Gefördert durch:

Bundesministerium
für Wirtschaft
und Energie

aufgrund eines Beschlusses
des Deutschen Bundestages

WHITEPAPER

ETHIK UND KÜNSTLICHE INTELLIGENZ:

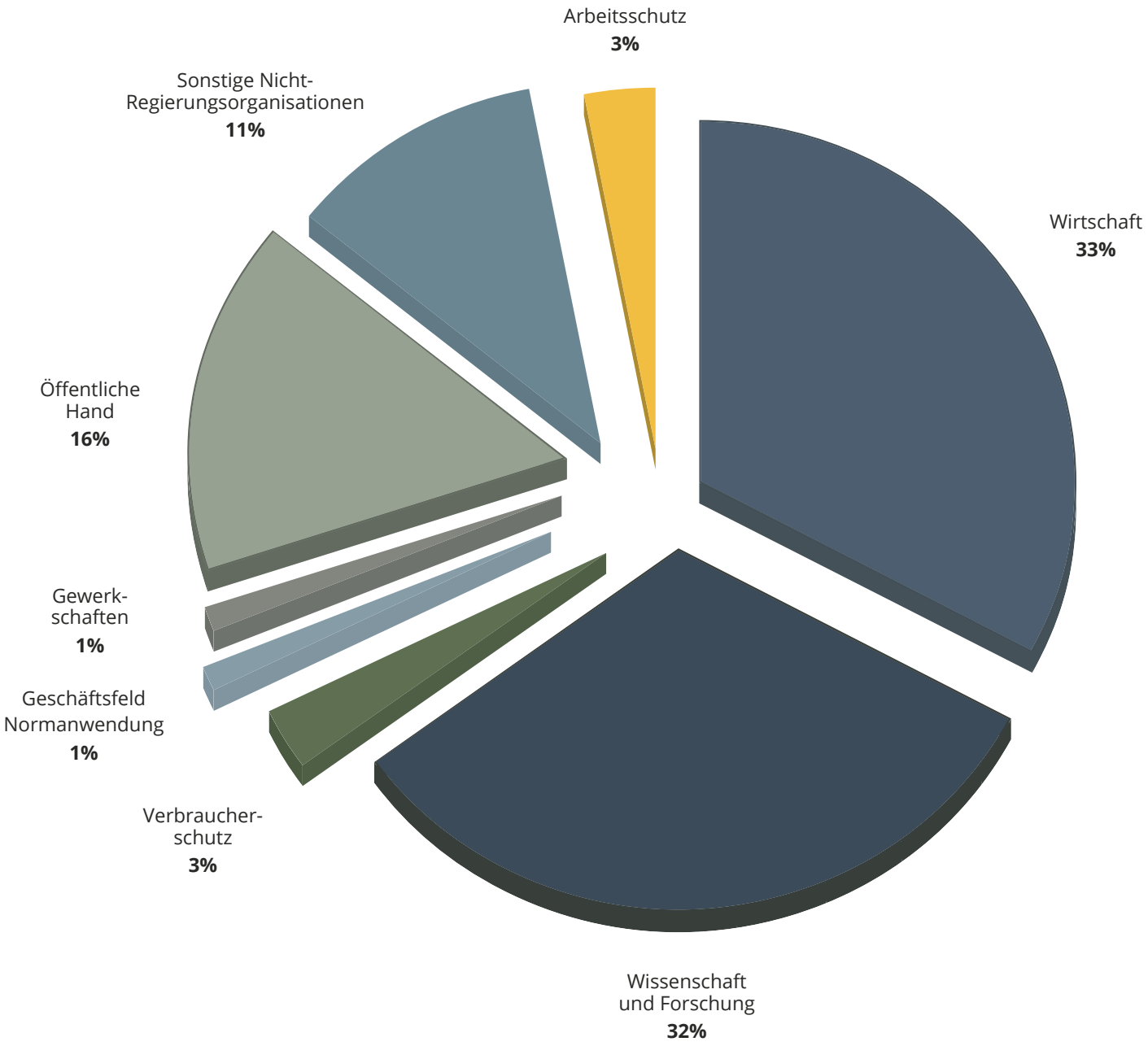
Was können technische Normen und Standards leisten?

DKE VDE DIN

DIN

Abbildung 1: Interessierte Kreise, die an dem Projekt mitgewirkt haben

2



PROJEKTBESCHREIBUNG

Wie können technische Standards ethisch wertvolles Verhalten einer autonomen Maschine, wie beispielsweise eines autonomen Fahrzeugs, die von einer KI gesteuert wird, sicherstellen?

Das vom Bundesministerium für Wirtschaft und Energie geförderte Projekt „Ethikaspekte in der Normung und Standardisierung für Künstliche Intelligenz in autonomen Maschinen und Fahrzeugen“ betrachtet die Zusammenhänge zwischen Ethik und Künstlicher Intelligenz (KI) und welche Rolle technische Normen und Standards in diesen Zusammenhang spielen können. Dabei fokussiert sich das Projekt auf die Bereiche autonome Maschinen und Fahrzeuge. Ziel des Projektes war es, zwischen November 2018 und April 2020 ein Projektergebnis zu erarbeiten, das den aktuellen Stand der Normung und Standardisierung im interdisziplinären Themenfeld KI und Ethik darstellt, sowie mögliche zukünftige Handlungsfelder der Normung und Standardisierung aufzeigt. Das Deutsche Institut für Normung, DIN und die Deutsche Kommission Elektrotechnik, DKE betreuten dieses Vorhaben gemeinschaftlich und erarbeiteten gemeinsam mit Fachexperten und Fachexpertinnen in Workshops die Inhalte des Dokuments.

