch wird sich die Anzahl der an Bord verbleibenden Besatzungsmitglieder verringern und die Schiffsbrücke wird im Normalbetrieb nicht zwingend besetzt sein. Der Fahrtverlauf wird zunehmend von Landinfrastruktur aus überwacht und aktiv gesteuert. Entsprechende Schiffe werden weite Strecken ganz ohne menschliches Eingreifen zurücklegen und innerhalb der ODD eigenständig nautische Entscheidungen treffen und ausführen. Die Möglichkeit der Fernsteuerung wird eine wichtige Rolle einnehmen und stellt zum einen eine Funktionsebene und zum anderen eine Rückfallebene dar. Eine wesentliche Grundvoraussetzung ist die sichere Erfassung des direkten Schiffsumfeldes im Nah- und Fernbereich, um stets ein aktuelles Lagebild verfügbar zu haben. Neben anderen Verkehrsteilnehmer*innen und deren Intension müssen auch im Wasser treibende Objekte und Seezeichen unter oft wechselhaften und schwierigen Umweltbedingungen sicher erkannt werden. Insbesondere Informationen aus Computer-Vision-Systemen und der dazugehörigen Sensorik werden das Lagebild vervollständigen und die Augen des

Nautikers ersetzen. Für eine sichere Navigation ist jedoch ein Situationsverständnis unerlässlich. Dies erlaubt zum einen eine KI-basierte und kontinuierliche Risikobewertung sowie eine dynamische und kollisionsverhütungsregel-(COLREG[93]) konforme Anpassung der Trajektorien, um potenziell gefährliche Schiffsbegegnungen zu minimieren. Zum anderen kann beim plötzlichen Verlassen der ODD oder bei OOD-Vorgängen die Fernsteuerung schnell an einen Remote Operator übergeben werden. Da Schiffe bei Fehlfunktionen oder Systemausfällen nicht immer einen sicheren Zustand annehmen können, müssen Redundanzen vorhanden sein.

Ein detailliertes Lagebild ist eine Grundvoraussetzung für das automatisierte Fahren in der Seeschifffahrt. Neben positions- und bewegungsspezifischen Daten müssen auch semantische Informationen durch die Sensorik erfasst werden. Zudem müssen Kollisionsverhütungsbestimmungen und Umweltbedingungen wie Wetter und Strömung mit in das Lagebild einfließen. Die KI-basierten Funktionen sind vielfältig und reichen von der Objekterkennung über Sensordatenfusion bis hin zur Auswertung. Derzeit existieren weder geeignete Datensätze noch etablierte Sensorkombinationen. Es fehlen spezifische Vorgaben hinsichtlich der Datenqualität und deren Umfang, um entsprechende Modelle zu entwickeln. Wie auch in den anderen Modalitäten werden zuverlässige und aussagekräftige Verifikations- und Validierungs- (V&V) Methoden benötigt.

Die genaue Schiffsroute liegt im Normalfall schon vor Reisebeginn fest. Situationsbedingt kommt es im Reiseverlauf aber immer wieder zu kleinen Abweichungen, beispielsweise bei Begegnungssituationen mit anderen Schiffen. Die Bestimmung der zum Teil komplexen Trajektorien erfolgt aktuell regelbasiert, jedoch wird es zukünftig mehr datenbasierte Ansätze geben, um Umwelteinflüsse und das Lagebild besser in den Entscheidungsprozess einzubetten. Es wird ein einheitliches Modell zur Beschreibung von maritimen Verkehrssituationen benötigt bzw. eine Beschreibungssprache mit entsprechenden Schnittstellen. Zur Abbildung relevanter Verkehrsszenarien müssen kritische Szenarien identifiziert werden und für die Erprobung werden Simulationen und virtuelle Testfelder benötigt. Eine „in situ"-Erprobung ist aufgrund der vielfältigen und kostenintensiven Rahmenbedingungen nur eingeschränkt möglich.

93 Convention on the international regulations for preventing collisions at sea.

Modalität Eisenbahn

Systemfremde Hindernisse verursachen oft unvermeidliche Kollisionen, da der Bremsweg der Triebfahrzeugführenden länger als die maximale geometrische Sichtweite sein kann. Die Sichtweite des Tf kann kürzer als die maximale geometrische Sichtweite sein – sie ist proportional zur Größe eines Hindernisses. Zusätzlich verringern Umweltbedingungen die Sichtweite des Tf. Dynamische Hindernisse können beliebig außerhalb und innerhalb des Bremswegs auftauchen und verschwinden. Auch wenn eine Kollision unvermeidbar ist, muss eine Schnellbremsung mit zusätzlichem Pfeifen zur Schadensreduktion ausgeführt werden. Es ist dabei nie zu spät für die Schnellbremsung [338]. Die Schadensreduktion entsteht durch eine Verzögerung der Kollision für dynamische Hindernisse, eine Wuchtreduktion und die Gefahrbereichsreduktion am Ort der Kollision. Selbst eine nach der Kollision ausgelöste Schnellbremsung verhindert die Weiterfahrt eines kollidierten Zuges. Die Maßnahme der Schnellbremsung ist im Eisenbahnbereich jedoch problematisch. Bei Fehlalarm kann sie nicht bei allen Zügen noch vor dem Stillstand wieder aufgelöst werden und ist daher mit einem wirtschaftlichen Schaden verbunden. Diese Sachlage stellt hohe Anforderungen an die Raten der falsch-negativen und der falsch-positiven Erkennungen eines im Eisenbahnbereich eingesetzten KI-Systems. Rechnerisch ergibt sich aus der Risikoabschätzung sogar eine viel kleinere Toleranz für falsch-positive als für falsch-negative Erkennungen.

Die meisten Toten im Eisenbahnkontext der EU sind keine Unfälle und entstehen durch widerrechtliches Betreten mit Selbstmordabsicht [339]. Unfälle durch widerrechtliches Betreten der Gleisanlagen ohne Selbstmordabsicht sind die größte Kategorie der für die Risikoabschätzung relevanten Toten. In beiden Fällen muss gebremst werden, auch wenn die Schadensreduktion innerhalb des Bremswegs gering ist. Die restlichen Unfälle sind erheblich seltener und können durch Sicherung und Reduktion der Zahl der Bahnübergänge stark verkleinert werden. Eine geringere Geschwindigkeit bei der Kollision mit schwereren Hindernissen und eine Vermeidung der Weiterfahrt nach einer Kollision reduzieren die Gefahr einer Entgleisung. Insassen von Straßenfahrzeugen überleben Kollisionen bei niedrigeren Geschwindigkeiten. Kollisionen mit leichten nicht-menschlichen Hindernissen (z. B. Vögel, kleine Äste und kleine Landtiere) werden ohne Reaktion hingenommen.

Zusätzlich zur Hinderniserkennung bedarf es der Kollisionserkennung und der Zustandsüberwachung. Die Kollisionserkennung verhindert die Weiterfahrt nach Kollisionen mit

nicht erkannten Hindernissen und ermöglicht die Wiederanfahrt nach einem Fehlalarm. Die Zustandsüberwachung des Fahrzeugs ist bei GoA4 eine Funktion des Tf und muss durch automatische Systeme abgebildet werden. Die Zustandsüberwachung geht nahtlos in Predictive Maintenance über, wobei die Fahrzeuge und Infrastruktur mithilfe von Sensoren und Algorithmen überwacht werden. Derzeit sind die Wartungsintervalle für sicherheitsrelevante Bauteile jedoch fest, d. h. unabhängig vom beobachteten Zustand. Da automatische Systeme den Tf nicht 1:1 ersetzen können, müssen sie menschliche Vorteile durch Fähigkeiten wie Long Range Obstacle Detection (LROD) kompensieren. Bei LROD soll ein System Hindernisse aus größerer Distanz als ein Tf erkennen. Wegen geringer Unfallhäufigkeit, hoher Hürden für den Zugang zur Infrastruktur und Auflagen für Experimente ist die Sammlung relevanter Perzeptionsdaten im Eisenbahnbereich erheblich erschwert. Durch Anwendung zustandsbasierter und vorausschauender Instandhaltung wird es möglich sein, eine Flexibilisierung von derzeit starren, im Regelwerk verankerten Fristen zur Inspektion und Wartung zu erreichen. Dafür sind allerdings gesicherte Aussagen über die Güte und Zuverlässigkeit von KI-Methoden notwendig, um entsprechende Regelwerke und Normen anpassen zu können.

Herausforderungen

Die Komplexität der – zum Teil auf KI-Verfahren basierenden – Funktionen in hochautomatisierten Mobilitätssystemen sowie insbesondere auch die Komplexität und Dynamik der Umwelt, in denen diese Systeme agieren müssen, bedingen zwangsläufig, dass eine vollständige Validierung und ein vollständiger Sicherheitsnachweis sämtlicher Verhaltensmöglichkeiten des Systems in allen denkbaren Szenarien unter allen möglichen Umweltbedingungen nicht näherungsweise realisierbar ist.

Im Eisenbahnbereich bedarf es für Automatisierungsgrade ab GoA3 nach § 45 I, § 3 I 1 EBO einer Ausnahmegenehmigung durch das Bundesministerium für Digitales und Verkehr (BMDV). Eine Zulassung von GoA3+ muss nach Common Safety Methods on risk assessment (CSM-RA) [340] erfolgen. In einfacher Darstellung kann die Zulassung nach CSM-RA auf einem der drei möglichen Wege erfolgen – durch eine Norm, nach harmonisierten Entwurfszielen und durch einen Vergleich mit dem Referenzsystem „Mensch“. Der erste Weg eignet sich wegen fehlender Normen derzeit nicht für KI-basierte Hinderniserkennungssysteme. Die vorhandene Norm DIN EN 50657:2017 [89] für Software auf Schienenfahrzeugen deckt lediglich konventionelle Software ab. Es gab weltweit mehrere Entwicklungen von Hinderniserkennungssystemen mithilfe der konventionellen Bildverarbeitungsmethoden [335]. Eines der ersten solcher Experimente war das Projekt KOMPASS vom Bundesministerium für Bildung und Forschung in 2003 [341]. Es kann derzeit nicht abgeschätzt werden, ob eine im Betrieb einsetzbare Hinderniserkennung mit konventioneller Software möglich ist. Daher erfordert der erste Weg Normen für KI-Systeme.

Sowohl im Automobil- als auch in der Schiff- und Luftfahrt müssen, wie oben bereits angedeutet, auf der einen Seite für jedes System Einsatz- bzw. Betriebsbereiche und -szenarien, die sogenannte Operational Design Domain (vgl. [342]) festgelegt werden, innerhalb derer das System eingesetzt werden darf, und im Rahmen der Typzulassung muss der Einsatz des Systems in diesen ODDs ausreichend geprüft und Trustworthiness-Eigenschaften wie insbesondere die funktionale Sicherheit nachgewiesen werden. Für Letzteres scheinen szenarienbasierte Testansätze eine ausreichende Sicherheit zu gewährleisten. Auf der anderen Seite müssen die Systeme so entwickelt werden, dass sie (a) zur Laufzeit kontinuierlich überprüfen, ob sie sich tatsächlich noch innerhalb der ODD befinden, ob sie die aktuelle Situation mit ausreichender Genauigkeit erkennen und ob in der realisierten Funktionalität Fehler auftreten; und dass sie (b) im Falle, dass der Laufzeittest fehlschlägt, auf eine sichere Rückfallebene zurückgreifen können, d. h. auf einen Betriebsmodus mit ggf. eingeschränkter Funktionalität, der trotz unbekannter Situation und Verlassen der ODD zumindest das Erreichen eines sicheren Zustands (z. B. „Anhalten am Straßenrand“) ermöglicht. Fehlgeschlagene Laufzeittests, die nicht aufgrund eines „absichtlichen“ Verlassens der ODD (z. B. das beabsichtigte Verlassen der ODD „Autobahn“, wenn die Zielausfahrt erreicht ist) verursacht werden, führen idealerweise dazu, dass aktuelle System- und Umgebungsdaten an den Herstellenden oder eine zentrale Stelle zurückgemeldet werden, und dort für die Verbesserung respektive Weiterentwicklung des Systems genutzt werden können.

Insgesamt müssen damit sowohl die Entwicklungsprozesse als auch die für eine Typzulassung bzw. Zertifizierung solcher Systeme notwendigen Analyse- und Testverfahren so erweitert werden, dass sie die kontinuierliche (Weiter-)Entwicklung solcher Systeme inklusive Updatefähigkeit, zugehöriger Laufzeittest sowie Angemessenheit und funktionale Sicherheit der gewählten Rückfallebenen erlauben und damit eine agile, kontinuierliche Zulassung bzw. Zertifizierung dieser Systeme (und ihrer Updates/Weiterentwicklungen) ermöglichen. Für KI-Systeme bzw. Systeme mit KI-basierten Komponenten ergibt sich hier insbesondere die Herausforderung der

Nachweismöglichkeiten der funktionalen Sicherheit dieser Komponenten in einer für die Typzulassung notwendigen Genauigkeit (und ggf. weiteren Qualitätseigenschaften wie Reproduzierbarkeit etc.) – dies gilt sowohl für die Typzulassung des Systems, die Zertifizierung von ggf. notwendigen Updates, die notwendigen Laufzeittests und die funktionale Sicherheit von innerhalb der Rückfallebenen realisierten KI-Funktionalitäten. Im Folgenden werden einige Teilherausforderungen, die mit dieser übergreifenden Methodenänderung verbunden sind, herausgestellt:

Wie für die Entwicklung jedes Mobilitätssystems notwendig, erfolgt auch für automatisierte Mobilitätssysteme zunächst die Anforderungserhebung und die System(-funktionalitäts)-beschreibung. Neue Herausforderungen ergeben sich in Komplexität und Umfang der Anforderungen, in denen das Systemverhalten nun in Bezug auf eine deutlich komplexere Umwelt und abhängig vom Verhalten anderer Verkehrsteilnehmer*innen in dieser Umwelt beschrieben werden muss. Eine weitere Neuerung ergibt sich aus der Notwendigkeit, die Verteilung der Fahraufgabe zwischen Mensch und System zu spezifizieren (Übergabezeitpunkte, Übergabemodalitäten, ggf. Überwachung der Aufmerksamkeit des Nutzenden bzw. der Fähigkeit und Bereitschaft des Nutzenden, die Fahraufgabe übernehmen zu können). Dies setzt sich fort in der Notwendigkeit, die Fahrzeugumgebung ausreichend gut und genau zu spezifizieren: Hierzu gehört die Beschreibung des geplanten bzw. erlaubten Betriebsbereichs (ODD) des zukünftigen Systems sowie eine Beschreibung der relevanten Objekte und Artefakte, die in diesem Betriebsbereich vorkommen können – inklusive der Aufstellung einer für die Objekterkennung nutzbaren Ontologie. Es ist zu erwarten, dass diese Beschreibungen nicht vollständig sein können, da eine vollständige Beschreibung der in der Realität vorkommenden Artefakte weder für die ODD-Beschreibung noch für die relevanten Objekte möglich ist; jedoch sollte ein systematischer Prozess auch herstellerübergreifend gefunden werden, der neben Kriterien für und Anforderungen an die Qualität und Vollständigkeit dieser Spezifikationen auch verbindliche Standardmengen für Objekte und Artefakte festlegt. Da ein vollständiger Test dieser Systeme aufgrund der Komplexität der Umgebung (bzw. der ODD) nicht mehr möglich ist, ist der Ansatz des szenariobasierten Testens solcher Systeme aktuell der vielversprechendste und derjenige, der bereits durch verschiedene bestehende und in Vorbereitung befindliche Standards gefordert wird. Hier ergibt sich als zusätzliche Herausforderung die Spezifikation und Beschreibung dieser Szenarien zunächst als Spezifikation der realisierten Systemfunktionalität, dann insbesondere jedoch auch als Grundlage für die durchzuführenden Tests und Validierungsverfahren. Für Letzteres besteht die Herausforderung zunächst darin, ausreichend viele Testfälle aus den Szenarien abzuleiten – so, dass ein möglichst großer Testraum innerhalb der ODD abgedeckt wird; dies erfordert insbesondere auch die Kombination und Rekombination der in einem Szenario beschriebenen Verhaltensweisen der Verkehrsteilnehmer*innen mit verschiedenen Umweltbedingungen wie Wetter, Lichtverhältnissen, Straßenbelagszuständen und vielem mehr. Weiterhin müssen aus diesen relevanten Szenarien auch die sogenannten „Edge Cases und Corner Cases“ identifiziert werden, also diejenigen Testfälle, die „am Rand der Systemleistung“ liegen und in denen daher das Auftreten von Fehlern am wahrscheinlichsten ist. Auch hier ist die Definition eines systematischen Prozesses zur Szenariensammlung, des Auffindens von für die jeweilige Anwendung spezifischen relevanten Szenarien nebst zugehörigen Edge Cases und Corner Cases wünschenswert. Aufgrund der Fülle möglicher Szenarien sind hierbei Prozesse zu bevorzugen, die diese Tätigkeiten – insbesondere die Sammlung von Szenarien sowie die anwendungsspezifische Erzeugung von Testdaten aus diesen Szenarien heraus – weitgehend automatisiert durchführen können; dabei ist insbesondere bei der Szenarien- und Testfallgenerierung darauf zu achten, dass Unzulänglichkeiten der zugrunde liegenden Daten (wie Bias oder Ähnlichem) entweder erkannt und berichtigt werden oder zumindest nicht zu entsprechenden Eigenschaften in der KI-Funktionalität führen. Die Sammlung dieser Szenarien in herstellerübergreifenden Datenbanken könnte dabei vorteilhaft sein, zum einen, um den Aufwand der Szenariensammlung nicht bei jedem Herstellenden wiederholt durchführen zu müssen, zum anderen auch, um einheitliche Testkriterien – also einheitliche Testszenarien pro Anwendungsfall – als Mindestanforderung festlegen zu können.

Bestimmte KI-Anwendungen, insbesondere sicherheitskritische, können zudem nicht vollständig im Feld bzw. mit Realdaten geprüft werden. Die kritischen Ereignisse, die beispielsweise bei einer Prüfung von Fahrautomatisierungsfunktionen höherer Autonomiestufen herangezogen werden müssen, sind zu unterschiedlich und können daher zu selten beobachtet werden. Hinzu kommt, dass aus offensichtlichen ethischen Gründen eine Prüfung durch Auftretenlassen kritischer Situationen („Kind läuft vor Auto“) nicht infrage kommt, auch wenn diese in dem Zusammenhang nicht mit entsprechenden Folgen hervorgerufen würden. Daher ist es notwendig, sogenannte synthetische Daten zur Hilfe zu nehmen, die aus einer Simulation („Digitaler Zwilling“) heraus erzeugt werden. In der Simulation können kritische Szenarien gezielt erzeugt

werden, jedenfalls diejenigen, die bekannt sind bzw. im Feld zumindest mit einer gewissen Häufigkeit vorkommen. Darüber hinaus gibt es in der Simulation die Möglichkeit, sämtliche relevanten Parameter bis an die Grenzen des physikalisch Erwartbaren zu setzen, sodass auch möglicherweise bislang gänzlich unbeobachtete und dennoch mögliche kritische Szenarien erzeugt werden können. Gegenüber der Prüfung im Feld hat diese Vorgehensweise darüber hinaus den Vorteil, dass Ereignisse reproduzierbar sind, was die Analyse der gewonnenen Resultate erheblich erleichtert.

Unabhängig davon, wie hoch der Automatisierungsgrad eines Systems ist, müssen geeignete Rückfallebenen vorhanden sein, die einen sicheren Betrieb oder einen sicheren Stopp des Systems auch für den Fall ermöglichen, dass das System seine Aufgabe nicht mehr funktional sicher durchführen kann. Dieser Fall kann sowohl in Systemen wie dem Autobahn-Chauffeur auftreten, wenn das System z. B. feststellt, dass es aufgrund von z. B. Witterungseinflüssen seine Umgebung nicht mehr richtig erkennen kann, es daher eine Übergabeaufforderung an den Fahrenden stellt, dieser die Fahraufgabe jedoch nicht übernimmt. Aber selbst in vollautomatisierten Fahrzeugen (SAE Level 5) kann durch den Ausfall von Teilsystemen – z. B. Steinschlag in die zur Umgebungswahrnehmung eingesetzte Kamera – oder andere Faktoren wie das überraschende Verlassen der ODD die Fahraufgabe ggf. nicht mehr vollständig durch das System durchgeführt werden. Sichere Rückfallebenen bestehen dabei aktuell typisch aus der Durchführung eines MRM (Minimum Risk Manouver; z. B. Fahren an den rechten Straßenrand und anhalten). Zukünftig können komplexere Rückfallebenen (z. B. Fortführung der Fahraufgabe mit ggf. stark reduzierter Geschwindigkeit) möglich sein. Neben der Herausforderung der kontinuierlichen Selbstüberwachung, die das System durchführen muss, um festzustellen, ob es die Fahraufgabe noch erfüllen kann, ist insbesondere die Definition geeigneter Rückfallebenen – die ja idealerweise abhängig vom konkret aufgetretenen Fehler sein sollten – sowie der Safety-Nachweis dieser Rückfallebenen ein aktuell ungelöstes Problem.

Auch für die Übergabe der Fahraufgabe sind besondere Vorkehrungen zu treffen. Dabei ist die Übergabe an das System typisch einfach und geschieht oft durch manuelles Einschalten der entsprechenden Automatisierungsfunktion durch die Fahrerin bzw. den Fahrer. Typischerweise testet das System, ob es sich innerhalb seines Betriebsbereichs befindet und funktionsfähig ist, übernimmt die Aufgabe und bestätigt diese Übernahme. Die Übergabe der Fahraufgabe an den Fahrer gestaltet sich aufwendiger. Normalerweise sind Zeitfenster vorgegeben, innerhalb derer der Fahrer oder die Fahrerin eine Aufforderung zur Übernahme akzeptieren muss (andernfalls wird ein MRM ausgelöst, siehe oben). Die Herausforderung besteht dann darin, während des Betriebs zu überwachen, dass die Nutzenden so weit aufmerksam sind, dass bei einer ggf. nötig werdenden Übergabeaufforderung die Annahme innerhalb dieses Zeitfensters möglich ist. Weiterhin muss das System innerhalb dieses Zeitfensters (also nachdem eine Situation eingetreten ist, die zur Aufforderung der Übernahme an den Fahrenden geführt hat) die Fahraufgabe weiterhin sicher ausführen können – oft sogar auch noch das danach ausgeführte Minimal-Risk-Manöver. Je nach der konkreten Situation und den konkreten Bedingungen, die zur Aufforderung der Übernahme geführt haben, ist die funktionale Sicherheit des Systems in diesen Zeiträumen oft nur sehr aufwendig sicherstellbar. Schließlich ist für die Übergabe der Fahraufgabe an den Fahrenden auch zu prüfen, ob eine langsame, transiente und teilassistierte Übergabe an den Fahrer – trotz Vorhandensein einer die Übergabeaufforderung auslösenden Situation – nicht die Gesamtsicherheit des Systems erhöhen würde.

Um dieser Fülle von Herausforderungen zu begegnen, bietet sich der Aufbau einer einheitlichen, ggf. sogar herstellerübergreifenden Infrastruktur zur Unterstützung der oben beschriebenen kontinuierlichen Weiterentwicklung hochautomatisierter Mobilitätssysteme an.

4.6.3 Normungs- und Standardisierungsbedarfe

Bedarf 06-01: Erfassung Stand Trustworthiness by design und Prüfbarkeit

Als Grundlage zur Operationalisierung des geplanten EU AI Act (aktuell 2. überarbeitete Entwurfsfassung) stellen umfassende Garantien im Hinblick auf die Vertrauenswürdigkeit von KI-Technologie im Bereich der Mobilität eine große Herausforderung dar und nötige rechtliche und organisatorische Rahmenbedingungen und technische Methoden und Werkzeuge stehen aktuell nicht hinreichend zum praxistauglichen Einsatz zur Verfügung. Der zu berücksichtigende Parameterraum ist hochdimensional, es sind u. a. zu betrachten: a) die Lebenszyklusphasen und die Einbettung des KI-Systems, b) die verschiedenen Trustworthiness-Aspekte (TW-Aspekte: Safety, Security, Robustheit, Transparenz, Fairness, Erklärbarkeit usw.), c) die verschiedenen KI-Modelle und -Lernverfahren und d) die verschiedenen Use Cases (Domänen, Modalitäten und Funktionalitäten) im Bereich Mobilität.

Der Stand der Entwicklung und Prüfbarkeit von KI-Systemen soll im Hinblick auf den oben genannten Parameterraum systematisch für relevante Anwendungen im Bereich Mobilität erfasst und über die Zeit nachverfolgt werden, um weitere Forschungs- und Entwicklungs (F&E)-Arbeiten, vor allem in den Bereichen „X-by-Design“, Prüfbarkeit und Nachweisbarkeit von System- und Komponenten-(Trustworthiness-) Eigenschaften wie IT-Sicherheit, Zuverlässigkeit, Erklärbarkeit und Introspektionsfähigkeit, sowie Maßnahmen zur Absicherung sinnvoll zu priorisieren.

Die zu etablierenden Normen und Standards sollten insbesondere umfassen:

→ Etablierung einer Methode, die den objektiven Vergleich der Trustworthiness-by-design-Entwicklung und der Prüfbarkeit im Hinblick auf verschiedene Anwendungen und über die Zeit erlaubt,
→ konkrete, praxistaugliche Bewertungskriterien für den gesamten relevanten Parameterraum,
→ Erläuterung der Methode anhand konkreter Beispiele.
→ Soweit möglich soll hierbei auf vorhandenen Normen und Standards aufgebaut werden, z. B. ISO 21448:2022 [90].

Aufgrund der großen Hebelwirkung (Vermeidung von Doppelarbeit, Nutzung von Synergien und Fokussierung auf wesentliche Arbeiten) sollten für diesen Bedarf von der Politik unbedingt die nötigen Ressourcen bereitgestellt werden.

Bedarf 06-02: Erarbeitung und praktische Umsetzung fehlender technischer, rechtlicher und organisatorischer Grundlagen

Bedingt durch die hohe Komplexität (vgl. Bedarf 06-01) sind insbesondere viele technische, aber auch rechtliche und organisatorische Grundlagen für eine Trustworthy-by-design-Entwicklung, die Prüfung und die Absicherung im Betrieb entweder gar nicht oder nicht hinreichend praktisch umsetzbar und verfügbar. Diese sind jedoch Voraussetzung für hinreichende Garantien hinsichtlich der Vertrauenswürdigkeit von KI-Systemen im Bereich Mobilität. Die bisher fehlenden oder nicht hinreichend praxistauglich umsetzbaren technischen, organisatorischen und rechtlichen Grundlagen in Bezug auf die Vertrauenswürdigkeit von KI-Systemen im Kontext Mobilität (vgl. Bedarf 06-01) sollen daher systematisch erarbeitet und praxistauglich umgesetzt werden. Hierzu gehören insbesondere geeignete Metriken (Key Trustworthiness Indicators) sowie die Definition von Mindestqualitäten auf Basis dieser Metriken (z. B. „akzeptables Restrisiko“), Vulnerabilitäten, Interpretationsmethoden, Absicherungsmaßnahmen, Verantwortlichkeiten und ihre jeweilige Abhängigkeit von den Randbedingungen (ODD).

Die zu etablierenden Normen und Standards sollten insbesondere umfassen:

→ Technische, organisatorische und rechtliche Grundlagen für alle praxisrelevanten Kombinationen von Lebenszyklusphase, Trustworthiness-Aspekt, Use Case / Funktionalität und KI-Technologie.
→ Randbedingungen, unter denen diese Grundlagen Gültigkeit haben, und praktische Hinweise, wie die Randbedingungen so angepasst werden können, dass die Trustworthiness gesteigert wird.
→ Herbei sollte auf relevante bestehende sektorübergreifende (v. a. DIN SPEC 13266:2020 [98]) und sektorspezifische Normen und Spezifikationen (z. B. ISO-26262-Reihe [455] für die Straße; DIN EN 50657:2017 [89], DIN VDE V 0831-101:2022 [344], DIN VDE V 0831-103:2020 [343] und DIN EN 62267:2010 [332] für die Eisenbahn) aufgebaut werden und sektorspezifische Normen und Spezifikationen sollten entsprechend erweitert werden (z. B. ISO 21448:2022 [90] Erweiterung auf den Bahnbereich und weitere Mobilitätsbereiche).

Ohne eine zügige Erarbeitung dieser Grundlagen wird eine fristgerechte Operationalisierung des AI Act ([4], aktuell 2. überarbeiteter Entwurf) nicht gelingen und daher sollten für diesen Bedarf von der Politik unbedingt die nötigen Ressourcen bereitgestellt werden.

Bedarf 06-03: Generalisierter und leicht auf spezifische Domänen und Use Cases anpassbarer Anforderungskatalog

Bedingt durch die Komplexität des Einsatzes von KI-Technologie im Bereich Mobilität können nicht im Vorhinein für alle Kombinationen von Lebenszyklusphasen, Trustworthiness-Aspekten, KI-Technologien und Anwendungen bzw. Funktionalitäten spezifische Anforderungen formuliert werden.

Ein generalisierter modularer Anforderungskatalog hinsichtlich technischer, organisatorischer und rechtlicher Aspekte soll zusammen mit praktischen Hinweisen und konkreten Beispielen zur Anpassung an beliebige Use Cases im Bereich Mobilität entwickelt werden. Spezifische Anpassungen an konkrete Anwendungen und Funktionalitäten sowie an Randbedingungen für diese Anwendungen (z. B. sicherheitskritische Anwendungen) sollen mit möglichst geringem Aufwand möglich sein.

Die zu etablierenden Normen und Standards sollten insbesondere umfassen:

- → einen umfassenden modularen und anwendungsagnostischen Anforderungskatalog hinsichtlich technischer, organisatorischer und rechtlicher Aspekte,
- → detaillierte Anweisungen und praktische Beispiele für eine anwendungsspezifische Anpassung des Anforderungskatalogs, wobei „anwendungsspezifisch“ sowohl die konkreten Anforderungen und Eigenschaften der Anwendung als auch der Einsatzumgebung der Anwendung umfasst (z. B. Sicherheitskritikalität),
- → Hinweise zur günstigen Gestaltung der Rahmenbedingungen, um einerseits die Trustworthiness zu steigern und andererseits die Entwicklungs- und Prüfaufwände zu reduzieren,
- → Architekturen und Architekturmuster zur Reduzierung der Fortpflanzung von Unsicherheiten und Steigerung der Zugänglichkeit für Nachweisverfahren,
- → Methoden zur Introspektion und zum Nachweis von Safety und Zuverlässigkeit von KI,
- → anwendungsspezifische Risikoakzeptanzkriterien.

Bedarf 06-04: Kontinuierliche (Weiter-)Entwicklung und Validierung im Betrieb

Wie in Kapitel 4.6.2 dargestellt, müssen sowohl die Entwicklungsprozesse für hochautomatisierte Mobilitätssysteme als auch die für eine Typzulassung bzw. für eine Zertifizierung solcher Systeme notwendigen Analyse- und Testverfahren so erweitert werden, dass sie die kontinuierliche (Weiter-)Entwicklung solcher Systeme inklusive Updatefähigkeit basierend auf im Feld gesammelten Daten, zugehörigem Laufzeittest sowie der Angemessenheit und funktionalen Sicherheit der gewählten Rückfallebenen erlauben. Für KI-Systeme bzw. Systeme mit KI-basierten Komponenten ergibt sich hier insbesondere die Herausforderung der Nachweismöglichkeiten der funktionalen Sicherheit dieser Komponenten in einer für die Typzulassung notwendigen Genauigkeit und Umfang – dies gilt sowohl für die Typzulassung des Systems, die Zertifizierung von ggf. notwendigen Updates, den notwendigen Laufzeittests und der funktionalen Sicherheit von innerhalb der Rückfallebenen realisierten KI-Funktionalitäten. Insgesamt müssen diese Prozesse und Verfahren somit eine dynamische, kontinuierliche (Re-)Zertifizierung bzw. Typzulassung im Sinne einer kontinuierlichen Systementwicklung erlauben.

Die zu etablierenden und bestehenden Standards müssen sowohl diese Entwicklungsprozesse unterstützen als auch Anforderungen an die Typzulassung definieren und dabei insbesondere die Updatefähigkeit sowie das über Laufzeitprüfung und Rückfallebenen realisierte Sicherungskonzept unterstützen. Neben der besonderen Berücksichtigung von KI-Komponenten im o. a. Sinne sollten hierbei auch Erkenntnisse in Form von kritischen Szenarien herstellerübergreifend katalogisiert werden können. Dies dient auf der einen Seite der kontinuierlichen Verbesserung der Systeme und auf der anderen Seite der Schärfung der Sicherheitsanforderungen (z. B. bei Domain-Shifts).

Die zu etablierenden Standards und Normen sollten daher insbesondere umfassen:

- → systematische Identifikationsprozesse für kritische Szenarien,
- → herstellerübergreifende Schnittstellen, Austauschprozesse und Vorgaben für ein Ökosystem mit unabhängigen Stellen (insbesondere für Szenarienkataloge),
- → Vorgaben zur Überwachung, Prüfung, Absicherung und Zertifizierung von Systemen mit KI-Komponenten innerhalb eines kontinuierlichen Entwicklungs- und Update-Prozesses,
- → Best Practices zur Mitigation von KI-System-Fehlfunktionen im Bereich Mobilität,
- → Leitlinien / Best Practice für eine Safe-/Trustworthy-by-design-Entwicklung für relevante Anwendungsfälle (s. Spalte Anwendungsfälle) bzw. idealerweise generalisierte Empfehlungen mit konkreten Hinweisen zur anwendungsspezifischen Anpassung,
- → Vorgaben zu sicheren Rückfallebenen einschließlich einer Kontrollübernahme durch den Menschen,
- → Handlungsempfehlungen zur Festlegung von Verantwortlichkeiten bei der Entwicklung, der Prüfung und dem praktischen Einsatz von KI-Technologie in der Mobilität.

Bedarf 06-05: Analyse-, Simulations- und Testmethoden sowie Testinfrastruktur

Die Komplexität der Anwendung von KI-Technologie in der Mobilität erfordert a) interdisziplinäres Wissen, b) standardisierte Methoden und Werkzeuge, c) große Mengen an qualitätsgesicherten Daten, die oftmals Beschränkungen der Nutzung z. B. hinsichtlich Datenschutz unterliegen, und d) große Rechenressourcen für Simulationen, Training und Prüfung. Dies erfordert zum einen den Einsatz simulativer Methoden, zum anderen eine Erprobungs- und Testinfrastruktur, deren Anforderungen nur durch sehr wenige Großkonzerne oder staatlichen Akteur*innen erfüllt werden können. Hier ist eine enge Kooperation mit vielen Partner*innen erforderlich, u. a. durch Informationsaustausch, gemeinsame Projekte, geteilte Daten- und Rechenressourcennutzung (vgl. auch [345]). Um eine solche Kooperation zu ermöglichen und auch

eine Vergleichbarkeit automatisierter Mobilitätssysteme und der eingesetzten KI-Komponenten insbesondere bezüglich Trustworthiness und Safety und deren Nachweisen zu etablieren, sind nicht nur die Mindestanforderungen an die Systeme zu standardisieren. Es ist vielmehr auch notwendig, geeignete Methoden zur Unterstützung der Entwicklung der Systeme und Überprüfung ihrer Eigenschaften zu definieren. Dabei stellen simulative Methoden eine kostengünstige und ungefährliche Möglichkeit zur Unterstützung der Entwicklung und Überprüfung von KI-Komponenten und Systemen mit KI-Komponenten für Mobilitätslösungen dar. Ohne diese zu standardisieren und durch Qualitätskriterien überprüfbar zu machen, kann eine Vergleichbarkeit der Ergebnisse nicht erreicht werden. Deshalb ist auch in vielen Anwendungsdomänen die Entwicklung und Zurverfügungstellung einer gemeinschaftlich nutzbaren (virtuellen oder physischen) Testinfrastruktur sinnvoll, um einen engen interdisziplinären und internationalen Informationsaustausch, die gemeinsame Nutzung von Daten und Rechenressourcen zur Entwicklung und Prüfung in Simulation und der physischen Welt und den Austausch von Methoden und Werkzeugen einfach zu ermöglichen. Bezüglich der funktionalen Sicherheit kommt der Qualität der Simulationsverfahren eine besondere Bedeutung zu; hier muss sichergestellt werden, dass die Simulation eine ausreichend hohe Übereinstimmung mit der Realität hat, um belastbare Aussagen für die Typzulassung und die Zertifizierung dieser Systeme zu erhalten. Aktuell existieren weder Methoden noch Argumentationsketten, die diese Übereinstimmung im ausreichend hohen Maße garantieren.

Die zu etablierenden Normen und Standards sollten insbesondere umfassen:

- → virtuelle Simulations- und Testmethoden, Prüfumgebungen und deren Qualität,
- → Verfahren zur Verifikation und Validierung (v. a. Erweiterung des „Sotif-Standards“ ISO 21448:2022 [90] auf andere Domänen wie z. B. Eisenbahn),
- → Leitlinien zur Zertifizierung von KI sowie Entwicklungs- und Testmethoden,
- → standardisierte Begrifflichkeiten zur effizienten Kommunikation,
- → standardisierte Schnittstellen zum Austausch von Daten, Modellen und Simulationen,
- → standardisierte Vorgehensweisen zur gemeinsamen Datenhaltung, Entwicklung und Prüfung von KI-Systemen.

Bedarf 06-06: Szenarien, Datensätze, Interoperabilität, Schnittstellen, Datenaustausch, Datenqualität, Digitale Zwillinge

Ein Austausch und eine Vergleichbarkeit von KI-Komponenten und deren vertrauenswürdiger und sicherer Einsatz in automatisierten Systemen erfordert standardisierte Schnittstellen und Mindestanforderungen; die konkrete Ausgestaltung dieser zu standardisierenden Schnittstellen und Mindestanforderungen ist dabei zumindest in Teilen aktiver Forschungsgegenstand. Diese Schnittstellen sollten vorrangig die Verwendung innerhalb von Testumgebungen und zusammen mit den Anforderungen auch den Betrieb innerhalb der eingesetzten Systeme für die Testdurchführung festlegen. Hier ist es notwendig, den Minimaleinsatzbereich für unterschiedliche Anwendungsfälle einheitlich zu definieren. Aufgrund der Komplexität im Mobilitätsbereich gelten Szenarien als Mittel der Beschreibung und Strukturierung des intendierten Einsatzbereichs (ODD) auf Systemebene. Um eine Vergleichbarkeit der Anforderungserfüllung zu ermöglichen, müssen auch die bei der Prüfung der Systeme verwendeten Kriterien für kritische Szenarien sowie (Feld-)Datensätze und Szenarienkataloge und zur Erreichung von Interoperabilität und Ermöglichung des Datenaustauschs weiterhin die Austauschformate für Felddaten, Szenarien und Datensätze standardisiert werden. Für standardisierte Komponententests insbesondere innerhalb der Perzeption sind weiterhin standardisierte Sensorkonfigurationen, welche von einer Testumgebung zur Verfügung gestellt werden müssen, notwendig. Zur kostengünstigen Ergänzung von Testdatensätzen sollten auch die Verwendung von Digitalen Zwillingen und Qualitätsanforderungen insbesondere zur Erzeugung synthetischer Daten standardisiert werden. Wo notwendig, müssen entsprechende Forschungsarbeiten die Standardisierungs- und Normungsarbeiten begleiten oder diesen vorangehen.

Die zu etablierenden Normen und Standards sollten insbesondere umfassen:

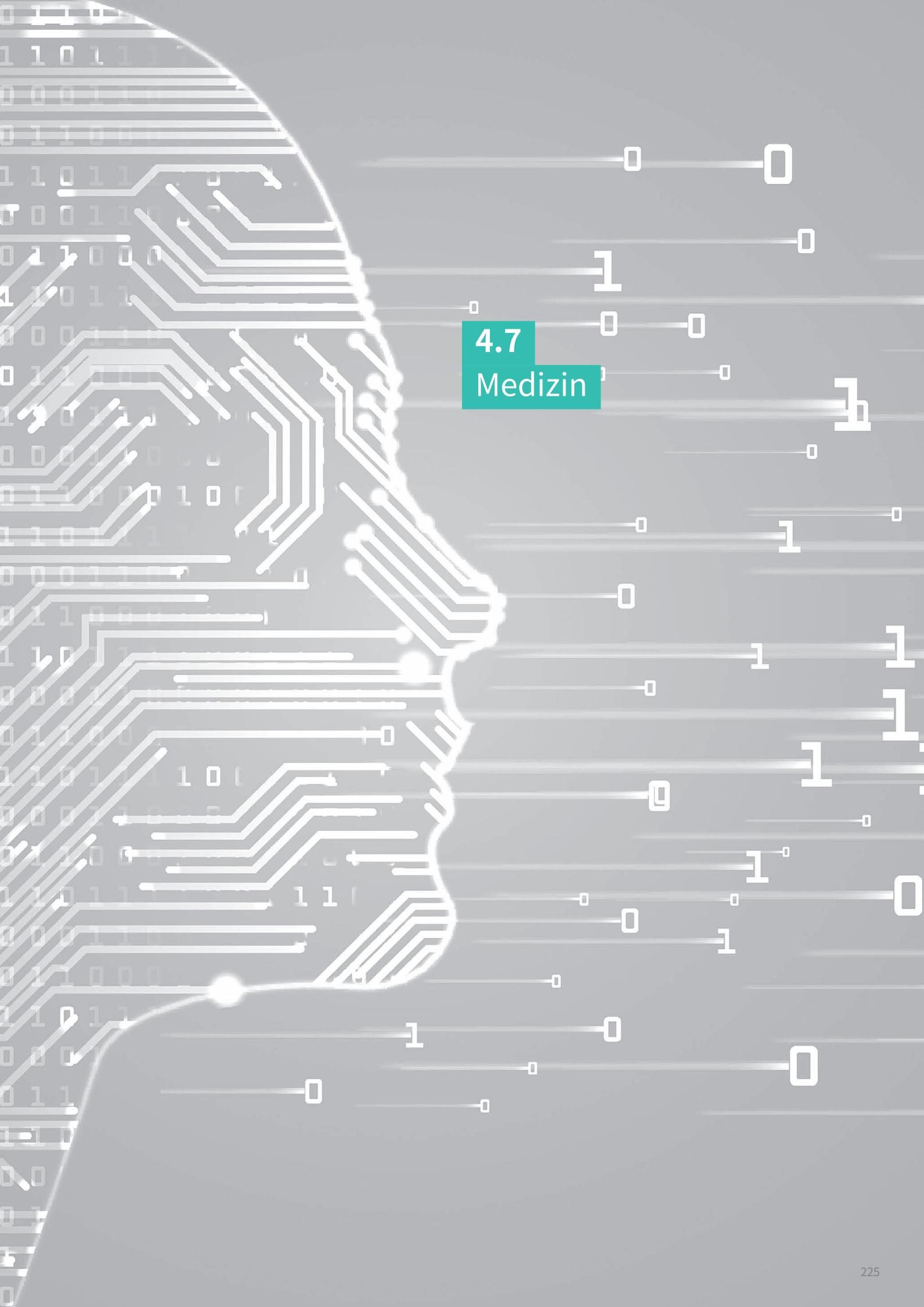
- → einheitliche Beschreibung der ODD, Szenarien und ggf. Schnittstellen für unterschiedliche Systeme,
- → Mindestanforderungen, Spezifikationen und unterstützende Austauschformate für Szenarien und Datensätze,
- → Kriterien zum Labeln von Daten und Szenarien sowie zur Überdeckung der ODD,
- → Kriterien für kritische Szenarien (insbesondere bezogen auf Safety, aber anwendungsbezogen auch auf alle Trustworthiness-Aspekte),
- → Standarddatensätze (inklusive Edge Cases und Corner Cases), Standardsensorkonfigurationen und Anforderungen an die Qualität der Datensätze,

→ Best Practices für die Erzeugung, Qualitätssicherung und den Einsatz synthetischer Daten,
→ Vorgaben zur Nachverfolgbarkeit von genutzten Daten z. B. zur Verhinderung eingeschleuster Hintertüren (Poisoning Backdoor Attacks).

Bedarf 06-07: Prüfung mit synthetischen Daten
Bezüglich der Normung und Standardisierung stellt sich die Frage, wie die Validität der synthetischen Daten, die für die Prüfung herangezogen werden, gewährleistet werden kann. Die Frage nach den Daten, die für das Training verwendet wurden, kann dabei zunächst als nachrangig angesehen werden, denn wenn die Prüfung rigoros, umfassend und valide ist, wird damit implizit auch die Qualität der Trainingsdaten aufgedeckt. Die synthetischen Prüfdaten müssen einen ausreichend geringen Unterschied zu den im Feld vorzufindenden Daten haben. Dieser Unterschied bezieht sich aber nicht auf den subjektiven Eindruck von „Echtheit“, den ein menschlicher Betrachter hat. Vielmehr müssen objektive und aufgabenspezifische Maße zugrunde gelegt werden. Selbst bei einer kamerabasierten Perzeptionskomponente steht nicht von vornherein fest, dass ein synthetisches Bild möglichst echt wirken muss, oder dass dies allein ausreichen würde, um die Validität zu gewährleisten. Bei einer Verhaltenskomponente (Fahrstrategie und Trajektorienplanung) liegen darüber hinaus synthetische wie reale Daten wesentlich abstrakter vor, z. B. als Positionsdaten von ausgerichteten Rechtecken, die sich über die Zeit verändern. Die Aufgabe der Normung und Standardisierung besteht also darin, für alle relevanten Komponenten (Anwendungsfälle) festzulegen, unter welchen Bedingungen synthetische Daten den Realdaten in hinreichendem Maße entsprechen. Die Validität der synthetischen Daten muss kontinuierlich überprüft werden, da eine Erweiterung der zugrunde liegenden Szenarien bzw. ein Angleichen an die sich verändernden tatsächlichen Begebenheiten (Straßentopografie etc.) unter Umständen Unterschiede hervorruft, die zuvor nicht aufgetreten waren. Die Prüfung der Validität muss zudem dem Umstand gerecht werden, dass kein 1:1-Vergleich aller Szenarien stattfinden kann. Es muss vielmehr sichergestellt werden, dass die Extrapolation (Szenarien, die ausschließlich synthetisch vorliegen) valide ist. Darüber hinaus muss festgelegt werden, wann ein Test eine hinreichende Abdeckung von kritischen Szenarien aufweist. Neben einer Normung der Herangehensweise bei der Erzeugung synthetischer Daten für die Prüfung von Automatisierungsfunktionen mit KI-Anteilen ist es daher denkbar, dass regelmäßig aktualisierte Szenarienkataloge von unabhängigen Stellen vorgehalten werden.

Die Arbeitsgruppe Mobilität hat die identifizierten Bedarfe nach der Dringlichkeit ihrer Umsetzung bewertet. Abbildung 42 zeigt die Dringlichkeit der Umsetzung, kategorisiert nach den Zielgruppen Normung, Forschung und Politik.

Abbildung 42: Priorisierung der Bedarfe aus Schwerpunkt Mobilität (Quelle: Arbeitsgruppe Mobilität)



4.7 Medizin

Die Nutzung von KI zur Verbesserung der medizinischen Versorgung ist einer der Anwendungsbereiche, welche die Europäische Union (EU) als ein zentrales Anwendungsfeld mit großem Potenzial sieht [7], [346]. Der Einsatz von KI in der Medizin zum Zwecke von Diagnosestellung, Screening, Therapie(-empfehlung), Monitoring, Triage und Prognose von Erkrankungen erfolgt sowohl in gering regulierten Bereichen zur Optimierung der Organisation von Gesundheitseinrichtungen, des Gesundheitssystems insgesamt oder von allgemeinen Gesundheits-Apps als auch in stark regulierten Bereichen von Medizinprodukten. Die hier dargelegten Themen gelten analog auch fortlaufend für das In-vitro-Diagnostikum.

KI-gestützte Algorithmen sind in der Lage, große Mengen an multimodalen Daten zu analysieren und hierbei in relativ kurzer Zeit Muster zu erkennen, wozu der Mensch nur eingeschränkt in der Lage wäre. So können KI-Systeme bereits heute menschlichen Expert*innen in einzelnen medizinischen Aufgabenstellungen überlegen sein (Beispiel Hautkrebs-Screening; [347]).

Besonders bei medizinischen Aufgabenstellungen müssen vor der Anwendung eines neuen Produkts am Menschen hohe Sicherheitsanforderungen erfüllt werden. Notwendigerweise ist damit die Entwicklung, Implementierung und das für den Marktzugang erforderliche Konformitätsbewertungsverfahren KI-basierter Medizinprodukte ein komplexer Prozess mit mannigfachen regulatorischen, ethischen, technischen und klinischen Anforderungen. Inzwischen haben eine Reihe medizinischer KI-Anwendungen erfolgreich derartige Konformitätsbewertungsverfahren durchlaufen oder sind bereits erfolgreich zugelassen worden (siehe z. B. [348] für KI-Medizinprodukte in der EU und USA sowie alle von der Food and Drug Administration (FDA) zugelassenen KI-Produkte [349]). Für die stark datengetriebenen Ansätze von KI- bzw. ML-basierten Systemen gibt es spezifische Aspekte, die gegenüber nicht KI-basierten Systemen auf neue bzw. erweiterte Weise berücksichtigt werden müssen, um das Konformitätsbewertungsverfahren erfolgreich zu durchlaufen: Beispiele sind die Qualität von Daten und Echtzeitentscheidungen, die Verlässlichkeit von Ergebnissen, die Komplexität der Modelle, die effektive Integration in bestehende klinische Abläufe und IT-Systeme.

Hierfür sind allgemeingültige Normen und Standards zu entwickeln, die für den Bereich KI-basierter Medizinprodukte Stand heute größtenteils weder auf nationaler noch auf europäischer oder internationaler Ebene existieren. Dabei ist eine generische, alle Facetten der Anwendung von KI im Bereich Medizin integrierende Betrachtung nur schwer möglich. Im Folgenden werden daher anhand von drei Use Cases aus den Bereichen medizinische Bildgebung, Zahnmedizin und Intensivmedizin konkrete Aufgabenstellungen behandelt, um Handlungsbedarfe für die Entwicklung geeigneter Vorgehensweisen und Normen abzuleiten.

4.7.1 Status quo

Die Konformitätsbewertung bei Medizinprodukten wird in der EU zentral über die Medical Device Regulation (MDR, [350], Stand 2021) reguliert. Für die dort beschriebenen Anforderungen gibt es bereits eine Reihe an Normen, die zentrale Aspekte wie Qualitätsmanagement [381], Risikomanagement ([351], [352]), Softwarelebenszyklus ([353], [354]) oder Gebrauchstauglichkeit ([355], [357]) abdecken und die seit Längerem in der Medizintechnikbranche etabliert sind. Diese Normen setzen allgemeine Anforderungen an Medizinprodukte um, enthalten aber keine spezifischen Anforderungen an KI-basierte Systeme. Parallel dazu gibt es horizontale, d. h. branchenübergreifend ausgerichtete Regelwerke zur Umsetzung KI-spezifischer Anforderungen wie die Normen der IEEE-7000er-Serie (2021) [10], [11], [12], [13] oder solche, die derzeit im ISO/IEC JTC1/SC42 erarbeitet werden. Diese betrachten jedoch keine speziellen Anforderungen an Medizinprodukte und können vorhandene Lücken bezüglich der erhöhten Anforderungen in der Medizin nur begrenzt schließen.

Um dennoch verlässliche Vorgehensweisen für die Umsetzung KI-basierter Medizinprodukte bzw. deren Konformitätsbewertung zu erhalten, hat z. B. die Interessengemeinschaft der Benannten Stellen für Medizinprodukte in Deutschland (IG-NB) darauf basierend einen Leitfaden „Künstliche Intelligenz bei Medizinprodukten“ [358] herausgegeben, welcher zentrale Anforderungen für KI-basierte Medizinprodukte systematisch erfasst und damit eine Hilfestellung für den Konformitätsbewertungsprozess liefert. Viele Benannte Stellen greifen bei der Prüfung von KI-Systemen auf diesen Fragebogen als zentrale Referenz zurück. Aktuell wird dabei davon ausgegangen, dass ein KI-basiertes System immer einen eingefrorenen Zustand hat, wenn es bewertet wird. Ein Weiterlernen nach der Inbetriebnahme beim Kunden würde somit eine erneute Konformitätsbewertung erfordern, sobald substanzielle Änderungen am KI-System vorgenommen werden. Es gibt aktuell dabei keine normativen oder regulatorischen Vorgaben, was substanzielle Änderung bedeutet. Ein ähnlicher Stand gilt in den USA, wo es ebenfalls keine

spezifischen Vorschriften für die Regulierung KI-basierter Medizinprodukte gibt. Dort hat die FDA im April 2019 mit [139] einen Vorschlag für die Regulierung KI-basierter Medizinprodukte gemacht, welcher jedoch wie in Europa noch nicht in konkrete Guidance-Dokumente umgesetzt wurde. Gleichwohl umfasst er auch KI-Systeme, die während des Betriebs kontinuierlich weiterlernen. Auf Basis eines fixierten Standes ist es jedoch bereits möglich, KI-basierte Medizinprodukte auf den Markt zu bringen. Das zeigt sich z. B. an der Liste an inzwischen über 300 allein in den USA zugelassenen Produkten (vgl. [349]). Auch in Europa bzw. Deutschland gibt es bereits Systeme, die auf dem Markt sind.

Durch den geplanten AI Act werden in den EU-Anforderungen formuliert, welche KI-spezifische Aspekte in Zukunft rechtlich verbindlich adressieren. Die zahlreichen Rückmeldungen relevanter Marktteilnehmer hat deutlich gemacht, dass ein weiterer Harmonisierungsbedarf mit bestehenden regulatorischen Anforderungen wie der Medical Device Regulation (MDR) besteht. Werden die Widersprüche beispielsweise zwischen MDR und geplantem AI Act nicht aufgelöst, ist mit Mehraufwänden zu rechnen oder werden gar Marktzugänge verweigert, da im Konformitätsverfahren gleichzeitig die Anforderungen der MDR und der Entwurf zum AI Act umgesetzt werden müssen (siehe auch Kapitel 1.4 sowie Anhang 13.1 Abschnitt „Exemplarische Darstellung am Beispiel Medizinprodukte“).

Grundsätzlich ist zu beachten, dass KI-basierte Medizinprodukte einige Besonderheiten aufweisen, die in anderen Einsatzbereichen nicht in gleicher Weise zum Tragen kommen und daher auch in der Normung gesondert zu betrachten sind. Dazu gehören die folgenden Kernaspekte:

→ HÖCHSTPERSÖNLICHE DATEN:
Medizinische Daten sind in der Regel stark personenbezogen und reichen oftmals in sensible Bereiche hinein. Für Europa und speziell Deutschland als wichtigem Medizintechnikstandort ist zudem zu beachten, dass der Zugang zu Daten aufgrund der bestehenden Datenschutzregeln und zusätzlicher Datenschutzgesetze auf Bundes- und Landesebene im Vergleich zu Ländern wie USA oder China stärker reglementiert ist. Dem gegenüber steht, dass gerade in der MDR umfassendes Datenmaterial für den Nachweis der Sicherheit gefordert wird. Diese Problematik hat inzwischen auch die EU aufgegriffen, indem sie z. B. im Entwurf für einen European Health Data Space (EHDS [359]) einen besseren Zugang zu medizinischen Daten ermöglichen will. Da sich der EHDS aktuell noch in Planung befindet, verbleiben einige Punkte noch ungeklärt. Das betrifft u. a. die Frage, wie für unmittelbar Betroffene und weitere Akteur*innen zukünftig differenzierte Zugangsmöglichkeiten zu Gesundheitsdaten im Einklang mit der DSGVO sichergestellt werden. Das beinhaltet, inwieweit und unter welchen Voraussetzungen ein Unternehmen für die Entwicklung kommerzieller Produkte einen Zugang zu medizinischen Daten zweckgebunden garantiert werden kann. Sehr wesentliche und komplexe Teilaspekte bilden die Verwendung anonymisierter vs. pseudonymisierter medizinischer Daten sowie das Problem einer Rückidentifikation persönlicher Informationen bei bestimmten Datentypen (insbesondere Bilddaten, z. B. bei der craniellen Bildgebung, sowie bei sehr individuellen Parametern, z. B. Personen, die mit einer seltenen Erkrankung in einer bestimmten Einrichtung untersucht wurden) bei eigentlich vorliegender Anonymisierung.

→ NUTZEN-RISIKO-ABWÄGUNG:
Entscheidende Zielgröße für ein Medizinprodukt ist vor der Inverkehrbringung immer eine klinische Bewertung der Erfüllung der spezifizierten Leistungsanforderungen, der Sicherheit und des Patientennutzens. Typische Bewertungskriterien, die in anderen Bereichen eingesetzt und nach wie vor in vielen medizinisch ausgerichteten KI-Publikationen herangezogen werden, können das in der Regel nicht leisten. Zum Beispiel führt bei einem diagnostischen Test die Minimierung übersehener Krankheiten (False Negatives mit z. T. schwerwiegenden Folgen) meist zu einer Erhöhung zu vieler fehldiagnostizierter Krankheiten (False Positives), die ebenfalls Schäden wie Verunsicherung der Patient*innen, unnötige Eingriffe usw. bewirken können. Es gilt, eine Balance zwischen diesen gegenläufigen Effekten zu finden, die jeweilige Wirkung der unterschiedlichen Fehlertypen eines KI-Verfahrens miteinzubeziehen und letztendlich eine Bewertung in Hinblick auf den klinischen Erfolg vorzunehmen sowie die Leistungsfähigkeit des Gesamtsystems zu optimieren. Im geplanten AI Act wird hingegen gefordert, dass zunächst vor allem die einzelnen Risiken reduziert werden müssen. Das Gesamtrisiko bzw. das Risiko-Nutzen-Verhältnis abbildende Ansätze sind im Entwurf des AI Act nicht in der Form vorhanden, wie sie speziell bei KI-basierten Systemen umgesetzt werden sollten.

→ EINGESCHRÄNKTE VERFÜGBARKEIT/MENGE UND HOHE KOMPLEXITÄT DER TRAININGSDATEN:
Qualitativ hochwertige Trainingsdatensätze sind entscheidend für die Leistungsfähigkeit eines KI-Systems bei ihrem Einsatz im vorhergesehenen klinischen Setting. Möglicherweise funktionieren KI-Systeme nicht einwandfrei,

wenn sie beispielsweise in unterschiedlichen Populationen oder in einem anderen Kontext (z. B. anderes Krankenhaus) eingesetzt werden und dort ggf. mit anderen Daten und Umständen konfrontiert sind als denen, mit denen sie trainiert wurden. Hinzu kommt, dass für manche Bereiche, beispielsweise in der Chirurgie, die Datenakquise z. B. über klinische Studien erschwert sein kann und nur wenige Fälle eingeschlossen werden können. Für bestimmte Arten von Behandlungen liegt somit nur eine sehr begrenzte Anzahl an hochwertigen Datensätzen vor, da diese aus dedizierten Studien in realen Anwendungsumgebungen stammen müssen. Zudem tragen häufig individuelle Faktoren und vielfältige Aspekte der Behandlungsumgebung zum Erfolg einer Behandlung bei. Bei Berücksichtigung derartiger Variationen ist darauf zu achten, dass in allen relevanten Bereichen (z. B. bezüglich Patientenpopulationen, Indikationen, aber auch unterschiedlicher Vorgehensweise der Ärzte und unterschiedlicher Krankenhausumgebungen) eine ausreichende statistische Zuverlässigkeit gewährleistet ist. Um gerade in einer stärker individuell ausgerichteten Behandlung bis hin zu einer personalisierten Medizin den Mehrwert von KI-basierten Verfahren zu realisieren, sind zusätzlich zu neuen Anforderungen an das Studiendesign jenseits des klassischen, statistischen Nachweises Wege zu finden, um Datensätze z. B. über die Generierung synthetischer Daten oder Methoden wie dem föderierten Lernen (KI-Modelle lernen aus dezentralisierten Trainingsdatensätzen, die Daten verbleiben z. B. im jeweiligen Krankenhaus) zur Verfügung zu stellen. Klare Vorgaben, wie vor allem deren Qualitätskontrolle umzusetzen ist, fehlen bisher.

→ FORMALISIERUNG VON PARAMETERN ZUR RISIKOQUANTIFIZIERUNG:
Bei KI-Systemen im Medizinbereich unterliegt die Formalisierung und Quantifizierung von Risikokriterien naturgemäß besonders hohen Ansprüchen. Dafür wiederum ist eine schlüssige Einordnung der Risiken erforderlich, die bei neuen Medizinprodukten oft nur schwer möglich ist, solange die Produkte noch nicht im regulären Betrieb eingesetzt wurden. Allerdings gibt es bei Medizinprodukten die Anforderung, dass genügend klinische Daten (d. h. Daten aus einer realen Anwendung) vorliegen bzw. über sogenannte klinische Prüfungen bereitgestellt werden müssen, bevor das Produkt auf den Markt gebracht werden kann. Deshalb erlauben bestehende Regularien (insbesondere DIN EN ISO 14971:2022 [351]) aus pragmatischen Gründen ein abgestuftes Vorgehen in Form einer semiquantitativen Abschätzung der Risiken.

→ UNTERSCHIEDLICHE AUTONOMIEGRADE UND ANFORDERUNGEN AN MENSCHLICHE AUFSICHT:
Es ist zu berücksichtigen, dass der Zweck von Medizinprodukten und das damit verbundene Risikopotenzial sehr unterschiedlich sein kann, je nachdem, wie hoch der Autonomiegrad eines KI-Systems ist – von einem reinen Unterstützungs- bis hin zu einem weitgehend autonomen System. Die meisten KI-basierten Systeme, die sich aktuell in der Entwicklung oder auch bereits im Einsatz befinden, sind im Bereich der Diagnostik bzw. Radiologie angesiedelt (z. B. Mammografie-Screening, Diagnostik von Augenerkrankungen oder Hautkrebs) [348]. Bei diagnostischen Anwendungen könnte z. B. ein menschlicher Betrachter immer als zusätzliche Kontrolle eingesetzt werden, bevor eine endgültige Entscheidung getroffen wird (Human-in-the-Loop). Bei anderen Systemen, z. B. einem Alarmsystem im Intensivbereich oder einem automatisiert funktionierenden Beatmungssystem, würde beim stärksten Autonomiegrad eine menschliche Kontrolle weitgehend entfallen und die KI als Closed-Loop-System fungieren. Derartige Aspekte müssten systematisch in die Risikobewertung einfließen. Der geplante AI Act enthält dazu auch die Anforderung, eine menschliche Aufsicht in die Produkte zu integrieren, die jederzeit in den Betrieb des Systems eingreifen kann. Er beschreibt jedoch nicht, was eine solche Aufsicht beinhalten kann oder muss. Zudem fehlen Vorgaben, welches Niveau an Erklärbarkeit KI-Systeme erreichen müssen, damit eine ausreichende Sicherheit gewährleistet werden kann.

4.7.2 Anforderungen und Herausforderungen

Grundanforderungen bezüglich der Umsetzung der Konformitätsbewertung bei KI-basierten Systemen – nach aktuellem Stand der Normung bzw. Gesetzgebung

Medizinprodukte, d. h. Instrumente, Geräte, Software o. Ä. mit einem dedizierten medizinischen Zweck, sind im Bereich der EU der MDR unterworfen und müssen damit vielfältige Anforderungen erfüllen. KI-basierte Anwendungen in Medizinprodukten fallen zumeist in die Kategorie Software und sind (gemäß MDR, Anhang VIII, Regel 11) bei geringerem potenziellem Schaden als IIa oder bei höherem potenziellem Schaden als IIb bis hin zur Risikoklasse III einzuordnen. In diesen Fällen sind die Produkte gemäß MDR einem Konformitätsbewertungsverfahren unter Einbezug einer Benannten Stelle zu unterziehen. Dadurch erfüllen sie auch das Kriterium im geplanten AI Act, das zu einer Einordnung in die Klasse der

Hochrisikoprodukte im Sinne des AI Act-Entwurfs führt (siehe dort Art. 6 bzw. Anhang II). Eine Vielzahl von KI-Anwendungen im Bereich der Medizin werden damit neben den bereits bestehenden Anforderungen der MDR in Zukunft zusätzliche Anforderungen aus dem geplanten AI Act erhalten. Herausforderungen, die sich in organisatorischer Hinsicht in Zukunft daraus ergeben könnten, sind in Kapitel 1.4 und speziell in Anhang 13.1 (Abschnitt „Exemplarische Darstellung am Beispiel Medizinprodukte“) dargestellt. Der Fokus dieses Kapitels liegt auf den Grundanforderungen, die ein KI-basiertes Medizinprodukt aufgrund der bestehenden Regularien (speziell in Bezug auf die MDR) zu erfüllen hat. Wie bereits geschildert, beinhalten weder die MDR noch die zugehörigen Normen spezielle Anforderungen für KI-basierte Systeme. Damit müssen sich die Herstellenden aktuell damit begnügen, inoffizielle Leitlinien wie den Fragebogen der IG-NB [358] heranzuziehen, um die Konformität eines KI-basierten Medizinprodukts nachzuweisen.

Damit die Leistungsfähigkeit und Sicherheit des Produkts für den gegebenen Anwendungszweck dargelegt werden kann, sind sowohl auf technischer als auch auf klinischer Seite entsprechende Anforderungen umzusetzen. Auf klinischer Seite schließt das einen Vergleich mit den bereits in Betrieb genommenen und erprobten Lösungen ein. Das beinhaltet neben dem technischen Vergleich der Leistungsfähigkeit eine positive Bewertung im Sinne des Nutzen-Risiko-Verhältnisses. Dazu ist zu definieren, in welchem Rahmen zusätzliche Risiken (im Verhältnis zu klassischen Methoden bzw. zum Stand der Technik) zulässig und akzeptabel sind und inwieweit diese Risiken durch einen entsprechenden klinischen Nutzen zumindest ausgeglichen werden können.

Mit zu berücksichtigen ist dabei die Interaktion zwischen den Anwender*innen und dem System, die sich gerade bei KI-basierten Systemen aufgrund von wechselseitigen Abhängigkeiten als komplex erweisen kann. Das gilt insbesondere, wenn sich auf der einen Seite der Benutzende verstärkt auf die Verlässlichkeit der Ergebnisse stützt und andererseits das System sein Verhalten an die jeweilige Anwendungsumgebung anpasst. Es können wesentliche Verschiebungen in der Risikobewertung entstehen, wenn sich die Benutzenden z. B. auf bestimmte, durch das KI-System ermittelte Diagnosen (siehe Use Case Bildgebung) oder Alarme (siehe Use Case Intensivmedizin) stützen. Selbst wenn die Ergebnisse eine bessere Genauigkeit erreichen, kann durch Sich-Verlassen auf die Ergebnisse des KI-Systems ein höheres Risikopotenzial gegeben sein. Daher ist z. B. zwischen technischen Performancekriterien (z. B. Erkennungsraten kritischer Situationen) auf der einen Seite und klinischen Parametern (z. B. Schaden für die Patient*innen durch fehlerhaft übersehene Diagnosen bzw. kritische Situationen) zu unterscheiden.

Auf der technischen Seite ist eine Reihe von Kriterien umzusetzen, die die grundlegende Sicherheit des KI-Systems gewährleisten. Dazu gehören die Ermittlung der Leistungsfähigkeit des Modells anhand technischer Kriterien, die Bereitstellung/Verfügbarkeit geeigneter Daten für das Training, die Erprobung (Validierung, Testung) des KI-basierten Modells, die Sicherstellung der Korrektheit und Robustheit (Fehlertoleranz) von Messgrößen, auf denen die KI-basierten Entscheidungen basieren, sowie Aspekte der Softwarearchitektur wie z. B. Einbindung von 3rd-Party-Komponenten für das Trainieren der Modelle (d. h. Komponenten von Fremdherstellern, die in das KI-basierte System eingebunden werden) und der Cybersecurity (z. B. spezielle Anforderungen an KI-Systeme in Hinblick auf den Schutz der Systeme vor Manipulationen). Die Offenlegung und Bewertung des klinischen Grundmodells, auf dem das KI-System aufbaut (welche individuellen Parameter haben welchen Einfluss auf die klinische Entscheidung) ist zudem im Sinne der Transparenz ein wesentlicher Faktor für die Akzeptanz und Umsetzung der Konformitätsbewertung bei solchen KI-Systemen.

Dabei ist zu beachten, wie die aufgestellten Kriterien der KI mit dem klinischen Nutzen zusammenspielen. Klassische Fehlerkriterien für das Trainieren von KI-Modellen wie „Accuracy“, Spezifität oder Sensitivität allein können noch keine unmittelbare Einordnung treffen, wie gut die Qualität der Modelle im klinischen Kontext ist. Zum Beispiel kann ein übersehener Alarm (False Negative) erheblich andere Auswirkungen haben als ein fehlerhaft ausgelöster und somit unnötiger Alarm (False Positive). Zudem ist entscheidend, wie zuverlässig das klinische Personal auf die Alarme bzw. Diagnosen reagiert und wie gut es diese einordnen und deren Ursachen verstehen kann.

Einzelanforderungen bezüglich der technischen und klinischen Bewertung des KI-basierten Systems

Modellbeschreibung und -selektion

Neben der Auswahl von Performancekriterien sind im medizinischen Kontext Bedingungen an das verwendete Modell zu knüpfen. So sind im Rahmen der Konformitätsbewertung

→ der gewählte Ansatz mit dem Stand in der Technik und Medizin sowie dem use-case-spezifischen etablierten Goldstandard zu vergleichen,

→ Robustheit, Fairness bzw. Reproduzierbarkeit nachzuweisen,
→ die verbleibende Unsicherheit der Vorhersage adäquat anzugeben
→ und die Transparenz des Vorhersageergebnisses zu fördern (Stichwort „Erklärbare KI“).

Der gewählte Ansatz des Medizinprodukts ist durch eine ausführliche Recherche zum Stand der Technik und zu etablierten Methoden im medizinischen Kontext zu belegen. Hierbei wird geraten, die unter Performancekriterien aufgeführten Metriken zum Vergleich der Modelle heranzuziehen.

Im Rahmen der verbleibenden Unsicherheitsuntersuchung soll gezielt nach Grenzen der Vorhersagbarkeit gesucht werden. Hier sind sowohl medizinische Randfälle zu recherchieren und mit dem vorliegenden Medizinprodukt zu testen als auch die Güte der Vorhersage bei Normalbefunden zu analysieren. Zu Nachweiszwecken ist eine statistische Untersuchung verschiedener Kriterien (Zufallsvariablen der Grundgesamtheit) und die Angabe von Konfidenzintervallen hilfreich. So lassen sich z. B. auch Robustheit und Fairness (im Sinne eines vorhandenen Bias) eines Vorhersageergebnisses quantitativ beschreiben. Die Robustheitsuntersuchung umfasst auch die Generalisierbarkeitsanalyse, d. h. die Anwendbarkeit der KI-Lösung auf Daten von Geräten unterschiedlicher Herstellender oder in unterschiedlichen Anwendungsumgebungen. Dies ist in Form einer Überprüfung auf Basis unabhängiger Daten durchzuführen, ggf. unter Nutzung einer Kreuzvalidierung (Cross Validation), falls nur eine geringe Menge an Daten vorliegt.

In der medizinischen Anwendung ist die Annahme eines KI-Medizinprodukts als Blackbox nur unter Erfüllung besonders hoher Anforderungen hinnehmbar. Eine einfache Nachvollziehbarkeit der Ergebnisse ist bei KI-Verfahren und insbesondere bei neuronalen Netzen oftmals eingeschränkt. Für konkrete Vorhersagen und Entscheidungsfindungen ist jedoch eine Begründung darzulegen, welche den Ansprüchen an ein Medizinprodukt genügt: Ein klinischer Anwender muss in die Lage versetzt werden, die Angemessenheit von Vorhersagen und Entscheidungen überprüfen und ggf. korrigieren zu können. Zur „Erklärbarkeit“ gehört ebenso das Visualisieren von Vorhersageentscheidungen mit ihren Verlässlichkeitswerten, basierend auf transparent dargestellten Grundregeln/Merkmalen der Bewertung, derer sich die KI bei solchen Entscheidungen bedient. Im Bereich Visualisierung während der Anwendung ist zusätzlich ein starker Fokus auf die Usability, eine eingängige Nutzerfreundlichkeit der grafischen Oberfläche, zu legen. Entsprechende Ansätze bedürfen aber ihrerseits einer – momentan nur in Ansätzen vorhandenen – Qualitätssicherung.

Performancekriterien

Um die Leistungsfähigkeit im Rahmen der technischen Validierung von KI-Modellen, die als Teilkomponenten eines KI-Systems eine bestimmte Vorhersage vornehmen, beurteilen zu können, wird ein unbekannter repräsentativer Datensatz, auch Testdatensatz genannt, verwendet, um die Modellvorhersagen mit den Annotationen (von menschlichen Expert*innen als „Goldstandard“ festgelegt) zu vergleichen. Dafür kommen in der Regel klassische Performancekriterien wie z. B. Accuracy, Spezifität/Sensitivität, Precision/Recall, F1-Score oder auch Receiver Operator Characteristic (ROC)-Kurven bzw. Area-under-the-Curve (AUC)-Werte zum Einsatz. Dabei können zudem die zugehörigen Detektionswahrscheinlichkeiten einfließen (Probability of Detection – POD – oder Probability of Classification – POC). Das betrifft z. B. die Frage, wie groß die Wahrscheinlichkeit ist, dass ein Tumor einer bestimmten Größe erkannt wird (POD) bzw. dass dieser Tumor korrekt als gut- oder bösartig klassifiziert wird (POC). Die hier genannten Kriterien beziehen sich hauptsächlich auf Aufgabenstellungen aus dem Bereich Klassifizierung und dem überwachten Lernen. Für andere Aufgabenstellungen sind entsprechend ausgerichtete Metriken zu verwenden. Das betrifft quantitative Abschätzungen aus dem Bereich der Regression, aber auch komplexere Szenarien mit dynamischen Aspekten, wie sie beispielsweise bei einer optimierten Therapieplanung zum Tragen kommen (siehe z. B. Use Case Intensivmedizin). Bei dynamischen und zeitkritischen Use Cases muss somit neben der Performance die Zeit mitberücksichtigt werden. Die Performancekriterien können z. T. durch Gewichtungen ergänzt werden, um spezifische Risikofaktoren zu berücksichtigen.

Datenmanagement

Die Bereitstellung der Datenbasis für das Trainieren und Testen des KI-Modells muss mehrere zentrale Qualitätskriterien erfüllen. Dazu gehört z. B. die bereits genannte Unabhängigkeit von Trainings-, Validierungs- und Testdaten. Jede einzelne dieser Gruppen muss zudem für den Anwendungszweck relevante Prüfdaten repräsentativ abdecken. Dies beinhaltet die Berücksichtigung unterschiedlicher Settings (z. B. unterschiedliches Equipment, unterschiedliche Qualifikation des Pflegeteams und Vorgehensweisen bei den Behandlungsprozessen, unterschiedliche Infrastruktur in Krankenhäusern, Arztpraxen) sowie eine entsprechende Bandbreite und Repräsentativität der Patientenpopulation (z. B. Alter,

Geschlecht, ethnische Zugehörigkeit). Diese Anforderung an die Trainings-, Validierungs- und Testdaten der KI gilt auch im Hinblick auf die ethischen Aspekte wie Fairness und Inklusivität gegenüber den unterschiedlichen Gruppen bzw. Nicht-Diskriminierung z. B. in Bezug auf Minderheiten. Hierbei besteht u. a. die Schwierigkeit, zur bestmöglichen individuellen Behandlung die unterschiedlichen Gruppen einerseits gleichwertig gut zu behandeln und andererseits das spezifische Optimierungspotenzial bei den einzelnen Gruppen möglichst umfänglich nutzen zu können. Um eine repräsentative Abbildung der durch die Zweckbestimmung definierten Patientengruppen zu ermöglichen, sollte ein gemeinsames Verständnis aufgebaut werden, welche demografischen Variablen für die konkrete Zweckbestimmung einen signifikanten Einfluss auf den klinischen Workflow haben und demzufolge in den Trainings-, Validierungs- und Testdaten entsprechend repräsentiert sein müssen.

Die Bereitstellung der Daten muss insgesamt auf einem sehr hohen Qualitätsniveau erfolgen, wobei bisher standardisierte Methodiken und Werkzeuge für die Bewertung von Datensammlungen fehlen. Beim überwachten Lernen muss z. B. sichergestellt werden, dass das Labeling der Daten durch entsprechend qualifiziertes Personal erfolgt. Dabei müssen u. U. mehrere Expert*innen unabhängig voneinander die Daten annotieren, um mögliche Verzerrungen (Bias) zu vermeiden. Zudem müssen klare Prozessvorgaben gemacht werden, um das Labeling auch bei der Ergänzung von Daten auf einem entsprechend hohen Niveau zu vollziehen.

KI-spezifische Fragen einer Risikoanalyse

Die DIN EN ISO 14971:2022 [351] fordert von Medizinprodukteherstellern einen Risikomanagementprozess, welcher sicherstellen soll, dass Risiken durch Medizinprodukte benannt, bewertet und beherrscht werden und dies immer im Verhältnis zum Nutzen akzeptabel ist. Dies gilt auch für Risiken, welche im Zusammenhang mit KI-gestützten Medizinprodukten entstehen. Allgemein sollte die Risikoanalyse ebenfalls folgende Punkte beachten:

→ Risiko-Nutzen-Abwägung zwischen dem Einsatz von KI oder der Verwendung von klassischen KI-freien Verfahren (hartkodierte Entscheidungsbäume).

→ Die Verstehbarkeit und klinische Bewertbarkeit des Vorhersageergebnisses durch das Vorhersagemodell (somit deren Ergebnisfindungsprozess) muss in geeignetem Maße sichergestellt werden. Dies gilt zum einen für die Überprüfung durch eine Benannte Stelle während des Konformitätsbewertungsverfahrens, aber auch im Falle einer Meldung im laufenden Betrieb einschließlich angemessener Eingriffsmöglichkeiten.

→ Es müssen Maßnahmen zur Sammlung von Log-Daten und Vitalitätsinformationen des KI-Systems eingeführt werden, die es ermöglichen, eine Bewertung der Funktionalität des KI-Modells zu erstellen und ggf. Fehlfunktionen identifizieren zu können.

→ Menge und Qualität der für das Training, die Validierung und Testung zur Verfügung stehenden Daten:
- Es ist zu prüfen, ob eine ausreichende Menge an für Training, Validierung und Testung des KI-Modells nutzbaren Daten vorhanden sind. Gegebenenfalls muss der Datenbestand durch synthetische Daten angereichert werden.
- Es ist der Nachweis zu führen und zu dokumentieren, dass die Daten möglichst frei von Bias sind. Falls nur eine begrenzte Kohorte Daten vorhanden ist, ist zu prüfen, ob dies Auswirkung auf den Intended Use des Produkts hat und dadurch ggf. die angezielte Patientengruppe angepasst/eingeschränkt werden muss.

→ Für zukünftige Umsetzung kontinuierlich- oder stufenweise lernender KI-Systeme:
- Die Risiko-Nutzen-Abwägung zum Einsatz eines offenen im Vergleich zu einem nicht kontinuierlich lernenden KI-Modell wird durchgeführt.
- Risiken, die für kontinuierlich lernende Systeme spezifisch sind, werden benannt und Maßnahmen zur Milderung werden umgesetzt.
- Das System kann auf einen bekannten Trainingszustand zurückgesetzt werden.

Klinische Bewertung

Die klinische Bewertung eines KI-basierten Medizinprodukts muss sicherstellen, dass das System beim Einsatz in einem komplexen klinischen Umfeld sicher, leistungsfähig und nützlich ist, keinen unvorhergesehenen Schaden anrichtet und dem professionellen Anwendenden stets ausreichende Eingriffsmöglichkeiten in die Auswahl- oder Entscheidungskriterien bietet. Dabei gilt es, die ethischen Aspekte ebenfalls miteinzubeziehen. Eine klinische Bewertung muss in allen Phasen entlang des Lebenszyklus des KI-basierten Medizinprodukts erfolgen. Zu Beginn der Entwicklungsphase sollte der Fokus auf einer Prüfung der Zweckbestimmung sowie des Standes der medizinischen Praxis im Anwendungsfeld des KI-basierten Medizinprodukts liegen. Hierzu gehört das Verständnis des medizinischen Problems, welches die KI-Anwendung zu lösen versucht, und ob sie hierfür geeignet ist. Zudem ist eine Beschreibung des beabsichtigten klinischen Nutzens gegenüber etablierten Methoden, der potenziellen

Risiken und Schäden, welche durch die KI verursacht werden könnten, und eine gute Dokumentation der Interoperabilität sowohl mit dem Nutzenden als auch beispielsweise dem IT-System inklusive einer Prüfung der „User Experience“ unter Einbeziehung von sicherheitsrelevanten Fragen entscheidend. Während der weiteren Entwicklung des dem Medizinprodukt unterliegenden KI-Modells ist es notwendig, die verwendeten Testdaten des Modells genau aufzuführen und die Modellperformance mit dem aktuellen Goldstandard zu vergleichen. Dabei ist zu beachten, dass für manche KI-Anwendungen ein entsprechender use-case-spezifischer Standard ggf. noch zu definieren ist.

Zur klinischen Bewertung zählt auch die Generierung klinischer Daten, mit denen die Leistungsfähigkeit, Sicherheit und der Patientennutzen bei bestimmungsgemäßem Gebrauch validiert werden. Die Leistung eines KI-Systems mag zwar unter Testbedingungen optimal sein, beim Einsatz im „echten Leben“ durch diverse menschliche und technische Einflussfaktoren jedoch dem beabsichtigten Nutzen nicht mehr entsprechen. Daher sollte die Generierung klinischer Daten als notwendiges Instrument zur Bewertung von KI-Technologien vor und nach ihrer Implementierung als wichtiger Faktor in die Entwicklung integriert werden. Das ist auch in der MDR verankert, die sehr deutlich die Verfügbarkeit klinischer Daten auf einem entsprechenden Niveau einfordert.

Bei der Durchführung klinischer Studien ist es wichtig, die Wirkung der KI-Intervention über den gesamten Behandlungspfad hinweg abzubilden.

Klinische Studien liefern die notwendige Evidenz für die Wirksamkeit und Sicherheit eines Medizinprodukts/KI-Systems. Es gibt verschiedene Studientypen, die sich in Umfang und Ablauf unterschieden und jeweils gewisse Vor- und Nachteile haben; zu den wichtigsten zählen beispielsweise randomisierte klinische Studien (randomisiert kontrollierte Studien (RCTs) oder Kohortenstudien) und retrospektive Fall-Kontroll-Studien. Die Auswahl eines passenden Studientyps richtet sich auch nach der expliziten Fragestellung. Anerkannte Regeln für die Planung und Durchführung klinischer Studien für Medizinprodukte (einschließlich der Erstellung von Prüfplänen etc.) finden sich beispielsweise in der Good Clinical Practice, DIN EN ISO 14155:2021 [360], MDR.

Insgesamt sollte aus dem Studienprotokoll hervorgehen, ob ein Ergebnis für einen spezifischen Endpunkt (klinisch oder für das System) robust und aussagefähig ist, und ein Studientyp (inklusive Studienprotokoll mit transparenter Berichterstattung) gewählt bzw. entwickelt werden, der durch die Minimierung von Verzerrungen (Bias) die nötige Evidenz für ein KI-System liefert und Vertrauen in die Ergebnisse schafft. Dies kann letztlich auch Entscheidungsträgern und den Nutzer*innen Sicherheit bieten.

Durchaus lassen sich die Grundprinzipien von guten klinischen Studienprotokollen unter Berücksichtigung spezifischer Anforderungen an ihre Bewertung in gleicher Weise auf KI-Systeme übertragen. Bisher entspricht das Niveau des Studiendesigns und der Berichterstattung von veröffentlichten KI-Studien hierbei häufig nicht den hohen Anforderungen (siehe z. B. [361]). Aus diesem Grund werden von verschiedenen Initiativen, wie beispielsweise auf internationaler Ebene von dem interdisziplinären „EQUATOR-Netzwerk“ [EQUAUTOR Network.org], Leitfäden zur Verbesserung der spezifischen Studiendesigns bei der Evaluierung von KI-Systemen entwickelt, für die Berichterstattung von Studienprotokollen z. B. „SPIRIT-AI“ [362] bzw. von Studienberichten „CONSORT-AI“ [363]. Bei der Planung bzw. Umsetzung einer klinischen Studie ist darüber hinaus ein Ethikvotum miteinzubeziehen. Die Studienergebnisse sind abschließend von unabhängigen Expert*innen zu begutachten.

Im weiteren Verlauf der Bewertung des KI-basierten Medizinprodukts ist während der Implementierungsphase des KI-Modells eine fortlaufende Prüfung der Leistungsfähigkeit und Sicherheit notwendig, um ggf. unerwartete Effekte, welche erst mit Einsatz in einem komplexen klinischen Umfeld auftreten könnten, zu registrieren und zu beheben. Dies schließt auch Versions-Updates der KI ein. Eine Möglichkeit, dies anzugehen, sind sogenannte AI-Audit-Verfahren, mit denen unerwartete Effekte aufgedeckt und genau analysiert werden können, wie beispielsweise in [364] beschrieben.

Benutzerinteraktion

Beim Einsatz von KI-Modellen in der Praxis ist die Interaktion zwischen den Benutzer*innen und dem System unabdingbar, da das medizinische Fachpersonal im Laufe der Zeit ihr eigenes Verhalten wahrscheinlich auf die automatisierte Unterstützung anpassen wird, z. B. indem es sich mehr und mehr auf das System verlässt. Das kann insbesondere dann zu Schwierigkeiten führen, wenn das Personal die Entscheidungen des Systems nicht ausreichend verstehen kann oder sich das System aufgrund häufiger Neu-Releases oder auch im Fall eines kontinuierlich lernenden Systems mit ständigen Anpassungen verändert. Die Benutzer*innen können sich dann evtl. nicht mehr ausreichend auf das neue Systemverhalten einstellen. Zudem ist die Frage zu beantworten, welche

und wie viele Informationen die Benutzer*innen benötigen, um die Entscheidungen des Systems nachvollziehen und richtig einordnen zu können. Eine Antwort darauf muss zusätzlich berücksichtigen, dass die Benutzer*innen in Bezug auf ihren Kenntnisstand, ihre persönlichen Einstellungen oder auch in Bezug auf die mit ihnen verbundene Krankenhausumgebung oft sehr heterogen sind.

Da derartige Effekte oft erst im realen Betrieb vollständig erfasst werden können, ist eine schlüssige Überwachung der Systeme auch in Verbindung mit einer systematischen Erfassung von Fehlerfällen, aber auch positiven Resultaten im Sinne einer Post Market Surveillance ein wichtiger Faktor. Das Erfordernis derartiger Schritte ist sowohl in der MDR als auch im geplanten AI Act verankert. Erfahrungen mit dem System sind gezielt zusammenzutragen und zu bewerten. Der geplante AI Act verlangt dabei, dass eine menschliche Aufsicht die Kontrolle über das System behält und das System ggf. rechtzeitig abschalten (bzw. in einen klassischen Modus umschalten) kann. Wie bereits angesprochen, ist es dabei wichtig, dass die Benutzer*innen ein ausreichendes Verständnis des Systems und von dessen Entscheidungsgrundlagen sowie eine ausreichende Kenntnis der darauf basierenden Entscheidungen erreichen können.

4.7.2.1 Anwendungsbeispiel: KI-assistierte 2-D-Röntgenbildanalyse zur Kariesdiagnostik in der Zahnmedizin

Auch in der Zahnheilkunde werden vermehrt KI-Softwareapplikationen in die Praxis eingeführt. Fokus der aktuellen Bemühungen sind vor allen Dingen das Maschinelle Sehen (Computer Vision), vor allem im Bereich der zahnärztlichen Röntgenbildanalyse (Diagnoseunterstützung), insbesondere im 2-D-Röntgenbereich, z. B. Analyse von Einzelbildern, Panoramaschichtbildern, Bissflügelaufnahmen und Fernröntgenseitenbildern. Dies liegt darin begründet, dass in der Zahnheilkunde eine große Zahl von Röntgenbildern angefertigt werden (in Deutschland mehr als 50 Millionen Bilder und weltweit ca. 520 Millionen pro Jahr), die Genauigkeit von Zahnärzt*innen bei der Diagnostik auf diesen Bildern begrenzt ist (beispielsweise liegt die Sensitivität für die Detektion früher Karies auf Röntgenbildern bei < 50 %) und eine systematische und umfängliche Befundung und Dokumentation der Diagnoseergebnisse aufwendig ist.

Die Analyse von 2-D-Röntgenbildern in der Zahnmedizin mittels KI kann helfen, die diagnostische Genauigkeit, Zuverlässigkeit, Effizienz und Kommunikation von Befunden zu verbessern. KI-gestützte Medizinprodukte müssen hierbei ebenfalls einem eingehenden Prüfungsprozess unterzogen werden, um ihre Sicherheit, Robustheit, Transparenz, Fairness, Inklusivität und (Kosten-)Effizienz zu gewährleisten. Das im Folgenden beschriebene Anwendungsbeispiel (siehe auch Anhang 13.5) einer KI-Komponente in der dentalen 2-D-Röntgenbildanalyse liegt im Kontext der digitalen intraoralen Röntgenbildgebung, konkret der Bissflügelaufnahmen. Hierbei wird mittels eines gerichteten Röntgenstrahlers der Seitenzahnbereich eines*r Patient*in von außen durchstrahlt und das Signal durch einen in der Mundhöhle des*r Patient*in platzierten digitalen Röntgensensor aufgenommen. Diese KI-Komponente wird als Backend Service ausgelegt: Die erstellte Bissflügelaufnahme sowie Metainformationen (z. B. Pixelgröße, Strahlendosis) werden an diesen Backend Service übermittelt, welcher auf einem Cloudserver läuft. Die Ausführung dieses Backend Service wird somit durch das System selbst und ohne menschliche Interaktion veranlasst. Die empfangenen Daten werden sodann durch ein vorab entsprechend trainiertes neuronales Netz automatisch analysiert und etwaige Detektionen von Karies in Form von Polygonzügen im Koordinatensystem der digitalen Bissflügelaufnahme ausgegeben. Beides (das nicht modifizierte digitale Röntgenbild und die Punkte entlang der Polygonzüge etwaiger Kariesdetektionen) werden über ein sicheres Netzwerkprotokoll an eine Workstation mit Software und entsprechender Benutzeroberfläche zur Befundung bereitgestellt.

Goldstandard

Das üblichste Verfahren zur KI-gestützten 2-D-Röntgenbildanalyse in der Zahnmedizin ist das überwachte ML (vgl. z. B. [365]); erste Ansätze haben allerdings auch nicht überwachtes (unsupervised) Lernen eingesetzt [366]. Im Rahmen des überwachten Lernens muss mittels des Annotationsprozesses ein Goldstandard (Referenztest) etabliert werden.

Für die 2-D-Röntgenbildanalyse in der Zahnmedizin existiert kein weithin akzeptierter Goldstandard; je nach Anwendungsfokus (Kariesdetektion, Detektion apikaler Läsionen, Vermessung des parodontalen Knochenabbaus) kommen verschiedene Referenztestverfahren zum Einsatz (u. a. alternative Bildgebungen mit hoher Sensitivität, z. B. 3-D-Röntgenaufnahmen wie digitale Volumentomogramme oder histologische Evaluationen, z. B. an extrahierten Zähnen, die zuvor röntgenologisch analysiert wurden). Für den Bereich von KI-Anwendungen ist die visuelle Begutachtung der Röntgenaufnahmen durch Zahnärzt*innen verbreitet, wobei in der Etablierung eines Goldstandards üblicherweise

mehrere Expert*innen einbezogen werden, um der Ungenauigkeit der einzelnen Befunde zu begegnen und die Validität des Goldstandards zu erhöhen. Wie genau aus den verschiedenen Befunden dann der Goldstandard konstruiert wird, ist ebenfalls nicht abschließend definiert (siehe z. B. [367], [368]); für Klassifikationsaufgaben kommen Mehrheitsvoten oder Konsensuspanels zum Einsatz, z. B. [369]; für Segmentationsaufgaben wurden u. a. hierarchische Verfahren (drei bis fünf unabhängige Expert*innen segmentieren, ein „Masterreviewer“ überarbeitet die Gesamtheit der Segmentationen) eingesetzt [366].

Modellbeschreibung und -selektion

Zur zahnmedizinischen Analyse von 2-D-Röntgenbildern werden vor allem Convolutional Neural Networks (CNN) eingesetzt, wobei je nach Aufgabe (Klassifikation, Detektion, Segmentierung) unterschiedliche Modellarchitekturen sowie unterschiedlich annotierte Daten (siehe hierzu Textpassage „Goldstandard“) zum Einsatz kommen (vgl. z. B. [365] oder [370]). Die eingesetzten Modellarchitekturen orientieren sich am generellen State-of-the-Art; für die Zahnmedizin werden keine speziellen Architekturen eingesetzt.

Für den Trainingsprozess werden die Daten wie üblich in unabhängige Trainings-, Validierungs- und Testdatensätze getrennt; hierbei ist relevant, die Kontamination der Datensätze (data snooping) zu vermeiden. Dies ist insbesondere in der Zahnmedizin zu beachten, da oft von ein und demselben Patienten mehrere 2-D-Röntgenbilder vorliegen. Das Splitten der Datensätze in Trainings-, Test- und Validierungsdatensatz sollte demnach auf Patientenebene, nicht Bildebene erfolgen. Auch weitere Parameter müssen im Rahmen der Partitionierung der Daten berücksichtigt werden.

Die Auswahl der passenden Modellarchitektur erfolgt oft empirisch; systematische Untersuchungen zur optimalen Modellwahl existieren in der Zahnmedizin kaum (vgl. [370]). Ebenso wird die Festlegung der sogenannten Hyperparameter während der Modellvalidierung (Hyperparameter Tuning) zurzeit vor allem empirisch durchgeführt (vgl. [372]). Die Möglichkeit des Vortrainierens der Modelle auf röntgenologischen Datensätze (z. B. frei verfügbare Datensätze mit Lungenröntgenbildern) wurde bereits demonstriert [366] und stellt eine Alternative gegenüber dem üblichen Vortrainieren auf allgemeinen, nicht-dentalen Datensätzen dar.

Für den Use Case der Karieserkennung in Bissflügelaufnahmen kommen ebenfalls CNNs zum Einsatz. Diese sind in der Lage, eine pixelbasierte Klassifikation zu liefern, sogenannte „semantic segmentation“. Hierzu werden zur Generierung von Trainings-, Validierungs- und Testdaten die innerhalb von markierten Polygonzügen befindlichen Pixel mit einem numerischen Label versehen, Fehlermetriken können somit pixelbasiert berechnet werden. Beim praktischen Einsatz solcher trainierter CNNs werden die pixelbasierten Ausgaben durch ein geeignetes Post-Processing in Polygonzüge umgerechnet.

Performancekriterien

Zur Erfassung der Performance der Modelle werden am Testdatensatz Metriken wie Genauigkeit, Sensitivität, Spezifität, F-1 sowie die Fläche unter der Receiver-Operating-Characteristics-Kurve bestimmt. Neben der metrischen Charakterisierung sollten weitere Qualitätskriterien wie die Robustheit des Modells, Fairness, „Erklärbarkeit“ und die Fähigkeit, die Ungenauigkeit der Vorhersage adäquat zu beschreiben, berücksichtigt werden.

Datenmanagement

Die Datenerhebung für den Trainingsprozess muss so gestaltet werden, dass die Daten von den im Feld gemäß Zweckbestimmung anzutreffenden Röntgengeräten stammen sowie die Population für das Einsatzgebiet repräsentieren. Die für das überwachte Lernen notwendigen Annotationen müssen durch qualifiziertes Personal vorgenommen und deren Qualität durch einen Reviewprozess sichergestellt werden (siehe hierzu Textpassage „Goldstandard“). Die Trainingsdaten werden über die Auslieferung des Produkts hinaus zur Dokumentation und für spätere Wiederholungen des Trainingsvorgangs archiviert. Üblicherweise werden dem KI-Modell zur Beurteilung der Leistungsfähigkeit im Rahmen der technischen Validierung Testdaten zugeführt und die Modellvorhersagen dann mit Labels bzw. Annotationen (von menschlichen Expert*innen als „Goldstandard“ s. o. festgelegt) verglichen.

KI-spezifische Fragen einer Risikoanalyse

Die Risikoanalyse richtet sich an den allgemeinen Prinzipien zur Risikoerfassung (Häufigkeit, Schweregrad) und der Ableitung von Mitigationsstrategien ab. Das in vielen Fällen schwerwiegendste nicht erwünschte Ereignis ist der Zahnverlust; nur für bestimmte (z. B. chirurgische) KI-Anwendungen auf 2-D-Röntgenbildern sind weitergehende Schäden zu erwarten.

Klinische Bewertung

In der klinischen Bewertung muss der Nutzen des KI-Systems evaluiert werden, um zu sehen, ob die KI im Zusammenspiel mit den Nutzer*innen im Einsatzumfeld, z. B. im Krankenhaus oder in einer zahnmedizinischen Praxis, wie beabsichtigt

funktioniert und das medizinische Personal, Patient*innen etc. von der Methode profitieren, sie also einen Mehrwert bringt. Fragen, die hierfür relevant sind, sind beispielsweise:

→ Ist die KI-Methode in den teils hochkomplexen realen Anwendungssituationen in der zahnmedizinischen Praxis sicher und wirksam?
→ Bringt die KI-Methode einen messbaren Nutzen über etablierte Methoden hinaus?

Optimalerweise sollten die klinischen Anforderungen bereits von Beginn an mitbedacht werden, d. h. Zahnärzt*innen bereits am Entwicklungsprozess beteiligt sein, und ein ständiger und fortlaufender Austausch zwischen Entwicklern, Klinikern, allen beteiligten Stakeholdern erfolgen. Die klinische Validierung beinhaltet idealerweise die Durchführung einer randomisiert kontrollierten Studie oder ein ähnliches Studiendesign. Hierbei sollten zudem Aspekte wie Akzeptanz, Implementierbarkeit und Aufrechterhaltung, aber auch der Einfluss auf den diagnostischen und therapeutischen Prozess (Therapieentscheid) und die sich ergebende Kostenwirksamkeit der KI berücksichtigt werden.

Benutzerinteraktion

Innerhalb der KI-Software haben behandelnde Zahnärzt*innen üblicherweise die Möglichkeit, die Aufnahme sowie die Ergebnisse der KI-Analyse darzustellen. Anwender*innen des Systems verwenden die Ergebnisse der KI folglich als zusätzliche Informationsquelle während der Befundung (Assistenz). Die durch die Software dargestellten Detektionen können üblicherweise durch Anwender*innen gelöscht, bearbeitet, oder neue Detektionen hinzugefügt und zu Dokumentationszwecken neben der Aufnahme in der digitalen Patientenakte gespeichert werden. Wie im Anwendungsbeispiel beschrieben, dient die KI-Komponente der assistierten Kariesdiagnostik. Der behandelnde Zahnarzt bzw. die behandelnde Zahnärztin kann die Ergebnisse der KI-Komponente ausblenden, modifizieren, falsch positive Detektionen löschen oder von der KI-Komponente übersehene (falsch negative) Kariesläsionen manuell nachtragen. Somit unterliegen alle Ergebnisse der menschlichen Aufsicht.

Fazit

Die Leistungsfähigkeit, Sicherheit und Effizienz KI-basierter Anwendungen zur Analyse von 2-D-Röntgenbildern in der Zahnmedizin muss u. a. durch Normungsprozesse gewährleistet werden. Bei der Umsetzung von Normungsaktivitäten sollten die spezifischen Herausforderungen in der Zahnmedizin (u. a. Vorhandensein oft mehrerer Bilder desselben Patienten bzw. derselben Patientin in einem Datensatz; Clustering von Pathologien und statistischen Einheiten auf Patienten- und Zahnebene: Patienten und Zähne weisen teilweise mehrere Pathologien auf, deren Auftreten oftmals nicht unabhängig voneinander ist; unerwünschte Gesundheitseffekte oftmals auf den einzelnen Zahn begrenzt) berücksichtigt werden, wird aber in weiten Teilen analog zu anderen Gesundheitsfeldern erfolgen.