Dabei fokussiert sich das Projektteam auf drei inhaltliche Schwerpunkte im Anwendungsfeld autonome Maschinen und Fahrzeuge:

1. Festlegen von klaren Anwendungsbereichen: Wie kann das Festlegen von klaren Anwendungsbereichen für eine autonome Maschine oder ein autonomes Fahrzeug zur Berücksichtigung von Ethik in technischen Normen und Standards führen?
2. Überpüfbarkeit von autonomen Maschinen/Fahrzeugen: Welche Voraussetzungen für ethisches Verhalten einer autonomen Maschine/Fahrzeug können überprüft werden?
3. Ausgestaltung der Mensch/Maschine-Schnittstelle mit dem Ziel Verantwortlichkeiten festzulegen bzw. Verständnis zwischen Mensch und Maschine sicherzustellen: Wie können Normen und Standards zu einer verbesserten Interaktion zwischen Mensch und Maschine beitragen?

Das Projekt startete Anfang April 2019 mit einer Auftaktveranstaltung. Im ersten Schritt analysierte das Projektteam Ethikrichtlinien aus Politik, Wissenschaft und Wirtschaft hinsichtlich ethischer Anforderungen an KI, welche Anforderungen Normung und Standardisierung adressieren und welche der Anforderungen relevant für automatisierte Maschinen und Fahrzeuge sind. Das Ergebnis dieser Analyse ist eine Liste mit zehn relevanten Attributen. Diese wurden im Juli 2019 im Rahmen des ersten Workshops mit Experten und Expertinnen diskutiert und im zweiten Schritt mit den Projektschwerpunkten verknüpft. Die Zwischenstände dieser Erarbeitung, wurden zur Kommentierung auf DIN.one[1] veröffentlicht. Der dritte finale Workshop fand im Januar 2020 stand. Hier wurden die eingegangenen Kommentare erneut in einem Kreis aus Expertinnen und Experten diskutiert. Anschließend begann die finale Ausarbeitungs- und Gestaltungsphase des Whitepapers. Insgesamt haben 77 Experten aus 8 verschiedenen interessierten Kreisen mitgewirkt. Am stärksten waren die Bereiche Wissenschaft und Forschung und die Wirtschaft mit jeweils ca. 30 Prozent vertreten. Auch die öffentliche Hand hat mit ca. 15 Prozent einen starken Beitrag geleistet. Die Vertretung weiterer interessierter Kreise ist im Schaubild links dargestellt.

Die Ergebnisse des vorliegenden Whitepapers werden auch in der KI-Normungsroadmap von DIN und DKE verarbeitet und fließen in die Arbeit der Arbeitsgruppe 'Ethik/ Responsible AI' ein. Die Veröffentlichung der KI-Normungsroadmap ist im Dezember 2020 geplant.

1 DIN.one ist eine von DIN zur Verfügung gestellte Arbeitsplattform zu der jede*r Interessierte nach vorheriger Registrierung Zugriff hat.

KURZFASSUNG

4

Dieses Whitepaper präsentiert die Ergebnisse des vom Bundesministerium für Wirtschaft und Energie geförderten Projekts „Ethikaspekte in der Normung und Standardisierung für Künstliche Intelligenz in autonomen Maschinen und Fahrzeugen“. Diese Ergebnisse basieren auf einer umfassenden Literaturrecherche, einer anschließenden Analyse durch das Projektteam sowie einer darauf basierenden Diskussion mit Expertinnen und Experten. Alle Texte standen auf der Plattform DIN.one zur Kommentierung durch die Öffentlichkeit bereit und alle eingegangenen Kommentare wurden mit den Kommentierenden sowie weiteren Experten und Expertinnen besprochen.

Aus diesem Vorgehen ergaben sich 8 Handlungsempfehlungen für künftige Normungs- und Standardisierungsaktivitäten. Sie sind in Kapitel „Handlungsempfehlungen“ dargestellt.

Die umfangreiche Literaturrecherche spiegelt sich in den Anfangskapiteln wider: Das Kapitel „KI, Ethik, Moral – Begriffsverständnisse und Ansichten“ klärt zunächst das Verständnis unterschiedlicher Begrifflichkeiten, auf denen dieses Whitepaper basiert. Anschließend gibt das Kapitel „Ethik und KI – Grundzüge der aktuellen Debatte“ einen Überblick über die Anforderungen an eine ethische KI, die in diversen Veröffentlichungen erschienen sind. Das Bild wird vervollständigt durch die Übersicht der bereits vorhandenen Normungs- und Standardisierungsaktivitäten im Kapitel „Aktuelle Normungs- und Standardisierungsaktivitäten“.