4.7.2.2 Anwendungsbeispiel: KI-basiertes Beatmungssystem in der Intensivmedizin

Während bei vielen aktuellen Entwicklungen von KI-basierten medizinischen Anwendungen (z. B. radiologische Untersuchungen, bildbasierte Detektion von Augen- oder Hauterkrankungen) der diagnostische Aspekt im Vordergrund steht, kommen bei intensivmedizinischen Anwendungen verstärkt Monitoring-Aufgaben bzw. sogar Aspekte der Steuerung des Therapieablaufs hinzu. Wie im Rahmen des folgenden Anwendungsbeispiels beschrieben, müssen hier spezielle Anforderungen beachtet werden, um die Sicherheit und Effektivität des Systems nachweisen zu können. Das Anwendungsbeispiel beinhaltet dabei verschiedene Abstufungen, wie z. B. bezüglich des Autonomiegrads (unterschiedliche Stufen der Automatisierung, des Umfangs der menschlichen Aufsicht bzw. der Interaktion zwischen Mensch und System, vgl. z. B. Stufen beim autonomen Fahren, siehe auch Kapitel 4.7.3, Handlungsbedarf 07-05 und [373]) und des Zeitpunkts des Maschinellen Lernens (einmaliger Lernvorgang vs. kontinuierliches Lernen unter Einbezug neuer Umgebungsdaten, siehe [374]). Im Kontext der Therapieunterstützung ist dabei zu beachten, dass nicht nur reine Klassifikations- oder Regressionsaufgaben durch die ML-Komponente umgesetzt werden. Es handelt sich vielmehr um ein dynamisch wirkendes System, das im Rahmen des Monitorings bzw. der Therapiesteuerung immer wieder Messungen an den Patient*innen durchführen muss, um Vorhersagen (Prädiktion) durchführen zu können bzw. ihr eigenes Verhalten anzupassen. Das System wechselwirkt dabei verstärkt mit den Patient*innen, indem es sie im Therapieverlauf unterstützt, ohne dass zwischenzeitlich eine behandelnde Person eingreift. Insofern handelt es sich um einen Closed-Loop-Ansatz (siehe [375]), wobei der Grad der Autonomie auch eingeschränkt sein kann, insbesondere wenn der Arzt / die Ärztin an bestimmten Stellen eingreifen muss. Die Beschreibung des Anwendungsbeispiels fokussiert sich auf erweiterte Anforderungen, die über die bisher beschriebenen Anwendungsbeispiele hinausgehen.

Konkretes Anwendungsbeispiel: Beatmungsgerät mit KI-gestützter Entwöhnung

Beatmungsgeräte dienen in erster Linie dazu, die Patient*innen bei einer Störung der Lungenfunktion mit ausreichend Sauerstoff zu versorgen und sie damit bei der Atmung zu unterstützen. Das gilt insbesondere für kritische Zustände, wie sie beispielsweise nach Unfällen oder bei einer Coviderkrankung mit schwerem Verlauf gegeben sein können. Die Behandlung umfasst dabei eine Reihe einzelner Schritte – von der Entscheidung, die Beatmung durchzuführen, über die Intubation und das Screening bis hin zur Entwöhnung und der abschließenden Extubation. Im Rahmen des hier vorliegenden Anwendungsbeispiels liegt der Fokus auf einer ML-basierten Entwöhnung, bei der die Beatmungsunterstützung schrittweise reduziert wird, um sie letztendlich ganz ausschalten zu können, wenn vorhersehbar ist, dass die Patient*innen mit der eigenen Atmung vollständig und dauerhaft auskommen können. Die zentralen Therapieentscheidungen zur Einleitung der Entwöhnungsphase oder auch zur Extubation muss von ärztlicher Seite (ausgebildeter Facharzt bzw. Fachärztin für Anästhesiologie/Intensivmedizin) vorgenommen werden (siehe auch Anhang 13.5).

Die Entwöhnung selbst soll durch ein ML-basiertes Verfahren gesteuert werden, indem es die Beatmungsparameter auf Basis einer kontinuierlichen Messung zentraler physiologischer Parameter und unter Verwendung trainierter neuronaler Netze dynamisch anpasst und die Patient*innen so in einem stabilen Zustand hält. Die Entwöhnung erfolgt dabei durch eine schrittweise Reduktion der Atmungsunterstützung unter Bewertung der jeweils aktuellen Situation. Dabei ist zu beachten, dass die Patient*innen während der Entwöhnung in unterschiedliche pathologische Zustände gelangen können (wie z. B. Hypo- oder Hyperventilation, Tachypnoe, …), in denen das System passend reagieren muss. Begleitend dazu sind Alarme auszulösen, durch die das Intensivpersonal über kritische Zustände informiert wird und erforderliche Maßnahmen auslösen kann, die nicht der Entscheidung der KI unterliegen.

Das System übernimmt damit sowohl Monitoring-Funktionen in Verbindung mit einer Alarmkomponente für kritische Zustände als auch eine Therapieunterstützung im Sinne eines Closed-Loop-Systems. Anhand einer entsprechenden Datenbasis aus realen Fällen sollen neuronale Netze so trainiert werden, dass sie einerseits Alarmsituationen erkennen und Alarme auslösen sowie andererseits erforderliche Änderungen der Beatmungsparameter automatisiert umsetzen können. Es wird zunächst von einer fixen Datenbasis und einem für die Konformitätsbewertung fixierten Stand des Modells ausgegangen. Erweiterungen in Richtung kontinuierlich lernende Systeme, indem das neuronale Netz während des Betriebs an Umgebungsparameter (wie z. B. die spezielle Krankenhausumgebung) oder an individuelle Parameter der Patient*innen angepasst wird, werden als Erweiterungsmöglichkeit betrachtet.

Auf Basis klassischer Logik bzw. physiologischer Modelle gibt es bereits Systeme auf dem Markt, die eine solche automatisierte Entwöhnung realisieren und ähnliche Funktionen übernehmen (Identifikation des aktuellen Zustands der Patient*innen sowie die erforderlichen Unterstützungsmaßnahmen im Sinne eines Closed-Loop-Systems mit einer zusätzlichen Alarmkomponente). Die Entscheidungen bauen hier jedoch auf festen Kriterien bezüglich der zentralen physiologischen Parameter auf, wie z. B. Spontanatemfrequenz, Tidalvolumen und endtidales CO_2. Die Systeme müssen durch entsprechend ausgebildete Intensivpfleger*innen bedient werden, wobei erneut für bestimmte Schritte (z. B. Entscheidung zum Start/Ende der Entwöhnung) ein Arzt bzw. eine Ärztin herangezogen werden muss. Diese bestehenden Systeme sind als Stand der Technik / Standard-of-Care bei der Entwicklung eines ML-basierten Systems zu betrachten.

Eine detaillierte Beschreibung des Anwendungsbeispiels ist in Tabelle 19 im Anhang 13.5 zu finden. Im Folgenden sind spezielle Anforderungen in Bezug auf das beschriebene Anwendungsbeispiel gelistet, die die in Kapitel 4.7.2 gelisteten Grundanforderungen ergänzen.

Spezielle Aspekte der Konformitätsbewertung im beschriebenen Anwendungsbeispiel „Intensivmedizin“

Auf der technischen Seite sind im vorliegenden Fall eine Reihe von Kriterien umzusetzen, die die grundlegende Sicherheit des Systems gewährleisten. Wir beschränken uns dabei auf Aspekte, die speziell mit den ML-basierten Komponenten des Beatmungsgeräts verbunden sind, wie z. B. der Vermessung der Leistungsfähigkeit des Modells, der Bereitstellung/Verfügbarkeit geeigneter Daten für das Training und der Erprobung (Validierung, Testung) des ML-basierten Modells im vorliegenden Fall. Auch die Sicherstellung der Korrektheit und Robustheit (Fehlertoleranz) von Messgrößen, auf denen die ML-basierten Entscheidungen basieren, z. B. physiologische Messgrößen bezüglich des Atmungszustands der Patient*innen, ist im vorliegenden Fall ein wichtiges Kriterium, um eine verlässliche Einordnung der Systemperformance geben zu können. Die Offenlegung und Bewertung des klinischen Grundmodells, auf dem das ML-System aufbaut (welche phy-

siologischen Parameter erfordern welche klinische Entscheidung im Sinne von „passender“ Beatmungsunterstützung), ist zudem im Sinne der Transparenz ein wesentlicher Faktor für die Konformitätsbewertung bei solchen ML-Systemen. Ein weiterer wichtiger Punkt ist mit zunehmendem Autonomiegrad intensivmedizinischer Geräte die jederzeitige Möglichkeit für den Anwendenden, Entscheidungen des ML-Systems abzuändern, um bessere oder spezifischere Therapieergebnisse zu erzielen oder um im Fehlerfall die Kontrolle an sich zu ziehen.

Bei den Bewertungskriterien für das ML-System ist zu beachten, wie sie mit den tatsächlichen klinischen Effekten zusammenspielen, d. h. welche Kriterien einen optimalen klinischen Outcome charakterisieren. Klassische Fehlerkriterien für das Trainieren von ML-Modellen wie Accuracy, Spezifität oder Sensitivität allein können noch keine unmittelbare Einordnung treffen, wie gut die Qualität der Modelle im klinischen Kontext ist. Zum Beispiel kann ein übersehener Alarm (falsch negativ) erheblich andere Auswirkungen haben als ein fehlerhaft ausgelöster und somit unnötiger Alarm (falsch positiv). Die aus den einzelnen Fehlerarten resultierenden Wirkungen sollten gezielt in die Fehlerkriterien integriert werden, um eine systematische Minimierung des Risikopotenzials zu erreichen.

Bei der Steuerung der Beatmung sind zudem die genannten klassischen Performancekriterien nicht anwendbar, da es sich um die Optimierung eines dynamischen Prozesses handelt, bei dem die Leistungsfähigkeit in anderer Weise gemessen werden muss. Dieser Effekt wird verstärkt, wenn ML-Modelle mit klassischen physiologischen Modellen kombiniert werden, indem z. B. Randbedingungen wie die Reaktion auf bekannte kritische Werte bezüglich Spontanatemfrequenz oder Tidalvolumen in klassischer Weise fest einkodiert und integriert werden, um bestimmte Risiken bei der Erkennung von pathologischen Zuständen zu vermeiden. Bei solchen hybriden Modellen gibt es aktuell keine klaren Vorgaben, wie bestehende klassische und lange etablierte Ansätze mit Reaktionen eines neuartigen ML-basierten Systems in Bezug auf ihre klinische Leistungsfähigkeit zu vergleichen sind.

Weiterhin ist zu beachten, wie zuverlässig das Intensivpersonal auf die Alarme reagiert und wie gut es die entstandenen Alarme einordnen und deren Ursachen verstehen kann. Gerade im Bereich der Intensivmedizin ist eine zuverlässige Umsetzung und gezielte Erprobung der Mensch-Maschine-Interaktion in Hinblick auf deren klinische Wirksamkeit ein wichtiger Schritt, um zuverlässig Entscheidungen treffen zu können, die in lebenskritische Bereiche hineinreichen. Dem Benutzenden ist dabei gezielt zu vermitteln, wie zuverlässig die Ergebnisse des KI-Systems einzuordnen sind (z. B. über Verlässlichkeits-Scores für die Alarme). Dabei ist zu berücksichtigen, dass die Benutzer*innen oftmals entweder zu stark oder zu schwach auf das System vertrauen und somit die Wirksamkeit zusätzlich verändert ist. Letztendlich kann das System nur im Echtzeitbetrieb umfassend erprobt werden, wobei es substanzielle Unterschiede in der Wahrnehmung bei verschiedenen Benutzergruppen und/oder -umgebungen geben kann.

Die ML-basierte Bestimmung optimaler Entscheidungen hängt dabei von vielen individuellen Parametern (Patient*innen unterschiedlichen Geschlechts, Alters oder ethnischer Herkunft können unterschiedliche Atemmuster aufweisen) sowie von der Komplexität und den Möglichkeiten der jeweiligen Umgebung (unterschiedliches Equipment, unterschiedliche Qualifikation des Pflegeteams, unterschiedliche Vorgehensweisen bei den Behandlungsprozessen) ab. Die Trainings-, Validierungs- und Testdaten müssen diese Szenarien möglichst umfassend und repräsentativ abdecken. Das ML-Modell selbst muss dann möglichst zuverlässig erfassen, welche Aktion für welche Patient*innen in welcher Umgebung die beste klinische Wirkung erreicht.

Das erfordert auf der einen Seite eine bessere Anpassung an die jeweilige Umgebung und Patientenpopulation. Auf der anderen Seite wäre zusätzlich ein kontinuierlicher Lernvorgang erforderlich, um das ML-System in geeigneter Weise anpassen zu können. Ein hoher Grad an Flexibilität in den Modellen kann sich aber wiederum auf die Sicherheit des Systems insbesondere in Verbindung mit der Benutzerinteraktion auswirken. Gerade wenn die Modelle häufig geändert werden, können sich Benutzer*innen potenziell nicht schnell genug an die neuen Gegebenheiten adaptieren. Es sollten daher Kontrollmechanismen in das System integriert werden, die derartige Model Drifts nicht nur auf rein technischer Ebene, sondern auch bezüglich ihrer klinischen Wirkung überprüfen. Einerseits wird das im geplanten AI Act im Bereich der menschlichen Aufsicht auch gefordert. Andererseits gibt es gerade für das sehr sicherheitskritische Umfeld im Bereich der Medizin bzw. Intensivmedizin dafür noch keine Vorgaben, wie das in entsprechenden Prozessen umzusetzen ist. In jedem Fall ist die Verfügbarkeit geeigneter Daten aus dem Echtzeitbetrieb ein wichtiger Faktor, um eine hohe Leistungsfähigkeit zu erreichen. Dafür ist eine umfassende Verwendbarkeit von Echtzeitdaten und eine inkrementelle Integration der Daten in das Modell sowie die dafür erforderliche Erprobung ein zentraler Aspekt für die Entwicklung KI-basierter Anwendungen im Bereich Intensivmedizin.

Fazit
Insgesamt betrachtet handelt es sich bei dem vorliegenden Anwendungsbeispiel um ein recht komplexes und bezüglich seiner Risikoaspekte kritisches Szenario, das in bestehenden Normen (z. B. DIN EN ISO 14971:2022 [351] bezüglich Risikomanagement, DIN EN 62304:2016 [353] bezüglich Softwarelebenszyklus, DIN EN 623661:2021 [355] bezüglich Gebrauchstauglichkeit/Mensch-Maschine-Interaktion) noch nicht ausreichend abgedeckt ist (siehe auch [356]). Es werden in dem Szenario einige Aspekte aufgezeigt, die in rein diagnostisch ausgerichteten Anwendungsbeispielen eine untergeordnete Rolle spielen, die aber in Zukunft adressiert werden sollten, um umfassende Vorgaben für die Umsetzung KI-basierter Systeme in den verschiedenen Anwendungsszenarien zu bekommen.

4.7.2.3 Anwendungsbeispiel: Segmentierung und Klassifikation von Gehirnarealen (inklusive Liquor) und deren Volumenbestimmung

Stand der Technik
Die Magnetresonanztomografie (MRT) hat sich als Standardverfahren in der neuroradiologischen Diagnostik etabliert. Insbesondere können hiermit verschiedene Gewebestrukturen untereinander sowie krankhafte Veränderungen von normalem Gewebe gut dargestellt werden. Neben der visuellen Auswertung von 3D-MRT-Daten ist für die Diagnose häufig eine quantitative Vermessung anatomischer Strukturen sowie ggf. deren zeitliche Veränderung, z. B. zur Kontrolle eines Therapieverlaufes, notwendig. Für eine konventionelle Volumenbestimmung müssen hierzu ausgewählte anatomische Strukturen manuell oder halbautomatisch (z. B. durch Kontrast- oder Kantenerkennung) segmentiert / im Bild markiert werden. Durch den hohen Zeitaufwand wird dieser Prozess im klinischen Alltag allerdings nur selten durchgeführt. Ersatzweise werden häufig einfache Längenmessungen vorgenommen, deren diagnostische Aussagekraft gegenüber der Volumenbestimmung im Allgemeinen deutlich eingeschränkt ist.

Konkretes Anwendungsbeispiel: Segmentierung und Klassifikation von Gehirnarealen (inklusive Liquor) und deren Volumenbestimmung
Bei dem beschriebenen Fallbeispiel erfolgt eine KI-gestützte, vollautomatische Segmentierung aller relevanten Hirnareale. Hierzu werden die 3D-MRT-Aufnahmen des Kopfes eines Patienten oder einer Patientin an ein zentrales Bildarchivierungs- und Kommunikationssystem (Picture Archiving and Communication System, kurz: PACS) geschickt und dort von einem*r Radiolog*in analysiert. Die 3D-MRT-Bilder werden automatisch analysiert, d. h. spezifische Regionen volumetrisch quantifiziert und dann für Radiolog*innen visualisiert (z. B. in einem übersichtlichen Report, der genaue quantitative Angaben enthält und bestimmte Läsionen markiert). Das aktuelle Fallbeispiel beschränkt sich hierbei auf eine KI-Komponente, welche die Befundung von neurologischen Erkrankungen in der Radiologie unterstützt. Diese KI-Komponente läuft auf einer separaten Datenverarbeitungsrecheneinheit, die lokal in die radiologische Infrastruktur integriert ist. Die Ausführung der Berechnung erfolgt durch Empfang der Daten aus dem PACS. Die ausgegebenen Inhalte (Reports und Visualisierungen) werden auf das PACS nach Beendigung der Berechnung zurückgespielt und sind zusammen mit den aufgenommenen 3D-MRT-Bildern für die Radiolog*innen verfügbar (siehe auch Anhang 13.5).

Der Volumetrie-Report kann zur Unterstützung der Diagnosen bei neurodegenerativen Erkrankungen (wie Alzheimer, Frontotemporaldemenz, multiple Sklerose und Parkinson-Formen) genutzt werden. Dabei können die Radiolog*innen die 3D-MRT-Aufnahmen entweder allein oder gemeinsam mit den Ergebnissen der KI-Komponente darstellen. Anwender*innen des Systems verwenden die Ergebnisse der KI-Komponente folglich als zusätzliche Informationsquelle während der Befundung und können so die rein qualitativ durchgeführte Diagnose mit quantitativen Informationen und zusätzlichen Visualisierungen ergänzen. Die etablierte diagnostische Befundung wird durch die KI-Komponente nicht ersetzt, sondern durch die ergänzend erzeugten Informationen angereichert. Durch die Visualisierung der KI-basierten Ergebnisse sind die Radiolog*innen in der Lage, die Korrektheit der durch die KI-Komponente generierten Informationen zu bewerten. Der Abgleich der durch die KI segmentierten Strukturen mit der in den Bilddaten vorliegenden anatomischen Realität erfolgt im Rahmen der professionellen Kompetenz der klinischen Anwender*innen und bedarf keiner spezifischen Schulung in Bezug auf die KI-Komponente.

Modellbeschreibung und -selektion
Die Segmentierung und Klassifikation von Regionen im Gehirn erfordert zunächst eine Aufbereitung der Daten zur Erstellung der 3-D-Datensätze des Gehirns. Für die semantische Segmentierung der Gehirnregionen werden CNN in Form von Encoder-Decoder-Architekturen eingesetzt. Dabei bietet der Stand der Wissenschaften eine Vielzahl von Architekturen der neuronalen Netze zur semantischen Segmentierung auf Basis von 3-D-Datensätzen wie das DeepLabV3+ [376] oder das U-Net [377].

Neben der Beschreibung der Architektur des eingesetzten KI-Modells ist der Trainings- und Validierungsprozess zu erläutern. Welche Auswahl der Modellparameter (Hyperparameter) und vorverarbeiteten Datensätze haben zum gewünschten Vorhersageergebnis geführt? Falls ein anschließendes Postprocessing notwendig wird, ist dies ebenfalls zu beschreiben.

Performancekriterien
Ein gängiges Performancekriterium zur Bewertung der Qualität einer dreidimensionalen semantischen Segmentierung im KI-Umfeld ist das Intersection over Union (IoU). Diese Metrik stellt das numerische Verhältnis der Schnitt- zur Vereinigungsmenge zwischen vorhergesagter und tatsächlicher Segmentierung dar. Das Ergebnis wird Voxel-weise ausgewertet. Ein weiteres Kriterium ist der F1-Score bzw. Dice Coefficient und Mean Average Precision.

Falls das Ergebnis von den Anwender*innen kurzfristig benötigt wird, wie beispielsweise während einer Operation, ist die Zeit bis zum Vorliegen des KI-Ergebnisses zu berücksichtigen und muss ebenfalls validiert werden.

Datenmanagement
Wie bereits im allgemeinen Kapitel 4.7.2 erwähnt, ist ein gemeinsames Verständnis erforderlich, welche demografischen, epidemiologischen und indikationsspezifischen Variablen für die medizinische Zweckbestimmung einen signifikanten Einfluss haben und demzufolge in den Daten entsprechend repräsentiert sein müssen. In den DICOM-Daten, die von Bildverarbeitungssystemen erzeugt und verwendet werden, können bestimmte demografische Attribute enthalten sein. Hier ist allerdings zu beachten, dass diese Daten nicht immer zur Verfügung stehen bzw. es mitunter kein DICOM-Attribut gibt, um entsprechende Daten zu erfassen. Es besteht darüber hinaus Klärungsbedarf, mit welchem Detaillierungsgrad demografische Attribute zu welchem Zweck (z. B. Trainieren von Machine-Learning-Modellen) DSGVO-konform verwendet werden dürfen.

Beim Labeling bzw. Annotieren der Trainingsdaten für überwachtes Lernen sind generelle Aspekte zu berücksichtigen, wie z. B. die Qualifikation des Personals und der Einsatz validierter Softwarewerkzeuge. Dies gilt sowohl für das Erzeugen der Annotationen als auch für die Prüfung der Annotationen durch eine zweite Person. Aus dem Anwendungsbeispiel ergeben sich spezifische Anforderungen an die für das Labeling verwendeten Softwarewerkzeuge: Die Software sollte Metadaten verarbeiten können und idealerweise über Funktionen verfügen, die den Annotationsvorgang unterstützen und effizienter machen, wie z. B. eine automatisierte Bestimmung des für die Annotation relevanten Sichtfeldes, um die Zeit zu verringern, die mit Scrollen und Zoomen verbracht wird. Darüber hinaus ist eine Bewertung bzw. Kommentierung der Qualität des Bilddatensatzes als Teil des Annotations- und Prüfungsprozesses empfehlenswert.

Ein weiterer Aspekt ist der Einsatz synthetischer Daten, um die Datenmenge durch neue Merkmale zu bereichern. Einen interessanten Ansatz stellen hierbei Generative Adversial Networks (GAN) dar, die bereits auf Basis von MRT-Daten zur Tumorsegmentierung im Gehirn angewendet wurden [378].

Risikomanagement
Für das beschriebene Anwendungsbeispiel ist das Verfahren des Risikomanagements insbesondere im Hinblick auf die Funktion im Rahmen der vermittelten bildbasierten Messfunktion einzuordnen. Es ergeben sich dabei folgende spezifische Herausforderungen.

Ermitteln und Bewerten von Fehlern, die bei der Berechnung des Volumetrie-Reports durch die Verwendung eines MRT entstehen. Zwischen den einzelnen Schichtaufnahmen des MRTs sind jeweils Abweichungen möglich. Ebenfalls muss die Qualität der Berechnung des Volumetrie-Reports hinsichtlich Auswirkungen durch Alter, Geschlecht, Ethnie sowie Vorerkrankungen betrachtet werden. Eine unterschiedliche Dichte oder Veränderung im zu messenden Gewebe aufgrund von Narben oder Schwellungen durch Hirnentzündungen, Hirnaneurysmen, Infarkt oder Adipositas kann Auswirkung auf die Berechnung des Volumetrie-Reports haben. Zudem sollte die Darstellung erlauben, die Verwertbarkeit der Ergebnisse zu prüfen. Dadurch, dass die Anwender*innen die anatomische Korrektheit der durch die AI erzeugten Segmentierungen bei geeigneter Visualisierung auf Basis ihres klinischen Wissens direkt beurteilen können, ist die Anwendungssicherheit der Lösung sichergestellt.

Klinische Bewertung
Die Software ist dazu bestimmt, Informationen zu liefern, die zu Entscheidungen für diagnostische oder therapeutische Zwecke herangezogen werden (Volumina, Anatomie), und übernimmt nicht die Diagnose. Die klinische Bewertung soll nach einem definierten und methodischen Verfahren durchgeführt werden. Wenn genügend belastbare wissenschaftliche Fachliteratur verfügbar ist, die eine Bewertung der Sicherheit, Leistung und Auslegungsmerkmale des Produkts zulässt, kann mit einer systematischen Übersichtsarbeit die Evidenz für die genannten Punkte zusammengestellt werden.

In dem Fall, in dem die KI-Komponente eine neue Technologie darstellt, zu der es keine bzw. nicht genug wissenschaftliche Fachliteratur gibt, ist eine klinische Prüfung notwendig, um genug Daten über die Sicherheit und die Performance des Produkts zu sammeln, insbesondere, um die Performance der KI-Komponente nachzuweisen. Derartige Studien sollten genug Daten produzieren, um eine verallgemeinerbare Aussage über die Segmentierung der Gehirnregionen und deren Volumenwerte zu ermöglichen.

Das Primärziel solcher Studien wäre es, die Performance des Produkts im Rahmen seiner geplanten Zweckbestimmung zu beweisen. Um dies zu erreichen, sollte ein Vergleich zwischen algorithmisch und durch Expert*innen bestimmten Volumenwerten stattfinden. Ein mögliches Maß für die Interrater-Reliabilität ist Cohens-Kappa. Hierfür wird in der folgenden Literaturquelle ein Mindestwert von $\kappa = 0.4$ gefordert, wobei $\kappa > 0.6$ als substanziell und $\kappa > 0.8$ als exzellentes Übereinstimmungsergebnis zählt [379], [380].

Der klinische Bewertungsbericht soll die Sicherheit und Leistung des Produkts untermauern. Hierfür werden die Ergebnisse der Studien als auch die von den nicht-klinischen Testmethoden (z. B. durch Usability, Verifikation) erzeugten nicht-klinischen Daten herangezogen.

Benutzerinteraktion
Wie beschrieben ist das KI-Medizinprodukt zur Unterstützung der klinischen Diagnostik gedacht. Die Verantwortung der Diagnose liegt beim auf die KI-Komponente geschulten klinischen Personal. Die Ergebnisse der KI-Software können der digitalen Patientenakte direkt hinzugefügt werden und so eine Zeitersparnis im klinischen Alltag sein.

Damit das KI-Medizinprodukt einen Mehrwert im klinischen Alltag bietet, ist eine Aufbereitung des Ergebnisses notwendig. Da sich das Gehirnvolumen (bzw. das Volumen einzelner Regionen) mit dem Alter verändert, wäre ein Vergleich mit der Alterskohorte sinnvoll sowie ein Vergleich des Volumens eines Patienten oder einer Patientin zu unterschiedlichen Zeitpunkten.

Fazit
Das beschriebene KI-gestützte Anwendungsbeispiel zur Volumetrie von Hirnregionen und Liquor soll Ärzt*innen bei diagnostischen Entscheidungen zu neurodegenerativen Erkrankungen unterstützen. Die von der KI bestimmten Ergebnisse werden in einem Report zusammengefasst und können visualisiert werden. Eine sinnvolle Ergänzung wäre der direkte Vergleich der bestimmten Volumenwerte mit beispielsweise der durchschnittlichen Alterskohorte bzw. ein Verlaufsdiagramm zur Änderungsverfolgung der Hirnvolumina bei einzelnen Patient*innen.

Um die Software zeitnah verbessern zu können, wäre die Möglichkeit einer Feedbackschleife für Fehldetektionen durch das klinische Personal wünschenswert. Eine solcher Bedarf kann durch ein kontinuierlich oder stufenweise lernendes System gedeckt werden, welcher in Kapitel 2 unter Handlungsempfehlung 4 näher erläutert wird.

4.7.3 Normungs- und Standardisierungsbedarfe

Bedarf 07-01: Nutzbarkeit und Verwertbarkeit von Daten für KI-basierte Systeme in der Medizin
KI-Systeme haben einen hohen Bedarf an Daten, um zuverlässige Aussagen ableiten zu können und auch im Hinblick auf ihre Validierung zuverlässige Bewertungen bezüglich ihrer Performance zu ermöglichen. Im medizinischen Bereich haben Daten einige spezielle Anforderungen, die im Rahmen der Entwicklung und Validierung KI-basierter Systeme zu beachten sind. Erstens handelt es sich typischerweise um **personenbezogene Daten**, die strenge Datenschutzanforderungen erfüllen müssen. Zweitens gibt es für die Erfassung der Daten im medizinischen Bereich hohe Standards (Umsetzung in dedizierten Studiendesigns, Einbindung einer Ethikkommission, hohe Standards in Bezug auf die statistische Auswertung), die dazu führen, dass der Zugang zu den Daten zusätzlich begrenzt ist. Da gemäß MDR oftmals klinische Prüfungen durchgeführt werden müssen, um ein Produkt auf den Markt bringen zu können (d. h. aufwendige und kostenintensive Studien in einem häufig relativ stark eingeschränkten Kontext auf Basis eines noch nicht zugelassenen Produkts), ist speziell die Sammlung von Daten eingeschränkt, die aus realen und breit gefächerten Anwendungsumgebungen kommen. Daher wäre drittens die Frage, welche weiteren Quellen an Daten zulässig wären (z. B. Real World Data aus ähnlichen Anwendungen oder dem Betrieb eines bereits bestehenden Systems, zur Verfügung gestellte Benchmarkdaten, zugängliche zentrale Datenbanken wie beispielsweise im geplanten EHDS vorgesehen, synthetische Daten) und wie sie erschlossen bzw. genutzt werden könnten, um mehr Daten aus realen Umgebungen zur Verfügung zu haben. Und viertens wäre zu klären, inwieweit die bestehenden, auf klassische Studiendesigns ausgerichteten statistischen Anforderungen anzupassen sind, wenn statt einzelne dedizierte Parameter bei Machine-Learning-Verfahren oftmals sehr viel komplexere

Szenarien abzudecken sind. Dabei wäre auch die Frage, ob eine randomisiert kontrollierte Studie für KI-basierte Systeme nach wie vor als Maß der Dinge betrachtet werden sollte oder ob hier Alternativen besser geeignet wären (z. B. bei Updates eines bestehenden Systems anhand von Real World Data). Übergeordnet ist dabei die Anforderung zu klären, wann eine Datenmenge und -auswahl für eine bestimmte Anwendung als ausreichend repräsentativ betrachtet werden kann und welche Daten auf welche Weise hierbei herangezogen werden können oder sogar müssen.

Zusammenfassend ergeben sich folgende Punkte zur Klärung bzw. Umsetzung in der Normung als Handlungsbedarfe:

→ Klärung von Anforderungen an das Datenmanagement und die damit verbundenen Prozesse inklusive von Vorgaben an die Datenakquise, an das Labeling, die Qualifikationen der beteiligten Personen und insgesamt an eine standardisierte Bewertung von Datensammlungen.

→ Klärung der Anforderungen an (klinische) Studiendesigns, die für die Validierung KI-basierter Medizinprodukte nutzbar sind – u. a. in Bezug auf die Rahmenbedingungen des Designs, den Umfang der Daten, Repräsentativität für den Anwendungsfall, Zugang zu den Daten (z. B. durch Benannte Stellen für eine Überprüfung wie im geplanten AI Act gefordert), Nutzbarkeit von Daten aus zentralen Datenbanken und Rahmenbedingungen in Bezug auf den Datenschutz.

→ Klärung der Anforderungen, inwieweit Real-World-Data für die Entwicklung und Erprobung KI-basierter Medizinprodukte herangezogen werden können. Das beinhaltet potenziell die Nutzung mitgeloggter Daten aus dem Betrieb einer bereits bestehenden Version des Systems, aber auch die Nutzung eines anderen Systems, das Daten aus der Betriebsumgebung systematisch erfasst.

→ Klärung der Anforderungen an die Nutzung synthetischer Daten für KI-basierte Medizinprodukte. Das beinhaltet die Anwendung generischer Verfahren im Kontext von Machine Learning wie z. B. GANs als auch spezifische Modelle, um neue Daten zu erzeugen, wie z. B. die künstlich erzeugte Transformation eines MRT-Datensatzes auf eine andere Parameterdarstellung bzw. ein neues Verfahren unter Bezug auf ein deterministisches Modell für die Übertragung von der bestehenden auf die neue Modalität. Im medizinischen Kontext ist bei der Nutzung synthetisierter Daten zu klären, in welchem Umfang die Einhaltung spezieller Anforderungen (z. B. die Verlässlichkeit der erzeugten Daten für die jeweilige Anwendung, datenschutzrechtliche Aspekte) nachgewiesen werden muss.

Bedarf 07-02: Gestaltung geeigneter Metriken für unterschiedliche Arten KI-basierter Medizinprodukte

Die Konformitätsbewertung bei Medizinprodukten erfordert eine systematische Überprüfung ihrer Leistungsfähigkeit und auch ihrer Sicherheit, die bei KI-Systemen in Form geeigneter Metriken entsprechend quantifiziert werden muss. Dabei gibt es einige Abweichungen in Vergleich zu anderen Branchen. Übergeordnet ist bei Medizinprodukten immer der klinische Outcome zu bewerten, der sowohl Risiken als auch den Nutzen für die Patient*innen beinhaltet. Deshalb erfordert die MDR in der klinischen Bewertung eine systematische Betrachtung des Risiko-Nutzen-Verhältnisses als zentralen Schritt im Konformitätsbewertungsverfahren. Zudem ist ein Abgleich mit Referenzverfahren wie dem etablierten Standard of Care und auch den sich bereits auf dem Markt befindlichen Produkten erforderlich. Das erfordert, dass definierte Referenzkriterien vorhanden sind, um nicht nur einzelne Systeme zu bewerten, sondern einen gezielten Vergleich zwischen Systemen umzusetzen inklusive Maßgaben, bis wann eine Äquivalenz der Systeme als gegeben angesehen werden kann. Idealerweise sollten für unterschiedliche Anwendungen Benchmarking-Datensätze vorliegen, um einen standardisierten Abgleich vollziehen zu können.

Hinzu kommt, dass im medizinischen Bereich die Bewertungen stark use-case-spezifisch umgesetzt werden müssen, damit der spezifische Nutzen bzw. die resultierenden Risiken gezielt bewertet werden können. Dabei ist zu berücksichtigen, dass es sehr unterschiedliche Anwendungsbereiche gibt, die z. B. Aufgabenstellungen in den Bereichen Diagnostik, Monitoring und Therapie umfassen und diese zusätzlich mit unterschiedlichen Autonomie- und Risikograden verbunden sein können. Gerade bei KI-basierten Systemen können dabei neben technischen Risiken auch Faktoren wie Transparenz oder Erklärbarkeit (wie kommt die KI grundsätzlich zu welcher Entscheidung, welche Grundannahmen legt sie dabei zugrunde, welche Schritte vollzieht die KI zum aktuellen Zeitpunkt) sowie die Eingriffsmöglichkeiten im Rahmen einer menschlichen Aufsicht als Kriterien einfließen. Diese sind ebenfalls im Sinne ihrer klinischen Wirksamkeit zu betrachten.

Um Risiken effektiv reduzieren und den Nutzen optimieren zu können, ist es im Grunde erforderlich, diese Faktoren in die Bewertungsmetriken einfließen zu lassen. Derartige Faktoren können oftmals gegenläufig sein (z. B. Transparenz vs. Genauigkeit, Sicherheit für die einzelnen Patient*innen vs. Nutzen für eine bestimmte Bevölkerungsgruppe), sodass Zielkonflikte entstehen, die das Bewertungskriterium als Ganzes

erfassen muss. Eine Reduktion einzelner Risiken, so wie es typischerweise in Risikomanagement-Normen anvisiert wird, ist gerade bei KI-basierten Systemen nur bedingt wirkungsvoll. Dabei ist zu beachten, dass die Integration von Risikoaspekten in die Metriken auf abgestufte Weise erfolgen sollte, da oftmals während der Entwicklung eine Quantifizierung dieser Punkte nicht vollständig möglich ist. Diese Integration sollte keine essenzielle Hürde für die erfolgreiche Bewertung der Konformität eines Medizinprodukts sein, aber das Verbesserungspotenzial der Systeme möglichst gezielt ausschöpfen.

Insofern gibt es einige Anforderungen, die über bestehende Ansätze hinausreichen und für die KI-spezifische Vorgaben zu entwickeln sind. Es ist zu klären, welche Metriken in Bezug auf KI-basierte Systeme relevant sind und wie diese in Hinblick auf die Konformitätsbewertung umzusetzen sind. Dies beinhaltet insbesondere die folgenden Aspekte.

→ Bereitstellung standardisierter Metriken, um einen systematischen Abgleich verschiedener Systeme für vergleichbare Anwendungsfälle umsetzen zu können.
→ Integration von KI-spezifischen Risikofaktoren und Aspekten des klinischen Nutzens in die Bewertungskriterien, sodass eine Optimierung des Risiko-Nutzen-Verhältnisses im Gesamtsystem in geeigneter Weise umgesetzt werden kann.
→ Festlegung von überprüfbaren, ggf. gestuften Anforderungen an die Transparenz und die Erklärbarkeit, die es Anwender*innen und Patient*innen erlauben, das grundlegende Wirkprinzip zu verstehen, und dem Anwendenden zugleich eine Orientierung bei der kritischen Bewertung von KI-basierten Entscheidungen ermöglicht.
→ Berücksichtigung unterschiedlicher Autonomiegrade und Anwendungsbereiche (z. B. Diagnostik vs. Monitoring vs. Therapie) sowie der jeweils damit verbundenen Qualitätskriterien. Klärung von deren Wechselwirkung mit Maßnahmen im Entwicklungs- und Lebenszyklusprozess von Medizinprodukten.

Bedarf 07-03: Gesellschaftliche und regulatorische Rahmenbedingungen für die Anwendung von KI in Medizinprodukten

Die Entwicklung der MDR bzw. In Vitro Diagnostic Medical Devices Regulation (IVDR) hat aufgezeigt, wie wichtig es ist, Regularien so zu gestalten, dass sie auch im gegebenen zeitlichen Rahmen gut umsetzbar sind und sich positiv auf die Gesundheitsversorgung auswirken. Zu hohe Hürden (z. B. Erforderlichkeit klinischer Studien für lange etablierte Bestandsprodukte mit begrenztem Risikopotenzial) und nicht verfügbare Infrastruktur (z. B. Verfügbarkeit zentraler Datenbanken wie Eudamed, fehlende Benannte Stellen und harmonisierte Normen) führen nicht nur zu Unsicherheiten im Entwicklungsprozess, sondern können auch Probleme für den Industriestandort Deutschland sowie für die Gesundheitsversorgung insgesamt (z. B. Fehlen wichtiger Nischenprodukte) bewirken.

Durch den geplanten AI Act der EU kommt eine zusätzliche Komplexitätsstufe hinzu, die Wechselwirkungen mit bestehenden Regularien beinhaltet und damit nochmals Mehrbelastungen in Bezug auf die Konformitätsbewertung bei KI-basierten Medizinprodukten bewirken kann (siehe auch Anhang 13.1, Abschnitt „Exemplarische Darstellung am Beispiel Medizinprodukte“). Um die Innovationskraft in diesem wichtigen Zukunftsbereich nicht unverhältnismäßig einzuschränken, sollten Inkonsistenzen zwischen dem geplanten AI Act und der MDR (z. B. im Bereich Risikomanagement, Datenbanken für Post Market Surveillance) beseitigt und Doppelaufwendungen minimiert werden. Begleitend zur Gestaltung des geplanten AI Act sollten die entsprechenden Normen vorbereitet werden, um für eine konsistente und möglichst effiziente Umsetzung der Vorgaben zu sorgen.

Zusätzlich sollte die Zugänglichkeit zu Daten verbessert werden, um die Hürden für die Entwicklung und Umsetzung neuer KI-basierter Anwendungen in der Medizin nicht unverhältnismäßig hoch zu setzen. Hier ist eine gute Balance zwischen den durch die DSGVO gegebenen Datenschutzaspekten und den medizintechnischen Anforderungen gefragt, die umfassende Daten brauchen, um ein passendes Sicherheitsniveau bzw. einen entsprechenden klinischen Nutzen zu erreichen. Die Bestrebungen in Richtung EHDS sind ein Ansatz in dieser Richtung. Es ist dabei aber zu beachten, dass auch Unternehmen Zugang zu entsprechenden Daten benötigen, um konkurrenzfähig und innovativ sein zu können.

Insgesamt sollte der Fokus in Zukunft stärker auf einer Bewertung liegen, inwieweit neue Produkte einen positiven Effekt auf die Gesamtversorgung haben. Das heißt, die Risiken eines einzelnen Produkts sollten stärker mit dem gesamtgesellschaftlichen Nutzen abgewogen werden können. Auch ein Produkt, das aufgrund zu hoher Hürden nicht auf dem Markt verfügbar ist und daher in der Gesundheitsversorgung fehlt, bewirkt einen Schaden. In diesem Zusammenhang sollte eine gezielte Evaluation erfolgen, die die Wirkung der Regularien selbst (und der zugehörigen Normen) nicht nur in Bezug auf ihre technische Umsetzung, sondern auch hinsichtlich ihrer Wirkung auf das Gesundheitssystem insgesamt bewertet.

Im Einzelnen entstehen folgende Bedarfe:

→ **Elimination von Inkonsistenzen und Doppelbelastungen zwischen geplantem AI Act und MDR** bzw. IVDR (z. B. im Bereich Risikomanagement, Datenbanken für Post Market Surveillance).
→ **Sicherstellung der Infrastruktur zur Umsetzung des geplanten AI Act**, z. B. in Hinblick auf die detaillierte Klärung der enthaltenen Anforderungen, Bereitstellung harmonisierter Normen, Verfügbarkeit zentraler Datenbanken (z. B. für Post Market Surveillance) sowie Verfügbarkeit von Benannten Stellen, die sowohl für den geplanten AI Act als auch für die MDR bzw. IVDR zertifiziert sind. Es ist darauf zu achten, dass die Übergangsfristen des geplanten AI Act entsprechend ausgelegt sind.
→ Verbesserung des Zugangs zu medizinischen Daten, um die Innovationskraft in Deutschland/Europa zu stärken und über KI-basierte Systeme auch die Gesundheitsversorgung insgesamt zu verbessern. Das sollte auch einen geeigneten Zugang zu Daten für Unternehmen beinhalten.
→ Stärkerer Einbezug der positiven Wirkungen von (KI-basierten) Medizinprodukten im Rahmen der Konformitätsbewertung bzw. Risiko-Nutzen-Bewertung.
→ Gezielte Evaluierung der Regularien und Normen (auf wissenschaftlicher und unabhängiger Basis) in Bezug auf deren Wirkung hinsichtlich der Gesundheitsversorgung insgesamt.

Bedarf 07-04: Autonomiegrade bei KI-basierten Systemen – verschiedene Stufen von Human-in-the-Loop bis hin zu Closed-Loop-Modellen

Im Bereich KI-basierter Anwendungen in der Medizin gibt es ein breites Spektrum an Autonomiegraden, die bei unterschiedlichen Aufgabenstellungen auftreten – von einem reinen Mitloggen der Daten über dedizierte diagnostische Entscheidungshilfen und Unterstützungssysteme im Bereich Monitoring (wie z. B. der Intensivmedizin) bis hin zu hochautomatisierten Systemen. Bei einem geringen Autonomiegrad müssen die Benutzer*innen (z. B. medizinisch geschultes Personal) die algorithmischen Ergebnisse in einer verlässlichen Weise überwachen können („Human/Clinician in the Loop"-Systeme). Das erfordert, dass die Benutzer*innen auch in dynamischen und komplexen Umgebungen ein ausreichendes Verständnis des Systems haben, um auf dessen Entscheidungen richtig reagieren zu können. Bei einem hochautomatisierten Ansatz – im Extremfall einem Closed-Loop-System – muss das zentrale Systemverhalten hingegen ohne einen Eingriff des Menschen gesteuert werden und dennoch sicher funktionieren. Im Gegensatz zu anderen Branchen (z. B. Automobilbereich mit Abstufungen von assistiertem Fahren bis zu autonomem Fahren) gibt es im Bereich der Medizintechnik keine konsequente Einteilung in Autonomiegrade, sondern nur sehr begrenzte Anhaltspunkte. Die PD IEC/TR 6060141:2017 [373] beinhaltet in Annex C Klassifizierungstabellen für Autonomiegrade, die jedoch keine KI-basierten Aspekte adressieren. Für Closed-Loop-Systeme, die klassische, regelbasierte Ansätze verwenden (z. B. anhand physiologischer Modelle) gibt es mit der DIN EN 60601-1-10:2021 [375] eine normative Grundlage, bei der jedoch ebenfalls der klassische physiologische Regelkreis im Fokus steht und nicht ein KI-basiertes System.

Daher besteht ein Bedarf an Klärung, welche Autonomiegrade in Bezug auf KI-basierte Systeme relevant sind und wie sich diese auf die Konformitätsbewertung bei KI-basierten Medizinprodukten und speziell auf die Interaktion mit dem Menschen (bei Human-in-the-Loop-Systemen) bzw. mit physiologischen Systemen (bei manchen Closed-Loop-Systemen) auswirken. Das beinhaltet insbesondere die folgenden Aspekte.

→ Definition unterschiedlicher Autonomiegrade und Klärung der daraus resultierenden Anforderungen hinsichtlich der Maßnahmen im Entwicklungs- und Lebenszyklusprozess von Medizinprodukten. Das betrifft insbesondere den Einfluss der Autonomiegrade in Bezug auf die Bewertung/Behandlung von Risiken, der Validierung der Systeme oder auch der Überwachung im Feld. Zudem ist die Wechselwirkung mit anderen Parametern der Risikobewertung, z. B. Schweregrad und Eintrittswahrscheinlichkeiten, zu berücksichtigen. Bei ML-basierten Systemen kommen weitere Parameter wie Komplexität und Interpretierbarkeit der Systeme hinzu. Insgesamt ist ein konsequent risikobasierter Ansatz zu entwickeln, der die Abstufungen bezüglich der Autonomiegrade entsprechend berücksichtigt und Anpassungen der damit verbundenen Anforderungen gezielt ermöglicht.
→ Speziell bei Human-in-the-Loop-Ansätzen: Klärung der Anforderungen in Bezug auf die Mensch-Maschine-Interaktion (siehe auch Anforderungen zur menschlichen Aufsicht im geplanten AI Act): Welche Informationen benötigen die Benutzer*innen in welcher Weise, um erforderliche Reaktionen umsetzen zu können, z. B. auch in Unterscheidung zwischen Alarmen (unmittelbare Erforderlichkeit einer Aktion) und Alerts (Aufmerksammachung zur Initiierung weiterer Klärungsschritte)? Das beinhaltet zudem Klärungen bezüglich der Anforderungen an Transparenz und Erklärbarkeit/Interpretierbarkeit der Systeme speziell auch in Bezug auf das sehr dynami-

sche Systemverhalten, das KI-basierte Systeme aufweisen können. Dabei ist zu klären, welche Maßnahmen die menschliche Aufsicht beinhalten kann, um Risiken wie ein zu starkes oder zu schwaches Verlassen auf die Entscheidungen des Systems oder einen Model Drift zu vermeiden.

→ Speziell bei Closed-Loop-Systemen: Klärung der Anforderungen an die Verlässlichkeit von KI-basierten Systemen oder Komponenten, die nicht wie bisherige Closed-Loop-Systeme hauptsächlich auf etablierten physiologischen Modellen basieren. Zudem Klärung, ab wann ein System nur eine Konfiguration von Parametern (z. B. KI-basierte Abschätzung / Anpassung individueller Parameter) und wann eine Veränderung des Closed-Loop-Systemverhaltens darstellt sowie Klärung der Frage, unter welchen Bedingungen / mit welchen Anforderungen Kombinationen aus KI-basierten und klassischen physiologischen Modellen, d. h. hybride Modelle, auf den Markt gebracht werden können.

Bedarf 07-05: Klärung der Abgrenzung zwischen Medizin- und Nicht-Medizinprodukten in Verbindung mit abgestuften Anforderungen bei KI-basierten Systemen im Gesundheitsbereich

KI-basierte Anwendungen in der Medizin können ein breites Spektrum an Systemen und Komponenten abdecken. Erstens gibt es den Bereich der Medizinprodukte, deren Anforderungen in Europa durch die MDR geregelt sind und für die in vielen Fällen umfangreiche zusätzliche Anforderungen aus dem geplanten AI Act hinzukommen werden, nachdem die MDR im Rahmen des geplanten AI Act als ein substanzieller Indikator für die Einordnung eines KI-basierten Systems als Hochrisikosystem betrachtet wird. Zweitens gibt es die In-vitro-Diagnostika, deren Konformitätsbewertung in Europa über die IVDR geregelt wird. Aufgrund der Verwandtschaft zwischen MDR und IVDR sowie der Tatsache, dass auch IDVR-Produkte explizit als Kandidaten für Hochrisikoprodukte gelistet sind, gelten zu einem nicht unerheblichen Anteil ähnliche Anforderungen wie bei Medizinprodukten.

Darüber hinaus gibt es KI-basierte Anwendungen mit Gesundheitsbezug, die nicht unter die MDR oder IVDR fallen. Dazu gehören allgemeine Gesundheitsanwendungen, um die eigene Gesundheit individuell zu managen, erhalten oder zu verbessern (z. B. durch „Smart Watches“ oder andere Anwendungen, die Gesundheitsparameter messen und KI-basiert verarbeiten). Hier greifen aktuell nur sehr bedingt spezifische Regularien oder Normen. Als Ausnahme liefert die DIN EN 82304-1:2018 [354] einige allgemeine Anforderungen, die aber keine speziellen Aspekte von KI-Systemen beinhalten. Da solche Werkzeuge in Zukunft an weiterer Bedeutung gewinnen werden und diese auch verlässlich arbeiten müssen, um Risiken zu vermeiden und einen positiven Effekt auf die Gesundheit des Einzelnen bewirken zu können, wären Vorgaben für die Umsetzung derartiger Systeme hilfreich. Dies wäre insbesondere durch die Bereitstellung passender Normen anzugehen. Um die Innovationskraft nicht unangemessen einzuschränken, sollten sich diese Vorgaben von den strengeren Anforderungen an Medizinprodukte abgrenzen. Dabei wäre es vorteilhaft, wenn auf internationaler Ebene Einigkeit erreicht werden könnte, wann ein KI-basiertes System eine allgemeine Gesundheitsanwendung und wann ein Medizinprodukt ist bzw. welche Einordnung bezüglich des Risikoniveaus das jeweilige System aufweist. Ähnliches gilt für Systeme in der Gesundheitsversorgung, die selbst keinen medizinischen Zweck aufweisen und damit auch kein Medizinprodukt darstellen, auf der anderen Seite aber Abläufe im Gesundheitssystem unterstützen, z. B. Optimierung von Prozessen in einem Krankenhaus oder einer Pflegeeinrichtung.

Eine letzte Art von KI-Anwendungen sind Komponenten und Werkzeuge, die die Entwicklung, Konformitätsbewertung oder den Betrieb von Medizinprodukten unterstützen, z. B. im Bereich der Qualitätssicherung, der Optimierung von Prozessen und Produkten oder der Auswertung im Bereich der Marktüberwachung nach der Inverkehrbringung. Diese Komponenten fallen in gewissem Maße in den Einflussbereich der MDR bzw. IVDR und müssen z. B. gemäß DIN EN ISO 13485:2021 [381] (Bereich „Computer System Validierung“) entsprechende Anforderungen erfüllen. Auch hier sind bisher aber keine KI-spezifischen Aspekte abgedeckt, sodass die aktuellen Normen entsprechend ergänzt werden sollten.

In allen Fällen von Nicht-Medizinprodukten bleibt im geplanten AI Act die Problematik, dass er zwar Hochrisiko-KI-Systeme sehr detailliert regelt, dass er auf der anderen Seite aber wenig Klärung beinhaltet, wie Systeme auf den Markt gebracht werden sollten, die geringere Risikoanforderungen beinhalten.

Zusammenfassend bleiben folgende zentrale Handlungsbedarfe:

→ **Verbesserte Abgrenzung zwischen Medizinprodukten und Nicht-Medizinprodukten bzw. konsistente Einordnung der jeweiligen KI-basierten Systeme bezüglich ihres Risikoniveaus** – idealerweise im internationalen Konsens, der jedoch mit den jeweiligen gesetzgeberischen Vorgaben abgestimmt sein muss.

→ **Definition von reduzierten Anforderungen** an KI-basierte Systeme oder auch Teilkomponenten, die selbst ein geringeres Risiko aufweisen, dennoch aber eine hohe Verlässlichkeit beinhalten sollten, um einen positiven Effekt auf die Gesundheitsversorgung zu haben. Hierzu zählen insbesondere allgemeine Gesundheitsanwendungen, KI-basierte Systeme zur Verbesserung der Prozesse im Bereich von Gesundheitseinrichtungen sowie Werkzeuge für die Entwicklung und Optimierung von Medizinprodukten und In-vitro-Diagnostika.

Bedarf 07-06: Anwendung von Assurance Cases zur Erbringung von Sicherheitsnachweisen bei KI-basierten Anwendungen im Bereich der Medizin

Alternativ zur Auslegung bestehender regelbasierter, aber den Bereich KI unzureichend adressierender Normen erscheint ein stärker zielorientiertes Vorgehen bei Sicherheitsnachweisen für KI-Komponenten mittels des in der ISO/IEC/IEE 150261:2019 [114] definierten Konzepts der Assurance Cases gerade im medizinischen Bereich als sinnvolle Grundlage und Brücke zu kommenden KI-Standards [382], die daher intensiver betrachtet werden sollte. In der Norm wird ein Assurance Case hierbei als begründbares und überprüfbares Artefakt verstanden, das die Annahme stützt, dass eine aufgestellte Behauptung (z. B. bezüglich der Sicherheit eines Medizinprodukts) erfüllt ist und dabei eine systematische Argumentation sowie die zugrunde liegenden Beweise und expliziten Annahmen, auf die sich die Behauptung stützt, umfasst ISO/IEC/IEE 15026-1:2019 [114].

Der Einsatz von Assurance Cases wird insbesondere dann empfohlen, wenn innovative Anwendungsfälle umgesetzt oder neuartige Technologien zum Einsatz gebracht werden sollen [383]. Beides liegt beim Einsatz von KI in Medizinprodukten gewöhnlich vor. Mittels Assurance Cases lässt sich so der Nachweis der Einhaltung von im jeweiligen Bereich akzeptierten Risikoakzeptanzkriterien (vgl. z. B. [384]) strukturiert auf die durch die Qualitätssicherung erbrachten Evidenzen herunterbrechen [385]. Hierdurch wird die Relevanz und der Beitrag der jeweiligen Maßnahmen bei der Absicherung der KI-Anteile des Produkts transparent aus der Sicherstellung eines akzeptablen Risiko-Nutzen-Verhältnisses sowie der Reduktion des Restrisikos begründbar.

Erfahrungen aus der Anwendung von Assurance Cases als strukturierte Argumentationen unterstützen zudem bei der Entwicklung von Normen mit begründbaren Anforderungen. Die Entwicklung und Konsolidierung geeigneter Argumentationsmuster beim Einsatz von KI in Medizinprodukten sowie ihre praktische Anwendung, beispielsweise im Rahmen von Experimentierräumen, sollte daher durch die Politik gefördert werden.

Empfehlungen:

→ Förderung der Anwendung von Assurance Cases als sinnvolle Grundlage und Brücke zu kommenden KI-Standards

Die Arbeitsgruppe Medizin hat die identifizierten Bedarfe nach der Dringlichkeit ihrer Umsetzung bewertet. Abbildung 43 zeigt die Dringlichkeit der Umsetzung, kategorisiert nach den Zielgruppen Normung und Politik.

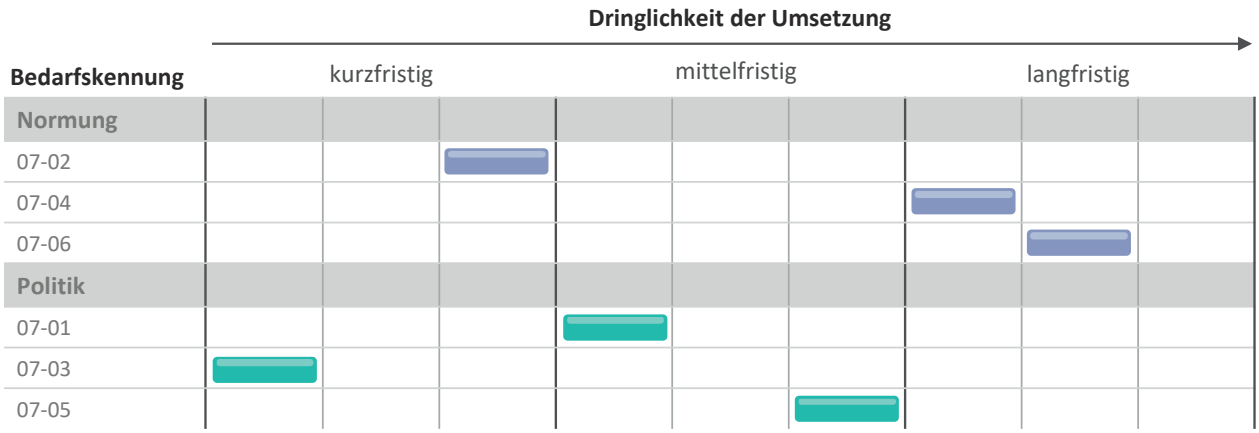


Abbildung 43: Priorisierung der Bedarfe aus Schwerpunkt Medizin (Quelle: Arbeitsgruppe Medizin)

4.8 Finanzdienstleistungen

Künstliche Intelligenz (KI) ist eine der Schlüsseltechnologien des 21. Jahrhunderts, die in den kommenden Jahren und Jahrzehnten unternehmerisches Handeln, aber auch das Leben von Bürger*innen in vielfacher Hinsicht beeinflussen wird. Mit Blick auf Unternehmen, insbesondere Finanzinstitute, werden nicht nur Geschäftsprozesse mit KI automatisiert, sondern auch gänzlich neue Geschäfts- und Betriebsmodelle entwickelt. Dies kann dazu führen, dass Unternehmen sich auf eine komplett neue Art organisieren und Wert schöpfen.

KI birgt jedoch auch Risiken, die insbesondere mit Blick auf die Gesellschaft zum Vorschein kommen. Kritische Entscheidungen einem KI-System zu überlassen kann dazu führen, dass bestimmte Bevölkerungsgruppen unfair behandelt oder gar diskriminiert werden. Dies gilt es zwingend zu vermeiden. Der Schutz der Grundrechte muss beim Einsatz von KI-Systemen oberste Priorität sein. Bürger*innen der Europäischen Union (EU) oder die hier lebenden Menschen müssen in allen Bereichen des Lebens vor den Risiken von KI-Systemen geschützt werden.

Die Finanzbranche ist bereits jetzt auf KI angewiesen und ist ohne den Einsatz von KI-Systemen nicht mehr vorstellbar, siehe Abbildung 44. Finanzinstitute sehen sich mit einem kontinuierlichen Wachstum an Daten konfrontiert, mit denen umgegangen werden muss. Darüber hinaus birgt KI Möglichkeiten zur Innovation, die jetzt noch nicht abzusehen sind. Start-ups oder FinTechs in der Finanzbranche zeigen, dass mit KI völlig neue Geschäftsmodelle entwickelt werden können, aber auch die Geschäftsmodelle der etablierten Banken befinden sich in einem durch datengetriebene Systeme geprägten Umbruch.

4.8.1 Status quo

KI kann in Finanzinstituten eingesetzt werden, um eine Reihe von Anwendungsfällen zu automatisieren (siehe Abbildung 44):

Grundsätzlich birgt der Einsatz von KI-Systemen drei Risiken, die von Finanzinstituten adressiert werden müssen:

- → **Hohe Komplexität:** Die in modernen KI-Systemen eingesetzten, durch Maschinelles Lernen (ML) erzeugten Modelle können eine deutlich höhere Komplexität haben, als dies bei klassischen KI-Systemen der Fall war, sodass es deutlich erschwert wird, die Systeme nachzuvollziehen und zu prüfen.
- → **Kurze Rekalibrierungszyklen:** Da die in KI-Systemen verwendeten ML-Modelle in kürzeren Abständen neu trainiert werden und sich dadurch stetig weiterentwickeln, muss auch die Validierung dynamisch erfolgen.
- → **Bias:** Aufgrund von möglichen schwer erkennbaren Verzerrungen in den verwendeten großen und komplexen Datenquellen steigt das Risiko eines verzerrten Ergebnisses sowie einer unfairen Behandlung von bestimmten Bevölkerungsgruppen.

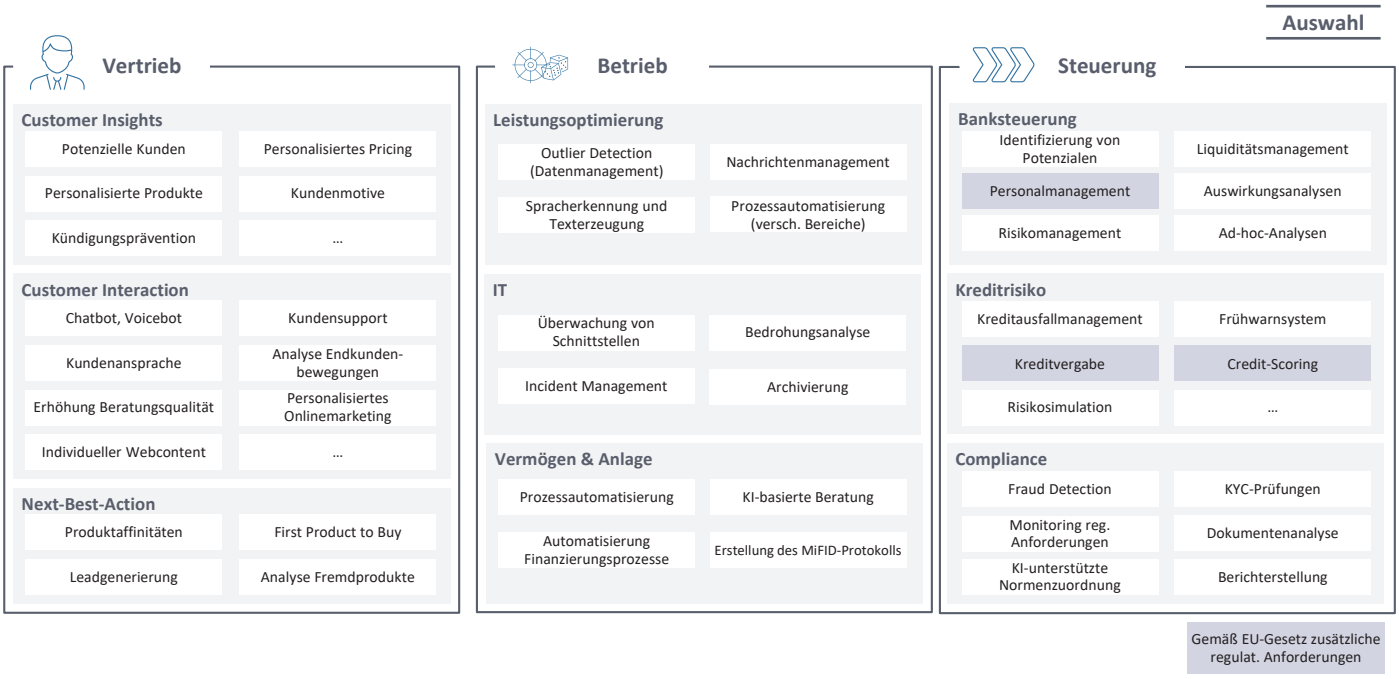


Abbildung 44: KI in der Finanzbranche (Quelle: Arbeitsgruppe Finanzdienstleistungen)

Im Entwurf der EU-Kommission zur Regulierung von KI wurden mehrere Anwendungsfälle genannt, die durch den Einsatz von KI hohe Risiken bergen, zwei davon sind für die Finanzbranche relevant: Kreditwürdigkeitsprüfungen bei der Kreditvergabe und Mitarbeitermanagementsysteme. Die Argumentation der EU-Kommission ist hier, dass durch einen Bias (also eine systematische Verzerrung in der Ausgabe gegenüber einem bekannten richtigen Ergebnis, die unterschiedliche Ursachen haben kann) im KI-System für bestimmte Bevölkerungsgruppen der Zugang zu finanziellen Mitteln oder Weiterentwicklungs- und Aufstiegsmöglichkeiten in ihrem Berufsleben erschwert wird. Daher müssen hier bestimmte Kontrollmechanismen eingerichtet werden, die einer unfairen Behandlung von bestimmten Menschen vorbeugen oder diese im Nachhinein korrigieren. Da Daten oft einfach die Gegebenheiten der Gesellschaft widerspiegeln, werden auch bestehende negative Tendenzen und die Schlechterstellung bestimmter Bevölkerungsgruppen übernommen. Außerdem können Verzerrungen dadurch entstehen, dass die Lerndaten in wesentlichen Aspekten einseitig lückenhaft sind.

Dennoch sind Finanzinstitute mit sehr komplexen Risikomanagementsystemen ausgestattet, wodurch es den Unternehmen möglich wird, neue Risiken zu identifizieren, diesen vorzubeugen oder zu mitigieren. Zudem sind die hier genannten Risiken den Finanzinstituten bereits bekannt und werden durch die Risikomanagementprozesse adressiert und gesteuert.

4.8.2 Anforderungen und Herausforderungen

In diesem Kapitel sollen die spezifischen Anforderungen behandelt werden, die sich durch die Anwendung von KI im Finanzsektor ergeben. Die Spezifik entspringt vor allem aus den folgenden zwei Umständen, einerseits aus Verbraucherinnen- und Verbrauchersicht, andererseits aus Institutssicht.

Aus Verbraucherinnen- und Verbrauchersicht
Aus Sicht der Kundinnen und Kunden handelt es sich bei KI-Anwendungen im Finanzbereich häufig um Anwendungen mit direktem Bezug zu Menschen, ähnlich den Anwendungen im Bereich soziotechnischer Systeme. Die folgenden Aspekte bergen spezielle Herausforderungen:

→ Häufig werden hier Modelle für menschliches Verhalten benötigt. Dieses ist in aller Regel variabel, zeitlich veränderlich und stark zwischen den Individuen vernetzt, deren Verhalten das Modell beschreibt.
→ Für die Erstellung dieser Modelle müssen häufig komplexe, vernetzte und unstrukturierte (das heißt interpretationsbedürftige) Daten berücksichtigt werden.
→ Es werden häufig Modelle für Gruppen erstellt, deren durchschnittliche Eigenschaften auf Individuen übertragen werden sollen.
→ Die von den Modellen gelieferten Ergebnisse haben unter Umständen Bezüge zu wichtigen Voraussetzungen gesellschaftlicher Teilhabe bis hin zu Grundrechten, etwa beim Zugang zu Krediten und anderen grundlegenden Finanzdienstleistungen.
→ Aufgrund der strengen Vertraulichkeit und Schutzbedürftigkeit der persönlichen Finanzdaten ergibt sich eine besondere Relevanz des Datenschutzes und der Datensicherheit.

Aus Institutssicht
Finanzmittel, insbesondere im Wortsinn Kredite, sind inhärent risikobehaftet, das geschäftliche Umfeld von Finanzdienstleistern ist stark von Risiken und Komplexität geprägt, Abhängigkeiten zwischen Systemen müssen berücksichtigt werden. Risikomanagement ist daher Teil des Kerngeschäftsmodells jedes Finanzunternehmens, und sowohl die Verwendung von ML-basierten KI-Systemen als auch das Management der damit verbundenen Modellrisiken sind seit Langem etablierter Standard in Finanzinstituten, der auch von den Aufsichtsbehörden ständig überwacht wird. Daraus ergeben sich im Umgang mit KI-Systemen besondere Herausforderungen:

→ Die bereits bestehende Systemlandschaft von KI-Modellen ist integraler Bestandteil des Geschäftsmodells, Eingriffe, auch aufgrund von neuen Standards, können daher große Auswirkungen z. B. auf die Kapitalbasis haben. Daher müssen alle Begrifflichkeiten präzise definiert sein und Standards quantitativ gut begründet.
→ Neue Validierungs- und Zertifizierungsprozesse müssen in den bestehenden Modellrisikomanagement-Rahmen eingebettet werden.
→ Der grundlegende Fokus aus Institutssicht liegt auf dem Management von Portfoliorisiken, nicht von Einzelrisiken, entsprechend den aufsichtsrechtlichen Anforderungen.

Vor diesem Hintergrund wurden die im Folgenden beschriebenen fünf Themenschwerpunkte zur KI-Normung im Finanzbereich identifiziert.

4.8.2.1 Besonderheiten des Finanzsektors

Gesetzliche und aufsichtsrechtliche Anforderungen an die Finanzbranche

Nur wenige Branchen unterliegen so vielen regulatorischen Anforderungen wie die Finanzbranche, insbesondere die Kreditwirtschaft und der Zahlungsverkehr, deren Einhaltung durch sektorspezifische Aufsichtsbehörden auf nationaler und europäischer Ebene streng überwacht wird.

Die regulatorischen Anforderungen, die für KI-Systeme besonders relevant sind, sind jene, die das Risikomanagementsystem und die Informationstechnologie von Finanzinstituten betreffen. Wichtig ist hier zu erwähnen, dass die Risikomanagementsysteme von Banken in der Lage sind, neue Risiken zu identifizieren, zu bewerten und zu mitigieren, sodass Risiken, die durch den Einsatz von KI-Systemen entstehen, angemessen gesteuert werden können.

Hier soll zusammenfassend aufgezeigt werden, welche Anforderungen für Systeme der Informationstechnologie im Finanzsektor gelten.

Europäische Anforderungen

DIGITAL OPERATIONAL RESILIENCE ACT

Am 24. September 2020 hat die EU-Kommission den Entwurf für die „Verordnung über die Betriebsstabilität digitaler Systeme des Finanzsektors [386] “ veröffentlicht. Die Verordnung hat das Ziel, sicherzustellen, dass alle Teilnehmenden des Finanzsystems über die notwendigen Sicherheitsvorkehrungen verfügen, um Cyberangriffen sowie anderen Risiken vorzubeugen und sie einzudämmen. Finanzaufsichtsbehörden sollen Zugang zu Informationen über IT-bezogene Vorfälle bekommen und sicherstellen, dass Finanzunternehmen die Wirksamkeit ihrer Präventiv- und Belastbarkeitsmaßnahmen beachten und Schwachstellen identifizieren und auflösen.

Nationale Anforderungen

ZWEITES GESETZ ZUR ERHÖHUNG DER SICHERHEIT INFORMATIONSTECHNISCHER SYSTEME (IT-SICHERHEITSGESETZ)

Der Finanzsektor unterliegt als sogenannte kritische Infrastruktur dem Anwendungsbereich des „Zweiten Gesetzes zur Erhöhung der Sicherheit informationstechnischer Systeme“ (IT-Sicherheitsgesetz), das Ende Mai 2022 in Kraft getreten ist. Das Gesetz zielt darauf ab, dass eine Verwendung bestimmter IT-Komponenten durch Betreibende kritischer Infrastrukturen nunmehr untersagt werden kann, wenn anzunehmen ist, dass der Einsatz dieser Komponenten die öffentliche Ordnung oder Sicherheit Deutschlands voraussichtlich beeinträchtigt. Die genauen Voraussetzungen werden durch die „Verordnung zur Bestimmung Kritischer Infrastrukturen“ weiter konkretisiert. Jeder Einsatz kritischer Komponenten kritischer Infrastrukturen ist beim Bundesministerium des Innern anzuzeigen und wird von diesem geprüft.

MINDESTANFORDERUNGEN FÜR DAS RISIKOMANAGEMENT FÜR DEUTSCHE KREDITINSTITUTE (MARISK) UND BANKAUFSICHTLICHE ANFORDERUNGEN AN DIE IT (BAIT)

Die MaRisk beinhalten ausführliche regulatorische Anforderungen vonseiten der Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin) für die Ausgestaltung des Risikomanagementsystems deutscher Kreditinstitute. Die MaRisk sind sogenannte normeninterpretierende Verwaltungsvorschriften und legen den § 25a Abs. 1 des Kreditwesengesetzes verbindlich aus. Ziel ist es, das Risikomanagementsystem angemessen auszugestalten und nachvollziehbar zu dokumentieren. Alle kritischen Bereiche eines Kreditinstituts sollen vom Risikomanagementsystem überwacht werden, u. a. auch der Kreditvergabeprozess.

Da die IT die Basisinfrastruktur für alle Prozesse eines Finanzinstituts bereitstellt, wurden die regulatorischen Anforderungen an die IT in den Bankaufsichtlichen Anforderungen an die IT (BAIT) ausgeführt. Ziel der BAIT ist eine angemessen technisch-organisatorische Ausstattung der IT-Systeme in Banken und legt einen Fokus u. a. auf die Anforderungen der Informationssicherheit und des Notfallmanagements.

Eine risikomanagementspezifische Betrachtung der gesetzlichen Anforderungen findet sich in Kapitel 4.8.2.5.

Weiterhin sind u. a. die folgenden Standards und Regulierungen relevant:

→ Anforderungen an die IT für weitere Finanzdienstleister: siehe Kapitel 4.8.2.4
→ IT-Sicherheit: BSI-Kritisverordnung, IT-Grundschutz des BSI, ISO/IEC 27001/Zertifizierungen)

MINDESTANFORDERUNGEN FÜR VERSICHERUNGEN UND ANDERE FINANZINSTITUTE

Im Nicht-Bankenbereich gehören zu den nationalen Anforderungen auch die Mindestanforderungen für Versicherungsunternehmen (MaGo (Mindestanforderungen an die Geschäftsorganisation von Versicherungsunternehmen) und VAIT (Versicherungsaufsichtliche Anforderungen an die IT)) und Kapitalverwaltungsgesellschaften (KAMaRisk sowie KAIT

(Kapitalverwaltungsaufsichtliche Anforderungen an die IT)) sowie Anforderungen an Zahlungs- und E-Geldinstitute (ZAIT (Zahlungsdiensteaufsichtliche Anforderungen an die IT)).

Grundkonzepte

ANONYMITÄT

Daten sind anonym, wenn sie sich nicht auf eine bestimmte natürliche Person beziehen lassen. Häufig trifft man die Ansicht an, dass dies lediglich bedeute, dass die einzelnen Datensätze keinen eindeutigen Schlüssel mehr enthalten. Ein eindeutiger Bezug kann jedoch auch durch die Einzigartigkeit eines Datensatzes entstehen, etwa wenn dieser aus einer historischen Zeitreihe von Zahlungen besteht. Dadurch ist in vielen Fällen die Anonymisierung von Daten sehr schwierig und die Verwendung anonymisierter Daten für das Training von Modellen nur sehr eingeschränkt möglich.

PERFORMANCE

In der Regel versteht man unter Performance eines KI-Systems oder eines ihm zugrunde liegenden Machine-Learning-Modells so etwas wie die Häufigkeit korrekter Vorhersagen auf einem Testdatensatz. Hier ist zu beachten, dass dieses Maß zwar in gewissem Sinne optimal ist, aber trotzdem kein absolute, sondern eine zufallsbehaftete und zeitlich möglicherweise variable Messgröße darstellt, da der Testdatensatz in der Regel zufällig gewählt wird und sich Daten zeitlich verändern können. Die Performance eine KI-Systems sollte daher auch im Sinne einer zeitlichen Stabilität und Robustheit bewertet werden. Da Transparenz und Erklärbarkeit von ML-Modellen diese befördern, löst sich auch der oft fälschlicherweise propagierte Widerspruch zwischen Performance und Erklärbarkeit auf.

WAHRSCHEINLICHKEIT

Zu selten wird immer noch die zentrale Rolle des Wahrscheinlichkeitsbegriffs bei KI-Anwendungen thematisiert. Dies wird umso problematischer, als dieser Begriff mitnichten klar definiert ist, und der einfache Ansatz, im KI-Umfeld anstelle von klassischen Messgrößen einfach Wahrscheinlichkeiten zu normen, ins Leere läuft. Streng definiert ist der Begriff nur als relative Häufigkeit im Falle unter exakt den gleichen Umständen wiederholten idealisierten Experimenten. Da Machine-Learning-Modelle unter den gleichen Annahmen aufgestellt und genutzt werden, ist dies konzeptionell nicht problematisch, aber die Gültigkeit der Annahme ist in der Regel nicht gegeben, sobald es um die Modellierung menschlichen Verhaltens wie typischerweise in Finanzmodellen geht.

Aus diesem Grunde ist der Bayes'sche Begriff der Wahrscheinlichkeit im Sinne subjektiver, weil von A-priori-Annahmen abhängiger Erwartungen für Anwendungen im Finanzbereich wesentlich geeigneter. Das hat weitreichende Konsequenzen für die Art von Normen, die für KI-Systeme aufgestellt werden, die auf ML-Modellen für Bayes'sche Wahrscheinlichkeiten beruhen.

ABGRENZUNG VON MODELLERGEBNIS / NUTZUNG DER ERGEBNISSE

Für die Bewertung der Nutzung von Modellergebnissen ist es wichtig, konzeptionell zwischen der produzierten Ausgabe des Modells und deren Nutzung für eine (ggf. automatisierte) Entscheidungsfindung zu unterscheiden. Oftmals kann beispielsweise für die Nutzung im Risikomanagement eine von einem Modell prognostizierte Wahrscheinlichkeit direkt verwendet werden, etwa die Ausfallwahrscheinlichkeit im Pricing. Die Korrektheit einer aus dieser Wahrscheinlichkeit abgeleiteten Entscheidung im Einzelfall spielt hier in der Regel keine Rolle.

In anderen Fällen, etwa bei der automatischen Ablehnung von Online-Kreditanträgen, rückt dagegen die Entscheidung im Einzelfall in den Fokus. Diese beinhaltet zusätzlich zum Ergebnis des verwendeten Prognosemodells einen Entscheidungsalgorithmus, der ebenfalls als Bestandteil des KI-Systems zu betrachten ist und eigene Anforderungen und Normen erfüllen muss.

Abgrenzung des Finanzsektors

Der Finanzsektor umfasst alle Unternehmen, die Dienstleistungen bei Geldangelegenheiten als Kerngeschäft haben. Dies schließt u. a. Banken oder Kreditinstitute, Versicherungsgesellschaften, Kapitalverwaltungs- und Kapitalanlagegesellschaften, Zahlungsverkehrsdienstleister oder Börsen ein. Finanzinstitute werden von der europäischen Bankenaufsicht (European Banking Authority) sowie von der BaFin beaufsichtigt. Die BaFin ist in Deutschland auch die Behörde, die die behördliche Erlaubnis zum Betreiben von Finanzinstituten vergibt.

Finanzdienstleistungsinstitute bieten Privatkunden, Firmenkunden, öffentlichen Stellen oder anderen Banken Finanzdienstleistungen an. Dazu gehören u. a.: Verwahrung und Anlage von Finanzen, Finanzierungen von Projekten über Kredite, Versicherungspolicen, Transaktionen, Wertpapiergeschäfte etc.

Künstliche Intelligenz wird im Finanzsektor bereits in vielen Prozessen eingesetzt. Dazu gehören Authentifizierungsver-

fahren von Neukunden, Sprachanalyse und Texterstellung, Risikoanalysen, Credit-Scoring in Kreditvergabeprozessen, KI-basierte Finanzanlagen sowie Erstellung individualisierter Versicherungspolicen.

4.8.2.2 Wissensdatenbanken/Suchmaschinen

Für die KI-Entwicklung im Finanzkontext werden häufig keine reinen Quelldaten verwendet. Stattdessen müssen die Daten vorbereitet und typischerweise aus strukturierten und unstrukturierten Quellen zusammengeführt werden, manchmal werden diese für die spätere Verwendung in Wissensdatenbanken („knowledge graphs“) vernetzt gespeichert. Dies kann in Batchverfahren, aber auch ad hoc über Suchmaschinen erfolgen. Letztere können öffentlich angebotene Websuchmaschinen oder auch selbst entwickelte Suchsysteme sein.

Zentraler Baustein ist hier das sogenannte Entity oder Identity Matching. Für das Matchen von Daten zur gleichen Identität in unterschiedlichen Datensätzen gibt es eine große Vielfalt an Verfahren.

Für die Qualität einer auf den verknüpften Daten basierenden KI ist der Prozess der Datenverknüpfung insbesondere zu Individuen entscheidend und sollte gewisse Standards erfüllen, wenn er nicht vollständig deterministisch ist (z. B. über eindeutige technische Kennzeichen). Der Fokus liegt hier auf der Datenverknüpfung. Das spätere Training von Modellen wird noch nicht betrachtet.

Normen für exaktes Entity Matching
Im einfachsten Fall kann die Zuordnung gleicher Identitäten über einen auf beiden Seiten vorhandenen technischen eindeutigen Schlüssel erfolgen. Abgesehen von Datenqualitätsproblemen ist dieses unkritisch und kann ohne weitere Kontrolle angewendet werden.

In einigen Fällen werden Schlüssel verwendet, die nicht technischer Natur sind, sondern typischerweise aus einer Kombination von Datensatzattributen bestehen, von der man annimmt, dass sie eine eindeutige Referenz auf die Entität herstellt, etwa Name, Vorname, Geburtsdatum und Wohnort bei natürlichen Personen, § 111 OWiG Gesetz über Ordnungswidrigkeiten (OWiG) § 111 Falsche Namensangabe (s. [387]) etwa enthält entsprechende Attribute. Auch diese Verfahren werden in der Regel ohne weitere Kontrolle verwendet und als exakt angesehen, auch wenn sie eine gewisse Fehlerwahrscheinlichkeit haben.

Je nach Kritikalität einer KI-Anwendung sollten gewisse Standards dafür definiert werden, welche Datenquellen und Attribute als für ein exaktes Matching ausreichend angesehen werden können, ggf. unterschieden nach natürlichen und juristischen Personen.

Normen für probabilistisches Entity Matching
Im Bereich der KI, insbesondere basierend auf Big Data häufig anzutreffen sind Matchingverfahren, die Machine-Learning-Modelle verwenden, um mit einer gewissen (hohen) Wahrscheinlichkeit identische Entitäten in unterschiedlichen Datenquellen zu finden. Da diese sogenannten Fuzzy Matches häufig weiterverwendet werden, ist eine Normung von deren Qualität besonders wichtig.

Daher sollten Standards dafür gelten, welche Datenquellen für Anwendungen unterschiedlicher Kritikalität überhaupt genutzt werden dürfen, und ab welcher Kritikalität ungenaue Matches generell nicht erlaubt sind, etwa bei Hochrisikoanwendungen wie der Kreditvergabe.

Zudem sollte es klare Regeln für Angaben zur Genauigkeit von Matches geben sowie zur späteren Reproduzierbarkeit von Datenverknüpfungen, die ggf. korrigiert wurden. Hier wäre ein Stufensystem für Identitätsmatching (eindeutige IDs, hohes Konfidenzniveau, mittleres Konfidenzniveau etc.) denkbar.

Normen für Suchmaschinen
Grundsätzlich beruhen auch Suchmaschinen auf dem Prinzip des ungenauen Entity Matchings. Jedoch findet der Matchingvorgang hierbei ad hoc statt und der Nutzende bestimmt in der Regel selbst, welche Suchergebnisse korrekte Matches darstellen. Trotz dieser Kontrollmöglichkeit durch einen menschlichen Nutzenden (sofern ein solcher involviert ist), ist auch hier die Qualität des modellbasierten Matchings entscheidend für die korrekte Zuordnung von Entitäten. Damit sollten hier die gleichen Normen greifen wie im Falle einer Batchverabeitung.

Zusätzlich ergeben sich zwei weitere Herausforderungen bei der Verwendung von Suchmaschinen: Da diese ad hoc genutzt werden, ändert sich das verwendete Matchingmodell ständig, was die Reproduzierbarkeit der Zuordnungen gefährdet. Zum anderen verändert auch die Nutzeraktivität selbst insbesondere bei kommerziellen Suchmaschinen das Suchergebnis. Gehäufte Auskunftsanfragen können in bestimmten Fällen den Kreditwürdigkeitsscore beeinflussen.

Um den Problemen der Veränderlichkeit von Suchergebnissen und der Rückkopplungsaffekte von Suchanfragen zu begegnen, sollten Normen für die Reproduzierbarkeit von Suchergebnissen und die Transparenz über Rückkopplungseffekte aufgestellt werden. Unter Umständen dürfen in bestimmten Kontexten nur Suchmaschinen oder -dienste verwendet werden, bei denen die Anfrage das Ergebnis künftiger Anfragen nicht beeinflussen kann.

Normen für Einfluss- und Kontrollmöglichkeiten
Neben der Normung für die technische Qualität und Transparenz von exakten und probabilistischen Matches könnten Normen für persönliche Kontrollmöglichkeiten für die unter der eigenen Identität verknüpften Daten sinnvoll sein.

Hierzu könnten Standardzugänge für die Kontrolle, Einwilligung zur Nutzung sowie Korrektur probabilistisch verknüpfter Daten gehören. Das sollte Informationen zur Herkunft der Matches oder auch die Festlegung eines gewünschten Konfidenzniveaus für das Matching etwa bei Suchmaschinen umfassen.

Eine Möglichkeit, dies zu erreichen, wäre die Verwendung von Identitätsnormungssystemen wie der SSI (self sovereign identity) oder einer Bank-ID wie in den nordischen Ländern[94]. Dieses könnte auch Normen für den Informationsaustausch zwischen den Plattformen inklusive der Mitbestimmung der Betroffenen bei der Portabilität beinhalten.

4.8.2.3 Individualisierung / Fairness

Finanzdienstleistungen beziehen sich ihrer ökonomischen Natur nach immer direkt oder indirekt auf den Menschen. Daher ergibt sich auch für KI-Anwendungen im Finanzbereich die Herausforderung, dass die ihnen in aller Regel zugrunde liegenden Machine-Learning-Modelle menschliches Verhalten statistisch abbilden sollen. Menschen sind jedoch in ihrem Verhalten im Gegensatz zu Maschinen, autonomen Fahrzeugen und sogar in gewissem Grade den medizinisch relevanten biologischen Prozessen im menschlichen Körper hochgradig individuell, unterschiedlich und schwer vorhersagbar. Das stellt die Anwendung statistischer Modellierung wie das Machine Learning für das Vorschlagen fairer Entscheidungen und Prognosen vor besondere Schwierigkeiten, aus denen sich spezifische Normungsbedarfe ergeben. Diese sollen im Folgenden näher beleuchtet werden.

Fairness
Fairness wird häufig als Operationalisierung von Nicht-Diskriminierung gesehen, geht jedoch oft auch darüber hinaus. In der Fachliteratur werden über 20 verschiedene grundlegende Methoden diskutiert Fairness zu messen, darüber hinaus gibt es von den meisten mehrere Variationen. Viele Fairnessmaße verfolgen unterschiedliche Philosophien von Fairness und stehen damit im Widerspruch zueinander. Wird das Ziel verfolgt, mehreren Fairnessphilosophien gerecht zu werden, gibt es verschiedene Möglichkeiten. Es können z. B. Mindestgrenzwerte für mehrere, einander widersprechende Maße definiert werden, oder mehrere Maße können gewichtet herangezogen werden, um einen Gesamtfairnesswert zu berechnen. Welche Fairnessmaße unter welchen Umständen angemessen sind und welche nicht, ist bisher nur für wenige Maße wissenschaftlich fundiert ermittelt. Es sind Anwendungen bekannt, bei denen die Wahl eines unangemessenen Maßes zu gesellschaftlichen Schäden geführt hat. Die meisten Fairnessmaße basieren auf Qualitätsmaßen (es wird die Qualität zweier Teilgruppen, die sich nur in einem sensitiven Attribut unterscheiden, auf irgendeine Art verglichen). Ein wichtiger Aspekt von Fairness in Bezug auf ML besteht darin, dass der Begriff immer in Bezug auf das Zielkriterium des Modells betrachtet wird. Ist dieses objektiv gegeben, sollte Fairness als Ziel der Entwicklung immer nachgelagert betrachtet werden.

Transparenz und Erklärbarkeit
Häufig werden Machine-Learning-Modelle als Blackboxes bezeichnet, worunter fälschlicherweise verstanden wird, die Ergebnisse solcher Modelle seien nicht im Detail nachvollziehbar oder gar zufällig. Beides ist in aller Regel falsch. Die Ergebnisse ergeben sich als mathematisch eindeutige Formeln, und zufällig ist allein die Richtigkeit der Vorhersage, nicht aber die deterministisch aus den Eingangswerten zu berechnende Vorhersage selbst. Ein Mangel an Transparenz entsteht nur durch die ggf. komplexen Formeln, die ein ML-Modell beschreiben, und der Tatsache, dass diese nicht aus tieferen Zusammenhängen und einer logischen Form abgeleitet sind, sondern eine allgemeine Formel über freie Parameter an Beispieldaten angepasst wurde. Es fehlt daher die Erklärung ex ante durch die zugrunde liegende Theorie, die nur durch Analyse des Modells im Nachgang aufgestellt werden kann.

94 s. auch https://www.crefotrust.de/

Hierbei ist entscheidend, dass Erklärbarkeit zwangsläufig auf (zumindest approximativer) Kausalität beruhen muss. Eine Erklärung, die nicht kausal und damit unter allen Umständen korrekt ist (also nicht nur zur Rechtfertigung ex post in Einzelfällen taugt), ist zumindest unvollständig und wird damit dem ureigensten Anspruch an eine Erklärung nicht gerecht. Daher sollte der Begriff Erklärbarkeit zugunsten des Transparentbegriffs vermieden werden und nicht in einer Normung auftauchen.

Welche statistischen Aussagen dürfen auf das Individuum übertragen werden?
Die Diskussion um Fairness nimmt häufig Ungleichbehandlungen in Bezug auf ein grobes Attribut in den Blick, z. B. das Geschlecht oder die Ethnie.

Allerdings stellt sich die Frage nach Fairness auch schon umgekehrt auf einer ganz elementaren Ebene: Von welchen Gruppenattributen aus ist es überhaupt fair, von der Gruppe auf das Individuum zu schließen? Alle statistischen, nicht kausalen Modelle, also auch alle ML-Modelle, aber auch alle regelbasierten Modelle (auch bei diesen ist die Richtigkeit der Entscheidung oft eine Zufallsgröße) entscheiden auf Grundlage von möglichst relevanten, aber niemals das Individuum (kausal) vollständig abbildenden Eingangsdaten. Das ist unvermeidlich und notwendig und sollte auch nicht infrage gestellt werden.

Auch hier ist wieder der Anwendungsfall von entscheidender Bedeutung. Insbesondere, wenn es um (Einschränkung von) Grundrechte(n) geht, sollte eine Möglichkeit bestehen, den individuellen Fall zu betrachten. Das bedeutet insbesondere, zumindest alle verfügbaren relevanten individuellen Faktoren einzubeziehen. Ein aktuelles Beispiel ist die Beurteilung einer Gefährdung durch die Infektion mit dem Coronavirus, sei es im Kontext gesellschaftlicher Einschränkungen oder, derzeit rein hypothetisch, etwa im Zusammenhang mit Versicherungstarifen. Dabei sollte das Recht bestehen, etwa den Nachweis von Antikörpern (bzw. anderer relevanter Faktoren) in das Entscheidungsmodell oder die Einzelentscheidung einzubeziehen. Eine Einschätzung der Gefährdung ohne Berücksichtigung eines derart relevanten Faktors kann nicht als fair gelten, zumindest, solange daraus kein direkter Nachteil für andere entsteht.

Das Problem lässt sich häufig dann beobachten, wenn sich die nach dem Entscheidungskriterium gebildeten Gruppen bezüglich des Zielkriteriums statistisch nur wenig unterscheiden, also in allen Gruppen eine große Varianz zu beobachten ist und sich die Wertebereiche überdecken. Zum Beispiel ist das Einkommen bei Frauen und Männern zwar im Mittel signifikant verschieden, die Verteilungen überlappen sich aber stark. Würde das Geschlecht für die Prognose des Gehalts und damit der Kreditwürdigkeit benutzt, wäre das für das Individuum nicht zu rechtfertigen. Technisch würde sich dies durch eine hohe Fehlerquote in der Vorhersage des Modells bemerkbar machen und sollte daher bei einer guten Modellvalidierung auffallen. Allerdings ist den Anwender*innen in einigen Fällen nicht bewusst, dass es sich auch bei der Verwendung von einfachen Mittelwerten für die Entscheidungsfindung um ein Prognosemodell handelt, etwa bei der oben erwähnten Gefährdungsbeurteilung. Für Kreditwürdigkeitsscores ist die Bildung von Kundengruppen sehr relevant. Hier ist jedoch anzumerken, dass Letztere in der Regel lediglich den Risikopreis und nicht die Kreditentscheidung beeinflussen und daher eher gerechtfertigt sind.

Um die faire Anwendung von Modellen für Gruppen menschlicher Individuen sicherzustellen, werden daher Normen benötigt, die klar definieren, wann von der vollständigen Erfassung aller wesentlichen Einflussfaktoren und, im Falle von Entscheidungssystemen, einer hinreichenden statistischen Trennung der Gruppen auszugehen ist.

Verwendungshinweis

Vollständigkeit von Trainingsdaten
Machine-Learning-Modelle sollen, wie alle Modelle, die Wirklichkeit in allen relevanten Aspekten approximieren. Sie sind also genau dann vollständig, wenn sie alle relevanten Zusammenhänge abdecken. Im Gegensatz zu naturwissenschaftlichen Modellen sind die grundlegenden Gesetzmäßigkeiten, deren Parameter aus den Daten gelernt werden sollen, bei Machine-Learning-Modellen z. B. für Probleme der KI oder für Finanzanwendungen in der Regel nicht bekannt. Sie müssen selbst aus den Daten gelernt werden, wofür in der Regel ein Kontinuum an Beispieldaten nötig wäre. Das heißt, dass Vollständigkeit immer nur pragmatisch definiert werden kann, aber niemals exakt, etwa bei der Frage, welche Faktoren und Beispieldatenpunkte die Rückzahlungswahrscheinlichkeit eines Kredits in allen denkbaren Situationen korrekt bestimmen lassen.

Nichtdiskriminierung bei Finanzentscheidungen
Als Pendant zur im ersten Unterkapitel betrachteten Problematik der unfairen Gleichbehandlung von Mitgliedern einer Gruppe schauen wir hier auf die häufiger diskutierte Frage unfairer Entscheidungen durch ungerechtfertigte Diskri-

minierung, also Ungleichbehandlung. Dabei ist der Begriff der „Gleichbehandlung“ problematisch und soll hier nicht thematisiert werden. Normen für eine Definition von Gleichbehandlung sind eine wichtige Voraussetzung für die Operationalisierung von Fairness.

Vorbetrachtung: Fairness bei KI-Anwendungen als nachgelagertes Konzept

Fairness ist im Kontext KI im Finanzsektor immer ein nachgelagertes Konzept, da die KI-Anwendung zunächst für ihren eigentlichen Zweck, häufig eine statistisch korrekte Risikobeurteilung, realisiert werden muss. Erst danach kann man sinnvoll über die Fairness der Anwendung sprechen und diese sicherstellen. Dies lässt sich wie folgt begründen.

Fairness als unscharfer Begriff

Fairness oder Gerechtigkeit sind keine fest definierten Begriffe. Typischerweise versteht man darunter, dass bestimmte Gruppen

(a) „gleich“ behandelt werden oder
(b) unter Berücksichtigung objektiv relevanter Kriterien gleichbehandelt werden.

Bereits diese beiden Sichtweisen (a) und (b) widersprechen sich, und es gibt viele mathematische Definitionen von Fairness, von denen man beweisen kann, dass sie niemals gleichzeitig erfüllt sein können. Fairness sollte daher als Eigenschaft eines KI-Systems weitestgehend nachgelagert betrachtet werden.

Die Aufgabe von Machine Learning und anderer Methoden, datengetriebene Entscheidungssysteme zu bauen

Typischerweise denken wir bei KI-Systemen in Bezug auf Fairness an ein System, das für gegebene Eingangsdaten eine(n) Entscheidung(svorschlag) in Bezug auf eine natürliche Person liefert, ggf. zusammen mit einer Wahrscheinlichkeit für die Richtigkeit der Entscheidung. Beruht der Vorschlag auf menschendefinierten Regeln, können diese direkt auf ihre Fairness hin betrachtet werden. Für uns relevanter sind Systeme, bei denen die Zuordnung von Eingangsdaten zu einer Entscheidung durch eine mathematische Funktion (ein „Machine-Learning-Modell“) geschieht, die anhand von Beispiel-/Trainingsdaten so ermittelt wurde, dass sie für neue, unbekannte Eingabedaten „bestmögliche“ Ausgaben produziert (hier kommt die Expertise des Modellierers zum Tragen).

Die optimale Abbildung der Trainingsdaten durch das ML-Modell passiert in der Regel ohne die Berücksichtigung von Fairnessaspekten. Das ist wichtig, da jeglicher Eingriff in den Lernprozess die oben genannte Aufgabe des Machine Leanings teilweise korrumpieren würde – allein schon vor dem Hintergrund der oben begründeten Unschärfe des Fairnessbegriffs.

Es gibt jedoch Methoden, die Fairness bei gleicher Vorhersagequalität unter Umständen verbessern können, sofern ein konkretes Fairnessmaß gegeben ist und das gefundene Modell dieses tatsächlich verletzt. Man unterscheidet dabei zwischen

→ **Preprocessing:** z. B. Modulation der Datensets auf Fairness bei gleichem Informationsgehalt
→ **Inprocessing:** z. B. Fairness als paralleles Lernziel (Teil der Zielfunktion)
→ **Postprocessing:** z. B. Ausgaben mit hoher Unsicherheit werden zur Fairnessoptimierung manipuliert

In der Praxis der Finanzanwendungen sind diese allerdings weniger relevant, da Fairness explizit realisiert werden muss.

Praktische Implementierung von Fairness

Wenn das KI-System mit dem enthaltenen ML-Modell fertig trainiert ist, muss eine Überprüfung erfolgen, inwiefern die Entscheidungen „fair“ nach einem festzulegenden Maß sind. Wird mangelnde Fairness beobachtet, kann dies mehrere Ursachen haben.

Finden sich Fehler im Modell, sollte das Modell geprüft werden, insbesondere im Hinblick auf seine Transparenz und Kausalität, auch im Hinblick auf die Vollständigkeit der betrachteten Merkmale („weitere/andere Spalten im Trainingsdatensatz“).

Ist bereits in den Trainingsdaten ein Mangel an Fairness zu beobachten und ist das Modell valide, gibt es offenbar objektive Ursachen für die Ungleichbehandlung. Möglicherweise sind die Trainingsdaten nicht repräsentativ für alle Gruppen, dann kann versucht werden, diese zu ergänzen („weitere/andere Zeilen im Trainingsdatensatz“). Falls dies nicht der Fall ist, kann eine definitiv objektive, aber nicht gewollte Ungleichbehandlung ex post ausgeglichen werden, indem die Ergebnisse über das relevante Gruppenattribut gemittelt werden. Dazu muss das entsprechende Attribut, z. B. das Geschlecht, bekannt und im Modell explizit enthalten sein.

Außerdem gibt es Fälle, in denen kein objektives Zielkriterium im Trainingsdatensatz verwendet wurde (z. B. echte Kreditausfälle), sondern menschliche Entscheidungen, die bereits verzerrt waren. Dies kann auch implizit dadurch entstehen,

dass z. B. bestimmte Kreditanträge von vornherein nicht angenommen werden und daher ihr Ausfallverhalten nicht beobachtet werden kann. Hier verwendbare Trainingsdaten zu erstellen, ist sehr schwierig.

Alle diese Maßnahmen greifen grundsätzlich nicht in den Modellierungsprozess ein, sondern betreffen entweder die Bereitstellung der Daten oder die Ex-post-Behandlung der Ergebnisse.

Ein weiterer wichtiger Aspekt ergibt sich daraus, dass Fairness nicht separat diskutiert werden kann, sondern eine Abhängigkeit zu den Themen Transparenz/Erklärbarkeit und Vollständigkeit der Trainingsdaten besteht. Insbesondere besteht kein Widerspruch oder „Trade-off" zwischen Fairness und Performance oder Fairness und Transparenz, ebenso wenig wie zwischen Fairness, Transparenz und Performance. Performance und Transparenz sowie Performance und Auswahl der Trainingsdaten bedingen einander und müssen gemeinsam optimiert werden.

Die Beurteilung der Fairness dagegen hängt kausal davon ab, dass Performance, Transparenz und korrekte Datenauswahl gegeben sind. Allerdings kann aus der Beobachtung einer Unfairness der Entscheidungen wie oben beschrieben geschlossen werden, dass es hier möglicherweise Defizite gibt. Dies muss aber nicht der Fall sein, ein direkter Eingriff in die Modellerstellung nur zur Erreichung von Fairness sollte nicht erfolgen.

Für die Normung ergeben sich daraus die Anforderungen, zum einen Fairnessmaße zu definieren, aber zumindest im Kontext von Finanzanwendungen mit Risikobezug keine Anforderungen zu stellen, die etwa die unverfälschte Abbildung der beobachteten Daten zugunsten von Fairness a priori gefährden würden.

Fairnessbegriffe

Warum ist Fairness bei Finanzentscheidungen von besonderer Relevanz?

JURISTISCHER FAIRNESSBEGRIFF

Speziell in den USA haben sich für die Versicherungsbranche die verschiedenen juristischen Fairnessbegriffe „Disparate Treatment" und „Disparate Impact" etabliert. Es muss definiert werden, was eine Proxy Diskrimination darstellt, also eine (un-)beabsichtige Diskriminierung durch eine Stellvertretervariable, wie z. B. Postleitzahl statt ethnischer Herkunft, bei der ein kausaler Zusammenhang oder eine hohe Korrelation mit dem verbotenen Merkmal vorhanden ist. Zusätzlich muss definiert werden, was unter risikoadäquater Differenzierung verstanden wird und was eine Diskriminierung darstellt, siehe [388].

MATHEMATISCHER FAIRNESSBEGRIFF

Neben juristischen Fairnessbegriffen gibt es auch mathematische Fairnessbegriffe. Eine Übersicht dazu findet man in [389]. Dort wird zwischen individuellen und Gruppen-Fairnessmaßen unterschieden (s. Auflistung unten).

Für den Einsatz eines mathematischen Fairnessmaßes müssen Auswahlkriterien festgelegt werden. Außerdem müssen Toleranzen festgelegt werden, da diese Maße im Allgemeinen nicht exakt erfüllbar sind. Außerdem muss klargestellt werden, dass mehrere Fairnessmaße sich im Allgemeinen nicht gleichzeitig erfüllen lassen. So lassen sich z. B. Independence, Separation und Sufficiency nur erfüllen, wenn die Daten in sich bereits „fair" sind (siehe [390], Kapital 2 etwa Proposition 2 für solche Aussagen).

FAIRNESS VON DATEN

Auch Daten als solche können schon unfair sein. So können etwa bestimmte Gruppen über- oder unterrepräsentiert sein. Auch können bestehende Benachteiligungen durch Verwendung von Daten weiterbestehen.

Es gibt Ansätze, sogenannte Debiasing-Techniken, Daten vor dem eigentlichen Machine-Learning so aufzubereiten, dass der Bias entfernt wird. Beispiele dafür sind Disparate Impact Remover oder Orthogonal Predictors (siehe [389]).

DEFINITION VON SCHÜTZENSWERTEN MERKMALEN

Wie werden schützenswerte Merkmale definiert? Wie ist mit Grenzfällen umzugehen wie z. B. Postleitzahl, Bildungsgrad, Score für Bonität (siehe dazu auch [388])?

Es gibt mehr als 20 Maße, die entweder auf einem oder mehreren Qualitätsmaßen (distributive Fairness, Gruppenfairness) oder auf einem Distanzmaß (individuelle Fairness) basieren. Darüber hinaus gibt es für die meisten Fairnessmaße eine Vielzahl an Variationen, Fairnessmaße können miteinander kombiniert werden und es ist möglich, eigene Fairnessmaße zu entwickeln/definieren.

Konflikte zwischen Fairnessmaßen

Bei der Wahl einer Fairnessoperationalisierung gibt es meist kein definitives Richtig (nur eventuell ein definitives Falsch). Verschiedene konkrete Anwendungen sowie verschiedene

Beteiligte können eine angemessene Wahl beeinflussen, jedoch können auch unterschiedliche Beteiligte unterschiedliche Ziele verfolgen und damit andere Maße bevorzugen. Insofern bieten sich ähnliche Fragestellungen sowie Lösungsansätze bei der Wahl einer Erklärung an.

Verschiedene Fairnessmaße repräsentieren verschiedene Vorstellungen von Fairness, viele davon können nicht gleichzeitig optimiert werden, da sie zu einem gewissen Grad im Widerspruch zueinander stehen. Wird gezielt für ein bestimmtes Fairnessmaß optimiert, werden damit die Ergebnisse anderer Fairnessmaße zwangsläufig reduziert. Dadurch kann Diskriminierung nach dem Verständnis der reduzierten Maße sogar erhöht werden. Darüber hinaus ist es möglich, gezielt Fairnessmaße anzugeben, die zwar einen sehr hohen Wert erreichen, sich jedoch nicht unbedingt mit bestimmten Fairnessvorstellungen decken. Deshalb ist es sinnvoll, die Wahl von Fairnessmaßen zu begründen, wenn die resultierenden Werte kommuniziert bzw. transparent gemacht werden. Wenn eine externe Partei (z. B. eine Prüfbehörde) kontrollieren möchte, ob die transparent gemachten Werte tatsächlich erreicht werden, bietet sich die Umsetzung von Nachvollziehbarkeitsmechanismen wie z. B. Assurance Cases an.

Als Prozesse zur Wahl geeigneter Fairnessmaße bieten sich Design thinking, Specification Workshop, Assurance Cafés oder die Einbeziehung von Betroffenen an. Die Wahl fällt potenziell auf konkurrierende Fairnessmaße, die nicht gleichzeitig optimiert werden können. Es gibt verschiedene Möglichkeiten, mit dieser Konfliktsituation umzugehen:

→ Gewichtung: Für jedes Maß wird eine „Wichtigkeit“ festgelegt, z. B. in Form eines Faktors (man spricht von einer Gewichtung). Addieren sich die Gewichte zu 1 auf, entsprechen sie einem Prozentwert, den das jeweilige Maß zur Gesamtbewertung beiträgt. Für diese Gesamtbewertung wird ein (argumentierter) Mindestwert festgelegt. Wird dieser Wert erreicht oder überschritten, gilt das System als fair.

→ Thresholds: Für jedes Maß wird ein (argumentierter) Mindestwert festgelegt. Wird für jedes Maß der spezifizierte Wert erreicht oder überschritten, gilt das System als fair.

Einbettung in den sozialen Prozess

Die Wahl von Fairnessmaßen und Konfliktlösungsstrategien kann willkürlich gefällt werden, im schlimmsten Fall sogar gezielt so, dass ein System möglichst gute Werte erreicht, unabhängig davon, wie aussagekräftig die gewählten Metriken tatsächlich sind. Damit wird unfaires (diskriminierendes) „Verhalten“ nicht nur nicht entdeckt, sondern sogar gezielt verschleiert. Umso wichtiger ist es, wie bereits erläutert, fundierte Entscheidungen bei der Wahl zu treffen sowie die Argumentation zu dokumentieren. Der Mehrwert besteht nicht nur in einer guten Argumentationsbasis im Falle einer Prüfung (z. B. Zertifizierung oder Rechtsstreit), sondern auch in der Option, sich gegenüber einem Kunden oder Betroffenen rechtfertigen zu können (z. B. als CDR-Maßnahme aber auch aus Marketinggründen).

Die Kommunikation von Fairnessmaßen gegenüber Nicht-Expert*innen kann sich schwierig gestalten, da ohne spezifische Vorkenntnisse ein gutes mathematisches Grundverständnis notwendig ist. Es ist also nicht ausreichend, die Wahl der Fairnessmaße und die jeweils erreichten Werte zu nennen, sondern diese müssen auch kunden- oder betroffenengerecht erläutert werden. Für viele grundlegende Maße liegen ausreichend wissenschaftliche Grundlagen vor, um bedarfsgerechte Erläuterungen in Normen verankern zu können.

Über die Wahl der Maße und die erreichten Werte hinaus spielt es auch eine Rolle, wie diese Informationen verwendet werden. Sie können im Rahmen eines KI-Trainingsprozesses als (zusätzliche) Zielfunktionen verwendet werden, im Rahmen eines Qualitätssicherungsprozesses als Mindestanforderung (requirement engineering) oder sogar als Kontrollvorgaben im Rahmen regelmäßiger automatischer Prüfungen im Einsatz. Die Wichtigkeit der Kommunikation über die Verwendung von Fairnessmaßen lässt sich an einem einfachen Beispiel verdeutlichen: Wie erläutert basieren die meisten Fairnessmaße auf einer Grundwahrheit (ground truth). Diese Grundwahrheit entspricht jedoch nicht notwendigerweise den realen Gegebenheiten, in denen ein KI-System eingesetzt wird. Das bedeutet, selbst wenn Fairnessmaße unter Laborbedingungen auf Basis von Testdaten ausreichend gute Werte erzielen, sagt dies nicht notwendigerweise etwas über die Fairnessperformance im realen Einsatz aus. Werden nun die gewählten Fairnessmaße im realen Einsatz regelmäßig automatisch berechnet und wird im Falle der Nichterfüllung ein Fehlverhalten gemeldet, entsteht ein viel wirkungsvollerer Qualitätssicherungsprozess.

Übersicht Fairnessmaße

Individuelle Fairnessmaße

→ Fairness By Awareness [391]

→ Fairness Through Awareness [393]

→ Counterfactual Fairness [394]

→ Controlling for the Protected Variable

Gruppen-Fairnessmaße
→ Demographic Parity (or Statistical Parity) (Independence)
→ Disparate Impact (the Four-Fifths Rule)
→ Conditional Demographic Parity
→ Separation
→ Sufficiency
→ Demographic Parity (or Statistical Parity) (Independence)
→ Disparate Impact (the Four-Fifths Rule)
→ Conditional Demographic Parity
→ Separation
→ Sufficiency

4.8.2.4 Informationssicherheit

IT-Sicherheit[95] als wichtiger horizontaler Baustein ist bereits in der ersten Ausgabe der Normungsroadmap KI (NRM KI Ausgabe 1) ein prominentes Thema. Trotz des in dieser 2. Ausgabe der Normungsroadmap KI (NRM KI Ausgabe 2)) erstmals aufgegriffenen vertikalen Bereichs der Finanzdienstleistungen und dess Unterbereich der Informationssicherheit[96] behalten alle Grundanforderungen auch hier weiter Bestand. Dabei ist das Ziel der IT-Sicherheit mit ihrem Bezug auf den Einsatz von Informationstechnik und der Schnittmenge mit der Informationssicherheit weiter der maximale Schutz gegenüber Bedienungsfehlern, technischem Versagen, katastrophenbedingten Ausfällen und absichtlichen Manipulationsversuchen ([63], S. 99). In die technische Betrachtung sollten dabei erweiternd auch unabsichtliche Manipulation als Teil der Fehlbedienung sowie neben den katastrophenbedingten Ausfällen ebenso Beschädigungen einbezogen werden. Die Schutzziele der Informationssicherheit, Vertraulichkeit, Integrität, Verfügbarkeit und Authentizität sind auch hier Grundlage für die weiteren Betrachtungen, siehe Abbildung 45. Der Aspekt der Sicherheit steht dabei im Vordergrund. Die im Besonderen auch in der Finanzindustrie wichtige Datenqualität ist für den erfolgreichen Einsatz von KI, aber auch für die Betrachtung der Informationssicherheit weiterhin essenziell. Allgemeine Maßnahmen zur Datensicherheit und zu Vertrauensstufen hinsichtlich der Datenqualität von Inputdaten sind bereits in der NRM KI Ausgabe 1 umfangreich beschrieben. Es ist darüber hinaus wichtig, Mechanismen zu definieren, die eine Aussage über die Datenqualität sowie die in Verbindung mit der Qualität dieser Daten stehenden Einsatzmöglichkeiten treffen. Die Verwendung der Daten ist in Abhängigkeit zur Qualität der Daten risikoorientiert auszugestalten. Risikoreiche Geschäftsprozesse sollten daher höheren Datenqualitätsansprüchen unterliegen als weniger risikoreiche Geschäftsprozesse.

Die Risikobetrachtung zur IT-Sicherheit, zu Angriffen sowie einer Auswahl an Verteidigungsmechanismen werden für den allgemeinen Fall ebenso thematisiert. Weiterhin sind in der NRM KI Ausgabe 1 Rechercheergebnisse zu Gesetzen, Normen und Standards naturgemäß mit Stand 2020 zu finden, die in der NRM KI Ausgabe 2 auf den aktuellen Stand gebracht wurden.

Vertrauen ist die Grundlage jeder Art von Geschäftsbeziehung. Daher besteht grundsätzlich, aber in der Finanzindustrie im besonderen Maße, die Herausforderung, Vertrauen in die IT-Sicherheit und Informationssicherheit des Anbietenden als auch in das KI-System herzustellen. Aus diesem Grund ist es einerseits notwendig, wie bereits in der NRM KI A1 ausgeführt, Überprüfbarkeit, Erklärbarkeit sowie einen Konformitätsnachweis (wie beispielsweise im AI Act-Entwurf für Hochrisikosysteme gefordert) sicherzustellen. Andererseits sollten bereits ganz am Anfang des Entstehungsprozesses einer KI alle relevanten Stakeholder nach Betroffenheit in die individuelle Risikobetrachtung der Informationssicherheit einbezogen werden. Zu den Stakeholdern gehören insbesondere der Vorstand eines Unternehmens, der Informationsrisikobeauftragte sowie das Risikomanagement. In jedem Finanzinstitut gibt es weiterhin für KI-Systeme und Daten direkte Verantwortliche, die die Risiken für die Institute selbst, aber insbesondere auch für die Kundinnen und Kunden identifizieren, überwachen und mitigieren. Hier sollte ein besonderer Fokus auf vulnerable Verbrauchergruppen gesetzt werden, insofern diese in dem jeweilig zu betrachtenden KI-System direkt oder indirekt eine Rolle spielen.[97]

95 IT-Sicherheit bezeichnet einen Zustand, in dem die Risiken, die beim Einsatz von Informationstechnik aufgrund von Bedrohungen und Schwachstellen vorhanden sind, durch angemessene Maßnahmen auf ein tragbares Maß reduziert sind. IT-Sicherheit ist also der Zustand, in dem Vertraulichkeit, Integrität und Verfügbarkeit von Informationen und Informationstechnik durch angemessene Maßnahmen geschützt sind. [BSI (2022)]

96 Informationssicherheit hat den Schutz von Informationen als Ziel. Dabei können Informationen sowohl auf Papier, in Rechnern oder auch in Köpfen gespeichert sein. Die Schutzziele oder auch Grundwerte der Informationssicherheit sind Vertraulichkeit, Integrität, Verfügbarkeit und Authentizität. Viele Anwender*innen ziehen in ihre Betrachtungen weitere Grundwerte mit ein. [BSI (2022)]

97 Vorschlag für die Definition: Verbraucherin und Verbraucher im Sinne des Digitalen Verbraucherschutzes des BSI ist jede natürliche Person, der bei der privaten Nutzung von Produkten, Dienstleistungen oder Anwendungen ein IT-Sicherheitsrisiko entsteht oder entstehen könnte. [BSI (2021c)]

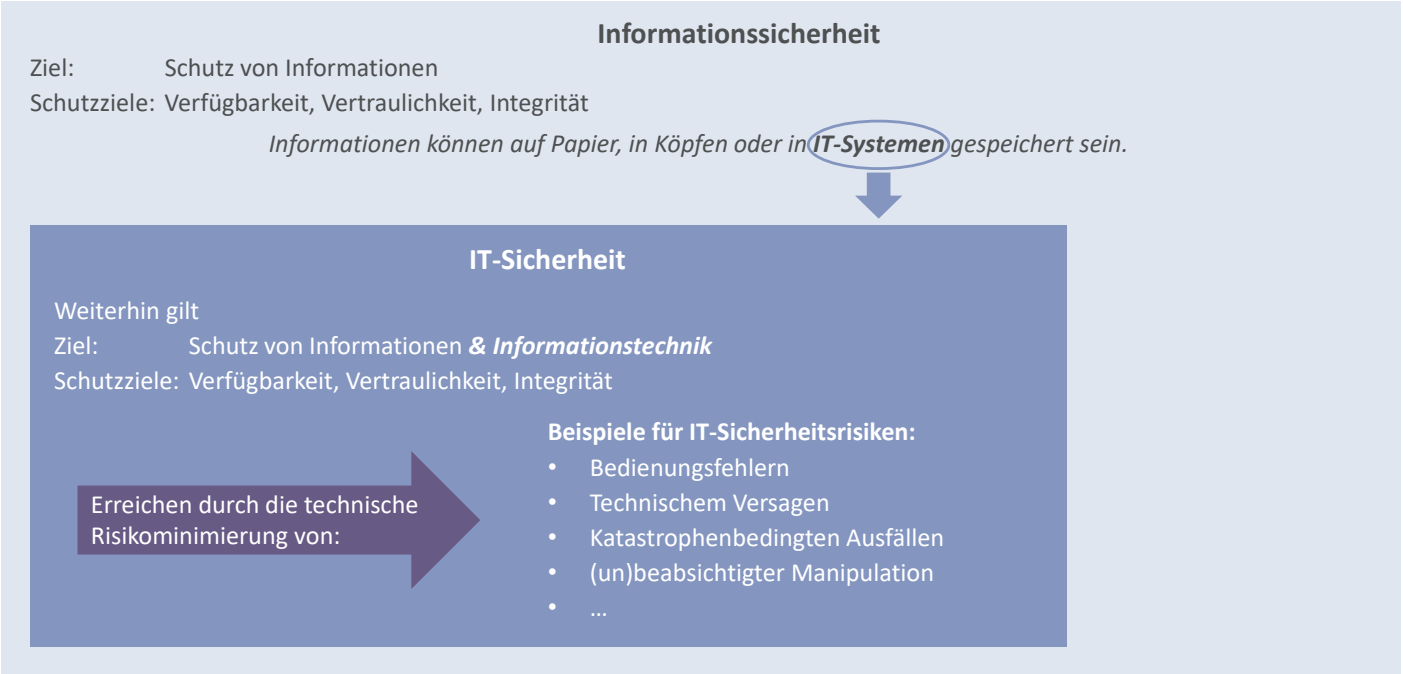


Abbildung 45: Informationssicherheit (Quelle: Arbeitsgruppe Finanzdienstleistungen)

Besondere Anforderungen an die Informationssicherheit

Eines der zentralen Schlüsselelemente der Geschäftsmodelle in der Finanzwirtschaft ist Vertrauen – Vertrauen in die organisatorische und materielle Leistungsfähigkeit eines Finanzinstituts sowie die Sicherheit (unter Einbezug sämtlicher Schutzziele) der Daten. In der Regel sind Finanz- und Versicherungsprodukte abstrakt, nicht direkt greifbar und für Nichtfachleute in deren Komplexität schwer nachzuvollziehen. Es werden dabei besonders schützenswerte und zum Großteil auch personenbezogene (Finanz-)Daten verarbeitet. So ergibt sich die besondere Situation, dass Kundinnen und Kunden ihren Finanzinstituten auf mehrfache Weise vertrauen müssen, beispielsweise bei der Berücksichtigung der individuellen Lebensumstände bei individuellen Anlageentscheidungen. Finanzinformationen in falscher Hand und/oder falsche Bonitätsentscheidungen können existenzbedrohende Auswirkungen haben, sowohl für Kunden als auch für Finanzinstitute. Finanzdienstleistern kommt daher eine besondere Verantwortung zu. Es muss darauf vertraut werden können, dass das Dienstleisterversprechen erfüllt wird, dass die dahinter liegenden IT- und KI-Systeme richtig funktionieren und dass angemessene Informationssicherheitsmaßnahmen eingerichtet worden sind. Dafür ist weiter wichtig, dass der Finanzdienstleister ein fehlerhaftes Ergebnis erkennen und dann auch unterscheiden kann, ob dieses im Modell(-risiko) begründet liegt oder aufgrund eines erfolgreichen Angriffs auf die KI erfolgt ist.

Maßnahmen in der Informationssicherheit dürfen nicht nur theoretisch zu mehr Sicherheit führen, sondern müssen konkret auch so umgesetzt werden, dass sie für den Bankmitarbeiter oder den Kunden auch handhabbar sind. Das betrifft den Einsatz von (Sicherheits-)Technologien ebenso wie Sicherheitsanforderungen (Managementanforderungen), sodass diese tatsächlich wie vorgesehen zum Einsatz kommen und nicht ausgelassen, umgangen oder falsch eingesetzt werden.

Usable Security im Sinne des Treffens richtiger und nötiger Maßnahmen wird durch die Schaffung von Transparenz, Nutzbarkeit, Barrierefreiheit, Zugänglichkeit und Akzeptanz erreicht. Ziel ist es, KI-Systeme aus der Sicht des Benutzenden zu entwickeln. Dazu gehört, dass der jeweilige Mitarbeiter oder Kunde mit einer verständlichen Benutzeroberfläche arbeitet, ausreichend informiert oder geschult ist und Sicherheitsprozesse ohne Eingriffe des Nutzenden ablaufen. Nutzungsfehler, die die Sicherheit kompromittieren können, werden so minimiert. Mitarbeitende des Finanzinstituts müssen in der Anwendung des KI-Systems angemessen geschult werden, sodass sie die Funktionsweise der Anwendung verstehen und nachvollziehen können, wie sie sich in den Gesamtprozess einfügt. Verbraucher*innen jedoch sind mit den internen Prozessen von Finanzinstituten grundsätzlich nicht vertraut und müssen dementsprechend ausreichend über die Nutzung des KI-Systems informiert werden. So führt Usable Security zu einem angemessenen Sicherheitslevel, aber auch zu einer höheren Effizienz und Performanz der Systeme.

KI-spezifische Herausforderungen

Für den Finanzsektor gibt es eine Reihe an regulatorischen Anforderungen an die IT, die für die Finanzinstitute verpflichtend sind. Dazu kommen gängige Normen und Standards, die in der Branche anzuwenden sind. Diese sind zwar nicht bindend, es wird jedoch von der Regulatorik (vgl. Kapitel 4.8.2.1) vorgegeben, sich an gängige Standards auszurichten. Diese Anforderungen gelten dementsprechend auch für KI. Die regulatorischen Anforderungen verpflichten Finanzdienstleister, ein risikoadäquates Internes Kontrollsystem (IKS) sowie ein „Information Security Management System“ (ISMS) zu etablieren und deren Funktionsfähigkeit, auch zum Bezug von entsprechenden Versicherungsdienstleistungen, nachzuweisen. Darüber hinaus müssen die IT-seitigen Maßnahmen dem Stand der Technik entsprechen. Angriffe auf KI-Systeme sind nicht spezifisch für den Bereich der Finanzdienstleistungen, jedoch bestehen für Finanzdienstleister bereits durch den hohen Schutzbedarf der Daten erhöhte Anforderungen an die Informationssicherheit. Krisen im Finanzsektor oder auch nur ein Vertrauensverlust in die Institute können weitreichende Folgen für die gesamte Wirtschaft haben. Die gesamtwirtschaftlichen Folgen können höher ausfallen als in anderen Branchen. Die Entscheidungen, die durch KI-Systeme von Finanzdienstleistern getroffen werden, betreffen Kund*innen (z. B. Kreditvergaben) und auch weitere Stakeholder. Stuft man also die Konsequenzen von Sicherheitsproblemen ein, dann sind all diese Personengruppen einzubeziehen.

Die Anforderungen zur Informationssicherheit, die aus aktuellen Normen und Standards für IT-Systeme resultieren, müssen auch in den technischen Prozessen umgesetzt werden. In der IT ist hier insbesondere eine geeignete Kombination aus statischen und dynamischen Analysen erforderlich, um potenzielle Sicherheitslücken und Schwachstellen frühzeitig (vor ihrer Ausnutzung) aufzudecken und zu beheben. Diese Analysen beinhalten u. a. Vulnerability Scans eingesetzter Third Party Libraries (inklusive Open Source), diverse Codeanalysen und Penetrationstests. Gleiches gilt für die Absicherung der Infrastruktur, organisatorische Abläufe, Prozesse etc. Hierfür wird kein spezieller Normungsbedarf gesehen, bestehende Normen schließen KI-Systeme implizit ein.

Im Fall von KI-Systemen sind jedoch zusätzliche Angriffstypen und Angriffsszenarien möglich, die besonders zu behandeln sind. Das Dokument „Sicherer, robuster und nachvollziehbarer Einsatz von KI“, welches vom Bundesamt für Sicherheit in der Informationstechnik (BSI) [83] veröffentlicht wurde, benennt u. a. Evasion / Adversarial Attacks, Data Poisoning Attacks, Privacy Attacks und Model Stealing Attacks als KI-spezifische Angriffstypen. Auf Evasion / Adversarial Attacks und Model Stealing Attacks wird bereits in der ersten Ausgabe der Normungsroadmap eingegangen (vgl. [1], S. 109 ff.), hier ergeben sich keine zusätzlichen Normungsbedarfe. Hinzu kommen Data Poisoning Attacks und Privacy Attacks. Diese sind gemäß BSI folgendermaßen definiert:

Data Poisoning Attacks

Durch eine Manipulation der Trainingsdaten des KI-Modells erwirken Angreifende, dass dieses auf (bestimmte) Eingaben nicht wie vom Entwickelnden vorgesehen reagiert. Aufgrund der vielen Daten und der mangelnden Transparenz sind diese Angriffe meist schwer detektierbar.

Privacy Attack

Angreifende extrahieren Informationen hinsichtlich der Trainingsdaten aus dem Modell. Model Inversion Attacks extrahieren Trainingsdaten und Membership Inference Attacks stellen fest, ob ein Datum zum Training verwendet wurde.

Somit besteht der Bedarf nach Normen für geeignete Maßnahmen, durch die diese Angriffsszenarien angemessen mitigiert werden. Regulatorische Anforderungen und gängige Standards, die für den Einsatz von IT im Allgemeinen bestehen, sind vor dem Hintergrund einer potenziell veränderten Risikosituation zu betrachten. Basierend hierauf sind zusätzliche Schutzmaßnahmen zu berücksichtigen, die auf die konkrete Bedrohungslage des Einsatzes von KI abzielen und dann in das bestehende Risikomanagementsystem integriert werden.

Für das Training und die Validierung der Modelle, die in KI-Systemen zum Einsatz kommen, werden häufig Produktivdaten (so weit möglich und sinnvoll anonymisiert) genutzt. Damit sind diese nicht gleichzusetzen (und vor allem nicht gleichzubehandeln) mit beispielsweise synthetischen Testdaten, die für die Qualitätssicherung von IT-Systemen genutzt werden. Daher gilt für diese Daten, welche während der Entwicklung des KI-Systems Verwendung finden, der gleiche Schutzbedarf wie für die Daten, welche später im produktiven Betrieb Verwendung finden.

Normungsbedarf besteht darin, die Nutzung von Produktivdaten für Trainings- und Validierungszwecke zuzulassen und andererseits aber den hohen Schutzbedarf durch geeignete Maßnahmen ausreichend zu berücksichtigen. Hier besteht eine Verbindung zum Thema „Data Governance“, die nachfolgend adressiert wird.

Schutzmaßnahmen (Informationsschutz)
Im Zuge der Betrachtung möglicher Angriffsvektoren auf die Informationssicherheit wurde ersichtlich, dass insbesondere bei Finanzdienstleistern ein großes Augenmerk auf diese gelegt werden muss. Dies ist weniger damit begründet, dass es eine größere Anzahl an möglichen Angriffsvektoren gibt, im Vergleich zu Nicht-Finanzdienstleistern, sondern vielmehr darin, dass der Schutzbedarf der Daten, welcher Verwendung im Rahmen von KI findet, grundsätzlich von vier Schutzbedarfsklassen (gering, mittel, hoch, sehr hoch) mindestens „hoch“ oder „sehr hoch“ ist. Dies rührt vor allem daher, dass KI-Anwendungsfälle in der Finanzwirtschaft meist durch eine geringe Distanz zum Endkunden geprägt sind und außerdem schützenswerte und vertrauliche Informationen wie beispielsweise Bonitäts- und Gesundheitsdaten verarbeitet werden. Dementsprechend ist davon auszugehen, dass der Schutzbedarf der meisten Daten beim Einsatz von KI im Finanzsektor hinsichtlich der Schutzziele Vertraulichkeit, Integrität und auch der Verfügbarkeit mindestens hoch ist und somit Schutzmaßnahmen getroffen werden müssen, die diesem Schutzbedarf gerecht werden.

Dabei ist dieser erhöhte Schutzbedarf der Daten, insbesondere im Finanzsektor, bereits in Form einer bestehenden und umfassenden Regulatorik adressiert. Daher ist es derzeit nicht notwendig, die bereits bestehende Regulatorik umfassend zu erweitern. Vielmehr genügt in der Regel eine zielgerichtete Präzisierung bzw. Konkretisierung und Verweis auf diese, um den spezifischen Eigenschaften einer KI bei der Entwicklung und dem produktiven Betrieb adäquat zu begegnen. Der technologieneutrale Ansatz der heutigen Aufsichtspraxis der BaFin ermöglicht auf Basis der MaRisk und deren Konkretisierungen hinsichtlich der IT BAIT, VAIT (Versicherungsaufsichtliche Anforderungen an die IT), ZAIT (Zahlungsdiensteaufsichtliche Anforderungen an die IT) und KAIT (Kapitalverwaltungsaufsichtliche Anforderungen an die IT) in der Regel einen risikoadäquaten Umgang mit der IT im Kontext KI.

Neben den bestehenden regulatorischen Anforderungen gibt es gängige Standards, die bei der Implementierung von Informationssicherheits- und Informationsrisikomanagementsystemen angewendet werden. Zu den wichtigsten Standards gehören die DIN-EN-ISO/IEC-27000-Reihe [479] sowie der BSI-Grundschutz des Bundesamtes für Sicherheit in der Informationstechnik. Neben diesen nationalen Anforderungen und Standards für die IT sowie einschließlich KI gibt es auf EU-Ebene weitere spezifische Initiativen hinsichtlich KI und Daten: z. B. die Verordnung für KI der EU (Artificial Intelligence Act, siehe Kapitel 1.4). Diese Verordnung soll branchenübergreifend und EU-weit einen Rechtsrahmen für die Entwicklung und den Einsatz von KI stellen. Weiterhin hat die EU-Kommission im Rahmen ihrer Datenstrategie zwei Initiativen in Bewegung gesetzt: den Data Governance Act, der dieses Jahr in Kraft getreten ist, sowie den Data Act, für den ein Gesetzesentwurf derzeit im Parlament besprochen wird. Beide Gesetzesinitiativen haben das Ziel, den Datenaustausch in der EU zu steigern, sicher zu gestalten und Innovationen durch verstärkte Nutzung von Daten zu fördern. Zusammenfassend kann gesagt werden, dass eine sinnvolle Ausrichtung bestehender Anforderungen im Kontext der Informationssicherheit auf die Technologie von KI benötigt wird.

Im Folgenden erfolgt daher keine isolierte Gegenüberstellung einzelner Angriffsvektoren sowie möglicher korrespondierender Schutzmaßnahmen, wie dies bereits in der Normungsroadmap Ausgabe 1 (siehe: [63], 4.4.2.2.2 Angriffsvektoren und Verteidigungsmechanismen) erfolgt ist, sondern es soll der Gedanke des holistischen Konzepts der „robusten KI-Plattform“ angestoßen werden. Gemäß diesem Konzept der „robusten KI-Plattform“ werden Data Governance und Informationssicherheitsmaßnahmen von Anfang an integriert, sodass Informationen im gesamten KI-Lebenszyklus angemessen geschützt sind. Eine KI-Plattform setzt sich dabei aus Systemkomponenten bzw. deren Subkomponenten zusammen, die die KI bereitstellen. Dazu kommen die zugehörigen Daten, Prozesse und Sicherheitsmaßnahmen, die sich in allen Lebenszyklusphasen der KI-Anwendung finden (siehe hierzu auch [63], 4.2.2.2 Herausforderung 1: „Definition von Schutzzielen auf der Ebene von Prozessen und Daten innerhalb der KI-Komponente“). Der Terminus „robust“ bezieht sich auf bestehende Regulatorik sowie individuelle Maßnahmen in dem jeweiligen Finanzinstitut. Robust ist eine KI-Plattform daher erst, wenn sie sich nachweisbar im Einklang mit bestehenden Anforderungen und Gesetzen befindet und sich den spezifischen Herausforderungen eines Finanzinstituts stellt, welche entsprechende Informationsrisiken adressieren.

Das Zielbild sollte daher eine allgemein akzeptierte Definition eines Katalogs an Mindestanforderungen und „Best Practice“-Vorgehensweisen an eine KI-Plattform je Komponente sein, welche durch Verweise und Konkretisierung bereits bestehender Regularien und Gesetze erfolgt. Da man bei der Umsetzung von KI-Anwendungsfällen in der Praxis häufig noch nicht auf etablierte Blueprints und Vorgehensweisen zurückgreifen kann, handelt es sich hier um ein Forschungsfeld von hoher Relevanz. Ziel ist eine beschleunigte, dabei aber konforme, sichere und verlässliche Implementierung von KI in Deutschland.

4.8.2.5 Risikomanagement

Finanzdienstleister sind professionelle Risikomanager und setzen seit Langem KI-Systeme ein, um diese Aufgabe zu erfüllen. Zudem haben sie eine lange Tradition im Management von Modellrisiken und werden diesbezüglich von den Finanzaufsichtsbehörden kompetent und anhand von gut etablierten Standards überwacht. Daher ist eine Einbettung einer KI-Normung in die bestehenden Risikomanagement- und Auditprozesse für den Sektor von höchster Bedeutung. Relevante Aspekte dazu sollen im Folgenden erörtert werden.

Individuelles Modellrisiko

VERSTÄNDNIS DER RISIKEN DES MODELLEINSATZES UND DER DAMIT VERBUNDENEN HERAUSFORDERUNGEN DES MASCHINELLEN LERNENS

KI-/ML-Modelle teilen die meisten Risiken mit herkömmlichen Modellen; diese Risiken sind aber schwieriger zu erkennen und zu bewerten. Vor allem die Qualität der Daten wirkt sich erheblich auf die Leistung von KI/ML-Modellen aus und kann als wichtigster Begrenzungsfaktor für KI/ML angesehen werden.

KI/ML-Modelle und herkömmliche Modelle unterscheiden sich in den Merkmalen der algorithmischen Risiken wie Erklärbarkeit, Verzerrung und Robustheit.

Dabei sind nicht nur algorithmische Risiken zu beachten, sondern auch rechtliche Risiken in Bezug auf Datenschutzrecht, Zivilrecht (beispielsweise Verantwortung bei Willenserklärungen durch KI), Anti-Diskriminierungsrecht.

Zur effektiven Reduzierung des Modellrisikos ist es entscheidend, dass das Modellrisiko angemessen in allen drei Lines of Defence verankert ist. Dafür müssen die Mitarbeiter*innen die entsprechenden Fähigkeiten haben und Verantwortlichkeiten müssen klar verteilt sein.

WIE LÄSST SICH KI IN BESTEHENDE RAHMENWERKE FÜR DAS MODELLRISIKOMANAGEMENT INTEGRIEREN?

(Regulierte) Finanzunternehmen müssen eine Governancestruktur einrichten, die sich mit den Risiken von KI/ML-Modellen befasst – idealerweise auf der Grundlage bestehender Governance-Rahmenwerke. Eine Governance – die für die gesamte Bank gilt – sollte dabei das allgemeine Modellrisikomanagement aller produktiven Modelle abdecken und um KI-spezifische Richtlinien erweitern.

Der erste Schritt auf dem Weg zu einem KI-spezifischen Modellrisikorahmen ist die Überarbeitung der bereits bestehenden Prozesse/Regelwerke. Der Gesamtansatz unterscheidet sich nicht vom Modellrisiko traditioneller Modelle, nämlich von der Entwicklung einer Risikostrategie mit entsprechender Risikotragfähigkeit und einem adäquaten Risikoappetit bis hin zu Risikominderungsmaßnahmen.

- → Entwicklungs- und Validierungsüberlegungen sollten von konzeptioneller Solidität, Daten- und Feature-Engineering, Training und Kalibrierung sowie Testen und Überwachung bestimmt sein. In diesem Zusammenhang sollte das Auffinden des geeignetsten Modells und seiner Parameter sowie die Risikobewertung, das Modelländerungsmanagement, die laufende Modifizierung, das Problemmanagement, der Softwareentwicklungsprozess sowie das Lieferantenmodellmanagement berücksichtigt werden.
- → Darüber hinaus müssen Modelleigentümer sowie Modellentwickler eine Bewertung vornehmen, ob ein KI-/ML-Modell die gewünschte Leistungssteigerung bringt oder ob ein traditionelles Modell ausreichend ist, um dem Risikoappetit der Bank hinsichtlich Modellrisiko zu entsprechen.
- → Es sollte gewährleistet sein, dass die Testdaten, insbesondere in Bezug auf die produktiven Daten, repräsentativ sind, um die Generalisierbarkeit messen zu können. Auch die Datenqualität der Trainingsdaten sollte sichergestellt werden, indem diese korrekt, vollständig und widerspruchsfrei sind. Im Bereich des Supervised Learning sollte außerdem sichergestellt sein, dass die Annotationen der Daten von hoher Qualität sind.
- → Die Modellkomplexität kann durch die Bewertung von drei Unterkategorien beurteilt werden, nämlich Modelleingabedaten, Annahmen & Theorie. Dabei sollte die Implementierung keinen Einfluss auf die Komplexität eines Modells haben, sondern die Zahl der freien Parameter, wobei diese nicht immer offensichtlich sind, wenn z. B. komplexe Features erzeugt werden.
- → Generell sollten Modelle verschiedenen Komplexitätsklassen mit entsprechenden Sorgfaltspflichten zugeordnet werden. Eine exakte Normung dessen scheint aus Komplexitätsgründen unrealistisch, daher sollte eine grobe qualitative Einteilung in wenige Klassen erfolgen.
- → Es ist ein kontinuierliches, weitgehend automatisiertes Monitoring zu definieren, das die bisherigen jährlichen Validierungszyklen durch anlassbezogene Überprüfungen ersetzt, die durch die automatischen Monitorings getriggert werden.

Wie herkömmliche Modelle müssen auch KI-/ML-Modelle einer ersten und regelmäßigen Überprüfung und Validierung unterzogen werden, je nach dem Grad des Modellrisikos.

Wie lässt sich KI in den bestehenden Rahmen für die Modellvalidierung integrieren?

KI/ML-Algorithmen bringen besonderen Herausforderungen mit sich, wie z. B. Erklärbarkeit und Robustheit, die es insbesondere für Banken in einem regulatorischen Internal ratings-based approach (IRBA) Model umzusetzen gilt. Der vollständige Datenverlauf (Data Lineage), die Modellparameter und generell alle Metadaten müssen zu einem späteren Zeitpunkt zugänglich sein. Der Prozess der Modellvalidierung muss erweitert werden, um den Besonderheiten der KI/ML gerecht zu werden. Wie bei traditionellen Modellen müssen Validierende sicherstellen, dass das ausgewählte Modell konzeptionell solide ist, indem sie prüfen, ob die einzelnen Merkmale generierbare Prädikatoren sind und ob sie aus betriebswirtschaftlicher Sicht sinnvoll sind, und ob das Modell nicht übermäßig an irrelevante Aspekte der Trainingsdaten angepasst ist. Der Prozess der Modellvalidierung muss verbessert werden, um den Besonderheiten von ML gerecht zu werden. Die Modellvalidierung muss sicherstellen, dass das Modell in unvorhergesehenen Situationen wie beabsichtigt Vorhersagen trifft, indem z. B. Stresstests durchgeführt werden. Die Modellvalidierung muss sicherstellen, dass das Modell für die verschiedenen Interessengruppen hinreichend transparent ist, d. h. dass es in der Lage ist, die Gründe für eine bestimmte Entscheidung zu erläutern. Möglichkeiten, Modellergebnisse a posteriori zu plausibilisieren, sind z. B. Explainable-AI-Verfahren wie: Perturbationsanalyse, Gradientenanalyse, Surrogatmodellierung sowie beispielbasierte Erklärungen. Bias innerhalb der Daten muss bereits in der Entwurfsphase berücksichtigt werden – beginnend mit der Auswahl der Eingangsgrößen. Diese erfolgt zwar weitgehend automatisch im Zuge der Modelloptimierung, aber nur in Bezug auf verfügbare Eingangsgrößen kann etwa ein Bias überhaupt festgestellt werden.

Risikomodellierung / Korrelationsmodell für Messung kumulierter Fehler

Zur erfolgreichen Risikomodellierung im Maschinellen Lernen / der künstlichen Intelligenz gilt es, die besonderen Risiken eines solchen Systems zu beachten. Häufig werden ML-Modelle untereinander verknüpft oder für ganze Portfolios angewandt und potenzieren damit mögliche Risiken – wie bei herkömmlichen Modellen, jedoch unterscheiden sich die Auswirkungen und Werkzeuge, die zur Reaktion im Fehlerfall zur Verfügung stehen. Die folgenden möglichen Risiken und Risikodimensionen sind hier zu betrachten.

Mögliche Risiken und Risikodimensionen

AUSGABEN VON KORREKT FUNKTIONIERENDEN EINZELMODELLEN WERDEN FEHLERHAFT IN ANDEREN SYSTEMEN VERWENDET

Werden Ergebnisse einzelner Modelle rein anhand ihrer Klassifizierung genutzt, kann dies zu unerwartetem Systemverhalten führen. Wenn z. B. die von einem Modell vorgeschlagene Klassifizierungsentscheidung in anderen Modellen weiterverwendet wird, etwa eine Ratingklasse und nicht der exakte Kreditwürdigkeitsscore, führt dies schon bei einer einfachen Mittelwertbildung zu Verzerrungen. Für die Verwendung von Modellausgaben sollte der ursprüngliche Kontext bekannt sein und berücksichtigt werden. Dies schließt ein: Modellentscheidungen und Abwägungen, Ursprung und Kontext der Trainingsdaten, Ursprung und Kontext der Echtzeiteingabedaten. Normung kann hier ansetzen und sicherstellen, dass keine relevanten Informationen während der Übernahme unbekannt bleiben.

KONTEXTSENSITIVE MODELLE WERDEN IN ANDEREN UMGEBUNGEN EINGESETZT / AUS VERSCHIEDENEN UMGEBUNGEN VERKNÜPFT

Im Gegensatz zu klassischen Modellen sind Entscheidungen Künstlicher Intelligenz nicht einfach nachzuvollziehen. Werden Modelle einem neuen Kontext zugeführt oder aus verschiedenen Kontexten verknüpft, kann dies zu grundlegend falschen Entscheidungen führen. Für die Verwendung anderer Modellausgaben sollte der ursprüngliche Kontext bekannt sein und berücksichtigt werden. Dies schließt ein: Modellentscheidungen und Abwägungen, Ursprung und Kontext der Trainingsdaten, Ursprung und Kontext der Echtzeiteingabedaten.

Eine mögliche Kostenersparnis durch Wiederverwendung sollte immer differenziert betrachtet werden, da ein augenscheinlich gutes Modell in einem neuen Kontext grundlegend andere Ergebnisse liefern kann.

Normung kann hier ansetzen und dazu beitragen, dass wesentliche Informationen genutzt werden.

KUMULATION VON ABWEICHUNGSFEHLERN

Manche Eingabewerte ändern sich mit der Zeit. Eine manuelle Neugewichtung der Parameter ist im Maschinellen Lernen in der Regel nicht praktikabel. Zur Anpassung des Modells können von Anfang an anpassbare Parameter integriert werden, um solchen Entwicklungen Rechnung zu tragen – eine weitere Möglichkeit ist eine regelmäßige Reevaluierung

anhand neuer Trainingsdaten und des zu Projektbeginn festgelegten Anforderungskatalogs. Da die Eingabewerte mit der Zeit immer stärker abweichen, steigt der Aufwand, je länger ein Modell nicht reevaluiert wurde. Um Hürden zu verringern, könnte eine Normung Voraussetzungen für eine regelmäßige und niedrigschwellige Reevaluierung schaffen.

Einordnung in die bestehende Regulierung und die laufenden Konsultationen der BaFin

Nachfolgend wird die Sicht des Regulators, namentlich der Europäische Bankenaufsichtsbehörde (EBA) sowie der BaFin und Bundesbank, dargelegt. Da die Konsultationsphase noch nicht abschließend eingearbeitet wurde, wird hierfür auf die aktuellen Diskussionspapiere zu diesem Thema zurückgegriffen. Hinsichtlich MaRisk wird erwartet, dass derzeit keine weiteren Regelungen hinsichtlich der Verwendung von Künstlicher Intelligenz nötig sind und entsprechende Modelle durch bereits existierende Regeln abgedeckt sind.

Das Papier der EBA zur Verwendung von Maschinellem Lernen im „Internal ratings-based“ (IRB)-Kontext enthält eine Reihe von prinzipienbasierten Empfehlungen, die sicherstellen sollen, dass Modelle des Maschinellen Lernens im Kontext des IRB-Rahmens die in der „Capital Requirements Regulation“ (CRR) festgelegten regulatorischen Anforderungen einhalten.

IRB-Modelle müssen gemäß Art. 179 CRR „intuitiv“ sein. Das bedeutet, es muss eine leicht verständliche Verbindung zwischen den Risikotreibern und dem Ausfallindikator für PD-Modelle (Probability of Default (Ausfallwahrscheinlichkeit)) geben. Herkömmliche Modelle erfüllen diesen Anspruch: Sie zeigen oft sehr klare und sofort quantifizierbare Beziehungen zwischen einem Risikotreiber (z. B. Loan-to-Income) und dem Ausfall „Ja/Nein“ auf.

Die EBA empfiehlt den Instituten, sicherzustellen, dass die Leitungsorgane in der Lage sind, Annahmen, Limitierungen und die Theorie des Modells zu verstehen, indem diesen eine angemessene Dokumentation zur Verfügung gestellt wird. Außerdem müssen die Mitarbeiter*innen in den Abteilungen für Modellentwicklung, Kreditrisikocontrolling und Validierung ausreichend qualifiziert sein. Dabei sind vor allem die Bausteine Fairness, Erklärbarkeit und Robustheit für die quantitative Modellvalidierung relevant.

Bei ML-basierten Modellen ist es in der Regel schwieriger, das Design, die funktionalen Details, die dem Modell zugrunde liegende Theorie und die Modellierungsannahmen zu dokumentieren (siehe CRR, Art. 175). Deshalb wird den Instituten empfohlen, übermäßig komplexe Modellierungsentscheidungen zu vermeiden, es sei denn, sie sind durch eine signifikante Verbesserung der Vorhersagefähigkeit gerechtfertigt. Vor allem soll Bias vermieden werden, sodass unternehmerische Entscheidungen nicht auf systematisch verzerrten Ergebnissen beruhen dürfen, wodurch einzelne Kundengruppen benachteiligt werden. Dies ist nötig, um vor allem Diskriminierungsverbote in EU-Gesetzen einzuhalten und die damit resultierenden Reputationsrisiken zu minimieren. Dies ist ein übergeordnetes Prinzip, da Verzerrungen sowohl Entwicklung als auch Anwendung betreffen können. Auch wird auf gesetzlich untersagte Differenzierung hingewiesen. So dürfen bestimmte Merkmale wie Herkunft, Geschlecht oder sexuelle Orientierung in der Risiko- und Preiskalkulation keine Berücksichtigung finden.

Während der Entwicklungsphase gilt es, folgende Punkte zu beachten (vgl. IRBA):

- → Geeignete Datenstrategie und Data Governance (inklusive Repräsentativität)
- → Beachtung von Datenschutzregeln
- → Korrekte, reproduzierbare und robuste Ergebnisse
- → Angemessene Dokumentation
- → Angemessene Validierungsprozesse

Während der Anwendungsphase gilt es, folgende Punkte zu beachten:

- → „Putting the human in the loop“
- → Intensive Freigabe- und Feedbackprozesse
- → Etablierung von Notmaßnahmen
- → Validierung, Evaluation und Anpassung

Um sicherzustellen, dass das Modell richtig interpretiert wird, wird den Instituten bei der Analyse der Modelle Folgendes empfohlen:

- → Auf statistische Weise die Beziehung jedes einzelnen Risikotreibers zur Ausgangsvariablen und das Gesamtgewicht jedes Risikotreibers bei der Bestimmung der Ausgangsvariablen zu analysieren.
- → Die ökonomische Beziehung jedes Risikotreibers mit der Output-Variablen zu bewerten, um sicherzustellen, dass die Modellschätzungen plausibel und intuitiv sind.
- → Ein zusammenfassendes Dokument zu erstellen, in dem das Modell auf der Grundlage der Analyseergebnisse auf einfache Weise erläutert wird.
- → Sicherzustellen, dass potenzielle Verzerrungen des Modells (z. B. Überanpassung an die Trainingsstichprobe) erkannt werden.

Um die Verwendung der Künstlichen Intelligenz in bereits anerkannte Strukturen zu bringen, fokussiert sich das gemeinsame Diskussionspapier [395] von BaFin und Bundesbank auf die Ergänzung, Präzisierung und Weiterentwicklung bestehender Regelungen und soll als Orientierungshilfe für beaufsichtigte Unternehmen dienen. Grundsätzlich sollen vorrangig bestehende Regelungen beachtet werden. Dabei werden insbesondere die internen Modelle für die Eigenmittelanforderungen nach Säule 1 und für das Risikomanagement nach Säule 2 in den Fokus gesetzt.

Bei der Betrachtung von ML gilt es immer, den gesamten Prozess und die konkrete Situation im Institut zu betrachten und nicht nur den Algorithmus allein. Die Angemessenheit eines Algorithmus hängt dann entsprechend vom konkreten Einsatz bzw. Entscheidungsprozess ebenso wie vom Umfang und der Qualität der Daten ab (siehe auch Kapitel 4.8.2.4).

Qua Vielfalt der ML-Ansätze gibt es auch keinen allgemeingültigen Abnahmeprozess. Es gilt, eine risikoorientierte Prüfung und Beanstandung algorithmenbasierter Entscheidungsprozesse vorzunehmen. Ausgenommen hiervon sind begründete Fälle, etwa bei internen Modellen, mit Fokus auf Methodik, Kalibrierung oder Validierung.

Für die Aufsicht wichtige Prinzipien bleiben die Risikoorientierung, Verhältnismäßigkeit und Technologieneutralität. Auch deswegen erfordert dies eine intensivere Überwachung bei der Anwendung von Algorithmen in kritischen Entscheidungsprozessen. Gleiches gilt bei der Berücksichtigung von Komplexität, Rekalibrierungsfrequenz und Automatisierungsgrad.

Die Verantwortung verbleibt auch in dieser neuen Hinsicht klar bei der Geschäftsleitung. Diese gibt Strategien und Leitlinien zum Einsatz algorithmenbasierter Entscheidungsprozesse vor, wobei Potenziale sowie Grenzen und Risiken solcher Prozesse berücksichtigt werden müssen. Hierfür wird ein angemessenes technisches Verständnis und adressatengerechte Kommunikation vorausgesetzt. Es wird Instituten empfohlen, ein übergreifendes Rahmenwerk zu schaffen, das eine Aufstellung aller algorithmenbasierten Entscheidungsprozesse beinhaltet (Modell Inventar) und deren wechselseitige Abhängigkeit berücksichtigt. Zudem wird vorgeschlagen, diesen Aspekt im Model Risk Management Framework zu berücksichtigen.

BaFin und Bundesbank sehen derzeit keine Erfordernisse für eine grundsätzlich neue Aufsichtspraxis für ML-Methoden. Es wird jedoch laufend untersucht, inwieweit Anpassungen an bestimmten Stellen erforderlich sind.

Die aufsichtliche Praxis, insbesondere geprägt durch die MaRisk und xAIT, hat Bestand. Die umfangreichen Regeln in Säule 1 und die prinzipienorientierten Anforderungen in Säule 2 liefern eine solide Grundlage. Der aufsichtliche Fokus liegt bei neuen oder deutlicher ausgeprägten Risiken in der Datengrundlage, Validierung, Modelländerung und bei der Steuerung. Eine Herausforderung für die Aufsicht stellt die Widerspruchsfreiheit zu KI-Verordnung und Verbraucherschutz dar. Drei wesentliche Punkte für die Bankenaufsicht sind:

1. Modelle laden zur Datengläubigkeit ein, wodurch die Gefahr des „Overfitting“ entsteht. Es werden vermeintliche Korrelationen auf Basis zufälliger Eigenschaften identifiziert. Daraus resultiert, dass die Sicherstellung der Datenqualität eine zentrale Aufgabe der beaufsichtigten Institute ist.
2. Die Erklärbarkeit der Modelle rückt anstelle der Nachvollziehbarkeit in den Fokus. Dabei schätzt die Bankenaufsicht Explainable AI als vielversprechend ein, allerdings ist zu berücksichtigen, dass sich hinter diesem Begriff ebenfalls Modelle verbergen.
3. Materielle Modellveränderungen sind schwerer zu erkennen (Adaptivität). Die Grenze zwischen Modellpflege und Modelländerungen ist fließend. Insbesondere besteht die Gefahr, dass sich die Modelle in kurzer Zeit vom Ursprungsmodell entfernen, ohne dass der Model Owner dies bemerkt.

Die Vorschläge scheinen aus heutiger Sicht ein guter Weg zu sein, den bestehenden Rahmen zur Regulierung von Modellrisiken für die Spezifika der für KI relevanten Methoden zu erweitern.

4.8.3 Normungs- und Standardisierungsbedarfe

Bedarf 08-01: Definition nachprüfbarer Antidiskriminierungsmetriken zum Nachweis der Diskriminierungsfreiheit einer KI-Lösung
KI soll eine möglichst positive Wirkung entfalten, muss auf der anderen Seite aber Regeln unterworfen sein. Dort, wo es im Finanzdienstleistungssektor um Menschen geht, ist eine wichtige Regel das Diskriminierungsverbot. Die Einhaltung der Regeln durch Anbieter, die Überprüfung durch Kontrollbehörden und die Darstellung gegenüber den Verbraucher*innen ist eine große Herausforderung, u. a. weil der Begriff Diskriminierung mehrdeutig und mit anderen Begriffen wie Fairness, Gerechtigkeit und Gleichbehandlung verwandt ist.

Im Folgenden wird Diskriminierung als ungerechtfertigte Benachteiligung oder Bevorzugung verstanden. (Im Sinne von Art. 3 Abs. 3 GG der Bundesrepublik Deutschland.) Eine – im besten Fall automatisierte – nachprüfbare Definition von Diskriminierung kann sich aus der Normung von Metriken diesbezüglich ergeben.

Hierbei gibt es einige Schwierigkeiten:
- → In der aktuellen Forschung werden Antidiskriminierungsmetriken oft als „Fairnessmetriken“ bezeichnet.
- → Darüber hinaus gibt es in der aktuellen Diskussion mehr als eine Diskriminierungsmetrik.
- → Nicht alle bisher bekannten Antidiskriminierungsmaße können gleichzeitig eingehalten werden.
- → Entwickelnde von KI-Lösungen müssen also die Möglichkeit der Auswahl haben.
- → Bei der Einhaltung von Metriken muss es erlaubte Toleranzen geben, wenn sich die Metriken in der Praxis nicht exakt einhalten lassen.

Eine vertrauensvolle KI kann eine Chance für Europa und europäische Firmen im Wettbewerb mit US-amerikanischen und chinesischen KI-Anbietern sein. Der Finanzsektor würde besonders von Maßen profitieren, da es hier weniger Vertrauenspersonen gibt als z. B. mit den Ärzt*innen im Medizinsektor. Denkbar wäre ein „Gütesiegel“ in Analogie zum „Blauen Engel“ oder dem Nutri-Score und/oder eine Bewertung im „S“-Teil von Environmental Social Governance (ESG-Scores) für Unternehmen.

Anbieter*innen und Entwickler*innen von KI-Lösungen profitieren von der Rechtssicherheit durch objektive und automatisiert nachprüfbaren Regeln.

Bedarf 08-02: Normung der für Nichtdiskriminierung relevanten Merkmale und des Umgangs damit
In den Gesetzen und Vorgaben zu Antidiskriminierung werden die relevanten Merkmale inkonsistent genannt.

Beispiele:

Charta der Grundrechte der EU (Art. 21 „Nichtdiskriminierung“): „Diskriminierungen, insbesondere wegen des Geschlechts, der Rasse, der Hautfarbe, der ethnischen oder sozialen Herkunft, der genetischen Merkmale, der Sprache, der Religion oder der Weltanschauung, der politischen oder sonstigen Anschauung, der Zugehörigkeit zu einer nationalen Minderheit, des Vermögens, der Geburt, einer Behinderung, des Alters oder der sexuellen Ausrichtung, sind verboten.“

Vertrag über die Arbeitsweise der EU: „... discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation.“

Eine einheitliche und abschließende Liste der Merkmale kann helfen, Aufwände bei der Erstellung von KI-Lösungen zu vermeiden bzw. die Leistungsfähigkeit einer KI-Lösung zu verbessern.

Darüber hinaus soll genormt werden, wie die relevanten Merkmale bei der Erstellung der KI-Lösungen berücksichtigt werden sollen. Ein genereller Ausschluss ist möglicherweise kontraproduktiv. Beispiel: Unter der Annahme, dass die Kreditwürdigkeit einer Person von der Dauer der bisherigen Bankverbindungen abhängt und zugleich historisch bedingt insbesondere ältere Frauen im Mittel kürzere Bankverbindungen haben, wäre eine gegebene Dauer einer Bankverbindung für eine Frau möglicherweise positiver zu werten als für einen Mann. Entfernt man das Merkmal „Geschlecht“ aus den Lerndaten, wären ältere Frauen systematisch benachteiligt.

Anbieter*innen und Entwickler*innen von KI-Lösungen profitieren von der Rechtssicherheit durch konsistente Regeln.

Bedarf 08-03: Normung der Berücksichtigung von Nichtdiskriminierungsaspekten bei der Erstellung einer KI-Lösung zum Nachweis der Diskriminierungsfreiheit
Eine weitere Möglichkeit des Nachweises der Diskriminierungsfreiheit einer KI-Lösung ist nicht das Produkt / der Service selbst, sondern, den Erstellungsprozess des Produkts/ Services in Hinblick auf die Berücksichtigung des Diskriminierungsverbots zu normen. Die in 08-01 angeforderten Metriken

können dabei eingebracht werden, dadurch dass ein genormter Prozess die Verwendung genormter Metriken vorschreibt. Dabei muss es möglich sein, den Einfluss der Einhaltung der Metriken auf die Gesamtperformance der KI-Lösung zu ermitteln.

Bedarf 08-04: Definition des Begriffs Fairness durch nachprüfbare Metriken
Fairness ist ein Begriff, der noch weniger definiert ist als Diskriminierung. Im Unterschied zu Diskriminierung ist er nicht gesetzlich geregelt und taucht nicht in der Charta der EU-Grundrechte und im Vertrag über die Arbeitsweise der EU nur im Zusammenhang Sport auf. Umso mehr bedarf es Normen analog zu denen in 08-01 und 08-03 in Bezug auf „Nichtdiskriminierung" genannten.

Auch eine „faire" KI – also das freiwillige Einhalten von Fairnessmetriken – kann im Sinne der Begründung von 08-01 ein Vertrauens- bzw. Verkaufsargument für KI-Lösungen im Finanzbereich sein.

Bedarf 08-05: Regeln für den Nachweis der Abdeckung aller relevanter Faktoren bei Gruppenbetrachtungen
Wenn KI-Systeme Aussagen über Gruppen machen, sind diese nicht notwendigerweise auf das Individuum übertragbar. Daher muss sichergestellt sein, dass entweder keine wesentlichen individuellen Faktoren im Modell fehlen oder eine Geltendmachung und Berücksichtigung grundsätzlich möglich ist, sofern sie nicht ethischen Grundsätzen widerspricht. Dies gilt insbesondere, wenn Grundrechte aufgrund von Modellen eingeschränkt werden, die Aussagen über Gruppen von Individuen machen.

Im Kontext von Finanzanwendungen, aber auch bei anderen sozioökonomischen Systemen, steht häufig eine Risikobetrachtung über die Gruppe im Vordergrund, etwa bei der Vorhersage des erwarteten Verlusts in einem Kreditportfolio oder bei der erwarteten Ausbreitung einer Krankheit. Eine korrekte Vorhersage für das Portfolio und entsprechende Risikopreise (oder, analog, entsprechende Gesundheitsschutzmaßnahmen), muss aber auch für das Individuum (dessen Grundrechte berührt werden) unter allen für es verfügbaren Informationen optimiert werden. Das heißt, es müssen je nach Schwere der Konsequenzen alle individuellen Faktoren berücksichtigt werden, die nachweislich einen signifikanten Einfluss auf die Prognose haben. Es braucht daher Regeln, nach denen die relevanten Faktoren bestimmt werden.

Bedarf 08-06: Erarbeitung und Definition von (Mindest-) Anforderungen an eine KI-Plattform
Es sind Leitplanken zur Ausgestaltung einer robusten KI-Plattform aus der Perspektive der Informationssicherheit notwendig. Davon betroffen sind nicht nur rein technische Aspekte einer entsprechenden IT-Plattform, sondern auch die prozessuale Ausgestaltung der Entwicklung und späteren Operationalisierung des KI-Systems. Der Begriff der KI-Plattform definiert sich hier aus der Summe der die KI bereitstellenden Systemkomponenten bzw. deren Subkomponenten sowie der zugehörigen Daten und Prozesse, die über die Lebenszyklusphasen der KI Anwendung finden.

Nicht nur der spätere Betrieb, sondern auch schon die Entwicklung stellt hohe Anforderungen an die Informationssicherheit einer KI-Plattform. Neben den direkten und erweiterten Schutzzielen der Informationssicherheit sind bezüglich der Mindestanforderungen für KI-Plattformen insbesondere auch Vorgaben aus dem Datenschutz zu berücksichtigen.

Vor allem im Bereich der Finanzdienstleistungen ist die Kritikalität dieser Aspekte besonders hoch. Dies ist vor allem damit begründet, dass KI-Use-Cases in der Finanzwirtschaft meist durch eine deutlich geringere Distanz zum Endkunden geprägt sind und so Informationen wie beispielweise Bonitäts- und Gesundheitsdaten Verwendung finden, welche eine erhöhte Sensibilität erfordern.

Deshalb sollten hohe Anforderungen an eine KI-Plattform im Finanzdienstleistungssektor gestellt werden. Diese müssen mit der bestehenden spezifischen Regulatorik (BAIT, VAIT, KAIT etc.) im Einklang sein. Um dem erhöhten Schutzbedarf gerecht zu werden, ist es aber nicht notwendig, die bereits bestehende Regulatorik umfassend zu erweitern. Vielmehr soll eine zielgerichtete, praxisnahe Präzisierung und ein Verweis in Form von Leitplanken und konkreten Vorgaben (i. S. v. Best Practices) erfolgen. Es gilt zu berücksichtigen, dass künftige Änderungen bzw. Ergänzungen der aufsichtlichen Anforderungen nicht auszuschließen sind. Dies gilt insbesondere mit Blick auf die aktuell laufenden internationalen Regulierungsvorhaben.

Bedarf 08-07: Rahmenbedingungen zum Umgang mit Trainingsdaten für KI-Modelle
Für Daten, welche zu Testzwecken in der Finanzwirtschaft verwendet werden, existieren umfangreiche (Verhaltens-)Anforderungen. Hinsichtlich Trainingsdaten für KI-Systeme sind die bestehenden Restriktionen hinsichtlich der Praktikabilität und Beibehaltung eines hohen Schutzbedarfs zu überprüfen.

Für das Training der Modelle, die in KI-Systemen zum Einsatz kommen, werden häufig Daten aus der Produktivumgebung (so weit möglich und sinnvoll anonymisiert) genutzt. Damit sind die Trainingsdaten nicht gleichzusetzen (und vor allem nicht gleichzubehandeln) beispielsweise mit synthetischen Testdaten, die für die Qualitätssicherung von IT-Systemen genutzt werden.

Synthetische Testdaten haben keinen Bezug zu realen Daten und lassen damit auch keine Rückschlüsse auf solche zu. Der Schutzbedarf synthetischer Testdaten ist daher in der Regel niedrig und entsprechend sind es auch die Anforderungen für den Umgang mit ihnen. Hier gibt es eher Vorgaben, dass reale Daten nicht für Tests genutzt werden dürfen.

Trainingsdaten für KI-Modelle müssen aber (in einem gewissen Rahmen) Rückschlüsse zulassen, damit die auf ihnen trainierten Modelle valide sind. Damit ist ihr Schutzbedarf deutlich höher als derjenige synthetischer Testdaten. Daher sind die (geringen) Auflagen für synthetische Testdaten nicht übertragbar auf Trainingsdaten, hier werden weitergehende Regelungen benötigt.

Die Trainings-, Validierungs- und Testdaten der KI-Modelle besitzen somit den gleichen Schutzbedarf wie die Produktivdaten. Im Fall von Finanzdienstleistungen besteht in der Regel mindestens hoher Schutzbedarf (die höchste Schutzbedarfsklasse für personenbezogene Daten). Hier müssen geeignete Rahmenbedingungen insbesondere mit Hinblick auf Informationssicherheit und Datenschutz geschaffen werden, die einerseits dem hohen Schutzbedarf Rechnung tragen und andererseits das Training der KI-Modelle zulassen.

Bedarf 08-08: KI-spezifische Angriffsszenarien und Schutzmaßnahmen
Durch KI entsteht eine neue Risikosituation in der Finanzwirtschaft zum einen durch die Veränderung der Intensität bestehender Risiken, aber auch durch neue Angriffsvektoren. Die veränderten Rahmenbedingungen sind in einer Normung zu berücksichtigen.

Durch den Einsatz von KI in IT-Systemen werden – unter dem Aspekt der Informationssicherheit – zusätzliche Angriffstypen und Angriffsszenarien möglich. Um das Risiko derartiger Angriffe angemessen zu reduzieren, sind diese im Rahmen von Maßnahmen zur Informationssicherheit zu beachten. Das Dokument „Sicherer, robuster und nachvollziehbarer Einsatz von KI“, welches vom BSI [83] veröffentlicht wurde, benennt u. a. Evasion/Adversarial Attacks, Data Poisoning Attacks, Privacy Attacks, Model Stealing Attacks.

In aktuellen Normen und Standards für IT-Systeme (ohne speziellen Fokus auf die Frage, ob KI zum Einsatz kommt) wird auf diese Angriffsszenarien bzw. entsprechende Maßnahmen nicht spezifisch eingegangen. In einer Norm für KI-Systeme sollte darauf jedoch eingegangen werden.

Dieser Bedarf wird im Kontext von Finanzdienstleistungen geäußert, da die Sicherheitsanforderungen hinsichtlich Vertraulichkeit, Verfügbarkeit und Integrität (mindestens) hoch sind. Dies zeigt sich auch in bestehenden Anforderungen und Normen für allgemeine IT-Systeme durch die regulatorischen Anforderungen der Bankenaufsicht. Regelungen, die für den Einsatz von IT-Systemen (ohne Künstliche Intelligenz) sind, sind vor dem Hintergrund einer potenziell veränderten Risikosituation zu betrachten. Basierend hierauf sind zusätzliche Schutzmaßnahmen zu implementieren, die auf die konkrete Bedrohungslage abzielen.

Bedarf 08-09: Festlegung von Kriterien, die für ein automatisches Entity-Matching ausreichend sind
Für kritische Systeme dürfen Identitäten in zwei unterschiedlichen Datensätzen nur gematcht werden, wenn sie zu 100 % übereinstimmen. Daher muss festgelegt werden, welche Kriterien hierfür ausreichend sind. Auch für nicht-kritische Systeme dient es der Qualität, wenn Daten den richtigen Identitäten zugeordnet werden.

Beispiel: Kundennummer ist nicht eindeutig zur Person zuzuordnen. Im Finanzsektor sind die Datensätze, die zum Training einer KI verwendet werden, nicht immer über eindeutige Identifizierungsmerkmale zugeordnet wie z. B. die Personalausweisnummer oder die Krankenversichertennummer im Gesundheitssektor.

Bedarf 08-10: Festlegung von Kriterien, wie die Verlässlichkeit von Matching mithilfe von statischen Modellen gemessen werden kann und welche Mindestwerte notwendig sind
Wenn Identitäten nur probabilistisch gematcht werden, muss gemessen werden können, wie verlässlich das Matching ist und für welche Art der Anwendung welche Mindestverlässlichkeiten gelten sollen.

Die falsche Zuordnung von Daten zu Entitäten ist ebenso eine Fehlerquelle für Training und Anwendung von KI wie die Fehlerhaftigkeit von korrekt zugeordneten Daten.

Bedarf 08-11: Festlegung von Mechanismen, mit denen die Nutzer*innen die Verwendung der eigenen Identität überwachen können
Die Nutzer*innen sollten die Möglichkeit haben, zu erfahren, welche Daten unter ihrer Identität zusammengefasst wurden. Das ist schon im Rahmen der Datenschutz-Grundverordnung (DSGVO) verpflichtend, allerdings ist unklar, ob das alle Daten umfasst, die durch Fuzzy-Matching hinzugezogen wurden, etwa Zeitungsartikel.

Der Prozess der unscharfen Zuordnung ist kaum vollständig sicher zu überwachen. Eine Beteiligung der Betroffenen würde helfen, die Qualität der Zuordnung signifikant zu steigern. Das ist im Finanzsektor seit jeher eine große Herausforderung.

Bedarf 08-12: Leitfaden Usable Security
Maßnahmen in der Informationssicherheit dürfen nicht nur theoretisch zu mehr Sicherheit führen, sondern müssen konkret auch aus Nutzersicht praktisch handhabbar/umsetzbar sein. Das betrifft den Einsatz von (Sicherheits-)Technologien ebenso wie Sicherheitsanforderungen (Managementanforderungen), sodass diese tatsächlich wie vorgesehen zum Einsatz kommen und nicht ausgelassen, umgangen oder falsch eingesetzt werden.

Usable Security im weitesten Sinne wird erreicht durch die Schaffung von Transparenz, Nutzbarkeit, Barrierefreiheit und Zugänglichkeit sowie von Akzeptanz. Nutzungsfehler, die die Sicherheit kompromittieren könnten, werden so vermieden. Betrachtet werden muss der Aspekt der Usable Security aufseiten von Verbraucher*innen, wenn sie mit Systemen interagieren. Betrachtet werden muss aber auch die Nutzung von KI-Systemen durch Anwendende wie beispielsweise Finanzberater*innen. Auch hier führt Usable Security zu einer höheren Effizienz und Performanz der Systeme.

Wird nicht nur in technischen Sicherheitsanforderungen gedacht, sondern der Nutzende einbezogen, kann sich einerseits das Sicherheitsniveau erhöhen und andererseits generell die Motivation, das Vertrauen und vor allem die Akzeptanz der Nutzer*innen für den Einsatz der KI erhöhen.

Bedarf 08-13: Vorgehensweise für die Sicherheitsbetrachtung relevanter Stakeholder
Die Mehrzahl der Fragestellungen rund um das Management der Informationssicherheit in Unternehmen, z. B. ISO/IEC 27001 bzw. IT-Grundschutz, hat auch einen unternehmensinternen Scope. Die Betrachtung des Schutzbedarfs zur Verfügung gestellter Produkte und Dienste in der Anwendung relevanter Stakeholder, insbesondere von Verbraucher*innen, wird in den genannten ISMS nicht betrachtet. Die zu diskutierende Vorgehensweise soll aufgrund der hohen Individualität jeder einzusetzenden KI und der damit verbundenen, immer wieder neuen Beurteilung der Kritikalität gerade im sehr sensiblen Bereich der Finanzdienstleistungen einen unterstützenden Leitfaden bieten.

Bereits in der Ideenphase einer neuen KI sind die relevanten Stakeholder zu ermitteln, deren Schutzbedarf ist festzustellen und entsprechende Maßnahmen „KI-Security by Design“ sind zu entwickeln.

Beispiel: Dem Kunden bzw. der Kundin werden Hard- oder Softwareschnittstellen zu KI-Systemen zur Verfügung gestellt (z. B. Software: Apps/Marketing für Empfehlungen von Geldanlagen; Hardware: Sensorik z. B. im Fahrzeug für Telematiktarife).

Stakeholder eines Unternehmens sind in Bezug auf die Informationssicherheit von Maßnahmen des Unternehmens abhängig und müssen darauf vertrauen. Gerade im Bereich KI ist dieses Vertrauen essenziell, da KI in der Regel individuell in Entstehung und Kritikalität ist. Vertrauen kann teilweise über die Zertifizierung von Managementanforderungen wie einem ISMS realisiert werden, wobei Produkte und Dienste, die sich z. B. direkt an Endverbraucher*innen richten, davon nicht erfasst sind. Eine standardisierte Vorgehensweise würde den Blick für alle Anspruchsgruppen öffnen und mit einer transparenten Vorgehensweise mehr Vertrauen in KI schaffen und das Sicherheitsniveau insgesamt erhöhen.

Besonders im Finanzbereich sind neben Betreibern und Verbraucher*innen zahlreiche (behördliche) Stakeholder zu erkennen und durch die Einstufung von Finanzdienstleistungen als Kritische Infrastrukturen haben vertrauensbildende Maßnahmen eine besondere Relevanz.

Bedarf 08-14: Normen für die Validierung des Modells, um bewerten zu können, ob das KI-System für den Einsatz in der produktiven Umgebung hinreichend überprüft wurde
Hinreichende Generalisierbarkeit eines KI-Systems muss gewährleistet sein, um somit in zukünftigen Situationen zuverlässig entscheiden zu können. KI-Systeme neigen zu Over- und Underfitting bei nicht adäquater Entwicklung; daher ist es von hoher Relevanz, das Modell hinreichend zu validieren, um einen zuverlässigen Betrieb in der Produktion sicherstellen zu können. So muss das Modell entsprechend

durch adäquate Methoden (u. a. Back-Testing, Stresstests, Adversarial Attacks) geprüft werden mit dem Ziel einer harmonisierten Richtlinie für die Überprüfung der KI-Systeme. Es muss gewährleistet sein, dass ML-Methoden, die Gegenstand aufsichtlicher Prüfungen und Erlaubnisverfahren sind (interne Modelle zur Berechnung der regulatorischen Eigenmittelanforderungen (Säule 1) oder im Risikomanagement in Säule 2), hinreichend validiert sind. Um eine entsprechende Qualität sicherstellen zu können, müssen adäquate Normen definiert sein, da existierende regulatorische Anforderungen derzeit noch nicht die besonderen und komplexen Eigenschaften, die KI- und Machine-Learning-Technologien künftig enthalten, berücksichtigen.

Die besondere Relevanz für den Finanzsektor ergibt sich daraus, dass sich die Modelle häufig auf menschliches Verhalten sowie veränderliche Umgebungen, z. B. Marktumfelder, beziehen und Stressperioden mit abdecken. Entsprechend robust müssen die Prognosen sein.

Bedarf 08-15: Normen für die Transparenz zur Fehlerkorrelation des Systems
Ein KI-System soll in standardisierter Weise transparent machen, wie die Korrelationsstruktur der statistischen Unsicherheiten aussieht. Statistische Unsicherheiten der Ausgaben eines KI-Systems sind nicht notwendigerweise unabhängig. Für das Risikomanagement möglicher Fehler des Systems ist eine Kenntnis der Abhängigkeitsstruktur entscheidend. Zudem muss definiert werden, inwiefern ein Input unter Unsicherheit erstellt wurde (durch ein vorgeschaltetes Modell oder einen Datensatz).

Bedarf 08-16: Definition hinreichender Maße für Transparenz, damit der Entwickler weiß, welche zusätzlichen Informationen bereitgestellt werden müssen, um die entsprechende Architektur des KI-Systems zu konstruieren
Die Entscheidung eines KI-Systems muss hinreichend nachvollziehbar sein, um die Entscheidungsfindung zu verstehen. Außerdem sollten Transparenzanforderungen in Normen aufgenommen werden, etwa indem Auftragnehmende verpflichtet werden, die Überprüfung durch Dritte sowie die Schaffung von Nachvollziehbarkeit aktiv zu unterstützen. Es sollte u. a. im Fokus stehen, zu verstehen, was genau Einfluss auf die resultierende Entscheidung hat, wie z. B., um einem Darlehenbewerber ggf. erklären zu können, aus welchem Grund dieser kein Darlehen gewährt bekommt.

Die besondere Relevanz für den Finanzsektor ergibt sich daraus, dass sich die Modelle häufig auf menschliches Verhalten sowie veränderliche Umgebungen, z. B. Marktumfelder, beziehen und Stressperioden mit abdecken. Entsprechend robust müssen die Prognosen sein.

Bedarf 08-17: Normung von Dokumentationspflichten zum Ursprungskontext von Modellen und (Trainings-) Daten
Der Ursprungskontext von (Trainings-)Daten sowie fertigen Modellen muss in jedem Schritt der Verwendung verfügbar sein, um eine Überprüfbarkeit zu gewährleisten.

Werden Ergebnisse einzelner Modelle rein anhand ihrer Klassifizierung genutzt, kann dies zu unerwartetem Systemverhalten führen. Aufgrund der Prozesse im Maschinellen Lernen können unwichtig erscheinende Randdaten zu unerwünschten Korrelationseffekten führen. Für die Verwendung anderer Modellausgaben sollte der ursprüngliche Kontext bekannt sein und berücksichtigt werden. Dies schließt ein: Modellentscheidungen und Abwägungen, Ursprung und Kontext der Trainings-, Validierungs- und Testdaten, Ursprung und Kontext der Echtzeiteingabedaten. Normung kann hier ansetzen und sicherstellen, dass keine relevanten Informationen während der Übernahme unbekannt bleiben.

Bedarf 08-18: Normen für die Transparenz zur Konfidenz und Modellrisiken von Einzelentscheidungen
Im Gegensatz zu Entscheidungen mit vorgegebenen Algorithmen gehört die Unsicherheit über die Richtigkeit der Entscheidungen zur Ausgabe des ML-basierten KI-Systems. Diese sollten daher in genormter Weise transparent gemacht werden, z. B. durch die Angabe entsprechender Wahrscheinlichkeiten der möglichen Entscheidungen.

Da im Risikomanagement oft mehrere Modelle verkettet werden, diese aber nichtlinear gekoppelt werden, ist die Kenntnis über die Fehlerverteilungen der Einzelsysteme für die Abschätzung der Fehlerverteilung des Gesamtsystems entscheidend. Das ist für Finanzdienstleister als native Risikomanager von grundlegender Bedeutung.

Die Arbeitsgruppe Finanzdienstleistungen hat die identifizierten Bedarfe nach der Dringlichkeit ihrer Umsetzung bewertet. Abbildung 46 zeigt die Dringlichkeit der Umsetzung, kategorisiert nach den Zielgruppen Normung und Forschung.

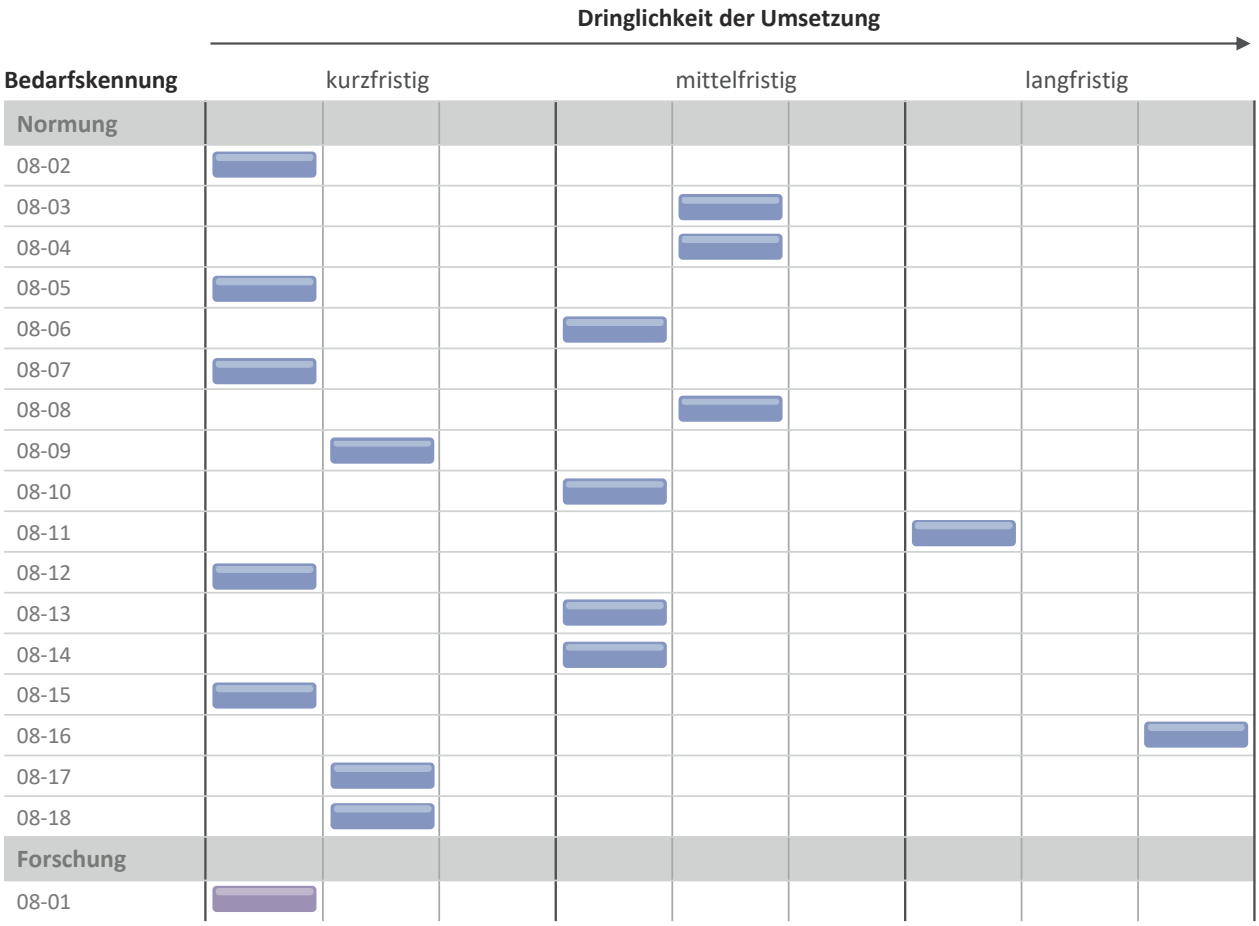


Abbildung 46: Priorisierung der Bedarfe aus Schwerpunkt Finanzdienstleistungen (Quelle: Arbeitsgruppe Finanzdienstleistungen)

4.9 Energie und Umwelt

Künstliche Intelligenz (KI) drängt in vielfältige Anwendungsbereiche vor. Im integrierten Bereich Energie und Umwelt ist mithin ein komplexes Gefüge aus domänenspezifischen und -übergreifenden Anwendungen festzustellen. Zugleich werden diese Anwendungen für eine zunehmende Bandbreite von Problemstellungen genutzt. Im spezifischen Aspekt der Energiesysteme und -technik entsteht die Fragestellung, inwieweit KI als Set neuer Technologien mit bestehenden Systemen verknüpfbar ist und diese verändert. In der Umwelttechnik kann KI die Entwicklung von Kreislaufprozessen und Dekarbonisierungsstrategien unterstützen sowie Konsument*innen Feedback zu Kaufentscheidungen geben. Bereichsübergreifend sind Umweltaspekte bezüglich der Weiterentwicklung von Energieeffizienz sowie der Feststellung von Energiebedarfen und Umweltwirkungen der KI-Methoden selbst relevant. Es gibt KI-Anwendungen, die explizit zu Energieeffizienz und Umweltschutz beitragen sollen, wie z. B. die Optimierung tribologischer Systeme (vgl. [396]). Gleichzeitig bedarf die Entwicklung und Anwendung von KI in jedem Anwendungsfeld Energie für die Rechenleistung der technischen Infrastruktur sowie spezifischer Materialien und Rohstoffe, die ihrerseits für Umweltwirkungen im Lebenszyklus verantwortlich zeichnen.

Um die Entwicklungen und Anwendungen der vielfältigen KI-Technologien integrativ, energiesparend und umweltschonend und zum menschlichen Nutzen zu gestalten, sind Normungsprozesse in vielen Bereichen erforderlich. Vorliegend kann nur eine Auswahl von Normungsbedarfen erfasst und beschrieben werden. Dieses Kapitel fokussiert die Bereiche Energietechnik und Umweltwirkungen. Es bringt KI als innovative, für den Energiesektor immer noch neue Technologie mit den bewährten Systemansätzen und Anwendungsmöglichkeiten der Normungsexpert*innen der Energietechnik zusammen. Die Normungsexpert*innen haben eine funktionierende Architektur erstellt, in der Normen und Standards Interoperabilität sicherstellen. Die Entwickler*innen der KI offerieren hingegen Ideen und Anwendungen, die diese Architektur erweitern. Für die Bestimmung von Umweltwirkungen bietet die KI ebenfalls neue Perspektiven, indem sie beim Management komplexer, domänenübergreifender Datensysteme unterstützt. Im umwelttechnischen Kontext stellt KI analog zur energietechnischen Domäne ein neuartiges Technologieset dar. Normungsexpert*innen und KI-Entwickler*innen haben eine gemeinsame Architektur entworfen, die die Interoperabilität und Notwendigkeit sektorenübergreifender Datenstandards verdeutlicht.

4.9.1 Status quo

Die Versorgung mit Energie ist nach wie vor ein Hauptthema der politischen Agenda. Mit der in Deutschland eingeleiteten Energiewende und den aktuellen, dramatischen weltpolitischen Veränderungen sollen unterschiedlichste Ziele wie Wirtschaftlichkeit, Versorgungssicherheit, Klimaschutz und die Umstellung auf erneuerbare Energien gleichzeitig erfüllt werden. Dabei spielt das Smart Energy Grid, die Verbindung von Energietechnik mit Informations- und Kommunikationstechnologien (IKT), eine entscheidende Rolle. Die Normung wiederum ist eine notwendige Voraussetzung für die technische Umsetzung und Investitionssicherheit in diesem Bereich. Die Einführung der Künstlichen Intelligenz hat nun einen immensen Effekt auf den Status quo in den entsprechenden Feldern der Smart-Energy-Grid-Normung. Hierzu gehören die Vielzahl der Akteur*innen, der regionalen und internationalen Aktivitäten sowie die enorme Geschwindigkeit der Entwicklung. Viele dieser Besonderheiten sind mittlerweile durch die Aktivitäten des System-Komitees „Smart Energy“ (DKE/K901) in der DKE seit über zehn Jahren adressiert. Die wesentlichen Auswirkungen der KI auf die etablierten Strukturen sollen hier untersucht werden.

In den letzten Jahren ist im Zusammenhang mit den Normungsaktivitäten im Bereich Smart Grid eine neue Herangehensweise an die Normung an sich etabliert worden, die den vielfältigen Herausforderungen in komplexen Systemen Rechnung trägt. Wesentlich ist dabei die Integration unterschiedlichster Teilgebiete und der entsprechenden betroffenen Fachkreise. Dies gelingt über die Ausrichtung der Aktivitäten auf die gewünschten oder geforderten Dienste, die das komplexe System Smart Grid anbieten soll. Auf der Basis dieser Dienste oder Funktionen untersucht man mithilfe eines generischen Modells (Smart Grid Architecture Model – SGAM) die Umsetzungsmöglichkeiten. Durch die Beschreibung der Dienste und die zunehmende Detaillierung in Anwendungsfällen, sogenannten Use Cases, auf Funktions-, Informations-, Kommunikations- und Komponentenebene schafft man die Voraussetzung, dass die unterschiedlichsten beteiligten Normungsgremien zusammen an einem gemeinsamen Ziel arbeiten – der Realisierung der gewünschten Dienste und Funktionen. Dieses Verfahren gewährleistet nicht nur eine kohärente Normungsarbeit, es liefert zudem die notwendige Grundlage für ein gemeinsames Verständnis und die Konsensbildung zwischen allen Parteien. Zudem ist es gelungen, die Sammlung der grundlegenden Dienste und Funktionen weit über den etablierten Teilnehmerkreis der Normung hinaus zu öffnen.

Die Auswirkungen von KI-Technologien auf Umwelttechnik stellen ein komplexes Gefüge dar. Die Optimierung von Systemen und Prozessen zur Maximierung der Energieeffizienz und Minimierung der Umweltwirkungen ist der Kern von KI-Anwendungen im Umweltbereich und eine Schlüsselkomponente in der Erreichung globaler und nationaler Klimaschutzziele. Dies betrifft beispielsweise die Minimierung von Reibungsverlusten (Tribologie), die Pfadoptimierung (Logistik) und die Bestimmung von Sanierungspfaden (Bauwesen). Zugleich stellt KI auf der Metaebene ein energie- und ressourcenintensives Technologieset dar. Der Energiebedarf bzw. -verbrauch und die Umweltwirkungen von KI-Anwendungen sind mithin ein prinzipielles Kriterium für die Beurteilung der Güte Künstlicher Intelligenz in allen Anwendungsbereichen. Die gemeinsamen Normungs- und Standardisierungsaktivitäten von DIN, DKE und VDI (Verein Deutscher Ingenieure) sowie die gemeinsamen Aktivitäten des Europäischen Komitees für Standardisierung (CEN) und des Europäischen Komitees für Elektrotechnische Standardisierung (CENELEC) haben eine Bandbreite an umweltbezogenen Anwendungsfällen der Künstlichen Intelligenz hervorgebracht. Diese werden im politischen Umfeld durch Positionspapiere, sektorübergreifende Publikationen und Regularien u. a. seitens des Bundesministeriums für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz (BMUV) und des Europäischen Parlaments flankiert. Hierbei wird KI auch als genereller Impulsgeber für die Umwelt- und Nachhaltigkeitsforschung betrachtet. Vorliegend soll für KI im Umweltbereich der Status quo in Form der wesentlichen Normungsbestrebungen, politischen Zielsetzungen, Verbandspositionen und Forschungsaktivitäten dargelegt werden. Im Hinblick auf das vielschichtige Spannungsfeld zwischen ökonomischen, ökologischen und sozialen Aspekten der Nachhaltigkeit erfolgt vorliegend keine Betrachtung soziotechnischer Systeme.

Energietechnik

Intelligente Energie und intelligente Energienetze müssen um Echtzeit-Datenerfassungs-, Kommunikations-, Überwachungs- und Steuerungsfunktionen erweitert werden, um Ausfälle zu beheben, die zunehmend dezentralisierte Stromerzeugung zu steuern, erneuerbare Energien und Energiespeicher hinzuzufügen und gleichzeitig strengere Emissionsziele einzuhalten. Eine weitgehend elektrifizierte und automatisierte Welt erfordert eine kontinuierliche, zuverlässige und nachhaltige Versorgung mit Strom. Dies wird durch ein Netz erreicht, das in der Lage ist, Informationen zu sammeln und zu kommunizieren. Idealerweise basiert es auf standardisierter Hardware, Software und Prozessen, die eine nahtlose Integration und Interoperabilität gewährleisten.

Elektrizität ist das ultimative Just-in-time-Produkt. Sie wird in dem Moment verbraucht, in dem sie erzeugt wird, und muss kontinuierlich geliefert werden. In Zeiten hoher Stromnachfrage werden die Anlagen extrem belastet. Viele der heutigen Stromnetze wurden in den 1960er-Jahren, manchmal sogar noch früher, gebaut und erreichen das Ende ihrer Nutzungsdauer. Die Modernisierung der Netze unter Einsatz der neuesten Technologien ist daher ein Muss. Sie trägt auch dazu bei, die Energieeffizienz zu verbessern und die Erzeugung, Übertragung und den Verbrauch von Energie nachhaltiger zu gestalten. Zu den Schlüsseltechnologien, die für intelligente Netze eingesetzt werden, gehören Sensoren, die die relevanten Parameter wie Temperatur, Spannung und Stromstärke messen; Kommunikationssysteme, die einen Zwei-Wege-Dialog mit einem Gerät ermöglichen; Steuersysteme, die es ermöglichen, ein Gerät aus der Ferne neu zu konfigurieren; Benutzerschnittstellen- und Entscheidungsunterstützungssysteme, die einen Überblick über den Zustand der Anlagen geben und fortgeschrittene Datenanalysen durchführen.

Mehrere Technische Komitees der IEC entwickeln die Normen, die dazu beitragen, die Anpassungsfähigkeit der Netze zu verbessern, damit sie mit Mehrwege-Stromflüssen, der Integration erneuerbarer Energiequellen und der Energiespeicherung zurechtkommen und kostengünstiger, sicherer, zuverlässiger und flexibler werden. IEC TC 57 entwickelt Schlüsselstandards für Smart-Grid-Technologien und deren Integration in bestehende Stromnetze. Viele andere IEC TCs tragen mit Normen für Sensoren, intelligenten Schaltern, automatisierten Umspannwerken oder intelligenten Zählern, um nur einige zu nennen, zu intelligenten Netzen bei. Solche Normen dienen auch als Grundlage für die Prüfung und Zertifizierung von Komponenten, Geräten und Systemen. IEC betreibt vier Konformitätsbewertungssysteme (CA), deren Mitglieder überprüfen, ob Geräte und Systeme die Anforderungen der IEC-Normen und -Spezifikationen erfüllen. Die IEC hat ein Systemkomitee, SyC Smart Energy, eingerichtet, um die Normung auf Systemebene für intelligente Energie und intelligente Netze zu gewährleisten. Das SyC hilft bei der Identifizierung aller relevanten Normen und koordiniert die Arbeit der vielen technischen Komitees, die an der Normung für intelligente Energie beteiligt sind. Die IEC hat eine Roadmap für die Normung intelligenter Netze veröffentlicht, die Leitlinien für die Auswahl der am besten geeigneten Normen und Spezifikationen enthält.

In dieses etablierte System aus Datenmodellen und Systemarchitekturen muss nun die neue Technologie KI integriert werden. Der Vorteil der vorhandenen Systeme ist, dass die

Schnittstellen und Prozesse bereits vorhanden sind und z. B. zur Entscheidungsfindung für KI-Systeme genutzt werden könnten.

Umwelt

Klimaschutz und Dekarbonisierung bedürfen umfassender Strategien zur Senkung von Energie- und Ressourcenverbräuchen und Emissionen (Umweltwirkungen). Aktuelle Lösungsansätze zur Bestimmung dieser Umweltwirkungen und deren Kommunikation an Marktakteur*innen und Verbraucher*innen sind charakterisiert durch einen hohen Datenbedarf und eine starke Datendynamik. Hier bestehen Chancen zur Maximierung der Energieeffizienz, Minimierung der ökologischen Implikationen und Begleitung nachhaltiger Konsumentscheidungen (vgl. [397]). Demgegenüber existieren in der Anwendung von KI immanente umweltbezogene Risiken (vgl. [223]). Aufbau und Anwendung von Künstlicher Intelligenz und Maschinellem Lernen (ML) sind prinzipiell gekennzeichnet durch eine hohe Rechen- und Ressourcenintensität, wodurch der Zusatznutzen umweltbezogener KI- und ML-Anwendungen im Spannungsfeld zu deren eigenen Umweltwirkungen steht (vgl. [398]). Insofern ist ein entschiedenes und kontinuierliches Monitoring prozessbezogener Daten entlang des vollständigen Lebenszyklus von Produkten und Dienstleistungen erforderlich. Dies gebietet einheitliche Datenstandards und Übersetzungen (sogenannte Mappings) für verschiedene Datenformate und -umgebungen.

Die Forschung identifiziert KI als bedeutendes Instrument im Umweltbereich zur Erreichung der durch die Vereinten Nationen (UN) definierten 17 Sustainable Development Goals (SDG) (vgl. [399], [400]) und zum Aufbau nachhaltiger Geschäftsmodelle (vgl. [401]). Auf der europäischen Politikebene werden KI-Nutzungspotenziale im Kontext des European Green Deal (vgl. [402]) gesehen. Das Policy Department for Economic, Scientific and Quality of Life Policies identifiziert sektorübergreifende und -spezifische Maßnahmen. Sektorübergreifend sollen mithilfe von KI Verhaltensempfehlungen für Marktakteur*innen und Verbraucher*innen zur Minimierung des ökologischen Fußabdrucks sowie Maßnahmen zur Verringerung der Umweltwirkungen von KI selbst entwickelt werden. Sektorspezifisch werden im Energie- und Gebäudesektor bislang ungehobene Potenziale in der Minimierung von Energieverbräuchen und zugehörigen Emissionen im Lebenszyklus gesehen. Im Mobilitätssektor soll KI zur Optimierung und Automation von Transportrouten sowie zum Vehikeldesign beitragen. Für den landwirtschaftlichen Bereich wird eine Fokuserweiterung in der KI-Nutzung von Produktivitätsmaximierung hin zur Reduktion von Düngemittel-, Wasser- und Flächennutzung empfohlen (vgl. [403]). Im europäischen Finanzwesen stellt die Offenlegungspflicht nachhaltigkeitsbezogener Investmentindikatoren „Environmental Social Governance“ (ESG) (vgl. [404]) im Kontext mit der Taxonomie nachhaltiger Aktivitäten (vgl. [405]) eine Herausforderung dar, zu deren Lösung ML- und KI-gestützte Systeme einen wesentlichen Beitrag leisten können (vgl. [406]). Das deutsche BMUV sieht KI im Umweltbereich als Werkzeug für Ressourceneffizienz in Industrie und Mittelstand sowie zur Verarbeitung großer Datenmengen in verschiedenen Wirtschaftssektoren. Weiterhin soll mithilfe von KI die ressourcenschonende Gestaltung von KI- und ML-Modellen begleitet und zur informativen Reichweite von Energie- und Umweltindikatoren von Produkten und Dienstleistungen genutzt werden (vgl. [407], [398]).

Im Rahmen der Normungsaktivitäten der CEN CENELEC ist eine Road Map on Artificial Intelligence (AI) entstanden, die eine Bandbreite an den Technical Committees (TC) zugehörigen Anwendungsfällen enthält.[98] Die Road Map ergänzt diese hochspezifischen Fragestellungen um die Notwendigkeit sektorübergreifenden Forschungsbedarfs für KI-Systemarchitektur, Algorithmik und das Themenfeld Ethik, insbesondere die Bereiche Datenschutz, -transparenz, Rechenschaft und Erklärbarkeit. Es wird eine übergreifende Kooperation mit weiteren Standardisierungsorganisationen (International Standardization Organisation (ISO), IEC) erörtert (vgl. [392]). Die gemeinsame Normungslandkarte zur Ressourceneffizienz von DIN, DKE und VDI identifiziert Normen und Standardisierungsaktivitäten zur Umsetzung des Deutschen Ressourceneffizienzprogramms (ProgRess III) (vgl. [408], [409]). Betroffen sind Wertschöpfungsprozesse und Prozessketten in allen Sektoren bezüglich der Produktion und Logistik, der Digitalisierung sowie der Kommunikation von Energieverbräuchen und Umweltwirkungen. Die DIN-Normenausschüsse NA 172 Grundlagen des Umweltschutzes (NAGUS) und NA 005 Nachhaltiges Bauen (NABau) haben bereits umfangreiche Spezifikationen zu Ökobilanzierungen, zugehörigen Datendokumentationsformaten und Produktdeklarationen sowie Kommunikationsanforderungen von Fußabdruckinformationen erarbeitet (vgl. [410], [411], [412], [413], [414]). Zugleich bearbeitet der Normenausschuss NA 043 Informationstechnik und Anwendungen (NIA) in verschiedenen Gremien und Ge-

98 JTC5 – Space, TC61- Safety of household and similar electrical appliances, TC64 – Electrical installations and protection against electric shock, TC134 – Resilient, textile and laminate floor coverings, TC248 – Textiles, TC307 – Blockchain and distributed ledger technologies, TC332 – Laboratory equipment, TC348 – Facility management

meinschaftsarbeitsausschüssen eine Bandbreite an technischen und organisatorischen Fragestellungen zu Architektur und Einsatz von KI.

Es gilt, die etablierten Regelungen zu Umweltschutz und Nachhaltigkeit in einen synergetischen Kontext mit KI-bezogenen Werkzeugen und Prozessen zu setzen. Hierbei kann ein Fokus auf Datenumgebungen und die Harmonisierung von Datenstandards unterschiedlicher Disziplinen einen Beitrag zur Hebung von Nutzenpotenzialen der KI-Anwendungen im Umweltbereich leisten.

4.9.2 Anforderungen und Herausforderungen

Der in Kapitel 4.9.1 dargelegte Stand der Technik zeigt die aktuelle Divergenz zwischen bereits etablierten Systemen sowie technisch prinzipiell umsetzbaren Lösungen einerseits und Sicherheitsarchitekturen sowie Datenstandards andererseits auf. Für die Entwicklung sicherer und effizienter Systeme in Energie und Umwelt bedarf es eines interdisziplinären Ansatzes, der gemeinsame Standards für Sicherheitssysteme und Datenformate etabliert. Hierbei muss die Kommunikation von und mit Akteur*innen durchgehend gewährleistet werden. Aufgrund der breit gefächerten Sachlage kann vorliegend nur eine Auswahl konkreter Anforderungen und Herausforderungen ausgeführt werden. Bei der Planung, Errichtung und dem Betrieb von neuen Energie- und Informationsstrukturen spielen Normung und Standardisierung eine gewichtige Rolle. In der Bestimmung und Kommunikation von Umweltwirkungen entstehen branchenübergreifende Normungsbedarfe. Existierende Normen und Spezifikationen aus ganz unterschiedlichen Technologiegebieten müssen zusammengeführt, auf Kompatibilität untersucht und interdisziplinär angewendet werden. Aufgrund neuer Marktanforderungen entstehen neue Funktionalitäten und Schnittstellen, die zu neuen Normen und Spezifikationen führen werden. Dies gilt nicht zuletzt für den Bereich der Interoperabilität im Bereich Energie, die Fachleuten und Laien gleichermaßen als Systemnutzer*innen einen Zugang zu den Optimierungsaufgaben bieten muss. Umweltwirkungen sollen im Einklang mit etablierten Methoden skalierbar, bestimmbar und kommunizierbar sein. Eine wichtige Rolle für die vom Menschen zu leistende Spezifikation von Funktionen und Schnittstellen spielt die Methodik der Anwendungsfälle, sogenannte Use Cases. Neben diversen Beschreibungsvorlagen für die Normungsgremien werden strukturierte Ablage- und Suchfunktionen für Use Cases bereitgestellt. Diese Methodik hat sich im internationalen Informationsaustausch zwischen den Normungsgremien wie z. B.: IEC TC 57 bereits bewährt und unterstützt das Ziel, durch internationale Normung und Standardisierung eine solide Basis für den Auf- und Ausbau von Smart Energy Grids zu schaffen. Die nachfolgende Auswahl an Use Cases (siehe Tabelle 10) wird in Anlehnung an DIN EN 62559-2:2016-05; VDE 0175-102:2016-05:2016 [415] strukturiert. Die jeweiligen tabellarischen Zusammenfassungen sind dem Anhang zu entnehmen (siehe Tabelle 21 bis Tabelle 26 in Anhang 13.6).