Die Analyse dieser Anforderungen in Bezug auf technische Normungs- und Standardisierungsmöglichkeiten sowie die Identifikation von Standardisierungspotentialen präsentiert das Kapitel „Welche neuen Normen und Standards werden zukünftig benötigt?“. Einige, der mit Expertinnen und Experten entwickelten Normungs- und Standardisierungspotentiale geben Antworten auf die Anforderungen an ethische KI, die durch die Literaturrecherche identifiziert wurden, und werden somit zu Handlungsempfehlungen an die Normungsgremien oder mögliche Interessierte an Standardisierungsaktivitäten.

INHALT

7

Abbildungsverzeichnis

Tabellenverzeichnis

8

Abkürzungsverzeichnis

Abkürzung	Definition	Abkürzung	Definition
AA	Arbeitsausschuss	**IST**	intelligent transportation system (dt. intelligenten Verkehrssysteme)
AI	Artificial Intelligence (dt. Künstliche Intelligenz)	**IT**	Informationstechnologie
AK	Arbeitskreis	**ITU**	International Telecommunication Union (dt. Internationale Fernmeldeunion)
AWI	Approved Work Item (dt. Genehmigtes Arbeitsprojekt)	**JTC**	Joint Technical Committee (dt. Gemeinschaftskommitte)
BMVI	Bundesministerium für Verkehr und digitale Infrastruktur	**KI**	Künstliche Intelligenz
BMWi	Bundesministerium für Wirtschaft und Energie	**NA**	Normenausschuss
BVDW	Bundesverband Digitale Wirtschaft	**NIA**	DIN-Normenausschuss Informationstechnik und Anwendungen
CASCO	ISO's Committee on Conformity Assessment (dt. ISO Ausschuss für Konformitätsbewertung)	**OECD**	Organisation for Economic Co-operation and Development (dt. Organisation für wirtschaftliche Zusammenarbeit und Entwicklung)
CEN	Comité Européen de Normalisation (dt. Europäisches Komitee für Normung)	**RTS**	Road Traffic Safety
CENELEC	Comité Européen de Normalisation Électrotechnique (dt. Europäisches Komitee für elektrotechnische Normung)	**SAE**	Society of Automotive Engineers
COM (2018) 237	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 zum Thema Artificial Intelligence	**SC**	Subcommittee
DIN	Deutsches Institut für Normung	**SEG**	Standardization Evaluation Groups (dt. Standardisierungs-Bewertungsgruppen)
DKE	Deutsche Kommission Elektrotechnik, Elektronik und Informationstechnik in DIN und VDE	**StVG**	Straßenverkehrsgesetz
DSGVO	Datenschutz-Grundverordnung	**TC**	Technical Committee
EU	Europäische Union	**TR**	Technical Report (dt. Technischer Report)
HLEG	High Level Expert Group (dt. Hochrangige Expertengruppe)	**TS**	Technical Specification (dt. Technische Spezifikation)
IEC	International Electrotechnical Commission (dt. Internationale Elektrotechnische Kommission)	**WG**	Working Group
IEEE	Institute of Electrical and Electronics Engineers	**XAI**	Explainable Artificial Intelligence
ISO	International Organization for Standardization (dt. Internationale Organisation für Normung)		

EINLEITUNG

Künstliche Intelligenz oder kurz „KI“ vereint Altes und Neues. Altes, weil die Grundzüge der wissenschaftlichen Diskussion zum Thema KI bis in die 1950er Jahre zurückreichen; die philosophischen sogar noch weiter, bis ins antike Griechenland (1). Neues, weil das Thema nach zwei Phasen der Ernüchterung – gemeinhin als „KI-Winter“ bezeichnet – nun dank großer technischer Fortschritte auf dem Gebiet der Datenverarbeitung wieder mächtig an Fahrt gewonnen hat. KI ist nicht mehr länger nur eine wissenschaftliche Disziplin, sondern auch Thema in Politik und Wirtschaft geworden (2).

Wissenschaftliche Beiträge, wie Roboter, die durch KI gesteuert werden, laufen lernen (3) oder Videoaufnahmen von (noch recht oberflächlichen) Konversationen zwischen Assistenzsystemen, wie z.B. Alexa oder Siri zeigen, dass zwischen der Vision von einer starken, denkenden, gar emotionalen KI und der Realität noch eine große Lücke klafft (und es erscheint fraglich, ob sie überhaupt je geschlossen werden kann). Sie zeigen jedoch auch, dass aktuelle KI-Systeme schon zu beachtlichen Leistungen fähig sind. Wie leistungsfähig KI-Systeme mittlerweile geworden sind lässt sich daran ermessen, dass sie Zusammenhänge aufzeigen können, wo Menschen den Überblick verlieren[2], Entscheidungen treffen können, wo Menschen erst realisieren[3] und künstliche Inhalte generieren können, deren vermeintliche Natürlichkeit sogar Menschen täuscht[4].

KI selbst ist dabei nicht inhärent ethisch oder unethisch – ihre Anwendung mag es jedoch sein. Bekanntgeworden sind bereits Fälle, in denen die Anwendung von KI zur Diskriminierung von Menschen auf Basis ihres Geschlechts (4) oder ihrer Hautfarbe (5) geführt hat. Hinzu kommen gesellschaftliche Befürchtungen, früher oder später einem allwissenden KI-System und dessen Beurteilung ausgeliefert zu sein. Es ist ersichtlich, dass ein begründetes Vertrauen in die Arbeit von KI ein wesentlicher Baustein für die zukünftige, breite Akzeptanz dieser Technologie sein wird.