Tabelle 10: Übersicht der Anwendungsfälle im Themenbereich Energie/Umwelt

Nummer	Name	Kurzbeschreibung
1	Autonomes Smart Grid Power Management and Consumption System	Power Management System (PMS) und Industrial Automation and Control Systems (IACS) werden je autonom entworfen und betrieben, erzeugen jedoch als gekoppeltes System wechselseitige Abhängigkeiten, die in Realzeit und kurzen Zeitspannen kontrolliert und ggf. ausgeglichen werden müssen.
2	Energieeffizienz in Gebäuden und Kopplung mit Energienetzen	Optimierte Anpassung des Strombedarfs von Gebäuden an prognostizierte Lastgänge in der Erzeugung
3	Personalisierte, KI-gestützte Empfehlungssysteme für nachhaltigen Konsum	Personalisierte, KI-gestützte Empfehlungssysteme für nachhaltigen Konsum gleichen Produkteigenschaften und individuelle Einstellungen ab und geben passgenaue Produktempfehlungen in verschiedenen Einkaufssituationen.
4	Skalierbare Bestimmung von Umweltwirkungen im Gebäudesektor	Bestimmung der Umweltwirkungen von Gebäuden und Quartieren mit angepasster Detailtiefe der Daten in der Ökobilanzierung
5	Ressourcenintensität von KI & ML	Integration einer Metrik/eines Referenzverfahrens für Umweltwirkungen von KI & ML-Modellen in deren Bewertung
6	Adversarial Resilience Learning – Marktlicher Eingriff durch Aggregatoren im Verteilnetz	Vermeidung potenzieller Angriffe auf das Netz im Rahmen des Engpassmanagements in volatilen Lastgängen

4.9.2.1 Anwendungsfall 1: Autonomes Smart Grid Power Management and Consumption System

Autonome Gridsysteme sind weit verteilte Systeme, die mit mobilen Daten, Dingen oder Energieflüssen belebt und mit stationären Gegenständen, Produktionsstätten und Gebäuden ausgerüstet werden. Autonomie in Smart Grids erfordert verfügbares, gespeichertes Wissen über mögliche kritische Zustände bzw. Situationen, die zu vermeiden sind, und Technologien zur dynamischen Steuerung und Regelung von Komponenten oder Teilnetzen (grids). Alle Subsysteme im Grid, z. B. Grid Power Management System (PMS) und Home, Building, IACS, werden jede für sich autonom entworfen und betrieben. Als gekoppeltes System erzeugen sie jedoch Abhängigkeiten voneinander, die in Realzeit und in kurzen Zeitspannen kontrolliert und ggf. ausgeglichen werden müssen. Diese Abhängigkeiten beeinflussen die Stabilität des Gesamtsystems, z. B. führen hoher Energiebedarf und niedrige Energieeinspeisung zur Destabilisierung. Hinzu kommt die Resilienz gegenüber Ausfällen von „Distributed Energy Resources"-Komponenten (DER), die sich zu unkontrollierbaren kaskadierenden Effekten hochschaukeln können. Aus der Architekturbeschreibung des UC SGAM, Smart Grid Reference Architecture Modell, ergeben sich **mindestens drei miteinander kooperierende Systeme** [PMS, System Interface (SIF), IACS], wobei jedes System als Vektor von Variablen dargestellt wird. Jede Belegung der Variablenvektoren beschreibt einen Systemzustand, der durch Inzidenzen verändert wird. In stabilen Systemzuständen von SGAM-Systemen transportieren und transformieren die Variablen Energie. Wenn das System instabil wird, droht eine **Blackout-Inzidenz**. Die Beobachtung von kritischen Systemzustandsveränderungen, z. B. vom Übergang vom stabilen in den instabilen Systemzustand, ist eine der Aufgaben des Digitalen Zwillings, der mit analytischen Fähigkeiten ausgestattet ist. Die Analyse der kritischen Inzidenzen ist datenbasiert. Alle kritischen Inzidenzen müssen vorher mit validen und zeitnahen Metadatenerhebungen transparent dokumentiert werden. Der Digitale Zwilling nutzt Analysewerkzeuge und kennt Maßnahmen zu geeigneten Reaktionen auf möglicherweise auftretende Blackout-Inzidenzen. Blackouts können durch Daten, die ontologisches Systemwissen repräsentieren, von der Systemkontrolle Digitaler Zwillinge genutzt und mögliches Fehlverhalten kann ggf. erkannt und vermieden werden. Mit der ML-Technologie können diese Daten nach Mustern der Instabilität durchsucht und identifiziert werden.

4.9.2.2 Anwendungsfall 2: Energieeffizienz in Gebäuden und Kopplung mit Energienetzen

Da erneuerbare Energien nicht konstant zur Verfügung stehen, erfordert der Ausbau der Erneuerbaren eine flexible Energienutzung. Mit über 40 % Anteil am Energieverbrauch bieten Gebäude großes Potenzial für flexible Energienutzung. So können Klimaanlagen, Heizungen, Warmwasserbereitung oder auch Ladestationen für Elektrofahrzeuge dazu genutzt werden, den Stromverbrauch im Gebäude zeitlich zu steuern und Energie vermehrt dann zu verbrauchen, wenn diese erzeugt wird. Damit können Energienetze stabilisiert und zeitgleich der CO_2-Fußabdruck von Gebäuden reduziert werden. Hierfür müssen einerseits Prognosen für das Energienetz und dessen CO_2-Faktor erstellt werden als auch individuelle Prognosen für die Energienutzung im Gebäude.

Auf Basis von historischen Wetter- und Gebäudedaten kann Künstliche Intelligenz eine Vorhersage für die Gebäudenutzung ermitteln. So können automatisch die Raumbelegungen sowie die Energieverbräuche vorhergesagt und in Zusammenhang mit Energienetzdaten ein optimierter Energienutzungsplan ermittelt werden. CO_2- und Energieeinsparungen von bis zu 40 % sind damit möglich [416] (vgl. auch Abbildung 47 und Abbildung 48).

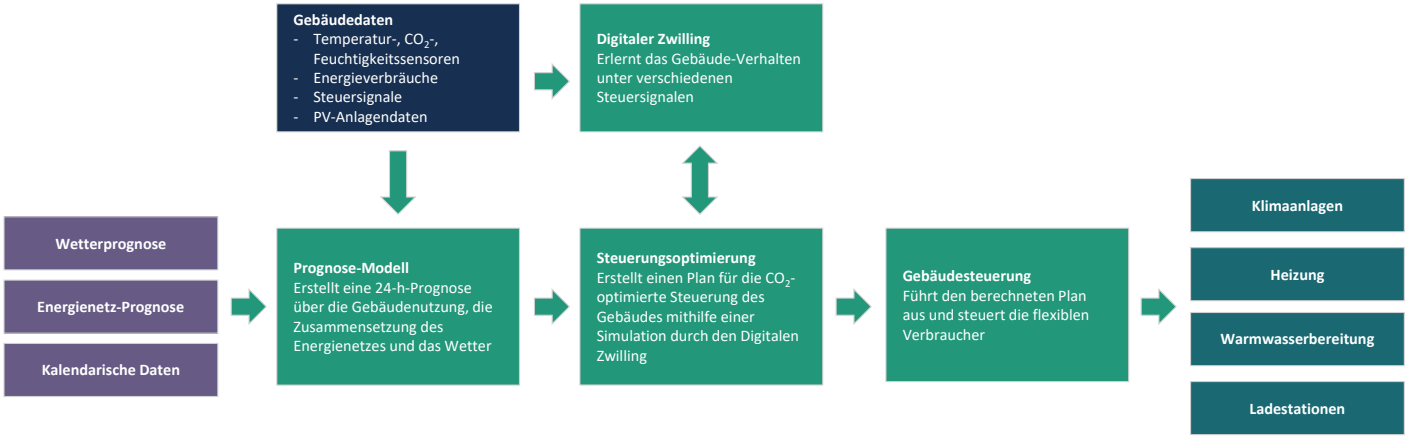


Abbildung 47: Schema zur Optimierung und Steuerung von Gebäuden (Quelle: Unetiq GmbH)

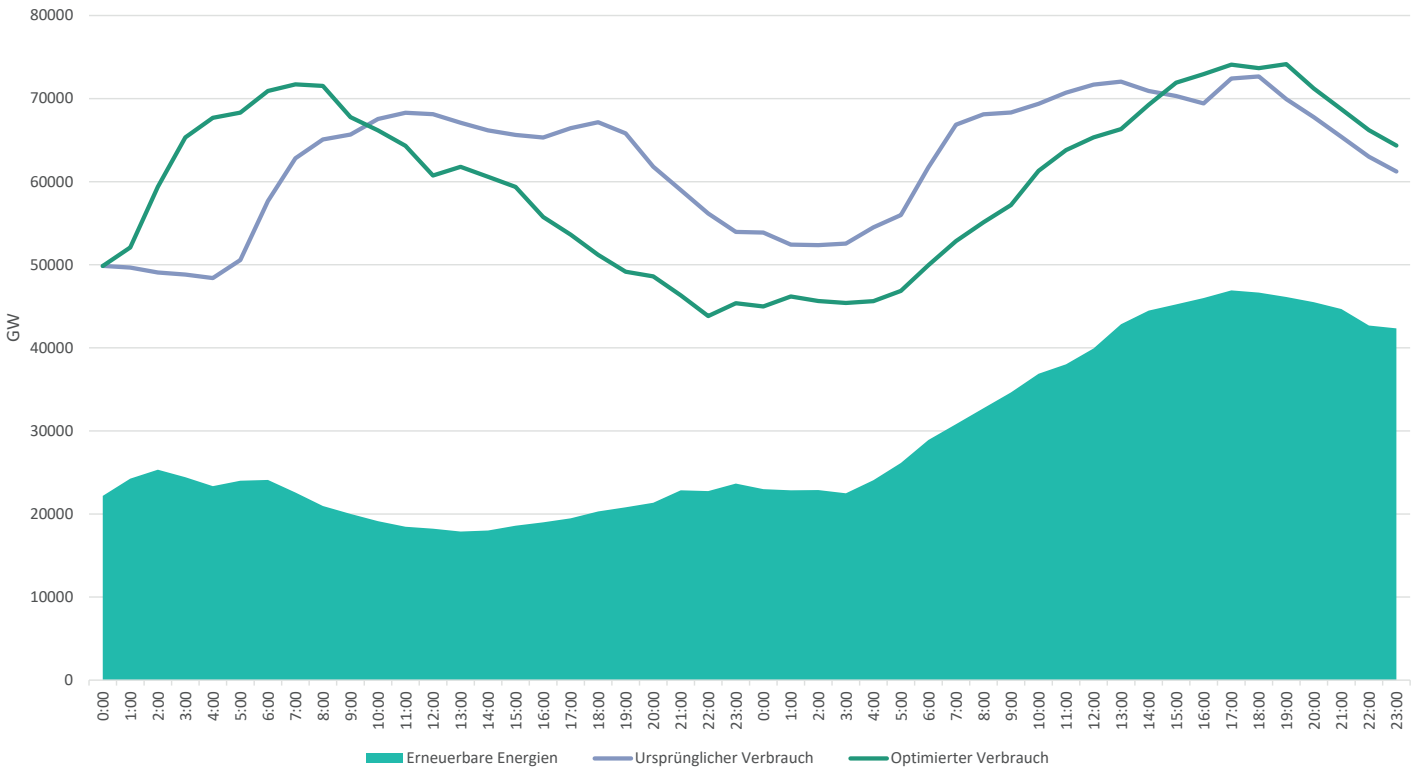


Abbildung 48: Zeitlicher Verlauf der erneuerbaren Energie, des ursprünglichen und des optimierten Verbrauchs in GWh (Quelle: Unetiq GmbH)

4.9.2.3 Anwendungsfall 3: Personalisierte, KI-gestützte Empfehlungssysteme für nachhaltigen Konsum

Der Konsum privater Haushalte und die damit einhergehende Güterproduktion verursachen einen erheblichen Anteil der weltweiten Treibhausgasemissionen sowie der globalen Energie und Rohstoffnutzung ([417]). Konsumentscheidungen haben folglich eine erhebliche Relevanz für Klima, Umwelt und Energie. Hindernisse für nachhaltigen Konsum sind die Intransparenz und Unübersichtlichkeit produktbezogener Nachhaltigkeitsinformationen beim Einkauf ([418], [420]).

Personalisierte, KI-gestützte Empfehlungssysteme für nachhaltigen Konsum adressieren dieses Problem, indem sie Produkteigenschaften und individuelle Einstellungen abgleichen und damit passgenaue Produktempfehlungen abgeben. Auf Basis einer kohärenten Datengrundlage, die sich z. B. über einen europaweit harmonisierten Digitalen Produktpass (DPP) [421], ([422]) ergeben könnte, ermöglicht die KI eine nachhaltigere Produktauswahl in verschiedenen Einkaufsentscheidungen ([423], [424]). KI könnte dann auch persönliche Einkaufsdaten im Zeitverlauf erfassen und auswerten, um wichtige Einsichten über das Einkaufsverhalten zu vermitteln und somit zur mittel- bis langfristigen bedarfsgerechten Optimierung von Konsummustern beizutragen.

Für eine breite Umsetzung solcher Empfehlungssysteme ergibt sich besonderer Normungsbedarf, der auf Kohärenz und Einheitlichkeit der Datengrundlagen und KI-Anwendungen für nachhaltigen Konsum abzielt. Dazu gehören die Vereinheitlichung der Umweltindikatoren, die im Rahmen der KI-Systeme genutzt werden, sowie die Gestaltung einheitlicher Datenmodelle und Schnittstellen für die Weitergabe von Umweltdaten zwischen Akteur*innen entlang von Produktketten. Zudem sollte versucht werden, die zu verwendenden Algorithmen zu harmonisieren und im Zuge dessen offene Schnittstellen und Interoperabilität zu gewährleisten. Weiterhin gilt es im Hinblick auf persönliche Daten innerhalb der KI-Anwendungen, Qualität und Schutz der Daten auf Basis von Datensicherheits-, Ethik- und Verbraucherschutznormen sicherzustellen.

4.9.2.4 Anwendungsfall 4: Skalierbare Bestimmung von Umweltwirkungen im Gebäudesektor

Die Klimaschutzziele erfordern die Entwicklung sektorspezifischer Dekarbonisierungsstrategien. Im Bausektor sind für den Gebäudebestand bauphysikalische und anlagentechnische Sanierungspfade zu definieren und für Neubauten klimaneutrale Standards festzulegen. Hierzu bedarf es einer umfassenden Nachhaltigkeitsbewertung und Abschätzung des Dekarbonisierungspotenzials von Bauteilen und Materialien. Lebenszyklusanalysen bzw. Ökobilanzen dienen der Ermittlung derartiger Energiebedarfe und Umwelteinträge. Es existieren eine breite Praxis und etablierte Normen zur Ökobilanzierung, die grundsätzlich durch einen hohen Informationsbedarf und Zeitaufwand geprägt sind (vgl. [425]). Gleichzeitig impliziert die lange Prozesskette im vollständigen Lebenszyklus mit vielen Einflüssen und Variablen eine hohe Volatilität der Bilanzierungsergebnisse.

Auf Basis von bauphysikalischen und anlagentechnischen Präferenzen können KI-Anwendungen prinzipiell Lösungsvorschläge für die klimaneutrale Planung von Gebäuden und Quartieren formulieren. Dazu braucht es im Hintergrund ein kontinuierlich lernendes System, das über die Verarbeitung relevanter Gebäude- und Quartiersdaten die Umwelteinträge im Lebenszyklus ermittelt.

4.9.2.5 Querschnitts-Anwendungsfall 5: Ressourcenintensität von KI & ML

Künstliche Intelligenz und Maschinelles Lernen dienen der Lösungsfindung und Effizienzsteigerung in vielfältigen Bereichen. Grundsätzlich sind KI- und ML-Modelle durch eine hohe Rechenlaufzeit und -leistung geprägt, die ihrerseits hohe Energieverbräuche und Umweltwirkungen implizieren (vgl. [398]). Der Zusatznutzen von KI- und ML-Anwendungen steht damit im Spannungsverhältnis zu deren Ressourcenverbräuchen. Dies gilt insbesondere für derartige Anwendungen, die der Steigerung der Ressourceneffizienz und der Minderung von Umwelteinträgen dienen sollen. Aus technischer Perspektive hängt der Ressourcenverbrauch prinzipiell von Datenbedarf und Laufzeit der Algorithmik ab.

Auf der übergeordneten Ebene kann KI Feedback zur Nachhaltigkeitsbeurteilung von KI- und ML-Anwendungen liefern. Hierzu bedarf es einer Metrik bzw. eines Referenzverfahrens zur Messung und Vergleichbarkeit der Performanz von Algorithmik.

4.9.2.6 Anwendungsfall 6: Adversarial Resilience Learning – Marktlicher Angriff durch Aggregatoren im Verteilnetz

In Verteilnetzen ist eine zukünftige, durch die Energiewende vor allem bedingte Herausforderung das sogenannte Engpassmanagement. Durch die Lastflussänderung kommt es dazu, dass auch auf unterster Ebene sogenannte Prosumer (elektrische) Energie nicht mehr nur verbrauchen, sondern auch aktiv in das Netz einspeisen. Der bisherige Ausbau der Netze sowie die Betriebsplanungen hatten diesen Aspekt nicht im ursprünglichen Fokus. Kommt es nun zu einer höheren Einspeisung als zum Verbrauch, treten Rückflüsse auf, die Engpässe verursachen können – umgekehrt können durch zahlreiche neue Verbraucher*innen auch Spannungsprobleme auftreten. Kurz, das Instrument des Engpassmanagements wird relevant. Diese Engpässe können tatsächlich zufällig, aber durch Absprachen auch gezielt auftreten und müssen behoben werden. Eine KI kann Angriffe und drohende Engpässe erkennen, Gamification identifizieren und sowohl als Instrument zur Angriffserkennnung als auch Netzplanung herangezogen werden.

4.9.3 Normungs- und Standardisierungsbedarfe

Die Anwendungsfälle in Kapitel 4.9.2 implizieren konkrete Bedarfe in der Normung und Standardisierung von Prozessen und Formaten. Mitunter besteht grundlegender Forschungsbedarf unter Einbezug verschiedener Disziplinen in Wirtschaft und Wissenschaft. Die nachfolgenden Ausführungen beschreiben Normungs- und Standardisierungsbedarfe, deren Erfüllung die dargelegten Anwendungsfälle flankiert und wesentlich zu deren Gelingen beitragen wird. Aufgrund der bereits zuvor getroffenen Auswahl besonders dringlicher Anforderungen und Herausforderungen können auch die nachfolgenden Aspekte nur einen Ausschnitt der vollständigen Bedarfsbandbreite darstellen. Die Anschlussfähigkeit zu weiteren Bedarfen ist mithin abhängig von der fortschreitenden Entwicklung in Politik, Forschung und Normung.

Bedarf 09-01: Interoperabilität von Terminologie, Semantik, Taxonomie und Daten

Materialwissenschaft und -wirtschaft sind mit grundlegenden Fragestellungen zur Erhöhung der Ressourcen- und Energieeffizienz konfrontiert. Dies betrifft insbesondere das Fachgebiet Tribologie, da Reibungs- und Verschleißoptimierung unmittelbare Auswirkungen auf den Material- und Energieaufwand haben. Durch viele involvierte Domänen entstehen in Charakterisierungs- und Modellierungsmethoden Inkonsistenzen in Begrifflichkeiten und Abhängigkeiten. Die **FAIR**-Prinzipien (Findable, Accessible, Interoperable, Reusable) müssen hierbei die Handlungsgrundlage bilden. Terminologien, Semantiken und Taxonomien domänenübergreifend zu erstellen bzw. zu harmonisieren, kann letztlich nur durch die Einbeziehung von Stakeholdern erfolgen und erfordert konsensorientierten Austausch. Weiterhin ist für die Bewertung der Zuverlässigkeit von KI-Entscheidungen die Integration geeigneter Metadaten (z. B. Sensortyp und Messungenauigkeit für Sensordaten) in die Datenmodelle zu erwägen. Folglich sollte dieser Prozess einem regelmäßigen Review unterliegen und normativ begleitet werden.

Bedarf 09-02: Schemata und Mapping für GIS-/BIM-Integration

Zur Bestimmung der Umweltwirkungen bzw. des **L**ife **C**ycle **A**ssessments (**LCA**) im Bauwesen entsteht auf Gebäude- und insbesondere auf Quartiersebene ein hoher Datenbedarf, der effizient bedient werden muss. **G**eografische **I**nformations**s**ysteme (**GIS**) und **B**uilding **I**nformation **M**odelling (**BIM**) weisen als geläufige Modellierungsmethoden Überschneidungen auf. Insbesondere GIS-basierte Gebäudemodelle in **L**evel **o**f **D**etail (**LoD**) 3 und 4 weisen qualitativ ähnliche Informationen auf wie detaillierte, BIM-basierte Gebäudemodelle. Eine Nutzung von Daten aus beiden Domänen kann eine signifikante Hebelwirkung in umweltbezogenen Anwendungen der Künstlichen Intelligenz und des Maschinellen Lernens entfalten. Hierzu bedarf es jedoch eines gemeinsamen Datenstandards in Form von Modellübersetzungen, Mappings von Datenformaten und Datenbankschemata. Ein derart gestalteter Datenstandard sollte kontinuierlich begleitet und infolge von Updates aus beiden Domänen (insbesondere **OGC** (**O**pen **G**eospatial **C**onsortium) für GIS und buildingSMART für BIM) regelmäßige Aktualisierungen erhalten.

Bedarf 09-03: Kohärenz und Einheitlichkeit der Datengrundlagen und KI-Anwendungen für nachhaltigen Konsum

Die einheitliche, branchenübergreifende bzw. -unabhängige Angabe von Umweltwirkungen und Kreislauffähigkeit von Gütern und Dienstleistungen erfordert ein gemeinsames Format zur Kommunikation. Dies beinhaltet einen gemeinsamen Datenstandard für die breit angelegte Bestimmung von Umweltwirkungen. Dieser Standard und integrative Datenformate vereinfachen den Aufbau KI-basierter Empfehlungssysteme für nachhaltigen Konsum. Es bedarf konkret einer Normung für Produktdatenbanken, zugehöriger Datenbankschemata und Datenmappings zur Sicherstellung der Interoperabilität. Weiterhin braucht es für ein lernendes Feedbacksystem bzw. die kontinuierliche Optimierung der Algorithmen eine datenschutzgerechte Formulierung der Nutzungsmöglichkeiten von Daten über das persönliche Konsumverhalten. Die dargelegten Aspekte betreffen eine Bandbreite an Stakeholdern aus Wirtschaft und Wissenschaft, die in die normativen Prozesse einbezogen werden sollten.

Bedarf 09-04: Methodik für die Bestimmung von Umweltwirkung und Performanz von Modellen der Künstlichen Intelligenz und des Maschinellen Lernens

Die grundsätzlich hohe Daten- und Rechenintensität von KI und ML-Modellen implizieren hohe Energieverbräuche und Umweltwirkungen, die im prinzipiellen Spannungsfeld zum Nutzen derartiger Modelle stehen. Weiterhin ist die Leistungsfähigkeit derartiger Modelle stark abhängig vom Anwendungsfall. Je nach Use Case weisen verschiedene Algorithmen unterschiedliche Laufzeiten und Genauigkeiten auf. Eine systematische Erfassung dieser Charakteristika als Metaparameter ermöglicht eine A-priori-Auswahl geeigneter Algorithmen für KI und ML-Anwendungen innerhalb der Kategorien überwacht/unüberwacht/reinforcement und sollte normativ begleitet werden. Zur Bestimmung der Nachhaltigkeitsgüte solcher Systeme braucht es einen einheitlichen Standard mit messbaren Bewertungskriterien. Bislang existieren hierzu noch keine genormten Verfahren. Es ist zunächst zu eruieren, ob eine absolute Metrik mit bestimmten Messkriterien oder ein Referenzverfahren mit standardisiertem Bezugssystem zu einer besseren Evaluierbarkeit und Vergleichbarkeit führen. Das gewählte Verfahren ist anschließend so weit auszuformulieren, dass ein KI-gestütztes Feedbacksystem zur Beurteilung der Laufzeit, Akkuranz und der Nachhaltigkeitsgüte von KI- und ML-Ansätzen aufgebaut werden kann.

Bedarf 09-05: Eingabeformate für lernende Systeme

Im Kontext domänenspezifischer Prozesse fällt immer wieder auf, dass Wissen mühevoll für die KI aufbereitet und umformatiert werden muss. Formate müssen als Standard etabliert werden, um eine breite Basis von Wissen so zur Verfügung zu stellen, dass sie in zahlreichen Anwendungen genutzt werden kann und somit „wachsendes“ Wissen etabliert wird. Eine vereinheitlichte Semantik als auch Syntax ermöglichen, ähnlich wie die Vereinbarung auf eine Geschäftssprache, schnellen Zugang zu dem dokumentierten Wissen sowie die bessere Wiederverwendung.

Bedarf 09-06: Übersicht und Referenzmodellbildung

Die Vereinheitlichung und Abstimmung von Inhalten in Normungsgremien bezüglich Definitionen, Taxonomien führt zu einer gemeinsamen Domänensemantik. Dazu sind für bestimmte Themenfelder und Inhalte führende Gremien zu definieren. Die Erstellung einer Normungslandkarte ermöglicht einen einfachen und schnellen Zugang zu den komplexen Abhängigkeiten der KI im Kontext der einzelnen führenden Gremien und kann daher die Verwendung von Normen unterstützen, indem die Ansprechpartner*innen und Wissensträger*innen besser zugänglich werden.

Bedarf 09-07: Dimensionierung und Begriffsbildung von I4.0 Referenzarchitekturmodellen (RAM)

Die RAMs für Smart Manufacturing (SM), Smart Grid (SG) und andere technische Infrastrukturen sind i. d. R. kubische Modelle, die vergleichbare Kategorien wie Kommunikationsschichten, Value-Stream-Zustände und Nutzungs- oder Produktionshierarchien verwenden. Die RAM-Begriffe und Konzepte sind jedoch aufgrund der disjunkten Anwendungsdomänen SM und SG nicht alle aufeinander abgestimmt. Es ergibt sich daraus ein Bedarf, die verwendeten RAM-Begriffe und Terminologien aus den Anwendungsdomänen semantisch, funktional, sicherheitspolitisch und ethisch miteinander zu vergleichen und abzustimmen.

Bedarf 09-08: Dynamisierung der statischen Referenzarchitekturmodelle (RAM)

Smart Manufacturing (SM), Smart Grid (SG) etc. sind Architekturmodelle. Daher sind keine Mittel vorgesehen, dynamische Abläufe als Teil der statischen Strukturen zu modellieren. Mithin gibt es in den heutigen RAM keine Begriffsdefinition einer Prozessvariablen. Ein System-von-Systemen (SvS) ist jedoch ein kommunizierendes Multi-Variablen-System, das die Ressourcen eines RAMs im Value Stream für die enthaltenen Transferfunktionen braucht. Der daraus ableitbare Handlungsbedarf besteht in der Integration aller Mittel zur Darstel-

lung architektureller (statischer) Strukturen und variablen (dynamischen) Verhaltens in dynamisierten RAM (dRAM).

Bedarf 09-09: Digitaler Zwilling zu Kontroll- und Prüfaufgaben in Smart Grid Architecture Model (SGAM)-Systemen

In SGAM-Systemen ist das Risiko von DER-Geräteausfällen (outages) besonders groß nach Naturereignissen wie beispielsweise Gewittern und Stürmen, da zumeist eine kaskadierende Lastverschiebung eintritt, die in den nicht erkannten Schwachpunkten zu einer Überlast führt und infolgedessen Ausfälle flächiger Netzteile verursacht. In diesem Kontext soll ein Digitaler Zwilling Betriebs- und Belastungsdaten über DER in einem SGAM- oder RAM-I4.0-System erheben, um sie in ein Betriebssystemmodell zu überführen. Hierin sollen gefährliche Inzidenzen analysiert werden. Zur vorausschauenden Vermeidung derartiger Inzidenzen könnte der Digitale Zwilling Lastverschiebungen zur Vermeidung von Überlast und Schwachpunkten gleichzeitig zum Wettergeschehen simulieren.

Bedarf 09-10: Berechnungsverfahren zur Ermittlung des CO_2-Faktors aus dem Strommix

Zur Ermittlung der CO_2-Emissionen aus dem Stromverbrauch zu einem gegebenen Zeitpunkt ist eine Zuordnung von Emissionen zu den erzeugten kWh erforderlich. Aktuell genormte Verfahren zur Allokation dieser Emissionen sehen eine statische Berechnung anhand eines festgelegten Faktors vor, der ggf. mit neuen Ausfertigungen der Norm aktualisiert wird. Diese Methodik trägt der Volatilität des Strommixes nicht hinreichend Rechnung, da witterungsbedingte Schwankungen in der Erzeugung aus erneuerbaren Quellen nicht berücksichtigt werden können. Es bedarf also eines agilen Berechnungsverfahrens mit höherer zeitlicher und ggf. geografischer Auflösung, um die Umweltwirkungen des Stromverbrauchs präziser zu ermitteln.

Die Arbeitsgruppe Energie und Umwelt hat die identifizierten Bedarfe nach der Dringlichkeit ihrer Umsetzung bewertet. Abbildung 49 zeigt die Dringlichkeit der Umsetzung, kategorisiert nach den Zielgruppen Normung und Forschung.

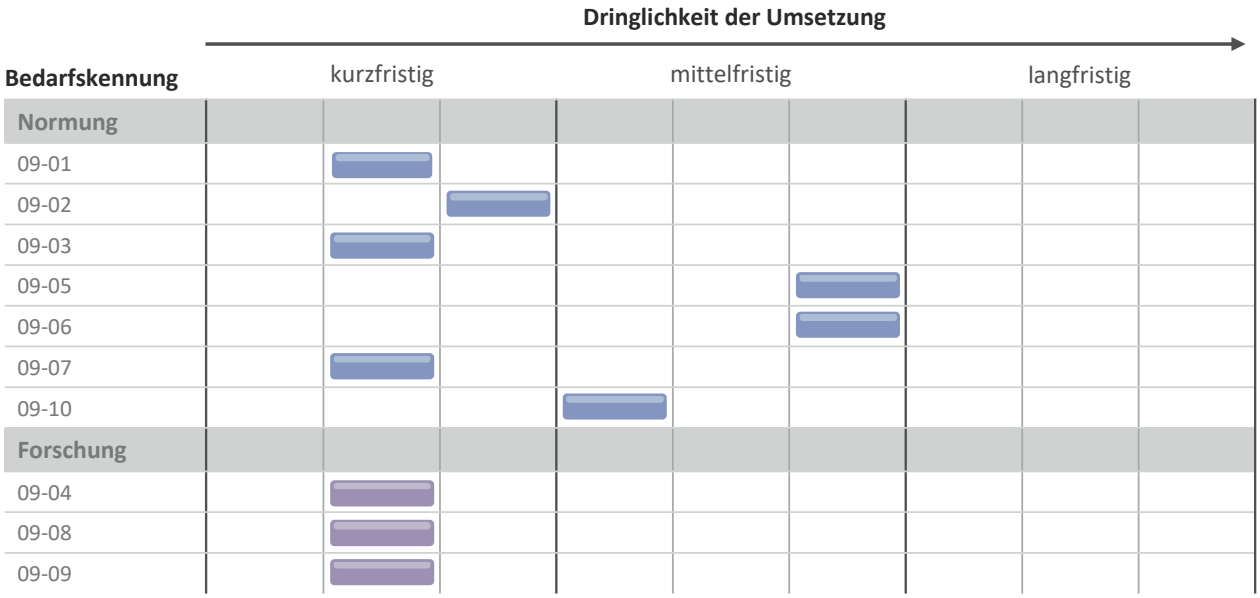


Abbildung 49: Priorisierung der Bedarfe aus Schwerpunkt Energie und Umwelt (Quelle: Arbeitsgruppe Energie und Umwelt)

5 Anforderungen an die Erarbeitung und Nutzung von Normen und Standards

Für die Erarbeitung und Nutzung von Normen und Standards stellt die rasant fortschreitende Technologieentwicklung und industriellen Anwendung von KI-Systemen gegenwärtig eine große Herausforderung dar. Abhängig vom Einsatzfeld der KI-Lösung verwenden verschiedene Branchen unterschiedliche und auf den Anwendungsfall bezogene KI-Technologien. Die Anwendungsspezifika werden dabei in den meisten Fällen von modernsten Ansätzen aus KI-Teildisziplinen erfüllt, die stetig angepasst und verfeinert werden. Folglich ist die Dynamik an der Schnittstelle zwischen KI-Forschung und industrieller Entwicklung und Anwendung besonders hoch.

Die Normung und Standardisierung muss diesem Spannungsbogen zwischen angewandter Forschung und industriereifer Entwicklung Rechnung tragen und neue Ansätze für die Analyse der Standardisierungsbedarfe, die Entwicklung marktreifer Normen und Standards sowie die Überprüfung und Anpassung existierender Normen und Standards verfolgen.

Um diesen Herausforderungen zu begegnen, werden derzeit verschiedene Ansätze verfolgt, die im Folgenden dargestellt werden.

5.1 KI-Tauglichkeit von Normen

Künstliche Intelligenz (KI) dringt in immer mehr Bereiche des Alltagslebens und in industrielle Anwendungen vor. Gleichzeitig ist davon auszugehen, dass KI nur dann ihr volles Potenzial entfalten kann, wenn ihr Einsatz nach anerkannten Qualitätsmaßstäben erfolgt, welche sicherstellen, dass KI-Anwendungen sicher und verlässlich sind und der Einsatz in Übereinstimmung mit gesellschaftlichen Wertevorstellungen erfolgt. Eine vielfach bewährte Möglichkeit, solche Qualitätsmaßstäbe zu realisieren, stellen Normen und Standards dar. Eine Marke „AI made in Europe“, welche auf hochwertigen Normen und Standards beruht, kann ein zentraler Wettbewerbsfaktor für die deutsche und europäische Wirtschaft darstellen. Hierfür ist eine Prüfung und ggf. Anpassung aller bestehenden Normen und Standards hinsichtlich KI nötig, wie sie auch explizit von der KI-Strategie der Bundesregierung[99] [2] in Handlungsfeld 10 gefordert wird.

Dazu ist im Januar 2022 unter der Leitung von DIN und im Auftrag des Bundesministeriums für Wirtschaft und Klimaschutz (BMWK) ein Projekt zur Evaluierung der „KI-Tauglichkeit von Normen“[100] gestartet. Mit einer Projektlaufzeit von zwei Jahren läuft das Projekt zunächst bis Dezember 2023. Beteiligt sind außerdem die Fraunhofer-Allianz „Big Data und Künstliche Intelligenz“, der Beuth Verlag sowie DIN Software.

Das Projekt verfolgt zwei Stoßrichtungen und lässt sich in folgende Aspekte unterteilen (siehe Abbildung 50):

1. der inhaltliche Bezug der Normen zu KI-Technologien
2. die Maschinenausführbarkeit/-lesbarkeit von Normen

Die Maschinenausführbarkeit von Normen wird bereits in der Initiative SMART Standards untersucht (siehe Kapitel 5.3). Das Projekt „KI-Tauglichkeit von Normen“ soll die Brücke zu KI-spezifischen Anwendungsfällen bauen und etwaige Anforderungen an die Maschineninterpretierbarkeit bzw. zukünftige Nutzungsmöglichkeiten ableiten. Die im Rahmen des Projekts gewonnenen Erkenntnisse sollen zusätzlich in das Netzwerk „Initiative Digitale Standards“ (IDiS) einfließen und Synergien schaffen.

Der andere und weitaus größere Fokus innerhalb des Projekts (Aspekt 1) liegt in der inhaltlichen Betrachtung der Normen. Ausgangspunkt ist die Annahme, dass Künstliche Intelligenz früher oder später für alle wirtschaftlichen und gesellschaftlichen Bereiche von großer Bedeutung sein wird. Auch Normen und Standards existieren für nahezu alle Wirtschaftsbereiche und Anwendungsfelder. Aktuell umfasst das deutsche Normenwerk mehr als 30.000 Normen (DIN, DIN EN, DIN EN ISO/IEC). Das bedeutet, dass ein Großteil der existierenden Normen Berührungspunkte zu KI-Technologien haben dürfte und daher entsprechend überprüft und um KI-Aspekte ergänzt werden muss. Gleichzeitig gibt es zum heutigen Standpunkt keinen zentralen Überblick, welche Normen für den Einsatz von KI-Technologien in ihrem Anwendungsgebiet ausgelegt sind.

Das Projekt soll eine Bestandsaufnahme über den Querschnitt des gesamten Normenwerks darstellen, um Fragen zu beantworten wie z. B.:

- → Wie viele und welche Normen haben Berührungspunkte zu KI-Technologien?
- → Welche dieser Normen sind schon auf den Einsatz von KI vorbereitet?
- → Welche Normen müssen dahingehend zeitnah überarbeitet werden und wie könnte diese Überarbeitung erfolgen?

99 https://www.ki-strategie-deutschland.de/home.html

100 https://www.din.de/de/din-und-seine-partner/presse/mitteilungen/din-startet-projekt-ki-tauglichkeit-von-normen--872810

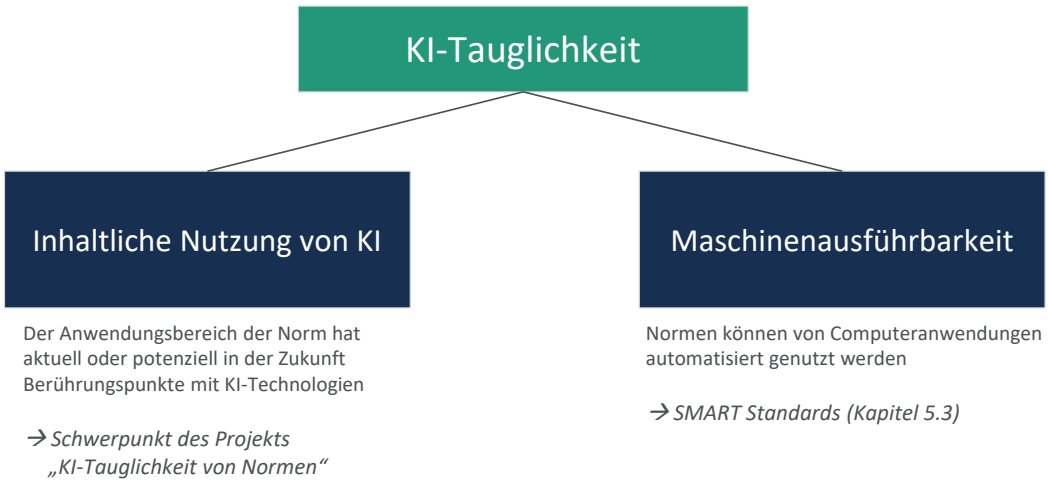


Abbildung 50: Struktur des Projekts „KI-Tauglichkeit von Normen“ (Quelle: DIN)

Ziel ist es, eine skalierbare Methodik und ein prototypisches KI-Werkzeug (Software) zu entwickeln, welche perspektivisch auf das gesamte Normenwerk anwendbar sind.

Vorgehen im Projekt

Im Rahmen des Projekts wird zunächst eine Definition von KI-Tauglichkeit (bezogen auf den inhaltlichen Aspekt) erarbeitet. Der Begriff der KI-Tauglichkeit wird geschärft und operationalisierbar gemacht, indem beispielsweise Bewertungskriterien zur Feststellung der KI-Tauglichkeit festgelegt werden. Zur Erarbeitung der Definition und Methodik sind zwei Aspekte besonders wichtig: das Fach- bzw. Normungswissen sowie die Expertise über KI-Methoden. Durch die beiden Gruppen der Fachexpert*innen der Normungsgremien auf der einen und die KI-Expert*innen der Fraunhofer-Gesellschaft auf der anderen Seite sollen beide Sichtweisen integriert werden. Das Projekt „KI-Tauglichkeit von Normen“ wurde bereits in den Beiratssitzungen einiger Normenausschüsse vorgestellt. Alle interessierten Fachbereiche sind stets eingeladen, sich an beliebiger Stelle in den Workshops zur Definition, Methodik oder in Pilotprojekten zu beteiligen.

Bei der Analyse der Beispiele wird ein besonderes Augenmerk auf den Fall gelegt, dass Normen und Standards den Einsatz von KI einschränken. Hier muss unterschieden werden zwischen grundsätzlichen Unzulänglichkeiten bestehender KI-Verfahren, welche einen Einsatz nach dem derzeitigen Stand der Technik verhindern, und dem Fall, dass eine inhaltliche Erweiterung bzw. Weiterentwicklung von Normen die Einschränkung überwinden kann. Ein mögliches Beispiel stellt hier das Thema „Safety“ dar, dessen Anforderungen den Einsatz von KI für bestimmte Anwendungsfelder aktuell verhindern.

Das erste zentrale Ergebnis ist eine sinnvolle Definition von „KI-Tauglichkeit von Normen“, bezogen auf den Inhalt. Damit diese in den verschiedenen Normungsbereichen anwendbar ist, folgt eine Operationalisierung in Form einer Arbeitshilfe, welche es entsprechenden fachkundigen Personen erlaubt, eigenständig eine Einschätzung vorzunehmen, ob eine Norm KI-tauglich ist.

Die Prüfung der Normen auf KI-Tauglichkeit soll zusätzlich durch einen maschinengestützten Prozess erleichtert werden. Dazu soll ein prototypisches KI-Werkzeug entwickelt werden, welches bei der Auswahl und Bewertung der Normen unterstützt. Der Unterstützungsgrad ist dabei abhängig von der Anzahl überprüfter Normen, mit denen das KI-Werkzeug trainiert wird. Je mehr Normen manuell überprüft werden, desto besser wird das KI-Tool die Bewertung vornehmen können.

Im Rahmen von Pilotprojekten und als Grundlage für die Entwicklung eines solchen KI-Werkzeugs wird im nächsten Schritt das Normenwerk hinsichtlich ausgewählter Fachbereiche mit besonderem Bezug zu KI (beispielsweise Branchen wie Maschinenbau, Automobil, Medizin oder Querschnittsthemen wie Ethik und Sicherheit, analog zu den Schwerpunktthemen der Deutschen Normungsroadmap KI) einer Beurteilung unterzogen. Anhand der erarbeiteten Arbeitshilfe analysieren verschiedene Branchen- und KI-Expert*innen in weiteren Schritten die Normen für diese ausgewählten Fachbereiche hinsichtlich ihrer KI-Tauglichkeit.

Als Projektergebnis ergeben sich:
- → eine Methodik zur Bewertung von KI-Tauglichkeit (ggf. erweiterbar auf weitere Anwendungsgebiete, „Klimaschutz-Tauglichkeit“ o. Ä.);
- → eine Liste der überprüften und bewerteten Normen;
- → Überarbeitungsempfehlungen zu den Normen, die als nicht KI-tauglich bewertet wurden;
- → das entwickelte prototypische KI-Werkzeug zur maschinengestützten Bewertung des restlichen Normenwerks.

Durch das Projekt werden die beteiligten Fachbereiche für das Thema KI sensibilisiert und können mithilfe der Unterstützung der KI-Expert*innen Anknüpfungspunkte zu ihren Prioritäten und ihrem Arbeitsprogramm identifizieren, bewerten und Impulse setzen. Die KI-Expert*innen stehen während der Laufzeit des Projekts zur Verfügung und können zur Beratung in Anspruch genommen werden. Durch die KI-Expert*innen und ihr Netzwerk können so auch neue KI-Expert*innen für die Ausschüsse gewonnen und damit ein wesentlicher Beitrag zur Berücksichtigung neuer technologischer Entwicklungen in der Normung geleistet werden. Mit dem KI-Werkzeug steht auch über die Projektlaufzeit hinaus zusätzlich eine wertvolle technische Möglichkeit zur Verfügung, um die Normenausschüsse zu unterstützen. Das Projekt kann somit ganz maßgeblich zur Stärkung des Wirtschaftsstandortes Deutschland beitragen.

5.2 Agile Entwicklung von Normen und Standards

Die hohe Dynamik in der Technologieentwicklung zu KI hat Auswirkungen auf die Anforderungen, die an die Erarbeitungsprozesse von Normen und Standards gestellt werden. Hier braucht es agile Ansätze und Prozesse, die bei der Gestaltung von Normen und Standards wechselseitige Impulse von Expert*innen stetig einbeziehen und eine kollaborative Entwicklung von Normen und Standards unterstützen.

Im Zentrum dieses Ansatzes steht das XML-Dateiformat, das sich als fester Bestandteil bei der Weiterverarbeitung von Normen und Normeninhalten etabliert hat und einen wesentlichen Grundpfeiler der Bestrebungen rund um SMART Standards (siehe Kapitel 5.3) darstellt. Die XML-first-Strategie der DIN-Gruppe sieht eine frühstmögliche Einbindung des XML-Dateiformats im Normenerstellungsprozess vor. Gemäß der Strategie sollen Normen und Standards in Zukunft direkt in XML erstellt werden und damit u. a. die sukzessive Ablösung konventioneller Textverarbeitungssoftware, nachgelagerter Konvertierungsprozesse und bestehender Medienbrüche ermöglichen. Auf Basis dieser Umstellung lassen sich Normenerstellungs-prozesse effizienter gestalten und potenzielle Fehlerquellen, die mitunter erhebliche Mehraufwände erzeugen, reduzieren.

Um dieses Vorhaben zu realisieren, braucht es ein geeignetes Tool. Mit FontoXML wird auf einen XML-Editor gesetzt, der sowohl die direkte Erstellung in XML als auch die kollaborative Bearbeitung von Inhalten ermöglicht, beides ein Novum für die Normenerstellung bei DIN und DKE.

Der XML-Editor wird den Normenersteller*innen eine technische Basis für die zukünftige Erfassung von Inhalten bieten und die Digitalisierung der Normenerstellung weiter vorantreiben. Gleichzeitig wird das kollaborative Erarbeiten von Inhalten die Transparenz des Normungsprozesses noch weiter stärken.

Die Expert*innen in den Normenausschüssen können parallel an Inhalten arbeiten. Die Prozesse zur Entwicklung und Überarbeitung von Normen und Standards werden dadurch maßgeblich an Agilität gewinnen. Kommentare zu Arbeitsständen können über das Tool abgegeben, von anderen direkt eingesehen und beurteilt werden. Aktuelle Arbeitsstände sind stets revisionssicher in der Cloud abgelegt, sodass die Normungsgremien nicht länger Offline-Kopien verwalten müssen. Inhalte gehen somit nicht verloren und auch das mühsame Verwalten mehrerer Word-Dokumente entfällt.

Der permanente Zugang zu aktuellen Arbeitsständen erlaubt u. a. ein schnelleres Reagieren auf sich ändernde Rahmenbedingungen im Laufe eines Normungsvorhabens und ermöglicht sowohl für die Expert*innen als auch für das Normungsgremium eine höhere Flexibilität in der Bearbeitung. Letztendlich können diese Vorteile auch zu einer kürzeren Erarbeitungs- bzw. Überarbeitungszeit und somit zu einer schnelleren Verfügbarkeit von Normen und Standards beitragen.

Auch die internationalen (ISO und IEC) und europäischen Normungsorganisationen (CEN und CENELEC) setzen zukünftig auf XML und planen, die Erarbeitung der Normeninhalte unter Nutzung von FontoXML direkt im XML-Format zu ermöglichen. In Pilotprojekten wird seit Ende 2020 der Editor erprobt und gemeinsam mit den Entwickler*innen weiterentwickelt.

5.3 SMART Standards

Neugestaltung von Normen und Standards zur Integration in KI-Anwendungsprozesse

Die direkte Weiterverwertung von Normen und deren Inhalte in nachgelagerten Prozessen gewinnt zunehmend an Aufmerksamkeit. Von Normbestandteilen (Wertetabellen, Teilebeschreibungen, 3-D-Modellen, Software, Anforderungsdefinitionen, Prüfverfahren), die von Maschinen direkt übernommen und ausgeführt werden können, versprechen sich Unternehmen zukünftig Effizienzgewinne[101].

Um dieses Ziel zu erreichen, arbeiten DIN und DKE seit einigen Jahren an den SMART Standards.

Unter einem SMART Standard wird eine Norm (Standard) verstanden, deren Inhalte für Maschinen, Software oder sonstige automatisierte Systeme anwendbar (applicable) und lesbar (readable) sind und darüber hinaus anwendungs-/nutzerspezifisch digital bereitgestellt werden können (transferable).

Im Folgenden wird aufgezeigt, welche Entwicklungen sich seit 2020 ergeben haben. Dazu werden jeweils die Ausgangssituation in 2020 und der aktuelle Stand (2022) dargestellt.

Ausgangssituation 2020

Der seit Jahrzehnten etablierte Workflow funktioniert erfolgreich und ausgewogen aufgrund vereinbarter Übereinkünfte der handelnden Prozesspartner*innen. Die zugrunde liegenden Prinzipien sind sorgfältig normungs- und rechtskonform aufeinander abgestimmt und garantieren ein zuverlässiges Management der Normungsergebnisse in kundenorientierten Systemen.

Die anstehenden tiefgreifenden prozessualen Veränderungen im Rahmen der SMART Standards Erarbeitung, des Content Managements, der Distribution und der Nutzung werden vor dem Hintergrund bestehender eingeführter und regulierter Vorgehensweisen abgegrenzt und neu definiert werden müssen. Der entscheidende Wert („Asset“) eines Normungsgegenstands muss erhalten bleiben.[102]

Aktueller Status 2022 und Weiterarbeit

Die wesentliche Anforderung an einen Workflow für SMART Standards besteht darin, strukturierte und semantisch angereicherte Inhalte zu entwickeln und bereitzustellen, die die Grundlage für eine maschinelle Verarbeitung, im Besonderen auch für KI-Anwendungen darstellen.

Die 1. Ausgabe der Normungsroadmap KI hat die wesentlichen Entwicklungsschritte eines zukünftigen Normungsprozesses (Content Creation) beschrieben und einen tiefgreifenden Veränderungsbedarf des Prozesses aufgezeigt.[103] Neben der Content Creation müssen auch die Prozessschritte Content Management und Content Delivery dahingehend weiterentwickelt werden, dass sie fragmentierte und semantisch angereicherte Norminhalte verarbeiten und an die Anwendung (Content Usage) ausliefern können.[104]

Die Projektvorhaben zu SMART Standards auf europäischer und internationaler Ebene haben sich in den vergangenen zwei Jahren intensiv mit den oben skizzierten Herausforderungen beschäftigt. Sie haben dazu jeweils eigene Projektstrukturen aufgesetzt, die entweder direkt zusammenarbeiten oder sich über formelle und informelle Kanäle austauschen und abstimmen:

Bei CEN-CENELEC werden im **Workstream 3** „Technical Solution“ zum einen das Informationsmodell und zum anderen die technologische Infrastruktur für SMART Standards entwickelt. Hierbei spielen die Werkzeuge zur Content-Erfassung (beispielsweise XML-Editoren für die Erstellung von Level 3 Content auf Basis der Technologie FONTO) eine zentrale Rolle, denn sie sind zwingende technologische Voraussetzung dafür, dass ein höherer Strukturierungsgrad bei gleichzeitig größtmöglicher Prozesseffizienz erreicht werden kann.

Workstream 4 „Operationalisation“ beschreibt den Prozess der Content Creation bis auf die Ebene der neuen Arbeitsabläufe mit einem zukünftigen Erfassungswerkzeug und legt die Anforderungen an die unterstützende Organisation fest. Diese drei zentralen Komponenten Prozess, Organisation und Technologie werden aktuell prototypisch entwickelt, sodass sie in 2023 in konkreten Normungsprojekten (in Abstimmung

101 Siehe C. Wischhöfer, P. Rauh, Standards of the Future – Stand der Arbeiten zum Thema maschinenausführbarer Normeninhalte. DIN-Mitteilungen, August 2019, S. 4–8.

102 Siehe Normungsroadmap KI (Ausgabe 1), Kapitel 5.2.2 [63].

103 Siehe Normungsroadmap KI (Ausgabe 1), Kapitel 11.4.3, Anhang „Top-down-Methode“ [63]

104 Siehe Normungsroadmap KI (Ausgabe 1), Abbildung 31 und Abbildung 37 [63]

mit **Workstream 1** „Standards User Engagement“) pilotiert und getestet werden können.

Im **Workstream 5** „Business Model“ werden neue Bereitstellungsformen (im Sinne von Content Delivery) entsprechend der aufgenommenen Anwendungsfälle (**Workstream 2** „Standards Maker Engagement“) evaluiert und Ableitungen hinsichtlich kommerzieller und legaler Aspekte (z. B. Lizenz- und Nutzungsbedingungen) gemacht.

Auf internationaler Ebene hat ISO mit den sogenannten „Subgroups“ innerhalb von ISO SMART bzw. IEC mit den „Taskforces“ der SG 12 analoge Arbeitsgruppen gebildet. Diese Arbeitsgruppen zeichnen sich sowohl inhaltlich als auch in Bezug auf die Projektteilnehmer*innen durch eine große Überschneidung mit dem europäischen Projekt aus. Hierdurch ist ein Know-how-Transfer von der europäischen zur internationalen Ebene (und vice versa) gewährleistet, der von entscheidender Bedeutung ist, wenn bis 2024 ein Modell für Level 3 für die produktive Nutzung bereitgestellt werden soll und in der Folge die gemeinsame Weiterentwicklung Richtung Level 4 Content erfolgen soll.

Ausgangssituation 2020
Eine Herausforderung wird die Konsolidierung eines gemeinsamen Verständnisses der Entwickler*innen und Anwender*innen von SMART Standards sein.

Das Stufenmodell muss verifiziert und an andere Modelle adaptiert werden können, z. B. Referenzarchitekturmodell Industrie 4.0 (RAMI4.0) /HHHD-17/.[105]

Aktueller Status 2022 und Weiterarbeit
Im Whitepaper „Szenarien zur Digitalisierung von Normung und Normen“ von IDiS (Initiative Digitale Standards) wurde das IEC Utility Model (auch Stufenmodell von SMART Standards) um ein sogenanntes Level 5 erweitert: „Maschinensteuerbare Inhalte. Die Inhalte einer Norm können durch Maschinen selbstständig angepasst und durch automatisierte (verteilte) Entscheidungsprozesse verabschiedet werden. Die so verabschiedeten Inhalte werden automatisiert geprüft und über die Veröffentlichungskanäle der Normungsorganisationen veröffentlicht.“

Weiter wird im Whitepaper ausgeführt, dass KI-Anwendungen von einer verbesserten Maschinenanwendbarkeit profitieren, weil somit die Interpretierbarkeit und Auswertbarkeit von normativen Inhalten und Sachverhalten steigt.

Somit unterstützt das Whitepaper die aufgestellte These, dass SMART Standards Regeln und Prozesse bei der Beschreibung von Inhalten einführen, die es KI-Anwendungen erleichtern, die so erfassten Inhalte besser verarbeiten zu können. Das oben erwähnte Level 5 geht dabei noch einen Schritt weiter und beschreibt die Möglichkeit, dass KI-Anwendungen selbst Bestandteil von Entscheidungsprozessen (z. B. des Normungsprozesses) werden können und somit als aktive Teilnehmer auftreten können.

In IDiS wurde Ende 2021 (Laufzeit ca. zwölf Monate) ein Pilotprojekt gestartet, welches die Entwicklung eines domänenspezifischen Language Models, basierend auf circa 40.000 DIN- und VDE-Normen, zum Ziel hat. Mithilfe dieses Language Models soll die Identifikation passender Textpassagen aus relevanten Normen zu einem Produkt vorgenommen werden, dessen Merkmale im Produktdatenstandard ECLASS definiert sind. Das Pilotprojekt untersucht dabei die generelle Eignung normativer Inhalte für die Anwendbarkeit von Machine-Learning-Verfahren und andererseits, inwieweit KI-Verfahren bei der Suche nach relevanten Daten unterstützen können.

Ausgangssituation 2020
In allen Gremien wird ein wichtiger neuer Aspekt immer wieder thematisiert: Wie bereiten wir die Handelnden entlang des gesamten Wertschöpfungsprozesses auf die neuen Anforderungen vor?

Ein weiterer für die Zukunft systemrelevanter Aspekt betrifft die Definition der Anforderungen an die veränderten Qualifikationen der externen, aber auch der DIN-internen „Akteure“ im Gesamtprozess. Die bestehenden Konzepte bilden die Grundlage zur Definition der Anforderungen an Stellen sowie Weiterbildungsmöglichkeiten und müssen folglich weiterentwickelt werden, um die in SMART-Standards-Prozessen neu entstehenden Aufgaben aller Prozessbeteiligten zu beschreiben.[106]

105 Siehe Normungsroadmap KI (Ausgabe 1), Kapitel 5.2.3 [63]

106 Siehe Normungsroadmap KI (Ausgabe 1), Kapitel 5.2.3 [63]

Aktueller Status 2022 und Weiterarbeit
Die prozessualen, technologischen und geschäftsmodellbezogenen SMART-Standard-Aspekte werden national bei DIN-DKE, europäisch bei CEN-CENELEC und international bei ISO/IEC in verschiedenen Arbeitsgruppen intensiv diskutiert und erste Lösungen erarbeitet. Derzeit werden die geänderten Anforderungen an die Prozessteilnehmer entlang des gesamten Wertschöpfungsprozesses

Content Creation → Content Management → Content Delivery → Content Usage

jedoch nicht adäquat in Betracht gezogen. Die alleinige Planung und Durchführung von Trainings werden den Change durch die Beteiligten nicht sicherstellen. In Bezug auf eine zukünftige Nutzung KI-gestützter Verfahren reicht es überdies nicht aus, nur wenigen Expert*innen eine Plattform für deren Überlegungen und Realisierungen zu bieten.

Die zukünftigen Anforderungen zur gänzlichen Durchführung der unterschiedlichen Teilprozesse und damit verbundenen Aufgaben sind unterschiedlich – und damit auch die Anforderungen an die Personen (bzw. dem Vorhandensein der dafür erforderlichen Kompetenzen in entsprechenden Ausprägungsgraden), die die Aufgaben zu bearbeiten haben. Die Methodik ist bekannt und wird derzeit in DIN weiter ausgearbeitet und realisiert: Sogenannte Funktionsbeschreibungen mit klar definierten Handlungsspielräumen und wertprägenden Aufgabenbeschreibungen sind eine Voraussetzung, um in Unternehmen die zukünftigen Anforderungen bei der Content-Erarbeitung, Ausspielung und Nutzung nachvollziehbar zu beschreiben und vorzugeben, notwendige Stellen im Unternehmen zu identifizieren und strategisch zu verankern, und Personen im Job im Sinne eines Kompetenzaufbaus weiterzubilden.

In der Industrie existieren bereits erste Funktionsbeschreibungen entlang der hier betrachteten Wertschöpfungskette. Die Erfahrungen zeigen, dass die Attraktivität dieser Stellen für externe Bewerber sehr hoch ist.

Ausgangssituation 2020
SMART-Standards sind eine von vielen Wissensdomänen und ermöglichen es KI-Systemen grundsätzlich, die in ihnen enthaltenen Informationen automatisch und optimal in den verschiedenen Teilprozessen in einem Unternehmen zu nutzen.

Die Konzeption der notwendigen Datenmodelle und Schnittstellen wird Teil dieses Projekts sein müssen und leistet damit einen wichtigen Beitrag zur weiteren Durchdringung der KI-Anwendungen in den Teilprozessen von Unternehmen.[107]

Aktueller Status 2022 und Weiterarbeit
Das IEC Utility Model mit seinen fünf Stufen (Level 0 – 4) wurde in ISO und IEC weiterdiskutiert und als gemeinsame Grundlage für die Beschreibung der grundlegenden Maschinenanwendbarkeit von SMART Standards akzeptiert.

Darüber hinaus wurden auf dieser Grundlage weitergehende Konzepte für die zukünftige Verwendung von SMART Standards entwickelt, welche aktuell in den Arbeitsgruppen von ISO und IEC (ISO SMART, IEC SG 12) weiterdiskutiert und entwickelt werden.

So gibt es beispielsweise erste Ideen eines SAM (Standard Architecture Model) und einer SAS (Standard Administration Shell). Beide Konzepte basieren auf Ideen der Industrie 4.0 (RAMI 4.0 (Referenzarchitekturmodell Industrie 4.0) und AAS) und sollen dabei helfen, die Funktionalitäten und Verantwortlichkeiten rund um SMART Standards besser einordnen und diskutieren zu können. Das SAM ordnet dabei, in Anlehnung an das RAMI Model, Aktivitäten und Funktionen von SMART Standards unterschiedlichen Dimensionen zu (Application Layer, Utility Level und Standard Life Cycle), um so das Verständnis und die Unterscheidbarkeit zwischen Application weiter zu verbessern. Das SAS hingegen ist mehr ein technisches Modell und beschreibt, wie Funktionen und Verantwortlichkeiten aufgeteilt werden können, um einen einheitlichen Zugriff auf Inhalte von SMART Standards zu ermöglichen. In IDiS (Initiative Digitale Standards), der nationalen Community für SMART Standards, startete Mitte 2021 ein erstes Pilotprojekt (Laufzeit ca. 15 Monate) zum Thema Verwaltungsschale und Teilmodel einer Digitalen Norm.

Ausgangssituation 2020
Der wirtschaftliche Nutzen der Standardisierung wird in einigen Ländern quantifiziert. In Deutschland erspart die Normung der Wirtschaft jährlich 17 Milliarden Euro. Die Bezifferung eines wirtschaftlichen Nutzens von SMART Standards liegt noch nicht vor und kann bisher nur qualitativ genannt werden.

107 Siehe Normungsroadmap KI (Ausgabe 1), Kapitel 5.2.4 [63].

Im Rahmen des Projekts muss eine wirtschaftliche Bewertung bezüglich Aufwand, Nutzen, Realisierungszeitraum, Qualität etc. der verschiedenen Vorgehensweisen erfolgen. Danach oder projektbegleitend kann eine Priorisierung der Vorgehensweisen vorgenommen werden.[108]

Aktueller Status 2022 und Weiterarbeit
Der Beitrag zum deutschen Wirtschaftswachstum des aktuellen Bestands der Normen beträgt jährlich ca. 17 Milliarden Euro, das entspricht in etwa 0,7 % des Bruttoinlandsprodukts.

Je mehr Norminhalte durch SMART Standards automatisiert erschlossen werden können, desto höher dürfte der aktuelle Anteil sein, der sich insbesondere für die Nutzungsphase im Wertschöpfungsprozess ergibt. Es ist offensichtlich, dass das Potenzial zur Effizienzsteigerung bei der Anwendung von Normen durch eine solche automatisierte und anwendungsspezifische Bereitstellung und Übertragung von Normeninformationen erheblich ist.

Da die Normungsorganisationen dieses Potenzial derzeit nicht quantifizieren können, planen DIN und DKE, dies innerhalb der Initiative Digitale Standards (IDiS) im Rahmen eines Projekts zu untersuchen.

Ausgangssituation 2020
Der Fokus auf IT-gestützte Verfahren und deren Weiterentwicklung im „Content Management“ und „Content Delivery“ bietet die Chance, schnell zu konkreten Lösungen zu gelangen, die für Level 4 (KI) einen wertvollen Input liefert.

Für nachgelagerte KI-Anwendungsprozesse bedeutet das: Eine Validierung der Genauigkeit der automatisiert ermittelten (Partial-)Informationen [heute: fragmentierte Norminhalte] muss vorgenommen werden. Erlerntes Erfahrungswissen kann die Bewertung essenziell unterstützen.

Allgemeine Regeln zur Beschreibung der fragmentierten Norminhalte in Normen sowie die methodische Erarbeitung der genauen Verwendungsorte (Wirkorte) liegen für diese Vorgehensweise bisher nicht vor und müssen erarbeitet werden. Um KI-Anwendungsprozesse skalierbar mit fragmentierten Norminhalten zu versorgen, sind entsprechende Festlegungen zu vereinbaren.[109]

108 Siehe Normungsroadmap KI (Ausgabe 1), Kapitel 5.2.5 [63].

109 Siehe Normungsroadmap KI (Ausgabe 1), Kapitel 11.4.1 [63].

Aktueller Status 2022 und Weiterarbeit
Nutzer*innen von Normendokumenten investieren häufig viel Zeit für die Recherche, um für sie relevante Informationen aus Normen zu extrahieren (z. B. Anforderungen, Formeln, Produkt- und Klassifizierungsmerkmale) und verwenden zu können. Die Vielzahl der potenziell relevanten Normen erschwert hierbei den Rechercheaufwand. Systeme wie das Semantische Normen-Informations-Framework (SNIF) bieten hierbei eine gute Unterstützung und erleichtern die Stichwort- und Themensuche. Anwendungen wie SNIF basieren jedoch auf festen Regeln und Schlüsselwörtern. Sie sind auf diese beschränkt und führen also zu vorher definierten Ergebnisbereichen.

Moderne Methoden aus dem Bereich der Künstlichen Intelligenz, konkret hier Natural Language Processing (NLP), sind die Basis für starke Verbesserungen des Sprachverständnisses von Maschinen in verschiedenen Domänen. Hierzu werden vortrainierte (pretrained) Sprachmodelle (z. B. German BERT) verwendet, die auf einer breiten Vielfalt von Texten trainiert werden. Durch das Pretraining der Modelle erhalten diese ein grundsätzliches Verständnis der Domäne, aus der die Texte und die in ihnen enthaltenen Informationen stammen.

In verschiedenen Projekten werden Sprachmodelle trainiert, die auf internationalen und deutschsprachigen Normen basieren. Diese vortrainierten Sprachmodelle können für verschiedene Use Cases verfeinert werden. Einer dieser Use Cases ist z. B. das Extrahieren relevanter Norminhalte (z. B. Anforderungen oder Produktmerkmale). Weiterhin werden Datensätze erstellt, die Fragen an Normen in Kombination mit relevanten Textpassagen der Normen als Antworten enthalten. Somit können vortrainierte Sprachmodelle in Form eines spezialisierten Modells so verfeinert werden, dass diese lernen, passende Textpassagen zu einer Frage zu identifizieren und zu extrahieren. Überdies können statistische Klassifikatoren relevante Normeninhalte auf Grundlage regelbasierter Ansätze identifizieren. Somit können voraussichtlich u. a. Textstellen zurückgewiesen werden, die z. B. inhaltlich keine Anforderungen darstellen.

Ausgangssituation 2020
Auf der Basis XML-konvertierter Dokumente und unter Einhaltung des NISO STS wurde der Service „con:text“ entwickelt, der an verschiedene Normenmanagementsysteme gekoppelt werden kann. Das Funktionsset zielt darauf ab, den Inhalt tiefergehend zu erfassen, Zusammenhänge simultan auszuspielen und für den Anwendenden anwendungsfreundlich über zahlreiche Funktionen sichtbar zu machen.

Das Funktionsset von con:text spiegelt die Anforderungen der Anwender*innen wider. Somit entsteht hier ein Anwendungs-Know-how, das für die Funktionsbildung von KI-Anwendungsprozessen relevant sein kann. Gleichzeitig kann die Anwendung con:text von den Ergebnissen der KI-Projekts profitieren. Die Zusammenarbeit im KI-Projekt soll ermöglicht werden.[110]

Aktueller Status 2022 und Weiterarbeit
Ziel des bisherigen Projekts war es, einen Online Editor zu entwickeln, der sich funktional möglichst nahtlos in die bestehende Splitscreen-Oberfläche von con:text integriert lässt.

Inhalte werden hier nun in einer HTML-Oberfläche erfasst und direkt in XML strukturiert, sodass die zugrunde liegenden Schemata (beispielsweise NISO-STS) nachfolgende aufwendige Konvertierungsprozesse überflüssig machen. Die Erstellung von Inhalten kann in Projektteams mit unterschiedlichen Zuständigkeiten (initiieren, editieren, freigeben etc.) erfolgen. Der Online Editor kann diese Rollen flexibel abbilden bzw. über eine Schnittstellenanbindung bestehende Rechte- und Rollenkonzepte aus Drittsystemen übernehmen.

Künftige KI-basierte Maßnahmen zur Anwenderunterstützung sind etwa das (teil-)automatisierte Überprüfen text- und datenbasierter Inhalte zum Abgleich von Anforderungen, Werten, Wertbereichen oder anderen Vorgaben; das Suchen und Auffinden prozessrelevanter Textstellen in der aktuell bearbeiteten Norm, in zitierten oder thematisch passenden Normen; das Selektieren und Extrahieren definierter Bestandteile wie mathematische Formeln, Tabellen oder Anforderungen und das Überführen solcher Suchergebnisse in strukturierte Ausgabeformate (u. a. ReqIF). Im Rahmen der KI-basierten Weiterentwicklung können die zuvor genannten Funktionen auch in mit con:text unterstützten Produkten bereitgestellt werden, etwa Normenmanagementlösungen, Onlinedienste oder Portalservices.

Ausgangssituation 2020
Die Lösung besteht in einer automatischen Extraktion von Norminhalten und deren Überführung in eine maschinenausführbare Wissensrepräsentationsform, auf die von unterschiedlichen Autorensystemen zugegriffen werden kann. Aus den Erkenntnissen, die bei der konkreten Konzeptumsetzung gewonnen werden können, lassen sich Anforderungen und Gestaltungsregeln auf eine höhere Abstraktionsebene der „Next Generation Norm“ ableiten.

Die Nachstrukturierung des existierenden, sehr großen Normenfundus stößt an kapazitive Grenzen und wäre nur für definierte Themenbereiche wirtschaftlich vertretbar. Hierbei ist ein Einsatz von Künstlicher Intelligenz in der Extraktionsphase des Bottom-up-Ansatzes zu untersuchen, um diesen Arbeitsschritt maschinell zu unterstützen.[111]

Aktueller Status 2022 und Weiterarbeit
Neben Normen-Volltexten wird auch immer häufiger die Bereitstellung von kunden- und anforderungsspezifischen Teilinhalten fokussiert. Dabei stellen unterschiedliche Normenelemente unterschiedliche Herausforderungen dar, die es individuell zu adressieren und zu erarbeiten gilt. Entscheidend ist hier die Aufbereitung in möglichst generischen bzw. breit einsetzbaren Datenstrukturen (d. h. möglichst vielseitig übertragbare Ansätze zur Ablage, Anordnung und Verknüpfung von Daten) sowie eine breit angelegte Metadatenstrategie (d. h. Definition einer umfassenden und zielführenden Beschreibung der relevanten Daten, um diese optimal für unterschiedliche Anwendungsfälle identifizierbar und auswählbar zu machen) – diese bildet die Basis für weitere Nutzungsszenarien.

Vor allem Formelinhalte, bei denen die Kennzeichnung, die individuelle Ausspielbarkeit und dokumentübergreifende Vernetzung von Bedeutung sind, bieten sich hierbei an. Da Formeln bereits als XML-Element gekennzeichnet sind, können sie schnell identifiziert werden. Dies ist eine gute Grundlage für die weitere Verarbeitung.

Formeln können über ihre eigene Darstellung hinaus außerdem mit zusätzlichen Informationen versehen sein. Bei der Darstellung von mathematischen, physikalischen oder chemischen Zusammenhängen durch Formeln sind im Kontext einer Norm zusätzliche Informationen oft außerhalb der Formel selbst dargestellt. Dies können zusätzliche erweiternde oder beschränkende Eigenschaften sein, die begleitend zur Formel im Umgebungstext platziert werden. Ebenso sind alternative oder ergänzende Beschreibungen desselben oder eines verwandten Sachverhalts möglich, die in derselben oder weiteren Normen vermerkt sind.

110 Siehe Normungsroadmap KI (Ausgabe 1), Kapitel 11.4.1 [63].

111 Siehe Normungsroadmap KI (Ausgabe 1), Kapitel 11.4.2 [63].

Diese Informationen können die Bedeutung einer Formel je nach Kontext erweitern oder konkretisieren. Ihre korrekte und fallbezogene Bewertung und Beachtung stellt dadurch eine große Herausforderung dar. Hieraus ergibt sich auch, dass Formeln in Normen zwar in sich selbst eine große Relevanz für das Verständnis der durch sie ausgedrückten Zusammenhänge entfalten, in ihrer aktuellen Darstellung in Normen aber nicht von dem sie umgebenden Kontext loslösbar sind.

Das Bedürfnis von Nutzer*innen, den Umgang mit Formeln aus Normen im Praxisalltag zu vereinfachen, wird durch die Vielzahl von Softwareangeboten deutlich, die sich auf die praxisbezogene Unterstützung der Formelnutzung für unterschiedlichste Anwendungsfälle spezialisiert haben.

Gelingt es, anhand von Formeln neue, möglichst automatisierte Extraktions- und Semantikprozesse zu erarbeiten, können diese auf weitere Inhaltsarten übertragen werden. Wichtig sind hierbei stringente Content-Publishing-Policies, möglichst flexible Datenstrukturen auf Basis von zunächst XML oder vergleichbaren Umfeldern, die die Erstellung skalierbare Inhaltsdatenbanken ermöglichen.

Die umfassende Betrachtung von Formeln, ihre optimale Ablage, kontexbezogene Verknüpfung und die prozess- und systemorientierte Ausspielung bildet einen wichtigen Ansatz für die Modellierung und Bereitstellung aller weiteren Norminhalte.

KI-gestützte Verfahren stellen hierbei den vielversprechendsten Ansatz dar, die benötigten Datenstrukturen anzulegen, da nur so die gewünschten SMARTen Norminhalte zeitnah zur Verfügung gestellt werden können.

Ausgangssituation 2020
Der weitgehend automatisierte und KI-gestützte Gesamtprozess erfordert ein integriertes übergreifendes Handeln der Prozessverantwortlichen, sodass bisherige Verantwortungsgrenzen überdacht und neu festgelegt werden müssen. Definitiv muss die Content-Verantwortlichkeit für die „Content Creation“ im Prozess der Entwicklung der Normen – der Primärinhalte – verortet sein. Ein Postprocessing i. S. einer nachträglichen Interpretation bzw. „Deutung“ von Inhalten für die Weiterverarbeitung darf es nicht mehr geben.

Normungsphase: Derzeit kann die Sprache (Prosa) der Fachexpert*innen nicht direkt in eine maschineninterpretierbare Form im Sinne von SMART Standards transformiert werden. Mit zukünftig vorliegender Erfahrung und gelerntem Wissen in KI-Anwendungsprozessen ist dennoch zu konzipieren, dass eine KI-orientierte Modellierung realisiert werden kann.

Formalisierung und IV. Modellierung: Die Transformation mittels „Semantischer Tripels“ kann eine direkte Schnittstelle zu KI-Prozesse darstellen. Eine enge Zusammenarbeit ist erforderlich.[112]

Aktueller Status 2022 und Weiterarbeit
Der Umfang, in dem die Inhalte von SMART Standards einer Maschineninterpretierbarkeit zugänglich gemacht werden können, hängt direkt davon ab, inwieweit es gelingt, die dafür notwendigen strukturierten Informationen bereits während des Normenerarbeitungsprozesses, also innerhalb der Gremienarbeit, zu erfassen.

Dabei legt wiederum die Art der Strukturierung den Schwierigkeitsgrad dieser Aufgabe fest. Hier kommt das Informationsmodell ins Spiel, das definiert, wie Normeninhalte fragmentiert, vernetzt und mit Metadaten versehen werden.

Die Größe der erzeugten Fragmente beeinflusst dabei einerseits maßgeblich, inwieweit die Inhalte einer zuverlässigen automatisierten Nutzung zugänglich gemacht werden können, andererseits aber auch den Aufwand, der bei ihrer Erstellung anfällt.

Mit abnehmender Größe der Fragmente wächst somit auch die Bedeutung einer benutzerfreundlichen Tool-Unterstützung, die den zusätzlichen Aufwand bei der Erfassung der Norminhalte minimiert. Hierbei kommen voraussichtlich KI-gestützte Systeme zum Einsatz, die auf Basis eines Pre-Processings Vorschläge für die Modellierung der Inhalte anzeigen, die von den Normenautoren bestätigt werden.

Bereits heute werden in der Normung XML-Dokumente (NISO-STS) erzeugt, die eine, wenn auch grobe, Fragmentierung besitzen, welche sich im Wesentlichen an den Layoutstrukturen orientiert. Für Systeme, die Normeninhalte verstehen sollen, ist jedoch eine entsprechende semantische Strukturierung erforderlich.

Die theoretische Grundlage bildet das im Projekt 2 bei CEN-CENELEC erarbeitete Informationsmodell, das derzeit bei der IEC weiterentwickelt wird.

112 Siehe Normungsroadmap KI (Ausgabe 1), Kapitel 11.4.3 [63].

Es definiert die „Provision“ als zentrales Element und Fragment. Dies steht im Einklang sowohl mit den geltenden Normungsregeln (ISO Directives Teil 2) als auch mit den wichtigsten der bisher identifizierten Anwendungsfälle, wie etwa dem des Anforderungsmanagements, und entspricht dem Level 3 des IEC Utility Models. Ziel ist es, den Normenanwender*innen bis Ende 2024 entsprechend fragmentierte bzw. modellierte Norminhalte zur Verfügung zu stellen.

Aber auch die nächsten Schritte zum Level 4 des Utility Models wurden bereits in Pilotprojekten seit 2020 erprobt. Dabei wurde die „Provision“, das kleinste Element der „Zwischenstufe“ Level 3, weiter zerlegt. Die dabei verwendeten Ansätze reichen von einer Fragmentierung mittels Schablonen in semantische „Gruppen“ (z. B. „condition“, „subject“, „action“, „object“ etc.) bis zur vollständigen Modellierung der natürlichen Sprache in semantischen Tripeln mittels RDF.

Zur Erreichung einer vollumfänglichen Maschineninterpretierbarkeit von Norminhalten stellen die dabei gesammelten Erfahrungen eine wichtige Grundlage für die Arbeiten an SMART Standards ab 2025 dar. Denn nur solche vollständig modellierten Inhalte erlauben eine sichere Anwendung aller Normeninhalte durch KI-gestützte Systeme.

6 Umsetzung der 1. Ausgabe der Normungsroadmap KI

Mit Veröffentlichung der 1. Ausgabe der Roadmap begann die Phase der Umsetzung und Verstetigung ihrer Ergebnisse. Dabei galt es, möglichst viele der identifizierten Handlungsempfehlungen unter Mitwirkung von Expert*innen aus Wirtschaft, Forschung und Zivilgesellschaft und mit Unterstützung der Bundesministerien rasch umzusetzen. Zentrales Ziel der Verstetigung ist es, die identifizierten Themen in den einschlägigen Normungsgremien einzugliedern und konkrete Umsetzungs- bzw. Normungs- und Standardisierungsaktivitäten möglichst europäisch oder international anzustoßen. Mithilfe der entstehenden Normen und Standards sollen die identifizierten Potenziale gehoben und die internationale Wettbewerbsfähigkeit der deutschen Wirtschaft unterstützt werden. Im Folgenden wird der aktuelle Stand zur Umsetzung der Ergebnisse der 1. Ausgabe dargestellt.

Die Ausgabe 1 der Normungsroadmap KI formuliert für sieben Schwerpunktthemen fünf übergreifende Handlungsempfehlungen und insgesamt 78 Bedarfe mit zum Teil sehr unterschiedlichem Charakter – beispielsweise was die Zielgruppe oder den Reifegrad angeht. Zur Umsetzung der Ergebnisse der Roadmap wurde im ersten Schritt ein Verstetigungskonzept entwickelt, das die systematische Verankerung der Bedarfe in den einschlägigen Normungsgremien und die Initiierung von Normungs- und Standardisierungsprojekten zum Ziel hat. Dafür wurden die 78 Handlungsbedarfe der Roadmap nach ihrer Zielgruppe analysiert und wie folgt kategorisiert:

→ Kategorie 1: Bedarf adressiert Normung und Standardisierung
→ Kategorie 2: Bedarf adressiert Forschung
→ Kategorie 3: Bedarf adressiert Politik/Gesetzgeber
→ Kategorie 4: Kein Bedarf (Hinweise, Anmerkungen)

Unter der Kategorie 1 sind Empfehlungen zusammengefasst, die Bedarfe für Normen aufzeigen und sich damit an die Normungsorganisationen richten. Die Bedarfe der Kategorie 2 hingegen betreffen Bereiche, die zum gegenwärtigen Zeitpunkt Gegenstand der Forschung sind. Ziel ist hierbei die Initiierung von Forschungsprojekten und eine frühzeitige entwicklungsbegleitende Normung. Die Bedarfe der Kategorie 3 adressieren in erster Linie die Politik bzw. den Gesetzgeber. Sie zielen beispielsweise darauf ab, Rechtsrahmen oder Vorschriften zu überarbeiten, oder zeigen auf, wo Unterstützung durch die Politik geboten ist. Kategorie 4 subsumiert Hinweise/Anmerkungen oder Vorschläge für Vorgehensweisen, die bei der Erarbeitung der vorliegenden Normungsroadmap Berücksichtigung gefunden haben.

Abbildung 51 zeigt die Verteilung der 78 Bedarfe der Normungsroadmap KI auf die vier Kategorien.

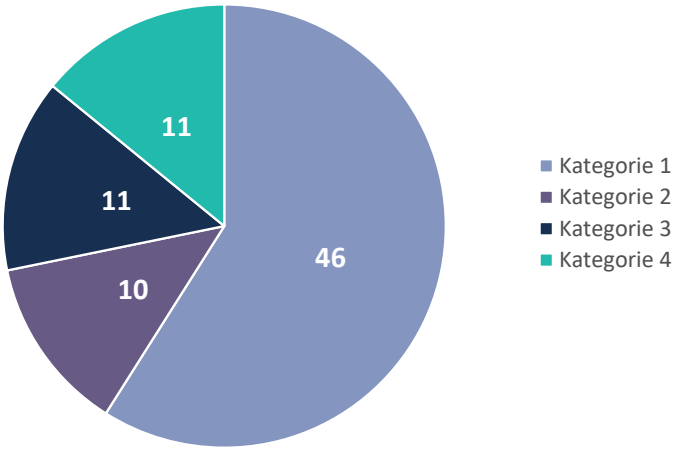


Abbildung 51: Verteilung der Bedarfe auf die Kategorien (Stand: Oktober 2022, Quelle: DIN)

6.1 Normungs- und Standardisierungsbedarfe

Erwartungsgemäß richtet sich die Mehrheit der identifizierten Bedarfe an die Normung. Um diese Kategorie-1-Bedarfe zeitnah in Normungs- und Standardisierungsprojekte zu überführen, wurden sie thematisch den einschlägigen Normenausschüssen zugeordnet und einer weiteren Analyse unterzogen. Zur Priorisierung wurden die Normungsbedarfe zunächst nach ihrem Reifegrad (Notwendigkeit der Konkretisierung oder Weiterentwicklung) und der Dringlichkeit ihrer Umsetzung bewertet und schließlich in einer Vielzahl von Fachworkshops und Sitzungen mit den Expert*innen der Ausschüsse diskutiert – stets mit dem Ziel, die Themen in den Arbeitsprogrammen der Ausschüsse einzugliedern und zeitnah konkrete Normungsprojekte anzustoßen.

Da in den betroffenen Ausschüssen nicht zwingend die notwendige KI-Expertise vorhanden ist, stellt die Gewinnung neuer Expert*innen für die Normungsarbeit einen kritischen Erfolgsfaktor bei der Umsetzung der Bedarfe dar. Interessierte Fachleute sind daher stets zur Mitarbeit in den entsprechenden Ausschüssen eingeladen.

Abbildung 52 zeigt die Verortung der Kategorie-1-Bedarfe in den Normenausschüssen. Da sich ein Bedarf oft in mehreren Ausschüssen thematisch verorten lässt, zeigt die Darstellung vereinfacht nur diejenigen Normenausschüsse, die als Hauptansprechpartner*innen geführt werden. Die Abbildung

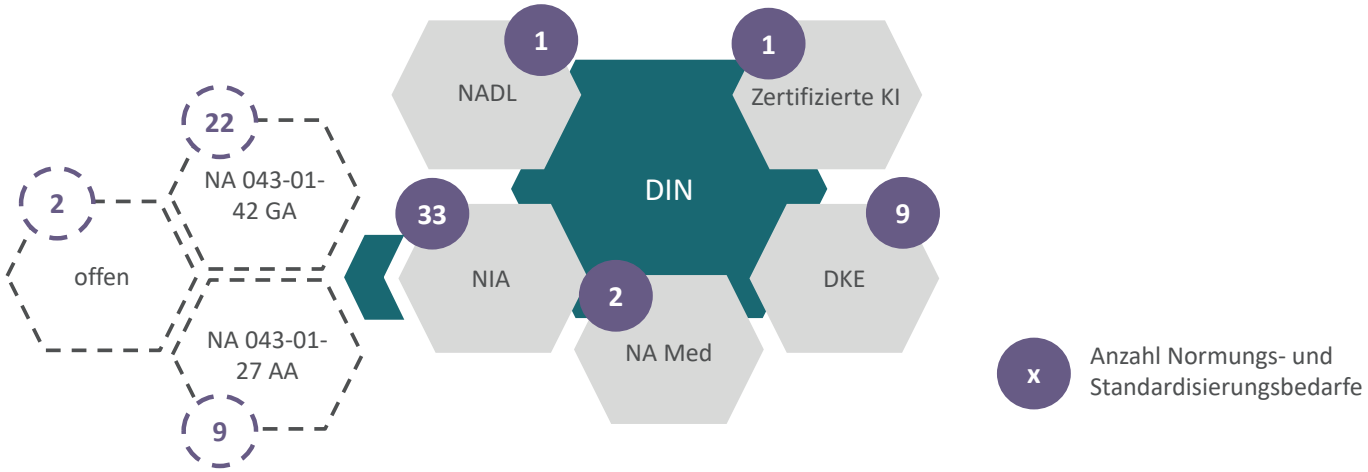


Abbildung 52: Verteilung der Normungsbedarfe auf die Normenausschüsse (Stand: Oktober 2022, Quelle: DIN)

macht deutlich, dass insbesondere der Normenausschuss für Informationstechnik und Anwendungen (NIA), in dem der DIN/DKE-Gemeinschaftsausschuss zu KI eingegliedert ist, aktuell den relevantesten Normenausschuss für die Umsetzung der Bedarfe darstellt.

Aus den vielfältigen Umsetzungsbestrebungen hat sich eine Vielzahl an Normungs- bzw. Standardisierungsaktivitäten ergeben, die in Abbildung 53 aufgezeigt werden.

Von den 46 identifizierten Normungs- und Standardisierungsbedarfen konnten 20 in bereits laufende Normungsprojekte integriert und 15 neue Normungsprojekte angestoßen werden. Tabelle 11 und Tabelle 12 zeigen die Normungsbedarfe, die in laufende Normungsprojekte überführt bzw. für die neue Normungsprojekte initiiert werden konnten.

Bei den verbleibenden elf Bedarfen der Kategorie 1 wurde in den Diskussionen mit den Expert*innen der Normenausschüsse konstatiert, dass eine Überführung in Normungsprojekte zum jetzigen Zeitpunkt nicht möglich ist. Gründe hierfür sind einerseits die fehlende KI-Expertise in den einschlägigen Normenausschüssen und andererseits die Notwendigkeit zur Weiterentwicklung bzw. Konkretisierung der Bedarfe, bevor sie an die relevanten Normenausschüsse übergeben und Normungsprojekte initiiert werden können.

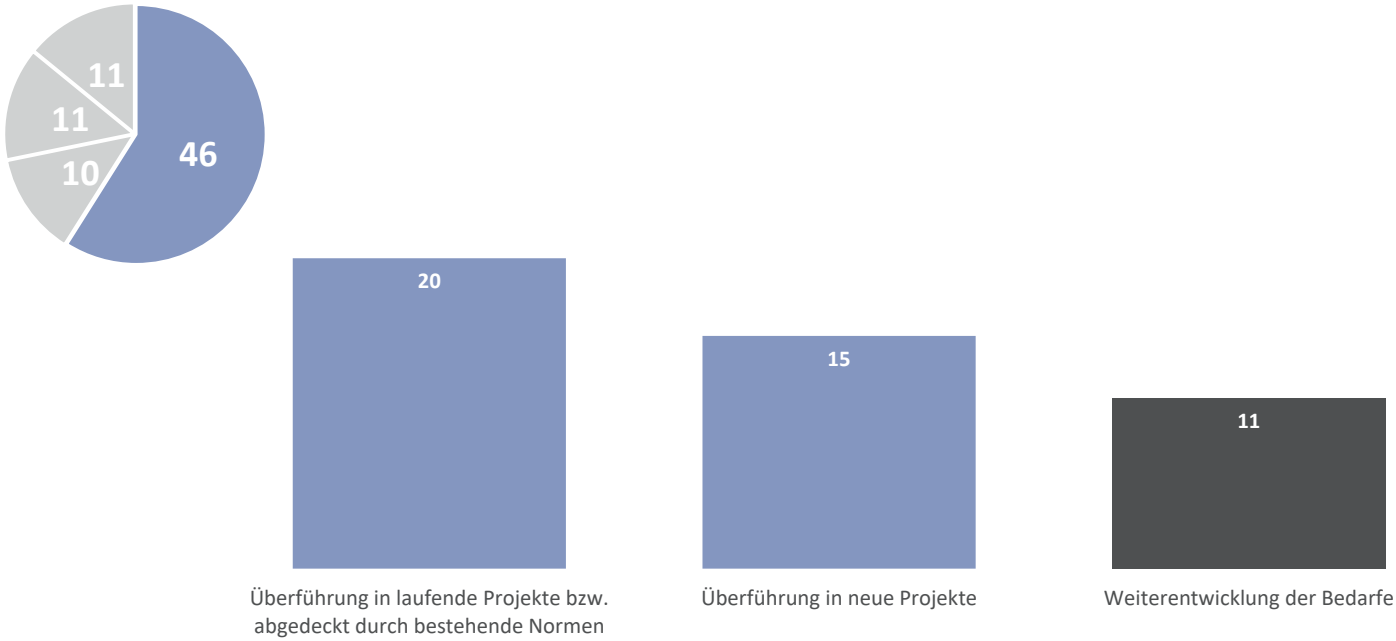


Abbildung 53: Überführung der Bedarfe in die Normung (Stand: Oktober 2022, Quelle: DIN)

Tabelle 11: Bedarfe überführt in laufende Normungsprojekte

Bedarf	Ausschuss	Norm
Definition von Daten und Verwendung festlegen	NA 043-01-42 GA	ISO/IEC 2382 [429]
Controls für IT-Sicherheit definieren	NA 043-01-27-01 AK	DIN EN ISO/IEC 27000 (Reihe) [479]
Risikobewertung von IT-Sicherheit hinsichtlich KI-Systeme	NA 043-01-27-01 AK	ISO/IEC 27005:2018 [161]
Normung eines Konzepts für privacy ethical design	NA 043-01-27-01 AK	DIN EN ISO/IEC 29100:2020 [133], DIN EN ISO/IEC 29134:2020 [134], DIN EN ISO/IEC 27701:2021 [128]
Datenqualitätsmanagement für KI	NA 043-01-42 GA	ISO/IEC 5259 (Reihe) [39]
Art und Qualität von Daten definieren	NA 043-01-42 GA	ISO/IEC 5259 (Reihe) [39]
Zweckbindung von Daten gestalten	NA 043-01-42 GA, NA 043-01-27-01 AK	ISO/IEC 5259 (Reihe) [39], DIN EN ISO/IEC 27701 [128]
Managementsystem für KI, das Anforderungen und Prozesse für Organisationen, die KI entwickeln oder nutzen, definiert (unter Berücksichtigung organisatorischer, technischer und prozessbezogener Prüfverfahren sowie Prüfschemata über den gesamten Lebenszyklus von KI-Systemen)	NA 043-01-42 GA	ISO/IEC 42001 [27]
Unterstützung der internationalen Standardisierungsarbeiten zu einem MSS (Managementsystemstandard) für KI	NA 043-01-42 GA	ISO/IEC 42001 [27]
Risikomanagement für KI	NA 043-01-42 GA	ISO/IEC 23894:2022 [25]
Reevaluierung von KI-Systemen gestalten	NA 043-01-42 GA	ISO/IEC 38507:2022 [26]
Einschränkungen bei Big Data festlegen	NA 043-01-42 GA	ISO/IEC TR 20547 (Reihe) [438], [439], [440], [441], [442]
Grundsätze für die Mensch-Maschine-Mensch-Interaktion im medizinischen Bereich	NA 063-07-02 AA	DIN EN IEC 81001-5-1:2022-01 – Entwurf, VDE 0750-103-5-1:2022 [430]
IT-Sicherheitsmetriken für Lernende Systeme und Adversarial Machine Learning (AML)	DIN SPEC 92001-2: 2020 WS	DIN SPEC 92001-2:2020 [240]
Kriterien zur Klassifikation von Systemen bzw. Komponenten im Rahmen der Künstlichen Intelligenz	NA 043-01-42 GA	ISO/IEC 5392 [32] ISO/IEC 42001 [27]
Erfassung von Begriffen aus unterschiedlichen Disziplinen (Glossar)	NA 043-01-41 AA; DKE	ISO/IEC 20924:2021 [431]
Standardisierte Aufbereitung von Use Cases	NA 043-01-42 GA, IEC TC65/WG23	ISO/IEC TR 24030 [293] PD IEC TR 63283-2 [294] ISO/IEC 22989:2022 [16]
Datenreferenzmodell für Interoperabilität schaffen	NA 043-01-42 GA	ISO/IEC 20547-3:2020 [440]
Funktionsreferenzmodell für Interoperabilität schaffen	NA 043-01-42 GA	ISO/IEC 42001 [27]
Verfahren für Datenaustausch festlegen	NA 043-01-32 AA	ISO/IEC 19763-3:2020 [426]

Tabelle 12: Bedarfe überführt in neue Normungs- und Standardisierungsprojekte

Bedarf	Ausschuss	Norm
Prüfkriterien und -methoden für technische Prüfungen von KI-Lösungen	NA 043-01-42 GA	DIN/TS 92004 [427]
Beziehung zwischen technischen Anforderungen einerseits und rechtlichen und ethischen Anforderungen andererseits	NA 043-01-42 GA	ISO/IEC TR 29119-11 [132]
Management von Transparenz und Vermeidung von Diskriminierung	NA 043-01-42 GA	ISO/IEC TS 12791 [38] ISO/IEC 12792 [238]
Quality Backward Chain im KI-Life-Cycle	NA 043-01-42 GA	ISO/IEC 5338 [30]
IT-Sicherheit von KI-Systemen bei mangelnder Verfügbarkeit von Ressourcen (Angriffsvektor)	NA 043-01-27-01 AK	ISO/IEC TR 27563 [138]
Designprinzipien für KI-Systeme	NA 043-01-42 GA	ISO/IEC TS 5471 [33][34] ISO/IEC 5338 [30]
KI-Security-by-Design und KI-Security-by-Default	NA 043-01-27-03 AK	ISO/IEC 7699 [428]
IT-Sicherheitskriterien für Lernende Systeme	NA 043-01-27-03 AK	ISO/IEC 7699 [428]
IT-Sicherheit der Training-Daten	NA 043-01-27-03 AK	ISO/IEC 7699 [428]
Kritikalitätsstufen und IT-Sicherheit	NA 043-01-42 GA	Ad Hoc Gruppe „KI-Klassifizierung“– Draft für den CEN/CLC JTC 21 Artificial Intelligence wird derzeit erarbeitet.
IT-Sicherheitskriterien für Trainingsmethoden	NA 043-01-42 GA	Ad Hoc Gruppe „KI-Klassifizierung“– Draft für den CEN/CLC JTC 21 Artificial Intelligence wird derzeit erarbeitet.
Fehlerklassifikationen, Fehleinordnungen und Lernen aus Fehlern definieren	NA 043-01-42 GA	ISO/IEC 42005 [432]
Prüfprozess zum Evaluieren vorhandener Prinzipien	NA 063-07-02 AA	Von NA 063-07-02 AA (wird in ISO/TC 215 WG 1/2 eingebracht)
Initiale Kritikalitätsprüfung von KI-Systemen schnell und einfach gestalten (Risikomatrix)	NA 043-01-42 GA	Ad Hoc Gruppe „KI-Klassifizierung“– Draft für den CEN/CLC JTC 21 Artificial Intelligence wird derzeit erarbeitet.
Explainable AI		DIN SPEC 92001-3 [117]

6.2 Forschungsbedarfe

Die Bedarfe der Kategorie 2 richten sich in erster Linie an die Forschungscommunity. Ausgabe 1 der Roadmap hat zehn Bedarfe dieser Art identifiziert. Ziel der Umsetzungsbestrebungen ist hierbei die Initiierung von Forschungsprojekten. Eine wesentliche Säule dieser Projekte stellt die entwicklungsbegleitende Normung dar, bei der die Projektergebnisse frühzeitig der Normung und Standardisierung zugeführt werden und damit ganz maßgeblich der Transfer wissenschaftlicher Ergebnisse in marktfähige Produkte und Dienstleistungen unterstützt wird. Normung und Standardisierung sind folglich ein Katalysator für Innovationen, der die Markterschließung, -durchdringung und Internationalisierung technologischer Neu- und Weiterentwicklungen begünstigt.

Aus den Ergebnissen der 1. Ausgabe der Roadmap konnten drei Forschungsprojekte angestoßen werden, bei denen vier der identifizierten Forschungsbedarfe aufgegriffen und umgesetzt werden: Zertifizierte KI, KI-Tauglichkeit von Normen und KIMEDS (siehe Kapitel 3.3.1 und Kapitel 3.3.2). Die verbleibenden sechs Forschungsbedarfe erfahren im Rahmen der vorliegenden Roadmap eine Konkretisierung bzw. Weiterentwicklung.

6.3 Politische Bedarfe

In der Ausgabe 1 der Roadmap wurden insgesamt elf Bedarfe formuliert, die sich an die Politik richten. Zur Umsetzung dieser Bedarfe hat sich DIN aktiv in den politischen Diskurs zu KI (beispielsweise zum AI Act) eingebracht: Die Ergebnisse der Normungsroadmap wurden sowohl der Europäischen Kommission und den EU-Parlamentariern als auch der Bundesregierung und Mitgliedern des Bundestages vorgestellt.

Aus den identifizierten Bedarfen sind schließlich politische Forderungen und Empfehlungen abgeleitet, in einem Positionspapier[113] zu KI zusammengefasst und an die Politik adressiert worden.

Insbesondere die Forderung nach der Anbindung an die internationale Normung und die finanzielle Unterstützung deutscher Expert*innen (insbesondere aus kleinen und mittleren Unternehmen, Wissenschaft und Zivilgesellschaft) zur aktiven Mitarbeit an internationalen und europäischen Normungsprojekten finden sich in dem Positionspapier wieder. Eine derartige Unterstützung gilt als essenziell, um die Berücksichtigung nationaler Interessen und europäischer Werte sicherzustellen.

113 www.din.de/resource/blob/857886/92863b23027a9737056f-6ca122035931/kurzpositionspapier-kuenstliche-intelligenz-data.pdf

6.4 Übergreifende Handlungsempfehlungen

In der Ausgabe 1 der Normungsroadmap KI wurden insgesamt fünf übergreifende Handlungsempfehlungen formuliert. Ihnen kommt eine besondere Bedeutung zu, da sie sämtliche Bereiche der 1. Ausgabe der Normungsroadmap KI betreffen und sich an Normung, Forschung und Politik gleichermaßen richten.

Die Handlungsempfehlungen werden derzeit umgesetzt – u. a. in Forschungs-, Normungs- und Umsetzungsprojekten. Im Folgenden wird der aktuelle Stand (Stand: Oktober 2022) der Umsetzung dargestellt.

1. **Datenreferenzmodelle für die Interoperabilität von KI-Systemen umsetzen**
   In Wertschöpfungsketten kommen viele unterschiedliche Akteur*innen zusammen. Damit auch die verschiedenen KI-Systeme dieser Akteur*innen automatisiert zusammenarbeiten können, ist ein Datenreferenzmodell nötig, um Daten sicher, zuverlässig, flexibel und kompatibel auszutauschen. Standards für Datenreferenzmodelle aus unterschiedlichen Bereichen schaffen die Grundlage für einen übergreifenden Datenaustausch und stellen damit weltweit die Interoperabilität von KI-Systemen sicher.
   Umsetzungsstand:
   Derzeit wird in der internationalen KI-Normung intensiv an dem Thema Daten gearbeitet. Hierzu sind folgende laufende Normungsprojekte zu nennen:
   - ISO/IEC 5259-1 „Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 1: Overview, terminology, and examples“ [40]
   - ISO/IEC 5259-2 „Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 2: Data quality measures“ [41]
   - ISO/IEC 5259-3 „Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 3: Data quality management requirements and guidelines“ [42]
   - ISO/IEC 5259-4 „Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 4: Data quality process framework“ [43]

- ISO/IEC 5259-5 „Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 5: Data quality governance“ [44]
- ISO/IEC 8183 „Informationstechnik – Künstliche Intelligenz – Framework für den Lebenszyklus von Daten“ [45]

Sie alle zahlen maßgeblich in die Umsetzung der vorliegenden Handlungsempfehlung ein.

2. **Horizontale KI-Basis-Sicherheitsnorm erstellen**
   KI-Systeme sind im Kern IT-Systeme – für Letztere gibt es bereits viele Normen und Standards aus verschiedensten Anwendungsbereichen. Um ein einheitliches Vorgehen beim Thema IT-Sicherheit von KI-Anwendungen zu ermöglichen, ist eine übergreifende „Umbrella-Norm“ sinnvoll, die vorhandene Normen und Prüfverfahren für IT-Systeme bündelt und um KI-Aspekte ergänzt. Diese Basis-Sicherheitsnorm kann dann durch Subnormen zu weiteren Themen ergänzt werden.
   Umsetzungsstand:
   Aktuell wird im KI-Umfeld intensiv an einer gemeinsamen internationalen und europäischen Standardisierungslandschaft gearbeitet, die auch das Thema Sicherheit im Sinne von IT-Security, Safety und Privacy beinhalten soll. Derzeit gibt es noch keine konkreten Projekte auf CEN/CENELEC- und ISO/IEC-Ebene. Um eine horizontale KI-Basis-Sicherheitsnorm zu ermöglichen, sind zwingend weitere aktive KI-Fachleute in der Normung sowie eine verstärkte Präsenz in internationalen KI-Normungsgremien erforderlich. Insbesondere ist hier noch Überzeugungsarbeit zu leisten.
3. **Praxisgerechte initiale Kritikalitätsprüfung von KI-Systemen ausgestalten**
   Wenn selbstlernende KI-Systeme über Menschen, deren Besitz oder Zugang zu knappen Ressourcen entscheiden, können ungeplante Probleme in der KI individuelle Grundrechte oder demokratische Werte gefährden. Damit sich KI-Systeme in ethisch unkritischen Anwendungsfeldern dennoch frei entwickeln lassen, sollte durch Normen und Standards eine initiale Kritikalitätsprüfung gestaltet werden – diese kann schnell und rechtssicher klären, ob ein KI-System solche Konflikte überhaupt auslösen kann.
   Umsetzungsstand:
   Im August 2022 wurde auf internationaler Ebene das Projekt ISO/IEC 42005 [432] „Information technology – Artificial intelligence – AI system impact assessment“ unter deutscher Leitung initiiert. Dieses Dokument ist ein Leitfaden für Organisationen, die KI-Systemfolgenabschätzungen für Einzelpersonen und Gesellschaften durchführen, welche von einem KI-System und seinen beabsichtigten und vorhersehbaren Anwendungen betroffen sein können. Es enthält Überlegungen dazu, wie und wann solche Abschätzungen durchzuführen sind und in welchen Phasen des Lebenszyklus eines KI-Systems sowie eine Anleitung zur Dokumentation der Folgenabschätzung für KI-Systeme. Darüber hinaus wird erläutert, wie der Prozess der Folgenabschätzung für KI-Systeme in das KI-Risikomanagement und das KI-Managementsystem einer Organisation integriert werden kann. Dieses Dokument ist für Organisationen gedacht, die KI-Systeme entwickeln, bereitstellen oder nutzen. Das Dokument ist auf jede Organisation anwendbar, unabhängig von Größe, Art und Beschaffenheit.
4. **Nationales Umsetzungsprogramm „Trusted AI“ zur Ertüchtigung der europäischen Qualitätsinfrastruktur initiieren**
   Bisher fehlen verlässliche Qualitätskriterien und Prüfverfahren für KI-Systeme – das gefährdet das wirtschaftliche Wachstum und die Wettbewerbsfähigkeit dieser Zukunftstechnologie. Es braucht ein nationales Umsetzungsprogramm „Trusted AI“, das die Basis für reproduzierbare und standardisierte Prüfverfahren legt, mit denen Eigenschaften von KI-Systemen wie Verlässlichkeit, Robustheit, Leistungsfähigkeit und funktionale Sicherheit geprüft und Aussagen über die Vertrauenswürdigkeit getroffen werden können. Normen und Standards beschreiben Anforderungen an diese und bilden so die Grundlage für die Zertifizierung und Konformitätsbewertung von KI-Systemen. Mit einer solchen Initiative hat Deutschland die Chance, ein weltweit erstes und international anerkanntes Zertifizierungsprogramm zu entwickeln.
   Umsetzungsstand:
   Im Normungsumfeld wurden verschiedene Initiativen zu „Trusted AI“ gestartet. Fokus ist die Erarbeitung von Managementsystemstandards zur Zertifizierung des vertrauenswürdigen Umgangs mit KI sowie die Festschreibung von Anforderungen an zertifizierende Organisationen. Zu nennen sind hier die Normungsprojekte:
   - ISO/IEC 42001 [27] „Information technology – Artificial intelligence – Management system“
   - ISO/IEC 23894 [25] „Information technology – Artificial intelligence – Guidance on risk management“
   - ISO/IEC 42005 [432] „Information technology – Artificial intelligence – AI system impact assessment“

   Zur Umsetzung der Handlungsempfehlung wurde darüber hinaus Anfang 2021 das Umsetzungsprojekt ZERTIFIZIERTE KI (siehe Kapitel 3.3.2) gestartet, bei dem Prüfkriterien, -methoden und -werkzeuge für

KI-Systeme entwickelt und standardisiert werden sollen, um so eine vergleichbare Bewertung von KI-Systemen zu ermöglichen. Durch einen breit angelegten Beteiligungsprozess soll zudem sichergestellt werden, dass sich die Verfahren zu allgemein akzeptierten Standards für KI-Systeme und deren Überprüfung entwickeln.

5. **Use Cases auf Normungsbedarf analysieren und bewerten**
   Die KI-Forschung sowie die industrielle Entwicklung und Anwendung von KI-Systemen sind hoch dynamisch. Bereits heute gibt es viele Anwendungsfälle in den verschiedenen Einsatzfeldern von KI. Über anwendungstypische und branchenrelevante Use Cases lassen sich Standardisierungsbedarfe für industriereife KI-Anwendungen ableiten. Um Normen und Standards zu gestalten, ist es wichtig, wechselseitige Impulse aus Forschung, Industrie, Gesellschaft und Regulierung einzubinden. Im Zentrum dieses Ansatzes sollten die entwickelten Standards entlang von Use Cases erprobt und weiterentwickelt werden. So lassen sich anwendungsspezifische Bedarfe frühzeitig erkennen und marktfähige KI-Standards realisieren.
   Umsetzungsstand:
   Der Technische Bericht ISO/IEC TR 24030:2021 [293] „Information technology – Artificial intelligence (AI) – Use cases“ wurde vom internationalen Normungsgremium ISO/IEC/JTC 1/SC 42/WG 4 „Use cases and applications“ erarbeitet und im Mai 2021 veröffentlicht. Das Dokument enthält eine Use-Case-Sammlung und stellt damit eine gute Grundlage für die o. g. Handlungsempfehlung dar. Darüber hinaus befasst sich auch die Technical Expert Group „Artificial Intelligence Applications in Industrie 4.0 / Intelligent Manufacturing“ (TEG AIAI2M) innerhalb der Deutsch-Chinesischen Kommission Normung (DCKN) mit dem Thema Use Cases.
   Auch auf europäischer Ebene innerhalb des CEN/CENELEC JTC 21 spielen Use-Case-Betrachtungen eine wichtige Rolle. Insbesondere stehen die Aktivitäten im Zusammenhang mit den geplanten europäischen Regulierungsvorhaben (vornehmlich Data Sovereignity Act und AI Act), da dort konkrete Anwendungen wiederum (exemplarisch) zur Einordnung, u. a. Kritikalität, Verwendung finden und damit sowohl ostentativ als auch komplementär zu oben genannten stehen.
   Darüber hinaus trägt auch das Umsetzungsprojekt ZERTIFIZIERTE KI (siehe Kapitel 3.3.2) maßgeblich zur Umsetzung der Empfehlung bei. In branchen- und technologiebezogenen Anwenderkreisen werden durch Beteiligte aus Wirtschaft und Wissenschaft konkrete industriespezifische Anwendungsfälle betrachtet, stets mit dem Ziel, Bedarfe zu definieren, Kriterien und Maßstäbe für eine Prüfung in der Praxis festzulegen und diese anhand branchentypischer Use Cases zu verifizieren. Die ermittelten Bedarfe und Erkenntnisse werden im nächsten Schritt in entsprechende Anforderungen an eine vertrauenswürdige Nutzung von KI übersetzt und schließlich der Normung zugeführt.

6.5 Gewinnung von Expert*innen für die Normung

Die Übersetzung der identifizierten Bedarfe in Normungs- und Standardisierungsprojekte und die daran anknüpfende Erarbeitung von Normen und Standards ist nur ein Ziel der Umsetzungsbestrebungen der Roadmap. Ein weiterer Fokus liegt auf der Gewinnung von Expert*innen für die Normungsarbeit. Die Normung ist eine Gemeinschaftsaufgabe und braucht für die Erarbeitung von Normen und Standards fachkundige KI-Expert*innen aus Wirtschaft, Wissenschaft und Zivilgesellschaft, die ihr Wissen aktiv in der Normung einbringen. Nur ein frühzeitiges Engagement von Fachleuten mit Erfahrungswerten und Einblicken aus der Praxis wird es ermöglichen, markt- und bedarfsgerechte Normen und Standards für KI zu erarbeiten. Aktuell sind diese KI-Expert*innen in den Normenausschüssen sehr rar vertreten, was die schnelle Umsetzung der Bedarfe in Normen und Standards deutlich erschwert.

Aus den bisherigen Verstetigungsaktivitäten konnten über zwei Dutzend neue fachkundige Expert*innen (Stand: Oktober 2022) gewonnen werden, die sich fortan in den Normungsgremien engagieren und ihr Know-how bei der Erarbeitung von Normen und Standards zu KI einbringen. Das stellt einen guten Startpunkt dar, ist aber im Hinblick auf die vielfältigen Potenziale und Bedarfe, die die Normungsroadmap KI aufzeigt, nicht ausreichend. Wenn Deutschland sicherstellen möchte, dass seine Interessen angemessen in internationalen KI-Standards Berücksichtigung finden, sind weitere aktive KI-Fachleute in der Normung erforderlich und eine verstärkte Präsenz in internationalen KI-Normungsgremien ist dringend angeraten.

6.6 Leuchtturmprojekte

Als eine weitere Verstetigungsmaßnahme wurde von der Koordinierungsgruppe „KI-Normung und -Konformität“ die Notwendigkeit sogenannter „Leuchtturmprojekte der Deutschen Normungsroadmap KI“ festgestellt. Unter einem Leuchtturmprojekt werden anwendungstypische und branchenrelevante Use Cases verstanden, die für KI-spezifische Anwendungen Anforderungen an die Normung und Standardisierung aufzeigen. Mithilfe der Leuchtturmprojekte sollen im jeweiligen Anwendungskontext praktische Erfahrungen gesammelt, konkrete Normungs- und Standardisierungsbedarfe abgeleitet und Erkenntnisse zur Qualitäts- und Konformitätsprüfung gewonnen werden. Ihnen kommt damit eine besondere Bedeutung bei der Umsetzung der Normungsroadmap KI zu, weshalb sie eine hohe Aufmerksamkeit bei den Normungsakteur*innen genießen sowie in Wirtschaft, Forschung und Politik große Sichtbarkeit und Strahlkraft besitzen. Das Konzept der Leuchtturmprojekte legt klare Rahmenbedingungen beispielsweise zum Auswahlprozess, zu Bewertungskriterien und Projektpatenschaften fest.

Die Bewertungskriterien berücksichtigen u. a.:

- → die ausgewogene Beteiligung aller relevanten Stakeholderkreise,
- → die strategische Bedeutung und Breitenwirksamkeit in gesamtwirtschaftlicher Hinsicht (z. B. Vorreiterrolle, Technologieführerschaft),
- → die europäische und/oder internationale Anschlussfähigkeit in der Normung,
- → regulatorische Vorgaben,
- → vorhandene Vorarbeiten aus Forschungs- und Umsetzungsprojekten sowie
- → soziotechnische Aspekte (wie beispielsweise humane Arbeitsgestaltung, organisationale Bedingungen etc.).

In einem Auswahlprozess wurden erste Leuchtturmprojekte bzw. Projekte mit Leuchtturmcharakter[114] durch die Koordinierungsgruppe „KI-Normung und -Konformität“ identifiziert und ausgezeichnet, die im Folgenden beschrieben werden: safe.trAIn, Medizinische Diagnose- und Prognosesysteme, Clouddienste sowie NDE4.0.

Safe.trAIn

Das Projekt safe.trAIn[115] (Sichere KI am Beispiel fahrerloser Regionalzug) ist das erste offizielle Leuchtturmprojekt der Normungsroadmap KI. Es wird vom BMWK gefördert und verfolgt seit 2022 das Ziel, KI-Verfahren mit den Anforderungen und Zulassungsprozessen im Bahnumfeld praktikabel zu verknüpfen. Der Fokus des Konsortiums liegt auf der Entwicklung standardisierter Prüfmethoden und -werkzeuge, um die zulassungsrelevante Produktsicherheit für einen breiten Einsatz vollautonomer Züge zu gewährleisten. Außerdem wird die Sicherheitsarchitektur am Beispiel des fahrerlosen Regionalzugs konkretisiert und ein vollautomatisiertes GoA4-System für diesen Anwendungsfall in einem virtuellen Testfeld konzeptionell entwickelt und validiert. Die Ergebnisse des Projekts sollen überführt werden in Normen und Standards. Diese spielen eine entscheidende Rolle für eine beschleunigte Markteinführung und die sichere, robuste sowie vertrauenswürdige Anwendung KI-basierter Methoden für den führerlosen Zugverkehr.

KI-Standards für medizinische Diagnose- und Prognosesysteme

Die Anwendung von KI-Systemen in medizinischen Diagnoseverfahren bietet großes Potenzial. Auch wenn die Zahl der KI-basierten Medizinprodukte auf dem Markt stetig steigt, ist der Prozess der Entwicklung, Herstellung und Markteinführung inklusive Prüfung durch bekannte Stellen bislang sehr aufwendig und kostspielig. Um den Einsatz KI-basierter Medizinprodukte zu erhöhen, müssen einerseits Akzeptanz und Vertrauen geschaffen und andererseits die Entwicklungs- und Zulassungsprozesse vereinfacht werden. Ziel des Projekts ist daher die Entwicklung standardisierter Prüfverfahren und -werkzeuge für medizinische, KI-gestützte Diagnose- und Prognosesysteme, die einen schnelleren und sicheren Marktzugang ermöglichen. Das Vorhaben, das von der Koordinierungsgruppe als „Projekt mit Leuchtturmcharakter“ ausgezeichnet wurde, bezieht explizit Marktteilnehmer*innen aus Industrie, Regulatorik, Forschung und Klinik mit ein, um marktfähige KI-Lösungen zu entwickeln und die Akzeptanz und das Vertrauen für KI-basierte Diagnose- und Prognosesysteme zu steigern.

114 Ein „Projekt mit Leuchtturmcharakter“ zeichnet sich dadurch aus, dass es die festgelegten Bewertungskriterien zwar hinreichend erfüllt, seine Finanzierung und damit auch seine Durchführung zum derzeitigen Moment jedoch nicht gesichert ist. Bis zu seiner Durchführung wird das Vorhaben daher zunächst als „Projekt mit Leuchtturmcharakter“ geführt. Mit Beginn der Projektdurchführung erhält es automatisch den Status „Leuchtturmprojekt der Deutschen Normungsroadmap KI“.

115 https://www.din.de/de/forschung-und-innovation/partner-in-forschungsprojekten/ki/safe-train-860442

Clouddienste

KI-Lösungen sind unter anderem deshalb eine Schlüsseltechnologie bei der Digitalisierung, weil sie skalierbare Cloudtechnologien nutzen. Sie bleiben dadurch in der Entwicklung und im Betrieb wirtschaftlich und sichern die internationale Wettbewerbsfähigkeit der Anwender*innen. Die Nutzung von Plattformen, Infrastrukturen und KI-Frameworks der großen Cloudprovider ermöglicht grundsätzlich den wirtschaftlichen Marktzugang auch für Marktteilnehmer*innen, die nicht über genügend eigene IT-Ressourcen und nur über wenig KI-Expertise verfügen. Durch die Anwendungsbreite kommt der Vertrauenswürdigkeit der hybriden und eingebetteten KI-Lösungen eine besondere Bedeutung zu, wobei ein Großteil der Verantwortung für die Vertrauenswürdigkeit der technischen KI-Komponenten bei den Cloudprovidern, den Entwickler*innen und Betreiber*innen der cloudbasierten KI-Services liegt. Das Vorhaben wurde von der Koordinierungsgruppe als „Projekt mit Leuchtturmcharakter“ ausgezeichnet.

Wesentliche Ziele des Projekts sind:

→ die Vertrauenswürdigkeit der KI-Lösungen in Entwicklung und Betrieb cloudbasierter KI-Services durch international anerkannte Konformitätsprüfungen transparent zu machen,
→ die dafür erforderlichen Prüfkriterien und Prüfverfahren zu entwickeln,
→ die Prüfgrundlagen als Grundlage für die anwendungsunabhängigen horizontalen KI-Standards auf europäischer Ebene einzuführen sowie
→ den Marktzugang zu vertrauenswürdiger KI auch für kleine und mittlere Unternehmen zu akzeptablen Kosten zu ermöglichen.

NDE 4.0

Die zerstörungsfreie Prüfung (ZFP; englisch: Nondestructive Evaluation, NDE) begleitet seit jeher den industriellen Fortschritt. Deutschland nimmt in diesem Feld seit vielen Jahrzehnten eine weltweit führende Rolle ein. NDE-Systeme sind ein zentrales Element in den Konzepten für Qualitäts- und Sicherheitstechnologien in der deutschen Wirtschaft. Dies gilt insbesondere mit Blick auf effiziente Produktionsprozesse, für den sicheren Betrieb technischer Systeme und Anlagen sowie für durchgehende Prozessmethodik. Höchste Relevanz haben NDE-Sensorsysteme im Rahmen von Freigabe-, Wartungs- und Instandhaltungsprozessen, traditionell vor allem im Bereich der kritischen oder resilienten Infrastruktur (z. B. Chemie- und Anlagensicherheit, Energieerzeugung und -verteilung, Transport und Verkehr, Bauinfrastruktur etc.). Durch die fortschreitende Digitalisierung und den Einzug von KI in der ZFP ist der heutige Normungsprozess veraltet und die etablierten Normen decken die rasanten Entwicklungen in der ZFP nicht mehr ab. Aus dieser Motivation heraus ist die Idee zum Vorhaben „Innovationsbeschleunigung durch flexibilisierte Validierungs- und Zertifizierungswege für NDE4.0“ entstanden, welches von der Koordinierungsgruppe zum „Projekt mit Leuchtturmcharakter“ ausgezeichnet wurde. Mithilfe des Projekts soll der Prozess der Zulassung und Normung beschleunigt und somit den Betroffenen beim Einsatz von NDE4.0 Rechtssicherheit gegeben werden.

7 Übersicht über relevante Dokumente, Aktivitäten und Gremien zu KI

Das nachfolgende Kapitel dient dazu, eine Übersicht zu bereits veröffentlichten Normen und Standards (Kapitel 7.1), laufenden Normungs- und Standardisierungsaktivitäten (Kapitel 7.2) und Gremien (Kapitel 7.3) mit Relevanz für KI zu geben. Die Darstellungen erheben keinen Anspruch auf Vollständigkeit.

7.1 Veröffentlichte Normen und Standards mit Relevanz für KI

Tabelle 13 gibt Informationen zu bereits veröffentlichten Normen und Standards mit KI-Bezug sowie zur Relevanz für die Arbeitsgruppen der Normungsroadmap.

Tabelle 13: Überblick über veröffentlichte Normen und Standards mit Relevanz für KI[116]

Dokument	Titel	Datum	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
VDE AR 2842-61 [105]	Entwicklung und Vertrauenswürdigkeit von autonom / kognitiven Systemen	2021	DKE/K 801: System Komitee AAL	X	X		X	X		X		X
ISO/IEC 23053 [24]	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)	2022	NA 043-01-42 GA			X		X		X	X	X
DIN EN 61508-3, VDE 0803-3 [103]	Funktionale Sicherheit sicherheitsbezogener elektrischer/elektronischer/programmierbarer elektronischer Systeme – Teil 3: Anforderungen an Software (IEC 61508-3:2010); Deutsche Fassung EN 61508-3:2010	2011	DKE/GK 914 Funktionale Sicherheit elektrischer, elektronischer und programmierbarer elektronischer Systeme (E, E, PES) zum Schutz von Personen und Umwelt		X			X				X
DIN EN 61508-5, VDE 0803-5 [433]	Funktionale Sicherheit sicherheitsbezogener elektrischer/elektronischer/programmierbarer elektronischer Systeme – Teil 5: Beispiele zur Ermittlung der Stufe der Sicherheitsintegrität (safety integrity level) (IEC 61508-5:2010); Deutsche Fassung EN 61508-5:2010	2011-02	DKE/GK 914 Funktionale Sicherheit elektrischer, elektronischer und programmierbarer elektronischer Systeme (E, E, PES) zum Schutz von Personen und Umwelt		X	X		X				X

116 Diese Übersicht erhebt keinen Anspruch auf Vollständigkeit.

Dokument	Titel	Datum	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
DIN EN 61511-1, VDE 0810-1 [434]	Funktionale Sicherheit – PLT-Sicherheitseinrichtungen für die Prozessindustrie – Teil 1: Allgemeines, Begriffe, Anforderungen an Systeme, Hardware und Anwendungsprogrammierung (IEC 61511-1:2016 + COR1:2016 + A1:2017); Deutsche Fassung EN 61511-1:2017 + A1:2017	2019-02	DKE/GK 914 Funktionale Sicherheit elektrischer, elektronischer und programmierbarer elektronischer Systeme (E, E, PES) zum Schutz von Personen und Umwelt		X			X				X
DIN SPEC 13266 [98]	Guideline for the development of deep learning image recognition systems	2020		X		X		X		X		
DIN EN IEC 62443 (alle Teile) [435]	Industrial communication networks – Network and system security	2020	DKE/UK 931.1 IT-Sicherheit in der Automatisierungstechnik			X		X		X		
ISO/IEC TR 24027 [436]	Information technology – Artificial intelligence (AI) – Bias in AI systems and AI aided decision making	2021	NA 043-01-42 GA	X		X	X	X		X	X	X
ISO/IEC TR 24372 [437]	Information technology – Artificial intelligence (AI) – Overview of computational approaches for AI systems	2021	NA 043-01-42 GA	X		X	X	X		X		X
ISO/IEC TR 24030 [293]	Information technology – Artificial intelligence (AI) – Use cases	2021	NA 043-01-42 GA	X		X		X		X	X	X
ISO/IEC 38507 [26]	Information technology – Governance of IT – Governance implications of the use of artificial intelligence by organizations	2022	NA 043-01-42 GA	X		X	X	X		X		
ISO/IEC TR 24368 [15]	Information technology – Artificial intelligence – Overview of ethical and societal concerns	2022	NA 043-01-42 GA	X			X	X		X	X	X
ISO/IEC TR 24028 [28]	Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence	2020	NA 043-01-42 GA	X		X	X	X		X	X	X
ISO/IEC TR 20547-1 [438]	Information technology – Big data reference architecture – Part 1: Framework and application process	2020	NA 043-01-42 GA			X						X

Dokument	Titel	Datum	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
ISO/IEC TR 20547-2 [439]	Informationstechnik – Big Data Referenzarchitektur – Teil 2: Anwendungsfälle und abgeleitete Anforderungen	2018	NA 043-01-42 GA			X						X
ISO/IEC 20547-3 [440]	Informationstechnik – Big-Data-Referenzarchitektur – Teil 3: Referenzarchitektur	2020	NA 043-01-42 GA			X						X
ISO/IEC 20547-4 [441]	Informationstechnik – Big Data Referenzarchitektur – Teil 4: Sicherheit und Datenschutz	2020	NA 043-01-42 GA			X	X					X
ISO/IEC TR 20547-5 [442]	Informationstechnik – Big Data Referenzarchitektur – Teil 5: Normungsroadmap	2018	NA 043-01-42 GA			X						X
ISO/IEC 20546 [443]	Informationstechnik – Big Data – Überblick und Begriffe	2019	NA 043-01-42 GA									X
ISO/IEC 33063 [444]	Informationstechnik – Prozessbewertung – Prozessbewertungsmodell für Software-Tests	2015	NA 043-01-07 AA					X		X		
DIN EN ISO/IEC 15408-1 [445]	Informationstechnik – IT-Sicherheitsverfahren – Evaluationskriterien für IT-Sicherheit – Teil 1: Einführung und allgemeines Modell (ISO/IEC 15408-1:2009); Deutsche Fassung EN ISO/IEC 15408-1:2020	2020	NA 043-04-27 AA			X		X		X		X
DIN EN ISO/IEC 15408-2 [446]	Informationstechnik – IT-Sicherheitsverfahren – Evaluationskriterien für IT-Sicherheit – Teil 2: Sicherheitsfunktionskomponenten (ISO/IEC 15408-2:2008); Deutsche Fassung EN ISO/IEC 15408-2:2020, nur auf CD-ROM	2020	NA 043-04-27 AA			X		X		X		X
DIN EN ISO/IEC 15408-3 [447]	Informationstechnik – IT-Sicherheitsverfahren – Evaluationskriterien für IT-Sicherheit – Teil 3: Komponenten zur Sicherheitskontrolle (ISO/IEC 15408-3:2008, korrigierte Fassung 2011-06-01); Deutsche Fassung EN ISO/IEC 15408-3:2020, nur auf CD-ROM	2021	NA 043-04-27 AA			X		X		X		X

Dokument	Titel	Datum	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
ISO/IEC 15408-4 [448]	Informationstechnik – IT-Sicherheitsverfahren – Evaluationskriterien für IT-Sicherheit – Teil 4: Rahmen für die Festlegung von Bewertungsmethoden und -tätigkeiten	2022	NA 043-04-27 AA			X		X		X		X
ISO/IEC 15408-5 [449]	Informationstechnik – IT-Sicherheitsverfahren – Evaluationskriterien für IT-Sicherheit – Teil 5: Vordefinierte Pakete von Sicherheitsanforderungen	2022	NA 043-04-27 AA			X		X				X
DIN EN ISO/IEC 18045 [75]	Information technology – Security techniques – Methodology for IT security evaluation	2021	NA 043-04-27 AA			X		X				X
DIN EN 62304 [353]	Medizingeräte-Software – Software-Lebenszyklus-Prozesse	2016	NA 063-01-13 AA		X			X		X		
DIN EN ISO 14971 [351]	Medical devices – Application of risk management to medical devices	2022	NA 063-01-13 AA	X	X			X		X		
ETSI TR 101 583 [450]	Methods for Testing and Specification (MTS); Security Testing; Basic Terminology	2015	European Telecommunications Standards Institute (ETSI)					X		X		X
DIN EN 61513, VDE 0491-2 [451]	Kernkraftwerke – Leittechnik für Systeme mit sicherheitstechnischer Bedeutung – Allgemeine Systemanforderungen (IEC 61513:2011); Deutsche Fassung EN 61513:2013	2013-09	DKE/UK 967.1 „Elektro- und Leittechnik für kerntechnische Anlagen“		X							X
DIN SPEC 91426[505]	Qualitätsanforderungen für videogestützte Methoden der Personalauswahl (VMP)	2020		X		X	X					
DIN EN 50128; VDE 0831-128 [452]	Bahnanwendungen – Telekommunikationstechnik, Signaltechnik und Datenverarbeitungssysteme – Software für Eisenbahnsteuerungs- und Überwachungssysteme; Deutsche Fassung EN 50128:2011	2012-03	UK 351.3 „Bahn-Signalanlagen“		X				X			
IEEE 7010 [453]	A New Standard for Assessing the Well-being Implications of Artificial Intelligence	2020	SMC/SC – Standards Committee	X			X			X		

Dokument	Titel	Datum	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
IEEE 2801 [454]	Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence	2022	IEEE EMB/Stds Com – Standards Committee					X		X		
DIN ISO 31000 [160]	Risikomanagement – Leitlinien (ISO 31000:2018)	2018	NA 175-00-04 AA	X		X	X	X		X	X	X
ISO/SAE 21434 [324]	Road vehicles – Cybersecurity engineering	2021	NA 052-00-32 AA			X						
ISO-26262-Reihe [455]	Straßenfahrzeuge – Funktionale Sicherheit		NA 052-00-32 AA		X	X						
ISO/TR 4804 [325]	Road vehicles – Safety and cybersecurity for automated driving systems – Design, verification and validation methods	2020	NA 052-00-33-17 AK			X						
DIN EN 62061 [456]	Sicherheit von Maschinen – Funktionale Sicherheit sicherheitsbezogener elektrischer, elektronischer und programmierbarer elektronischer Steuerungssysteme	2016	DKE/K 225 „Elektrotechnische Ausrüstung und Sicherheit von Maschinen und maschinellen Anlagen“		X	X		X				X
DIN EN ISO 12100 [517]	Sicherheit von Maschinen – Allgemeine Gestaltungsleitsätze – Risikobeurteilung und Risikominderung (ISO 12100:2010); Deutsche Fassung EN ISO 12100:2010	2011	NA 095-01-01 GA		X		X	X				X
DIN CEN ISO/TR 22100-1 [457]	Sicherheit von Maschinen – Beziehung zu ISO 12100 – Teil 1: Wie ISO 12100 und Typ-B- und Typ-C-Normen zusammenhängen	2021	NA 095-01-01 GA		X			X				X
DIN ISO/TR 22100-2, DIN SPEC 33887 [458]	Sicherheit von Maschinen – Beziehung zu ISO 12100 – Teil 2: Wie ISO 12100 und ISO 13849-1:2021 zusammenhängen	2014	NA 095-01-01 GA		X			X				X
DIN ISO/TR 22100-3, DIN SPEC 33888 [459]	Sicherheit von Maschinen – Beziehung zu ISO 12100 – Teil 3: Implementierung ergonomischer Grundsätze in Sicherheitsnormen	2017	NA 095-01-01 GA		X		X	X				X

Dokument	Titel	Datum	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
DIN CEN ISO/TR 22100-4 [460]	Sicherheit von Maschinen – Zusammenhang mit ISO 12100 – Teil 4: Leitlinien für Maschinenherstellende zur Berücksichtigung der damit verbundenen IT-Sicherheits-(Cybersicherheits-)Aspekte	2020	NA 095-01-01 GA		X	X		X				X
ISO/TR 22100-5 [461]	Sicherheit von Maschinen – Beziehung zu ISO 12100 – Teil 5: Auswirkungen von maschinellem Lernen mit künstlicher Intelligenz	2021	NA 095-01-01 GA		X	X	X	X				X
DIN EN ISO 13849-1 [109]	Sicherheit von Maschinen – Sicherheitsbezogene Teile von Steuerungen – Teil 1: Allgemeine Gestaltungsleitsätze	2016	NA 095-01-03 GA		X	X		X				X
DIN EN ISO 13849-2 [462]	Sicherheit von Maschinen – Sicherheitsbezogene Teile von Steuerungen – Teil 2: Validierung	2013	NA 095-01-03 GA		X	X		X				X
ISO/IEC 25012 [463]	Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model	2008	NA 043-01-07 AA	X		X		X		X		X
ISO/IEC/IEEE 29119-1 [464]	Software and systems engineering – Software testing – Part 1: General concepts	2022	NA 043-01-07 AA	X		X		X		X		X
ISO/IEC/IEEE 29119-2 [465]	Software and systems engineering – Software testing – Part 2: Test processes	2021	NA 043-01-07 AA	X		X		X		X		X
ISO/IEC/IEEE 29119-3 [466]	Software and systems engineering – Software testing – Part 3: Test documentation	2021	NA 043-01-07 AA	X		X		X		X		X
ISO/IEC/IEEE 29119-4 [467]	Software and systems engineering – Software testing – Part 4: Test techniques	2021	NA 043-01-07 AA	X		X		X		X		X
ISO/IEC/IEEE 29119-5 [468]	Software and systems engineering – Software testing – Part 5: Keyword-Driven Testing	2016	NA 043-01-07 AA	X						X		X

Dokument	Titel	Datum	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
IEEE 1012 [469]	Standard for System, Software, and Hardware Verification and Validation	2016	IEEE C/S2ESC – Software & Systems Engineering Standards Committee	X		X		X		X		X
IEEE 3333.1.3 [470]	Standard for the Deep Learning-Based Assessment of Visual Experience Based on Human Factors	2022	IEEE C/SAB – Standards Activities Board				X					
ANSI/UL 4600 [471]	Standard for Safety for the Evaluation of Autonomous Products	2022	American National Standards Institute (ANSI)		X			X				X
ISO/IEC/IEEE 12207 [148]	Systems and software engineering – Software life cycle processes	2017	NA 043-01-07 AA	X				X		X		X
ISO/IEC 25000 [472]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Guide to SQuaRE	2014	NA 043-01-07 AA	X		X		X		X		X
ISO/IEC 25024 [473]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of data quality	2011	NA 043-01-07 AA	X		X		X		X		X
ISO/IEC 25020 [474]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality measurement framework	2019	NA 043-01-07 AA	X		X		X		X		X
ISO/IEC 25010 [152]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models	2011	NA 043-01-07 AA	X		X		X		X		X
ISO/IEC 25021 [475]	Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality measure elements	2012	NA 043-01-07 AA	X		X		X		X		X
DIN EN ISO 25119-1 [112]	Tractors and machinery for agriculture and forestry – Safety-related parts of control systems	2021	NA 060-16-12 AA		X			X				

Dokument	Titel	Datum	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
DIN SPEC 2343 [476]	Transmission of language-based data between artificial intelligences – Specification of parameters and formats	2020		X								X
ISO/TS 17033 [477]	Ethische Behauptungen und unterstützende Informationen – Grundsätze und Anforderungen	2019	NA 147-00-03 AA	X			X					
DIN EN ISO 26000 [478]	Leitfaden zur gesellschaftlichen Verantwortung	2021	NA 175-00-03 AA	X			X					
IEEE 7000 [64]	IEEE Standard Model Process for Addressing Ethical Concerns during System Design	2021	IEEE C/S2ESC – Software & Systems Engineering Standards Committee	X			X			X		
IEEE 7001 [10]	Standard for Transparency of Autonomous Systems	2021	IEEE VT/ITS – Intelligent Transportation Systems	X			X	X				X
IEEE 7002 [11]	Standard for Data Privacy Process	2022	IEEE C/S2ESC – Software & Systems Engineering Standards Committee	X			X	X		X	X	X
IEEE 7007 [12]	Ontological Standard for Ethically driven Robotics and Automation Systems	2021	IEEE RAS/SC – Standing Committee for Standards	X			X			X		
IEEE 7005 [13]	Transparent Employer Data Governance	2021	IEEE C/S2ESC – Software & Systems Engineering Standards Committee	X			X	X				
DIN EN ISO/IEC 27000 [479]	Informationstechnik – Sicherheitsverfahren – Informationssicherheitsmanagementsysteme – Überblick und Terminologie	2020	NA 043-04-27-01 AK				X	X		X		X
DIN EN ISO/IEC 27001 [480]	Informationstechnik – Sicherheitsverfahren – Informationssicherheitsmanagementsysteme – Anforderungen	2017	NA 043-04-27-01 AK	X		X		X		X		X
DIN EN ISO/IEC 27002 [481]	Informationssicherheit, Cybersicherheit und Schutz der Privatsphäre – Informationssicherheitsmaßnahmen (ISO/IEC 27002:2022); Deutsche und Englische Fassung prEN ISO/IEC 27002:2022	2017	NA 043-04-27-01 AK			X		X		X		X

Dokument	Titel	Datum	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
DIN EN ISO/IEC 27701 [128]	Sicherheitstechniken – Erweiterung zu ISO/IEC 27001 und ISO/IEC 27002 für das Management von Informationen zum Datenschutz – Anforderungen und Leitlinien	2021	NA 043-04-27-05 AK	X		X		X		X		X
DIN EN ISO/IEC 17000 [147]	Conformity assessment	2020	NA 147-00-03 AA	X		X	X	X		X		X
ITU-T Y.qos-ml-arc [482]	Architecture of machine learning based QoS assurance for the IMT-2020 network	2017	ITU-T SG 13 – Future networks	X								X
ETSI TS 103 195-2 [483]	Autonomic network engineering for the self-managing Future Internet (AFI); Generic Autonomic Network Architecture; Part 2: An Architectural Reference Model for Autonomic Networking, Cognitive Networking and Self-Management	2018	ETSI „Autonomic network engineering for the self-managing Future Internet (AFI)“									X
DIN EN ISO/IEC 17011 [159]	Konformitätsbewertung – Anforderungen an Akkreditierungsstellen, die Konformitätsbewertungsstellen akkreditieren	2018	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17020 [157]	Konformitätsbewertung – Anforderungen an den Betrieb verschiedener Typen von Stellen, die Inspektionen durchführen	2012	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17021-1 [22]	Konformitätsbewertung – Anforderungen an Stellen, die Managementsysteme auditieren und zertifizieren – Teil 1: Anforderungen	2015	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17021-2 [484]	Konformitätsbewertung – Anforderungen an Stellen, die Managementsysteme auditieren und zertifizieren – Teil 2: Anforderungen an die Kompetenz für die Auditierung und Zertifizierung von Umweltmanagementsystemen	2019	NA 147-00-03 AA	X		X						

Dokument	Titel	Datum	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
DIN EN ISO/IEC 17021-3 [485]	Konformitätsbewertung – Anforderungen an Stellen, die Managementsysteme auditieren und zertifizieren – Teil 3: Anforderungen an die Kompetenz für die Auditierung und Zertifizierung von Qualitätsmanagementsystemen	2019	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 7024 [155]	Konformitätsbewertung – Allgemeine Anforderungen an Stellen, die Personen zertifizieren	2012	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17025 [156]	Allgemeine Anforderungen an die Kompetenz von Prüf- und Kalibrierlaboratorien	2018	NA 147-00-03 AA	X		X		X				
DIN EN ISO/IEC 17029 [158]	Konformitätsbewertung – Allgemeine Grundsätze und Anforderungen an Validierungs- und Verifizierungsstellen	2020	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17030 [486]	Konformitätsbewertung – Allgemeine Anforderungen an Konformitätszeichen einer dritten Seite	2021	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17040 [487]	Konformitätsbewertung – Allgemeine Anforderungen an die Begutachtung unter gleichrangigen Konformitätsbewertungsstellen und Akkreditierungsstellen	2005	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17043 [488]	Konformitätsbewertung – Allgemeine Anforderungen an die Kompetenz von Anbietern von Eignungsprüfungen	2022	NA 147-00-03 AA	X		X						
DIN EN ISO/IEC 17050-1 [489]	Konformitätsbewertung – Konformitätserklärung von Anbietern – Teil 1: Allgemeine Anforderungen	2010	NA 147-00-03 AA	X		X		X				
DIN EN ISO/IEC 17050-2 [490]	Konformitätsbewertung – Konformitätserklärung von Anbietern – Teil 2: Unterstützende Dokumentation	2005	NA 147-00-03 AA	X		X		X				
DIN EN ISO/IEC 17065 [17]	Konformitätsbewertung – Anforderungen an Stellen, die Produkte, Prozesse und Dienstleistungen zertifizieren	2013	NA 147-00-03 AA	X		X						

Dokument	Titel	Datum	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
DIN EN ISO/IEC 17067 [18]	Konformitätsbewertung – Grundlagen der Produktzertifizierung und Leitlinien für Produktzertifizierungsprogramme	2013	NA 147-00-03 AA	X		X		X		X		
ITU-T F.AI-DLFE [491]	Deep Learning Software Framework Evaluation Methodology	2021	ITU-T SG 16 – Multimedia	X						X		X
ITU-T Y.3173 [492]	Framework for evaluating intelligence level of future networks including IMT-2020	2020	ITU-T SG 13 – Future networks									X
ISO/IEC 27034-1 [122]	Informationstechnik – IT Sicherheitsverfahren – Sicherheit von Anwendungen – Teil 1: Überblick und Konzept	2011	NA 043-04-27 AA			X		X				X
ISO/IEC 27034-2 [123]	Informationstechnik – IT Sicherheitsverfahren – Sicherheit von Anwendungen – Teil 2: Organisation des normativen Rahmen	2015	NA 043-04-27 AA			X		X				X
ISO/IEC 27034-3 [124]	Informationstechnik – Sicherheit von Anwendungen – Teil 3: Managementprozess für die Sicherheit von Anwendungen	2018	NA 043-04-27 AA			X		X				X
ISO/IEC 27034-5 [125]	Informationstechnik – IT Sicherheitsverfahren – Sicherheit von Anwendungen – Teil 5: Protokolle und Datenstruktur zur Kontrolle der Anwendungssicherheit	2017	NA 043-04-27 AA			X		X				X
ISO/IEC 27034-6 [126]	Informationstechnik – IT Sicherheitsverfahren – Sicherheit von Anwendungen – Teil 6: Fallstudien	2016	NA 043-04-27 AA			X						X
ISO/IEC 27034-7 [127]	Informationstechnik – IT Sicherheitsverfahren – Sicherheit von Anwendungen – Teil 7: Model zur Voraussage der Zusicherung von Sicherheitsanwendungen	2018	NA 043-04-27 AA			X		X				X
DIN EN ISO/IEC 29101 [493]	Informationstechnik – Sicherheitstechniken – Architekturrahmenwerk für Datenschutz	2022	NA 043-04-27 AA			X		X				X

Dokument	Titel	Datum	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
DIN EN ISO/IEC 29134 [134]	Informationstechnik – Sicherheitsverfahren – Leitlinien für die Datenschutz-Folgenabschätzung	2020	NA 043-04-27 AA			X		X				X
DIN EN ISO/IEC 29147 [494]	Informationstechnik – Sicherheitstechniken – Offenlegung von Schwachstellen	2020	NA 043-04-27 AA			X		X				X
DIN EN ISO/IEC 29151 [135]	Informationstechnik – Sicherheitsverfahren – Leitfaden für den Schutz personenbezogener Daten	2022	NA 043-04-13 GA			X		X				X
DIN EN ISO/IEC 29100 [133]	Informationstechnik – Sicherheitsverfahren – Rahmenwerk für Datenschutz	2020	NA 043-04-27 AA					X		X		X
ITU-T F.AI-DLPB [495]	Metrics and evaluation methods for deep neural network processor benchmark	2020	ITU-T SG 16 – Multimedia							X		X
ITU-T Y.3170 [496]	Requirements for machine learning – based quality of service assurance for the IMT-2020 Network	2018	ITU-T SG 13 – Future networks				X					X
ETSI DGR SAI 002 [497]	Securing Artificial Intelligence (SAI); Data Supply Chain Report	2021	ETSI „Securing Artificial Intelligence (SAI)“			X		X		X	X	X
ETSI DGS SAI 003 [336]	Securing Artificial Intelligence (SAI); Security Testing of AI	2022	ETSI „Securing Artificial Intelligence (SAI)“			X		X		X	X	X
ETSI TS 103 296 [498]	Speech and Multimedia Transmission Quality (STQ); Requirements for Emotion Detectors used for Telecommunication Measurement Applications; Detectors for written text and spoken speech	2016	ETSI „Speech and Multimedia Transmission Quality (STQ)“				X					
ETSI GR ENI 004 [499]	Experiential Networked Intelligence (ENI); Terminology for Main Concepts in ENI Disclaimer	2019	ETSI „Experiential Networked Intelligence (ENI)“							X		
ISO/TR 24291 [501]	Medizinische Informatik – Anwendungen von Technologien des maschinellen Lernens für die künstliche Intelligenz in der Medizin	2021	ISO TC 215					X		X		

Dokument	Titel	Datum	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
ISO/TR 3985 [502]	Biotechnologie – Datenveröffentlichung – Vorüberlegungen und Konzepte	2021	ISO TC 276							X		
ISO/TS 22756 [503]	Medizinische Informatik – Anforderungen an eine Wissensbasis für medizinische Entscheidungsunterstützungssysteme von medikationsbezogenen Prozessen	2020	ISO TC 215							X		
DIN SPEC 92001-1 [162]	Life Cycle Prozesse und Qualitätsanforderungen – Teil 1: Qualitäts-Meta-Modell	2019	DIN SPEC Konsortium	X			X	X		X	X	
DIN SPEC 92001-2 [240]	Life Cycle Prozesse und Qualitätsanforderungen – Teil 2: Robustheit	2020	DIN SPEC Konsortium	X		X	X	X		X		X
DIN SPEC 13288 [506]	Leitfaden für die Entwicklung von Deep-Learning-Bilderkennungssystemen in der Medizin; Text Deutsch und Englisch	2021	DIN SPEC Konsortium					X		X		
ISO/TS 5346 [507]	Medizinische Informatik – Kategoriale Struktur zur Darstellung des klinischen Entscheidungsunterstützungssystems der Traditionellen Chinesischen Medizin	2022	ISO/TC 215							X		
Serie DIN EN ISO 11073 [508]	Medizinische Informatik – Kommunikation von Geräten für die persönliche Gesundheit		ISO/TC 215							X		
DIN CEN ISO/TS 22703 [509]	Medizinische Informatik – Anforderungen an Arzneimittel-Warnmeldungen (ISO/TS 22703:2021); Deutsche Fassung CEN ISO/TS 22703:2021	2022	ISO/TC 215		X					X		
ISO/TR 19669 [510]	Medizinische Informatik – Wiederverwendbare Komponenten-Strategie für die Entwicklung von Use-Cases	2017	ISO/TC 215							X		
IEEE P2802 [511]	Standard for the Performance and Safety Evaluation of Artificial Intelligence Based Medical Device: Terminology	2022	IEEE AIMDWG – Artificial Intelligence Medical Device Working Group				X	X		X		X

Dokument	Titel	Datum	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
DIN EN ISO 13485 [381]	Medizinprodukte – Qualitätsmanagementsysteme – Anforderungen für regulatorische Zwecke (ISO 13485:2016); Deutsche Fassung EN ISO 13485:2016 + AC:2018 + A11:2021	2021	NA 063-01-13 AA							X		
DIN EN 62366-1 [355]	Medizinprodukte – Teil 1: Anwendung der Gebrauchstauglichkeit auf Medizinprodukte (IEC 62366-1:2015 + COR1:2016 + A1:2020); Deutsche Fassung EN 62366 1:2015 + AC:2015 + A1:2020	2021	UK 811.4		X					X		
DIN EN 82304-1 [354]	Gesundheitssoftware – Teil 1: Allgemeine Anforderungen für die Produktsicherheit	2018	DKE/UK 811.3		X					X		
DIN EN 60601-1-10 [375]	Medizinische elektrische Geräte – Teil 1-10: Allgemeine Festlegungen für die Sicherheit einschließlich der wesentlichen Leistungsmerkmale – Ergänzungsnorm: Anforderungen an die Entwicklung von physiologischen geschlossenen Regelkreisen (IEC 60601 1-10:2007 + A1:2013 + A2:2020); Deutsche Fassung EN 60601-1-10:2008 + A1:2015 + A2:2021	2021	DKE/K 811		X					X		
IEC/TR 60601-4-1 [373]	Guidance and interpretation – Medical electrical equipment and medical electrical systems employing a degree of autonomy	2017	TC 62/SC 62A							X		
ISO/TR 24971 [352]	Medical devices – Guidance on the application of ISO 14971:2022	2020	ISO/TC 210							X		
IEC/TR 62366-2 [357]	Medical devices – Part 2: Guidance on the application of usability engineering to medical devices	2021	ISO/TC 210							X		
DIN EN 62267, VDE 0831-267 [332]	Bahnanwendungen – Automatischer städtischer schienengebundener Personennahverkehr (AUGT) – Sicherheitsanforderungen (IEC 62267:2009); Deutsche Fassung EN 62267:2009	2010	DKE/UK 351.3		X				X			

Dokument	Titel	Datum	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
DIN VDE V 0831-103 [343]	Ermittlung von Sicherheitsanforderungen an technische Funktionen in der Eisenbahnsignaltechnik	2020	DIN und VDE		X				X			
DIN VDE V 0831-101 [344]	Semi-quantitative Verfahren zur Risikoanalyse technischer Funktionen in der Eisenbahnsignaltechnik	2022	DIN und VDE		X				X			
ISO 22737 [327]	Intelligent transport systems – Low-speed automated driving (LSAD) systems for predefined routes – Performance requirements, system requirements and performance test procedures	2021	ISO/TC 204		X				X			
VDE SPEC 90012 [242]	VCIO based description of systems for AI trustworthiness characterisation	2022					X			X		
VDI-MT 7001 [512]	Kommunikation und Öffentlichkeitsbeteiligung bei Bau- und Infrastrukturprojekten – Standards für die Leistungsphasen der Ingenieure	2021					X					
DIN EN ISO 26800 [239]	Ergonomie – Genereller Ansatz, Prinzipien und Konzepte	2011					X					
DIN EN ISO 6385 [235]	Grundsätze der Ergonomie für die Gestaltung von Arbeitssystemen	2016					X					
DIN EN ISO 10075 (alle Teile) [513]	Ergonomische Grundlagen bezüglich psychischer Arbeitsbelastung						X					
DIN EN ISO 11064 [243]	Ergonomische Gestaltung von Leitzentralen	2011					X					
DIN EN ISO 9241 (alle Teile) [514]	Ergonomie der Mensch-System-Interaktion						X					
DIN EN 614-1 [180]	Sicherheit von Maschinen – Ergonomische Gestaltungsgrundsätze – Teil 1: Begriffe und allgemeine Leitsätze; Deutsche Fassung EN 614-1:2006+A1:2009	2009			X		X					

Dokument	Titel	Datum	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
DIN EN 614-2 [181]	Sicherheit von Maschinen – Ergonomische Gestaltungsgrundsätze – Teil 2: Wechselwirkungen zwischen der Gestaltung von Maschinen und den Arbeitsaufgaben; Deutsche Fassung EN 614-2:2000+A1:2008	2008			X		X					
DIN EN 894 (alle Teile) [515]	Sicherheit von Maschinen – Ergonomische Anforderungen an die Gestaltung von Anzeigen und Stellteilen				X		X					
DIN EN 16710-2 [516]	Verfahren der Ergonomie – Teil 2: Eine Methode für die Arbeitsanalyse zur Unterstützung von Entwicklung und Design; Deutsche Fassung EN 16710-2:2016	2016					X					
ISO/TR 16982 [518]	Ergonomie der Mensch-System-Interaktion – Methoden zur Gewährleistung der Gebrauchstauglichkeit, die eine benutzer-orientierte Gestaltung unterstützen	2002					X					
DIN EN ISO 27500 [271]	Die menschzentrierte Organisation – Zweck und allgemeine Grundsätze (ISO 27500:2016); Deutsche Fassung EN ISO 27500:2017	2017					X					
VDI/VDE-MT 7100 [241]	Lernförderliche Arbeitsgestaltung – Ziele, Nutzen, Begriffe	2022					X					
DIN EN 15804 [413]	Nachhaltigkeit von Bauwerken – Umweltproduktdeklarationen – Grundregeln für die Produktkategorie Bauprodukte; Deutsche Fassung EN 15804:2012+A2:2019 + AC:2021	2022	NA 005-01-31 AA									X
DIN EN ISO 14044 [412]	Umweltmanagement – Ökobilanz – Anforderungen und Anleitungen (ISO 14044:2006 + Amd 1:2017 + Amd 2:2020); Deutsche Fassung EN ISO 14044:2006 + A1:2018 + A2:2020	2021	NA 172-00-03-AA									X

Dokument	Titel	Datum	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
DIN EN ISO 14040 [411]	Umweltmanagement – Ökobilanz – Grundsätze und Rahmenbedingungen (ISO 14040:2006 + Amd 1:2020); Deutsche Fassung EN ISO 14040:2006 + A1:2020	2021	NA 172-00-03-AA									X
DIN EN ISO 14026 [410]	Umweltkennzeichnungen und -deklarationen – Grundsätze, Anforderungen und Richtlinien für die Kommunikation von Fußabdruckinformationen (ISO 14026:2017); Deutsche und Englische Fassung EN ISO 14026:2018	2018	NA 172-00-03-AA									X
ISO 21930 [519]	Nachhaltigkeit von Bauwerken – Grundregeln für die Umweltdeklaration von in Bauwerken verwendeten Bauprodukten und technischen Anlagen	2017	ISO/TC 59, Building and civil engineering works, Subcommittee SC 17, Sustainability in buildings and civil engineering works									X
ISO/TS 14048 [521]	Umweltmanagement – Ökobilanz – Datendokumentationsformat	2002	ISO/TC 207, Environmental management, Subcommittee SC 5, Life cycle assessment									X
CWA 17284 [522]	Materials modelling – Terminology, classification and metadata	2018	CEN/CENELEC, WS									X
CWA 17815 [523]	Materials characterisation – Terminology, metadata and classification	2021	CEN/CENELEC, WS									X
DIN IEC/TS 62998-1: 2021-10, VDE V 0113-998-1 [520]	Sicherheit von Maschinen – Sicherheitsrelevante Sensoren für den Schutz von Personen (IEC TS 62998-1:2019)	2021	IEC/TC 44 Sicherheit von Maschinen – Elektrotechnische Aspekte		X							
ISO/IEC 22989 [16]	Artificial intelligence – Concepts and terminology	2022-07	NA 043-01-42 GA	X	X	X	X	X	X	X		X
ISO/IEC 23894 [25]	Information Technology – Artificial Intelligence – Risk Management	2022	NA 043-01-42 GA	X		X	X	X	X	X	X	X
ISO/IEC 19763-3 [426]	Informationstechnik – Metamodell-Rahmenwerk für die Interoperabilität (MFI) – Teil 3: Metamodell für die Registrierung von Ontologien	2020	NA 043-01-32 AA	X								

7.2 Laufende Normungs- und Standardisierungsaktivitäten mit Relevanz für KI

Tabelle 14 gibt eine Auswahl an laufenden Aktivitäten zum Thema KI wieder. Hierbei erhebt weder die Tabelle noch die Zuordnung Anspruch auf Vollständigkeit.

Tabelle 14: Überblick über laufende Normungs- und Standardisierungsaktivitäten zu KI

Dokument	Titel	Kurzbeschreibung	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
IEEE P2846	A Formal Model for Safety Considerations in Automated Vehicle Decision Making	Technologieneutrales mathematisches Modell und Prüfverfahren für automatische Entscheidungsfindung in Fahrzeugen	IEEE VT/ITS – Intelligent Transportation Systems		X			X	X			
ISO/IEC 5259-2	Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 2: Data quality measures		NA 043-01-42 GA	X		X		X		X	X	X
ISO/IEC 5259-5	Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 5: Data quality governance	This document provides a data quality governance framework for analytics and machine learning to enable governing bodies of organizations to direct and oversee the implementation and operation of data quality measures, management, and related processes with adequate controls throughout the data life cycle. This document can be applied to any analytics and machine learning. This document does not define specific management requirements or process requirements specified in 5259-3 and 5259-4 respectively.	NA 043-01-42 GA	X		X	X	X		X	X	X

Dokument	Titel	Kurzbeschreibung	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
ISO/IEC TR 5469	Artificial intelligence – Functional safety and AI systems	Das Dokument soll Eigenschaften, relevante Risikofaktoren, verwendbare Methoden und Prozesse für die Anwendung von KI in sicherheitsrelevanten Funktionen zur Kontrolle von KI-Systemen und für die Anwendung von KI in der Entwicklung sicherheitsrelevanter Funktionen beschreiben. Es wird in Zusammenarbeit mit IEC SC65A (der Standardisierungsgruppe, die für IEC 61508 verantwortlich ist) erstellt werden.	NA 043-01-42 GA		X	X	X	X	X	X	X	X
ISO/IEC TS 5471	Artificial intelligence – Quality evaluation guidelines for AI systems		NA 043-01-42 GA	X		X	X	X		X		X
ISO/IEC 24029-2	Artificial intelligence (AI) – Assessment of the robustness of neural networks – Part 2: Methodology for the use of formal methods		NA 043-01-42 GA	X		X		X	X	X	X	X
ISO/IEC TR 24029-1	Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview	Betrachtet die Robustheit von KI-Systemen und bietet einen Überblick über die verfügbaren Ansätze und Methoden zur Bewertung von Problemen und Risiken im Zusammenhang mit Robustheit. Ein besonderer Schwerpunkt liegt auf neuronalen Netzen, deren Funktionsweise und Verwendbarkeit.	NA 043-01-42 GA	X		X		X	X	X	X	X
ISO/IEC 5259-1	Data quality for analytics and ML – Part 1: Overview, terminology, and examples	Datenqualitätsmanagement für maschinelles Lernen: Überblick, Terminologie und Beispiele	NA 043-01-42 GA	X		X	X	X	X	X	X	X
ISO/IEC 5259-3	Data quality for analytics and ML – Part 3: Data Quality Management Requirements and Guidelines	Datenqualitätsmanagement für maschinelles Lernen: Anforderungen und Richtlinien	NA 043-01-42 GA	X		X	X	X	X	X	X	X

Dokument	Titel	Kurzbeschreibung	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
ISO/IEC 5259-4	Data quality for analytics and ML – Part 4: Data quality process framework	Datenqualitätsmanagement für maschinelles Lernen: Prozesse	NA 043-01-42 GA	X		X	X	X	X	X	X	X
ISO/IEC TS 8200	Information technology – Artificial intelligence – Controllability of automated artificial intelligence systems	This document defines a basic framework with principles, characteristics and approaches for the realization and enhancement for automated artificial intelligence (AI) systems controllability. The following areas are covered: – State observability and state transition – Control transfer process and cost – Reaction to uncertainty during control transfer – Verification and validation approaches This document is applicable to all types of organizations (e. g. commercial enterprises, government agencies, not-for-profit organizations) developing and using AI systems during their whole life cycle.	NA 043-01-42 GA	X		X	X	X		X		X
ISO/IEC 8183	Information technology – Artificial intelligence – Data life cycle framework	This document provides an overarching data life cycle framework that is instantiable for any AI system from data ideation to decommission. This document is applicable to the data processing throughout the AI system life cycle including the acquisition, creation, development, deployment, maintenance and decommissioning. This document does not define specific services, platforms or tools. This document is applicable to all organizations, regardless of type, sizes and nature, that use data in the development and use of AI systems.	NA 043-01-42 GA	X		X		X		X	X	X
ISO/IEC 42001	Information Technology – Artificial intelligence – Management system		NA 043-01-42 GA	X		X	X	X		X		X

Dokument	Titel	Kurzbeschreibung	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
ISO/IEC TS 6254	Information technology – Artificial intelligence – Objectives and approaches for explainability of ML models and AI systems	This document describes approaches and methods that can be used to achieve explainability objectives of stakeholders with regards to ML models and AI systems' behaviours, outputs, and results. Stakeholders include but are not limited to, academia, industry, policy makers, and end users. It provides guidance concerning the applicability of the described approaches and methods to the identified objectives throughout the AI system's life cycle, as defined in ISO/IEC 22989:2022.	NA 043-01-42 GA	X		X	X	X		X		X
ISO/IEC TR 29119-11	Information technology – Artificial intelligence – Testing for AI systems – Part 11:	This document describes testing techniques (including those described in ISO/IEC/IEEE 29119-4: 2021) applicable for AI systems in the context of the AI system life cycle model stages defined in ISO/IEC 22989:2022. It describes how AI and ML assessment metrics can be used in the context of those testing techniques. It also maps testing processes, including those described in ISO/IEC/IEEE 29119-2:2021, to the verification and validation stages in the AI system life cycle.	NA 043-01-42 GA	X		X	X	X		X		X
ISO/IEC 12792	Information technology – Artificial intelligence – Transparency taxonomy of AI systems	This document defines a taxonomy of information elements to assist AI stakeholders with identifying and addressing the needs for transparency of AI systems. The document describes the semantics of the information elements and their relevance to the various objectives of different AI stakeholders. This document uses a horizontal approach and is applicable to any kind of organization and application involving AI. V02/	NA 043-01-42 GA	X			X	X		X	X	X

Dokument	Titel	Kurzbeschreibung	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
ISO/IEC TS 12791	Information technology – Artificial intelligence – Treatment of unwanted bias in classification and regression machine learning tasks	This document provides mitigation techniques that can be applied throughout the AI system life cycle in order to treat unwanted bias. This document describes how to address unwanted bias in AI systems that use machine learning to conduct classification and regression tasks. This document is applicable to all types and sizes of organization.	NA 043-01-42 GA	X		X	X			X		X
ISO/IEC FDIS 24668	Information technology – Artificial intelligence –Process management framework for Big data analytics	Management für Datenanalysen im Bereich Big Data	NA 043-01-42 GA				X				X	X
ISO/IEC 5338	Information technology – Artificial intelligence – AI system life cycle processes	Terminologiestandard zu Lebenszyklusprozessen von KI-Systemen (in Abstimmung)	NA 043-01-42 GA	X		X		X		X	X	X
ISO/IEC TS 4213	Information technology – Artificial Intelligence – Assessment of machine learning classification performance	Metriken zur Leistungsfähigkeit von KI	NA 043-01-42 GA	X		X	X	X	X	X		X
ISO/IEC 5339	Information Technology – Artificial Intelligence – Guidelines for AI Applications	Richtlinien zur Anwendung von KI-Systemen (in Abstimmung)	NA 043-01-42 GA	X			X	X		X		X
ISO/IEC 5394	Information Technology – Artificial intelligence – Management System	Managememententsystemstandard für KI	ISO/IEC JTC 1/SC 32			X	X	X		X		X
ISO/IEC 5392	Information technology – Artificial intelligence – Reference Architecture of Knowledge Engineering	Referenzarchitektur für wissensbasierte Systeme	NA 043-01-42 GA	X		X	X	X		X	X	X

Dokument	Titel	Kurzbeschreibung	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
ISO/IEC TS 24462	Ontology for ICT Trustworthiness Assessment	Neues Projekt für eine Technische Spezifikation. Bearbeitet in ISO/IEC JTC 1/WG 13 „Trustworthiness“.	ISO/IEC JTC 1/SC 27	X			X	X		X		X
ISO 24089	Road vehicles – Software update engineering	neuer Standard – in Bearbeitung	ISO/TC 22/SC 32						X			
ISO/IEC 25059	Software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) Quality Model for AI-based systems	Feststellung von Qualität für Systeme auf KI-Basis	NA 043-01-42 GA	X		X	X	X	X	X		X
IEEE P7003	Algorithmic Bias Considerations		IEEE C/S2ESC – Software & Systems Engineering Standards Committee	X			X			X		X
IEEE P7006	Standard on Personal Data AI Agent Working Group			X			X					
IEEE P7008	Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems		IEEE RAS/SC – Standing Committee for Standards	X			X	X		X		X
IEEE P7009	Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems		IEEE RAS/SC – Standing Committee for Standards	X			X	X				X
IEEE P7011	Standard for the Process of Identifying & Rating the Trustworthiness of News Sources		IEEE SSIT/SC – Social Implications of Technology Standards Committee	X			X	X				X
IEEE P7012	Standard for Machine Readable Personal Privacy Terms		IEEE SSIT/SC – Social Implications of Technology Standards Committee	X			X			X		X

Dokument	Titel	Kurzbeschreibung	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
IEEE P7014	Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems		IEEE SSIT/SC – Social Implications of Technology Standards Committee	X			X			X		
NISTIR 8269	A Taxonomy and Terminology of Adversarial Machine Learning	Die Taxonomie ordnet verschiedene Arten von Angriffen, Verteidigungen und Konsequenzen. Die Terminologie definiert Schlüsselbegriffe im Zusammenhang mit der Sicherheit von ML in KI-Systemen		X						X	X	X
ISO/IEC 27005	Informationssicherheit, Cybersicherheit und Datenschutz – Leitfaden zur Handhabung von Informationssicherheitsrisiken		NA 043-04-27-01 AK				X	X		X	X	X
ETSI DTR INT 008 (TR 103 821)	Autonomic network engineering for the self-managing Future Internet (AFI); Artificial Intelligence (AI) in Test Systems and Testing AI models.	Testframework für Systeme der Netzwerkautomatisierung wie z. B. ETSI GANA (Generic Autonomic Networking Architecture)										X
ISO/IEC TR 17866	Artificial intelligence – Best practice guidance for mitigating ethical and societal concerns		NA 043-01-42 GA	X						X		
ISO/IEC 42005	Information technology – Artificial intelligence – AI system impact assessment		NA 043-01-42 GA	X				X		X		
ISO/IEC NP TS 17847	Information technology – Artificial intelligence – Verification and validation analysis of AI systems		NA 043-01-42 GA	X				X		X		

Dokument	Titel	Kurzbeschreibung	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
ISO/IEC TR 17903	Information technology – Artificial intelligence – Overview of machine learning computing devices		NA 043-01-42 GA					X		X		
ISO TS 23543	Guidance for developing cybersecurity requirements in anaesthetic and respiratory equipment standards	This document is intended to provide guidance for the application of cybersecurity in safety standards for anaesthetic and respiratory equipment. It is intended to assist each committee in identifying, assessing, and addressing cybersecurity risks, and in the preparation of corresponding requirements in an appropriate and consistent way. This document is applicable to particular device standards for anaesthetic and respiratory equipment with external (accessible) data interfaces (Signal Input/Output Part (SIP/SOP)).	ISO TC 121							X		
DIN SPEC 92001-3	Life Cycle Prozesse und Qualitätsanforderungen – Teil 3: Erklärbarkeit	branchenunabhängiger Leitfaden zu geeigneten Ansätzen und Methoden für die Förderung von Erklärbarkeit im gesamten Lebenszyklus eines KI-Modells	DIN SPEC Konsortium	X				X		X		

Dokument	Titel	Kurzbeschreibung	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
ISO/TS 9491	Biotechnology – Recommendations and requirements for predictive computational models in personalized medicine research – Part 1: Guidelines for constructing, verifying and validating models	This document defines challenges and requirements for predictive computational models constructed for research purposes in personalized medicine. It specifies recommendations and requirements for the setup, formatting, validation, simulation, storing and sharing of such models, as well as their application in clinical trials and other research areas. It summarizes specific challenges regarding data input, as well as verifying and validating of such models that can be considered as best practices for modelling in research and development in the field of personalized medicine. This document also specifies recommendations and requirements for data used to construct or needed for validating models, including rules and requirements for formatting, description, annotation, interoperability, integration, accessing, as well as recording and documenting the provenance of such data. This document does not provide specific rules or requirements for the use of computational models in the clinical routine, or for diagnostic or therapeutic purposes.	ISO/TC 276							X		

Dokument	Titel	Kurzbeschreibung	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
PT 63450	Artificial Intelligence-enabled Medical Devices – Methods for the Technical Verification and Validation	This document establishes methods for medical device manufacturers to verify and validate artificial intelligence / machine learning-enabled medical devices (AI/ML-MD), i. e. medical devices that use artificial intelligence, in part or in whole, to achieve their intended medical purpose. This includes verification and validation activities for the model of the artificial intelligence as well as selection, metrological characterization and management of the data sets. Such activities are implemented at various stages of the medical device lifecycle, especially including design control, monitoring and design change. This document is also applicable to any hardware or software utilizing artificial intelligence that impacts the intended use of a medical device.	IEC/TC 62						X	X		
ISO PAS 8800	Road vehicles – Safety and AI	Dieses Dokument definiert sicherheitsrelevante Eigenschaften und Risikofaktoren, die sich auf die unzureichende Leistung und das Fehlverhalten von Künstlicher Intelligenz (KI) in einem Straßenfahrzeugkontext auswirken. Es beschreibt einen Rahmen, der alle Phasen des Lebenszyklus von Entwicklung und Einsatz berücksichtigt. Dazu gehören die Ableitung geeigneter Sicherheitsanforderungen an die Funktion, die Betrachtung der Datenqualität und -vollständigkeit, architektonische Maßnahmen zur Kontrolle und Abschwächung von Fehlern, Werkzeuge zur Unterstützung der KI, Verifizierungs- und Validierungstechniken sowie die Nachweise, die erforderlich sind, um die Gesamtsicherheit des Systems zu gewährleisten.	ISO/TC 22/SC 32		X				X	X		

Dokument	Titel	Kurzbeschreibung	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
ISO/DIS 34501	Road vehicles – Terms and definitions of test scenarios for automated driving systems	Definiert grundlegende Begriffe zu Szenarien und szenarienbasiertem Testen	ISO/TC 22/SC 33						X	X		
ISO TS 5083	Road vehicles – Safety for automated driving systems – Design, verification and validation	This document provides an overview and guidance of the steps for developing and validating an automated vehicle equipped with a safe automated driving system. The approach is based on top level safety goals and basic principles derived from worldwide applicable publications. It considers safety by design, verification and validation methods for automated driving focused on SAE level 3 and level 4 vehicles according to ISO/SAE PAS 22736. In addition, it outlines cybersecurity considerations throughout all described steps. The document is intended to be applied to road vehicles (incl. trucks and busses, i. e. road vehicles > 3,5 to) excluding motorcycles.	ISO/TC 22/SC 32		X					X		
DIN EN ISO 22057	Nachhaltigkeit von Bauwerken – Datenvorlagen für die Verwendung von EPDs für Bauprodukte in BIM	Formale Integration von Bauproduktdaten in BIM-Prozesse; KI-Bezug als Basis für Datenframework für ML-/KI-Modelle	NA 005-01-31-AA									X

Dokument	Titel	Kurzbeschreibung	Gremium	Relevanz für Schwerpunktthemen								
				Grundlagen	Sicherheit	Prüfung und Zertifizierung	Soziotechnische Systeme	Industrielle Automation	Mobilität	Medizin	Finanzdienstleistungen	Energie und Umwelt
DIN/TS 92004	Künstliche Intelligenz – Qualitätsanforderungen und -prozesse – Risikoschema für KI-Systeme im gesamten Lebenszyklus	Dieses Dokument enthält ein KI-Risikoschema, das die Risiken entlang des gesamten Lebenszyklus von Systemen der Künstlichen Intelligenz (KI) abdeckt, die Komponenten des Maschinellen Lernens (ML) enthalten. Das Schema unterscheidet zwischen acht KI-Risikokategorien, d. h. Zuverlässigkeit, Fairness, Autonomie und Kontrolle, Transparenz, Erklärbarkeit, Sicherheit in den Bereichen Safety und Security sowie Datenschutz und ordnet jeder Kategorie entsprechende Risikoursachen zu. Dieses Dokument soll für alle Entwickler, Anbieter und Betreiber von KI-Systemen anwendbar sein. Es soll als Grundlage für die Identifikation der KI-Risiken dienen, die in einem bestimmten KI-System vorhanden sind, sowie für deren Analyse, und damit die Teile des Risikomanagementprozesses innerhalb einer Organisation informieren, die für die Identifikation und Analyse von Risiken zuständig sind.	NA 043-01-42-01 AK	X	X	X	X	X	X	X	X	X
ISO/IEC PWI 7699	Guidance for addressing security threats and failures in artificial intelligence		ISO/IEC JTC 1/SC 27 NA 043-04-27 AA	X	X	X	X	X	X	X	X	X

7.3 Gremien zu KI

Tabelle 15 gibt einen Überblick über relevante Normungs- und Standardisierungsgremien im Kontext KI.

Tabelle 15: Überblick über wichtige KI-Normungs- und Standardisierungsgremien[117]

	Gremium	Spiegelgremium[118]
International	IEC/SyC AAL „System Komitee AAL“	DKE/K 801
	IEC/TC 9 „Electrical equipment and systems for railways“	DKE/UK 351.3
	IEC/TC 44 „Safety of machinery – Electrotechnical aspects“	DKE/K 225
	IEC/SC 45A „Instrumentation, control and electrical power systems of nuclear facilities“	DKE/UK 967.1
	IEC/TC 62 „ Medical equipment, software, and systems“	DKE/K 810
	IEC/TC 62/SC 62A „Common aspects of medical equipment, software, and systems“	DKE/UK 811.4
	IEC/TC 65/WG 10 „Security for industrial process measurement and control – Network and system security“	DKE/UK 931.1
	IEC/TC 65/SC 65A „System aspects“	DKE/GK 914
	ISO/CASCO „Committee on conformity assessment“	NA 147-00-03 AA
	ISO/IEC JTC 1/SC 7 „Software and systems engineering“	NA 043-01-07 AA
	ISO/IEC JTC 1/SC 27 „Information security, cybersecurity and privacy protection“	NA 043-04-27 AA
	ISO/IEC JTC 1/SC 32 „Data management and interchange“	NA 043-01-32 AA
	ISO/IEC JTC 1/SC 38 „Cloud computing and distributed platforms“	NA 043-01-38 AA
	ISO/IEC JTC 1/SC 41 „Internet of things and digital twin“	NA 043-01-41 AA
	ISO/IEC JTC 1/SC42 „Artificial Intelligence“	NA 043-01-42 GA

117 Die Tabelle erhebt keinen Anspruch auf Vollständigkeit.

118 NA 005 DIN-Normenausschuss Bauwesen (NABau)
NA 023 DIN-Normenausschuss Ergonomie (NAErg)
NA 043 DIN-Normenausschuss Informationstechnik und Anwendungen (NIA)
NA 052 DIN-Normenausschuss Automobiltechnik (NAAutomobil)
NA 053 DIN-Normenausschuss Rettungsdienst und Krankenhaus (NARK)
NA 060 DIN-Normenausschuss Maschinenbau (NAM)
NA 063 DIN-Normenausschusses Medizin (NAMed)
NA 095 DIN-Normenausschusses Sicherheitstechnische Grundsätze (NASG)
NA 105 DIN-Normenausschuss Terminologie (NAT)
NA 147 DIN-Normenausschusses Qualitätsmanagement, Statistik und Zertifizierungsgrundlagen (NQSZ)
NA 172 DIN-Normenausschuss Grundlagen des Umweltschutzes (NAGUS)
NA 175 DIN-Normenausschuss Organisationsprozesse (NAOrg)

	Gremium	Spiegelgremium
	ISO/TC 22/SC 32 „Electrical and electronic components and general system aspects“	NA 052-00-32 AA
	ISO/TC 22/SC 33 „Vehicle dynamics and chassis components“	NA 052-00-33 AA
	ISO/TC 23/SC 19 „Agricultural electronics“	NA 060-16-12 AA
	ISO/TC 37/SC 4 „Language resource management“	NA 105-00-06 AA
	ISO/TC 59/SC 17 „Sustainability in buildings and civil engineering works“	NA 005-01-31 AA
	ISO/TC 68 „Financial services“	NA 043-03-02 AA
	ISO/TC 121/SC1 „Breathing attachments and anaesthetic machines“	NA 053-03-01 AA
	ISO/TC 159/SC 1 „General ergonomics principles“	NA 023-00-01 GA
	ISO/TC 159/SC 3 „Anthropometry and biomechanics“	NA 023-00-03 GA
	ISO/TC 159/SC 4 „Ergonomics of human-system interaction“	NA 023-00-04 GA
	ISO/TC 163 „Thermal performance and energy use in the built environment“	NA 005-12-01 GA
	ISO/TC 176/SC 3 „Supporting technologies“	NA 147-00-01 AA
	ISO/TC 199 „Safety of machinery	NA 095 BR
	ISO/TC 204 „Intelligent transport systems“	NA 052-00-71 GA
	ISO/TC 207/SC 5 „Life cycle assessment“	NA 172-00-03 AA
	ISO/TC 210 „Quality management and corresponding general aspects for medical devices“	NA 063-01-13 AA
	ISO TC 215 „Health informatics“	NA 063-07-01 AA
	ISO/TC 262 „Risk management“	NA 175-00-04 AA
	ISO TC 276 „Biotechnology“	NA 063-09-02
	ISO/TC 299 „Robotics“	NA 060-38-01 AA
	ITU-T SG 13 „Future networks“	
	ITU-T SG 16 „Multimedia“	
Europäisch	CEN/CLC/JTC 13 „Cybersecurity and data protection“	NA 043-04-13 GA
	CEN/CLC/JTC 21 „Artificial Intelligence“	NA 043-01-42 GA
	CEN/TC 114 „Safety of machinery“	NA 095 BR
	CEN/TC 251 „Health Informatics“	NA 063-07-01 AA
	CLC/TC 62 „Electrical equipment in medical practice“	DKE/K 801

	Gremium	Spiegelgremium
	ETSI „Methods for Testing and Specification (MTS)“	
	ETSI „Securing Artificial Intelligence (SAI)“	
	ETSI „Speech and Multimedia Transmission Quality (STQ)“	
	ETSI „Experiential Networked Intelligence (ENI)“	
National	NA 159-07-01 AA „Finanzdienstleistungen für den Privathaushalt“	
	NA 175-00-03 AA „Gesellschaftliche Verantwortung von Organisationen“	
	DKE/K 811 „Allgemeine Bestimmungen für elektrische Einrichtungen in medizinischer Anwendung“	
	DKE/UK 931.1 „IT-Sicherheit in der Automatisierungstechnik“	
	DIN SPEC 2343 „Übertragung von sprachbasierten Daten zwischen Künstlichen Intelligenzen – Festlegung von Parametern und Formaten“	
	DIN SPEC 13266 „Leitfaden für die Entwicklung von Deep-Learning-Bilderkennungs-systemen“	
	DIN SPEC 92001 „Künstliche Intelligenz – Life Cycle Prozesse und Qualitäts-anforderungen“	
	DIN SPEC 91426 „Qualitätsanforderungen für video-basierte Methoden der Personalauswahl“	
	DIN SPEC 92001-3 „Life Cycle Prozesse und Qualitätsanforderungen – Teil 3: Erklärbarkeit“	
Konsortien	IEEE AIMDWG „Artificial Intelligence Medical Device Working Group“	
	IEEE C/S2ESC – Software & Systems Engineering Standards Committee	
	IEEE C/SAB – Standards Activities Board	
	IEEE EMB/Stds Com – Standards Committee	
	IEEE RAS/SC – Standing Committee for Standards	
	IEEE SMC/SC – Standards Committee	
	IEEE SSIT/SC – Social Implications of Technology Standards Committee	
	IEEE VT/ITS „Intelligent Transportation Systems“	

8 Abkürzungsverzeichnis

Abkürzung	Ausschreibung
AAS	Asset Administration Shell
ADM	Algorithmic Decision Making
AG	Arbeitsgruppe
AGV	Automated Guided Vehicles
AI	Artificial Intelligence
AI Act	Artificial Intelligence Act
AIM	AI Machine
AIMS	AI Management System
ALKS	Automated Lane Keeping System
API	Application Programming Interface
AR	Augmented Reality
ArbMedVV	Arbeitsmedizinische Vorsorge Verordnung
ArbStättV	Arbeitsstättenverordnung
ASR	Automatic Speech Recognition
ATDD	Acceptance-Test-Driven Development
AUC	Area under the Curve
AUGT	Automatischer städtischer schienen-gebundener Personennahverkehr
AV	Aerial Vehicles
AVP	Valet-Parking-Systeme
B2B	Business to Business
BaFin	Bundesanstalt für Finanzdienstleistungs-aufsicht
BAIT	Bankaufsichtliche Anforderungen an die IT
BetrSichV	Betriebssicherheitsverordnung
BIM	Building Information Modelling
BMAS	Bundesministerium für Arbeit und Soziales
BMBF	Bundesministerium für Bildung und Forschung

Abkürzung	Ausschreibung
BMUV	Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz
BMVI	Bundesministerium für Verkehr und digitale Infrastruktur
BMWK	Bundesministerium für Wirtschaft und Klimaschutz
BSI	Bundesamt für Sicherheit in der Informations-technik
CC	Common Criteria
CC-King	Competence Center KI-Engineering
CCAM	Cooperative, Connected und Automated Mobility
CCRA	Common Criteria Recognition Arrangement
CNN	Convolutional Neural Networks
COLREG	Convention on the international regulations for preventing collisions at sea
CPU	Central processing units
CRR	Capital Requirements Regulation
CSM-RA	Common Safety Method for Risk Evaluation and Assessment
D&A	Detect and Avoid
DER	Distributed Energy Resources
DGUV	Deutsche Gesetzliche Unfallversicherung
DICOM	Digital Imaging and Communications in Medicine
DKE	Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE
DL	Deep Learning
DPP	Digitaler Produktpass
DSGVO	Datenschutz-Grundverordnung

Abkürzung	Ausschreibung
DSO	Distribution System Operator (Verteilnetzbetreiber)
EAD	Ethically Aligned Design
EAL	Evaluation Assurance Levels
EBA	Europäische Bankenaufsichtsbehörde
EHDS	European Health Data Space
EHF	Ergonomics/Human Factors
ELGI	Ethische Leitlinien der Gesellschaft für Informatik e. V.
ENISA	European Union Agency for Cybersecurity
EOSC	European Open Science Cloud
ESG	Environmental Social Governance
EU	Europäische Union
EV	Electric Vehicle
F&E	Forschungs- und Entwicklungsarbeiten
FAIR	Findable, Accessible, Interoperable, Reusable
FDA	Food and Drug Administration
GAN	Generative Adversial Networks
GCP	Good Clinical Practice
GefStoffV	Gefahrstoffverordnung
GIS	Geografische Informationssysteme
GoA	Grade of Automation
GPU	Graphics processing units
HAS	Harmonised Standards
hEN	Harmonisierte Europäische Norm
HIC	Human in Command
HITL	Human-in-the-Loop
HLEG	High Level Expert Group

Abkürzung	Ausschreibung
HMI	Mensch-Maschine-Schnittstelle
HOTL	Human-on-the-Loop
HR	Human Resources
IACS	Industrial Automation and Control Systems
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IG-NB	Interessensgemeinschaft der Benannten Stellen
IKT	Informations- und Kommunikations-technologien
IML4E	Industrial Grade Machine Learning for Enterprises
IMO	International Maritime Organization
IoU	Intercsection over Union
IRB(A)	Internal ratings-based (approach)
ISMS	Information Security Management System
IVD	In-vitro-Diagnostikum
IVDR	In-vitro-Diagnostic Medical Devices Regulation
KAIT	Kapitalverwaltungsaufsichtliche Anforde-rungen an die IT
KAMaRisk	Mindestanforderungen an das Risiko-management von Kapitalverwaltungs-gesellschaften
KI	Künstliche Intelligenz
KPIs	Key Performance Indicators
KRITIS	Kritische Infrastrukturen
KTIs	Key Trustwothiness Indicators
LCA	Life Cycle Assessment
LoD	Level of Detail

Abkürzung	Ausschreibung
LROD	Long Range Obstacle Detection
LSA	Lichtsignalanlagensteuerung
MaRisk	Mindestanforderungen für das Risiko-management für deutsche Kreditinstitute
MDR	Medical Device Regulation
ML	Maschinelles Lernen/Lernverfahren
MRT	Magnetresonanztomografie
MSS	Managementsystemstandards
MTO	Mensch, Technik und Organisation
NFDI	Nationale Forschungsdateninfrastruktur
NRM KI	Normungsroadmap Künstliche Intelligenz
ODD	Operational Design Domain
OEM	Original Equipment Manufacturer
OGC	Open Geospatial Consortium
OMG	Object Management Group
OWL	Web Ontology Language
PACS	Picture Archiving and Communication System
PMS	Power Management System
POC	Probability of Classification
POD	Probability of Detection
QML	Quantum Machine Learning
RAM	Referenzarchitekturmodell
RAMI 4.0	Referenzarchitekturmodell Industrie 4.0
RDF	Resource Description Framework
SAE	Society of Automotive Engineers
SG	Smart Grid
SGAM	Smart Grid Architecture Model
SIF	System Interface
SM	Smart Manufacturing
SOTIF	Safety of the intended Function
TAI	Trusworthy Artificial Intellligence
Tf	Triebfahrzeugführenden
TRM	Trustworthiness Readiness Matrix
UAM	Urban Air Mobility
UML	Unified Modeling Language
VAIT	Versicherungsaufsichtliche Anforderungen an die IT
VNB	Verteilnetzbetreiber
VWS	Verwaltungsschale
W3C	World Wide Web Consortium
XAI	Explainable AI
ZAIT	Zahlungsdiensteaufsichtliche Anforderungen an die IT
ZFP	Zerstörungsfreie Prüfung

9 Glossar

Begriff (de/en)	Bedeutung und Verwendungen
adverser Angriff adversarial attack	Ein adverser Angriff ist der gezielte Versuch, mithilfe von adversen Eingaben (adversarial examples) Fehler zu verursachen. Insbesondere künstliche neuronale Netze gelten als besonders anfällig für diese Art von Angriffen. Erwähnt in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Agency agency	Agency ist ein „Inkraftsetzen“ („enactment“) politisch-ethischer Formen von Subjektivität. Der Fokus liegt auf der Prozessualität iterativer Praktiken („doing“). Agency bleibt nicht lediglich Menschen vorbehalten, sondern kann auch nicht-menschlichen Entitäten zugesprochen werden.
Agent agent	Im Kontext von KI versteht man unter einem Agenten ein entscheidendes und handelndes System, das mit seiner Umgebung und anderen Agenten interagieren kann. Erwähnt in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16].
Akkreditierung	Bestätigung durch eine dritte Seite, die formal darlegt, dass eine Konformitätsbewertungsstelle die Kompetenz, Unparteilichkeit sowie einheitliche Arbeitsweise besitzt, bestimmte Konformitätsbewertungstätigkeiten durchzuführen.
Akkreditierungsstelle	Befugte Stelle, die Akkreditierungen durchführt.
Aktualität currentness	Grad der zeitlichen Gültigkeit von Daten mit Relevanz für einen bestimmten Anwendungskontext.
allgemeine KI general AI	KI, die über das gesamte Spektrum der kognitiven KI-Fähigkeiten hinweg intelligentes Verhalten zeigt, das mit dem eines Menschen vergleichbar ist (Synonym: starke KI). Erwähnt in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Annotation label	Als Labels oder Annotationen werden im Maschinellen Lernen die Teile des Trainingsdatensatzes bezeichnet, die für Trainingszwecke die gewünschte ideale Ausgabe des Modells für eine entsprechende Eingabe angeben. Im weiteren Sinne werden auch die tatsächlichen Ausgaben eines Modells im Betrieb so bezeichnet. Erwähnt in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16].
Anpassungsfähigkeit adaptability	Fähigkeit eines Systems, auf Veränderungen in seiner Umgebung zu reagieren, um weiterhin sowohl funktionale als auch nicht-funktionale Anforderungen zu erfüllen. Erwähnt in ISO/IEC TR 29119-11 [132].
API (Anwendungsprogrammierschnittstelle) Application Programming Interface (API)	Eine Menge an Kommunikationsprotokollen, Code und Werkzeugen, die es einer Menge an Softwarekomponenten ermöglicht, entweder mit einem Menschen oder mit einer anderen Menge an Softwarekomponenten zu interagieren. Erwähnt in ETSI GR ENI 004 V2.2.1 [499].
Arbeitsorganisation work organisation	Interagierende Arbeitssysteme, deren Zusammenwirken ein bestimmtes Gesamtergebnis erzielt. Erwähnt in DIN EN ISO 6385:2016 [235].

Begriff (de/en)	Bedeutung und Verwendungen
Audit	Prozess zum Erlangen relevanter Informationen über einen Gegenstand der Konformitätsbewertung und zu deren objektiver Auswertung, um zu ermitteln, inwieweit die festgelegten Anforderungen erfüllt sind Anmerkung 1 zum Begriff: Beispiele für Gegenstände eines Audits sind Managementsysteme […]. Anmerkung 2 zum Begriff: Es werden nur Organisationen auditiert, keine Produkte oder Dienstleistungen.
autonomes System **autonomous system**	Ein System, das über längere Zeiträume ohne menschlichen Eingriff funktioniert. Erwähnt in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Autonomie **autonomy**	Autonomie ist die Abwesenheit der Fremdbestimmung. Bezogen auf Menschen bedeutet Autonomie den freien Willen und entspricht einem Grundprinzip der Digital-Ethik. Erwähnt in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Bayes'sches Netz **Bayesian network**	Ein Bayes'sches Netz ist ein gerichteter, zyklenfreier Graph. Während im Graph die Knoten Variablen mit Wertebereichen abbilden, stellen die Kanten bedingte Wahrscheinlichkeiten dar. Erwähnt in ISO/IEC 22989:2022 [16].
Beanspruchung (personenbezogen) **stress (person-related)**	Innere Reaktion einer Person auf Belastung, abhängig von derer individuellen Eigenschaften (z. B. Körpergröße, Alter, Fähigkeiten, Begabungen, Fertigkeiten usw.). Anmerkung 1: In DIN EN ISO 6385:2016 [235] ist „Beanspruchung" als „Arbeitsbeanspruchung" ausgewiesen. Anmerkung 2: Der Begriff „Beanspruchung" ist neutral. Deren Auswirkungen können positiv, neutral oder negativ sein. Erwähnt in DIN EN ISO 26800:2011 [239].
bestärkendes Lernen **reinforcement learning**	Einsatz von Software-Agenten zur Durchführung von Aktionen in einer Umgebung mit dem Ziel, eine kumulative Belohnung zu maximieren. Erwähnt in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Bestätigung	Erstellen einer Aussage auf der Grundlage einer Entscheidung, dass die Erfüllung festgelegter Anforderungen dargelegt wurde.
Bias **bias**	Allgemein: Die Abweichung von einem Referenzwert bzw. dem tatsächlichen Wert. Im Kontext von KI wird Bias oft als systematische Abweichung verstanden, die nicht der tatsächlichen oder gewünschten Verteilung entspricht. Bei KI-Anwendungen wird ein vorhandener Bias oft als ungerecht gegenüber einer bestimmten Person oder Gruppe angesehen. Ein Bias kann seine Ursache in Daten, einem Algorithmus selbst, soziokulturellen Einflüssen oder einer beliebigen Kombination der vorgenannten Ursachen haben. Vor diesem Hintergrund gehören auch kognitive Verzerrungen des Menschen zum Bias-Begriff. Erwähnt in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Big Data	→ Siehe Massendaten.
Building Information Modelling (BIM)	Arbeitsmethode für vernetzte Planung und Bau von Gebäuden mithilfe von informationsbasierten Modellen.

Begriff (de/en)	Bedeutung und Verwendungen
Closed-Box-Test **closed-box testing**	Ein Closed-Box-Test (auch Blackbox-Test) ist ein Testverfahren, bei dem einem Tester keine Interna (wie insbesondere das KI-Modell) des KI-Systems für Testszenarien zur Verfügung stehen. Dagegen stehen üblicherweise ausschließlich Eingaben an das KI-System mit den dazugehörigen Ausgaben vom KI-System zur Verfügung. Erwähnt in ISO/IEC TR 29119-11 [132].
computerbasiertes Sehen **computer vision**	KI-Fähigkeit einer funktionalen Einheit, visuelle Daten zu erfassen, zu verarbeiten und zu interpretieren. Computerbasiertes Sehen beinhaltet die Nutzung von Sensoren, um ein digitales Abbild einer visuellen Szene zu erstellen. Siehe Kapitel 4.1.2.6. Erwähnt in ISO/IEC 22989:2022 [16].
Computerlinguistik **computer linguistics**	Die Computerlinguistik untersucht, wie natürliche Sprache in Form von Text- oder Sprachdaten mithilfe des Computers algorithmisch verarbeitet werden kann. Sie ist Schnittstelle zwischen Sprachwissenschaft und Informatik.
Data Mining **data mining**	Computergestütztes Verfahren, bei dem mittels der Analyse quantitativer Daten aus verschiedenen Dimensionen Muster extrahiert, kategorisiert sowie potenzielle Beziehungen und Folgen erkannt werden. Erwähnt in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16].
Data Poisoning **data poisoning**	Die absichtliche und bösartige Manipulation von Trainings-, Validierung-, Test- oder Eingabedaten für KI-Modelle. Erwähnt in ISTQB – CTAI Syllabus v1.0 [137].
Datenqualität **data quality**	Grad, zu dem Charakteristiken von Daten explizit spezifizierte oder implizite Anforderungen für einen gegebenen Anwendungsfall erfüllen.
Datensatz **dataset**	Sammlung von Daten mit einem gemeinsamen Format und zielrelevantem Inhalt. Im Idealfall repräsentieren die so ausgewählten Daten den größeren Datensatz bzw. die angenommene reale Charakteristik. Anmerkung: Datensätze können zum Training, zur Validierung sowie zum Testen eines KI-Modells verwendet werden. Im Kontext des überwachten Maschinellen Lernens stellen Datensätze eine Grundlage für das Training des Lernalgorithmus dar. Beispiel 1: Mikroblogging-Beiträge vom Juni 2020, die mit den Hashtags #rugby und #football verknüpft sind. Beispiel 2: Makrofotos von Blumen mit der Größe 256x256 Pixel. Erwähnt in ISO/IEC 22989:2022 [16], ISTQB – CTAI Syllabus v1.0 [137].
Dialogsystem **chatbot**	Eine Anwendung aus der Computerlinguistik zur Führung einer textbasierten Konversation auf Text oder Synthese natürlicher Sprache. Erwähnt in ISTQB – CTAI Syllabus v1.0 [137].
Digitaler Zwilling **digital twin**	Virtuelle digitale Darstellung eines physischen Objekts oder Systems über seinen Lebenszyklus hinweg unter Verwendung von Echtzeitdaten. Das digitale Abbild kann als Grundlage für Nachvollziehbarkeit, Training und Inferenz von KI-Modellen einbezogen werden. Erwähnt in ETSI GR ENI 004 V2.2.1 [499].

Begriff (de/en)	Bedeutung und Verwendungen
Ergonomie **ergonomics**	Wissenschaftliche Disziplin, die sich mit dem Verständnis der Wechselwirkungen zwischen menschlichen und anderen Elementen eines Systems befasst. Des Weiteren auch ein Berufszweig, der Theorie, Grundsätze, Daten und Verfahren auf die Gestaltung von Arbeitssystemen anwendet mit dem Ziel, das Wohlbefinden des Menschen und die Leistung des Gesamtsystems zu optimieren. Anmerkung: Diese Definition stimmt mit der durch die International Ergonomics Association festgelegte Definition überein. Erwähnt in DIN EN ISO 26800:2011 [239].
erklärbare KI **explainable AI (XAI)**	Ein Forschungs- und Anwendungsbereich, der sich mit dem Verständnis der Faktoren befasst, die Ergebnisse von KI-Systeme beeinflussen. Erwähnt in ISTQB – CTAI Syllabus v1.0 [137].
Erklärbarkeit **explainability**	Angestrebte Eigenschaft eines KI-Systems, dass Faktoren, die zu einer automatisierten Entscheidung des Systems geführt haben, durch einen Menschen „verstanden“ werden können. Erwähnt in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Erklärung (i. S. d. Prüfung und Zertifizierung)	Bestätigung durch eine erste Seite (beispielsweise Selbsterklärung Herstellende).
Ethik **ethics**	Grundsätze, die das moralische Verhalten eines Menschen oder einer Maschine bestimmen (nach ETSI). Domänenübergreifend ist Ethik die wissenschaftliche Beschäftigung mit der Moral. Sie reflektiert und philosophiert über diverse Moralvorstellungen, sie analysiert und systematisiert, sie untersucht und hinterfragt ihre Begründungen und Prinzipien. Es gibt verschiedene Moralvorstellungen, Normensysteme, Prinzipien, Werte oder Dispositionen, die alle für sich den Anspruch erheben, die Grundlage richtigen Handelns zu sein. Erwähnt in ETSI GR ENI 004 V2.2.1 [499].
Expertensystem **expert system**	Häufig regelbasiertes System, das auf symbolischer Wissensverarbeitung beruht. Beispiel: Wenn-dann-Regeln. Anmerkung: Z. B. symbolische, formale Repräsentation von Wissen in KI-Systemen mit der Eigenschaft, mittels Schlussfolgerung auf Grundlage von Logik aus formalem Wissen neues Wissen herzuleiten. Erwähnt in ISO/IEC 22989:2022 [16], ISTQB – CTAI Syllabus v1.0 [137].
Fairness **fairness**	Beim Einsatz algorithmischer und soziotechnischer Systeme im weiteren und maschinell lernender Systeme im engeren Sinn beschreibt Fairness als ethisches Prinzip den reproduzierbaren Grad der Gleichbehandlung verschiedener Personen in allen Stufen des Lifecycles des Systems. Dieses Prinzip ist ebenfalls auf nicht-menschliche (z. B. Tiere, Umwelt, Natur) bzw. insgesamt auf natürliche Akteur*innen anwendbar.
falsch negativ **false negative (FN)**	Eine Modellvorhersage, bei der das Modell einer binären Klassifikation fälschlicherweise negativ vorhersagt, obwohl positiv richtig wäre. Erwähnt in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].

Begriff (de/en)	Bedeutung und Verwendungen
falsch positiv false positive (FP)	Eine Modellvorhersage, bei der das Modell einer binären Klassifikation fälschlicherweise positiv vorhersagt, obwohl negativ richtig wäre. Erwähnt in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
festgelegte Anforderung	Erfordernis oder Erwartung, das oder die niedergelegt ist.
Gebrauchstauglichkeit usability	Ausmaß, in dem ein System, ein Produkt oder eine Dienstleistung durch bestimmte Benutzer*innen in einem bestimmten Nutzungskontext genutzt werden kann, um bestimmte Ziele effektiv, effizient und zufriedenstellend zu erreichen. Anmerkung 1: Die „bestimmten“ Benutzer*innen, „bestimmten“ Ziele und der „bestimmte“ Nutzungskontext beziehen sich auf die jeweilige Kombination aus Benutzer*innen, Zielen und Nutzungskontext, denen eine Gebrauchstauglichkeit unterstellt wird. Anmerkung 2: Das Wort „Gebrauchstauglichkeit“ wird auch als Qualifizierungsmerkmal verwendet, um auf Gestaltungskenntnisse, -kompetenzen, -aktivitäten und -attribute zu verweisen, die zur Gebrauchstauglichkeit beitragen, wie Gebrauchstauglichkeits-Fachkenntnisse und -Fachleute, gebrauchstauglichkeitsorientierte Entwicklung, Verfahren und Evaluierung sowie Gebrauchstauglichkeitsheuristik. Erwähnt in DIN EN ISO 9241-210:2020 [183].
Genauigkeit (im Kontext Klassifikation) accuracy (in the context of classification)	Im Kontext von Klassifikation im Bereich KI stellt die Genauigkeit (accuracy) eine Metrik zur Messung der Qualität zumeist binärer Klassifikationen dar. Sie wird berechnet als Anteil der korrekten Klassifikationen an sämtlichen Klassifikationen. Erwähnt in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Gewicht weight	Als „Gewichte“ werden in einem weiten Sinne Parameter eines Modells bezeichnet, in der Regel Faktoren, die spezifische Einträge mehrdimensionaler Eingaben individuell skalieren („gewichten“). In künstlichen neuronalen Netzen beispielsweise skalieren Gewichte die Eingabewerte eines künstlichen Neurons. Im Maschinellen Lernen werden typischerweise die Gewichte eines Modells trainiert. Erwähnt in ISTQB – CTAI Syllabus v1.0 [137].
Glassbox-Test glassbox testing	Ein Glassbox-Test (auch Whitebox-Test) ist ein Testverfahren, bei dem einem Tester Interna (wie beispielsweise das KI-Modell) des KI-Systems zur Generierung der Testfälle zur Verfügung stehen. Erwähnt in ISO/IEC TR 29119-11 [132].
Graph graph	Ein mathematisches Modell, das Verbindungsstrukturen auf abstrakte Weise darstellt. Es besteht aus „Knoten“ sowie aus „Kanten“, die Verbindungen zwischen diesen Knoten darstellen. Sowohl Knoten als auch Kanten können je nach Anwendung Werte zugeordnet werden, beispielsweise Gewichte, Kosten oder Distanzen. Erwähnt in ETSI GR ENI 004 V2.2.1 [499].
Graphikprozessor graphical processing unit (GPU)	Eine anwendungsspezifische integrierte Schaltung mit optimierter Speicherausnutzung, um die Erzeugung von Bildern in einem Bildpuffer zu beschleunigen. Erwähnt in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Groundtruth ground truth	Informationen, die durch direkte Beobachtung und Messung gewonnen werden und von denen angenommen wird, dass sie real oder wahr sind. Erwähnt in ISO/IEC 22989:2022 [16], ISTQB – CTAI Syllabus v1.0 [137].
Grundwahrheit	→ Siehe Groundtruth.

Begriff (de/en)	Bedeutung und Verwendungen
Hyperparameter **hyperparameter**	Hyperparameter bezeichnen im Maschinellen Lernen meist alle Parameter, die nicht unmittelbar durch den Trainingsprozess definiert oder beeinflusst werden. Dazu zählen Modellparameter wie die Anzahl an Schichten eines neuronalen Netzes oder die Schrittweite des Trainingsprozesses, nicht jedoch etwa die gelernten Gewichte. Grundsätzlich können Hyperparameter nach algorithmisch und modellspezifisch differenziert werden. Algorithmische Hyperparameter beeinflussen die Performance des Lernalgorithmus; hingegen beeinflussen modellspezifische Hyperparameter das im Lernprozess verwendete mathematische oder statistische Modell. Erwähnt in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Informationssicherheit	→ Siehe Sicherheit (Security im Sinne von IT).
Inspektion	Untersuchung eines Gegenstands der Konformitätsbewertung und Ermittlung seiner Konformität mit detaillierten Anforderungen oder, auf der Grundlage einer sachverständigen Beurteilung, mit allgemeinen Anforderungen.
Internet der Dinge (IoT) **internet of things**	Das Internet der Dinge (IoT) vernetzt eine Vielzahl von vielfältigen (Edge-)Geräten (siehe auch IoT-Gerät) und zentralen Datenplattformen miteinander und verbindet so Systeme, Dienste, Menschen und Informationen aus der physischen und virtuellen Welt miteinander. Dies hat neben neuen Anwendungen und Dienstleistungen auch die Entwicklung neuer Geschäftsmodelle ermöglicht. Erwähnt in ISO/IEC 22989:2022 [16].
Interpretierbarkeit **interpretability**	Der Grad der Nachvollziehbarkeit der Funktionsweise einer zugrunde liegenden (KI-)Technologie. Erwähnt in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
KI-Fähigkeit **capability**	Fähigkeit wie „Wahrnehmen“, „Handeln“ oder „Kommunizieren“, die auf Grundlage von Methoden der Künstlichen Intelligenz umgesetzt wird. Siehe Kapitel 4.1.1.1.
KI-Komponente **AI component**	Eine Komponente, die Methoden der KI beinhaltet. Erwähnt in ISTQB – CTAI Syllabus v1.0 [137].
KI-Modul **AI module**	Softwaremodul, in welchem KI-Methoden implementiert sind. KI-Dienste als Bausteine in einer Kette von Lieferbeziehungen mit mehreren IT-Komponenten oder KI-Diensten (siehe Kapitel 4.3.2.1).
KI-System **AI system**	System, das Künstliche Intelligenz benutzt.
Klassifikation (Maschinelles Lernen) **classification (Machine Learning)**	Aufgabenstellung, mittels welcher die Ausgabeklasse für eine bestimmte Eingabe vorhergesagt wird. Erwähnt in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Klassifizierer **classifier**	Ein Verfahren/System, das zur Umsetzung einer Klassifikationsaufgabe dient. Erwähnt in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Kognition **cognition**	Das Verstehen von Daten und Informationen und die Erzeugung neuer Daten, Informationen sowie von neuem Wissen. Erwähnt in ETSI GR ENI 004 V2.2.1 [499].