Parallel zu den rein technischen Aspekten finden daher auf nationaler wie auch internationaler Ebene Diskussionen über die Gestaltung von „ethischer“ KI statt, das heißt über die Anforderungen, die ein KI-System erfüllen muss, um ethischen Prinzipien gerecht zu werden. Technische Normen und Standards können hier einen wertvollen Beitrag leisten, indem sie als Basis zur Vereinheitlichung dieser technischen Anforderungen, Prozesse und Terminologien sowie zur Sicherung der Qualität dienen. Normen und Standards werden von Gruppen bestehend aus Experte und Expertinnen, d.h. Menschen im Konsens erarbeitet. In gewisser Weise kann technischen Normen und Standards daher eine Moral zugrunde gelegt werden. Gleichwohl verfolgt die technische Normung, insbesondere im Kontext von „Ethischer KI“, ausdrücklich nicht das Ziel, bestimmte kulturell verankerte Ansichten, einen Wertekanon oder bestimmte Moralen zu unterstützen. Normen und Standards können so den Vorbehalten gegenüber (normengerechter) KI positiv entgegenwirken.

Das vorliegende Dokument zeigt zukünftige Standardisierungsbedarfe auf dem Feld der ethischen KI auf, insbesondere im Kontext von autonomen bzw. vollautomatischen Maschinen und Fahrzeugen. Anhand verschiedener Attribute, die zum ethisch wertvollen Verhalten einer KI beitragen können, wird herausgestellt, wie das jeweilige Attribut normativ gefasst werden könnte und welche Grenzen dabei bestehen, sowie die Frage nach der Überprüfbarkeit des Attributs und der passenden Ausgestaltung einer Mensch-Maschine-Schnittstelle.

2 Beispielsweise werden im Rahmen des Projekts OceanMind Satellitenbilder von einer KI ausgewertet, um illegalem Fischfang auf die Spur zu kommen.
3 KI-Systeme kommen zunehmend im Bereich des Hochfrequenzhandels zum Einsatz, wo innerhalb von Sekundenbruchteilen auf Marktentwicklungen reagiert werden muss.
4 Vgl. Deepfake – eine KI-gestützte Art der Medienmanipulation, welche beispielsweise das nachträgliche Verändern von Videoaufnahmen ermöglicht.

ZIELE

Der Erfolg eines breiten Einsatzes in den unterschiedlichsten Anwendungsbereichen von Künstlicher Intelligenz (KI) hängt maßgeblich von der Akzeptanz dieser Technologie ab. Wie im nachfolgenden Kapitel „KI, Ethik, Moral - Begriffsverständnisse und Ansichten“ vorgestellt wird, gibt es eine wachsende Diskussion um ethische Fragestellungen beim Einsatz von KI. Die Ergebnisse der ethischen Diskussion und die darauf basierenden Handlungen werden die Zukunft der KI definieren, insbesondere, wenn Menschen in direkten Kontakt mit damit verbundenen Technologien kommen.

Normen und Standards fördern die Zusammenarbeit auf dem Markt. Durch deren flächendeckende, konsensbasierte Erarbeitung und Anwendung bauen sie Handelshemmnisse ab, was den Warenverkehr erleichtert und den Export fördert. Zudem dienen sie der Sicherheit von Produkten und sorgen für ein Mindestmaß an Sicherheit seitens der Produktanwendung durch Verbraucher und Verbraucherinnen. Hierbei sollen Risiken und nicht-intendierte Folgen für Verbraucher und Verbraucherinnen und die Umwelt minimiert werden und ein einheitliches Verständnis über die Ziele des Einsatzes einer Technologie geschaffen werden. Dieser Prozess steht beim Einsatz der KI noch am Anfang, wird aber durch den vermehrten Einsatz von KI zukünftig noch stärker in den gesellschaftlich-politischen Fokus rücken.

Das vorliegende Whitepaper verfolgt das Ziel, den aktuellen Stand der Diskussion darzustellen und Normierungs- und Standardisierungsbedarfe festzustellen. Die zentrale Leitfrage ist dabei, welchen Beitrag die Normung und Standardisierung leisten kann, um die Anwendung von KI bei automatisierten Fahrzeugen und Maschinen unter ethischen Aspekten zu ermöglichen.

Das Whitepaper soll Anstoß von Normungs- und Standardisierungsarbeiten im interdisziplinären Bereich Ethik und KI sein. Dieses Whitepaper berücksichtigt durch Einbindung verschiedener Fachkreise und deren Experten und Expertinnen – Industrie, Forschung, Wissenschaft und Verbraucherschutz – verschiedene Perspektiven, Aktivitäten und Strategien und möchte einen Beitrag für die weitere Diskussion einer ethisch orientierten KI leisten.

Unter dem Begriff KI werden eine Reihe von heterogenen Systemen und Anwendungen verstanden (vgl. Kapitel „Was zeichnet eine Künstliche Intelligenz aus?“). Im Folgenden werden grundlegende Annahmen zu dem Verständnis von Künstlicher Intelligenz getroffen, die der Analyse von Normungs- und Standardisierungsbedarfen in Kapitel „Welche neuen Normen und Standards werden zukünftig benötigt?“ zugrunde liegen.