Begriff (de/en)	Bedeutung und Verwendungen
Konformitätsbewertung	Darlegung, dass festgelegte Anforderungen erfüllt sind.
Konformitäts-bewertungsstelle	Stelle, die Konformitätsbewertungstätigkeiten durchführt, jedoch keine Akkreditierung.
kontinuierliches Lernen **continual learning**	Im Kontext von KI versteht man unter kontinuierlichem Lernen das Training eines KI-Systems, das parallel zu dessen Betrieb iterativ und inkrementell stattfindet. Erwähnt in ISO/IEC 22989:2022 [15], [16].
Kontrollierbarkeit **controllability**	Eigenschaft, mittels derer ein Mensch oder ein anderer externer Agent unmittelbar und unverzüglich in die laufende Funktion des Systems eingreifen kann. Erwähnt in ISO/IEC 22989:2022 [16].
Kritikalität **criticality**	Maß für potenzielle Gefahren, die vom Einsatz eines KI-Systems in einem spezifischen Anwendungskontext ausgehen können. Der Begriff wird oft in ähnlicher Weise wie Risiko verwendet, wobei Kritikalität stärker auf eine Bewertung des Gesamtsystems abzielt.
Künstliche Intelligenz (KI) **artificial intelligence (AI)**	Der Begriff wird in verschiedenen Disziplinen aus unterschiedlichen Perspektiven diskutiert. Aufgrund des KI-Effekts entwickelt sich der Begriff stetig weiter. Folgend sind drei Definitionen ausgewiesen: Definition 1: Fähigkeit eines technischen Systems, Wissen und Kompetenzen zu erwerben, zu verarbeiten und anzuwenden (ISO/IEC TR 29119-11 [132]). Definition 2: Ein computergestütztes System, das kognitiv arbeitet, um Informationen zu verstehen und Probleme zu lösen (ISO/IEC 22989:2022 [16]). Definition 3: Ku$^{\mu\tau}$nstliche Intelligenz bezeichnet eine Reihe von Technologien, [...] die im Hinblick auf eine Reihe von Zielen, die vom Menschen festgelegt werden, Ergebnisse wie Inhalte, Vorhersagen, Empfehlungen oder Entscheidungen hervorbringen können, die das Umfeld beeinflussen, mit dem sie interagieren (Europäisches KI-Gesetz im Entwurf, [4]). Erwähnt in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
künstliches neuronales Netz(werk) (KNN) **artificial neural network (ANN)**	KNN sind Netze aus künstlichen Neuronen und haben ein biologisches Vorbild. Angelehnt an die Biologie ist ein künstliches Neuron ein Objekt, welches auf einen oder mehrere Reize reagiert, je nachdem, wie stark es aktiviert bzw. der Reiz gewichtet ist. Ein KNN besteht grundsätzlich aus einer Eingangs- und einer Ausgangsschicht. Dazwischen liegen verborgene Schichten (Hidden Layers) oder Aktivitätsschichten. KNN müssen in der Regel immer trainiert werden, bevor sie Problemstellungen lösen können. Dabei gewichtet ein bestimmter Algorithmus bzw. das neuronale Netz die Verbindungen der Neuronen anhand von vorgegebenem Lernmaterial und Lernregeln, bis es ein bestimmtes Lernziel erreicht bzw. entwickelt hat. Erwähnt in ETSI GR ENI 004 V2.2.1 [499].
Label	→ Siehe Annotation.
Lebenszyklus **life cycle**	Zeitlicher Verlauf zur Charakterisierung eines Systems, Produkts, einer Dienstleistung, eines Projekts oder einer anderen vom Menschen geschaffenen Einheit von der Konzeption bis zur Stilllegung. Erwähnt in ISO/IEC 22989:2022 [16]

Begriff (de/en)	Bedeutung und Verwendungen
Lernalgorithmus **learning algorithm**	Ein Algorithmus, der ein ML-Modell auf Grundlage von Charakteristiken der Trainingsdatensätze erstellt. → Vgl. Lernendes System. Erwähnt in ISTQB – CTAI Syllabus v1.0 [137].
Lerndaten	→ Siehe Trainingsdaten.
Lernendes System **learning system**	Lernende Systeme sind Maschinen, Roboter und Softwaresysteme, die abstrakt beschriebene Aufgaben auf Basis von Daten, die ihnen als Lerngrundlage dienen, selbstständig erledigen, ohne dass jeder Schritt spezifisch vom Menschen programmiert wird. Um Aufgaben zu lösen, setzen sie von Lernalgorithmen trainierte Modelle ein. Mithilfe des Lernalgorithmus können viele Systeme im laufenden Betrieb weiterlernen (kontinuierliches Lernen): Sie verbessern die vorab trainierten Modelle und erweitern ihre Wissensbasis. → Vgl. Lernalgorithmus.
maschinelle Übersetzung **machine translation**	Automatische Übersetzung von gesprochener oder geschriebener natürlicher Sprache in eine andere Sprache durch ein KI-System. Erwähnt in ISO/IEC 22989:2022 [16].
Maschinelles Lernen (ML) **machine learning (ML)**	ML als Teilgebiet der KI und Oberbegriff für die „künstliche“ Generierung von Wissen setzt computergestützte Techniken ein, um Systeme in die Lage zu versetzen, aus Daten oder Erfahrungen zu lernen. Ein solches System kann das erworbene Wissen nach Beendigung der Lernphase verallgemeinern, indem es aus den Lerndaten Muster und Gesetzmäßigkeiten erkennt und diese auf unbekannte Daten überträgt (Lerntransfer). Erwähnt in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Massendaten **big data**	Daten, deren Merkmale in Bezug auf Volumen, Komplexität, Änderungsdynamik und/oder mangelnder Struktur spezielle Technologien, Techniken und Methoden zur Verarbeitung erfordern. Erwähnt in ISTQB – CTAI Syllabus v1.0 [137].
Maßnahme (im Kontext von Zertifizierung sowie Sicherheits-, Safety- oder Datenschutz) **control (in the context of certification as well as security, safety or data protection)**	Verfahren (technisch, organisatorisch, rechtlich, physisch), mit denen Risiken für Sicherheits-, Safety- oder Datenschutzprobleme verringert werden.
menschzentrierte Gestaltung **human-centred design**	Herangehensweise bei der Gestaltung und Entwicklung von Systemen, die darauf abzielt, interaktive Systeme gebrauchstauglicher zu machen, indem sie sich auf die Verwendung des Systems konzentriert und Kenntnisse und Techniken aus den Bereichen der Arbeitswissenschaft/Ergonomie und der Gebrauchstauglichkeit anwendet. Erwähnt in DIN EN ISO 6385:2016 [235], ISO 9241-210:2020 [183].
Merkmal **feature**	Individuell messbare Eigenschaft eines zu beobachtenden Objekts. Erwähnt in ETSI GR ENI 004 V2.2.1 [499], ISTQB – CTAI Syllabus v1.0 [137].

Begriff (de/en)	Bedeutung und Verwendungen
Metrik **metric**	Eine Metrik ist ein Maß, um die Eigenschaft eines Objekts zu quantifizieren. Im Bereich von KI bzw. Machine Learning wird es eingesetzt, um die Charakteristika eines KI-Systems zu messen und sie damit in möglichst aufschlussreichen Kennzahlen abzubilden. Die Kennzahlen können sich auf Gütekriterien beziehen wie z. B. eine Falsch-Positiv-Rate für Klassifikationsausgaben oder den mittleren quadratischen Fehler bei Regressionsaufgaben. Zudem kann sie weiterführende Bewertungskriterien wie z. B. die Stärke des Bias zwischen Geschlechtern darstellen. Die Metrik sollte über ein algorithmisch umsetzbares Messprinzip definiert sein, sodass es für konkrete Problemstellungen anwendbar ist. Anmerkung: Metriken, die bewerten, wie gut ein KI-Systems seine Aufgabe oder Funktion erfüllt, werden auch als funktionale Leistungsmetriken bezeichnet. Erwähnt in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC TR 29119-11 [132].
ML-Modell **ML model**	Ein mathematisches Konstrukt, das auf der Grundlage von Eingangsdaten eine Schlussfolgerung oder Vorhersage trifft. Erwähnt in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16].
ML-System **ML system**	Ein System, das ML-Modelle integriert. Erwähnt in ISTQB – CTAI Syllabus v1.0 [137].
Modell **model**	Physikalische, mathematische oder anderweitig logische Darstellung eines Systems, einer Einheit, eines Phänomens, eines Prozesses oder von Daten einschließlich ihrer Beziehungen und Abhängigkeiten unter Verwendung eines festgelegten Satzes von Regeln und Konzepten. Erwähnt in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132].
Modul	→ Siehe Komponente.
Nachtrainieren **retraining**	Aktualisierung eines trainierten Modells durch erneutes Training mit anderen Trainingsdaten. Erwähnt in ISO/IEC 22989:2022 [16].
Neuron	→ Siehe künstliches Neuron.
Ökobilanzierung **life cycle assessment (LCA)**	Bestimmung des Inventars und zugehöriger Umweltwirkungen eines Produkts/einer Dienstleistung.
Ontologie **ontology**	Als Ontologie wird einerseits eine philosophische Disziplin bezeichnet, die sich damit befasst, Konzepte der Welt in möglichst sinnerhaltende Kategoriensysteme einzuordnen. Andererseits bezeichnet man in der Informatik konkret solche Kategoriensysteme beispielsweise aus Begriffen und Relationen zur algorithmischen Nutzung als „Ontologien“. Erwähnt in ETSI GR ENI 004 V2.2.1 [499].
Parameter **parameter**	Im Kontext des Maschinellen Lernens: interne Variable eines Modells, die sich auf die Berechnung von Ergebnissen auswirkt. Erwähnt in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132].
Pfadoptimierung **path optimisation**	Algorithmik zur Identifikation des kürzesten/günstigsten Pfades in einem Graphen aus Knoten und Kanten.
Planung **planning**	KI-Methode, die einen Arbeitsablauf aus einer Reihe von Aktionen zusammenstellt, um ein bestimmtes Ziel zu erreichen. Erwähnt in ISO/IEC 22989:2022 [16].

Begriff (de/en)	Bedeutung und Verwendungen
Präzision (im Kontext Klassifikation) **precision (in the context of classification)**	Als Präzision bzw. „positiver Vorhersagewert“ wird im Kontext von Klassifikation der Anteil von richtig-positiven Ausgaben eines Systems an seinen gesamten positiven Ausgaben bezeichnet. Erwähnt in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Prüfen (i. S. d. Prüfung und Zertifizierung)	Ermittlung eines oder mehrerer Merkmale an einem Gegenstand der Konformitätsbewertung nach einem Verfahren.
Rechenschaft **accountability**	Beschreibt eine Beziehung zwischen einem Akteur*innen und einem Forum, in welcher der Akteur*innen seine Haltung erläutern und rechtfertigen muss. Das Forum hat das Recht, die Erläuterungen des Betreibers zu hinterfragen (zur Klärung und für zusätzliche Erklärungen) und ein Urteil abzugeben. Grundsätzlich sollten dem Akteur*innen Konsequenzen angekündigt werden, sodass die Rechenschaftspflicht von dem Akteur*innen wahrgenommen und umgesetzt wird. Erwähnt in ISO/IEC 22989:2022 [16].
Regression **regression**	ML-Methode, die zu einem quantitativen Ausgabewert für eine gegebene Eingabe führt. Erwähnt in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137], ISO/IEC 23053:2022 [24].
Retraining	→ Siehe Nachtrainieren.
richtig negativ **true negative (TN)**	Eine Vorhersage, bei der das Modell die negative Kategorie korrekt vorhersagt. Erwähnt in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
richtig positiv **true positive (TP)**	Eine Vorhersage, bei der das Modell die positive Kategorie korrekt vorhersagt. Erwähnt in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Risiko **risk**	Der Begriff Risiko bezeichnet in der Regel unerwünschte Ereignisse mit noch unsicherem Eintreten, die mit einem Produkt oder Prozess verbunden sind; in formalen Definitionen werden sie meist als Kombination aus Schadenshöhe und Eintrittswahrscheinlichkeit für einen Schaden charakterisiert. Die quantitative Gesamteinordnung ergibt sich meist aus dem Erwartungswert für den (jeweils näher zu spezifizierenden) Schaden, also als Produkt aus Schadenshöhe und Eintrittswahrscheinlichkeit. Manchmal werden weitere Parameter integriert (z. B. Entdeckungswahrscheinlichkeit bzw. Vermeidbarkeit durch menschlichen Eingriff). Erwähnt in ISO/IEC 22989:2022 [16].
Roboter **robot**	Ein Roboter ist ein technisches System, das über Sensoren zur Wahrnehmung seiner Umwelt, eine zweckorientierte Verarbeitungseinheit und Effektoren zur Veränderung seiner räumlichen Relation in der Umwelt oder der Umwelt selbst verfügt. Erwähnt in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132].
Robotik **robotics**	Disziplin, die sich mit der Konstruktion von Robotern beschäftigt. Erwähnt in ISO/IEC 22989:2022 [16].
Robustheit **robustness**	Fähigkeit eines Systems, seine Funktion unter beliebigen Umständen zu erfüllen. Erwähnt in ISO/IEC 22989:2022 [16].
Safety	→ Siehe Sicherheit (Safety).

Begriff (de/en)	Bedeutung und Verwendungen
Sanierungspfad refurbishment track	Optimierte Reihenfolge von Sanierungen im Bausektor.
Semantik semantics	Forschungsfeld über die Analyse der Bedeutung von etwas (z. B. eines Satzes oder einer Beziehung in einem Modell). Erwähnt in ETSI GR ENI 004 V2.2.1 [499].
Sensitivität recall	Als Sensitivität bzw. „richtig-positiv-Rate“ wird in formalen Kontexten der Anteil von richtig-positiven Ausgaben eines Systems unter den tatsächlich positiven Sollausgaben bezeichnet (siehe richtig positiv sowie falsch negativ). Erwähnt in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Sicherheit (Safety) safety	Als Safety (nur unzureichend mit „Sicherheit“ zu übersetzen) wird meistens spezifisch die Abwesenheit von Risiken für Leib und Leben durch ein System bezeichnet. Im weiteren Sinne werden auch psychische Gesundheit sowie die Unversehrtheit von Umwelt und weiteren Werten zur Safety gezählt. Erwähnt in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Sicherheit (Security im Sinne von Informationssicherheit) security	Der Begriff Informationssicherheit (der Begriff „Sicherheit“ gibt dies nur unzureichend wieder) bezeichnet die Fähigkeit eines Systems, über dessen Lebenszyklus u. a. wichtige Informationen vor unerlaubtem Zugriff zu schützen, dessen Verfügbarkeit sicherzustellen oder dessen Vertraulichkeit, Integrität, Authentizität, Rechenschaft und Zuverlässigkeit zu bewahren, auch für die Funktionalität. Erwähnt in DIN EN ISO/IEC 27000er Reihe [131].
Smart Grid Architekturmodell smart grid architecture model	Generisches Modell eines Smart Grids, mithilfe dessen die Umsetzungsmöglichkeiten verschiedener Dienste oder Funktionen untersucht werden.
Smart Grid smart grid	Verbindung von Energietechnik mit Informations- und Kommunikationstechnologie zur Optimierung von Energieerzeugung, -transport und -nutzung.
soziotechnisches System socio-technical system	Soziotechnische Systeme beinhalten die Subsysteme Mensch und Technik, die miteinander verknüpft sind und in Wechselwirkung zueinander stehen oder stehen sollten. Die KI-Technologie steht dabei im Kontext zum Menschen, dem organisatorischen Umfeld und der Gesellschaft als Ganzes.
Spracherkennung speech recognition	Eine KI-Fähigkeit, die mittels der Umwandlung eines Sprachsignals von Sprache in Text den Inhalt der Sprache darstellt. Erwähnt in ISO/IEC 22989:2022 [16].
strukturierte Daten structured data	Informationen, die auf eine bestimmte Art und Weise (ein festes Format, Datenmodell oder Schema) in einem Datensatz oder einer Datei organisiert sind. Erwähnt in ETSI GR ENI 004 V2.2.1 [499].
Stützvektormaschine support vector machine (SVM)	Methode des Maschinellen Lernens, die Entscheidungsgrenzen mit maximalem Grenzwert findet. Erwähnt in ISO/IEC 22989:2022 [16], ISTQB – CTAI Syllabus v1.0 [137].

Begriff (de/en)	Bedeutung und Verwendungen
subsymbolische KI **subsymbolic AI**	Typ von KI-Methoden, die auf Modellen mit numerischer Darstellung und impliziter Informationskodierung basieren. Erwähnt in ISO/IEC 22989:2022 [16].
Support-Vector-Maschine (SVM)	→ Siehe Stützvektormaschine.
symbolische KI **symbolic AI**	Ein Typ von KI-Methoden, der auf der Verarbeitung von Symbolen und Strukturen basiert. Erwähnt in ISO/IEC 22989:2022 [16].
Syntax **syntax**	Menge an Regeln, die bestimmen, wie Elemente einer Aussage strukturiert sind. Erwähnt in ETSI GR ENI 004 V2.2.1 [499].
Taxonomie **taxonomy**	Verfahren zur Klassifikation von Objekten nach bestimmten Kriterien.
teilüberwachtes Lernen **semi supervised ML**	Mischform von überwachtem und unüberwachtem Lernen, wobei die Trainingsdaten sowohl aus markierten als auch aus unmarkierten Daten bestehen. Erwähnt in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16].
Testdaten (im KI-Kontext) **test data**	Daten, die zur Bewertung der Leistung eines finalen KI-Modells (im Allgemeinen) oder Maschinellen Lernmodells (im Speziellen) vor dessen Inbetriebnahme verwendet werden. Anmerkung: Grundsätzlich sollen Testdaten disjunkt zu Trainingsdaten und Validierungsdaten sein. Erwähnt in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132].
tiefes Lernen **deep learning (DL)**	Deep Learning bezeichnet eine Klasse von Optimierungsmethoden künstlicher neuronaler Netze (siehe Artificial Neural Network), die zahlreiche verborgene Schichten (hidden layers) zwischen Eingabeschicht und Ausgabeschicht haben und dadurch eine umfangreiche innere Struktur aufweisen. Als Erweiterung der Lernalgorithmen für Netzstrukturen mit sehr wenigen oder keinen Zwischenlagen ermöglichen die Methoden des Deep Learnings auch bei zahlreichen Zwischenlagen einen stabilen Lernerfolg. Erwähnt in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
tiefes neuronales Netz **deep neural network**	Neuronales Netzwerk, das neben der Eingabe- und Ausgabeschicht noch über weitere, sogenannte versteckte Schichten von Knoten verfügt (vgl. „tiefes Lernen“). Erwähnt in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Trainingsdaten **training data**	Daten, die im Trainingsprozess zum Erstellen eines KI-Modells verwendet werden können. Erwähnt in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132].
Trainingsprozess **training**	Prozess der Vermittlung einer Reihe von Kenntnissen, KI-Fähigkeiten, Verfahren und/oder Verhaltensweisen an eine Entität. Erwähnt in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16].
Transformer **transformer**	Im Maschinellen Lernen gehören Transformer und ihre Architekturen zu neuronalen Modellen, die u. a. für zahlreiche sprachtechnologische Aufgaben eingesetzt werden. Transformer gehören zu den Architekturen des tiefen Lernens (deep learning).

Begriff (de/en)	Bedeutung und Verwendungen
Transparenz transparency	Verfügbarkeit einer offenen, verständlichen und zugreifbaren Darstellung von Informationen zu funktionalen Aspekten eines KI-Systems. Dies beinhaltet u. a. die Erklärbarkeit das KI-Systems (z. B. neuronale Netze), die Nachvollziehbarkeit des Datenschutzkonzepts sowie Informationen zu Qualitätssicherungsprozessen während der Entwicklung. Erwähnt in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Überanpassung overfitting	Von Overfitting spricht man, wenn ein ML-Modell so stark auf den Trainingsdatensatz ausgerichtet ist, dass es nur noch schwer auf neue Daten verallgemeinert werden kann. Erwähnt in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Überprüfbarkeit examinability	Die Möglichkeit, Aussagen nachzuvollziehen, indem beispielsweise Zugriffe auf Daten, Dokumente oder (KI-)Systeme gewährt werden.
überwachtes ML supervised ML	Als überwachtes Lernen bezeichnet man im engeren Sinne Verfahren des Maschinellen Lernens, die mit konkret spezifizierten Sollausgaben trainiert werden (sogenannte „Labels“). Im weiteren Sinne werden Verfahren dazu gezählt, deren Lernziel durch konkrete Vorgaben bestimmt ist, wenn auch nicht auf der Ebene von Einzelausgaben. Dieser weitere Sinn schließt Verfahren wie etwa GANs und bestärkendes Lernen ein. Erwähnt in ISO/IEC 22989:2022 [16].
Unteranpassung underfitting	Die Schaffung eines ML-Modells, das den zugrunde liegenden Trend des Trainingsdatensatzes nicht widerspiegelt, was zu einem Modell führt, das nur schwer genaue Vorhersagen machen kann. Erwähnt in ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
unüberwachtes (Maschinelles) Lernen unsupervised ML	Als unüberwachtes Lernen werden Verfahren des Maschinellen Lernens bezeichnet, die eine Funktion erlernen, ohne auf konkret spezifizierte Zielvorgaben (beispielsweise „Labels“) angewiesen zu sein. Es existieren unterschiedliche Auffassungen, ab welchem Grad an Konkretheit externer Zielvorgaben nicht mehr von „unüberwachtem Lernen“ gesprochen werden kann. Erwähnt in ETSI GR ENI 004 V2.2.1 [499], ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132], ISTQB – CTAI Syllabus v1.0 [137].
Validierung (im Kontext System- und Produktentwicklung) validation (in the context of system or product development)	Validierung im Kontext von System- und Produktentwicklung ist die Bestätigung durch die Erbringung eines objektiven Nachweises, dass die Anforderungen an einen bestimmten Verwendungszweck oder eine bestimmte Anwendung erfüllt worden sind. Sie grenzt sich damit ab von der Verifikation sowie von der Validierung im Kontext von ML, die im Rahmen des Trainingsprozesses auf eine Optimierung von Hyperparametern bzw. Auswahl eines geeigneten Modells abzielt (→ siehe Validierung im Kontext ML). Erwähnt in ISO/IEC 22989:2022 [16].
Validierung (im Kontext von ML) validation (in the context of ML)	Validierung, auch bezeichnet als ML-Modelloptimierung oder -Model-Tuning, bezeichnet im Kontext von ML die Prüfung trainierter ML-Modelle durch Validierungsdaten. Dadurch lässt sich die Güte der trainierten ML-Modelle erkennen, vergleichen und optimieren (→ siehe Hyperparameter). Insbesondere lässt sich meist erkennen, ob das ML-Modell auf unbekannte Daten generalisieren kann oder auf die Trainingsdaten übertrainiert wurde (Überanpassung), vergleichbar mit einem „Auswendiglernen“ aller Trainingsfragen samt korrekter Antworten. Dieser Schritt grenzt sich von der Definition der Validierung im Kontext der System- und Produktentwicklung ab, da es sich bei der Validierung im ML-Kontext lediglich um einen Zwischenschritt im Trainingsprozess und nicht um eine unmittelbare Überprüfung des finalen Modells bzw. der System- bzw. Produktanforderungen handelt (→ Validierung im Kontext System- und Produktentwicklung). Erwähnt in ISTQB – CTAI Syllabus v1.0 [137].

Begriff (de/en)	Bedeutung und Verwendungen
Validierungsdaten **validation data**	Im Kontext von ML werden Validierungsdaten zur Überprüfung von trainierten ML-Modellen herangezogen (→ siehe Validierung im Kontext ML). Validierungsdaten dürfen im Allgemeinen nicht Teil der Trainingsdaten sein. Erwähnt in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132].
Verfügbarkeit **availability**	Die Eigenschaft, bei Bedarf durch Befugte zugänglich und nutzbar zu sein. Charakterisiert durch den Grad kann das Ausmaß der Verfügbarkeit von Merkmalen wie Aktualität, Interpretierbarkeit sowie Vollständigkeit von Informationen abhängen. Erwähnt in ISO/IEC 22989:2022 [16].
Verifizierung **verification**	Bestätigung durch objektive Nachweise, dass die festgelegten Anforderungen erfüllt wurden. Anmerkung: Die Verifizierung legt ausschließlich dar, dass ein Produkt mit seiner Spezifikation übereinstimmt. Erwähnt in ISO/IEC 22989:2022 [16].
Verständlichkeit **understandability**	Die Eigenschaft einer Entität, eines Systems oder eines Prozesses, nachvollziehbar zu sein.
Vertrauenswürdigkeit **trustworthiness**	Fähigkeit, Erwartungen nachweislich zu erfüllen. Anmerkung 1: Je nach Kontext oder Sektor und auch je nach dem spezifischen Produkt oder der Dienstleistung, den Daten und der verwendeten Technologie unterscheiden sich Merkmale, die überprüft werden müssen, um sicherzustellen, dass die Erwartungen der Interessengruppen erfüllt werden. Anmerkung 2: Zu den Merkmalen der Vertrauenswürdigkeit gehören z. B. Zuverlässigkeit, Verfügbarkeit, Belastbarkeit, Sicherheit (im Sinne von Security und Safety), Datenschutz, Verantwortlichkeit, Transparenz, Integrität, Authentizität, Qualität und Benutzerfreundlichkeit. Anmerkung 3: Vertrauenswürdigkeit ist ein Attribut, das sich auf Dienstleistungen, Produkte, Technologie, Daten und Informationen sowie – im Kontext der Governance – auf Organisationen anwenden lässt. Erwähnt in ISO/IEC 22989:2022 [16].
Verzerrung	→ Siehe Bias.
Vollständigkeit **completeness**	Grad, in dem Daten, die mit einer Entität assoziiert sind, Werte für alle Attribute dieser Entität sowie für zu dieser in Beziehung stehende Entitäten aufweisen.
Vorhersagbarkeit **predictability**	Eigenschaft eines KI-Systems, die verlässliche Annahmen über Ergebnisse bzw. die Güte der Vorhersagbarkeit ermöglicht. Mittels des Grades kann das Ausmaß von zutreffender Spekulation über eingetretene Zustände und Prozesse beschrieben werden. Erwähnt in ISO/IEC 22989:2022 [16].
Vorhersage **prediction**	Funktion eines ML-Modells, die zu einem vorhergesagten Zielwert für eine gegebene Eingabe führt. Erwähnt in ISO/IEC 22989:2022 [16], ISO/IEC TR 29119-11 [132].
Widerstandsfähigkeit **resilience**	Resistenz gegenüber Störungen und Ausfällen mit der damit verbundenen Fähigkeit, parasitäre Einflüsse auf den Betrieb zu verhindern. Erwähnt in ISO/IEC 22989:2022 [16].

Begriff (de/en)	Bedeutung und Verwendungen
Wissensrepräsentation **knowledge representation**	Repräsentation von Wissen, die für ein KI-System, z. B. ein Expertensystem, nutzbar ist. Erwähnt in ETSI GR ENI 004 V2.2.1 [499].
Zertifizierung	Bestätigung durch eine dritte Seite, bezogen auf einen Gegenstand der Konformitätsbewertung, ausgenommen Akkreditierung.
Zielpopulation **target population**	Personengruppe, für die etwas gestaltet wird, beschrieben anhand von relevanten Merkmalen. Erwähnt in DIN EN ISO 26800:2011 [239].
Zugänglichkeit **accessibility**	Ausmaß, in dem Produkte, Systeme, Dienstleistungen, Umgebungen und Einrichtungen durch Menschen aus einer Population mit dem weitesten Umfang an Benutzererfordernissen, Merkmalen und Fertigkeiten genutzt werden können, um identifizierte Ziele in identifizierten Nutzungskontexten zu erreichen. Anmerkung zur Terminologie: Der Nutzungskontext umfasst die direkte Nutzung oder durch Assistenztechnologien unterstützte Nutzung. Anmerkung zur Übersetzung: Die Begriffe „Barrierefreiheit“ und „Zugänglichkeit“ werden häufig synonym genutzt. „Barrierefreiheit“ ist mehr als nur die physische Zugänglichkeit, schließt diese jedoch mit ein. Deshalb wird z. B. im baulichen Bereich der Begriff „Zugänglichkeit“ und im IKT-Bereich der Begriff „Barrierefreiheit“ bevorzugt genutzt. Erwähnt in DIN EN ISO 9241-210:2020 [183], ISO 9241-112:2017 [249].
Zulassung	Erlaubnis, ein Produkt, eine Dienstleistung oder einen Prozess zum angegebenen Zweck oder unter angegebenen Bedingungen auf den Markt zu bringen oder zu nutzen.
Zuverlässigkeit **reliability**	Die Eigenschaft, ein vertrauenswürdiges Verhalten aufweisen zu können, sowie die Eigenschaft, konsistent ein beabsichtigtes Verhalten und Ergebnisse aufzuweisen. Erwähnt in ISO/IEC 22989:2022 [16].

10 Quellen- und Literaturverzeichnis

[1] Blind Prof. Dr. Knut; Jungmittag Prof. Dr. Andre; Mangelsdorf Dr. Axel, Der gesamtwirtschaftliche Nutzen der Normung. Eine Aktualisierung der DIN-Studie aus dem Jahr 2000, 2000, verfügbar unter: https://www.din.de/resource/blob/79542/946e70a818ebdaacce9705652a052b25/gesamtwirtschaftlicher-nutzen-der-normung-data.pdf (letzter Zugriff: 2022-09-26)

[2] Bundesministerium für Wirtschaft und Technologie (BMWK), Strategie Künstliche Intelligenz der Bundesregierung, 2018, verfügbar unter: www.ki-strategie-deutschland.de (letzter Zugriff: 2022-08)

[3] Bundesregierung, Die entscheidende Zukunftstechnologie des 21. Jahrhunderts, 2020, verfügbar unter: https://www.bundesregierung.de/breg-de/suche/fortschreibung-ki-strategie-1824340 (letzter Zugriff: 2022-09-26)

[4] European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts Com/2021/206 Final, 2021, verfügbar unter: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 (letzter Zugriff: 2022-08-29)

[5] European Commission, Mitteilung der Kommission an das EUROPÄISCHE PARLAMENT, DEN EUROPÄISCHEN RAT, DEN RAT, DEN EUROPÄISCHEN WIRTSCHAFTS- UND SOZIALAUSSCHUSS UND DEN AUSSCHUSS DER REGIONEN, Koordinierter Plan für künstliche Intelligenz, 2018, verfügbar unter: https://eur-lex.europa.eu/resource.html?uri=cellar:22ee84bb-fa04-11e8-a96d-01aa75ed71a1.0003.02/DOC_1&format=PDF (letzter Zugriff: 2022-06-30)

[6] Independent High-Level Expert Group on AI set up by the European Commission, Policy and Investment Recommendations for Trustworthy Artificial Intelligence, 2019, verfügbar unter: https://futurium.ec.europa.eu/en/european-ai-alliance/open-library/policy-and-investment-recommendations-trustworthy-artificial-intelligence (letzter Zugriff: 2022-09-26)

[7] European Commission, Europäische Kommission, WHITE PAPER, On Artificial Intelligence ENA European approach to excellence and trust, 2020, verfügbar unter: commission-white-paper-artificial-intelligence-feb2020_en.pdf (europa.eu) (letzter Zugriff: 2022-0829)

[8] Independent High-Level Expert Group on AI set up by the European Commission, Ethics Guidelines for Trustworthy AI, 2019

[9] European Commission, Regulatory framework proposal on artificial intelligence, Juni 2022, verfügbar unter: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (letzter Zugriff: 2022-08-29)

[10] IEEE 7001:2021, Standard for Transparency of Autonomous Systems

[11] IEEE 7002:2022, Standard for Data Privacy Process

[12] IEEE 7007:2021, Ontological Standard for Ethically driven Robotics and Automation Systems

[13] IEEE 7005:2021, Transparent Employer Data Governance

[14] ISO/IEC JTC1/SC 42, Artificial intelligence, verfügbar unter: https://www.iso.org/committee/6794475.html (letzter Zugriff: 2022-09-26)

[15] ISO/IEC TR 24368:2022, Information technology – Artificial intelligence – Overview of ethical and societal concerns, verfügbar unter: https://www.iso.org/standard/78507.html (letzter Zugriff: 2022-09-26)

[16] ISO/IEC 22989:2022, Informationstechnik – Künstliche Intelligenz – Konzepte und Terminologie der Künstlichen Intelligenz

[17] DIN EN ISO/IEC 17065:2013, Konformitätsbewertung – Anforderungen an Stellen, die Produkte, Prozesse und Dienstleistungen zertifizieren (ISO/IEC 17065:2012); Deutsche und Englische Fassung EN ISO/IEC 17065:2012

[18] DIN EN ISO/IEC 17067:2013, Konformitätsbewertung – Grundlagen der Produktzertifizierung und Leitlinien für Produktzertifizierungsprogramme (ISO/IEC 17067:2013); Deutsche und Englische Fassung EN ISO/IEC 17067:2013

[19] ISO/IEC TR 17026:2015, Konformitätsbewertung – Beispiel für ein Produktzertifizierungsprogramm für materielle Produkte

[20] ISO/IEC TR 17028:2017, Konformitätsbewertung – Leitlinien und Beispiele für ein Zertifizierungsprogramm für Dienstleistungen

[21] ISO/IEC TR 17032:2019, Konformitätsbewertung – Leitlinien und Beispiele für ein Zertifizierungsprogramm für Prozesse, verfügbar unter: https://www.iso.org/standard/29355.html (letzter Zugriff: 2022-09-26)

[22] DIN EN ISO/IEC 17021-1:2015, Konformitätsbewertung – Anforderungen an Stellen, die Managementsysteme auditieren und zertifizieren – Teil 1: Anforderungen

[23] Tambiama Madiega, Anne Louise Van De Pol, European Parliamentary Research Service, Artificial intelligence act and regulatory sandboxes, 2022, verfügbar unter: https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733544/EPRS_BRI(2022)733544_EN.pdf (letzter Zugriff: 2022-09-09)

[24] ISO/IEC 23053:2022, Framework für Systeme der Künstlichen Intelligenz (KI) basierend auf maschinellem Lernen (ML), 2022

[25] ISO/IEC DIS 23894:2022, Informationstechnik – Künstliche Intelligenz – Risikomanagement

[26] ISO/IEC 38507:2022, Informationstechnik – IT-Governance – Governance-Auswirkungen der Nutzung von KI in Organisationen, verfügbar unter: https://www.iso.org/standard/56641.html (letzter Zugriff: 2022-09-26)

[27] ISO/IEC DIS 42001, Information Technology – Artificial intelligence – Management system

[28] ISO/IEC TR 24028:2020, Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence

[29] ISO/IEC PRF TS 4213, Information technology – Artificial Intelligence – Assessment of machine learning classification performance

[30] ISO/IEC DIS 5338, Information technology – Artificial intelligence – AI system life cycle processes

[31] ISO/IEC CD 5339, Information Technology – Artificial Intelligence – Guidelines for AI Applications

[32] ISO/IEC CD 5392, Information technology – Artificial intelligence – Reference Architecture of Knowledge Engineering

[33] ISO/IEC DTR 5469, Artificial intelligence – Functional safety and AI systems

[34] ISO/IEC AWI TS 5471, Artificial intelligence – Quality evaluation guidelines for AI systems

[35] ISO/IEC DIS 25059:2022-07 – Entwurf, System- und Software-Engineering – Qualitätskriterien und Bewertung von Systemen und Softwareprodukten (SQuaRE) – Qualitätsmodell für KI-Systeme

[36] ISO/IEC AWI TS 6254, Information technology – Artificial intelligence – Objectives and approaches for explainability of ML models and AI systems

[37] ISO/IEC AWI TS 8200, Information technology – Artificial intelligence – Controllability of automated artificial intelligence systems

[38] ISO/IEC AWI TS 12791, Information technology – Artificial intelligence – Treatment of unwanted bias in classification and regression machine learning tasks

[39] ISO/IEC 5259 (alle Teile), Artificial intelligence – Data quality for analytics and machine learning (ML)

[40] ISO/IEC CD 5259-1, Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 1: Overview, terminology, and examples

[41] ISO/IEC AWI 5259-2, Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 2: Data quality measures

[42] ISO/IEC CD 5259-3:2022, Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 3: Data quality management requirements and guidelines

[43] ISO/IEC CD 5259-4, Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 4: Data quality process framework

[44] ISO/IEC AWI 5259-5, Artificial intelligence – Data quality for analytics and machine learning (ML) – Part 5: Data quality governance

[45] ISO/IEC CD 8183, Informationstechnik – Künstliche Intelligenz – Framework für den Lebenszyklus von Daten

[46] INDEPENDENT HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE SET UP BY THE EUROPEAN COMMISSION, A definition of AI: Main capabilities and scientific disciplines, 2019, verfügbar unter: https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines (letzter Zugriff: 2022-09-27)

[47] T. Schmid, W. Hildesheim, T. Holoyad, K. Schumacher, The AI Methods, Capabilities and Criticality Grid – A Three-Dimensional Classification Scheme for Artificial Intelligence Applications, 2021, verfügbar unter: https://doi.org/10.1007/s13218-021-00736-4 (letzter Zugriff: 2022-09-26)

[48] T. Schmid; W. Hildesheim; T. Holoyad; K. Schumacher, Managing and Understanding Artificial Intelligence Solutions – The AI-Methods, Capabilities and Criticality Grid and its Value for Decision Makers, Developers and Regulators, Künstliche Intelligenz managen und verstehen – Der Praxis-Wegweiser für Entscheidungsträger, Entwickler und Regulierer, 1. Auflage, Beuth-Verlag, Berlin, 2020, https://www.beuth.de/de/publikation/kuenstliche-intelligenz-managen-und-verstehen/359390396 (letzter Zugriff: 2022-09-26)

[49] Goertzel, Ben, Perception Processing for General Intelligence: Bridging the Symbolic/Subsymbolic Gap. Artificial General Intelligence, 2012

[50] Hammer Barbara, Hitzler Pascal, Perspectives of Neural-Symbolic Integration. Studies in Computational Intelligence, 2007

[51] Russell Stuart J., Norvig Peter, Artificial Intelligence: a modern approach. 3. Ed., 2014

[52] Horvitz Eric J.; Breese, John S.; Henrion, Max, Decision theory in expert systems and artificial intelligence. International Journal of Approximate Reasoning 2 (3), 1988

[53] Martin, Andreas; Hinkelmann, Knut; Gerber, Aurona; Lenat, Doug; van Harmelen, Frank; Clark, Peter, Proceedings of the AAAI 2019 Spring Symposium on Combining Machine Learning with Knowledge Engineering, 2019

[54] McGarry, Kenneth; Wermter, Stefan; MacIntyre, John, Hybrid neural systems: from simple coupling to fully integrated neural networks. Neural Computing Surveys 2 (1), 1999

[55] Méhaut, Philippe; Winch, Christopher, The European qualification framework: Skills, competences or knowledge?, 2012

[56] Bloom, Benjamin, Taxonomy of educational objectives, Vol. 1: cognitive domain, 2016

[57] Ritchie, J. Brendan, Carruthers, P, The bodily senses. In: Matthen M. (Ed.), The Oxford handbook of the philosophy of perception, 2015

[58] Macpherson, Fiona, Individuating the senses. In: Macpherson F. (ed.), The senses: classic and contemporary philosophical readings, 2011

[59] Krathwohl, David R., A revision of Bloom's taxonomy. Theory Pract 41, 2002

[60] Davidson, Donald, Essay III. In: Davidson D (ed) Essays on actions and events, 1980

[61] Shannon, Claude E., A mathematical theory of communication, 1948

[62] Luhmann, Niklas, What is communication? Communication Theory 2 (3), 1992

[63] Deutsches Institut für Normung, Deutsche Kommission Elektrotechnik Elektronik Informationstechnik, Herausgeber: Wahlster & Winterhalter, Deutsche Normungsroadmap Künstliche Intelligenz, 2020

[64] IEEE 7000:2021, IEEE Standard Model Process for Addressing Ethical Concerns during System Design, verfügbar unter: https://standards.ieee.org/ieee/7000/6781/# (letzter Zugriff: 2022-09-26)

[65] Immanuel Kant, Grundlegung zur Metaphysik der Sitten, 1785

[66] Enquete-Kommission (Deutscher Bundestag – Enquete-Kommission „Künstliche Intelligenz"), Bericht der Enquete-Kommission Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale, 2020, verfügbar unter: Drucksache 19/23700 (bundestag.de) (letzter Zugriff: 2022-09-26)

[67] Jobin, Anna; Ienca, Marcello; Vayena, Effi, The global landscape of AI ethics guidelines, 2019

[68] Heesen, J. et al., Ethik-Briefing. Leitfaden für eine verantwortungsvolle Entwicklung und Anwendung von KI-Systemen, 2020, verfügbar unter: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_Whitepaper_EB_200831.pdf (letzter Zugriff: 2022-09)

[69] Sartori, Laura; Thedorou, Andreas, A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. In Ethics and Information Technology. In: Ethics and Information Technology. 24:4, 2022, verfügbar unter: https://doi.org/10.1007/s10676-022-09624-3 (letzter Zugriff: 2022-09)

[70] Birhane, Abeba, Algorithmic injustice: a relational ethics approach. Patterns. 2, 2021, verfügbar unter: https://doi.org/10.1016/j.patter.2021.100205 (letzter Zugriff: 2022-09)

[71] Bratteteig, Tone; Verne, Guri, Does AI make PD obsolete? Exploring Challenges from Artificial Intelligence to Participatory Design. In Proceedings of PDC 2018, Belgium, August 2018, 5 pages, 2018, verfügbar unter: Does AI make PD obsolete? | Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial – Volume 2 (acm.org) (letzter Zugriff: 2022-09)

[72] Friedman, Batya et al., Value Sensitive Design and Information Systems. In: Doorn, Neelke; Schuurbiers, Daan; van de Poel, Ibo & Gorman, Michael E. (Hrsg.): Early Engagement and New Technologies: Opening Up the Laboratory. Springer VS, Wiesbaden, S. 55–96, 2013

[73] DIE ETHISCHEN LEITLINIEN DER GESELLSCHAFT FÜR INFORMATIK E. V., 2018, verfügbar unter: Unsere Ethischen Leitlinien – Gesellschaft für Informatik e. V. (gi.de) (letzter Zugriff: 2022-09-26)

[74] Künstliche Intelligenz im Dienste der Diversität, 2022, verfügbar unter: https://kidd-prozess.de/ (letzter Zugriff: 2022-08-12)

[75] DIN EN ISO/IEC 18045:2021, Informationstechnik – Sicherheitstechniken – Methodik für die Bewertung der IT-Sicherheit (ISO/IEC 18045:2008); Deutsche Fassung EN ISO/IEC 18045:2020, nur auf CD-ROM

[76] Arrangement on the Recognition of Common Criteria Certificates in the field of Information Technology Security, 1998, verfügbar unter: Arrangement on the Recognition of Common Criteria Certificates (commoncriteriaportal.org) (letzter Zugriff: 2022-09-26)

[77] Gemeinsame Kriterien für die Prüfung und Bewertung der Sicherheit von Informationstechnik, verfügbar unter: https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/Zertifizierung-und-Anerkennung/Zertifizierung-von-Produkten/Zertifizierung-nach-CC/IT-Sicherheiskriterien/CommonCriteria/commoncriteria_node.html (letzter Zugriff: 2022-09-26)

[78] ISO/IEC 38500:2015, Information security, cybersecurity and privacy protection – Evaluation criteria for IT security – Methodology for IT security evaluation

[79] M. Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio & Patrick J. Coles, Variational quantum algorithms, 2021, verfügbar unter: https://www.nature.com/articles/s42254-021-00348-9 (letzter Zugriff: 2022-08-12)

[80] Prasanna Date, Davis Arthur & Lauren Pusey-Nazzaro, QUBO formulations for training machine learning models, 2021, verfügbar unter: https://www.nature.com/articles/s41598-021-89461-4 (letzter Zugriff: 2022-08-12)

[81] Bundesamt für Sicherheit in der Informationstechnik (BSI), Quantum Machine Learning in the Context of IT Security, 2022, verfügbar unter: https://www.bsi.bund.de/DE/Service-Navi/Publikationen/Studien/QML/QML_node.html (letzter Zugriff: 2022-05-24)

[82] Bundesamt für Sicherheit in der Informationstechnik (BSI), AI Cloud Service Compliance Criteria Catalogue (AIC4), 2021, verfügbar unter: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.html (letzter Zugriff: 2022-05-10)

[83] Bundesamt für Sicherheit in der Informationstechnik, Sicherer, robuster und nachvollziehbarer Einsatz von KI Probleme, Maßnahmen und Handlungsbedarfe, 2021, Bonn, verfügbar unter: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Herausforderungen_und_Massnahmen_KI.pdf?__blob=publicationFile&v=6 (letzter Zugriff: 2022-0801)

[84] D. C. Ciresan, U. Meier, J. Schmidhuber, Multi-Column Deep Neural Networks for Image Classification. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012

[85] Oliver Zendel; Markus Murschitz; Martin Humenberger; Wolfgang Herzner, „Cv-hazop: Introducing test data validation for computer vision.” Proceedings of the IEEE International Conference on Computer Vision, 2015

[86] Andreas Geiger, Philip Lenz, Raquel Urtasun, Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012

[87] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The Cityscapes Dataset for Semantic Urban Scene Understanding. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

[88] Tagiew, R.; Buder, T.; Tilly, R.; Hofmann, K.; Klotz, C., Datensätze für das autonome Fahren als Grundlage für GoA3+. In: ETR – Eisenbahntechnische Rundschau, 2021, H. 9, S. 10–14

[89] DIN EN 50657:2017, Bahnanwendungen – Anwendungen für Schienenfahrzeuge – Software auf Schienenfahrzeugen

[90] ISO 21448:2022, Straßenfahrzeuge – Sicherheit der beabsichtigten Funktionalität

[91] ISO/IEC TR 24029-1:2021, Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview

[92] ISO/IEC DIS 24029-2, Artificial intelligence (AI) – Assessment of the robustness of neural networks – Part 2: Methodology for the use of formal methods

[93] C. Hasterok, J. Stompe, et al., PAISE – Das Vorgehensmodell für KI-Engineering, 2021. Verfügbar unter https://www.ki-engineering.eu/content/dam/iosb/ki-engineering/downloads/PAISE(R)_Whitepaper_CC-KING.pdf (letzter Zugrif 2022-10-21)

[94] Trapp, Mario; Schneider, Daniel; Weiss, Gereon, Towards safety-awareness and dynamic safety management. 2018 14th European Dependable Computing Conference (EDCC). IEEE

[95] Ruf, M. et al., Comparison of local vs. global optimization for trajectory planning in automated driving. 10. Workshop Fahrerassistenzsysteme, 2015

[96] Prof. Dr. Dr. Udo Di Fabio, Prof. Dr. Dr. h.c. Manfred Broy, Renata Jungo Brüngger, Dr. Ulrich Eichhorn, Prof. Dr. Armin Grunwald, Prof. Dr. Dirk Heckmann, Prof. Dr. Dr. Eric Hilgendorf, Prof. Dr. rer. Nat. Dr.-Ing. E. h. Henning Kagermann, Weihbischof Dr. Dr. Anton Losinger, Prof. Dr. Dr. Matthias Lutz-Bachmann, Prof. Dr. Christoph Lütge, Dr. August Markl, Klaus Müller, Kay Nehm, Bericht der Ethikkommission zum automatisierten und vernetzten Fahren, BMVI, 2017

[97] N. Heide, A Step towards Explainable Artificial Neural Networks in Image Processing by Dataset Assessment. Forum Bildverarbeitung, 2020

[98] DIN SPEC 13266:2020, Leitfaden für die Entwicklung von Deep-Learning-Bilderkennungssystemen

[99] ISO/IEC Guide 51:2014, Sicherheitsaspekte – Leitfaden für deren Aufnahme in Normen, verfügbar unter: https://www.beuth.de/de/technische-regel/iso-iec-guide-51/205060593 (letzter Zugriff: 2022-09-26)

[100] DIN CLC IEC/TR 63069:2021, Industrielle Prozess-Leittechnik, Steuerungs- und Automatisierungstechnik – Rahmenbedingungen für Funktionale Sicherheit und IT-Sicherheit (IEC/TR 63069:2019); Deutsche Fassung CLC IEC/TR 63069:2020, verfügbar unter: https://www.beuth.de/de/norm/din-clc-iec-tr-63069/332535627 (letzter Zugriff: 2022-09-26)

[101] DIN EN 61508-1:2011, VDE 0803-1:2011, Funktionale Sicherheit sicherheitsbezogener elektrischer/elektronischer/programmierbarer elektronischer Systeme – Teil 1: Allgemeine Anforderungen (IEC 61508-1:2010); Deutsche Fassung EN 615081:2010

[102] DIN EN 61508-2:2011, VDE 0803-2:2011-02, Funktionale Sicherheit sicherheitsbezogener elektrischer/elektronischer/programmierbarer elektronischer Systeme – Teil 2: Anforderungen an sicherheitsbezogene elektrische/elektronische/programmierbare elektronische Systeme (IEC 61508-2:2010); Deutsche Fassung EN 615082:2010

[103] DIN EN 615083:2011, VDE 0803-3:2011, Funktionale Sicherheit sicherheitsbezogener elektrischer/elektronischer/programmierbarer elektronischer Systeme – Teil 3: Anforderungen an Software (IEC 61508-3:2010); Deutsche Fassung EN 61508-3:2010

[104] NIST Special Publication 1011-I-2.0:2008, Autonomy Levels for Unmanned Systems (ALFUS) Framework, Volume I – Terminology

[105] VDE-AR-E 2842-61-2 Anwendungsregel:2021-06, Entwicklung und Vertrauenswürdigkeit von autonom/kognitiven Systemen

[106] ISO 21815-1:2022, Earth-moving machinery – Collision warning and avoidance – Part 1: General requirements

[107] ISO/TS 21815-2:2021, Earth-moving machinery – Collision warning and avoidance – Part 2: Onboard J1939 communication interface

[108] ISO/DIS 21815-3:2022- Entwurf, Earth-moving machinery – Collision warning and avoidance – Part 3: Risk area and risk level – Forward/reverse motion

[109] DIN EN ISO 13849-1:2016, Sicherheit von Maschinen – Sicherheitsbezogene Teile von Steuerungen – Teil 1: Allgemeine Gestaltungsleitsätze (ISO 138491:2015); Deutsche Fassung EN ISO 138491:2015

[110] ISO/AWI PAS 8800, Road Vehicles – Safety and artificial intelligence

[111] DIN EN ISO 138491:2021 – Entwurf, Sicherheit von Maschinen – Sicherheitsbezogene Teile von Steuerungen – Teil 1: Allgemeine Gestaltungsleitsätze (ISO/DIS 13849-1.2:2021); Deutsche und Englische Fassung prEN ISO 138491:2021

[112] DIN EN ISO 25119-1:2022-02 – Entwurf, Traktoren und Maschinen für die Land- und Forstwirtschaft – Sicherheitsbezogene Teile von Steuerungen – Teil 1: Allgemeine Gestaltungs- und Entwicklungsleitsätze (ISO 251191:2018); Deutsche und Englische Fassung prEN ISO 251191:2021

[113] ISO/IEC/IEEE DIS 150263:2022, Systeme und Software-Engineering – System- und Softwaresicherheit – Teil 3: System-Integritätslevel

[114] ISO/IEC/IEE 150261:2019, International Standard – Systems and software engineering–Systems and software assurance – Part 1: Concepts and vocabulary, verfügbar unter: https://ieeexplore.ieee.org/document/8657410 (letzter Zugriff: 2022-09-26)

[115] EU Observatory for ICT Standardisation, verfügbar unter: https://www.standict.eu/euos (letzter Zugriff: 2022-09-26)

[116] EU Observatory for ICT Standardisation, Report of TWG AI: Landscape of AI Standards, verfügbar unter: https://zenodo.org/record/5011179#.YhvgLOjMK5c (letzter Zugriff: 2022-09-26)

[117] DIN SPEC 92001-3, Künstliche Intelligenz – Life Cycle Prozesse und Qualitätsanforderungen – Teil 3: Explainability

[118] Agentur der Europäischen Union für Cybersicherheit, Über ENISA – die Agentur der Europäischen Union für Cybersicherheit, 2022, verfügbar unter: https://www.enisa.europa.eu/about-enisa/about/de (letzter Zugriff: 2022-09-22)

[119] Agentur der Europäischen Union für Cybersicherheit, Securing Machine Learning Algorithms, 2021, verfügbar unter: https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms (letzter Zugriff: 2022-09-22)

[120] Poretschkin, Maximilian; Schmitz, Anna; Akila, Maram; Adilova, Linara; Becker, Daniel; Cremers, Armin B.; Hecker, Dirk; Houben, Sebastian; Mock, Michael; Rosenzweig, Julia; Sicking, Joachim; Schulz, Elena; Voss, Angelika; Wrobel, Stefan, Leitfaden zur Gestaltung vertrauenswürdiger künstlicher Intelligenz (KI-Prüfkatalog), 2021, Sankt Augustin

[121] ISO/IEC AWI 27090, Cybersecurity – Artificial Intelligence – Guidance for addressing security threats and failures in artificial intelligence systems

[122] ISO/IEC 27034-1:2011, Informationstechnik – IT Sicherheitsverfahren – Sicherheit von Anwendungen – Teil 1: Überblick und Konzept, verfügbar unter: https://www.iso.org/standard/44378.html (letzter Zugriff: 2022-09-26)

[123] ISO/IEC 27034-2:2015, Informationstechnik – IT Sicherheitsverfahren – Sicherheit von Anwendungen – Teil 2: Organisation des normativen Rahmen

[124] ISO/IEC 27034-3:2018, Informationstechnik – Sicherheit von Anwendungen – Teil 3: Managementprozess für die Sicherheit von Anwendungen

[125] ISO/IEC 27034-5:2017, Informationstechnik – IT Sicherheitsverfahren – Sicherheit von Anwendungen – Teil 5: Protokolle und Datenstruktur zur Kontrolle der Anwendungssicherheit

[126] ISO/IEC 27034-6:2016, Informationstechnik – IT Sicherheitsverfahren – Sicherheit von Anwendungen – Teil 6: Fallstudien

[127] ISO/IEC 27034-7:2018, Informationstechnik – IT Sicherheitsverfahren – Sicherheit von Anwendungen – Teil 7: Model zur Voraussage der Zusicherung von Sicherheitsanwendungen

[128] DIN EN ISO/IEC 27701:2021, Sicherheitstechniken – Erweiterung zu ISO/IEC 27001 und ISO/IEC 27002 für das Management von Informationen zum Datenschutz – Anforderungen und Leitlinien (ISO/IEC 27701:2019); Deutsche Fassung EN ISO/IEC 27701:2021

[129] Zweites Gesetz zur Anpassung des Datenschutzrechts an die Verordnung (EU) 2016/679 und zur Umsetzung der Richtlinie (EU) 2016/680 (Zweites Datenschutz-Anpassungs- und Umsetzungsgesetz EU – 2. DSAnpUG-EU), 2019, https://www.bgbl.de/xaver/bgbl/media/FEE38353527D8A83E26346A53BE44BD7/bgbl119s1626_77927.pdf (letzter Zugriff: 2022-09-26)

[130] DIN EN ISO/IEC 27037:2016, Informationstechnik – IT-Sicherheitsverfahren – Leitfaden für die Identifikation, Mitnahme, Sicherung und Erhaltung digitaler Beweismittel (ISO/IEC 27037:2012); Deutsche Fassung EN ISO/IEC 27037:2016

[131] DIN EN ISO/IEC 27000er Folge, Fortlaufend ergänzte ISO Standardfolge zum Thema Information Security, Auszug: DIN EN ISO/IEC 27000:2014 Information technology – Security techniques – Information security management systems – Overview and vocabulary, DIN EN ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection – Information security management systems – Requirements, DIN EN ISO/IEC 27002:2022 Information security, cybersecurity and privacy protection – Information security controls

[132] ISO/IEC AWI TS 29119-11, Information technology – Artificial intelligence – Testing for AI systems – Part 11

[133] DIN EN ISO/IEC 29100:2020, Informationstechnik – Sicherheitsverfahren – Rahmenwerk für Datenschutz (ISO/IEC 29100:2011, einschließlich Amd. 1:2018); Deutsche Fassung EN ISO/IEC 29100:2020, verfügbar unter: https://www.beuth.de/de/norm/din-en-iso-iec-29100/325198919 (letzter Zugriff: 2022-09-26)

[134] DIN EN ISO/IEC 29134:2020, Informationstechnik – Sicherheitsverfahren – Leitlinien für die Datenschutz-Folgenabschätzung (ISO/IEC 29134:2017); Deutsche Fassung EN ISO/IEC 29134:2020

[135] DIN EN ISO/IEC 29151:2022, Informationstechnik – Sicherheitsverfahren – Leitfaden für den Schutz personenbezogener Daten (ISO/IEC 29151:2017); Deutsche Fassung EN ISO/IEC 29151:2022

[136] Bitkom Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e. V. Machine Learning und die Transparenzanforderungen der DS-GVO Leitfaden, 2018, https://www.bitkom.org/sites/default/files/file/import/180926-Machine-Learning-und-DSGVO.pdf (letzter Zugriff: 2022-09-26)

[137] ISTQB, Lehrplan zum Certified Tester AI Testing, verfügbar unter: https://www.istqb.org/certifications/artificial-inteligence-tester (letzter Zugriff: 2022-09-26)

[138] ISO/IEC CD TR 27563:2022, Security and privacy in artificial intelligence use cases, verfügbar unter: https://www.iso.org/standard/80396.html (letzter Zugriff: 2022-09-26)

[139] U.S. Food & Drug Administration, 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, 2019, verfügbar unter: https://www.fda.gov/media/122535/download (letzter Zugriff: 2022-08-17)

[140] Francesco Croce, Matthias Hein, Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, 2020, verfügbar unter: https://proceedings.mlr.press/v119/croce20b.html (letzter Zugriff: 2022-09-26)

[141] Mock, M.; Scholz, S.; Blank, F.; Hüger; F., Rohatschek, A.; Schwarz, L.; Stauner, T., SAFECOMP Workshops, Springer, An Integrated Approach to a Safety Argumentation for AI-Based Perception Functions in Automated Driving, 2021

[142] M. Mock, A. Schmitz, Fraunhofer IAIS, Management System Support for Trustworthy Artificial Intelligence, 2021

[143] Lernende Systeme – Die Plattform für Künstliche Intelligenz, Lernende Systeme im Gesundheitswesen. Grundlagen, Anwendungsszenarien und Gestaltungsoptionen, 2019, verfügbar unter: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG6_Lernende_Systeme_im_Gesundheitswesen_web_final.pdf (letzter Zugriff: 2022-09-26)

[144] Beschluss Nr. 768/2008/EG des Europäischen Parlaments und des Rates vom 9. Juli 2008 über einen gemeinsamen Rechtsrahmen für die Vermarktung von Produkten und zur Aufhebung des Beschlusses 93/465/EWG des Rates (Text von Bedeutung für den EWR), verfügbar unter: https://eur-lex.europa.eu/legal-content/de/ALL/?uri=CELEX:32008D0768 (letzter Zugriff: 2022-09-26)

[145] Verordnung (EG) Nr. 765/2008 des Europäischen Parlaments und des Rates vom 9. Juli 2008 über die Vorschriften für die Akkreditierung und Marktüberwachung im Zusammenhang mit der Vermarktung von Produkten und zur Aufhebung der Verordnung (EWG) Nr. 339/93 des Rates (Text von Bedeutung für den EWR)

[146] EU-Parlament, Richtlinie 2006/42/EG des Europäischen Parlaments und des Rates vom 17. Mai 2006 über Maschinen und zur Änderung der Richtlinie 95/16/EG (Neufassung) (Text von Bedeutung für den EWR)

[147] DIN EN ISO/IEC 17000:2020, Konformitätsbewertung – Begriffe und allgemeine Grundlagen (ISO/IEC 17000:2020); Dreisprachige Fassung EN ISO/IEC 17000:2020

[148] ISO/IEC/IEEE 12207:2017, System und Software-Engineering – Prozesse im Lebenszyklus von Software, verfügbar unter: https://www.iso.org/standard/63712.html (letzter Zugriff: 2022-09-26)

[149] ISO/IEC CD 5394:2021, Information Technology – Artificial intelligence – Management System

[150] ISO/IEC WD TS 24462:2022, Ontology for ICT Trustworthiness Assessment

[151] ISO/DIS 24089, Road vehicles – Software update engineering

[152] ISO/IEC 25010:2011, Software-Engineering – Qualitätskriterien und Bewertung von Softwareprodukten (SquaRE) – Qualitätsmodell und Leitlinien, verfügbar unter: https://www.iso.org/standard/35733.html (letzter Zugriff: 2022-09-26)

[153] N. Beck, C. Martens, K.H. Sylla, D. Wegener, A. Zimmermann, ZUKUNFTSSICHERE LÖSUNGEN FÜR MASCHINELLES LERNEN, MACHINE LEARNING OPERATIONS (MLOPS) – PROZESSE FÜR ENTWICKLUNG, INTEGRATION UND BETRIEB, 2021

[154] S. Beck, Plattform lernende Systeme, Künstliche Intelligenz und Diskriminierung. Herausforderungen und Lösungsansätze. Plattform Lernende Systeme (Hrsg.), 2019, verfügbar unter: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_Whitepaper_250619.pdf (letzter Zugriff: 2022-08-24)

[155] DIN EN ISO/IEC 17024:2012, Konformitätsbewertung – Allgemeine Anforderungen an Stellen, die Personen zertifizieren (ISO/IEC 17024:2012); Deutsche und Englische Fassung EN ISO/IEC 17024:2012

[156] DIN EN ISO/IEC 17025:2018, Allgemeine Anforderungen an die Kompetenz von Prüf- und Kalibrierlaboratorien (ISO/IEC 17025:2017); Deutsche und Englische Fassung EN ISO/IEC 17025:2017

[157] DIN EN ISO/IEC 17020:2012, Konformitätsbewertung – Anforderungen an den Betrieb verschiedener Typen von Stellen, die Inspektionen durchführen (ISO/IEC 17020:2012); Deutsche und Englische Fassung EN ISO/IEC 17020:2012

[158] DIN EN ISO/IEC 17029:2020, Konformitätsbewertung – Allgemeine Grundsätze und Anforderungen an Validierungs- und Verifizierungsstellen (ISO/IEC 17029:2019); Deutsche und Englische Fassung EN ISO/IEC 17029:2019

[159] DIN EN ISO/IEC 17011:2018, Konformitätsbewertung – Anforderungen an Akkreditierungsstellen, die Konformitätsbewertungsstellen akkreditieren (ISO/IEC 17011:2017); Deutsche und Englische Fassung EN ISO/IEC 17011:2017

[160] DIN ISO 31000:2018, Risikomanagement – Leitlinien (ISO 31000:2018)

[161] ISO/IEC 27005:2018, Informationstechnik – IT-Sicherheitsverfahren – Informationssicherheits-Risikomanagement

[162] DIN SPEC 92001-1:2019, Künstliche Intelligenz – Life Cycle Prozesse und Qualitätsanforderungen – Teil 1: Qualitäts-Meta-Modell, verfügbar unter: https://www.beuth.de/de/technische-regel/din-spec-92001-1/303650673 (letzter Zugriff: 2022-09-26)

[163] Verordnung (EU) 2019/881 des Europäischen Parlaments und des Rates vom 17. April 2019 über die ENISA (Agentur der Europäischen Union für Cybersicherheit) und über die Zertifizierung der Cybersicherheit von Informations- und Kommunikationstechnik und zur Aufhebung der Verordnung (EU) Nr. 526/2013 (Rechtsakt zur Cybersicherheit) (Text von Bedeutung für den EWR)

[164] Verordnung (EU) Nr. 526/2013 des Europäischen Parlaments und des Rates vom 21. Mai 2013 über die Agentur der Europäischen Union für Netz- und Informationssicherheit (ENISA) und zur Aufhebung der Verordnung (EG) Nr. 460/2004 Text von Bedeutung für den EWR

[165] KI-LOK sicher KI für die Schiene, verfügbar unter: https://ki-lok.itpower.de/ (letzter Zugriff: 2022-09-26)

[166] DSGVO, Datenschutz-Grundverordnung 2016, verfügbar unter: https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=celex%3A32016R0679, (letzter Zugriff: 2022-09-26)

[167] ProdSG, Gesetz über die Bereitstellung von Produkten auf dem Markt (Produktsicherheitsgesetz), verfügbar unter: https://www.gesetze-im-internet.de/prodsg_2021/ (letzter Zugriff: 2022-09-26)

[168] The Fraunhofer Institute for Open Communication Systems FOKUS, Industrial Grade Machine Learning for Enterprises, 2022, verfügbar unter: https://iml4e.org/ (letzter Zugriff: 2022-09-26)

[169] Verordnung (EU) Nr. 1025/2012 des Europäischen Parlaments und des Rates vom 25. Oktober 2012 zur europäischen Normung, zur Änderung der Richtlinien 89/686/EWG und 93/15/EWG des Rates sowie der Richtlinien 94/9/EG, 94/25/EG, 95/16/EG, 97/23/EG, 98/34/EG, 2004/22/EG, 2007/23/EG, 2009/23/EG und 2009/105/EG des Europäischen Parlaments und des Rates und zur Aufhebung des Beschlusses 87/95/EWG des Rates und des Beschlusses Nr. 1673/2006/EG des Europäischen Parlaments und des Rates Text von Bedeutung für den EWR

[170] Horizontal Harmonization Committee, EA-1/06 A-AB:2022, EA Multilateral Agreement. Criteria for signing. Policy and procedures for development, 2022, verfügbar unter: https://european-accreditation.org (letzter Zugriff: 2022-09-26)

[171] International Accreditation Forum, IAF PR4: 2015, Structure of the IAF MLA and List of IAF Endorsed Normative Documents, verfügbar unter: https://european-accreditation.org (letzter Zugriff: 2022-09-26)

[172] Artificial Intelligence and Data Act, verfügbar unter: https://www.osler.com/en/resources/regulations/2022/government-of-canada-s-artificial-intelligence-and-data-act-brief-overview (letzter Zugriff: 2022-09-26)

[173] T. Hagendorff, The Ethics of AI Ethics. An Evaluation of Guidelines, Minds and Machines, p. 122, 2020, verfügbar unter: https://arxiv.org/pdf/1903.03425.pdf (letzter Zugriff: 2022-09-26)

[174] Zweig, K. A., Krafft, T. D., Klingel, A., & Park, E., Sozioinformatik: ein neuer Blick auf Informatik und Gesellschaft. Carl Hanser Verlag GmbH Co KG, 2021

[175] Schlick, Christopher; Bruder, Ralph; Luczak, Holger, Arbeitswissenschaft. Springer-Verlag Berlin Heidelberg. 3. Auflage, 2010

[176] Suchman, L., Human–Machine Reconfigurations. Plans and Situated Actions, 2nd Edition. Cambridge: Cambridge University Press, 2007

[177] Emery, Frederick E./Trist, Eric L., Socio-technical Systems. In: Frederick E. Emery (Hrsg.): Systems Thinking. Harmondsworth, 1969, S. 281–295

[178] Sydow, J., Der soziotechnische Ansatz der Arbeits- und Organisationsgestaltung, 1985

[179] Lee, J.D., Wickens, C.D., Liu, Y. & Ng Boyle, L., Designing for People: An Introduction to Human Factors Engineering. Charlston: CreateSpace, 2017

[180] DIN EN 614-1:2009, Sicherheit von Maschinen – Ergonomische Gestaltungsgrundsätze – Teil 1: Begriffe und allgemeine Leitsätze; Deutsche Fassung EN 614-1:2006+A1:2009

[181] DIN EN 614-2:2008, Sicherheit von Maschinen – Ergonomische Gestaltungsgrundsätze – Teil 2: Wechselwirkungen zwischen der Gestaltung von Maschinen und den Arbeitsaufgaben; Deutsche Fassung EN 614-2:2000+A1:2008

[182] DIN CEN/TR 614-3:2011, Sicherheit von Maschinen – Teil 3: Ergonomische Grundsätze für die Gestaltung von mobilen Maschinen; Deutsche Fassung CEN/TR 614-3:2010

[183] DIN EN ISO 9241-210:2020, Ergonomie der Mensch-System-Interaktion – Teil 210: Menschzentrierte Gestaltung interaktiver Systeme (ISO 9241-210:2019); Deutsche Fassung EN ISO 9241-210:2019, verfügbar unter: https://www.iso.org/standard/77520.html (letzter Zugriff: 2022-09-26)

[184] Raisch S. & Krakowski S., Artificial Intelligence and Management: The Automation-Augmentation Paradox. Academy of Management Review, 2020

[185] Floridi L. & Sanders J., On the Morality of Artificial Agents. Minds and Machines, 14, 349–379, 2004

[186] MAKARIUS, E. E., MUKHERJEE, D., FOX, J. D. & FOX, A. K., Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. Journal of Business Research, 120, 262–273, 2020

[187] ŁAPIŃSKA, J., ESCHER, I., GÓRKA, J., SUDOLSKA, A. & BRZUSTEWICZ, Employees' Trust in Artificial Intelligence in Companies: The Case of Energy and Chemical Industries in Poland. Energies, 14, 1942, 2021

[188] S. Kugele, A. Petrovska, and I. Gerostathopoulos, „Towards a Taxonomy of Autonomous Systems", in Software Architecture, vol. 12857, S. Biffl, E. Navarro, W. Löwe, M. Sirjani, R. Mirandola, and D. Weyns, Eds. Cham: Springer International Publishing, 2021, pp. 37–45. Doi: 10.1007/978-3-030-86044-8_3, 2021

[189] Weyer, Johannes, Die Kooperation menschlicher Akteure und nicht-menschlicher Agenten. Ansatzpunkte einer Soziologie hybride Systeme (= Arbeitspapier 16 der Wirtschafts- und Sozialwissenschaftlichen Fakultät), Dortmund: Technische Universität Dortmund, 2006

[190] Elisabeth André & Wilhelm Bauer, Kompetenzentwicklung für Künstliche Intelligenz – Veränderungen, Bedarfe und Handlungsoptionen, Whitepaper aus der Plattform Lernende Systeme, München, 2021, verfügbar unter: https://doi.org/10.48669/pls_2021-2 (letzter Zugriff: 2022-09-26)

[191] HÖDDINGHAUS, M., SONDERN, D. & HERTEL, G., The automation of leadership functions: Would people trust decision algorithms? Comput. Hum. Behav., 116, 106635, 2021

[192] HUANG, M.-H., RUST, R. & MAKSIMOVIC, V., The Feeling Economy: Managing in the Next Generation of Artificial Intelligence (AI). California Management Review, 61, 43-65, 2019

[193] BRYNJOLFSSON, E., MITCHELL, T. & ROCK, D., What Can Machines Learn, and What Does It Mean for Occupations and the Economy? AEA Papers and Proceedings, 108, 43–47, 2018

[194] Moray, in J. Noyes & M. Bransby (Eds.), People in control. Human factors in control room design (101–115), Human and machines: Allocation of functions, 1989

[195] W. Hacker; P Sachse, Allgemeine Arbeitspsychologie. Psychische Regulation von Tätigkeiten. Göttingen: Hogrefe, 2014

[196] Watzlawick, P., Beavin, J.H., Jackson, D.D., Menschliche Kommunikation: Formen, Störungen, Paradoxien. Bern: Huber, 2011

[197] Cherns A., The Principles of Sociotechnical Design. Human Relations 29 (8), 783–792. DOI: https://doi.org/10.1177/001872677602900806, 1976

[198] Cherns A., Principles of sociotechnical design revisited. Human Relations, 40 (3), 153–161. DOI: https://doi.org/10.1177/001872678704000303, 1987

[199] Ulich, Eberhard, Arbeitssysteme als Soziotechnische Systeme – eine Erinnerung. Journal Psychologie des Alltagshandelns / Psychology of Everyday Activity, Vol. 6 / No. 1, ISSN 1998-9970, S. 4–12, 2013

[200] BENJAMIN, RUHA, Race after Technology. Abolitionist Tools for the New Jim Code. Cambridge/Medford: Polity Press, 2019

[201] Pentenrieder, Annelie; Weber, Jutta, Lucy Suchman (geb. 1951). In: Heßler, Martina & Liggieri, Kevin (Hrsg.). Technikanthropologie: Handbuch für Wissenschaft und Studium. Nomos: Baden-Baden. S. 215–224, 2020

[202] Dr. Norbert Huchler et al., Kriterien für die Mensch-Maschine-Interaktion bei KI. Ansätze für die menschengerechte Gestaltung in der Arbeitswelt, 2020, verfügbar unter: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG2_Whitepaper2_220620.pdf (letzter Zugriff: 2022-09-26)

[203] Sascha Stowasser & Oliver Suchy et al. (Hrsg.), Einführung von KI-Systemen in Unternehmen. Gestaltungsansätze für das Change-Management. Whitepaper aus der Plattform Lernende Systeme, München 2020

[204] Zink, K. J., Soziotechnische Ansätze. In H. Luczak & M. Volpert (Hrsg.), Handbuch Arbeitswissenschaft (S. 74–77). Stuttgart: Schäffer-Poeschel, 1997

[205] Ulich, Eberhard, Arbeitspsychologie, 2011

[206] Bergmann, B. & Richter, P. (Hrsg.), Die Handlungsregulationstheorie: Von der Praxis einer Theorie. Göttingen: Hogrefe, 1994

[207] Bendel A. & Latniak E., Soziotechnisch – agil – lean: Konzepte und Vorgehensweisen für Arbeits- und Organisationsgestaltung in Digitalisierungsprozessen. Gr Interakt Org (2020) 51:285–297, verfügbar unter: https://doi.org/10.1007/s11612-020-00528-8 (letzter Zugriff: 2022-09-26)

[208] KAHNEMAN, D. & TVERSKY, Intuitive prediction: Biases and corrective procedures. In: TVERSKY, A., KAHNEMAN, D. & SLOVIC, P. (eds.) Judgment under Uncertainty: Heuristics and Biases. Cambridge: Cambridge University Press, 1982

[209] WACHTER, S. & MITTELSTADT, B., A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI Columbia Business Law Review, 2, 2019

[210] Shin, D. und Y.J. Park, Role of fairness, accountability, and transparency in algorithmic affordance [online]. Computers in Human Behavior, 98, 277-284. ISSN 07475632, verfügbar unter: http://doi:10.1016/j.chb.2019.04.019, 2019

[211] Ostrom, A.L., D. Fotheringham und M.J. Bitner, Customer Acceptance of AI in Service Encounters: Understanding Antecedents and Consequences. In: P.P. Maglio, C.A. Kieliszewski, J.C. Spohrer, K. Lyons, L. Patrício und Y. Sawatani, Hrsg. Handbook of Service Science, Volume II. Cham: Springer International Publishing, S. 77–103. ISBN 978-3-319-98511-4, 2019

[212] Jeffrey Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, 2018, verfügbar unter: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (letzter Zugriff: 2022-08-12)

[213] Simbeck K., Diskriminiert durch Künstliche Intelligenz – Ethische Aspekte beim Einsatz von analytischen, datengetriebenen Verfahren im Personalmanagement. In: Zukunft der Arbeit. Soziotechnische Gestaltung der Arbeitswelt im Zeichen von „Digitalisierung“ und „Künstlicher Intelligenz“, S. 199–210, 2020

[214] BONNEFON, Jean-François; SHARIFF, Azim; RAHWAN, Iyad., The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. Proceedings of the IEEE, 2019, 107. Jg., Nr. 3, S. 502–504, 2019

[215] Alexander G. Mirnig and Alexander Meschtscherjakov, Trolled by the Trolley Problem: On What Matters for Ethical Decision Making in Automated Vehicles. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI ‘19). Association for Computing Machinery, New York, NY, USA, Paper 509, 1–10. https://doi.org/10.1145/3290605.3300739, 2019

[216] European Parliament and the Council, Directive 2006/42/EC on machinery, and amending Directive 95/16/EC, Mai 2006, verfügbar unter: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:157:0024:0086:en:PDF (letzter Zugriff: 2022-09-26)

[217] RICHTLINIE 2009/127/EG DES EUROPÄISCHEN PARLAMENTS UND DES RATES vom 21. Oktober 2009 zur Änderung der Richtlinie 2006/42/EG betreffend Maschinen zur Ausbringung von Pestiziden, 2009, verfügbar unter: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:310:0029:0033:de:PDF (letzter Zugriff: 2022-09-26)

[218] Deutsche Gesetzliche Unfallversicherung, Industrie 4.0: Herausforderungen für die Prävention – Positionspapier der gesetzlichen Unfallversicherung. Positionspapier 02/2017, verfügbar unter: https://www.dguv.de/medien/inhalt/praevention/arbeitenvierpunktnull/pospap-2-2017.pdf (letzter Zugriff: 2022-08-12)

[219] Karl Ludwig von Bertalanffy, General System Theory. In: Biologia Generalis. 1/1949, S. 114–129

[220] Karl Ludwig von Bertalanffy, The Theory of Open Systems in Physics and Biology. In: Science. Band 111, 1950, S. 23–29

[221] Talcott Parsons, The Social System. Free Press, New York, 1951

[222] Luhmann, Niklas, Soziale Systeme. 1. Auflage. Suhrkamp, Frankfurt am Main 1984, ISBN 3-518-28266-2

[223] Rohde, Friederike; Wagner, Josephin; Reinhard, Philipp; Petschow, Ulrich; Meyer, Andreas; Voß, Marcus; Mollen, Anne, Nachhaltigkeitskriterien für künstliche Intelligenz – Entwicklung eines Kriterien- und Indikatorensets für die Nachhaltigkeitsbewertung von KI-Systemen entlang des Lebenszyklus, 2021, verfügbar unter: https://www.ioew.de/publikation/nachhaltigkeitskriterien_fuer_kuenstliche_intelligenz (letzter Zugriff: 2022-07-12)

[224] Selma Muhammad, The Fairness Handbook, 2022, verfügbar unter: https://www.amsterdamintelligence.com/resources/the-fairness-handbook (letzter Zugriff: 2022-09-26)

[225] Saleiro, P., Rodolfa, K. T., & Ghani, R., Dealing with Bias and Fairness in Data Science Systems: A Practical Hands-on Tutorial, 2020, verfügbar unter: https://doi.org/10.1145/3394486.3406708 (letzter Zugriff: 2022-09-26)

[226] Buolamwini, J., & Gebru, T., Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research, 81, 77–91, 2018, verfügbar unter: http://proceedings.mlr.press/v81/buolamwini18a.html (letzter Zugriff: 2022-09-26)

[227] Silberg, J., & Manyika, J., Notes from the AI frontier: Tackling bias in artificial intelligence (and in humans). Mckinsey Global Institute, 1–8, 2019, verfügbar unter: https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans# (letzter Zugriff: 2022-09-26)

[228] Weerts, H. J. P., An Introduction to Algorithmic Fairness. 1–18, 2021, verfügbar unter: http://arxiv.org/abs/2105.05595 (letzter Zugriff: 2022-09-26)

[229] Simonsen, J., & Robertson, T., Routledge international handbook of participatory design (Vol. 711). New York: Routledge, 2013

[230] Puntschuh, M., & Fetic, L., Praxisleitfaden zu den Algo.Rules. Orientierungshilfen für Entwickler:innen und ihre Führungskräfte. Bertelsmann Stiftung, Gütersloh, 2020, verfügbar unter: https://doi.org/10.11586/2020029 (letzter Zugriff: 2022-09-26)

[231] VDE, Bertelsmann Stiftung (Hrsg.), From Principles to Practice – An interdisciplinary framework to operationalise AI ethics. AI Ethics Impact Group, VDE Association for Electrical Electronic & Information Technologies e. V., Bertelsmann Stiftung, 1–56, 2020, verfügbar unter: https://doi.org/10.11586/2020013 (letzter Zugriff: 2022-09-26)

[232] Puntschuh, M., & Fetic, L., Handreichung für die digitale Verwaltung. Algorithmische Assistenzsysteme gemeinwohlorientiert gestalten. Bertelsmann Stiftung, Gütersloh, 2020, verfügbar unter: https://doi.org/10.11586/2020060 (letzter Zugriff: 2022-09-26)

[233] Mökander, J., Sheth, M., Watson, D. W., & Floridi, L., Models for Classifying AI Systems: the Switch, the Ladder, and the Matrix. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FaccT'22), June 21–24, 2022, Seoul, Republic of Korea. ACM, New York, NY, USA, 1 page, verfügbar unter: https://doi.org/10.1145/3531146.3533162 (letzter Zugriff: 2022-09-26)

[234] IG Metall Vorstand (2019), Ressort Zukunft der Arbeit (Hrsg.), KOMPASS DIGITALISIERUNG. Eine Gestaltungshilfe für gute digitale Arbeit. 1. Auflage 2019

[235] DIN EN ISO 6385:2016, Grundsätze der Ergonomie für die Gestaltung von Arbeitssystemen (ISO 6385:2016); Deutsche Fassung EN ISO 6385:2016

[236] Ulich, Eberhard, Mensch, Technik, Organisation: ein europäisches Produktionskonzept. In O. Strohm & E. Ulich (Hrsg.), Unternehmen arbeitspsychologisch bewerten (S. 5–17). Schriftenreihe Mensch, Technik, Organisation, Band 10 (Hrsg. E. Ulich). Zürich: vdf Hochschulverlag, 1997

[237] Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (Hrsg.), Rechtliche Rahmenbedingungen für die Bereitstellung autonomer und KI-Systeme. Bericht, F 2432. 1. Auflage 2021, verfügbar unter: https://www.baua.de/DE/Angebote/Publikationen/Berichte/F2432.pdf?__blob=publicationFile&v=5 (letzter Zugriff: 2022-09-26)

[238] ISO/IEC WI 12792, Information technology – Artificial intelligence – Transparency taxonomy of AI systems

[239] DIN EN ISO 26800:2011, Ergonomie – Genereller Ansatz, Prinzipien und Konzepte (ISO 26800:2011); Deutsche Fassung EN ISO 26800:2011

[240] DIN SPEC 92001-2:2020, Artificial Intelligence – Life Cycle Processes and Quality Requirements – Part 2: Robustness

[241] VDI/VDE-MT 7100 – Entwurf, Lernförderliche Arbeitsgestaltung – Ziele, Nutzen, Begriffe

[242] VDE SPEC 90012:2022, VCIO based description of systems for AI trustworthiness characterization

[243] DIN EN ISO 11064:2001, Ergonomische Gestaltung von Leitzentralen – Teil 1: Grundsätze für die Gestaltung von Leitzentralen (ISO 11064-1:2000); Deutsche Fassung EN ISO 11064-1:2000

[244] DIN EN 894-1:2009, Sicherheit von Maschinen – Ergonomische Anforderungen an die Gestaltung von Anzeigen und Stellteilen – Teil 1: Allgemeine Leitsätze für Benutzer-Interaktion mit Anzeigen und Stellteilen; Deutsche Fassung EN 894-1:1997+A1:2008

[245] DIN EN ISO 9241-11:2018, Ergonomie der Mensch-System-Interaktion – Teil 11: Gebrauchstauglichkeit: Begriffe und Konzepte (ISO 9241-11:2018); Deutsche Fassung EN ISO 9241-11:2018

[246] DIN EN ISO 9241-110:2020, Ergonomie der Mensch-System-Interaktion – Teil 110: Interaktionsprinzipien (ISO 9241-110:2020); Deutsche Fassung EN ISO 9241-110:2020

[247] ISO/IEC 29138-1:2018, Informationstechnik – Barrierefreie Benutzungsschnittstellen – Teil 1: Barrierefreiheitserfordernisse der Benutzer

[248] VDI/VDE 3850-1:2014, Gebrauchstaugliche Gestaltung von Benutzungsschnittstellen für technische Anlagen – Konzepte, Prinzipien und grundsätzliche Empfehlungen

[249] DIN EN ISO 9241-112:2017, Ergonomie der Mensch-System-Interaktion – Teil 112: Grundsätze der Informationsdarstellung (ISO 9241-112:2017); Deutsche Fassung EN ISO 9241-112:2017

[250] Hacker, Software-Ergonomie: Gestalten rechnergestützter geistiger Arbeit?! In W. Schönpflug & M. Wittstock (Hrsg.), Software-Ergonomie, 87. Nützen Informationssysteme dem Benutzer? (S. 31–54). Stuttgart: Teubner, 1987

[251] Hacker, Software-Gestaltung als Arbeitsgestaltung. In K.-P. Fälmrich (Hrsg.), Software-Ergonomie, State of the Art (S. 29–42). München: Oldenburg, 1987

[252] Böde, Eckard; Hartmann, Ernst A.; Lüdtke, Andreas (u. a.), Mensch-Technik-Interaktion. Leitfaden für Hersteller und Anwender. Band 3. Hrsg.: Bundesministerium für Wirtschaft und Technologie (BMWi), 2013, verfügbar unter: https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/autonomik-Leitfaden3.pdf?__blob=-publicationFile&v=3 (letzter Zugriff: 2022-09-26)

[253] ISO 9241-2, Ergonomische Anforderungen für Bürotätigkeiten mit Bildschirmgeräten – Teil 2: Anforderungen an die Arbeitsaufgaben – Leitsätze

[254] Ulich, Eberhard, Aufgabengestaltung. In H. Schmidt & U. Kleinbeck (Hrsg.), Enzyklopädie der Psychologie (Arbeitspsychologie) (S. 581–622), 2010

[255] Deutsche Gesetzliche Unfallversicherung, Softwareergonomie, 2021

[256] Sheridan, TB, Humans and Automation – System Design and Research Issues. Wiley, New York, 2002

[257] P.M. Fitts, Human engineering for an effective air-navigation and traffic-control system. Washington, DC: National Research Council, 1951

[258] Kraiss KF, Schmidtke H, Funktionsteilung Mensch-Machine. In: Bundesamt für Wehrtechnik und Beschaffung (Hrsg.), Handbuch der Ergonomie. Carl Hanser, München, 2002

[259] Jordan, N., Allocation of functions between man and machines in automated systems. Journal of Applied Psychology 47, 161–165, 1963

[260] Dekker, S., Woods, D., MABA-MABA or Abracadabra? Progress on Human–Automation Coordination . Cognition, Technology & Work 4, 240–244, 2002

[261] P. A. Hancock, Automation: how much is too much? Ergonomics 57, 449–454, 2014

[262] L. Bainbridge, Ironies of Automation. In: Rasmussen J, Duncan K, Leplat J (Hrsg.) New Technology and Human Error, John Wiley, New York, 1987

[263] DIN EN ISO 9001:2015, Qualitätsmanagementsysteme – Anforderungen (ISO 9001:2015); Deutsche und Englische Fassung EN ISO 9001:2015

[264] DIN EN ISO 9000:2015, Qualitätsmanagementsysteme – Grundlagen und Begriffe (ISO 9000:2015); Deutsche und Englische Fassung EN ISO 9000:2015

[265] DIN EN ISO 14001:2015, Umweltmanagementsysteme – Anforderungen mit Anleitung zur Anwendung (ISO 14001:2015); Deutsche und Englische Fassung EN ISO 14001:2015

[266] DIN EN ISO 50001:2018, Energiemanagementsysteme – Anforderungen mit Anleitung zur Anwendung (ISO 50001:2018); Deutsche Fassung EN ISO 50001:2018

[267] DIN ISO 45001:2018, Managementsysteme für Sicherheit und Gesundheit bei der Arbeit – Anforderungen mit Anleitung zur Anwendung (ISO 45001:2018); Text Deutsch und Englisch

[268] DIN ISO 21500:2016, Leitlinien Projektmanagement (ISO 21500:2012)

[269] DIN 69901 (alle Teile), Projektmanagement – Projektmanagementsysteme

[270] DIN 69909 (alle Teile), Multiprojektmanagement – Management von Projektportfolios, Programmen und Projekten

[271] DIN EN ISO 27500:2017, Die menschzentrierte Organisation – Zweck und allgemeine Grundsätze (ISO 27500:2016); Deutsche Fassung EN ISO 27500:2017

[272] „Forum Soziale Technikgestaltung", Projekt „Der mitbestimmte Algorithmus", Projekt PROTIS-BIT, Kriterien zur Gestaltung algorithmischer Steuerungs- und Entscheidungssysteme, Schröter: Welf: Der mitbestimmte Algorithmus. Gestaltungskompetenz für den Wandel der Arbeit, 2019, verfügbar unter: www.blog-zukunft-der-arbeit.de (letzter Zugriff: 2022-09-26)

[273] Jessica Heesen, Jörn Müller-Quade, Stefan Wrobel et al., Kritikalität von KI-Systemen in ihren jeweiligen Anwendungskontexten – Ein notwendiger, aber nicht hinreichender Baustein für Vertrauenswürdigkeit. Whitepaper aus der Plattform Lernende Systeme, München 2021

[274] Herrmann, Thomas, Socio-technical design of hybrid Intelligence systems- the case of predictive maintenance, 2020, verfügbar unter: https://link.springer.com/chapter/10.1007/978-3-030-50334-5_20 (letzter Zugriff: 2022-09-26)

[275] Endsley, Mica R., From Here to Autonomy: Lessons Learned From Human–Automation Research, 2016, verfügbar unter: https://journals.sagepub.com/doi/abs/10.1177/0018720816681350 (letzter Zugriff: 2022-09-26)

[276] Shneiderman, Ben, Human-Centered Artificial Intelligence: Three Fresh Ideas, 2020, verfügbar unter: https://aisel.aisnet.org/thci/vol12/iss3/1/ (letzter Zugriff: 2022-09-26)

[277] Legg, Phil; Smith, Jim; Downing, Alexander, Visual analytics for collaborative human-machine confidence in human-centric active learning tasks, 2019, verfügbar unter: https://hcis-journal.springeropen.com/articles/10.1186/s13673-019-0167-8 (letzter Zugriff: 2022-09-26)

[278] Verband der Chemischen Industrie e. V., Industrieland Deutschland, 2022, verfügbar unter: https://www.vci.de/ergaenzende-downloads/industrieland-deutschland-daten-fakten-bedeutung-deutsche-industrie.pdf (letzer Zugriff: 2022-08)

[279] VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik, VDI-Statusreport Industrie 4.0 Wertschöpfungsketten, 2014, verfügbar unter: https://www.vdi.de/ueber-uns/presse/publikationen/details/industrie-40-wertschoepfungsketten (letzter Zugriff: 2022-09-26)

[280] Positionspapier; Plattform Industrie 4.0, Der Datenraum Industrie 4.0: Die Plattform Industrie 4.0 lädt ein, die digitalen Ökosysteme von morgen zu gestalten, 2021, verfügbar unter: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/PositionPaper-DataSpace.pdf?__blob=publicationFile&v=7 (letzter Zugriff: 2022-08)

[281] Anderl, R.; Bauer, K.; Bauernhansel, T.; Diegner, B.; Diemer, J.;Fay, F.; Goericke, D.; Grotepass, J.; Hilger, C.; Jasperneite, J.; Kalhoff, J.; Jubach, U.; Löwen, U.; Menges, G.; Michels, J.S.; Schmidt, F.; Stiedl, T.; ten Hompel, M.; Zeidler, C., Fortschreibung der Anwendungsszenarien der Plattform Industrie 4.0, November 2016, verfügbar unter: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/fortschreibung-anwendungsszenarien.html (letzter Zugriff: 2022-09-26)

[282] Bundesministerium für Wirtschaft und Technologie (BMWi), Technologieszenario „Künstliche Intelligenz in der Industrie 4.0, 2019, verfügbar unter: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/KI-industrie-40.html (letzter Zugriff: 2022-08)

[283] Geschäftsstelle Plattform Industrie 4.0, Multilaterales Datenteilen in der Industrie: Zielbild am Beispiel des „Collaborative Condition Monitorings“ als Basis für neue Geschäftsmodelle, 2022, verfügbar unter: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/Multilaterales_Datenteilen.pdf?__blob=publicationFile&v=8 (letzter Zugriff: 2022-09-26)

[284] Anderl, R.; Bauernhansel, T.; Broy, M.; Bullinger-Hoffmann, A.; Eckert, C.; Fay, A.; Gausemeier, J.;Hirsch-Kreinsen, H.; Hornung, G.; Lanza, G.; Liggesmeyer, P.; Nebel, W.; Pfeiffer, S.; Piller, F.; Schildhauer, T.; ten Hompel, M.; Wahlster, W.; Bauer, K.; Bauer, W.; Bond, J., Creutz, S.-M..; Fabian, J.-H., Fehring, A.; Frank, U.; Goericke, D.; Hamann, S.; Kubach, W.; Post, P.; Schöning, H.; von Wichert, G., Themenfelder Industrie 4.0: Forschungs- und Entwicklungsbedarfe zur erfolgreichen Umsetzung von Industrie 4.0, 2019, verfügbar unter: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/acatech-themenfelder-industrie-4-0.pdf?__blob=publicationFile&v=12 (letzter Zugriff: 2022-08)

[285] acatech HORIZONTE, KI in der Industrie, 2020, https://www.acatech.de/publikation/acatech-horizonte-ki-in-der-industrie/, (letzter Zugriff: 2022-09-26)

[286] Geschäftsstelle Plattform Industrie 4.0, Mit Normen und Standards Industrie 4.0 gestalten, Mai 2022, verfügbar unter: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/Normen-und-Standards.pdf?__blob=publicationFile&v=4 (letzter Zugriff: 2022-08)

[287] Geschäftsstelle Plattform Industrie 4.0, Industrie 4.0 gestalten. Resilient, nachhaltig, wettbewerbsstark, Mai 2022, verfügbar unter: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/2022-fortschrittsbericht.pdf?__blob=publicationFile&v=14 (letzter Zugriff: 2022-08)

[288] Bundesministerium für Wirtschaft und Technologie (BMWi), KI in der Industrie 4.0: Orientierung, Anwendungsbeispiele, Handlungsempfehlungen, 2019, verfügbar unter: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/ki-in-der-industrie-4-0-orientierung-anwendungsbeispiele-handlungsempfehlungen.html (letzter Zugriff: 2022-08)

[289] Jay Lee, Industrial AI: Applications with Sustainable Performance, 2020

[290] Big Data Value Association, Franco-German Position Paper on „Speeding up Industrial AI and Trustworthiness”, 2021, verfügbar unter: https://www.bdva.eu/speeding-industrial-ai-and-trustworthiness-position-paper-0 (letzter Zugriff: 2022-09)

[291] DIN und DKE, Deutsche Normungsroadmap Industrie 4.0, März 2020, https://www.dke.de/resource/blob/778174/cf0125ab96499cb80621518ca642d818/deutsche-normungs-roadmap-industrie-4-0-version-4-data.pdf (letzter Zugriff: 2022-09-26)

[292] DIN und DKE, Deutsche Normungsroadmap Industrie 4.0 – Fortschrittsbericht, April 2022, verfügbar unter: https://www.din.de/resource/blob/868856/5c6022929e31cfccc1efc5a1eed3e59f/nrm-industrie-4-0-fortschrittsbericht-web-data.pdf (letzter Zugriff: 2022-09-26)

[293] ISO/IEC TR 24030:2021, Information technology – Artificial intelligence (AI) – Use cases

[294] PD IEC TR 63283-2:2022, Industrial-process measurement, control and automation. Smart manufacturing. Use cases, verfügbar unter: https://www.vde-verlag.de/iec-normen/250750/iec-tr-63283-2-2022.html (letzter Zugriff: 2022-09-26)

[295] Plattform Industrie 4.0, Asset Administration Shell Reading Guide, Januar 2022, verfügbar unter: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/AAS-ReadingGuide_202201.pdf?__blob=publicationFile&v=4 (letzter Zugriff: 2022-09-26)

[296] Bundesministerium für Wirtschaft und Klimaschutz (BMWK), Plattform Industrie 4.0, Details of the Asset Administration Shell: Part 1 – The exchange of information between partners in the value chain of Industrie 4.0 (Version 3.0RC02), Mai 2022, verfügbar unter: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/Details_of_the_Asset_Administration_Shell_Part1_V3.html (letzter Zugriff: 2022-08)

[297] ZVEI, AI in Industrial Automation, April 2021, verfügbar unter: https://www.zvei.org/fileadmin/user_upload/Presse_und_Medien/Publikationen/2021/April/AI_in_Industrial_Automation/AI-in-Industrial-Automation-White-Paper-NEU.pdf (letzter Zugriff: 2022-09-26)

[298] Hyde M. Merril et al. IEEE Power & Energy Magazine, PEM pgs. 64 75., Nipping Black outs in the Bud – Introducing a Novel Cascading Failure Network, August 2020, verfügbar unter: https://ieeexplore.ieee.org/document/9120298 (letzter Zugriff: 2022-09-26)

[299] Rodrigo Moreno et al. IEEE Power & Energy Magazine, PEM, From Reliability to Resilience – Planning the Grid against the Extremes, August 2020, verfügbar unter: https://ieeexplore.ieee.org/document/9120304 (letzter Zugriff: 2022-09-26)

[300] Pfrommer, Julius, Usländer, Thomas und Beyerer, Jürgen, KI-Engineering – AI Systems Engineering: Systematic development of AI as part of systems that master complex tasks" at – Automatisierungstechnik, vol. 70, no. 9, 2022, pp. 756-766, 2022, verfügbar unter: https://doi.org/10.1515/auto-2022-0076 (letzter Zugriff: 2022-09)

[301] Patrik Haslum Australian National University, Nir Lipovetzky University of Melbourne, Daniele Magazzeni King's College London, Christian Muise IBM Research, An Introduction to the Planning Domain Definition Language – Synthesis Lectures on Artificial Intelligence and Machine Learning, 2019, verfügbar unter: https://www.morganclaypool.com/doi/abs/10.2200/S00900ED2V01Y201902AIM042 (letzter Zugriff: 2022-09-26)

[302] International Electrotechnical Commission, Semantic interoperability: challenges in the digital transformation age, 2019, verfügbar unter: https://www.iec.ch/basecamp/semantic-interoperability-challenges-digital-transformation-age (letzter Zugriff: 2022-09)

[303] Scientific Data 3, The FAIR Guiding Principles for scientific data management and stewardship, Dezember 2016, verfügbar unter: http://www.nature.com/articles/sdata201618 (letzter Zugriff: 2022-08)

[304] KI für Europa – Eine Europäische KI Strategie, COM(2018)237final, 2018, verfügbar unter: https://eur-lex.europa.eu/legal-content/de/TXT/?uri=CELEX:52018DC0237 (letzter Zugriff: 2022-09-26)

[305] KI-Fachkonferenz von DIN, DKE, Bitkom, VDMA und ZVEI Austausch zum AI Act, 22.11.2021, Fachkonferenz zum Austausch über den AI Act, verfügbar unter: https://www.din.de/de/din-und-seine-partner/presse/mitteilungen/ki-fachkonferenz-von-din-dke-bitkom-vdma-und-zvei–826868 (letzter Zugriff: 2022-06-30)

[306] EU COM(2020)767final – KI für Europa, EU Data Governance Act, 2020

[307] EU COM(2020)825final – DSA – Ensuring a safe and accountable online environment, EU Digital Services Act, 2020

[308] ISO/IEC SC41/WG6 IoT and Digital Twin, WG6 N089 2nd PWI on Guidelines for IoT and Digital Twin Use Cases

[309] VDE DKE/AK931.0.14_2022-003, SemNorm – Ergebnisse des DINCONNECT Projekts Semantische Normen, 2021

[310] DIN IEC 63351:2022-07 – Entwurf, VDE 0491-61:2022-07, Human Factors Engineering (HFE)-Anwendung auf die Auslegung von Mönsch-Maschine-Schnittstellen

[311] Dede, G., Hamon, R., Junklewitz, H., Naydenov, R., Malatras, A. and Sanchez Martin, J.I., Cybersecurity challenges in the uptake of Artificial Intelligence in Autonomous Driving, EUR 30568 EN, 2021

[312] Christian Berghoff, Jona Böddinghaus, Vasilios Danos, Gabrielle Davelaar, Thomas Doms, Heiko Ehrich, Alexandru Forrai, Radu Grosu, Ronan Hamon, Henrik Junklewitz, Matthias Neu, Simon Romanski, Wojciech Samek, Dirk Schlesinger, Jan-Eve Stavesand, Sebastian Steinbach, Arndt von Twickel, Robert Walter, Johannes Weissenböck, Markus Wenzel and Thomas Wiegand, Towards Auditable AI Systems – From Principles to Practice, Whitepaper, 2022, verfügbar unter: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Towards_Auditable_AI_Systems_2022.html (letzter Zugriff: 2022-09-26)

[313] Jeannette M. Wing, Trustworthy AI, 2021, Communications of the ACM, Vol. 64 No. 10, Pages 64-71, 10.1145/3448248

[314] Li, Bo and Qi, Peng and Liu, Bo and Di, Shuai and Liu, Jingen and Pei, Jiquan and Yi, Jinfeng and Zhou, Bowen, Trustworthy AI: From Principles to Practices, 2021, arXiv:2110.01167v2, verfügbar unter: https://arxiv.org/abs/2110.01167v2 (letzter Zugriff: 2022-08-15)

[315] Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi, Trustworthy Artificial Intelligence: A Review, 2023, ACM Comput. Surv. 55, 2, Article 39, verfügbar unter: https://doi.org/10.1145/3491209 (letzter Zugriff: 2022-08-15)

[316] Fredrik Heintz, Michela Milano and Barry O'Sullivan, Trustworthy AI – Integrating Learning, Optimization and Reasoning, 2021, Lecture Notes in Computer Science, Springer Nature Switzerland AG, verfügbar unter: https://doi.org/10.1007/978-3-030-73959-1 (letzter Zugriff: 2022-08-15)

[317] Stanton, B. and Jensen, T., Trust and Artificial Intelligence, 2021, NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD, verfügbar unter: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931087 (letzter Zugriff: 2022-08-15)

[318] Christian Berghoff, Battista Biggio, Elisa Brummel, Vasilios Danos, Thomas Doms, Heiko Ehrich, Thorsten Gantevoort, Barbara Hammer, Joachim Iden, Sven Jacob, Heidy Khlaaf, Lars Komrowski, Robert Kröwing, Jan Hendrik Metzen, Matthias Neu, Fabian Petsch, Maximilian Poretschkin, Wojciech Samek, Hendrik Schäbe, Arndt von Twickel, Martin Vechev and Thomas Wiegand, Towards Auditable AI Systems – Current status and future directions, Whitepaper, 2021, Bonn, Berlin, verfügbar unter: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Towards_Auditable_AI_Systems.pdf (letzter Zugriff: 2022-08-15)

[319] Bundesamt für Sicherheit in der Informationstechnik (BSI), Bundesministerium für Digitales und Verkehr (BMDV), Kraftfahrt-Bundesamt (KBA) und Bundesanstalt für Straßenwesen (BASt), AI-relevant use cases in automotive engineering, 2022, UNECE, 2nd GRVA Workshop on AI and Vehicle Regulations, May 9th 2022, verfügbar unter: https://unece.org/sites/default/files/2022-05/AI-relevant%20use%20cases%20in%20automotive%20engineering.pdf (letzter Zugriff: 2022-08-21)

[320] Berghoff C, Neu M and von Twickel A, Vulnerabilities of Connectionist AI Applications: Evaluation and Defense, 2020, verfügbar unter: https://doi.org/10.3389/fdata.2020.00023 (letzter Zugriff: 2022-09-26)

[321] Yuan Yang, James C Kerce and Faramarz Fekri, LOGICDEF: An Interpretable Defense Framework Against Adversarial Examples via Inductive Scene Graph Reasoning, 2022

[322] Bundesgesetzblatt, Gesetz zur Änderung des Straßenverkehrsgesetzes und des Pflichtversicherungsgesetzes – Gesetz zum Autonomen Fahren, Juli 2021, verfügbar unter: https://www.bgbl.de/xaver/bgbl/start.xav?startbk=-Bundesanzeiger_BGBl&jumpTo=bgbl121s3108.pdf#__bgbl__%2F%2F*%5B%40attr_id%3D%27bgbl121s3108.pdf%27%5D__1661512907226 (letzter Zugriff: 2022-09-26)

[323] Bundesgesetzblatt, Autonome-Fahrzeuge-Genehmigungs-und-Betriebs-Verordnung, Juni 2022, verfügbar unter: https://www.bgbl.de/xaver/bgbl/start.xav#__bgbl__%2F%2F*%5B%40attr_id%3D%27bgbl122s0986.pdf%27%5D__1661513096702 (letzter Zugriff: 2022-09-26)

[324] ISO/SAE 21434:2021, Road vehicles – Cybersecurity engineering

[325] ISO/TR 4804:2020, Road vehicles – Safety and cybersecurity for automated driving systems – Design, verification and validation

[326] ISO/AWI TS 5083, Road vehicles – Safety for automated driving systems – Design, verification and validation
-

[327] ISO 22737:2021, Intelligente Verkehrssysteme – Automatisches Fahrsystem für niedrige Geschwindigkeiten (LSAD) für beschränkte Bereiche – Leistungsanforderungen, Systemanforderungen und Prüfprozeduren

[328] European Union Aviation Safety Agency (EASA), Hrsg., Artificial Intelligence Roadmap 1.0: A human-centric approach to AI in aviation, 2020

[329] European Union Aviation Safety Agency, Concepts of Design Assurance for Neural Networks, März 2020

[330] European Union Aviation Safety Agency, Concepts of Design Assurance for Neural Networks (CoDANN) II, Mai 2020

[331] Bürkle, A., Segor, F. & Kollmann, M. J Intell Robot Syst 61, 339–353, Towards Autonomous Micro UAV Swarms, 2011, verfügbar unter: https://doi.org/10.1007/s10846-010-9492-x (letzter Zugriff: 2022-09-26)

[332] DIN EN 62267:2010-07, VDE 0831-267:2010-07, Bahnanwendungen – Automatischer städtischer schienengebundener Personennahverkehr (AUGT) – Sicherheitsanforderungen (IEC 62267:2009); Deutsche Fassung EN 62267:2009

[333] Dr. Rainer Müller, Automatische U-Bahn für Nürnberg: Voraussetzungen und Realisierungskonzept der VAG, 1999

[334] Railway Gazette International, Rio Tinto to test Rail Vision collision-avoidance technology, 2021

[335] Ristić-Durrant, Danijela, Marten Franke, and Kai Michels. Sensors 21.10 (2021): 3452, A review of vision-based on-board obstacle detection and distance estimation in railways, 2021

[336] ETSI DGS SAI 003, Securing Artificial Intelligence (SAI); Security Testing of AI

[337] Yvonne Prieur, Andreas Sesing-Wagenpfeil, Christian Müller, Datenschutz beim hochautomatisierten Fahren der Zukunft, in: Tagungsband des Internationalen Rechtsinformatik-Symposiums IRIS 2022

[338] Rustam Tagiew, Thomas Buder, Kai Hofmann, Christian Klotz; eb Ausgabe 6-7, „Risikoanalyse der Schnellbremsung bei frontaler Kollisionsgefahr“, 2022

[339] Railway Safety Statistics in the EU, 2022, verfügbar unter: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Railway_safety_statistics_in_the_EU (letzter Zugriff: 2022-08-23)

[340] Common Safety Methods – Risk Assessment oder auch für die europäische Durchführungsverordnung Nr. 402/2013 für Allgemeine Sicherheitsmethoden und Risikobewertung, 2013

[341] Bundesministerium für Bildung und Forschung, Ein Multisensorsystem für die Hinderniserkennung Fahrweg: Abschlussbericht; Forschungsvorhaben Komponenten Automatisierter Schienenverkehr (KOMPAS), Phase 1; Arbeitspaket AP 320 Entwicklung Hinderniserkennung, 2003

[342] SAE:J3016:2021, SAE International, Surface Vehicle Recommended Practice – Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles

[343] DIN VDE V 0831-103:2020-09, Elektrische Bahn-Signalanlagen – Teil 103: Ermittlung von Sicherheitsanforderungen an technische Funktionen in der Eisenbahnsignaltechnik

[344] DIN VDE V 0831-101:2022-08, Elektrische Bahn-Signalanlagen – Teil 101: Semi-quantitative Verfahren zur Risikoanalyse technischer Funktionen in der Eisenbahnsignaltechnik

[345] Plattform Lernende Systeme, Kompetenzentwicklung für Künstliche Intelligenz – Veränderungen, Bedarfe und Handlungsoptionen, 2021, verfügbar unter: https://www.acatech.de/publikation/kompetenzentwicklung-fuer-ki-veraenderungen-bedarfe-und-handlungsoptionen/ (letzter Zugriff: 2022-09-26)

[346] Europäische Kommission – COM(2021) 205 final, Annexes to the Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions – Fostering a European approach to Artificial Intelligence, 2021, verfügbar unter: https://ec.europa.eu/newsroom/dae/redirection/document/75787 (letzter Zugriff: 2022-08-17)

[347] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S., Dermatologist-level classification of skin cancer with deep neural networks, 2017, verfügbar unter: https://pubmed.ncbi.nlm.nih.gov/28117445/ (letzter Zugriff: 2022-09-26)

[348] Muehlematter, UJ, Daniore, P, Vokinger, KN, Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. The Lancet Digital Health 2021; 3: e195–e203, 2021, verfügbar unter: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30292-2 (letzter Zugriff: 2022-09-26)

[349] U.S. Food & Drug Administration, Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices, 2022, verfügbar unter: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (letzter Zugriff: 2022-08-17)

[350] European parliament, 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, 2017, verfügbar unter: https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=uriserv:OJ.L_.2017.117.01.0001.01.DEU&toc=OJ:L:2017:117:FULL (letzter Zugriff: 2022-09-26)

[351] DIN EN ISO 14971:2022, Medizinprodukte – Anwendung des Risikomanagements auf Medizinprodukte (ISO 14971:2019); Deutsche Fassung EN ISO 14971:2019 + A11:2021

[352] ISO/TR 24971:2020, Medizinprodukte – Leitfaden für die Anwendung von ISO 14971

[353] DIN EN 62304:2016, Medizingeräte-Software – Software-Lebenszyklus-Prozesse (IEC 62304:2006 + A1:2015); Deutsche Fassung EN 62304:2006 + Cor.:2008 + A1:2015

[354] DIN EN 82304-1:2018, Gesundheitssoftware – Teil 1: Allgemeine Anforderungen für die Produktsicherheit (IEC 82304-1:2016); Deutsche Fassung EN 82304-1:2017

[355] DIN EN 62366-1:2021, Medizinprodukte – Teil 1: Anwendung der Gebrauchstauglichkeit auf Medizinprodukte (IEC 62366-1:2015 + COR1:2016 + A1:2020); Deutsche Fassung EN 623661:2015 + AC:2015 + A1:2020

[356] IG-NB, Leitfaden Künstliche Intelligenz, Version 03.12.2021, Fragenkatalog „Künstliche Intelligenz bei Medizinprodukten“, 2021

[357] ABNT IEC/TR 62366-2:2021, Medical devices – Part 2: Guidance on the application of usability engineering to medical devices

[358] Interessengemeinschaft der Benannten Stellen für Medizinprodukte in Deutschland (IG-NB), Questionnaire „Artificial Intelligence (AI) in medical devices“ (Version 4, 09.06.2022), verfügbar unter: https://www.ig-nb.de/index.php?eID=dumpFile&t=f&f=2618&token=010db38d577b0bfa3c909d6f1d74b19485e86975 (letzter Zugriff: 2022-08-17)

[359] EU, Vorschlag für eine Verordnung (EU) des Europäischen Parlaments und des Rates über den europäischen Raum für Gesundheitsdaten, 2022, verfügbar unter: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_de (letzter Zugriff: 2022-08-17)

[360] DIN EN ISO 14155:2021, Klinische Prüfung von Medizinprodukten an Menschen – Gute klinische Praxis (ISO 14155:2020); Deutsche Fassung EN ISO 14155:2020

[361] Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, Topol EJ, Ioannidis JPA, Collins GS, Maruthappu M., Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies, 2020, verfügbar unter: BMJ. 2020 Mar 25;368:m689, https://www.bmj.com/content/368/bmj.m689 (letzter Zugriff: 20222022-09-26)

[362] Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group; SPIRIT-AI and CONSORT-AI Steering Group; SPIRIT-AI and CONSORT-AI Consensus Group, Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension, 2020, verfügbar unter: https://pubmed.ncbi.nlm.nih.gov/32908284/ (letzter Zugriff: 2022-09-26)

[363] Liu X, Cruz Rivera S, Moher D, Calvert M, Denniston A, The SPIRIT-AI and CONSORT-AI Working Group, Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension, 2020, verfügbar unter: https://www.nature.com/articles/s41591-020-1034-x (letzter Zugriff: 2022-09-26)

[364] Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L., The medical algorithmic audit, 2022, verfügbar unter: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(22)00003-6/fulltext (letzter Zugriff: 2022-09-26)

[365] Falk Schwendicke, Tatiana Golla, Martin Dreher, Joachim Krois, Convolutional neural networks for dental image diagnostics: A scoping review, 2019, verfügbar unter: J Dent. 2019 Dec;91:103226. Doi: https://doi.org/10.1016/j.jdent.2019.103226 (letzter Zugriff: 2022-08-19)

[366] Taleb A, Rohrer C, Bergner B, De Leon G, Rodrigues JA, Schwendicke F, Lippert C, Krois J., Self-Supervised Learning Methods for Label-Efficient Dental Caries Classification, 2022, verfügbar unter: Diagnostics. 2022; 12(5):1237. https://doi.org/10.3390/diagnostics12051237 (letzter Zugriff: 2022-08-19)

[367] M. Wenzel and T. Wiegand, Toward Global Validation Standards for Health AI, 2020, verfügbar unter: IEEE Communications Standards Magazine, vol. 4, no. 3, pp. 64-69, September 2020; doi 10.1109/MCOMSTD.001.2000006 (letzter Zugriff: 2022-09-26)

[368] Dudgeon SN, Wen S, Hanna MG, Gupta R, Amgad M, Sheth M, Marble H, Huang R, Herrmann MD, Szu CH, Tong D, Werness B, Szu E, Larsimont D, Madabhushi A, Hytopoulos E, Chen W, Singh R, Hart SN, Sharma A, Saltz J, Salgado R, Gallas BD, A Pathologist-Annotated Dataset for Validating Artificial Intelligence: A Project Description and Pilot Study, 2021, verfügbar unter: J Pathol Inform. 2021 Nov 15;12:45; doi: 10.4103/jpi.jpi_83_20 (letzter Zugriff: 2022-09-26)

[369] Guillaume Chassagnon, Maria Vakalopoulou, Enzo Battistella, Stergios Christodoulidis, Trieu-Nghi Hoang-Thi, Severine Dangeard, Eric Deutsch, Fabrice Andre, Enora Guillo, Nara Halm, Stefany El Hajj, Florian Bompard, Sophie Neveu, Chahinez Hani, Ines Saab, Aliénor Campredon, Hasmik Koulakian, Souhail Bennani, Gael Freche, Maxime Barat, Aurelien Lombard, Laure Fournier, Hippolyte Monnier, Téodor Grand, Jules Gregory, Yann Nguyen, Antoine Khalil, Elyas Mahdjoub, Pierre-Yves Brillet, Stéphane Tran Ba, Valérie Bousson, Ahmed Mekki, Robert-Yves Carlier, Marie-Pierre Revel, Nikos Paragios, AI-driven quantification, staging and outcome prediction of COVID-19 pneumonia, 2021, verfügbar unter: Medical Image Analysis, Volume 67, 2021, 101860, ISSN 1361-8415, https://doi.org/10.1016/j.media.2020.101860. (letzter Zugriff: 2022-09-26)

[370] Roman Rischke, Lisa Schneider, Karsten Müller, Wojciech Samek, Falk Schwendicke, Joachim Krois, Federated Learning in Dentistry: Chances and Challenges, 2022, verfügbar unter: https://doi.org/10.1177/00220345221108953 (letzter Zugriff: 2022-09-26)

[371] Ruiyang Ren, Haozhe Luo, Chongying Su, Yang Yao, Wen Liao, Machine learning in dental, oral and craniofacial imaging: a review of recent progress, 2021, verfügbar unter: PeerJ 9:e11451 https://doi.org/10.7717/peerj.11451 (letzter Zugriff: 2022-08-19)

[372] Jose E Cejudo, Akhilanand Chaurasia, Ben Feldberg, Joachim Krois, Falk Schwendicke, Classification of Dental Radiographs Using Deep Learning, 2021, verfügbar unter: https://doi.org/10.3390/jcm10071496 (letzter Zugriff: 2022-09-26)

[373] PD IEC/TR 60601-4-1:2017, Medical electrical equipment – Part 4-1: Guidance and interpretation – Medical electrical equipment and medical electrical systems employing a degree of autonomy

[374] M. Haimerl, Validation of Continuously Learning AI/ML Systems in Medical Devices – A Scenario-based Analysis. Upper Rhine Artificial Intelligence (URAI), 2020

[375] DIN EN 60601-1-10:2021, Medizinische elektrische Geräte – Teil 1-10: Allgemeine Festlegungen für die Sicherheit einschließlich der wesentlichen Leistungsmerkmale – Ergänzungsnorm: Anforderungen an die Entwicklung von physiologischen geschlossenen Regelkreisen (IEC 606011-10:2007 + A1:2013 + A2:2020); Deutsche Fassung EN 60601-1-10:2008 + A1:2015 + A2:2021

[376] A.R. Choudhury, R. Vanguri, S.R. Jambawalikar and P. Kumar, Segmentation of Brain Tumors Using DeepLabv3+, 2019, verfügbar unter: https://link.springer.com/chapter/10.1007/978-3-030-11726-9_14 (letzter Zugriff: 2022-0813)

[377] S. Raina, A. Khandelwal, S. Gupta and A. Leekha, Brain Tumor Segmentation Using Unet, 2021, verfügbar unter: https://link.springer.com/chapter/10.1007/978-981-16-1480-4_39 (letzter Zugriff: 2022-08-13)

[378] G. Neelima, D. R. Chigurukota, B. Maram, B. Giriajan, Optimal DeepMRSeg based tumor segmentation with GAN for brain tumor classification, 2022, verfügbar unter: https://www.sciencedirect.com/science/article/abs/pii/S1746809422000593 (letzter Zugriff: 2022-09-26)

[379] Cohen, J., Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit, 1968, verfügbar unter: https://doi.org/10.1037/h0026256 (letzter Zugriff: 2022-09-26)

[380] Landis, J.R., Koch G.G., The Measurement of Observer Agreement for Categorical Data, 1977, verfügbar unter: https://doi.org/10.3389/fnins.2012.00171 (letzter Zugriff: 2022-09-26)

[381] DIN EN ISO 13485:2021, Medizinprodukte – Qualitätsmanagementsysteme – Anforderungen für regulatorische Zwecke (ISO 13485:2016); Deutsche Fassung EN ISO 13485:2016 + AC:2018 + A11:2021

[382] Leonie Beining, KI in der Industrie absichern & prüfen. Was leisten Assurance Cases?, 2021, verfügbar unter: https://www.stiftung-nv.de/sites/default/files/ki_in_der_industrie_sichern_und_prufen.pdf (letzter Zugriff: 2022-09-26)

[383] Habli I., Alexander R., Hawkins R. D., Safety Cases: An Impending Crisis?, 2021, verfügbar unter: https://eprints.whiterose.ac.uk/169183/ (letzter Zugriff: 2022-09-26)

[384] Marhavilas & Koulouriotis, Risk-Acceptance Criteria in Occupational Health and Safety Risk-Assessment–The State-of-the-Art through a Systematic Literature Review, 2021, verfügbar unter: https://www.mdpi.com/2313-576X/7/4/77/htm (letzter Zugriff: 2022-09-26)

[385] Kläs M., Adler R., Jöckel L., Groß J., Reich J., Using Complementary Risk Acceptance Criteria to Structure Assurance Cases for Safety-Critical AI Components, 2021, verfügbar unter: http://ceur-ws.org/Vol-2916/paper_9.pdf (letzter Zugriff: 2022-09-26)

[386] European commission, Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on digital operational resilience for the financial sector and amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014 and (EU) No 909/2014 COM/2020/595 final, 2020, verfügbar unter: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0595 (letzter Zugriff: 2022-09-26)

[387] Bundesministerium für Justiz, Gesetz über Ordnungswidrigkeiten (OwiG) § 111 Falsche Namensangabe, 2021, verfügbar unter: https://www.gesetze-im-internet.de/owig_1968/__111.html (letzter Zugriff: 2022-09-21)

[388] CASUALTY ACTUARIAL SOCIETY, CAS RESEARCH PAPER SERIES ON RACE AND INSURANCE PRICING UNDERSTANDING POTENTIAL INFLUENCES OF RACIAL BIAS ON P&C INSURANCE: FOUR RATING FACTORS EXPLORED Members of the 2021 CAS Race and Insurance Research Task Force, 2022, https://www.casact.org/sites/default/files/2022-03/Research-Paper_Understanding_Potential_Influences.pdf?utm_source=Website&utm_medium=Press+Release&utm_campaign=RIP+Series (letzter Zugriff: 2022-09-26)

[389] Xi Xing, Fei Huang, Anti-Discrimination Insurance Pricing: Regulations, Fairness Criteria, and Models, 2022, verfügbar unter: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3850420 (letzter Zugriff: 2022-09-26)

[390] Barocas, S.; Hardt, M.; Narayanan, A., Fairness in Machine Learning. Limitations and Opportunities, 2017, verfügbar unter: Fairness in Machine Learning. Limitations and Opportunities. (letzter Zugriff: 2022-09-26)

[391] Hoffmann, Hannah; Vogt, Verena; Hauer, Marc P. et al., Fairness by awareness? On the inclusion of protected features in algorithmic decisions, Preprint 2022, verfügbar unter: Fairness by awareness? (letzter Zugriff: 2022-09-26)

[392] Europäisches Komitee für Standardisierung (CEN), Europäisches Komitee für Elektrotechnische Standardisierung (CENELEC), Focus Group Report: Road Map on Artificial Intelligence (AI), 2020, verfügbar unter: https://ftp.cencenelec.eu/EN/EuropeanStandardization/Sectors/AI/CEN-CLC_FGR_RoadMapAI.pdf (letzter Zugriff: 2022-07-12)

[393] Dwork, Cynthia; Hardt, Moritz; Pitassi, Tonnian et al., Fairness Through Awareness, 2011, verfügbar unter: Fairness Through Awareness. (letzter Zugriff: 2022-09-26)

[394] Kusner M. J., Loftus J. R., Russell C., Silva R., Counterfactual Fairness, 2017, verfügbar unter: https://arxiv.org/abs/1703.06856 (letzter Zugriff: 2022-09-26)

[395] Deutsche Bundesbank, Bundesanstalt für Finanzdienstleistungsaufsicht, Maschinelles Lernen in Risikomodellen – Charakteristika und aufsichtliche Schwerpunkte Konsultationspapier, 2021, verfügbar unter: https://www.bundesbank.de/de/startseite/maschinelles-lernen-in-risikomodellen-charakteristika-und-aufsichtliche-schwerpunkte-670944 (letzter Zugriff: 2022-09-26)

[396] Kenneth Holmberg & Ali Erdemir, Influence of tribology on global energy consumption, costs and emissions, 2017, verfügbar unter: https://doi.org/10.1007/s40544-017-0183-5 (letzter Zugriff: 2022-09-26)

[397] Geibler, Justus von; Gnanko, Toni, Nachhaltige Konsumentscheidungen durch Künstliche Intelligenz und den Digitalen Produktpass – Forschungsbericht zum Roadmapping der Forschungslinie „Transparente Wertschöpfungsketten“ im CO:DINA Projekt, 2022, verfügbar unter: https://codina-transformation.de/forschungsbericht_nachhaltige-konsument-scheidungen-durch-kuenstliche-intelligenz-und-den-digitalen-produktpass/ (letzter Zugriff: 2022-07-12)

[398] Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz, Fünf-Punkte-Programm „Künstliche Intelligenz für Umwelt und Klima“, 2021, verfügbar unter: https://www.bmuv.de/download/fuenf-punkte-programm-kuenstliche-intelligenz-fuer-umwelt-und-klima (letzter Zugriff: 2022-07-12)

[399] United Nations, Sustainable Development Goals (SDG): 17 Goals to Transform Our World, 2019, verfügbar unter: https://www.un.org/sustainabledevelopment/ (letzter Zugriff: 2022-07-12)

[400] Vinuesa, Ricardo; Azizpour, Hossein; Leite, Iolanda; Balaam, Madeline; Dignum, Virginia; Domisch, Sami; Felländer, Anna; Langhans, Simone Daniela; Tegmark, Max; Nerini, Francesco Fuso, The role of artificial intelligence in achieving the Sustainable Development Goals, 2020, verfügbar unter: https://www.nature.com/articles/s41467-019-14108-y (letzter Zugriff: 2022-07-12)

[401] Boll, Susanne; Schnell, Markus; Dowling, Michael; Faisst, Wolfgang; Mordvinova, Olga; Pflaum, Alexander; Rabe, Martin; Veith, Eric; Nieße, Astrid; Gülpen, Christian; Terzidis, Orestis; Riss, Uwe, Mit Künstlicher Intelligenz zu nachhaltigen Geschäftsmodellen – Nachhaltigkeit bon, durch und mit KI. Whitepaper aus der Plattform Lernende Systeme, 2022, verfügbar unter: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG4_WP_KI_und_Nachhaltigkeit.pdf (letzter Zugriff: 2022-07-12)

[402] Europäische Kommission, Europäischer Grüner Deal – Erster klimaneutraler Kontinent werden, 2019, verfügbar unter: https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_de (letzter Zugriff: 2022-07-12)

[403] Gailhofer, Peter; Herold, Anke; Schemmel, Jan Peter; Scherf, Cara-Sophie; Urrutia, Cristina; Köhler, Andreas R.; Braungardt, Sibylle, The role of Artificial Intelligence in the European Green Deal, 2021, verfügbar unter: https://op.europa.eu/en/publication-detail/-/publication/2c3de271-525a-11ec-91ac-01aa75ed71a1 (letzter Zugriff: 2022-07-12)

[404] European Parliament, REGULATION (EU) 2019/2088 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 November 2019 on sustainability-related disclosures in the financial services sector

[405] European Parliament, REGULATION (EU) 2020/852 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 18 June 2020 on the establishment of a framework to facilitate sustainable investment, and amending Regulation (EU) 2019/2088

[406] De Lucia, Caterina; Pazienza, Pasquale; Bartlett, Mark, Does Good ESG Lead to Better Financial Performance by Firms? Machine Learning and Logistic Regression Models of Public Enterprises in Europe, 2020

[407] Umweltbundesamt, Jetzke, Tobias; Richter, Stephan; Ferdinand, Jan-Peter; Schaat, Samer, Künstliche Intelligenz im Umweltbereich – Anwendungsbeispiele und Zukunftsperspektiven im Sinne der Nachhaltigkeit, 2019, verfügbar unter: https://www.umweltbundesamt.de/sites/default/files/medien/1410/publikationen/2019-06-04_texte_56-2019_uba_ki_fin.pdf (letzter Zugriff: 2022-07-12)

[408] Deutsches Institut für Normung, Deutsche Kommission Elektrotechnik Elektronik, Informationstechnik, Verein Deutscher Ingenieure, Normungslandkarte zur Ressourceneffizienz – Beitrag zu ProgRess III von DIN, DKE und VDI, 2021, verfügbar unter: https://www.din.de/resource/blob/797734/48f084aacb96a7970dd16bfcc88bf53c/normungslandkarte-fuer-ressourceneffizienz-data.pdf (letzter Zugriff: 2022-07-12)

[409] Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz, Deutsches Ressourceneffizienzprogramm III – 2020 bis 2023, 2020, verfügbar unter: https://www.bmuv.de/fileadmin/Daten_BMU/Pools/Broschueren/ressourceneffizienz_programm_2020_2023.pdf (letzter Zugriff: 2022-07-12)

[410] DIN EN ISO 14026:2018, Umweltkennzeichnungen und -deklarationen – Grundsätze, Anforderungen und Richtlinien für die Kommunikation von Fußabdruckinformationen (ISO 14026:2017); Deutsche und Englische Fassung EN ISO 14026:2018

[411] DIN EN ISO 14040:202102, Umweltmanagement – Ökobilanz – Grundsätze und Rahmenbedingungen (ISO 14040:2006 + Amd 1:2020); Deutsche Fassung EN ISO 14040:2006 + A1:2020

[412] DIN EN ISO 14044:2021, Umweltmanagement – Ökobilanz – Anforderungen und Anleitungen (ISO 14044:2006 + Amd 1:2017 + Amd 2:2020); Deutsche Fassung EN ISO 14044:2006 + A1:2018 + A2:2020

[413] DIN EN 15804:2022, Nachhaltigkeit von Bauwerken – Umweltproduktdeklarationen – Grundregeln für die Produktkategorie Bauprodukte; Deutsche Fassung EN 15804:2012+A2:2019 + AC:2021

[414] DIN EN ISO 22057:2022, Nachhaltigkeit von Gebäuden und Ingenieurbauwerken – Datenvorlagen für die Verwendung von Umweltproduktdeklarationen (EPDs) für Bauprodukte in der Bauwerksinformationsmodellierung (BIM) (ISO 22057:2022); Deutsche Fassung EN ISO 22057:2022

[415] DIN EN 62559-2:2016-05; VDE 0175-102:2016-05, Anwendungsfallmethodik – Teil 2: Definition der Anwendungsfallvorlage, Akteurliste und der Anforderungsliste

[416] Clauß, John; Finck, Christian; Vogler-Finck, Pierre; Beagon, Paul, Control strategies for building energy systems to unlock demand side flexibility – A review, 2017

[417] Bundesministerium für Umwelt, Naturschutz und nukleare Sicherheit (BMU), Nationales Programm für nachhaltigen Konsum. Gesellschaftlicher Wandel durch einen nachhaltigen Lebensstil, 2019, verfügbar unter: https://nachhaltigerkonsum.info/sites/default/files/medien/dokumente/nachhaltiger_konsum_broschuere_bf.pdf (letzter Zugriff: 2022-08-11)

[418] Gossen, Maike, Jankowski, Patricia, Driving Sustainable Behavior with Persuasive Technology: The Green Consumption Assistant, 2022. Ökologisches Wirtschaften, 2.2022 (37)

[419] Kahl, G.; Herbig, N.; Erdmann, L.; Stadler, K.; Peters, A., Ergebnisdokumentation des Praxisprojekts „Kundenführung am Point of Sale“: Arbeitspapier im Arbeitspaket 4 (AP 4.4) des INNOLAB Projekts. Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI GmbH), Saarbrücken 2017

[420] Altmeyer, Maximilian; Schubhan, Marc; Kerber, Frederic, Automatisieren Personalisieren, Optimieren: Chancen & Herausforderungen von KI-Anwendungen auf Basis des Digitalen Produktpasses im Handel, 2022, verfügbar unter: https://codina-transformation.de/kurzstudie/ (letzter Zugriff: 2022-08-11)

[421] Bundesministerium für Umwelt, Naturschutz und nukleare Sicherheit (BMU), Digitaler Produktpass, 2021, verfügbar unter: https://www.bmu.de/faqs/umweltpolitische-digitalagenda-digitaler-produktpass/ (letzter Zugriff: 2022-08-11)

[422] Götz, Thomas; Berg, Holger; Jansen, Maike; Adisorn, Thomas; Cembrero, David; Markkanen, Sanna; Chowdhury, Tahmid, Digital Product Passport: the ticket to achieving a climate neutral and circular European economy?, 2022, verfügbar unter: https://www.corporateleadersgroup.com/files/cisl_digital_products_passport_report_v6.pdf (letzter Zugriff: 2022-08-11)

[423] Geibler, Justus von; Gnanko, Toni, Künstliche Intelligenz für nachhaltigen Konsum. Ansatzpunkte und Herausforderungen für nachhaltige Konsumentscheidungen auf Basis künstlicher Intelligenz, 2021, verfügbar unter: https://codina-transformation.de/wp-content/uploads/CO-DINA-Positionspapier-7-KI-und-Nachhaltiger-Konsum-1.pdf (letzter Zugriff: 2022-08-11)

[424] Lasarov, Wassili, Nachhaltiger Konsum im digitalen Zeitalter, 2022, In: Bruhn, M., Hadwich K. (Hrsg.), Künstliche Intelligenz im Dienstleistungsmanagement. Springer Fachmedien Wiesbaden GmbH, 235–262

[425] Schneider-Marin, Patricia; Harter, Hannes; Tkachuk, Konstantin; Lang, Werner, Uncertainty Analysis of Embedded Energy and Greenhouse Gas Emissions Using BIM in Early Design Stages, 2020

[426] ISO/IEC 19763-3:2020, Informationstechnik – Metamodell-Rahmenwerk für die Interoperabilität (MFI) – Teil 3: Metamodell für die Registrierung von Ontologien

[427] DIN/TS 92004, Künstliche Intelligenz – Qualitätsanforderungen und -prozesse – Risikoschema für KI-Systeme im gesamten Lebenszyklus

[428] ISO/IEC PWI 7699, Guidance for addressing security threats and failures in artificial intelligence

[429] ISO/IEC 2382:2015, Informationstechnologie – Vokabularien

[430] DIN EN IEC 81001-5-1:2022-01 – Entwurf, VDE 0750-103-5-1:2022-01, Sicherheit, Effektivität und Sicherheit von Gesundheitssoftware und Gesundheits-IT-Systemen – Teil 5-1: Sicherheit – Aktivitäten im Produktlebenszyklus (IEC 62A/1419/CDV:2020); Deutsche und Englische Fassung prEN IEC 81001-5-1:2020

[431] ISO/IEC 20924:2021, Informationstechnik – Internet der Dinge (IoT) – Vokabular

[432] ISO/IEC AWI 42005, Information technology – Artificial intelligence – AI system impact assessment

[433] DIN EN 61508-5:2011-02, VDE 0803-5:2011-02, Funktionale Sicherheit sicherheitsbezogener elektrischer/elektronischer/programmierbarer elektronischer Systeme – Teil 5: Beispiele zur Ermittlung der Stufe der Sicherheitsintegrität (safety integrity level) (IEC 61508-5:2010); Deutsche Fassung EN 61508-5:2010

[434] DIN EN 61511-1:2019-02, VDE 0810-1:2019-02, Funktionale Sicherheit – PLT-Sicherheitseinrichtungen für die Prozessindustrie – Teil 1: Allgemeines, Begriffe, Anforderungen an Systeme, Hardware und Anwendungsprogrammierung (IEC 61511-1:2016 + COR1:2016 + A1:2017); Deutsche Fassung EN 61511-1:2017 + A1:2017

[435] DIN EN IEC 62443 (alle Teile), IT-Sicherheit für industrielle Automatisierungssysteme

[436] ISO/IEC TR 24027:2021, Information technology – Artificial intelligence (AI) – Bias in AI systems and AI aided decision making

[437] ISO/IEC TR 24372:2021, Information technology – Artificial intelligence (AI) – Overview of computational approaches for AI systems

[438] ISO/IEC TR 20547-1:2020, Information technology – Big data reference architecture – Part 1: Framework and application process

[439] ISO/IEC TR 20547-2:2018, Informationstechnik – Big Data Referenzarchitektur – Teil 2: Anwendungsfälle und abgeleitete Anforderungen

[440] ISO/IEC 20547-3:2020, Informationstechnik – Big-Data-Referenzarchitektur – Teil 3: Referenzarchitektur

[441] ISO/IEC 20547-4:2020, Informationstechnik – Big Data Referenzarchitektur – Teil 4: Sicherheit und Datenschutz

[442] ISO/IEC TR 20547-5:2018, Informationstechnik – Big Data Referenzarchitektur – Teil 5: Normungsroadmap

[443] ISO/IEC 20546:2019, Informationstechnik – Big Data – Überblick und Begriffe

[444] ISO/IEC 33063:2015, Informationstechnik – Prozessbewertung – Prozessbewertungsmodell für Software-Tests

[445] DIN EN ISO/IEC 15408-1:2020, Informationstechnik – IT-Sicherheitsverfahren – Evaluationskriterien für IT-Sicherheit – Teil 1: Einführung und allgemeines Modell (ISO/IEC 15408-1:2009); Deutsche Fassung EN ISO/IEC 15408-1:2020

[446] DIN EN ISO/IEC 15408-2:2020, Informationstechnik – IT-Sicherheitsverfahren – Evaluationskriterien für IT-Sicherheit – Teil 2: Sicherheitsfunktionskomponenten (ISO/IEC 154082:2008); Deutsche Fassung EN ISO/IEC 15408-2:2020, nur auf CD-ROM

[447] DIN EN ISO/IEC 15408-3:2020, Informationstechnik – IT-Sicherheitsverfahren – Evaluationskriterien für IT-Sicherheit – Teil 3: Komponenten zur Sicherheitskontrolle (ISO/IEC 15408-3:2008, korrigierte Fassung 2011-06-01); Deutsche Fassung EN ISO/IEC 154083:2020, nur auf CD-ROM

[448] ISO/IEC 15408-4:2022, Informationstechnik – IT-Sicherheitsverfahren – Evaluationskriterien für IT-Sicherheit – Teil 4: Rahmen für die Festlegung von Bewertungsmethoden und -tätigkeiten

[449] ISO/IEC 15408-5:2022, Informationstechnik – IT-Sicherheitsverfahren – Evaluationskriterien für IT-Sicherheit – Teil 5: Vordefinierte Pakete von Sicherheitsanforderungen

[450] ETSI TR 101 583:2015, Methods for Testing and Specification (MTS) – Security Testing – Basic Terminology

[451] DIN EN 61513:2013-09, VDE 0491-2:2013-09, Kernkraftwerke – Leittechnik für Systeme mit sicherheitstechnischer Bedeutung – Allgemeine Systemanforderungen (IEC 61513:2011); Deutsche Fassung EN 61513:2013

[452] DIN EN 50128:2012-03; VDE 0831-128:2012-03, Bahnanwendungen – Telekommunikationstechnik, Signaltechnik und Datenverarbeitungssysteme – Software für Eisenbahnsteuerungs- und Überwachungssysteme; Deutsche Fassung EN 50128:2011

[453] IEEE 7010:2020, A New Standard for Assessing the Well-being Implications of Artificial Intelligence

[454] IEEE P2801:2021, Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence

[455] ISO 26262 (alle Teile), Straßenfahrzeuge – Funktionale Sicherheit

[456] DIN EN 62061:2016, Sicherheit von Maschinen – Funktionale Sicherheit sicherheitsbezogener elektrischer, elektronischer und programmierbarer elektronischer Steuerungssysteme (IEC 62061:2005 + A1:2012 + A2:2015); Deutsche Fassung EN 62061:2005 + Cor.:2010 + A1:2013 + A2:2015

[457] DIN CEN ISO/TR 22100-1:2021, Sicherheit von Maschinen – Beziehung zu ISO 12100 – Teil 1: Wie ISO 12100 und Typ-B- und Typ-C-Normen zusammenhängen (ISO/TR 22100-1:2021); Deutsche Fassung CEN ISO/TR 22100-1:2021

[458] DIN ISO/TR 22100-2, DIN SPEC 33887, Sicherheit von Maschinen – Beziehung zu ISO 12100 – Teil 2: Wie ISO 12100 und ISO 13849-1 zusammenhängen

[459] DIN ISO/TR 22100-3, DIN SPEC 33888, Sicherheit von Maschinen – Beziehung zu ISO 12100 – Teil 3: Implementierung ergonomischer Grundsätze in Sicherheitsnormen

[460] DIN CEN ISO/TR 22100-4:2020, Sicherheit von Maschinen – Zusammenhang mit ISO 12100 – Teil 4: Leitlinien für Maschinenhersteller zur Berücksichtigung der damit verbundenen ITSicherheits- (Cybersicherheits-) Aspekte (ISO/TR 22100-4:2018); Deutsche Fassung CEN ISO/TR 22100-4:2020

[461] ISO/TR 22100-5:2021, Sicherheit von Maschinen – Beziehung zu ISO 12100 – Teil 5: Auswirkungen von maschinellem Lernen mit künstlicher Intelligenz

[462] DIN EN ISO 13849-2:2013, Sicherheit von Maschinen – Sicherheitsbezogene Teile von Steuerungen – Teil 2: Validierung (ISO 13849-2:2012); Deutsche Fassung EN ISO 13849-2:2012

[463] ISO/IEC 25012:2008, Software-Engineering – Qualitätskriterien und Bewertung von Softwareprodukten (SquaRE) – Modell der Datenqualität

[464] ISO/IEC/IEEE 29119-1:2022, Software- und Systemengineering – Software-Test – Teil 1: Allgemeine Konzepte

[465] ISO/IEC/IEEE 29119-2:2021, Software- und Systemengineering – Software-Test – Teil 2: Testprozesse

[466] ISO/IEC/IEEE 29119-3:2021, Software -und Systemengineering – Software-Test – Teil 3: Testdokumentation

[467] ISO/IEC/IEEE 29119-4:2021, Software- und Systemengineering – Software-Test – Teil 4: Testtechniken

[468] ISO/IEC/IEEE 29119-5:2016, Software- und Systemengineering – Software-Test – Teil 5: Keyword-driven Testen

[469] IEEE 1012:2016, Standard for System, Software, and Hardware Verification and Validation

[470] IEEE 3333.1.3:2022, Standard for the Deep Learning-Based Assessment of Visual Experience Based on Human Factors

[471] ANSI/UL 4600:2022, Standard for Safety for the Evaluation of Autonomous Products

[472] ISO/IEC 25000:2014, System und Software-Engineering – Qualitätskriterien und Bewertung von System- und Softwareprodukten (SquaRE) – Leitfaden für SquaRE

[473] ISO/IEC 25024:2015, System und Software-Engineering – Qualitätskriterien und Bewertung von System- und Softwareprodukten (SquaRE) – Messung der Datenqualität

[474] ISO/IEC 25020:2019, Software und System-Engineering – Software- und Systemqualitätsanforderungen und -bewertung (SquaRE) – Qualitätsmessungsrahmen

[475] ISO/IEC 25021:2012, Systems and software engineering – Systems and software Quality Requirements and Evaluation (SquaRE) – Elemente der Qualitätsmessung

[476] DIN SPEC 2343:2020, Übertragung von sprachbasierten Daten zwischen Künstlichen Intelligenzen – Festlegung von Parametern und Formaten

[477] ISO/TS 17033:2019, Ethische Behauptungen und unterstützende Informationen – Grundsätze und Anforderungen

[478] DIN EN ISO 26000:2021, Leitfaden zur gesellschaftlichen Verantwortung (ISO 26000:2010); Deutsche Fassung EN ISO 26000:2020

[479] DIN EN ISO/IEC 27000:2020, Informationstechnik – Sicherheitsverfahren – Informationssicherheitsmanagementsysteme – Überblick und Terminologie

[480] DIN EN ISO/IEC 27001:2017, Informationstechnik – Sicherheitsverfahren – Informationssicherheitsmanagementsysteme – Anforderungen

[481] DIN EN ISO/IEC 27002:2022-08 – Entwurf, Informationssicherheit, Cybersicherheit und Schutz der Privatsphäre – Informationssicherheitsmaßnahmen (ISO/IEC 27002:2022); Deutsche und Englische Fassung prEN ISO/IEC 27002:2022

[482] ITU-T Y gos-ml-arc, Architecture of machine learning based QoS assurance for the IMT-2020 network, 2019 draft

[483] ETSI TS 103 195-2:2018, Autonomic network engineering for the self-managing Future Internet (AFI); Generic Autonomic Network Architecture; Part 2: An Architectural Reference Model for Autonomic Networking, Cognitive Networking and Self-Management, verfügbar unter: https://portal.etsi.org/webapp/WorkProgram/Report_WorkItem.asp?WKI_ID=50970 (letzter Zugriff: 2022-09-26)

[484] DIN EN ISO/IEC 17021-2:2019, Konformitätsbewertung – Anforderungen an Stellen, die Managementsysteme auditieren und zertifizieren – Teil 2: Anforderungen an die Kompetenz für die Auditierung und Zertifizierung von Umweltmanagementsystemen (ISO/IEC 17021-2:2016); Deutsche und Englische Fassung EN ISO/IEC 17021-2:2018

[485] DIN EN ISO/IEC 17021-3:2019, Konformitätsbewertung – Anforderungen an Stellen, die Managementsysteme auditieren und zertifizieren – Teil 3: Anforderungen an die Kompetenz für die Auditierung und Zertifizierung von Qualitätsmanagementsystemen (ISO/IEC 17021-3:2017); Deutsche und Englische Fassung EN ISO/IEC 17021-3:2018

[486] DIN EN ISO/IEC 17030:2021, Konformitätsbewertung – Allgemeine Anforderungen an Konformitätszeichen einer dritten Seite (ISO/IEC 17030:2021); Deutsche und Englische Fassung EN ISO/IEC 17030:2021

[487] DIN EN ISO/IEC 17040:2005, Konformitätsbewertung – Allgemeine Anforderungen an die Begutachtung unter gleichrangigen Konformitätsbewertungsstellen und Akkreditierungsstellen (ISO/IEC 17040:2005); Deutsche und Englische Fassung EN ISO/IEC 17040:2005

[488] DIN EN ISO/IEC 17043:2022, Konformitätsbewertung – Allgemeine Anforderungen an die Kompetenz von Anbietern von Eignungsprüfungen (ISO/IEC DIS 17043:2022); Deutsche und Englische Fassung prEN ISO/IEC 17043:2022

[489] DIN EN ISO/IEC 17050-1:2010, Konformitätsbewertung – Konformitätserklärung von Anbietern – Teil 1: Allgemeine Anforderungen

[490] DIN EN ISO/IEC 17050-2:2005, Konformitätsbewertung – Konformitätserklärung von Anbietern – Teil 2: Unterstützende Dokumentation (ISO/IEC 17050-2:2004); Deutsche und Englische Fassung EN ISO/IEC 17050-2:2004

[491] ITU-T F.AI-DLFE, Deep Learning Software Framework Evaluation Methodology, 2021

[492] ITUT Y.3173, Framework for evaluating intelligence level of future networks including IMT-2020, 2020

[493] DIN EN ISO/IEC 29101:2022, Informationstechnik – Sicherheitstechniken – Architekturrahmenwerk für Datenschutz (ISO/IEC 29101:2018); Deutsche Fassung EN ISO/IEC 29101:2021

[494] DIN EN ISO/IEC 29147:2020, Informationstechnik – Sicherheitstechniken – Offenlegung von Schwachstellen (ISO/IEC 29147:2018); Deutsche Fassung EN ISO/IEC 29147:2020

[495] ITU-T F.AI-DLPB, Metrics and evaluation methods for deep neural network processor benchmark, 2020

[496] ITU-T Y.3170, Requirements for machine learning – based quality of service assurance for the IMT-2020 Network, 2018

[497] ETSI DGR SAI 002, Securing Artificial Intelligence (SAI); Data Supply Chain Report, 2021,

[498] ETSI TS 103 296, Speech and Multimedia Transmission Quality (STQ); Requirements for Emotion Detectors used for Telecommunication Measurement Applications; Detectors for written text and spoken speech, 2016

[499] ETSI GR ENI 004, Experiential Networked Intelligence (ENI); Terminology for Main Concepts in ENI Disclaimer, 2019

[500] ETSI GR NFV 003, Network Functions Virtualisation (NFV); Terminology for Main Concepts in NFV, 2020

[501] ISO/TR 24291:2021, Medizinische Informatik – Anwendungen von Technologien des maschinellen Lernens für die künstliche Intelligenz in der Medizin

[502] ISO/TR 3985:2021, Biotechnologie – Datenveröffentlichung – Vorüberlegungen und Konzepte

[503] ISO/TS 22756:2020, Medizinische Informatik – Anforderungen an eine Wissensbasis für medizinische Entscheidungsunterstützungssysteme von medikationsbezogenen Prozessen

[504] ITU-T F.VS-AIMC, Use cases and requirements for multimedia communication enabled vehicle systems using artificial intelligence, 2021

[505] DIN SPEC 91426:202012, Qualitätsanforderungen für video-gestützte Methoden der Personalauswahl (VMP)

[506] DIN SPEC 13288:2021, Leitfaden für die Entwicklung von Deep-Learning-Bilderkennungssystemen in der Medizin; Text Deutsch und Englisch

[507] ISO/TS 5346:2022, Medizinische Informatik – Kategoriale Struktur zur Darstellung des klinischen Entscheidungsunterstützungssystems der Traditionellen Chinesischen Medizin

[508] DIN EN ISO 11073 Reihe, Medizinische Informatik – Kommunikation von Geräten für die persönliche Gesundheit

[509] DIN CEN ISO/TS 22703:2022, Medizinische Informatik – Anforderungen an Arzneimittel-Warnmeldungen (ISO/TS 22703:2021); Deutsche Fassung CEN ISO/TS 22703:2021

[510] ISO/TR 19669:2017, Medizinische Informatik – Wiederverwendbare Komponenten-Strategie für die Entwicklung von Use-Cases

[511] IEEE P2802:2022, Standard for the Performance and Safety Evaluation of Artificial Intelligence Based Medical Device: Terminology

[512] VDI-MT 7001:2021, Kommunikation und Öffentlichkeitsbeteiligung bei Bau- und Infrastrukturprojekten – Standards für die Leistungsphasen der Ingenieure

[513] DIN EN ISO 10075 (alle Teile), Ergonomische Grundlagen bezüglich psychischer Arbeitsbelastung

[514] DIN EN ISO 9241 (alle Teile), Ergonomie der Mensch-System-Interaktion

[515] DIN EN 894 (alle Teile), Sicherheit von Maschinen – Ergonomische Anforderungen an die Gestaltung von Anzeigen und Stellteilen

[516] DIN EN 16710-2:2016, Verfahren der Ergonomie – Teil 2: Eine Methode für die Arbeitsanalyse zur Unterstützung von Entwicklung und Design; Deutsche Fassung EN 16710-2:2016 Verfahren der Ergonomie (Feedbackmethode zur Arbeitsweise von Menschen mit Maschinen, Arbeitsanalyse zur Unterstützung von Entwicklung und Design)

[517] DIN EN ISO 12100:2011, Sicherheit von Maschinen – Allgemeine Gestaltungsleitsätze – Risikobeurteilung und Risikominderung (ISO 12100:2010); Deutsche Fassung EN ISO 12100:2010

[518] ISO/TR 16982:2002, Ergonomie der Mensch-System-Interaktion – Methoden zur Gewährleistung der Gebrauchstauglichkeit, die eine benutzer-orientierte Gestaltung unterstützen

[519] ISO 21930:2017, Nachhaltigkeit von Bauwerken – Grundregeln für die Umweltdeklaration von in Bauwerken verwendeten Bauprodukten und technischen Anlagen

[520] DIN IEC/TS 62998-1:2021-10, VDE V 0113-998-1:2021, Sicherheit von Maschinen – Sicherheitsrelevante Sensoren für den Schutz von Personen (IEC TS 62998-1:2019)

[521] ISO/TS 14048:2002, Umweltmanagement – Ökobilanz – Datendokumentationsformat

[522] CWA 17284:2018, Materials modelling – Terminology, classification and metadata

[523] CWA 17815:2021, Materials characterization – Terminology, metadata and classification

[524] Geibler, Justus von; Riera, Nuria; Echternacht, Laura; Björling, Sten-Eric.; Domen, Tom et al., myEcoCost. Forming the Nucleus of a Novel Environmental Accounting System: Vision, prototype and way forward, 2015, Wuppertal Institute for Climate, Environment and Energy, Wuppertal, verfügbar unter: https://epub.wupperinst.org/frontdoor/index/index/docId/6009 (letzter Zugriff: 2022-07-12)

11 Autorenverzeichnis

Dr.-Ing. Rasmus Adler, Fraunhofer-Institut für Experimentelles Software Engineering (IESE)

Araceli Alcala, Carl Zeiss Meditec AG

Marie Anton, Bundesverband der Arzneimittel-Hersteller e. V. (BAH)

Tristan Armbruster, PricewaterhouseCoopers GmbH

Stefan Arntzen, Huawei Technologies Düsseldorf GmbH

Prof. Dr. Doris Aschenbrenner, Hochschule Aalen

Klaus-Dieter Axt, EUnited (European Engineering Industries Association)

Dr. Renate Baumgartner, Eberhard Karls Universität Tübingen

Nikolas Becker, Gesellschaft für Informatik e. V. (GI)

Rebecca Beiter, Cyber Valley

Thomas Bendig, adesso SE

Ralf Benecke, sonnen GmbH

Dr. Philipp Benner, Bundesanstalt für Materialforschung und -prüfung (BAM)

Bastian Bernhardt, IABG mbH

Paul Beyer, FSD Fahrzeugsystemdaten GmbH

Karsten Bich, Deutsches Institut für Normung e. V. (DIN)

Jan Biehler, Plattform Lernende Systeme / acatech

Lukas Bieringer, QuantPi GmbH

Dr. Andreas Binder, SAMSON Pilotentwicklung GmbH

Dr. Sylwia Birska, BG ETEM

André Bluhm, ai.dopt GmbH

John Böhm, T-Systems Multimedia Solutions GmbH

Dr. Jürgen Bohn, Schaeffler AG

Dr.-Ing. Patrick Bollgrün, Plattform Lernende Systeme / acatech

Dr.-Ing. Mikko Börkircher, Verband der Metall- und Elektroindustrie Nordrhein-Westfalen e. V. (METALL NRW)

Kevin Borowski, HASPA

Oliver Bracht, eoda GmbH

Matthias Brand, MBDA Deutschland GmbH

Katharina Buchsbaum, Deutsches Forschungsinstitut für öffentliche Verwaltung (FÖV)

Lena Marie Budde, Bund für Umwelt und Naturschutz Deutschland (BUND)

Dr. Joachim Bühler, TÜV-Verband

Dr. Andreas Bunte, Fraunhofer IOSB-INA (IOSB-INA)

Dr. Aljoscha Burchardt, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

Prof. Dr. Simon Burton, Fraunhofer-Institut für Kognitive Systeme (IKS)

Tim Büttel, TÜV Nord Mobilität GmbH & Co. KG IFM

Ulla Coester, xethix Empowerment

Prof. Dr. Armin B. Cremers, b-it Emeritus Research Group AI Foundations, Universität Bonn

Damian A. Czarny, Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE (DKE)

Lucas da Silva, Ingrano Solutions GmbH

Dr. David Dang, Deloitte GmbH

Klaus Däßler, Gesellschaft für Mathematische Intelligenz (GMI)

Jan de Meer, Hochschule für Technik und Wirtschaft Berlin (HTW)

Axel Demel, qdive GmbH

Dr. Peter Deussen, Microsoft Deutschland GmbH

Ernestine Dickhaut, Universität Kassel

Eckhard Dittrich, Privatperson

Alexander Dittrich, Deloitte GmbH

Lilian Do Khac, Philipps-Universität Marburg und adesso SE

Felix Dotzauer, SPECTARIS – Deutscher Industrieverband für Optik, Photonik, Analysen- und Medizintechnik e. V.

Gilbert Drzyzga, Technische Hochschule Lübeck

Prof. Dr. Martin Ebers, Robotics & AI Law Society (RAILS)

Filiz Elmas, Deutsches Institut für Normung e. V. (DIN)

Dr. Stefan Elmer, Festo SE & Co. KG

Dr.-Ing. Marko Esche, Physikalisch-Technische Bundesanstalt (PTB)

Benjamin Fehlandt, SALT AND PEPPER Technology

Leander Féret, JUMO GmbH & Co. KG

Lajla Fetic, Bertelsmann Stiftung

Marc Fliehe, TÜV-Verband e. V.

Dr. Julia Fligge-Niebling, Deutsches Zentrum für Luft- und Raumfahrt e. V. (DLR)

Werner Flögel, GEMÜ Gebr. Müller Apparatebau GmbH & Co. KG

Christopher Frank, Deutsche Gesetzliche Unfallversicherung (DGUV)

Matthias Frank, Brose Fahrzeugteile SE & Co. KG

Annika Franken, Forschungsinstitut für Rationalisierung (FIR) e. V. an der RWTH Aachen

Prof. Dr. Martin Fränzle, Universität Oldenburg

Christian Fraunholz, Fraunholz Technologies UG

Charlotte Frierson, Forschungsinstitut für Rationalisierung (FIR) e. V. an der RWTH Aachen

Florian Gauer, PricewaterhouseCoopers GmbH

Prof. Dr. Clemens Gause, Verband für Sicherheitstechnik e. V. (VFS)

Antoine Gautier, QuantPi GmbH

Dr. Marc Gebauer, Brandenburgische Technische Universität Cottbus-Senftenberg

Dr. Bernd Geiger, semafora systems GmbH

Dr. Sergio Genovesi, Universität Bonn

Dr. Detlef Gerst, IG Metall

Simon Geschwill, Schwarz Dienstleistung KG

Prof. Dr. Dagmar Gesmann-Nuissl, Technische Universität Chemnitz (TUC)

Nora Helena Glasmeier, Bundesverband der Deutschen Volksbanken und Raiffeisenbanken e. V. (BVR)

Dr. Ludwig Glatzner, Büro für Umwelt, Qualität, Sicherheit

Jens Gnaudschun, TÜV Nord Mobilität GmbH & Co. KG

Marius Goebel, Spherity GmbH

Dominik Grau, Beuth Verlag

Viacheslav Gromov, AITAD GmbH

Dr.-Ing. Jürgen Großmann, Fraunhofer-Institut für Offene Kommunikationssysteme (FOKUS)

Prof. Dr. Jürgen Grotepass, Huawei Technologies Düsseldorf GmbH

Yvonne Gruchmann, Wirtschaftsförderung Land Brandenburg GmbH

Norman Günther, Technische Hochschule Wildau (TH Wildau)

Prof. Dr. Martin Haimerl, Hochschule Furtwangen (HFU)

Christian Hattenkofer, Bank-Verlag GmbH

Marc Hauer, Algorithm Accountability Lab der TU Kaiserslautern (AAL TUK)

Prof. Dr. Stefan Haufe, Physikalisch-Technische Bundesanstalt (PTB) und Technische Universität Berlin

Elias Heider, MARELLI

Jürgen Heiles, Siemens AG

Dr. Tobias Heimann, Siemens Healthineers

Tabea Hein, Stadtverwaltung Frankfurt am Main

Claudia Heinemann, Selbständige Beraterin

Christoph Henseler, Deutsches Institut für Gutes Leben GmbH (difgl)

Dr. Wolfgang Hildesheim, IBM Deutschland GmbH

Barbara Hilgert, Fortbildungsakademie der Wirtschaft (FAW) gemeinnützige Gesellschaft mbH

Dr. Lukas Höhndorf, IABG mbH

Taras Holoyad, Bundesnetzagentur

Dr. Maximilian Hösl, Plattform Lernende Systeme / acatech

Alexander Jaschke, Fraunhofer-Institut für Integrierte Schaltung (IIS)

Dr. Barbara Jung, Physikalisch-Technische Bundesanstalt (PTB)

Dr. Vanessa Just, KI Bundesverband e. V.

Agnieszka Kacyniak, consileo GmbH & Co. KG

Thomas Kaiser, Kodex AI GmbH

Dr. Leo Kärkkäinen, Huawei Technologies Deutschland GmbH

Jan Kiefer, Bundesanstalt für Finanzdienstleistungsaufsicht (BaFin)

So-Jin Kim, Deutsches Institut für Normung e. V. (DIN)

Roland Kirsch, Bundesamt für Sicherheit in der Informationstechnik (BSI)

Nils-Olaf Klabunde, 4PL Intermodal GmbH

Dr. Michael Kläs, Fraunhofer-Institut für Experimentelles Software Engineering (IESE)

Philip Kleen, Fraunhofer-Institutsteil für industrielle Automation (INA) des Fraunhofer IOSB

Prof. Dr. Annette Kleinfeld, Hochschule für Technik, Wirtschaft und Gestaltung (HTWG) Konstanz

Anita Klingel, PD – Berater der öffentlichen Hand

Dr. habil. Jürgen Klippert, IG Metall

Julia Kloiber, Superrr Lab

Mirko Knaak, IAV GmbH

David Knauer, T-Systems Multimedia Solutions GmbH

Ricardo Knauer, Hochschule für Technik und Wirtschaft Berlin

Andrea Knaut, Institut für Sozialarbeit und Sozialpädagogik e. V./ Geschäftsstelle Dritter Gleichstellungsbericht der Bundesregierung

Franz Knecht, Connexis AG

Mario Knicker, SHARP Electronics GmbH

Marco Knödler, YNCORIS GmbH & Co. KG / Interessensgemeinschaft Regelwerke Technik e. V. (IGR)

Harry Knopf, High Knowledge GmbH

Dr. Martin F. Köhler, Rechtsanwalt

Christian Kolf, TÜV AI Lab

Dr. Georgios Kolliarakis, Deutsche Gesellschaft für Auswärtige Politik (DGAP)

Christopher Koska, dimension2 economics & philosophy consult GmbH

Sebastian Kosslers, Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE (DKE)

Roland Kossow, CyberTribe® – Das dezentrale Systemhaus

Dr. Stefan Kothe, Physikalisch-Technische Bundesanstalt (PTB)

Sebastian Kotte, neurocat GmbH

Tobias Krafft, Trusted AI GmbH

Prof. Dr.-Ing. Klaus Kratzer, Technische Hochschule Ulm

Dr. Tom Kraus, VDI/VDE Innovation und Technik GmbH

Prof. Markus Krebsz, UNECE GRM & The Human-Ai.Institute

Tim Kremer, Deutscher Sparkassen und Giroverband e. V.

Sebastian Kriegsmann, Deutsches Institut für Normung e. V. (DIN)

Mirco Kröll, Bundesanstalt für Materialforschung und -prüfung (BAM)

Prof. Dr. Antonio Krüger, Deutsches Forschungszentrum für Künstliche Inteligenz (DFKI)

Katja Krüger, Deutsches Institut für Normung e. V. (DIN)

Jacques Kruse Brandao, SGS

Dr. Tanja Kubes, Freie Universität Berlin

Susanne Kuch, Deutsche Akkreditierungsstelle GmbH (DAkkS)

Stefan Kunkel, sagena Innovationsgesellschaft mbH

Benjamin Küttner, Deutsche Bundesbank

Dr. Jens F. Lachenmaier, Universität Stuttgart

Holger Laible, Siemens AG

Joel Lakermann, TÜV Nord Mobilität GmbH & Co. KG

Fredi Lang, Berufsverband Deutscher Psychologinnen und Psychologen e. V.

Dr. Erich Latniak, Universität Duisburg-Essen

Elisa Lederer, PricewaterhouseCoopers GmbH

Dr.-Ing. Christoph Legat, HEKUMA GmbH

Lorenz Lehmhaus, Aleph Alpha GmbH

Dr. Mahei Manhai Li, Universität Kassel

Michael Lipka, Huawei Technologies Deutschland GmbH

Daniel Loevenich, Bundesamt für Sicherheit in der Informationstechnik (BSI)

PD Dr. habil. Jeanette Miriam Lorenz, Fraunhofer-Institut für Kognitive Systeme (IKS)

Prof. Dr. Ulrich Löwen, Siemens AG

Dr. Jackie Ma, Fraunhofer-Institut für Nachrichtentechnik, Heinrich-Hertz-Institut (HHI)

Dr.-Ing. Stefan Maack, Bundesanstalt für Materialforschung und -prüfung (BAM)

Manuela Mackert, Privatperson

Sabine Mahr, word b sign Sabine Mahr

Maximilian Margreiter, Deloitte GmbH

Karla Markert, Fraunhofer-Institut für Angewandte und Integrierte Sicherheit (AISEC)

Dr. Oliver Maspfuhl, Deutsche Bank AG

Isabel Matthias, Universität Bremen

Gerd Matzke, Drägerwerk AG & Co. KGaA

Henri Meeß, Fraunhofer-Institut für Verkehrs- und Infrastruktursysteme (IVI)

Iris Merget, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

Ralf Meschede, Bundesanstalt für Straßenwesen (BASt)

Martin Meyer, Siemens Healthineers

Olga Meyer, Fraunhofer-Institut für Produktionstechnik und Automatisierung (IPA)

Dr.-Ing. Sascha Meyne, Physikalisch-Technische Bundesanstalt (PTB)

Alexander Mihatsch, Plattform Lernende Systeme / acatech

Dr. Alexander G. Mirnig, Paris Lodron Universität Salzburg & AIT Austrian Institute of Technology GmbH

Prof. Dr. Andreas Mockenhaupt, Hochschule Albstadt-Sigmaringen

Dr.-Ing. Eike Möhlmann, Deutsches Zentrum für Luft- und Raumfahrt e. V. (DLR)

Michael Mörike, Integrata-Stiftung

Andreas Müller, Schaeffler AG

Dr. Christian Müller, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

Tomislav Nad, SGS

Gert Nahler, Samson AG

Dr. Andreas Nawroth, Münchener Rückversicherungs-Gesellschaft (Munich RE)

Jens Neuhüttler, Fraunhofer-Institut für Arbeitswirtschaft und Organisation (IAO)

Dr. Matthias Neumann-Brosig, IAV GmbH

Dr. Marc Neveling, Deloitte GmbH

Dr. Peter Nickel, Institut für Arbeitsschutz der Deutschen Gesetzlichen Unfallversicherung (IFA)

Jürgen Niehaus, SafeTRANS

Reimund Nienaber, EDLIGO

Johannes Nöbel, KPMG AG Wirtschaftsprüfungsgesellschaft

Dr. Antje Nowack, Verband der Vereine Creditreform e. V.

Dr. Shane O'Sullivan, Universidade de São Paulo

Otto Obert, Main DigitalEthiker GmbH

Dr. Ursula Ohliger, Plattform Lernende Systeme / acatech

Rebecca Page, Endress+Hauser Flowtec AG

Dr. Jochen Papenbrock, NVIDIA GmbH

Ludwig Pechmann, UniTransferKlinik Lübeck GmbH

Yannick Peifer, Institut für angewandte Arbeitswissenschaft e. V. (ifaa)

Robin H. Pekerman, swiss4digital

Dr. Annelie Pentenrieder, Institut für Innovation und Technik Berlin

Dr. Christoph Peters, Universität Kassel – ITeG

Fabian Petsch, Bundesamt für Sicherheit in der Informationstechnik (BSI)

Katharina Petschick, WBS GRUPPE

Dr. Christoph Peylo, Robert Bosch GmbH

Daniel Pflumm, TÜV-Verband e. V.

Patrick Philipp, Fraunhofer Institut für Optronik, Systemtechnik und Bildauswertung (IOSB)

Dr. Henrich C. Pöhls, Universität Passau

Frank Poignée, infoteam Software AG

Dr. Maximilian Poretschkin, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS)

Martin Portier, Bundesamt für Seeschifffahrt und Hydrographie

Dr.-Ing. Jens Prager, Bundesanstalt für Materialforschung und -prüfung (BAM)

Dr. Henrik J. Putzer, fortiss

Alexander Rabe, eco – Verband der Internetwirtschaft e. V.

Peter Rauh, Deutsches Institut für Normung e. V. (DIN)

Hendrik A. Reese, PricewaterhouseCoopers GmbH

Prof. Dr. Georg Rehm, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

Dr. Claudia Reinel, Deutsches Institut für Normung e. V. (DIN)

Axel Rennoch, Fraunhofer-Institut für Offene Kommunikationssysteme (FOKUS)

Klaus Roleff, Wintegral GmbH

Karsten Roscher, Fraunhofer-Institut für Kognitive Systeme (IKS)

Michael Rosenthal, regio iT gesellschaft für informationstechnologie mbh

Jan Rösler, Deutsches Institut für Normung e. V. (DIN)

Nils Röttger, imbus AG

Dr. Gerhard Runze, imbus AG

Martin Ruskowski, DFKI

Dr. Martin Saerbeck, TÜV SÜD

Peter K. Sanner, areasix GmbH

Ingo Sawilla, TRUMPF Werkzeugmaschinen SE + Co.KG

Dr.-Ing. Mario Schacht, Deutsches Institut für Normung e. V. (DIN)

Dr. Peter Schemel, Deloitte GmbH

Kim Marvin Scheurenbrand, Deloitte GmbH

Maximilian Schildt, Lehrstuhl für Energieeffizientes Bauen (E3D) RWTH Aachen University

David Schirgi, Siemens AG

Dr.-Ing. Miriam Schleipen, EKS InTec GmbH

Dr. Dirk Schlesinger, TÜV AI Lab

Nadine Schlicker, Institut für künstliche Intelligenz in der Medizin, Universitätsklinikum Gießen und Marburg GmbH, Philipps-Universität Marburg

Jun.-Prof. Dr. Thomas Schmid, Martin-Luther-Universität Halle-Wittenberg

Michael-Christian Schmidt, ESKITEC GmbH

Dr. Thomas Schmidt, acatech-Deutsche Akademie der Technikwissenschaften

Christoph Schmidt, Deutsches Institut für Normung e. V. (DIN)

Jörg Schmidtke, VIVAVIS AG

Frank Schmiedchen, Vereinigung Deutscher Wissenschaftler e. V. (VDW)

Thorsten Schmitz, EKS InTec GmbH

Jonas Schneider, EFS – Elektronische Fahrwerksysteme GmbH

Detlef Schoepe, Zentrum für Digitalisierung Bundeswehr – Kompetenzzentrum KI

Prof. Dr. Wolfgang M. Schröder, Julius-Maximilians-Universität Würzburg (JMU)

Welf Schröter, Forum Soziale Technikgestaltung

Dr. med. Stephan Schug, Deutsche Gesellschaft für Gesundheitstelematik (DGG) e. V.

Tim Schüßler, Amprion GmbH, Universität Siegen Lehrstuhl Embedded Systems

Jan Fiete Schütte, dimension2 economics & philosophy GmbH

Prof. Dr. Falk Schwendicke, Charité – Universitätsmedizin Berlin

Adrian Seeliger, Deutsches Institut für Normung e. V. (DIN)

Prof. Dr. Eberhard K. Seifert, VDW-Studiengruppe Digitalisierung, Senior Fellow IASS-Potsdam

Jan Seitz, Technische Hochschule Wildau (TH Wildau)

Annegrit Seyerlein-Klug, neurocat GmbH

Fatemeh Shahinfar, Institut für angewandte Arbeitswissenschaft e. V. (ifaa)

Prof. Dr. Katharina Simbeck, Hochschule für Technik und Wirtschaft Berlin (HTW)

Andreas Skuin, Orban Consulting Holding GmbH

Ariana Sliwa, TÜV AI Lab

Dr. Reiner Spallek, IABG mbH

Dr. Felix Spangenberg, msg systems ag

Philip Sperl, Fraunhofer-Institut für Angewandte und Integrierte Sicherheit (AISEC)

Patrick Spitzer, Deloitte GmbH

Lucas Spreiter, Unetiq GmbH

Prof. Dr. André Steimers, Hochschule Koblenz, Institut für Arbeitsschutz der DGUV

Rosmarie Steininger, CHEMISTREE GmbH

Jannis Steinke, Technische Universität Braunschweig

Mira Stemmer, Deutsches Institut für Normung e. V. (DIN)

Dr.-Ing. Patricia Stock, REFA-Institut e. V.

Julia Stoll, Deutsche Akkreditierungsstelle GmbH (DAkkS)

Dr. Christina Strobel, KI Bundesverband e. V.

Dr. Oliver Stuch, Verband der Vereine Creditreform e. V.

Johannes Stürenburg, Bundesamt für Sicherheit in der Informationstechnik (BSI)

Alexandra Surdina, Deutsche Bahn AG

Ernö Szivek, Deutsche Bundesbank

Dr. Rustam Tagiew, Deutsches Zentrum für Schienenverkehrsforschung (DZSF)

Neal Ternes, ERGO Digital Ventures AG

Sebastian Terstegen, Institut für angewandte Arbeitswissenschaft e. V. (ifaa)

Martin Tettke, Berlin Cert Prüf- und Zertifizierstelle für Medizinprodukte GmbH

Ralph Traphöner, Empolis Information Management GmbH

Holk Traschewski, Your Expert Cluster GmbH

Dr. Volker Treier, Deutscher Industrie- und Handelskammertag (DIHK)

Thorsten Trippel, Universität Tübingen

Kristina Unverricht, Bundesamt für Sicherheit in der Informationstechnik (BSI)

Dr.-Ing. Thomas Usländer, Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung (IOSB)

Dr.-Ing. Mathias Uslar, OFFIS – Institut für Informatik

Tobias van Hasselt, TÜV Nord Mobilität GmbH & Co. KG IFM

Dr.-Ing. Eric MSP Veith, OFFIS – Institut für Informatik

Sharan Vijayagopal, bauforumstahl e. V. (BFS)

Dr. Silvia Vock, Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA)

Thomas Vollmer, Philips

Dr. Justus von Geibler, Wuppertal Institut für Klima, Umwelt, Energie

Dr. Arndt von Twickel, Bundesamt für Sicherheit in der Informationstechnik (BSI)

Sabine Waechter, Datev eG

Kirsten Wagner, Deutscher Verein des Gas- und Wasserfaches e. V. (DVGW)

David Wagner-Stürz, SAMSON AG

Prof. Dr. Siegfried Wahl, Carl Zeiss Vision International GmbH

Prof. Dr. rer. nat. Dr. h.c. mult. Wolfgang Wahlster, Plattform Lernende Systeme / Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

Robert Walter, TÜV AI Lab

Dr. Thomas Waschulzik, Siemens Mobility GmbH

Steffen Waurick, Bundesamt für Sicherheit in der Informationstechnik (BSI)

Dr. Marco Wedel, Technische Universität Berlin

Prof. Dr. Dieter Wegener, Siemens AG / Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE (DKE)

Eva Weicken, Fraunhofer-Institut für Nachrichtentechnik, Heinrich-Hertz-Institut (HHI)

Felix Wenzel, ERGO Digital Ventures AG

Martin Westhoven, Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA)

Bernd Wildpanner, Imabicon UG

Dorothea Winter, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

Dr. Johannes Winter, L3S AI Research Center, ehemals: Plattform Lernende Systeme / acatech

Prof. Dr. Mario Winter, German Testing Board e. V. (GTB)

Christoph Winterhalter, Deutsches Institut für Normung e. V. (DIN)

Dr. Oliver Wirjadi, Dentsply Sirona

Sebastian Wohlrapp, Field 33 GmbH

Susanna Wolf, Datev eG

Prof. Dr. Stefan Wrobel, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS)

Dr. Alexander Wunderle, infologistix GmbH

Michael Wutz, Vitesco Technologies

Jason YiJunsong, Huawei Technologies

Prof. Dr.-Ing. Sebastian Zaunseder, Fachhochschule Dortmund

Dr. Meike Zehlike, Zalando SE

Dr. Stephan Zidowitz, Fraunhofer-Institut für Digitale Medizin (MEVIS)

Dr. Wolfgang Ziegler, z-rands

Jens Ziehn, Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung (IOSB)

Sonja Zillner, Siemens AG

Dr. Bettina Zucker, Drägerwerk AG & Co. KGaA

12 Weitere Mitglieder der Arbeitsgruppen

Dr.-Ing. Mohamed Abdelaal, Software AG (SAG)

Katja Anclam, female.vision e. V.

Lisa Auer, RWTH Aachen

Benedikt Auth, Leistritz Group

Dr. Gergana Baeva, iRights.Lab GmbH

Adam Bahlke, Motor AI

Michael Barth, Privatperson

Jens Bauch, Deutsche Bahn AG

Jessica Bauer, FOM Hochschule für Ökonomie & Management

Stephan Bautz, PricewaterhouseCoopers GmbH

Judit Bayer, Universität Münster

Justus Benning, FIR e. V. an der RWTH Aachen

Torsten Berge, Deloitte GmbH

Marc Bergenthal, Brainlab AG

Irvin Bislimi, Aesculap AG

René Böhm, Vitesco Technologies

Andre Bojahr, IAV GmbH

PD Dr. med. Ulrich Bork, Universitätsklinikum Carl Gustav Carus, Technische Universität Dresden

Dr. Mathis Börner, SAP SE

Thomas Boué, BSA

Robert Brunner, AI4SMB GbR – AI for SMBs in Logistics & Healthcare

Aaron Butler, Universität Luzern

Ralf Casperson, Bundesanstalt für Materialforschung und -prüfung (BAM)

Chih-Hong Cheng, Fraunhofer-Institut für Kognitive Systeme (IKS)

Vasilios Danos, TÜViT

Dr. Werner Daum, Bundesanstalt für Materialforschung und -prüfung (BAM)

Dr. med. Björn Diem, BIOTRONIK SE & Co. KG

Verena Dietrich, imbus AG

Marina Dolokov, FSD Fahrzeugsystemdaten GmbH

Jannis Dörhöfer, TÜV-Verband e. V.

Dr. Patrick Draheim, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

Heiko Ehrich, TÜV NORD

Claudia D. Eich, B ‘IMPRESS

Kentaro Ellert, PricewaterhouseCoopers GmbH

Jens Elsner, Munich Innovation Labs GmbH

Dr. Rainer Engels, GIZ

Prof. Dr. Kurt Englmeier, Hochschule Schmalkalden

Dr. Nico Erdmann, Deloitte GmbH

Eva Daria Ernst, Deutsches Institut für Normung e. V. (DIN)

Dr. Matthias Fabian, Privatperson

Daniel Fehrenbacher, e:fs TechHub GmbH

Jörn Fiedler, Bundesministerium der Verteidigung (BMVg)

Philip Finkler, Deloitte GmbH

Kerstin Franzl, nexus Institut

Saskia Fruth, Industrie-Anlagen-Betriebs-Gesellschaft

David Fuhr, HiSolutions AG

Christopher Ganz, C. Ganz Innovation Services

Dr. Jens Gayko, Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE (DKE)

Prof. Dr. Raimund Geene, Alice Salomon Hochschule Berlin

Regina Geierhofer, Siemens Healthcare GmbH

Sebastian Giera, Robert Bosch GmbH

Dr. Patrick Gilroy, TÜV-Verband e. V.

Lea Gimpel, GIZ

Dr. Robert Ginthör, Know-Center GmbH

Rebekka Görge, Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS)

Dr. Alexander Goschew, Deutsches Institut für Normung e. V. (DIN)

Dr. Maximilian Grabowski, BASt

Marian Gransow, VIVE-MedTech GmbH

Stephan Griebel, Siemens Mobility GmbH

Claudia Großmann, Modis GmbH

Tanja Hagemann, Deutsche Telekom AG

Andreas Halbleib, B. Braun Gruppe

Prof. Anselm Haselhoff, Hochschule Ruhr West

Dr. Vahid Hashemi, Audi AG

Thomas Heckel, Bundesanstalt für Materialforschung und -prüfung (BAM)

Manfred Hefft, Domino Deutschland GmbH

Dr. Jens Heidrich, Fraunhofer-Institut für Experimentelles Software Engineering (IESE)

Mathias Heiles, LIME medical GmbH

Martina Heim, Alcon

Jana Heinrich, Fraunhofer-Institut für Experimentelles Software Engineering (IESE)

Paul Hellwig, BG Kliniken IT-Services GmbH

Andreas Hepfner, Neo Q Quality in Imaging GmbH

Karol Tatiana Puscus Hernandez, RWTH Aachen

Dr.-Ing. Stefan Hillmann, Technische Universität Berlin

Karl Peter Hoffmann, Stadtwerke Sindelfingen

Reiner Hofmann, Universität Bayreuth, Medizincampus Oberfranken

Christoph Hohenberger, retorio GmbH

Dr. Johannes Hüdepohl, Berufsgenossenschaft Energie Textil Elektro Medienerzeugnisse – BG ETEM

Marc Jopek, Lyniate

Juliane Jungk, Freudenberg & Co. KG

Dr. Michael Karner, SETLabs Research GmbH

Sven Kasan, Digithurst Bildverarbeitungssysteme

Klaus Kaufmann, Mittelstand 4.0 Kompetenzzentrum eStandards

Dr. Hubert B. Keller, ci-tec GmbH

Dr. Christian Kellermann, Vereinigung Deutscher Wissenschaftler e. V. (VDW)

Michael Kieviet, innotec GmbH

Dorian Knoblauch, Fraunhofer-Institut für Offene Kommunikationssysteme (FOKUS)

Dr. Gesine Knobloch, Bayer AG

Sabine Knör, Atos

Johannes Koch, Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE (DKE)

Philipp Koch, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

Michael Kolain, Deutsches Forschungsinstitut für öffentliche Verwaltung (FÖV)

Dr. Sergii Kolomiichuk, Fraunhofer-Institut für Fabrikbetrieb und -automatisierung (IFF)

Roman Konertz, FernUniversität in Hagen

Dr.-Ing. Dietmar Köring, Arphenotype

Stephan Krähnert, Verband der Automobilindustrie e. V. (VDA)

Dr. rer. nat. Joachim Krois, Charité – Universitätsmedizin Berlin

Christian Kruschel, IAV GmbH

Prof. Dr. Kai-Uwe Kühnberger, Universität Osnabrück

Mark Küller, TÜV-Verband e. V.

Dr. Kai Kümmel, Privatperson

Philipp Lämmel, Fraunhofer-Institut für Offene Kommunikationssysteme (FOKUS)

Sebastian Land, Old World Computing GmbH

Claus Lang, Kodex AI GmbH

Yves Leboucher, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ), ehemals: Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE (DKE)

Dr. Andreas Lemke, mediaire GmbH

Johann Letnev, JUMO GmbH & Co. KG

Ulli Leucht, PricewaterhouseCoopers GmbH

David Lewenko, Deloitte GmbH

Matthias Lieske, Hitachi Europe GmbH

Thomas Linner, OTH Regensburg

Alina Lorenz, IT-Systemhaus der Bundesagentur für Arbeit

Mihai Maftei, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

Hans-Christian Mangelsdorf, Auswärtiges Amt

Angelina Marko, Bitkom e. V.

Dr.-Ing. Erik Marquardt, VDI Verein Deutscher Ingenieure e. V.

Björn Matthias, ABB AG

PD Dr. Matthias May, Universitätsklinikum Erlangen

Benjamin Meier, Curalie GmbH

Andreas Meisenheimer, Bundeswehr

Christian Meyer, msg-systems AG

Stephan Mietke, Bundesverband deutscher Banken

Olaf Minkwitz, Marelli Automotive Lighting

Dr. Klaus Möller, DEFINO Institut für Finanznorm AG

Dr. Julia Maria Mönig, Universität Bonn

Bernhard Mühlbauer, Energie Baden-Württemberg AG (EnBW)

Dr. Frank Müller, Heidelberg Engineering GmbH

Wolfgang Müller, Zentralverband der Augenoptiker und Optometristen

Corinna Mutter, SPECTARIS – Deutscher Industrieverband für Optik, Photonik, Analysen- und Medizintechnik e. V.

Florian Neumeier, M3i Industrie-in-Klinik-Plattform

Dr. Jens Niederhausen, Physikalisch-Technische Bundesanstalt (PTB)

Franziska Noack, Privatperson

Jan Noelle, RKiSH gGmbH

Alexander Nollau, Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE (DKE)

Prof. Dr. Dirk Nowotka, Christian-Albrechts-Universität zu Kiel

Karolina Ochs, Christian-Albrechts-Universität zu Kiel

Stefan Otterbach, Bundesministerium der Verteidigung (BMVg)

Dr. Daniel Paulus, Acosu

Juliane Pfeil, Technische Hochschule Wildau (TH Wildau)

Dr. Christian Piovano, ZF Friedrichshafen AG

Dr. Axel Plinge, Fraunhofer-Institut für Integrierte Schaltung (IIS)

Bernd Püttmann, TÜV NORD CERT GmbH

Dr. Frederic Raber, Bundesamt für Sicherheit in der Informationstechnik (BSI)

Myriam Raboldt, TU Berlin

Dr. Hans Rabus, Physikalisch-Technische Bundesanstalt (PTB)

Felix Rau, Universität Köln

Lukas Rauh, Fraunhofer-Institut für Produktionstechnik und Automatisierung (IPA)

Martin Reich, MORE THAN CAPITAL

Dr. Alexander Reiprich, KARL STORZ Endoskopie Berlin GmbH

Ina Reis, Senatskanzlei Hamburg, Amt für IT und Digitalisierung

Luca Rettenberger, Karlsruher Institut für Technologie (KIT)

Christian Richter, Verwaltungs-Berufsgenossenschaft (VBG)

Dr. Patrick Riordan, Siemens AG

Renato Rodrigues, DB Netz

Maximilian Rohleder, Friedrich-Alexander-Universität Erlangen-Nürnberg

Christian Rudolf, MHP

Ingo Rütten, Strategieberatung Zielwerk GmbH

Peter Salathe, m.Doc GmbH

Sophia Saller, SMF

Friedrich Sanzi, Leuze electronic GmbH + Co. KG

Christian Schaaf, Universitätsklinikum Heidelberg

Daniel Schäfer, Hermann Bock GmbH

Stefan Schaffer, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

Michaela Schierholz, Deutsches Institut für Normung e. V. (DIN)

Dr. Jasmine Schirmer, Carl Zeiss Meditec AG

Hans-Dieter Schmees, Verein Deutscher Werkzeugmaschinenfabriken e. V. (VDW)

Andreas Schmidt, ZF Friedrichshafen AG

Jonas Schmidt, ZF Friedrichshafen AG

Christian Schmitz, Novar GmbH a Honeywell Company

Dr. med. ETH Rüdiger Schmitz, Universitätsklinikum Hamburg-Eppendorf

Dr.-Ing. Fabian Schnabel, Fachverband des Tischlerhandwerks Nordrhein-Westfalen

Frank Schneider, TÜV-Verband e. V.

Mark Schutera, ZF Friedrichshafen AG

Daniel Schwabe, Physikalisch-Technische Bundesanstalt (PTB)

Dr. Joachim Seeler, HSP Hamburg Invest GmbH

Roman Senderek, FIR e. V. an der RWTH Aachen

Aydin Enes Seydanlioglu, Robert Bosch GmbH

Dr. Georgy Shakirin, Carl Zeiss Meditec AG

Ankur Sharma, Bayer AG

Kris Shrishak, Irish Council for civil liberties

Tomasz Soltysinski, QuIP GmbH

Georg Peter Sotiriadis, Phantasma Labs GmbH

Dirk Spaltmann, Bundesanstalt für Materialforschung und -prüfung (BAM)

Florian Stark, Industrial Analytics IA GmbH

Christina Stathatou, Kugler Maag CIE

Jan Stodt, Hochschule Furtwangen (HFU)

Christian Stohs, Union Investment

Prof. Dr.-Ing. habil. Sascha Stowasser, Institut für angewandte Arbeitswissenschaft e. V. (ifaa)

Volker Sudmann, mdc medical device certification GmbH

Dima Taleb, TÜV Rheinland

Dr.-Ing. Nikolay Tcholtchev, Fraunhofer-Institut für Offene Kommunikationssysteme (FOKUS)

Dr. habil. Florian Thiel, Physikalisch-Technische Bundesanstalt (PTB)

Heike Thomas, UL International Germany GmbH

Jack Thoms, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

Verena Till, Think Tank iRights.Lab

Hauke Timmermann, eco Verband der Internetwirtschaft e. V.

Mario Tokarz, RightMinded AI GmbH

Kevin Trelenberg, Hochschule Ruhr West

Merle Uhl, Bitkom e. V.

Dr. Thomas Unger, KraussMaffei Extrusion GmbH

René Urban, Unitransferklinik Lübeck

Bhaskar Vanamali, Kugler Maag CIE

Sonja Verschitz, Digital Humanities – Konzept und Strategie: Daten – Information – Wissen

Annette von Wedel, female.vision e. V.

Ronny Wegner, PAUL HARTMANN AG

Christoph Wehner, Otto-Friedrich-Universität Bamberg

Prof. Dr. Joh Wilh Weidringer, Privatperson

Reinhard Weissinger, ISO

Frank Werner, Software AG (SAG)

Lucas Weyrich, Robo Test

Dr. Sebastian Wieczorek, SAP SE

Rick Wilming, TU Berlin

Fabian Witt, MATHEMA GmbH

Sebastian Witte, Bundesverband Digitale Wirtschaft

Dr. Nicole Wittenbrink, VDI/VDE Innovation + Technik GmbH

Georg Woditsch, Alexianer GmbH

Thorsten Wujek, SALT AND PEPPER Technology

Semih Yalcin, TakeAway Express GmbH (Lieferando)

Marc Zeller, Siemens AG

Jing Zhang, Huawei Technologies

Klaus Ziegler, Internationaler Verband der Konferenzdolmetscher

13 Anhang

13.1 Anhang Artificial Intelligence Act (AI Act)

Die folgende Tabelle 16 liefert für die in Abbildung 6 dargestellten EU-Gesetze eine kurze Beschreibung der Inhalte und Bezüge zum AI Act. Zudem sind weitere Details wie Art der Gesetzgebung, verbundene Gesetze auf deutscher Ebene und auch Stand der Gesetzgebung gelistet.

Tabelle 16: EU-Gesetze mit verstärktem Bezug zum AI Act

1a:

EU-Grundrechtecharta

Offizieller deutscher Titel: Charta der Grundrechte der Europäischen Union

Stand: rechtsverbindlich seit dem 1. Dezember 2009

Beschreibung:

Die Charta der Grundrechte der Europäischen Union kodifiziert Grund- und Menschenrechte. In sechs Titeln (Würde des Menschen, Freiheit, Gleichheit, Solidarität, Bürgerrechte und justizielle Rechte) fasst die Charta die allgemeinen Menschen- und Bürgerrechte und die wirtschaftlichen und sozialen Rechte in einem Dokument zusammen. Die Charta enthält einige wesentliche Grundsätze, an die sich vor allem der europäische Gesetzgeber zu halten hat. In 50 Artikeln werden umfassende Rechte anerkannt, für deren Durchsetzung nicht nur der Europäische Gerichtshof in Luxemburg, sondern vorab sämtliche nationalen Richter – gewissermaßen als Unionsrichter – zuständig sind. In Art. 1 der Charta heißt es wie in Art. 1 Abs. 1 des Grundgesetzes der Bundesrepublik Deutschland: „Die Würde des Menschen ist unantastbar“. Dabei sind auch Schutzbereiche geregelt, die das deutsche Grundgesetz nicht ausdrücklich erwähnt, wie den Schutz personenbezogener Daten, das Recht auf Bildung, die Rechte von Kindern, Menschen mit Behinderung und älteren Menschen, das Recht auf eine gute Verwaltung oder die Gewährleistungen im Arbeitsrecht. Weiterhin wird der Verbraucherschutz, die Unverletzlichkeit der Wohnung, das Telekommunikationsgeheimnis, „würdige Arbeitsbedingungen“ und eine kostenlose Arbeitsvermittlung garantiert. Zudem ist die Charta von der Antidiskriminierung durchdrungen. In Art. 21 mit „Diskriminierungen insbesondere wegen des Geschlechts, der Rasse, der Hautfarbe, der ethnischen oder sozialen Herkunft, der genetischen Merkmale, der Sprache, der Religion oder der Weltanschauung, der politischen oder sonstigen Anschauung, der Zugehörigkeit zu einer nationalen Minderheit, des Vermögens, der Geburt, einer Behinderung, des Alters oder der sexuellen Ausrichtung, sind verboten.“

Die Grundrechtecharta gilt auch für Anwendungen der KI und ist die Basis für die technische Realisierung zur Vermeidung ungewollter Diskriminierung.

1b:

Produkthaftungsrichtlinie

Offizieller deutscher Titel: Richtlinie 85/374/EWG zur Angleichung der Rechts- und Verwaltungsvorschriften der Mitgliedstaaten über die Haftung für fehlerhafte Produkte

Stand: in Kraft getreten, in Deutschland als Produkthaftungsgesetz (Gesetz über die Haftung für fehlerhafte Produkte) umgesetzt

Beschreibung:

Die Kommission ist besorgt, dass die Undurchsichtigkeit und Komplexität sowie der hohe Grad an Autonomie einiger KI-Systeme es Geschädigten erschweren könnte, die Fehlerhaftigkeit eines Produkts bzw. das Verschulden sowie den ursächlichen Zusammenhang mit dem Schaden zu beweisen. Es könne auch ungewiss sein, ob und inwieweit nationale Vorschriften über die verschuldensunabhängige Haftung (z. B. für gefährliche Tätigkeiten) auf die Nutzung KI-gestützter Produkte oder Dienste Anwendung finden.

Diesen Gefahren will die Kommission ggf. durch verschiedene Maßnahmen wie etwa Beweislasterleichterungen oder eine verschuldensunabhängige Haftung des Herstellenden begegnen.

1c:

Rahmenrichtlinie zur Sicherheit und Gesundheitsschutz bei der Arbeit

Offizieller deutscher Titel: Richtlinie 89/391/EWG über die Durchführung von Maßnahmen zur Verbesserung der Sicherheit und des Gesundheitsschutzes der Arbeitnehmer bei der Arbeit

Stand: in Kraft getreten, in Deutschland als Arbeitsschutzgesetz umgesetzt

Beschreibung:

Ziel der Richtlinie ist es, für alle Arbeitnehmer ein vereinheitlichtes Schema hinsichtlich Gesundheitsschutz und Sicherheit zu schaffen. Danach werden Arbeitgeber gesetzlich verpflichtet, geeignete Präventivmaßnahmen zur Verbesserung von Sicherheits- und gesundheitsfördernden Maßnahmen zu ergreifen. Ein Angelpunkt der Richtlinie entspricht der Gefährdungsbeurteilung mit der Herausstellung u. a. folgender Themen:

→ Identifikation von Gefahren am Arbeitsplatz sowie deren schädlichen Auswirkungen
→ Geeignete Maßnahmen zur Bekämpfung potenzieller Risiken
→ Vorgehen bei der Dokumentation

1d:

Maschinenrichtlinie bzw. -verordnung

bisher als EU-Richtlinie:

Offizieller deutscher Titel: Richtlinie 2006/42/EG des Europäischen Parlaments und des Rates vom 17. Mai 2006 über Maschinen und zur Änderung der Richtlinie 95/16/EG

Stand: 2006 in Kraft getreten, in Deutschland als Produktsicherheitsgesetz umgesetzt

in Zukunft als EU-Verordnung:

Offizieller deutscher Titel: Verordnung über Maschinenprodukte

Stand: in Planung

Beschreibung:

Mit der EU-Maschinenrichtlinie werden einheitliche Anforderungen für Maschinen und unvollständige Maschinen für ein einheitliches Schutzniveau zur Unfallverhütung beim Inverkehrbringen derselben geregelt. In Deutschland wurde die Richtlinie in das Produktsicherheitsgesetz (ProdSG) und die darauf gestützte Maschinenverordnung (9. ProdSV) umgesetzt. Folgende Anforderungen müssen umgesetzt werden (Auszug):

→ Die Maschine muss mechanisch und elektrisch sicher gestaltet und die funktionale Sicherheit (z. B. sichere Steuerkreise) muss umgesetzt werden,
→ zum Zeitpunkt des Inverkehrbringens ist die Maschine sicher und eine sichere Bedienung ist gewährleistet,
→ Sicherheits- bzw. Schutzeinrichtungen der Maschine können nicht einfach umgangen werden,
→ Konformitätsbewertungsverfahren mit Risikobeurteilung (§ 158 ff.) werden durchgeführt,
→ nach erfolgreicher Bewertung erfolgt die Konformitätserklärung und das Anbringen des CE-Kennzeichens,
→ Erstellen einer technischen Dokumentation und Betriebsanleitung, die Benutzer*innen und Bedienende der Maschine deutlich auf die gekennzeichneten vorhandenen Restrisiken aufmerksam macht.

Da, wo KI-Komponenten in oder für „Maschinen“ verbaut werden, gelten die Anforderungen der Maschinenrichtlinie. Spezielle Erwägungen zu Risiken aus KI und zugehörigen Maßnahmen sind jedoch dort nicht enthalten. Ähnlich wie bei den sektoralen Harmonisierungsvorschriften (z. B. Medizinprodukteverordnung) erfolgt bisher die Vergabe des CE-Zeichens über die Maschinenrichtlinie.

Am 21.04.2021 legte die EU-Kommissionen einen Vorschlag vor, die Maschinenrichtlinie in eine Maschinenverordnung zu überführen (Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates über Maschinenprodukte, Brüssel, den 21.4.2021, [346] 202 final 2021/0105 (COD)), die in das New Legislative Framework (NLF, 768/2008/EC) eingebettet ist. Sie strebt volle Kompatibilität mit dem AI Act an, greift den Begriff „Künstliche Intelligenz“ explizit auf und weist vergleichbar zum AI Act Hochrisikosysteme aus, u. a. „Maschinen, in die Sicherheitsfunktionen wahrnehmende KI-Systeme integriert sind“.

1e:

Medizinprodukteverordnung (Medical Device Regulation (MDR)) als Beispiel für sektorspezifische Regularien der Sicherheit von Produkten in den jeweiligen Anwendungsbereichen

Offizieller deutscher Titel: Verordnung 2017/745 über Medizinprodukte

Stand: 2017 in Kraft getreten

Beschreibung:

Die MDR regelt die Zulassung und den Betrieb von Medizinprodukten. Sie stellt zentrale Anforderungen an deren Sicherheit und Wirksamkeit und schließt dabei Anforderungen an den Entwicklungsprozess von Medizinprodukten sowie alle weiterführenden Maßnahmen zur Gewährleistung einer sicheren Herstellung, Inbetriebnahme und des Betriebs. Die MDR enthält keine spezifischen Anforderungen an KI-basierte Systeme, die ein Medizinprodukt darstellen oder eine Komponente davon sind. Der AI Act versucht, diese Lücke in den sektoralen Harmonisierungsvorschriften durch grundlegende Anforderungen an KI-Systeme in einem horizontalen Ansatz zu schließen. Die Konsistenz zwischen AI Act (horizontal) und MDR (sektoral) sollte gewährleistet sein, um die Umsetzung KI-basierter Medizinprodukte zu ermöglichen und nicht zu behindern.

2a:

Datenschutz-Grundverordnung (DSGVO)

Offizieller deutscher Titel: Verordnung 2016/679 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten

Stand: 2016 in Kraft getreten

Ergänzende deutsche Gesetze: Bundesdatenschutzgesetz, Landesdatenschutzgesetze

Beschreibung:

Die DSGVO regelt die Verarbeitung personenbezogener Daten. Art. 5, 24, 25 und 32 enthalten Verantwortlichkeiten, die Erstellung einer Datenschutzfolgeabschätzung (Risikobetrachtung) und Anforderungen an eine datenschutzfreundliche und sichere Technik sowie Organisation (u. a. Pseudonymisierung und Verschlüsselung).

Für automatisierte Entscheidungsfindung z. B. aus Machine-Learning (ML)-Modellen, die Personen betreffen, ist folgender Passus entscheidend: „Werden personenbezogene Daten […] erhoben, so teilt der Verantwortliche […] Folgendes mit: das Bestehen einer automatisierten Entscheidungsfindung […] und […] aussagekräftige Informationen über die involvierte Logik […].“

Zur Bestimmung der Risiken betroffener Personen haben sich die Datenschutz-Aufsichtsbehörden europaweit auf neun Kriterien geeinigt:

1. Bewerten oder Einstufen,
2. automatische Entscheidungsfindung,
3. systematische Überwachung,
4. vertrauliche oder höchst persönliche Daten,
5. Datenverarbeitung im großen Umfang,
6. Abgleichen oder Zusammenführen von Datensätzen,
7. Daten zu schutzbedürftigen Betroffenen,
8. innovative Nutzung oder Anwendung neuer technologischer oder organisatorischer Lösungen,
9. Betroffene werden an der Ausübung eines Rechts oder der Nutzung einer Dienstleistung bzw. Durchführung eines Vertrags gehindert.

Die genannten Risikokriterien und deren Bewertung sind bei Verwendung einer KI relevant, wenn personenbezogene Daten verwendet werden. Über die verwendete Logik müssen aussagekräftige Informationen vorliegen, d. h. Transparenz über die Entstehung der Entscheidung einer KI. In der „Hambacher Erklärung zur Künstlichen Intelligenz“ [43] nehmen die deutschen Datenschutzaufsichtsbehörden zu den Anforderungen der DSGVO in Bezug auf KI konkret Stellung.

2b:

Network Information Security (NIS)-Richtlinie

Offizieller deutscher Titel: Richtlinie 2016/1148 über Maßnahmen zur Gewährleistung eines hohen gemeinsamen Sicherheitsniveaus von Netz- und Informationssystemen in der Union

Stand: 2016 in Kraft getreten

Beschreibung:

Das Ziel ist ein gleichmäßig hohes Sicherheitsniveau von Netz- und Informationssystemen in der gesamten EU durch eine erhöhte Kapazität der Cybersicherheit auf nationaler Ebene, verstärkte Zusammenarbeit auf EU-Ebene und Verpflichtungen für Betreiber wesentlicher Dienste und Anbieter digitaler Dienste, Mindestsicherheitsanforderungen für Risikovorsorgen und Aufrechterhaltung von wesentlichen Diensten sowie Meldepflichten. Es wurden Sektoren als kritische Infrastruktur definiert wie etwa Energie, Verkehr, Gesundheit und digitale Infrastruktur sowie Sanktionen. die NIS-Richtlinie ist in Deutschland mit dem IT Sicherheitsgesetz 1 und 2 umgesetzt

2c:

Cybersecurity Act (CSA)

Offizieller deutscher Titel: Verordnung 2019/881 über die ENISA (Agentur der Europäischen Union für Cybersicherheit) und über die Zertifizierung der Cybersicherheit von Informations- und Kommunikationstechnik

Stand: 2019 in Kraft getreten

Beschreibung:

Ziel des Cybersecurity Act ist es, die IT-Sicherheit EU-weit mit einheitlichen Regularien zu etablieren und für sogenannte informations- und kommunikationstechnische (IKT) Systeme, Dienste und Prozesse zu stärken.