Dabei ist es nicht Ziel dieses Whitepapers, ethische Werte zu definieren, sondern den KI-Einsatz bereits vorhandenen Grundsätzen und systematischen Anforderungen zuzuordnen. Die Bewertung über ethisches Verhalten einer KI bedarf eines gesellschaftlichen und politischen Konsenses, der derzeit in Deutschland beispielsweise durch die Enquete-Kommission im Bundestag erarbeitet und auch europäisch und international diskutiert wird. Normen und Standards können einen solchen Prozess unterstützen und politisch definierte Vorgaben umsetzen.

Die in Kapitel „Welche neuen Normen und Standards werden zukünftig benötigt?“ analysierten Grundsätze einer ethischen KI gelten über alle Domänen und KI-Anwendungen hinweg, legen jedoch aufgrund des Projektschwerpunktes einen Fokus auf automatisierte Fahrzeuge und Maschinen. Dies schließt auch Industrie- und Produktionsanwendungen ein, die kein intensives menschliches Handeln erfordern oder keine Schnittstelle mit dem Menschen aufweisen. Dies schließt nicht aus, dass bereits existierende Normen und Standards oder technische Richtlinien, wie z.B. die der ‚Arbeitssicherheit‘ oder der ‚Funktionalen Sicherheit‘ bereits Kriterien implizieren können, die Anwendungen vertrauenswürdig oder ethisch machen.

METHODISCHES VORGEHEN

Um möglichst objektiv zu den in Kapitel "Welche neuen Normen und Standards werden zukünftig benötigt?" betrachteten Attributen zu gelangen, wurde bis Juli 2019 eine umfangreiche Recherche von Ethikrichtlinien durchgeführt. Eine Übersicht der Dokumente ist in Anhang 1 dargestellt. Zudem wurden die aktuellen Normungs- und Standardisierungsaktivitäten im Bereich KI betrachtet, die sich weitestgehend mit dem Bereich Ethik beschäftigen.

Diese Dokumente wurden vom Projektteam auf ethische Anforderungen an KI analysiert. Diese Analyse ist in den Kapiteln „Ethik und KI – Grundzüge der aktuellen Debatte“ und „Aktuelle Normungs- und Standardisierungsaktivitäten“ zu finden. Nachdem eine Vielzahl von Attributen identifiziert wurde, erfolgte eine Gruppierung derselben nach Ähnlichkeit. Zudem wurde die Liste reduziert, indem sie nur Attribute berücksichtigt, die bereits direkt mittels technischer Normen und Standards adressiert werden können. Einige Anforderungen an KI, wie beispielsweise die Achtung der Rechtsstaatlichkeit, sollten zunächst durch den Gesetzgeber konkretisiert werden.

In einem Workshop mit Expertinnen und Experten wurden die identifizierten Attribute diskutiert. Es wurde geklärt, ob die damit verbundenen Anforderungen technisch umsetzbar sind und ob sich daraus Vorschläge für zukünftige Normen und Standards ergeben.

Die Ergebnisse der Analyse und des Workshops wurden anschließend auf DIN.one zur Kommentierung bereitgestellt. In einem abschließenden Workshop wurden alle eingegangen Kommentare mit den Expertinnen und Experten diskutiert und Teile des Dokuments entsprechend angepasst. Abschließend wurden die Normungs- und Standardisierungspotentiale, die einer Anforderung aus den Ethik-Richtlinien begegnen als Handlungsempfehlungen für künftige Normung- und Standardisierungsaktivitäten formuliert.

KI, ETHIK, MORAL
— Begriffsverständnisse und Ansichten

13

Dieses Kapitel gibt Antworten auf grundlegende Fragen, zum Unterschied von Moral und Ethik über subsymbolische und symbolische KI bis hin zu schwachen und starken KI-Systemen.

Darüber hinaus wird ein Überblick über die politischen und wirtschaftlichen Initiativen, sowohl auf nationaler als auch auf internationaler Ebene hinsichtlich des Themas künstlicher Intelligenz, gegeben.

Welche Einschätzungen trifft beispielweise die deutsche Bundesregierung in ihrem Strategiepapier Künstliche Intelligenz?

ETHISCHE ODER MORALISCHE KI?

Ethik und Moral sind eng miteinander verknüpft. Genau genommen handelt es sich bei der Moral um einen Teilbereich der Ethik, während die Ethik selbst einen Teilbereich der Philosophie bildet. Eine Moral beschreibt in ihrem Kern, welche Reaktion von einem Individuum in einer bestimmten Situation erwartet wird. Da eine Moral von gesellschaftlichen, politischen oder religiösen Ansichten beeinflusst werden kann, können verschiedenen Ausprägungen derselben existieren bzw. es können mehrere Moralen koexistieren, die ihrerseits wiederum mit der Zeit Veränderungen erfahren mögen.

Ethik dagegen steht als Reflexionsstufe über der Moral bzw. den koexistierenden Moralen und analysiert, systematisiert und hinterfragt diese (6). Ethik kann somit Handlungsregeln, Empfehlungen, Gebote und Verbote formulieren, die auf einer Vielzahl moralischer Vorstellungen gründen und durch gesellschaftliche Prozesse definiert werden können. Vor diesem Hintergrund erscheint es daher zielführender, grundlegende ethische Grundsätze für die Entwicklung und den (möglicherweise weltweiten und damit interkulturellen) Einsatz von Künstlicher Intelligenz zu formulieren, anstatt zu untersuchen, wie eine KI moralisch handeln kann. Diese Herangehensweise schließt moralische Überlegungen nicht per se aus, geht aber in ihren Auswirkungen weit darüber hinaus.

WAS ZEICHNET EINE KÜNSTLICHE INTELLIGENZ AUS?

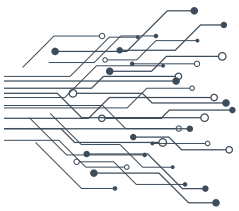
Es ist eine anspruchsvolle Aufgabe, eine passende Definition für KI zu finden. Ursächlich dafür ist nicht zuletzt das Fehlen einer eindeutigen und allgemein anerkannten Definition von Intelligenz. Mit Aufnahme der Arbeiten am Dokument ISO/IEC 22989 haben es sich die Internationale Organisation für Normung, ISO und die Internationale Elektrotechnische Kommission, IEC unter anderem zur Aufgabe gemacht, eine Definition von KI zu erarbeiten. Aktuell nähert man sich hier einer Definition für KI über die Fähigkeit eines solchen Systems, sich Wissen und Fähigkeiten anzueignen und diese auch anzuwenden. Bis zur endgültigen, publizierten Fassung der Norm ISO/IEC 22898 mögen sich hier jedoch noch Änderungen ergeben.

Ein anderer Ansatz, KI zu definieren, besteht darin, die Definitionen der einzelnen Begriffsbausteine heranzuziehen. Der Begriff „künstlich" beschreibt gemäß dem Wörterbuch Duden die Nachahmung eines natürlichen Vorgangs, während „Intelligenz" definiert wird als die „Fähigkeit [des Menschen], abstrakt und vernünftig zu denken und daraus zweckvolles Handeln abzuleiten". In diesem Sinne kann KI abstrakt als der Versuch verstanden werden, eine Nachahmung der geistigen Fähigkeiten des Menschen zu schaffen.

Diese Definition von KI findet sich in ähnlicher Form auch auf dem Gebiet der Computerwissenschaften wieder, mit denen KI aus öffentlicher Sicht gewöhnlich assoziiert wird. So betont etwa die Definition von KI als das „Unterfangen, Computern das Denken beizubringen" (7) die Nachahmung der geistigen Leistungsfähigkeit des Menschen durch Systeme, wie zum Beispiel Maschinen oder Computer. Diese Sichtweise spiegelt sich beispielsweise in dem, vom britischen Wissenschaftler Alan M. Turing entwickelten „Turing-Test" wieder (8), den eine KI nur bestehen kann, wenn sie mit einem Menschen in dessen natürlicher Sprache kommunizieren, sich Wissen aneignen und repräsentieren, logisch Schlussfolgern und sich an veränderte Umstände anpassen kann (1).

In einer anderen, ebenfalls gebräuchlichen Definition von KI wird diese als „die Studie des Entwurfs intelligenter Agenten" (9) definiert. Im Vergleich zur zuvor angesprochenen Variante, steht hier weniger die Imitation des Menschen im Vordergrund, als vielmehr das angestrebte Verhaltensideal des Systems bzw. des Agenten, rational zu handeln, das heißt gemäß seinen Kenntnissen das „Richtige" zu tun. Bemerkenswert an dieser Definition ist ihre Offenheit – das Handeln des Agenten kann, muss aber nicht durch logische Schlussfolgerungen bestimmt sein, um dennoch „richtig" zu sein und richtet sich nicht explizit nach einer menschlichen Intelligenz als Vorbild (1).

Dieses Dokument legt keine Definition von KI fest – diese Aufgabe wird der Normung vorbehalten –, sondern verwendet stattdessen die von der europäischen HLEG vorgeschlagene Definition:



„Systeme der Künstlichen Intelligenz (KI-Systeme) sind vom Menschen entwickelte Softwaresysteme (und gegebenenfalls auch Hardwaresysteme, die in Bezug auf ein komplexes Ziel auf physischer oder digitaler Ebene handeln, indem sie ihre Umgebung durch Datenerfassung wahrnehmen, die gesammelten strukturierten oder unstrukturierten Daten interpretieren, Schlussfolgerungen daraus ziehen oder die aus diesen Daten abgeleiteten Informationen verarbeiten, und über das bestmögliche Handeln zur Erreichung des vorgegebenen Ziels entscheiden. KI-Systeme können entweder symbolische Regeln verwenden oder ein numerisches Modell erlernen, und sind auch in der Lage, die Auswirkungen ihrer früheren Handlungen auf die Umgebung zu analysieren und ihr Verhalten entsprechend anzupassen.“ (10)

Dasselbe aber nicht das Gleiche – Arten künstlicher Intelligenz

KI-Systeme lassen sich anhand verschiedener Merkmale unterscheiden. Bezüglich des Vorgangs, wie ein KI-System auf Basis bekannter Fakten eine Schlussfolgerung generiert – auch als Interferenz bezeichnet – kann zwischen induktiven und deduktiven KI-Systemen unterschieden werden. **Induktive Systeme** analysieren einzelne Beispiele auf übergreifend anwendbare Muster. **Deduktive Systeme** dagegen gelangen anhand feststehender Regeln zu einem Ergebnis.