Kernelemente des CSA sind ein permanentes Mandat für die europäische Cyber-Sicherheitsagentur ENISA („European Union Agency for Cybersecurity“) sowie die Einführung eines einheitlichen europäischen Zertifizierungsrahmens für IKT-Produkte, -Dienstleistungen und -Prozesse. Diese sollen gemäß definierter Sicherheitslevel in „niedrig“, „mittel“ und „hoch“ nach unterschiedlichen Vorgaben zertifiziert werden.

Mögliche IT-Sicherheitsrisiken aus KI sind nicht speziell beschrieben oder berücksichtigt. Eine Prüfung, inwieweit eventuell Ergänzungen erforderlich wären, ist empfehlenswert. KI-basierte Produkte unterliegen aber als IT-System-Cybersecurity-Anforderungen und müssen diese entsprechend umsetzen.

2d:

Cyber Resilience Act (CRA)

Stand: in Planung

Beschreibung:

Der CRA regelt Cybersicherheitsanforderungen für ein breites Spektrum digitaler Produkte und zugehöriger Nebendienstleistungen. Gegenstand des Gesetzes sind materielle digitale Produkte und nicht eingebettete Software in ihrem gesamten Lebenszyklus. Damit erfasst das Gesetz Hardware und Software gleichermaßen.

Die Gesetzesinitiative definiert die drei folgenden Hauptziele:

→ Gewährleistung eines gleichbleibend hohen Cybersicherheitsniveaus für digitale Produkte und Nebendienstleistungen
→ Erhöhung der Transparenz der Cybersicherheitsmerkmale
→ gleiche Wettbewerbsbedingungen für Anbieter digitaler Produkte und Nebendienstleistungen

3a:

Data Governance Act (DGA)

Offizieller deutscher Titel: Verordnung des Europäischen Parlaments und des Europäischen Rates über Europäische Daten-Governance und zur Änderung der Verordnung (EU) 2018/1724 (Daten-Governance-Rechtsakt)

Stand: 2022 in Kraft getreten

Beschreibung:

Der DGA soll europaweit Impulse zur besseren Nutzung wertvoller Daten schaffen. Es geht dabei ebenso um Daten der öffentlichen Hand wie um Datenschätze von Unternehmen.

Beispiele hierfür sind u. a. Umweltdaten aus Smart-Home-Geräten, die bei der Bekämpfung des Klimawandels helfen könnten, aber auch die stärkere Nutzung von Gesundheitsdaten zu Forschungszwecken.

Die Verordnung soll den Zugang sowohl zu persönlichen Daten als auch zu nicht-persönlichen Daten erleichtern. Sie ergänzt die im Vorjahr beschlossene Open-Data-Richtlinie der EU. Die Verordnung soll helfen, Daten einfach und rechtssicher verfügbar zu machen.

Die Verordnung schafft auch einen rechtlichen Rahmen für sogenannte Datenintermediäre. Dabei handelt es sich um neutrale Vermittlungsstellen, die den Austausch zwischen Datenquellen und interessierten Parteien ermöglichen soll. Personen, deren persönliche Daten genutzt werden, sollen sich hingegen künftig in Datengenossenschaften organisieren können. Erleichtern möchte die EU-Kommission zudem unter dem Stichwort Datenaltruismus Datenspenden für gemeinnützige Zwecke, wie sie etwa schon in der deutschen Corona-Datenspende-App geschehen.

Der DGA zielt nicht darauf ab, wesentliche Rechte auf den Zugang zu Daten und deren Nutzung zu gewähren, zu ändern oder zu beseitigen.

3b:

Digital Services Act (DSA)

Offizieller deutscher Titel: Verordnung des Europäischen Parlaments und des Europäischen Rates über einen Binnenmarkt für digitale Dienste (Gesetz über digitale Dienste) und zur Änderung der Richtlinie 2000/31/EG

Stand: in Planung, kurz vor Abschluss des Gesetzgebungsverfahrens; Geltung voraussichtlich spätestens ab dem 1. Januar 2024

Beschreibung:

Der DSA soll als eine Art „Charta des Internets“ den digitalen Raum gegen die Verbreitung illegaler Inhalte schützen und die Grundrechte der Nutzer*innen gewährleisten. Er soll die Verbreitung von Hatespeech und Desinformation verhindern, den Verbraucherschutz im Netz stärken und Transparenz darüber schaffen, wie digitale Dienste funktionieren.

Der DSA folgt dabei dem Grundsatz, dass das, was offline illegal ist, auch online illegal sein muss.

Im Wesentlichen müssen Online-Plattformen einschließlich sozialer Medien und Marktplätze Maßnahmen ergreifen, um die Nutzer*innen vor illegalen Inhalten, Waren und Dienstleistungen zu schützen. Der DSA wird für alle Online-Vermittler gelten, die in der EU Dienste anbieten, aber sehr große Online-Plattformen („very large online-platforms“, VLOPs) und sehr große Online-Suchmaschinen („very large online-search engines“, VLOSEs), also Dienste mit mehr als 45 Millionen aktiven Nutzer*innen in der EU, werden strengeren Anforderungen unterliegen als Kleinst- und Kleinunternehmen, die von einigen der Verpflichtungen ausgenommen sind.

Auf Anfrage der zuständigen Behörde müssen besonders große Online-Plattformen der zuständigen Behörde Zugang zu den Daten geben, die notwendig sind, um die Einhaltung des DSA zu überwachen.

3c:

Digital Markets Act

Offizieller deutscher Titel: Verordnung des Europäischen Parlaments und des Europäischen Rates über bestreitbare und faire Märkte im digitalen Sektor (Gesetz über digitale Märkte)

Stand: in Planung

Beschreibung:

Einige wenige, sehr große Online-Plattformen machen einen sehr großen Teil der digitalen Wirtschaft in der EU aus. Ihre wirtschaftliche Macht und Kontrolle über ganze Plattform-Ökosysteme machen es für Konkurrenten oder neue Marktteilnehmer oft unmöglich, im Wettbewerb zu bestehen. Der Digital Markets Act zeigt Möglichkeiten für die Regulierung großer, als „Gatekeeper“ fungierender Online-Plattformen auf.

Gemäß Art. 19 des DMA kann die Europäische Kommission durch einfaches Auskunftsverlangen oder im Wege eines Beschlusses auch Zugang zu Datenbanken und Algorithmen von Unternehmen verlangen und diesbezügliche Erläuterungen anfordern.

3d:

Data Act

Offizieller deutscher Titel: Verordnung des Europäischen Parlaments und des Europäischen Rates über harmonisierte Vorschriften für einen fairen Datenzugang und eine faire Datennutzung (Datengesetz)

Stand: in Planung

Beschreibung:

Durch „vernetzte“ Produkte und Dienstleistungen, das sogenannte Internet of Things (IoT), werden Daten in erheblichem Ausmaß und von erheblichem Wert erzeugt, beispielsweise beim Fahren des eigenen Autos oder der Steuerung der eigenen Heizung. Der weit überwiegende Teil dieser Daten wird gegenwärtig entweder überhaupt nicht genutzt oder es profitieren nur wenige sehr große Unternehmen davon. Der Data Act zielt auf eine gerechtere Verteilung der mit Daten verbundenen Wertschöpfung und soll durch die Wiederverwendbarkeit von Daten Wettbewerbsfähigkeit und Innovation in der Europäischen Union fördern. Durch die neuen Vorschriften sollen mehr Daten für die Weiterverwendung zur Verfügung stehen, und es wird erwartet, dass hierdurch bis 2028 ein zusätzliches BIP in Höhe von 270 Milliarden Euro entsteht.

Hierzu schafft der Verordnungsentwurf bestimmten privaten und öffentlichen Akteur*innen ein neues Recht auf Datenzugang und Datennutzung.

3e:

European Health Data Space (EHDS) als Beispiel für sektorspezifische Regularien bezüglich Zugang zu Daten in den jeweiligen Anwendungsbereichen

Offizieller deutscher Titel: Verordnung des Europäischen Parlaments und des Europäischen Rates über den europäischen Raum für Gesundheitsdaten

Stand: in Planung

Beschreibung:

Der EHDS soll den Zugang zu Gesundheitsdaten regeln. Das betrifft einerseits Maßnahmen zur Kontrolle der Einzelpersonen bezüglich ihrer eigenen Daten. Andererseits fördert er die Nutzung von Gesundheitsdaten, um eine bessere medizinische Versorgung insbesondere für die Bereiche Forschung, Innovation und Politikgestaltung zu ermöglichen. Er versucht dabei, das Potenzial von Austausch, Nutzung und Weiterverwendung der Daten unter dem Maßstab der Interoperabilität, aber auch unter Gewährleistung eines gesicherten Zugangs umfassend auszuschöpfen.

In Hinblick auf den geplanten AI Act stellt er ein zentrales Element dar, damit KI- bzw. Machine-Learning-basierte Ansätze im Gesundheitswesen umgesetzt werden können. Beide Verordnungen beinhalten Anforderungen in Hinblick auf den Zugang bzw. den Umgang mit den für die KI-Systeme erforderlichen Daten, die im medizinischen Umfeld zunächst oftmals personenbezogene Daten darstellen und daher in geeigneter Weise vorverarbeitet werden müssen, insbesondere mit Methoden zur Anonymisierung bzw. Pseudonymisierung.

Exemplarische Darstellung am Beispiel Medizinprodukte
KI-basierte Medizinprodukte gehören zu den Bereichen, die nach Geltungsbeginn des AI Act zwei Harmonisierungsvorschriften erfüllen müssen. Neben dem AI Act ist dies noch die Verordnung 2017/745 über Medizinprodukte, im Folgenden MDR (Medical Device Regulation), als weitere EU-Verordnung. Medizinprodukte gelten dabei gemäß Art. 6 Abs. 1 und Anhang II kategorisch als Hochrisikoprodukte im Sinne des AI Act, sobald sie gemäß der MDR einer Konformitätsbewertung unterzogen werden. Die MDR selbst beinhaltet eine eigene Risikoklassifizierung, die die Klassen I, IIa, IIb und III umfasst, wobei der Schweregrad bei einer potenziellen Schädigung der Patient*innen oder Benutzer*innen miteinfließt. Nach Anwendung der Klassifizierungsregel 11 (siehe Anhang VIII der MDR) wird praktisch jede medizinische Stand-alone-Software mindestens in Klasse IIa eingeordnet, bei höherem Gefährdungspotenzial auch IIb oder III. Aufgrund der Anforderungen der MDR durchläuft das Produkt in der Folge ein überwachtes Konformitätsbewertungsverfahren gemäß MDR und ist damit ein Hochrisikosystem i. S. d. Art. 6 Abs. 1 des AI Act-Entwurfs.

Damit gelten die entsprechenden Anforderungen des AI Act, z. B. in Bezug auf Informationssicherheit (Cybersecurity), Umsetzung eines Risikomanagementsystems, Post-Market Surveillance, Meldesystem, technische Dokumentation, Kennzeichnung, QM-System und den Eintrag in eine Produktdatenbank. Gleiches fordert auch die MDR. Allerdings weichen die beiden Verordnungen in einigen Punkten voneinander ab bzw. enthalten Inkonsistenzen, die ausgeräumt werden sollten, damit die Produkte beiden Verordnungen entsprechend auf den Markt gebracht werden können. So ist das für die Einhaltung der grundlegenden Sicherheits- und Leistungsanforderungen der MDR anzuwendende Qualitätsmanagementsystem gemäß DIN EN ISO 13485:2021 [381] zwar grundsätzlich mit den Anforderungen des Entwurfs zum geplanten AI Act kompatibel. Folgende Forderungen aus diesem Entwurf werden aber nicht berücksichtigt:

- → Spezifisches Verfahren zur Verwaltung der Daten, die zum Trainieren des Geräts vor und zum Zweck des Inverkehrbringens erforderlich sind
- → Anpassen des Verfahrens für die Kommunikation mit den Marktbehörden: Zugang zu Daten
- → Anpassen des Design- und Entwicklungsprozesses, um die Anforderungen von Anhang VI zu erfüllen (z. B. Schulung des KI-Systems, menschliche Aufsicht)
- → Anpassen des Risikomanagementsystems. Gemäß Art. 9 Abs. 8 ist u. a. zu berücksichtigen, ob das Hochrisiko-KI-System wahrscheinlich für Kinder zugänglich ist oder Auswirkungen auf Kinder hat.

Zusätzlich gibt es in der MDR gemäß Art. 33 eine eigene Produktdatenbank (Eudamed), die für vielfältige Zwecke eingesetzt wird. Dazu gehören die Registrierung sowie eine grundlegende Beschreibung der Produkte inklusive Informationen über den Herstellenden und andere relevante Wirtschaftsakteur*innen, zum Produkt gehörige Leistungsnachweise (inklusive klinische Prüfungen) sowie gesammelte Informationen bezüglich Vigilanz und Marktüberwachung. Es bleibt unklar, ob die gemäß Art. 60 AI Act geforderte Produktdatenbank bereits durch die Eudamed-Datenbank gegeben ist oder ob es sich um eine eigenständige Datenbank handelt. In letzterem Fall würde das eine Doppelung des Aufwands bezüglich der Pflege der Daten bedeuten mit dem zusätzlichen Risiko von Inkonsistenzen bei der Meldung von Vorkommnissen, z. B. aufgrund unterschiedlicher Anforderungen.

Eine weitere Herausforderung sind widersprüchliche Anforderungen an das Risikomanagement. Während die MDR eine Risiko-Nutzen-Abwägung (vgl. Art. 2 Nr. 24 MDR) erlaubt, nach der ein Medizinprodukt in Verkehr gebracht werden darf, wenn der Nutzen eines Produkts schwerer wiegt als die damit verbundenen Risiken (potenzielle Schädlichkeit, vgl. Anhang I Nr. 8 MDR), verfolgt der Entwurf des geplanten AI Act einen ALAP-Ansatz (As Low As Possible), wonach Risiken unabhängig vom Nutzen so weit wie möglich gemindert werden müssen. Da die Konformität mit beiden Harmonisierungsvorschriften gewährleistet sein muss, würde das bedeuten, dass immer die strengeren Regeln beachtet werden müssten. Die Anforderungen des geplanten AI Act sind jedoch, wie im Falle des Risikomanagements, für den spezifischen Anwendungsbereich u. U. nicht sinnvoll. Im Bereich von Medizinprodukten ist die Abwägung von Risiken und Nutzen ein zentrales Merkmal für die Konformitätsbewertung.

Weitere Überschneidungen und auch Inkonsistenzen der Anforderungen können zu Problemen im Zulassungsprozess führen. Nach Erwägungsgrund 63 und Art. 43 Abs. 3 des Entwurfs zum geplanten AI Act soll es zwar – um Doppelungen zu vermeiden – ausreichen, für Hochrisiko-KI-Systeme nur ein Konformitätsbewertungsverfahren nach einer in Anhang II gelisteten anwendbaren Vorschrift zu durchlaufen. Hierbei wird allerdings davon ausgegangen, dass die Benannte Stelle auch für den AI Act zertifiziert ist. Wird die Zertifizierung durch die Benannte Stelle aber nicht angestrebt, muss eine weitere Benannte Stelle für die Überwachung der Einhaltung des AI Act herangezogen werden. Da die Verfügbarkeit für die MDR bei Benannten Stellen ohnehin nach wie vor sehr knapp ist, kommt hier nochmals eine zusätzliche Einschränkung hinzu. In Verbindung mit dem Fachkräftemangel in diesem Feld,

das durch die Einbindung von KI weitere Komplexität erhält, dürfte es erhebliche Engpässe geben.

Aus der Perspektive der Medizintechnik wird der AI Act also eher zur Innovationsbremse als zur Förderung. Herstellende von Medizinprodukten dürfen KI nur zur Verbesserung von Sicherheit, Leistungsfähigkeit und Wirksamkeit von Produkten einsetzen (Anhang I Kapitel I MDR). Die Ergebnisse einer von SPECTARIS veröffentlichten Unternehmensumfrage (vgl. Erste Bilanz der deutschen Herstellenden von Medizinprodukten nach Geltungsbeginn der EU-Medizinprodukteverordnung (MDR); hrsg. v.: Deutscher Industrie- und Handelskammertag e. V., MedicalMountains GmbH, SPECTARIS. Deutscher Industrieverband für Optik, Photonik, Analysen- und Medizintechnik e. V.; Berlin, Tuttlingen; April 2022) zeigen eine deutliche Verlängerung der Konformitätsbewertungsverfahren unter Einbindung einer Benannten Stelle von durchschnittlich 45 %. In der Risikoklasse III hat sich die Dauer der Konformitätsbewertungsverfahren sogar mehr als verdoppelt (101 %). Dazu kommt, dass jetzt schon zahlreiche Produkte vom Markt genommen werden, viele Innovationsprodukte auf Eis liegen und die meisten Bestandsprodukte noch nicht in die MDR überführt wurden.

Die Gefahr einer weiteren Verzögerung durch zusätzliche Anforderungen aus einer KI-Verordnung werden nicht viele Herstellende eingehen.

Eine weitere Inkonsistenz ergibt sich aus der Anwendung des geplanten AI Act parallel zu den Forderungen aus der DSGVO. Art. 64 Abs. 1 des Entwurfs zum AI Act fordert „uneingeschränkten Zugang zu den von den Anbietern genutzten Trainings-, Validierungs- und Testdatensätzen, auch über Anwendungsprogrammierschnittstellen (engl. Application Programming Inerface, API) oder sonstige für den Fernzugriff geeignete technische Mittel und Instrumente“ für Marktüberwachungsbehörden. Es ist nicht auszuschließen, dass zum Training, zur Validierung und zum Testen eines KI-Systems vertrauliche Patientendaten herangezogen werden. Diese per Remote-Zugriff zugreifbar zu machen, steht im Widerspruch zu den Regeln der DSGVO, die Gesundheitsdaten als besonders schützenswerte personenbezogene Daten definiert. Entweder verstößt man also gegen die eine oder gegen die andere Vorschrift. Neben dem geplanten EU Data Act wäre dies eine weitere Vorschrift zum Umgang mit Daten.

13.2 Anhang Sprachtechnologien

Existierende spezielle Normen und Standards im Bereich Sprachtechnologie i. w. S. auf nationaler, europäischer und internationaler Ebene

Design

→ ISO 9241: Ergonomics of human-system-interaction – Part 110: Dialogue principles
- Last reviewed and confirmed in 2018. → Now under review
- DIN EN ISO 9241-110:2020 → Published in May 2020

→ ISO 9241: Ergonomics of human-system-interaction – Part 154: Interactive voice response (IVR) applications)
- Last reviewed and confirmed in 2020. → Now confirmed

→ ISO 9241: Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts
- Published on 2018-04-04

→ ISO 9241-210: Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems
- Published on 2019-07-04

→ ISO 9241: Ergonomics of human-system interaction – Part 171: Guidance on software accessibility
- International standard confirmed on 2018-12-08

→ AS 5061
- Withdrawn 2019

Voice interaction

→ ETSI ES 202 076 V2.1.1

→ ISO/IEC 30122: Information technology – User interfaces – Voice commands – Part 1: Framework and general guidance
- ISO/IEC 30122-1:2016 → 08-2016

→ ISO/IEC 30122: Information technology – User interfaces – Voice commands – Part 2: Constructing and testing
- ISO/IEC 30122-2:2017 → 02-2017
- 15.01.2022 Under systematic review

→ ISO/IEC 30122: Information technology – User interfaces – Voice commands – Part 3: Translation and localization
- ISO/IEC 30122-3:2017 → 02-2017
- 15.01.2022 Under systematic review

→ Voice Control API (VOCAPI)

→ Web Speech API
- Draft Community Group Report, 18 August 2020

NLP

→ ISO 24617-2: Language resource management – Semantic annotation framework (SemAF) – Part 2: Dialogue acts
- Under review, it will be replaced by ISO/DIS 24617-2

- ISO 24617-2:2020: Language resource management – Semantic annotation framework (SemAF) – Part 2: Dialogue acts → 02.12.2020
- Speech Recognition Grammar Specification (SRGS); 16 March 2004
- Semantic Interpretation for Speech Recognition (SISR); 5 April 2007

→ **ISO standards:**
Foundational and terminological standards:
- ISO/IEC 2382: Information technology – Vocabulary
- ISO/IEC 22989:2022: Information technology – Artificial intelligence – Artificial intelligence concepts and terminology
- ISO/IEC 24029-2: Information technology – Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 2: Methodology for the use of formal methods

Natural language data

→ ISO 5127: Information and documentation – Foundation and vocabulary

ISO/TC 37 projects:

→ ISO 639 series: Codes for the representation of names of languages
→ ISO/TR 20694: A typology of language registers
→ ISO/TR 21636: Identification and description of language varieties
→ ISO 24611: Language resource management – Morpho-syntactic annotation framework (MAF)
→ ISO 24612: Language resource management – Linguistic annotation framework (LAF)
→ ISO 24614 series: Language resource management – Word segmentation of written texts
→ ISO 24615 series: Language resource management – Syntactic annotation framework (SynAF)
→ ISO 24617 series: (especially parts 2 and 4): Language resource management – Semantic annotation framework (SemAF)
→ ISO 24624: Language resource management – Transcription of spoken language
→ ISO 24619: Language resource management – Persistent identification and sustainable access (PISA)
→ ISO 20539: Translation, interpreting and related technology – Vocabulary
→ ISO 17100: Translation services – Requirements for translation services

ISO/TC 159 PROJECTS:

→ ISO 24551: Ergonomics – Accessible design – Spoken instructions of consumer products
→ ISO 9241-154: Ergonomics of human-system interaction – Part 154: Interactive voice response (IVR) applications
→ ISO/TR 19358: Ergonomics – Construction and application of tests for speech technology

ITU-T standards:

Projects from SG16 „Multimedia coding, systems and applications”:

→ ITU-T F.745: Functional requirements for network-based speech-to-speech translation services
→ ITU-T F.746.5: Framework for a language learning system based on speech and natural language processing (NLP) technology
→ ITU-T F.746.10: Architecture for a spontaneous dialogue processing system for language learning
→ ITU-T H.625: Architecture for network-based speech-to-speech translation services
→ ITU-T H.862.5 (ex F.EMO-NN): Emotion enabled multimodal user interface based on artificial neural network
→ ITU-T F.746.11 (ex F.IQAS-INT): Interfaces for intelligent question answering system
→ ITU-T F.AI-FASD: Framework for audio structuralizing based on deep neural network (see work programme)
→ ITU-T F.AI-SCS: Use cases and requirements for speech interaction of intelligent customer service (see work programme)
→ ITU-T F.REAIOCR: Requirements and evaluation methods for AI-based optical character recognition service (see work programme)
→ ITU-T F.AI-RMCDP: Requirements of multimedia composite data preprocessing (see work programme)
→ ITU-T FSTP-ACC-AI: Guideline on the use of AI for ICT accessibility (see work programme)

Quality assessment

Projects from SG12 „Performance, quality of service and quality of experience”:

→ ITU-T P.1130: Subsystem requirements for automotive speech services
→ ITU-T P.1140: Speech communication requirements for emergency calls originating from vehicles
→ ITU-T P.1150: In-car communication audio specification
→ ITU-T P.59: Artificial conversational speech
→ ITU-T P.85: A method for subjective performance assessment of the quality of speech voice output devices

→ ITU-T P.807: Subjective test methodology for assessing speech intelligibility
→ ITU-T Rec. P.851: Subjective quality evaluation of telephone services based on spoken dialogue systems
→ ITU-T P.Sup24: Parameters describing the interaction with spoken dialogue systems

W3C Community Groups

→ Voice Interaction Community Group
- JSON Representation of Semantic Information → last modified: February 12, 2019
- Intelligent Personal Assistant Architecture → Architecture and Potential for Standardization Version 1.0 → Last modified: March 24, 2020
- Intelligent Personal Assistant Architecture → Architecture and Potential for Standardization Version 1.2 → Last modified: July 19, 2021

→ Conversational Interfaces Community Group
- Dialogue Manager Programming Language (DMPL) → Final Community Group Report 13 April 2020
- DM Script (DMS) → Final Community Group Report 13 April 2020

→ Voice Assistant Standardisation Community Group → nothing new
→ The Voice Browser Working Group
→ Closed on 2015-10-12.
→ Multimodal Interaction Working Group
→ Closed in February 2017
→ https://www.w3.org/community/mqmcg/
→ https://www.astm.org/workitem-wk46396

W3C standards

→ Voice Extensible Markup Language (Voice XML) Version 2.0
- W3C Recommendation 16 March 2004
- VoiceXML Version 3.0 → W3C Working Draft 16 December 2010

→ Speech Synthesis Markup Language (SSML) Version 1.1
- W3C Recommendation 7 September 2010

→ Pronunciation Lexicon Specification (PLS) Version 1.0
- W3C Recommendation 14 October 2008

→ EMMA: Extensible MultiModal Annotation markup language
- W3C Recommendation 10 February 2009
- EMMA: Extensible MultiModal Annotation markup language Version 2.0 → W3C Working Group Note 2 February 2017

Other projects

→ COMPRISE: D5.1 Data protection and GDPR requirements

Other regulations

→ Interstate Media Treaty (Medienstaatsvertrag – MStV) → new regulations on firms or technologies that serve as intermediaries to online media services. → 7 November 2020 (https://www.die-medienanstalten.de/fileadmin/user_upload/Rechtsgrundlagen/Gesetze_Staatsvertraege/Interstate_Media_Treaty_en.pdf
→ European Data Protection Board: „Guidelines 02/2021 on virtual voice assistants" → February 2021 (https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-022021-virtual-voice-assistants_en)

Associations

→ Open Voice Network (OVON)

De facto standards:

[tokenization, PoS tagging, dependency parsing] Universal Dependencies guidelines

[language identification] Formats and metrics of the NIST LRE challenge series (documentation)

[speaker detection] Formats and metrics of the NIST SRE challenge series (documentation)

[machine translation] .sgm format (based on SGML), OPUS data formats (documentation)

[machine translation] NIST and BLEU evaluation metrics, sacreBLEU evaluation tool

[automatic summarization] ROUGE evaluation metrics
[word embeddings] word2vec format (space-separated), GloVe format (without header), fastText binary format

[named entity recognition] CoNLL-03 format, BIO/BILOU versions

[entity detection] ACE EDT guidelines (documentation)

[entity link tracking] ACE LNK guidelines (documentation)

[entity linking] TAC KBP EDL guidelines (documentation)

[relation extraction, event extraction] TAC Rich ERE guidelines (documentation)

[relation extraction] TACRED annotation scheme (documentation)

[entity tagging, values, relations, event extraction] ACE English guidelines for Entities, Values, Relations, Events

13.3 Anhang Sicherheit

Tabelle 17: Beispiele für existierende Prüfungen und Zertifizierungen für Safety/Security/Privacy

	Safety Produkt/System	Safety-Prozess	Security Produkt/System	Security prozessorientiert	Privacy Produkt/System	Privacy prozessorientiert
Prüfziele	Sicherheit und Gesundheit (Anhang I MRL, insbesondere: Anforderung an die Steuerung (Zuverlässigkeit (Kriterien siehe Normen zur Gestaltung von Steuerungen (z. B. DIN EN ISO 138491:2016 [109] und maschinentypspezifische Typ C-Normen))), ggf. bei Vorhandensein einer Mensch-Maschine-Schnittstelle Ergonomie) Reliability, Availability, Maintainability, Safety	Risikobeurteilung und Risikominderung (Prozess nach DIN EN ISO 12100: 2011 [517] (maschinentyp-spezifische Typ B und C-Normen)) Software – Entwicklung nach V-Modell (DIN EN 61508:2011 [101], [102], [103], [433]) Berücksichtigung von Hardware-Anforderungen	→ Confidentiality → Integrity → Accountability → Authenticity → Availabiltiy → Non-repudiation → Security by design and default → Security over LifeCycle	→ Confidentiality → Integrity → Accountability → Authenticity → Availabiltiy → Non-repudiation → Information-Security Management System (ISMS) → inklusive Risikobeurteilung	**Datenschutz Privacy by design and default** → Datenschutzfolgeabschätzung (Impact/Risiko) **Datensicherheit** → Confidentiality → Integrity → Accountability → Authentizität → Availability → Non-repudiation	**Datenschutzprozesse/ISMS** → Datenschutzfolgeabschätzung (Impact/Risiko) **Datensicherheit** → Confidentiality → Integrity → Accountability → Authentizität → Availability → Non-repudiation
Arten von Bewertung/Prüfungen Operationalisierung	MRL: Eigenerklärung EU-Baumusterprüfung (1 Baumuster wird geprüft – Herstellender gewährleistet selbst, dass Herstellung gemäß Baumuster erfolgt) CE- Kennzeichen (Conformité Européenne; Europäische Konformität) Selbsterklärung	MRL: Eigenerklärung (Modul A – interne Fertigungskontrolle (sowohl Produkt als auch Prozess))	Selbsterklärung Kennzeichen	Zertifikat	Zertifikat	Zertifikat
Prüfvorgaben aus	MRL: (Modul A – interne Fertigungskontrolle (sowohl Produkt als auch Prozess))	MRL: Qualitätssicherung (Anhang X MRL: Prüfung ob Entwicklungs-, Herstellungs- und Prüfungsprozess den Anforderungen genügt, Kriterien aus Modul H)		z. B. ISO/IEC 27001 [480] ff. DIN EN IEC 62443 (alle Teile) [435]		DIN EN ISO/ IEC 29100:2020 Privacy framework [133] ISO/IEC 27701 [128]
Zertifizierung für	MRL: maschinentypspezifische Typ C-Normen harmonisierte Normen Anhang I MRL	MRL: keine Normen vorhanden, die allein als ausreichend angesehen werden (z. B. DIN EN ISO 9001: 2015 [263])		NIS/CSA z. B. mit ISO 27001 ff. oder z. B. DIN EN IEC 62443 (alle Teile) [435]	DSGVO Zertifizierung in Arbeit aber noch nicht verabschiedet.	DSGVO
	MRL: GS (Geprüfte Sicherheit – national → freiwillige Einbeziehung einer Drittstelle für Maschinen)					

13.4 Anhang Mobilität

Trustworthiness-Readiness-Matrizen: ausgewählte Ergebnisse

Erfassung von Relevanzen und Operationalisierungsstand und Ableitung von Handlungsbedarfen mittels

Trustworthiness-Readiness-Matrix (TRM)
In jeweils zwei Workshops mit Fachexpert*innen wurden für die drei Use Cases

1. Ausweichmanöver als komplexes Fahrmanöver beim automatisierten Fahren
2. Ridesharing als Mobilitätsdienst/Mobilitätskette
3. Use Case Verkehrsoptimierung/Verbesserung der LSA in der Verkehrsinfrastruktur

jeweils
→ Relevanzen und
→ Operationalisierungsstand

für den zweidimensionalen Raum mit den Dimensionen
→ Einbettung und Lebenszyklusphasen und
→ Vertrauenswürdigkeitsaspekte (nachstehend auch nur als „TW Aspect“ bezeichnet)

erarbeitet [312]. Für jede Zelle der Matrix wurden zu diesem Zweck Punkte auf einer Skala von 0 bis 10 vergeben und entsprechend der Bedeutung farblich markiert (grün = führt eher nicht zu großem Handlungsbedarf, gelb = führt ggf. zu Handlungsbedarf, rot = führt wahrscheinlich zu großem Handlungsbedarf). Aus den sich danach ergebenen Matrizen für Relevanz und Operationalisierungsstand wurden nach der Formel
→ Handlungsbedarf = 1.5 * Relevanz * (10-Operationalisierung) / 10
die jeweiligen Handlungsbedarfe abgeleitet.

Abschließend wurden die genauen Punktwertungen entfernt, um einer Scheingenauigkeit bzw. diesbezüglichen Fehlinterpretationen vorzubeugen. Denn die jeweiligen Bewertungen beruhen zwar auf den Erfahrungen ausgewählter Expert*innen, wurden jedoch in einem kleinen Kreis abgestimmt. Mithin fehlen bislang streng belastbare und breit abgestimmte Kriterien für die jeweiligen Punktwertungen (diese werden aktuell außerhalb der 2. Ausgabe der Normungsroadmap erarbeitet).

Im Folgenden werden die erarbeiteten Matrizen – als Ergänzung zur obigen Zusammenfassung in Textform (vgl. Kapitel 4.6) – zur Verfügung gestellt.

Use Case Ausweichmanöver als komplexes Fahrmanöver beim automatisierten Fahren

Abbildung 54 zeigt die Relevanzen der Kombinationen von TAI-Aspekten (Trusworthy Artificial Intellligence) und Lebenszyklusphasen bzw. Einbettungsaspekten für den Use Case Ausweichmanöver beim automatisierten Fahren

Abbildung 55 zeigt den Stand der Operationalisierung der Kombinationen von TAI-Aspekten und Lebenszyklusphasen bzw. Einbettungsaspekten für den Use Case Ausweichmanöver beim automatisierten Fahren.

Abbildung 56 zeigt die Handlungsbedarfe bezüglich der Kombinationen von TAI-Aspekten und Lebenszyklusphasen bzw. Einbettungsaspekten für den Use Case Ausweichmanöver beim automatisierten Fahren.

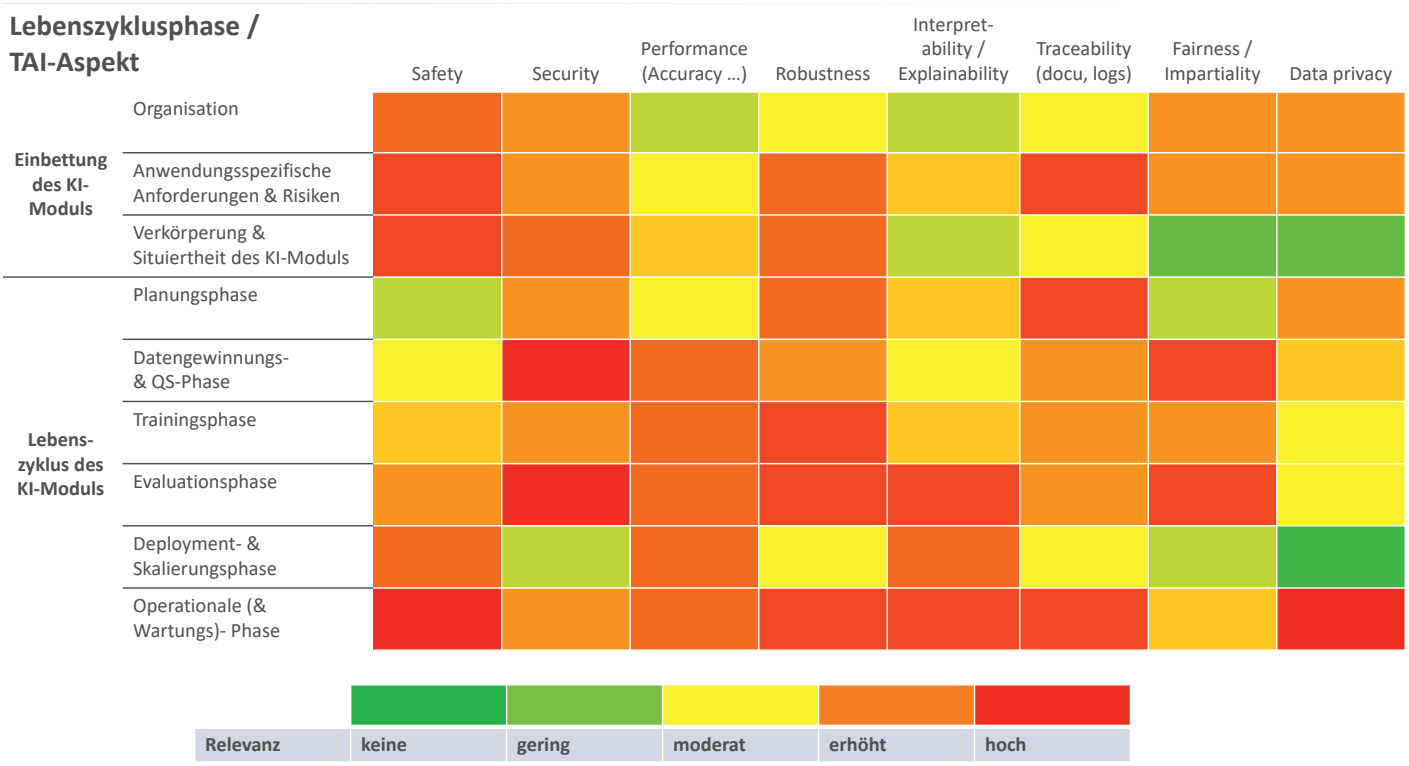


Abbildung 54: Relevanzen der Kombinationen für den Use Case Ausweichmanöver (Quelle: Arndt von Twickel, Martin F. Köhler)

Abbildung 55: Stand der Operationalisierung für den Use Case Ausweichmanöver (Quelle: Arndt von Twickel, Martin F. Köhler)

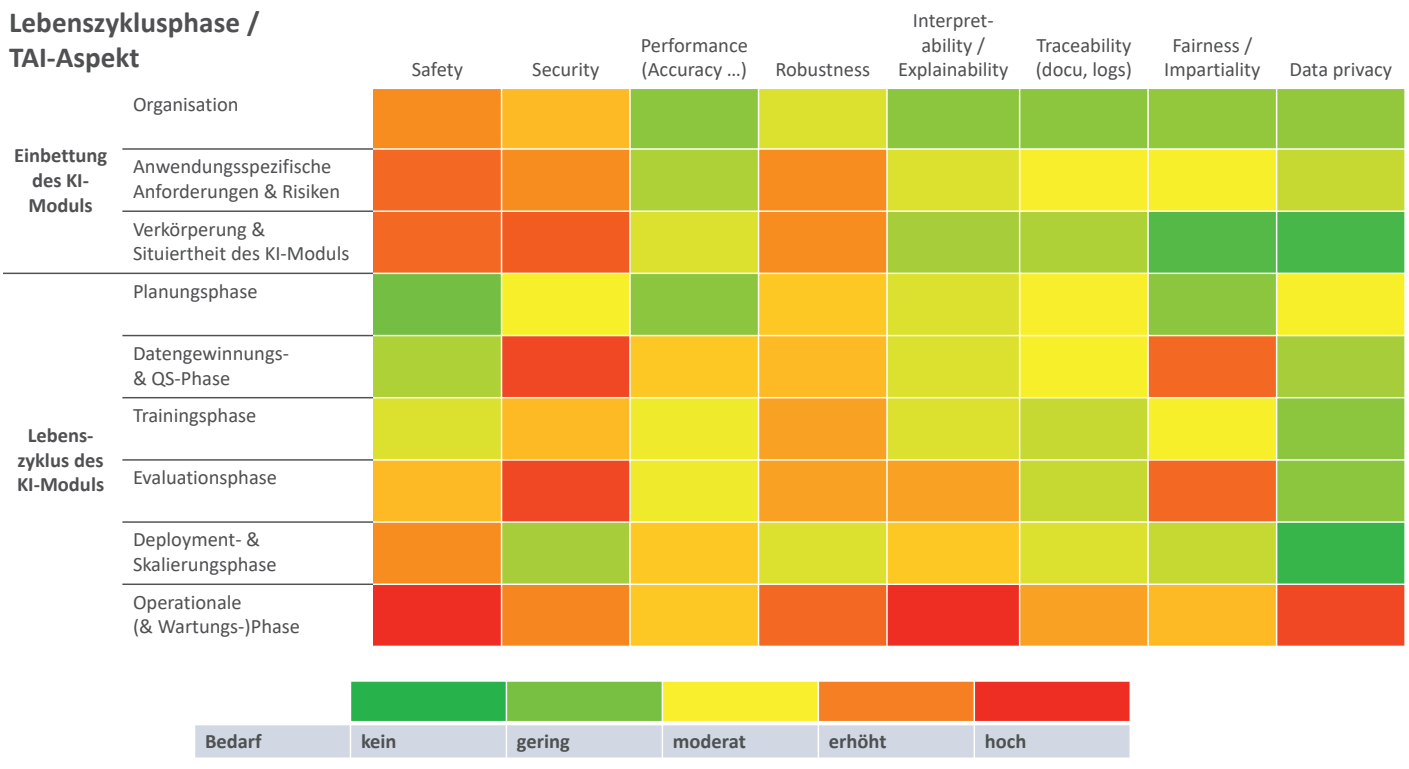


Abbildung 56: Handlungsbedarfe für den Use Case Ausweichmanöver (Quelle: Arndt von Twickel, Martin F. Köhler)

Use Case Ridesharing als Mobilitätsdienst (Mobilitätskette)

Bei der Erfassung des aktuellen Trustworthiness-Readiness-Standes beim Ridesharing dient der aktuelle Stand im Bereich automatisiertes Fahren als Grundlage und dieser wird um Ridesharing-Aspekte erweitert. Daher sind hier nur die zusätzlichen Relevanzen, Operationalisierungsstände und Handlungsbedarfe im Verhältnis zum automatisierten Fahren aufgeführt. Die Zellen ohne wesentliche Änderungen sind grau markiert, die mit Änderungen im bekannten Farbcode.

Abbildung 57 zeigt die Relevanzen der Kombinationen von TAI-Aspekten und Lebenszyklusphasen bzw. Einbettungsaspekten für den Use Case Ridesharing als Mobilitätsdienst (Mobilitätskette) – Ergänzungen im Vergleich zum automatisierten Fahren. Zellen mit unveränderten Relevanzen sind ausgegraut, die restlichen im bekannten Farbschema markiert.

Abbildung 58 zeigt den Stand der Operationalisierung der Kombinationen von TAI-Aspekten und Lebenszyklusphasen bzw. Einbettungsaspekten für den Use Case Ridesharing als Mobilitätsdienst (Mobilitätskette). Es wurden Ergänzungen im Vergleich zum automatisierten Fahren vorgenommen. Zellen mit unveränderten Operationalisierungsständen sind ausgegraut, die restlichen Zellen sind im bekannten Farbschema markiert.

Abbildung 59 zeigt die Handlungsbedarfe bezüglich der Kombinationen von TAI-Aspekten und Lebenszyklusphasen bzw. Einbettungsaspekten für den Use Case Ridesharing als Mobilitätsdienst (Mobilitätskette). Es wurden Ergänzungen im Vergleich zum automatisierten Fahren vorgenommen. Zellen mit unveränderten Bedarfen sind ausgegraut, die restlichen Zellen sind im bekannten Farbschema markiert.

Abbildung 57: Relevanzen der Kombinationen für den Use Case Ridesharing (Quelle: Arndt von Twickel, Martin F. Köhler)

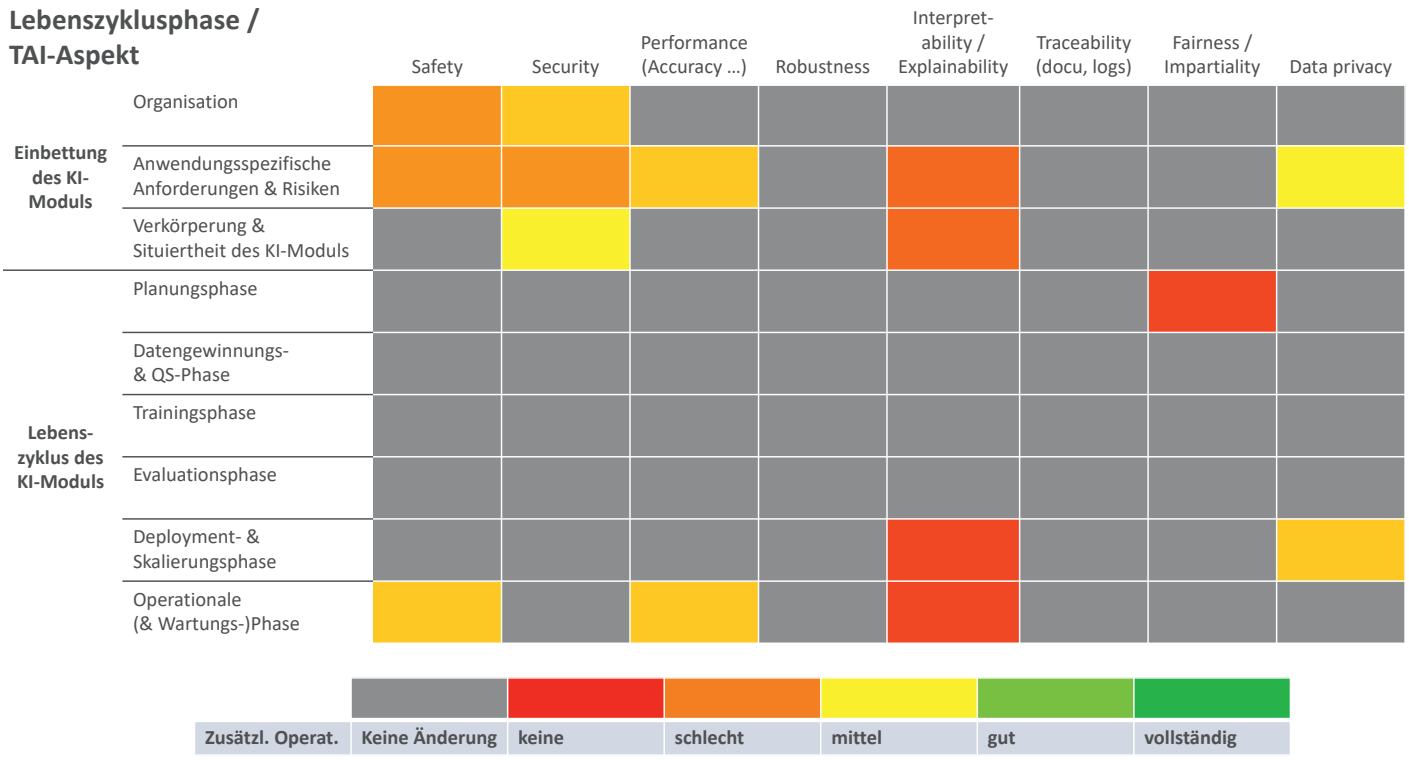


Abbildung 58: Stand der Operationalisierung für den Use Case Ridesharing (Quelle: Arndt von Twickel, Martin F. Köhler)

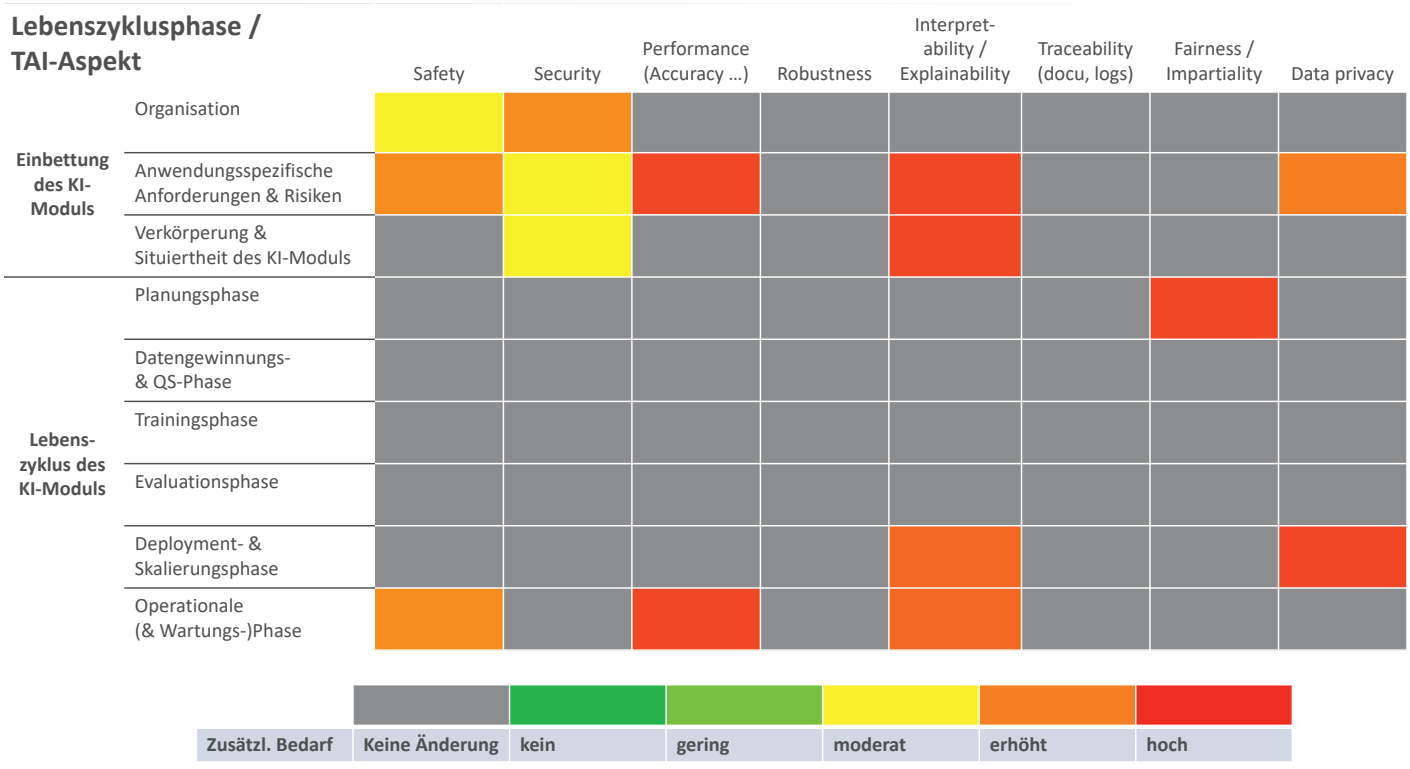


Abbildung 59: Handlungsbedarfe für den Use Case Ridesharing (Quelle: Arndt von Twickel, Martin F. Köhler)

Use Case Verkehrsoptimierung / Verbesserung der Lichtsignalanlagensteuerung (LSA) in der Verkehrsinfrastruktur

Abbildung 60 zeigt die Relevanzen der Kombinationen von TAI-Aspekten und Lebenszyklusphasen bzw. Einbettungsaspekten für den Use Case Lichtsignalanlagensteuerung (LSA) in der Verkehrsinfrastruktur.

Abbildung 61 zeigt den Stand der Operationalisierung der Kombinationen von TAI-Aspekten und Lebenszyklusphasen bzw. Einbettungsaspekten für den Use Case Lichtsignalanlagensteuerung (LSA) in der Verkehrsinfrastruktur.

Abbildung 62 zeigt die Handlungsbedarfe bezüglich der Kombinationen von TAI-Aspekten und Lebenszyklusphasen bzw. Einbettungsaspekten für den Use Case Lichtsignalanlagensteuerung (LSA) in der Verkehrsinfrastruktur.

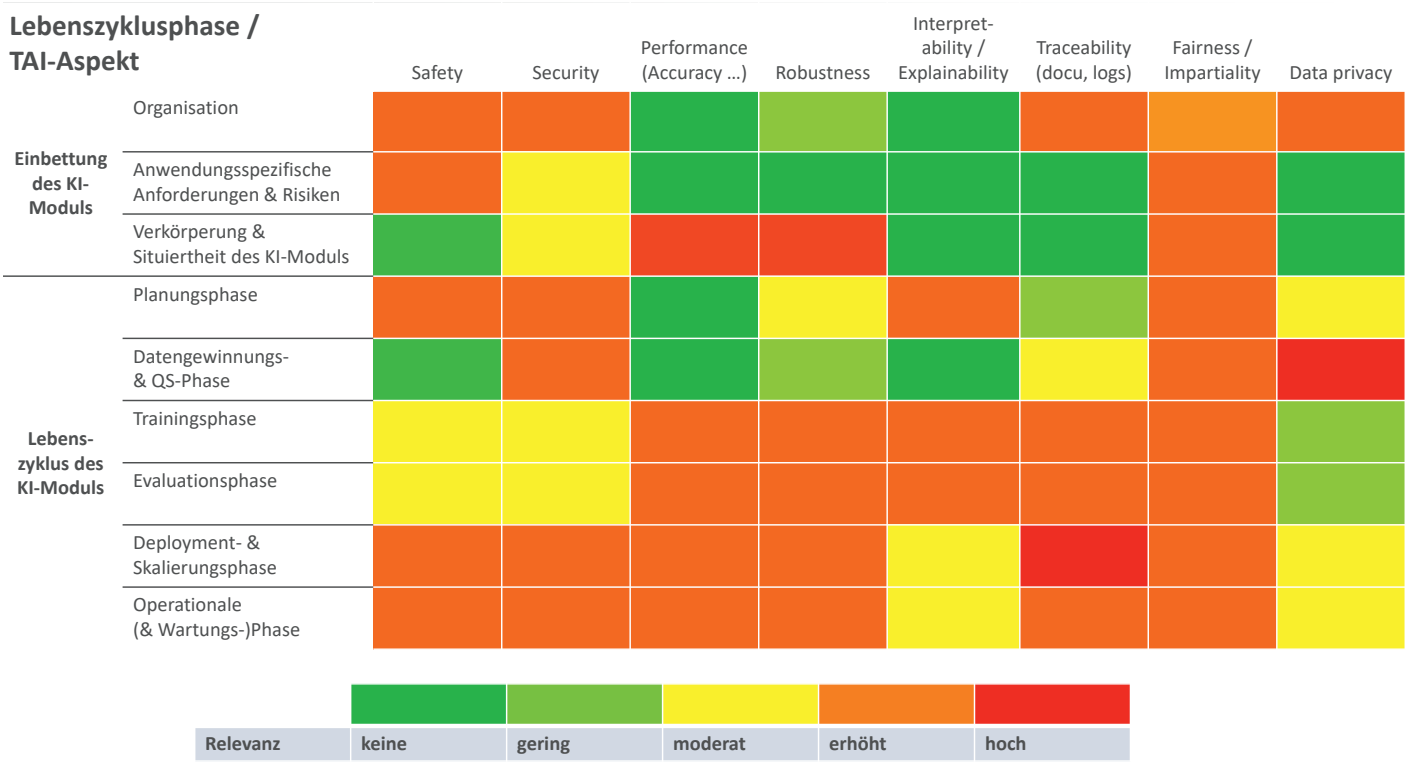


Abbildung 60: Relevanzen der Kombinationen für den Use Case Lichtsignalanlagensteuerung (Quelle: Arndt von Twickel, Martin F. Köhler)

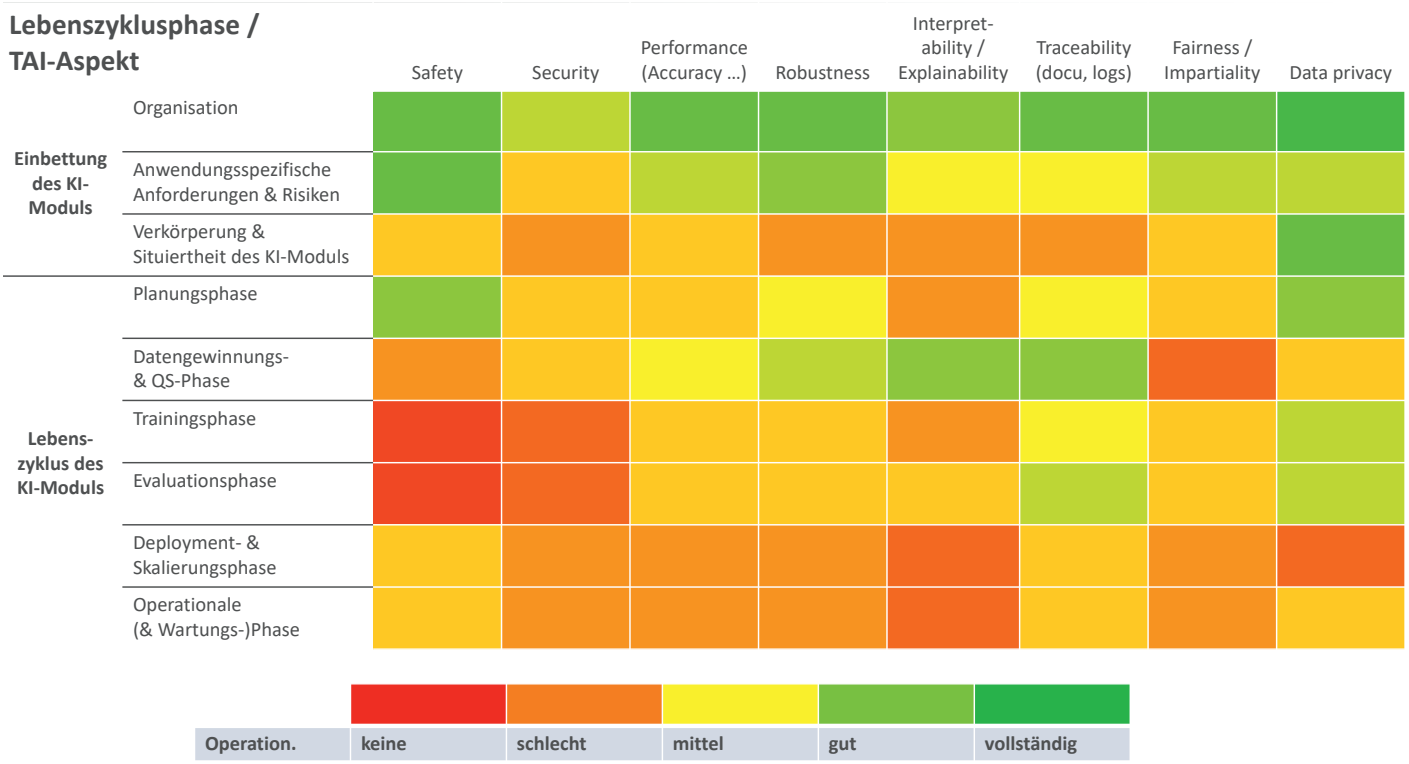


Abbildung 61: Stand der Operationalisierung für den Use Case Lichtsignalanlagensteuerung (Quelle: Arndt von Twickel, Martin F. Köhler)

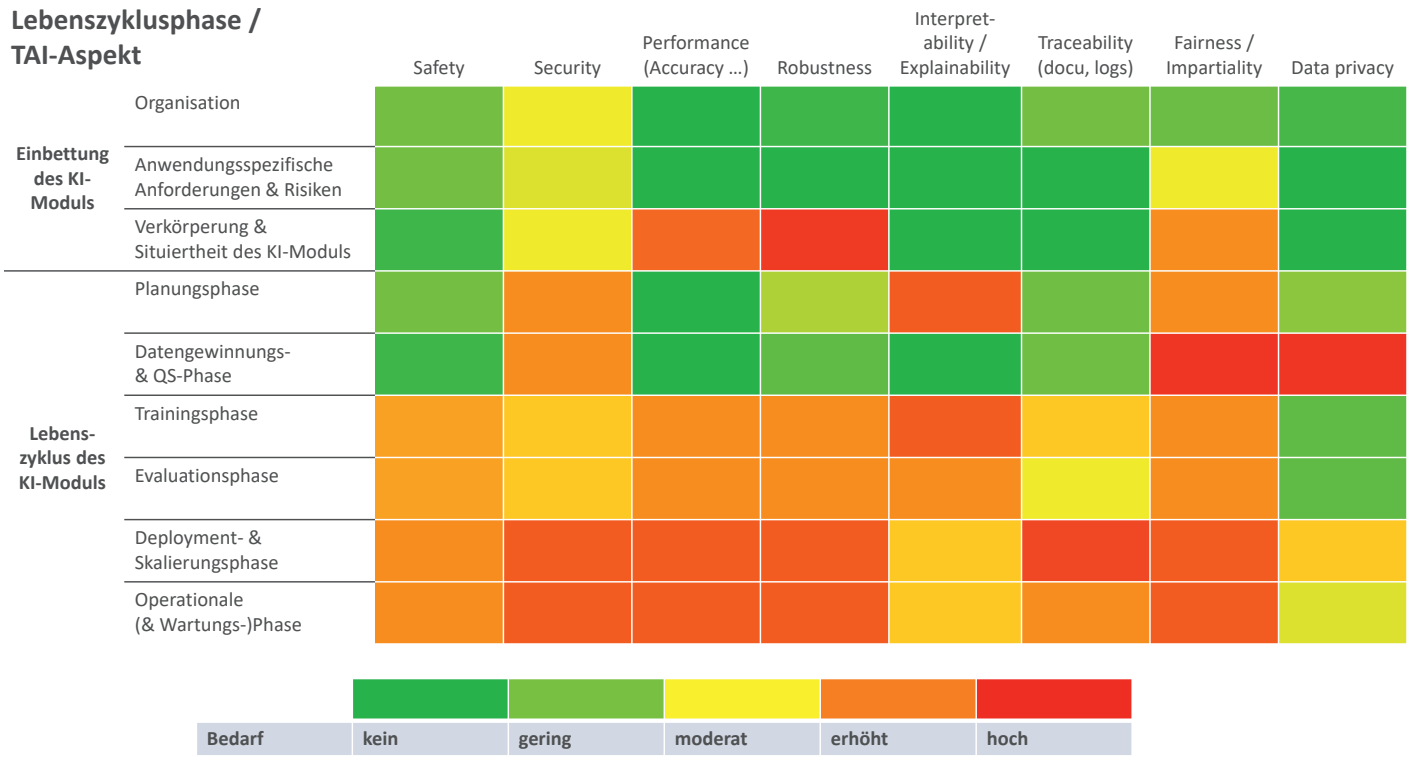


Abbildung 62: Handlungsbedarfe für den Use Case Lichtsignalanlagensteuerung (Quelle: Arndt von Twickel, Martin F. Köhler)

13.5 Anhang Medizin

Um einen Überblick über die Anwendungsbeispiele der Kapitel 4.7.2.1 bis 4.7.2.3 zu geben, wurden die drei KI-basierten medizinischen Anwendungen nach den folgenden vergleichenden Kriterien untersucht:

→ Akteur*innen (beteiligte Personen)
→ Ziel (Beschreibung des durch das Medizinprodukt gelöste Problem)
→ System (Beschreibung der Wirkungsweise des Medizinprodukts)
→ Voraussetzung (technische, organisatorische oder infrastrukturelle Voraussetzungen zur Leistungserbringung)
→ Auslöser (wodurch wird die Anwendung ausgelöst?)
→ Stakeholder (weitere an der KI-Anwendung interessierte Parteien)

Tabelle 18: Anwendungsfall 1: KI-assistierte 2-D-Röntgenbildanalyse zur Kariesdiagnostik in der Zahnmedizin

Akteur*innen	→ behandelnde Zahnärztin oder Zahnarzt → ggf. Spezialist*in (z. B. diagnostische Radiologie, Mund-Kiefer-Gesichtschirurgie) → ggf. überweisender Arzt/Ärztin → Medizinisch-technische*r Assistent*in → Patient*in → Entwickler*in → Gesundheitssystem ist indirekt beteiligt → Krankenkassen sind indirekt beteiligt
Ziel	Assistierte 2-D-Röntgendiagnostik mit dem Ziel der Zeitersparnis und erhöhter Reproduzierbarkeit diagnostischer Arbeitsabläufe. Ggf. zudem verbesserte Diagnostik und Therapiemöglichkeiten durch KI-gestützte 2-D-Röntgenbildanalyse (Benefit für alle beteiligten Akteur*innen)
System	Das System besteht in erster Linie aus einer Softwarekomponente, welche anatomische und ggf. pathologische Merkmale in 2-D-Röntgenbildern detektiert und für Zahnärzte visualisiert, d. h. auf dargestellten Bildern markiert. Eine solche Komponente kann beispielsweise als Backend Service in einer größeren Softwarearchitektur eingebunden werden. Die Eingabedaten für diese Komponente bestehen aus einem 2-D-Röntgenbild sowie Meta-informationen (z. B. Pixelgröße, Strahlendosis); die Komponente liefert Konturen (2-D-Polylines) und Annotationen für jede Kontur (je nach Objekt, z. B. Zahnnummer). Die Ausführung der Komponente wird durch das radiologische System veranlasst (Berechnung nach Verfügbarkeit eines neuen 2-D-Röntgenbildes), die numerischen Ergebnisse werden in einer Daten-bank gemeinsam mit den Patienten- und Bilddaten abgelegt. Die visuelle Darstellung dieser Ergeb-nisse erfolgt nach Öffnen des Datensatzes durch das behandelnde ärztliche Personal an einem an das System angeschlossenen Arbeitsplatz. Der/die behandelnde Arzt/Ärztin oder Spezialist*in begutachtet das dargestellte Ergebnis und nimmt ggf. manuelle Korrekturen vor. Derartige Korrekturen werden an die o. g. Datenbank über-tragen. Die nachgelagerte Befundung wird nach zahnärztlichem Standard auf Basis der Bilddaten und unter Zuhilfenahme der (ggf. korrigierten) KI-assistierten Informationen durchgeführt.
Voraussetzungen	→ Für dentale Röntgendiagnostik qualifizierte Praxis mit entsprechendem Personal und technischer Ausstattung /Infrastruktur. Dies beinhaltet: → Röntgengerät → MTA*in zur Durchführung des Röntgenbildes → Software mit „KI-Komponente“ → Arzt/Ärztin zur Befundung der 2-D-Röntgenbilder, der/die mit dem System vertraut ist

Auslöser	Die Komponente wird nach Bereitstellung eines neuen Datensatzes durch das System automatisch aufgerufen.
Stakeholder	→ Datenschutzbeauftragte*r: Die Komponente verändert den Patientendatensatz; Sicherheit der übertragenen und in der Datenbank abgelegter Daten muss anwendbaren Datenschutzrichtlinien genügen. → Entwickler*in → Krankenhaus mit IT-Abteilung → Krankenkassen → Benannte Stellen bzgl. Umsetzung der Konformitätsbewertung

Tabelle 19: Anwendungsfall 2: Beatmungsgerät mit KI-gestützter Entwöhnung

Akteur*innen	→ Patient*innen (hier: Mensch, pädiatrisch 15–35 kg, adult 35–200 kg) → Facharzt/Fachärztin für Anästhesiologie/Intensivmedizin → Intensivpfleger*in → Medizintechniker*in → Herstellender
Ziel	Anwendungsbeispiel: Beatmungsgerät mit KI-gestützter Entwöhnung Gerade die Coronazeit hat gezeigt, dass eine schonende, an die Patient*innen angepasste Entwöhnung entscheidend ist für die Rehabilitation und das nachhaltige Wohlbefinden. Ein weiterer Effekt ist die Reduzierung der Arbeitslast in der Intensivpflege, denn bei der üblichen klinischen Prozedur müssen die Beatmungsparameter bei der Entwöhnung je nach Zustand der Patient*innen immer wieder manuell angepasst werden. Bei einem automatisierten System werden geeignete Anpassungen in kürzeren Zeitabständen gemacht, dadurch wird die Entwöhnung insgesamt verkürzt und der Patient situativ besser unterstützt. Auch die Anzahl der körpernahen Kontakte von Intensivpflegekräften mit möglicherweise infektiösen Intensivpatient*innen wird reduziert. Im Vergleich zum vorexistierenden automatischen System, das auf klassischer KI-basiert, bietet das neue, auf neuronalen Netzen basierende System den Vorteil, von Intensivmediziner*innen und -pfleger*innen lernen zu können, um eine angemessene Reaktion auf immer mehr Ausnahmesituationen sicherzustellen und somit die derzeitige Flut von Alarmen auf die wirklich wichtigen Alarmsituationen zu reduzieren.
System	**1. Systembeschreibung** Das Entwöhnungssystem ist im Beatmungsgerät integriert und als neuronales Netz (NN) realisiert. Es handelt sich um ein „eingefrorenes“ NN, d. h. die Lernphase ist beendet, bevor das Gerät in den Markt gebracht wird. Die Aufgabe des Entwöhnungssystems ist, intubierte Patient*innen bei der Entwöhnung vom Beatmungsgerät adaptiv zu unterstützen. Die Patient*innen sind schon fähig, spontane Atemzüge zu initiieren, jedoch nicht kräftig genug, um gegen den Widerstand des Tubus anzukämpfen und ausreichend Luft zu bekommen. Deshalb werden sie mit einem positiven Atemwegsdruck unterstützt. Die Unterstützung soll nach und nach reduziert werden, um die Patient*innen an eine normale Atmung zurückzuführen. Sollten jedoch eine Verschlechterung der Atmung oder Stresssymptome eintreten, muss die Unterstützung wieder verstärkt werden. Das Entwöhnungssystem muss in der Lage sein, den Zustand der Patient*innen in Bezug auf die Beatmungsbedürfnisse korrekt einzuordnen (diagnostische Funktion) und die Atmungsunterstützung daran anzupassen (therapeutische Funktion).

Das NN besteht aus drei Layern: 1) Input-Layer, 2) Diagnose-Layer, 3) Output-Layer.

Im Input-Layer gibt es Knoten für Parameter, die am Anfang der Entwöhnung eingestellt werden, wie Patientenklasse (pädiatrisch/adult), Gewicht, Körpergröße, Anamnese (z. B. bei Patient*in mit COPD oder neurologischer Störung). Drei Input-Knoten werden mit Parametern gefüttert, die kontinuierlich vom Beatmungsgerät gemessen werden: f_{spn} (Frequenz der Spontanatmung), V_T (Tidalvolumen), $etCO_2$ (endtidale Konzentration von Kohlendioxid).

Im Diagnose-Layer gibt es acht Knoten, die jeweils mit einer Klassifizierung des Patientenzustands bezüglich der Atmung korrelieren. Die acht Zustände sind: normale Ventilation, Hyperventilation, Tachypnoe, schwere Tachypnoe, insuffiziente Ventilation, Hypoventilation, zentrale Hypoventilation, unerklärte Hyperventilation.

Im Output-Layer wird aufgrund der Diagnose die therapeutische Entscheidung getroffen. Hier gibt es drei Knoten: a) Verminderung der Druckunterstützung, b) Erhöhung der Druckunterstützung oder c) Alarmierung ohne Änderung der Druckunterstützung.

Nach einer Alarmierung speichert das System das Ausmaß der erforderlichen Druckkorrektur, die durch das Krankenhauspersonal vorgenommen wurde. Auch der Zustand aller Parameter bei der Alarmierung und nach der Korrektur wird gespeichert. Diese Daten werden dem Herstellenden über eine Datenschnittstelle direkt oder indirekt durch das Krankenhauspersonal verfügbar gemacht.

Die bei einer Alarmierungssituation von den Geräten im Feld erhobenen Parametersätze werden zum Anlernen bzw. Testen eines neuen, verbesserten Netzes NN2 beim Herstellenden genutzt. NN2 wird dann vom Herstellenden analysiert, mit NN verglichen, und wenn es nach gründlicher Nutzen-Risiko-Bewertung als geeigneter befunden wird, eingefroren und ggf. für eine neue Version des Systems genutzt. Ziele der Änderung können sein: Verkürzung der Entwöhnung, weniger und geringere Schwankungen der Performance, Reduzierung der Situationen, die zum Alarm führen, bessere Anpassung an ungewöhnliche Situationen, Beseitigung von festgestelltem Bias.

2. Prozess der Leistungserbringung

Ein*e langzeitbeatmete*r Patient*in ist so weit stabil, dass der Facharzt bzw. die Fachärztin den Anfang der Entwöhnung anordnet. Die automatische Entwöhnung wird manuell gestartet (Arzt/Ärztin/Pfleger*in auf ärztliche Anweisung).

Das System geht in die Phase „Anpassung“. Hier wird das NN verwendet, um regelmäßige Anpassungen der Atemwegs-Druckunterstützung nach unten (a) oder situativ bedingt nach oben (b) vorzunehmen.

Nur beim Output c) (siehe Systembeschreibung) wird alarmiert, dann interveniert der Intensivpfleger bzw. die Intensivpflegerin. Die relevanten Parameter und die manuell vorgenommene Druckkorrektur werden im System gespeichert und können für die Optimierung des NN an den Herstellenden anonymisiert weitergeleitet werden.

Unterschreitet die Druckunterstützung eine bestimmte Schwelle, die für die Tubuskompensation benötigt wird, geht das System in die Phase „Beobachtung“, die abhängig vom initialen Niveau der Druckunterstützung ein bis zwei Stunden andauert. Diese Phase entspricht einem automatisierten Spontanatmungstest. Die Anpassung der Druckunterstützung nach unten darf den Schwellwert nicht weiter unterschreiten. Die erfolgreiche Entwöhnung der Patient*innen meldet das System, wenn respiratorische Instabilitäten unterhalb 20 % der Beobachtungszeit, also unterhalb von 12 bis 24 Minuten, bleiben. Andernfalls geht das System zurück in die Phase „Anpassung“.

	Bei erfolgreicher Phase „Beobachtung“ wechselt das System in die Phase „Erhaltung“. Die Patient*innen werden weiterhin mit konstanter geringer Druckunterstützung beatmet, wobei kleinere Instabilitäten wie in der Phase „Beobachtung“ ausgeglichen werden. Nur bei häufigen oder längeren Instabilitäten wird die Meldung der Entwöhnung zurückgezogen und das System automatisch wieder in die Phase „Anpassung“ zurückversetzt. Dies soll aus Transparenzgründen und Gründen der „situational awareness“ ebenfalls gemeldet / haltend angezeigt werden. Während der Phase „Erhaltung“ wird empfohlen, dass der Arzt bzw. die Ärztin die Extubation zu einem beliebigen Zeitpunkt anordnet.
Voraussetzungen	→ Datenport am Beatmungsgerät und Internetanbindung des Krankenhauses, um relevante Parameter des Systems wie bei Alarmierungssituationen zum Herstellenden übertragen und als Datensätze für zukünftige Lern- oder Testphasen abspeichern zu können, sowie zur Auswertung von möglichem unerwünschtem Systemverhalten (BIAS). → Bereitschaft des Krankenhauses, die anonymisierten Benutzungsdaten und Parameter dem Herstellenden für die Weiterentwicklung des Systems zur Verfügung zu stellen. → Falls die Anonymisierung der Daten nicht ausreichend möglich ist, muss auch eine Einwilligung der Patient*innen bzw. ihrer Verwandten für die Nutzung der Daten eingeholt werden und der Herstellende die Vertraulichkeit der nicht anonymisierbaren Daten sicherstellen.
Auslöser	→ ärztliche Entscheidung, die Entwöhnung zu starten (kein Automatismus)
Stakeholder	→ Regulator → Datenschutzbeauftragte → IT-Direktor*in des Krankenhauses → Träger des Krankenhauses → Krankenkasse

Tabelle 20: Anwendungsfall 3: Segmentierung und Klassifikation von Gehirnarealen (inklusive Liquor) und deren Volumenbestimmung

Akteur*innen	direkt: Ärzt*innen (aus den Bereichen Radiologie, Neurochirurgie, Neurologie) (lösen Analyse aus, wertenErgebnisse aus, treffen darauf basierend Entscheidungen, Prognosen, Diagnosen) indirekt: Patient*in
Ziel	Das gewählte Anwendungsbeispiel löst das Problem zeitaufwendiger, manueller oder einfacher teilunterstützter Segmentierung von Strukturen im Bild. Ursprünglich sehr zeitaufwendige manuelle Arbeiten werden automatisiert und im klinischen Kontext mit höherer Genauigkeit und Wiederholbarkeit durchgeführt. Dies betrifft insbesondere auftretende inter- und intraindividuelle Abweichungen bei Wiederholungen. Es dient Ärzt*innen zur Unterstützung bei Diagnosen zu neurodegenerativen Erkrankungen.
System	Das Anwendungsbeispiel beschreibt eine KI-gestützte, vollautomatische Segmentierung aller relevanten Hirnareale anhand von 3-D-MRT-Daten (Magnetresonanztomografie). Die segmentierten Regionen werden volumetrisch quantifiziert und visualisiert. Die Ausführung der Berechnung erfolgt durch Empfang der Daten aus der radiologischen Infrastruktur (Picture Archiving and Communication System, PACS).

Voraussetzungen	technisch: Server-Infrastruktur, Verbindung zu Bild-Workstation bzw. PACS geeignete Geräte, um die 3-D-Bilddaten zu generieren (wie z. B. beim MRT 1,5 Tesla (T))
Auslöser	Auslöser der Segmentierung und Volumenbestimmung ist die Übertragung der Bilddaten an die radiologische Infrastruktur (PACS).
Stakeholder	→ Herstellender → Benannte Stelle (Konformitätsbewertung) → Aufsichtsbehörden → IT-Abteilung (technische Umsetzung) → Datenschutzbeauftragte (Übergabe der Daten an das System) → Krankenkassen (ggf. frühzeitige Erkennung von Krankheiten)

13.6 Anhang Energie/Umwelt

Um einen Überblick über die Anwendungsbeispiele der Kapitel 4.9.2.1 bis 4.9.2.6 zu geben, wurden die Anwendungsfälle wie folgt systematisch analysiert und aufgebaut:

→ Akteur*innen (beteiligte Personen)
→ Ziel (Beschreibung des zu lösenden Problems)
→ System (Beschreibung der Wirkungsweise)
→ Voraussetzung (technische, organisatorische oder infrastrukturelle Voraussetzungen zur Leistungserbringung)
→ Auslöser (wodurch wird die Anwendung ausgelöst?)
→ Stakeholder (weitere an der KI-Anwendung interessierte Parteien)

Tabelle 21: Anwendungsfall 1: Autonomes Smart Grid Power Management and Consumption System

Akteur*innen	→ Electrical Power Management System (PMS) Energy Provider → Electrical System Interface (SIF) Manager → Distributed Energy Resource (DER) Manager → Industrial Automation and Control System (IACS) Energy Consumer → Layered Communication IT Operator (Communication) → Value Stream Life Cycle Operator (Semantics) → AAS Asset Operator (Physics) → Digital Twin Operator (Analytics, Ursachen) → Data Manager (Learning, Wirkung)
Ziel	→ an Kundenwünsche oder ethische Anforderungen anpassbare (parametrisierbare) Produktion und Produkte (Asset/Value-Stream-Operatorrollen) → flexible (smarte) Erzeugung, Übertragung, Verteilung und Konsum von Energie (PMS-/IACS-/SIF-/DER-Managerrollen) → Sammeln, Darstellen und Erwerben von Wissen (Analystenrolle) → Strukturierung von Asset-Datenräumen (Datenmanagementrolle)

System	Industrielle Referenzmodelle wie SGAM oder RAMI4.0 beschreiben die Struktur von Systemen-von-Systemen. Zur Systemstruktur gehören a) die Ontologien der strukturellen (syntaktischen) Interoperabilität, b) die semantische Interoperabilität im Value Stream zwischen semantischen Domänen (conduits genannt) während des Life Cycle und c) die physischen Hierarchien, d. h. Nutzungsstruktur (Zonen genannt) des betrachteten Assets und seiner AAS. Die Leistung in SGAM-Systemen ist die effektive Energieversorgung industrieller Produktionsanlagen (nach RAMI4.0) oder individueller Verbraucher*innen. Von der Erzeugung bis zum Verbrauchenden nimmt die Energie im Value Stream (SGAM x-Achse) verschiedene heterogene Formen, je nach Medium, das sie tragen muss, an. Man spricht hier auch von heterogenen Modellen, die semantisch interagieren müssen. Diese Energieträgermedien können das Wettergeschehen sein, falls Wind und Sonne als volatile Quellen infrage kommen. Energieträger- bzw. Energieerzeugermedien sind mechanischen Windgeratoren oder Fotovoltaik-Geräte, zur elektrischen Energieübertragung werden weitreichende Hochspannung-Gleichstrom-Übertragungsnetze oder lokale Wechselspannungsnetze benötigt. Sogenannte verteilte Energieressourcen (DER) dienen der Energieverteilung im SGAM-Netz. Und am Schluss steht die Abhängigkeit der Verbraucher*innen von seinen Produktionsanlagen, von seinem Verhalten beim Energiekonsum, von der Verfügbarkeit der elektrischen Energie und nicht zuletzt von der „ethischen Qualität“ der gehandelten Energie. Die Leistung in RAMI4.0-Systemen ist die effiziente Produktion eines Produkts. Im Value Stream (x-Achse RAMI4.0) zur Herstellung eines Produkts nimmt das Produkt, ähnlich der Energie, verschiedene heterogene Formate, je nach Life-Cycle-Zustand, an. Diese Formate unterscheiden sich grob in Typisierung und Instanziierung des Produkts. Beide, Typisierung und Instanziierung, zeichnen sich durch eine Entwicklungs- und Benutzungsphase aus. Diese sequenziellen Produktionszustände können mit verschiedenen Modellen, die semantisch interagieren müssen, modelliert werden. Es gibt also aus semantischer Sicht durchaus eine Vergleichbarkeit des Value Stream in den Referenzarchitekturmodellen bei der Modellierung der Zustände und Transformation der Eigenschaften von Produktentwicklung oder Energieversorgung.
Voraussetzungen	Um weitgehend autonom (bzw. automatisch) zu funktionieren, sind sowohl für Produktionsanlagen als auch für Energieversorgungssysteme Infrastrukturmaßnahmen erforderlich, die es ermöglichen, Information aus der Umgebung (von außen) als auch von der Einbettung (von innen) zu erhalten und verarbeiten zu können. Während die äußere Umgebung unbekannt ist und daher erst gelernt werden muss, ist die Einbettung a priori bekannt und kann als Modell in den Value Stream eingespeist werden. Bekannte Verfahrensmodelle können automatisiert, unbekannte Verfahrensmodelle können mit geeigneten ML-Methoden gelernt und zur Autonomisierung der Energieübertragung oder Produktion verwendet werden. Automatisierung und Autonomisierung können beide zur Systemkontrolle miteinander kombiniert werden. Dabei stellt die Automatisierung einen geschlossenen und die Autonomisierung einen offenen Kreislauf dar. Die Adjektive „offen“ und „geschlossen“ bezeichnen Kreisläufe, die nach außen als auch in ihre Umgebung offen oder geschlossen sind bezüglich der Aufnahme von Informationen.
Auslöser	In SGAM-Systemen ist die Triggering-Funktion z. B. die überraschende Verfügbarkeit volatiler Energie aufgrund des Wettergeschehens. Vor der Zuschaltung der zusätzlichen Energie müssen die Netzwerkmanager die Stabilität des elektrischen Versorgungsnetzes unter den gegebenen und den veränderten Zuständen zeitgleich messen, analysieren und entscheiden. Diese Entscheidung hat Auswirkungen auf die nachgelagerten Netze bzw. „Energieträgermedien“ zu Erzeugung, Übertragung, Verteilung, Verbrauch. Alle Netze bzw. Medien müssen harmonisieren, d. h. aufeinander abgestimmt werden, um Instabilitäten zu vermeiden.

Stakeholder	Die automatisierte oder autonomisierte Kontrolle von Energieversorgungsnetzen als kritische Infrastrukturen von Krankenhäusern, öffentlicher Sicherheit, Verkehrskontrolle, Wettervorhersagen etc. liefert eine Fülle weiterer Abhängigkeiten von Stakeholdern kritischer Infrastrukturen von der zuverlässigen und ethisch korrekten Energieversorgung.

Tabelle 22: Anwendungsfall 2: Energieeffizienz in Gebäuden und Kopplung mit Energienetzen

Akteur*innen	→ Gebäudebetreiber*innen → KI-Entwickler*innen
Ziel	→ Durch den stärkeren Einsatz von Erneuerbaren schwankt die Stromproduktion mit dem Wetter. → Energie muss stärker dann verbraucht werden, wenn sie erzeugt wird. → Gebäude sollen als flexibler Energieverbraucher im Stromnetz genutzt werden. → Energienutzung im Gebäude soll optimiert werden (40 % der Energie wird in Gebäuden genutzt). → intelligente Steuerung von Klimaanlage, Heizung, Warmwasserbereitung und Ladestationen
System	Das KI-System wird an die Gebäudesteuerung sowie an Wetter- und Energienetzdaten angebunden. Mithilfe dieser Daten generiert es täglich eine Vorhersage über die Gebäudenutzung sowie über die Verfügbarkeit von erneuerbaren Energien. Ein intelligenter Algorithmus steuert dann das Gebäude so, dass es Energie hauptsächlich in Zeiten mit hoher Verfügbarkeit von Erneuerbaren nutzt. **Komponenten:** Im Gebäude → Temperatur-, Feuchtigkeit-, CO2-Sensoren → Building Management System → Steuergeräte von Klimaanlagen, Heizung, Ladestationen etc. Cloud/Internet → Dashboard-App: zeigt Sensorwerte an, nimmt User-Input auf, um die Grenzen des KI-Systems festzulegen → KI-Backend: ermittelt Prognosen und Steuerbefehle → Energienetzdatenschnittstelle → Wetterdatenschnittstelle Prozess: → Jeden Tag um Mitternacht: → Abrufen des Wetterberichts → Abrufen der Energienetzvorhersage über Verfügbarkeit von Erneuerbaren → Durchführen der Gebäudenutzungsprognose (basierend auf historischen Nutzungsdaten) → Simulation und Optimierung der Gebäudeenergienutzung → Erstellen eines optimierten 24-h-Plans für die Steuerung von flexiblen Geräten Während des Tages: → Steuerung des Gebäudes basierend auf dem erstellten Plan → Anpassung des Plans an Echtzeitänderungen
Voraussetzungen	→ Integration der Software in das Gebäude, Schaffung von Schnittstellen zum Gebäudemanagementsystem → Integration von Energienetz und Wetterdaten → Trainieren der ML-Modelle auf die Gebäudedaten → Aufsetzen der Software auf einem Cloudsystem

Auslöser	→ automatischer Trigger jeden Tag um Mitternacht, um einen neuen Plan zu erstellen
Stakeholder	→ Energienetzbetreiber → Datenschutzbeauftragte (Vorhersage der Raumbelegung teilweise kritisch)

Tabelle 23: Anwendungsfall 3: Personalisierte, KI-gestützte Empfehlungssysteme für nachhaltigen Konsum

Akteur*innen	Endkonsument*innen, Handel und jegliche Akteur*innen der Wertschöpfungs- und Lieferkette, Datenhub zum DPP und Betreibende der KI
Ziel	→ **Problem:** Die Intransparenz und Unübersichtlichkeit produktbezogener Nachhaltigkeitsinformationen beim Einkauf (z. B. „Labeldschungel“) stellen ein Hindernis für nachhaltigen Konsum dar ([418], [419]). Konsument*innen wählen hierdurch Produkte aus, die nicht mit ihren individuellen Einstellungen, u. a. Nachhaltigkeitspräferenzen, übereinstimmen. Diese Problematik betrifft den stationären Handel und den Onlinehandel. → **Lösungsteil a) Personalisierte Empfehlungssysteme für nachhaltigen Konsum durch KI-gestützte Assistenzsysteme bei konkreten Einkaufsentscheidungen** ([424], 54) (u. a. auf Basis umweltbezogener Lebenszyklusdaten, ggf. als Daten des DPP und persönliche Präferenzen). Hierdurch kann die persönliche Relevanz von Produkten im Sinne einer „**bedeutsamen Produktberatung**“ ([420], 12) erhöht werden, da die KI persönliche Einstellungen, Präferenzen und Bedürfnisse mit den Produkteigenschaften abgleichen könnte ([424], 255, [524], 18). Ein Abgleich der individuellen Nachhaltigkeitspräferenzen mit einer Produktdatenbank und daraus erzeugten personalisierten Produktempfehlungen ermöglicht u. a. eine Nachhaltigkeitsoptimierung der Produktauswahl. → **Lösungsteil b) Strategische Empfehlungen bei der Konsum-/Einkaufsplanung** zur mittel- bis langfristigen **bedarfsgerechten Optimierung von Konsummustern** (aufbauend auf **a)**). Durch Erfassung und Interpretation persönlicher Einkaufsdaten könnte das Empfehlungssystem wichtige **Einsichten über das Konsument*innenverhalten** gewinnen. Die KI könnte Produktnachfragen prognostizieren, Vielverbrauch flaggen und daraus abgeleitete relevante Alternativen vorschlagen. • Beispiel: Kauft Konsument*in A zwei Pakete Butter pro Woche, so könnte ein Alternativvorschlag des Empfehlungssystems sein, bei zukünftigen Einkäufen die Großpackung zu wählen (weniger Verpackungsmaterial, preisgünstiger, Verminderung der Einkäufe). → Zudem können Akteur*innen der Lieferkette weitere Informationen (z. B. Nährwertangaben, empfohlene Tagesbedarfe, Haltbarkeit) in das Empfehlungssystem einspeisen, was eine **Optimierung der Konsummuster** ermöglicht ([420], 24). Beispiel: Konsument*in A aus dem vorigen Beispiel konsumiert vergleichsweise CO_2-intensivere Produkte und würde zudem von einer weniger fettigen Nahrung gesundheitlich profitieren. • Beispiel: Das KI-gestützte Empfehlungssystem könnte vorschlagen, nun anstelle der zwei Kleinpackungen Butter eine Großpackung pflanzlicher Margarine zu wählen. In Summe können so Konsummuster beeinflusst und nachhaltig verändert werden.

System	**Technisches System:** → Emissionen, Ressourcen – Prädiktion: ermittelt mithilfe von regressiven neuronalen Netzen den Verbrauch von Ressourcen und die Emissionen für ein Produkt → Empfehlungssystem: empfiehlt Nutzer*innen, basierend auf ihrem und dem Verhalten anderer, nachhaltige Produkte Personalisierte, KI-gestützte Empfehlungssysteme könnten u. a. nach der Logik folgender Verhaltensänderungsmechanismen funktionieren: → **Coercive Intervention/Auswahleinschränkung**: Ausschluss nicht nachhaltiger Optionen durch KI mit dem Nachteil, dass keine Selbstreflexion der Konsument*innen unterstützt wird ([423], 12) → **Persuasive Intervention/Nudging:** Anstöße zur Entscheidung für nachhaltige Produkte/Dienstleistungen innerhalb der bestehenden Auswahl [(Thorun et al. (2017), 48f.)], wobei allerdings die moralische Entscheidungsmacht über die Auswahl zu empfehlender Produkte den Designer*innen/Forscher*innen obliegt ([423], 12). Im Projekt Green Consumption Assistant werden Scraping-Technologien genutzt, um eine Datenbank mit nachhaltigen Produkten zu entwickeln [418]. **Reflektierte Intervention/Feedback:** Nachhaltige Konsumentscheidungen werden aufgrund zurückgespielter Daten/Feedbacks (z. B. CO_2-Emissionen des letzten Einkaufs) getroffen ([423], 12).
Voraussetzungen	**Technische Voraussetzungen:** Datenverfügbarkeit, standardisierte Schnittstellen und strukturierte Datenformate, Entwicklung der Algorithmen, Cloudinfrastruktur zur Ausführung, App zur Anzeige der Empfehlungen **Organisatorische Voraussetzungen:** Einweisung von Kund*innen zur Nutzung, Einweisung von Akteur*innen der Lieferkette zur Einspeisung von Informationen, Klärung der Frage, welche Institution die KI betreibt **Datengrundlage:** 1. produktbezogen (ggf. DPP): nachhaltigkeitsbezogene Eigenschaften (Regionalität, Labels, Obsoleszenzdaten, Fischfangmethoden, Inhaltsstoffe, Allergene, Mutterunternehmen) ([420], 20). 2. konsument*innenbezogen: Clustering von Konsument*innen (als KI-Trainingsgrundlage), Kaufverhalten, Lebensstil, Ernährungsweise
Auslöser	Trigger für das KI-System können z. B. sein: → Scannen / Kauf eines Artikels → Tägliches Retraining der Algorithmen, basierend auf neuen Daten
Stakeholder	Use Case Akteur*innen, Krankenkassen, Forschungseinrichtungen, Umwelt, Verbraucherschutz, Datenschutzbeauftragte

Tabelle 24: Anwendungsfall 4: Skalierbare Bestimmung von Umweltwirkungen im Gebäudesektor

Akteur*innen	→ Anwendende in • Gebäude- und Quartiersplanung • Formulierung politischer Rahmenbedingungen für bauliche Fördermaßnahmen → Server- und DB-Maintainer (Datenbank)
Ziel	Die ökologische Lebenszyklusanalyse von Gebäuden und Quartieren erfordert eine hohe Informationsbreite und -tiefe. Dies impliziert einen hohen Zeit- und Rechenaufwand in der Bestimmung der Umweltwirkungen (vgl. [425]). Mit Methoden des Maschinellen Lernens können Richtwerte ermittelt und genutzt werden, die eine hinreichende Aussage zum Footprint des Gebäudes/Quartiers sowie mögliche Umweltoptimierungen liefern. Hieraus resultiert eine signifikante Zeitersparnis.
System	→ Frontend • zum User-Input (Anwendende) von Eckdaten zu Gebäude/Quartier, die in frühen Planungsphasen grundsätzlich vorhanden/öffentlich verfügbar sind und im ML-Modell höchste Entropie/Informationsdichte aufweisen • zur Anzeige des ermittelten Footprints und der statistischen Unsicherheit → Backend • zur Ermittlung des Footprints anhand des übertragenen User-Inputs mittels Abgleich mit DB (Model Output) • zur Anonymisierung der Inputdaten bei hohem Detailgrad und Einbindung als Trainingsdaten in das ML-Modell • Systemkomponente zur Ermittlung von Lern-/Skaleneffekten im „Goal“ • Systemkomponente zu Tracking und ggf. Konformitätsprüfung der Datennutzungsdauer • einheitliche Datensystematik (Ontologie, Semantik)

Voraussetzungen	→ technisch/infrastrukturell
	• Server(s) mit vortrainiertem ML-Modell und DB für Model- Input/-Output
	• REST-API (Datenschnittstelle)
	  • für User Input Queries
	  • ggf. Anbindung an ML-Server
	  • Rückspielen des passenden Outputs aus DB an Frontend
	• Frontend als Webservice/Plugin/...
	→ organisatorisch
	• prozessuale Einbindung des Use Cases in den Planungsprozess von Gebäuden/Quartieren bzw. Neubauten/Sanierungen

Auslöser	→ Eingabe/Upload von Eckdaten/Dateien in Frontend
Stakeholder	→ Datenschutzbeauftragte bezüglich Anonymisierungs- und Aggregationsgrad des ML-Inputs und Outputs → Data-Science-Akteur*innen → Simulation Scientists

Tabelle 25: Anwendungsfall 5: Ressourcenintensität von KI & ML

Akteur*innen	→ Anwendende von Systemen mit KI-Elementen
Ziel	Künstliche Intelligenz und Modelle des Maschinellen Lernens benötigen per definitionem eine signifikante Menge an Daten(-verarbeitung) zur Mustererkennung. Dies bedingt eine tendenziell hohe Rechenlaufzeit und -leistung. Hieraus resultieren ein hoher Energieverbrauch und zugehörige Umweltwirkungen (vgl. [398]). Für die beabsichtigte Anwendung von KI-Elementen soll daher anhand eines Metasystems geprüft werden, ob das konzipierte KI-System/ML-Modell Optimierungsbedarf hinsichtlich der benötigten Datenmenge und Algorithmik bzw. Laufzeit hat.
System	→ Frontend ● zum User Input (Anwendende) über die angedachte Systematik (Datenmenge//Algorithmik//technisches Setup, ML-Server//...) ● für Feedback • prognostizierte Umweltwirkung des Systems/KI-Elements im System • Optimierungsvorschläge zur Systematik auf Basis von Ensemble-Learning-Ergebnissen/gesicherten Forschungsergebnissen → Backend ● Ensemble-Learning-Server ● Abgleich der Eingabeparameter mit DB • zur Umweltwirkungsprognose • zur Findung von Optimierungsvorschlägen auf Basis bisheriger Ensemble-Learning- / gesicherter Forschungsergebnisse
Voraussetzungen	→ technisch/infrastrukturell ● (Server für) DB mit kategorisierten Forschungsergebnissen/Vergleichswerten, Umweltindikatoren zur Berechnung der Umweltwirkung ● Frontend als Webservice/Plugin/... ● REST-API (Datenschnittstelle) • für User Input Queries • Rückspielen des Outputs aus Backend an Frontend • ggf. Anbindung an Ensemble-Learning-Server → organisatorisch ● prozessuale Einbindung der Use Cases in Konzeption von KI-Anwendungen ● gesteigerte Aufmerksamkeit für Ressourcenintensität von KI-Anwendungen ● Aufbau einer einheitlichen Metrik und/oder eines Referenzsystems zur Vergleichbarkeit
Auslöser	→ Eingabe/Upload von User Input in Frontend (manuell/automatisiert)
Stakeholder	→ Datenschutzbeauftragte → Data-Science-Akteur*innen

Tabelle 26: Anwendungsfall 6: Adversarial Resilience Learning – Marktlicher Angriff durch Aggregatoren im Verteilnetz

Akteur*innen	→ Verteilnetzbetreiber: bietet einen lokalen Markt, um Netzengpässe aufzulösen. → Konsumenten/Prosumer: (reguläre) Teilnehmende am lokalen Energiemarkt → „Angreifer“: Konsumenten/Prosumer, die Marktregeln zu ihrem Vorteil nutzen wollen; zur Effektivitätssteigerung konzentriert an einem Strang (ggf. auch KI-basiert als „automatischer Angreifer“ am Markt) → „Verteidiger“: Lernendes Agentensystem zur Detektion marktbasierter Angriffe
Ziel	Im Kontext des Verteilnetzbetriebs kann es in marktlicher Situation dazu kommen, dass Absprachen und Koordination von Anlagen mit dem Wissen über den Netzzustand so verschränkt werden, dass Gamification auftreten kann: Die Aggregatoren optimieren ihr koordiniertes Verhalten so, dass sie gegenüber den Netzbetreibern einen Engpass erschaffen, den sie, natürlich gegen entsprechende Vergütung durch Dritte als Incentive, auch selber wieder auflösen können. Die Marktregeln werden also so „ausgenutzt“, dass künstliche Probleme bedingt durch den Anreiz der Vergütung ausgelöst und dann selber gegen „Entgelt“ beseitigt werden. Angebote und Nachfrage werden hier künstlich geschaffen. Ziel: KI als Angreifer (Marktteilnehmende mit Malicious Intent Behaviour) simuliert die Gamification des Marktes, zeigt Schwachstellen auf und hilft, das Risiko zu ermitteln. Ziel: KI als Verteidiger lernt, die Verhaltensmuster der Aggregatoren/bösartigen Marktteilnehmenden zu entdecken, und kann so (1) entweder als Assistenzsystem/Detektor fungieren oder (2) Maßnahmen einleiten, um die Gamification zu verhindern (als direkter Akteur*innen Clearing House; je nach Marktentwurf).
System	Netzengpässe sind ein mittlerweile dauerhaft auftretendes Problem der VNB (Verteilnetzbetreiber). Redispatch-Prozesse greifen zunehmend auf Klein(st)anlagen zurück (aktuell ab 100 kW), während kaum ein Verteilnetz ausreichend mit Sensorik und Aktorik ausgestattet ist bzw. die IKT der VNB nicht gerüstet ist, um eine zentral gesteuerte Auflösung dieser zunehmend komplexen Situation zu erreichen. Als Alternative werden zunehmend lokale Märkte in Betracht gezogen bzw. sogar implementiert, die durch eine Form der Selbstorganisation das Komplexitätsproblem adressieren. Im einfachen Marktentwurf ermittelt der VNB eine Engpasssituation an einem Strang (Radial Feeders, nicht vollvermaschtes Netz). Er nutzt den Flexibilitätsmarkt, um durch finanzielle Anreize die lokalen Konsumenten zur Lastreduktion/Lastverschiebung zu animieren. In einer Variante des Anwendungsfalls kann das Ziel die möglichst hohe Eigenversorgung des Strangs sein, sodass Prosumer zur entsprechenden Einspeisung über Preissignale animiert werden. Der Netzbetreiber kann bilaterale Absprachen zwischen den lokalen Teilnehmenden typischerweise nicht detektieren. Lastrampen können von den Teilnehmenden künstlich ausgelöst/forciert werden, indem beispielsweise „Electric Vehicle“ zu bestimmten Zeitpunkten geladen werden. Somit wird zwar der Engpass detektiert, der Auslöser ist dem VNB aber offiziell unbekannt. Da jeder Marktteilnehmer aber natürlich das Laden seines EV (bzw. seinen Lastbedarf generell) genauso auch wieder reduzieren kann, kann eine Koalition von Marktteilnehmern fast beliebig Geld vom VNB beziehen, ohne dafür eine echte Gegenleistung als „Opfer“ in Form einer „echten“ incentivierten Verhaltensänderung erbringen zu müssen. Da sich die Koalitionen meist dynamisch bilden und die Veränderungen in den Randverteilungen nicht detektierbar sind, insbesondere, weil der Grund unbekannt ist – es existieren natürlich auch genauso viele valide, kontextfreie Gründe –, ist dieser Angriff zulasten des VNB nicht mit bisherigen Mitteln in der Leittechnik detektierbar.

Voraussetzungen	Erfassen der Messdaten, Szenariendefinition, Algorithmik über die Marktprozesse, Topologiedaten
Auslöser	Engpässe im Netz treten unerwartet auf und werden nach Preissignalen am Markt plötzlich wieder gegen Incentive aufgelöst. Gehäufte vermutete Gamification, KI reagiert autonom auf Basis eines festgelegten Thresholds an Ereignissen bzw. Grenzwerten in einem Log eines Systems.
Stakeholder	Distribution System Operator (Verteilnetzbetreiber), Netzbetreiber, Flexanbieter, Aggregatoren

Abbildungsverzeichnis

Tabellenverzeichnis

DIN e. V.

Am DIN-Platz
Burggrafenstraße 6
10787 Berlin
Tel.: +49 30 2601-0
E-Mail: presse@din.de
Internet: www.din.de

DKE Deutsche Kommission Elektrotechnik
Elektronik Informationstechnik in DIN und VDE
Merianstraße 28
63069 Offenbach am Main
Tel.: +49 69 6308-0
Fax: +49 69 6308-9863
E-Mail: dke@vde.com
Internet: www.dke.de

Stand: Dezember 2022