Eine weitere Unterscheidung wird hinsichtlich des Ansatzes praktiziert, wie sich ein KI-System der angestrebten Intelligenzleistung annähert. Die **symbolische KI** basiert auf der grundlegenden Annahme, dass Information durch explizite Symbole wie z.B. Zahlen repräsentiert und intelligentes Verhalten (weitgehend) durch mathematische Manipulation dieser Symbole (das heißt logische Verknüpfungen der Art „WENN A > B, DANN C = 1, SONST C = 0“) nachgeahmt werden kann (14). Symbolische KI nähert sich der Intelligenzleistung von der begrifflichen Ebene her bzw. verfolgt einen Top-Down-Ansatz. Mitunter wird sie auch als „klassische“ KI bezeichnet, um sie gegenüber der (vermeintlich moderneren) subsymbolischen KI abzugrenzen. Die **subsymbolische KI** basiert demgegenüber auf der Annahme, dass sich ein konnektionistisches Modell erstellen lässt, mit dessen Hilfe ähnliche Eingabemuster auf bestimmte Ausgabemuster abgebildet werden können. Subsymbolische KI nähert sich der Intelligenzleistung von einer impliziten Darstellung der Information – verfolgt also einen Bottom-Up-Ansatz.

KI-Systeme lassen sich prinzipiell auch hinsichtlich ihrer Trainingsmethode unterscheiden. Die derzeit vorherrschende Trainingsmethode ist das **maschinelle Lernen**, welche sich wiederum in weitere Arten unterteilen lässt. Beim **überwachten Lernen** (Supervised Learning) werden dem KI-System – beispielsweise einem künstlichen neuronalen Netzwerk – ausgewählte Trainingsdaten vorgelegt, die zuvor von Menschen hinsichtlich ihres Inhaltes ausgewertet bzw. gekennzeichnet (annotiert) wurden. Auf Basis dieser Trainingsdaten führt das KI-System eine Funktionsapproximation durch, das heißt, es entwickelt eine möglichst genaue Beschreibung einer unbekannten (und möglicherweise impliziten) Funktion, welche die Beziehung zwischen Eingangsdaten und gewünschtem Ergebnis beschreibt (Regression) bzw. eine Zuordnung von Eingangsdatensätzen zu zuvor bereits festgelegten Kategorien herstellt (Klassifizierung). Die Herausforderung besteht darin, das Fehlermaß bezogen auf die Trainingsdaten zu minimieren und dennoch ausreichend zu generalisieren, um auch für Eingangsdaten, die nicht Teil der Trainingsdaten sind, sinnvolle Ausgaben zu erzeugen.

Im Gegensatz dazu wird das KI-System beim **unüberwachten Lernen** (Unsupervised Learning) mit Trainingsdaten versorgt, die nicht zuvor von Hand ausgewertet (annotiert) wurden. Dem KI-System steht folglich keine Übersicht der möglichen Ergebnisse, nebst Beziehung zu den Trainingsdaten zur Verfügung. Stattdessen soll das KI-System im Verlauf des Trainings selbst ein Modell entwickeln, das die Trainings- bzw. Eingangsdaten strukturiert, das heißt auf Gruppen (Cluster) abbildet. Methoden des unüberwachten Lernens bieten gegenüber denen des überwachten Lernens einen Vorteil, wenn die Trainingsdaten sehr umfangreich sind (unter anderem weil der arbeitsintensive Schritt der händischen Auswertung und Kennzeichnung der Trainingsdaten entfällt) bzw. wenn bislang unbekannte Zusammenhänge in den (Trainings-)Daten aufgespürt werden sollen.

Eine besondere Form des überwachten Lernens ist **bestärktes Lernen** (Reinforcement Learning). Dem KI-System werden bei dieser Form des Trainings keine Trainingsdatensätze im eigentlichen Sinn zur Verfügung gestellt, sondern eine Lernumgebung (beispielsweise Spielregeln oder Straßenverkehrsszenarien), sowie ein Ziel, das es zu erreichen gilt. Es wird keine Strategie zur Zielerreichung vorgegeben – diese soll das KI-System im Verlauf des Trainings selbst entwickeln. Dafür erhält es mittels einer Bewertungsfunktion Rückmeldung, indem einzelne Schritte des Lernprozesses „belohnt" oder auch „bestraft" werden können, je nachdem ob sie in die (vermutlich) richtige bzw. falsche Richtung führen. Ein prominentes Beispiel für ein, mittels bestärktem Lernen trainierten KI-System ist das 2017 vorgestellte AlphaGo Zero, das nicht nur als das aktuell leistungsfähigste KI-System im Bereich des Go-Spiels gilt, sondern bei dessen Training auch völlig auf die Analyse menschlicher Spielzüge verzichtet wurde.

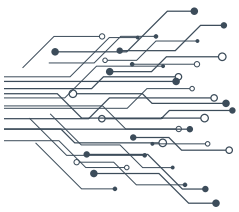
Im Vergleich zum bestärkten Lernen soll das KI-System beim **nachgeahmten Lernen** (Imitation Learning) ebenfalls eine Strategie zur Bewältigung einer Herausforderung entwickeln, dies jedoch in Anlehnung an ein Strategie-Vorbild. Je nach Zielsetzung soll das KI-System lernen, das Strategie-Vorbild zu imitieren oder es lediglich latent, das heißt in seinen Grundzügen übernehmen. Die Methode des nachgeahmten Lernens könnte für vollautomatisierte Fahrzeuge relevant werden, da sie das Erlernen einer gängigen Praxis zur konkreten Auslegung von Verkehrsregeln in bestehenden regionalen Ökosystemen ermöglicht und mit Blick auf die Vorlieben des Kunden Variationen des Fahrstils zulässt.

Schließlich lassen sich KI-Systeme auch nach ihrer Leistungsfähigkeit und Anwendungsdomäne differenzieren. Insbesondere in der KI-Forschung ist die abstrakte Unterscheidung in so genannte „starke" und „schwache" KI gebräuchlich. Die Grundlage dieser Unterscheidung ist philosophischer Natur und basiert auf zwei Hypothesen: der schwächeren Hypothese, dass sich ein System (z.B. eine Maschine) intelligent verhalten kann und der stärkeren Hypothese, dass ein solches System einen Verstand haben kann. Entsprechend würde ein **starkes KI-System** intelligentes Verhalten zeigen, weil es *wirklich* denkt, während sich ein **schwaches KI-System** lediglich so verhält, *als ob* es intelligent wäre (1). Die Leistungsfähigkeit eines starken KI-Systems wäre der des menschlichen Gehirns ebenbürtig oder würde sie sogar übertreffen. Dagegen ist ein schwaches KI-System auf die Bearbeitung einzelner Aufgaben spezialisiert und dient viel mehr dem Zweck, den Menschen in seiner (Denk-)Arbeit zu unterstützen, als ihn darin zu ersetzen (11) – ein Verständnis, das sich auch im aktuellen KI-Strategiepapier der Bundesregierung widerspiegelt (2).

Während Fragen zu einem möglichen Bewusstsein von starker KI, sowie deren genereller Realisierbarkeit nach wie vor kontrovers diskutiert werden (11), sind schwache KI-Systeme bereits Realität und seit einigen Jahren erfolgreich im Einsatz. Neben dem bereits erwähnten AlphaGo Zero sind Assistenzsysteme, wie sie z.B. auf vielen Smartphones vorzufinden sind, weitere bekannte Beispiele. Unlängst demonstrierte etwa das Unternehmen Google mit Google Duplex einen weiteren Entwicklungsschritt seines Assistenzsystems und ließ das KI-System telefonisch unter anderem einen Termin für einen Haarschnitt bei einer Mitarbeiterin eines Frisörsalons vereinbaren, ohne sich dabei als KI-System zu erkennen zu geben (12). Jahrzehnten zuvor war es diversen

Chatbots oder auch ELIZA[5] bereits gelungen, Menschen über die Tatsache hinweg zu täuschen, dass sie ihre (textbasierte) Konversation in Wahrheit mit einer Maschine führten (1), was die prinzipielle Leistungsfähigkeit schwacher KI-Systeme nur allzu unterstreicht.

Die HLEG kommt hier in ihrem zweiten Teil der KI Definition zu folgendem Schluss:



„Als wissenschaftliche Disziplin umfasst die KI mehrere Ansätze und Techniken wie z. B. maschinelles Lernen (Beispiele dafür sind „Deep Learning" und bestärkendes Lernen), maschinelles Denken (es umfasst Planung, Terminierung, Wissensrepräsentation und Schlussfolgerung, Suche und Optimierung) und die Robotik (sie umfasst Steuerung, Wahrnehmung, Sensoren und Aktoren sowie die Einbeziehung aller anderen Techniken in cyber-physische Systeme)." (10)

KI benötigt (die richtigen) Grenzen

KI im Licht der Ethik zu betrachten impliziert, den Fokus nicht ausschließlich darauf zu richten, was entwickelt werden könnte, sondern auch, wie sie entwickelt und eingesetzt werden sollte. Entgegen der öffentlichen Wahrnehmung besteht die Problematik bei KI-Systemen weniger darin, dass diese die ihnen auferlegten Grenzen übertreten würden, als vielmehr, dass sie diese Grenzen exakt einhalten werden (3). Die Beachtung von ethischen Grundsätzen ist daher vor allem bei der Entwicklung von subsymbolischer KI eine Herausforderung. Hierbei werden dem KI-System, je nach Wahl der Trainingsmethode und -daten, mehr oder minder weitläufige Freiräume zur Entwicklung eines Modells gewährt – Freiräume, deren Ausreizung letztlich in einem unbeabsichtigten Verhalten des KI-Systems resultieren können.

Eine Grundvoraussetzung für ethisches Verhalten von, mittels überwachtem Lernen trainierten KI-Systemen ist daher, dass bereits die Zusammenstellung der Trainingsdatensätze nach ethischen Gesichtspunkten und (möglichst) verzerrungsfrei erfolgt. Dies allein ist jedoch noch kein Garant dafür, dass entsprechende KI-Systeme später tatsächlich ethisches Verhalten zeigen (13). Gerade bei kritischen Anwendungen sollten diese KI-Systeme zuvor auf ethisches Verhalten geprüft werden. Letzteres trifft umso mehr auf KI-Systeme zu, die mittels unüberwachtem Lernen trainiert worden sind. Dass das KI-System hier während des Trainings tendenziell freier ist und die Trainingsdaten zuvor nicht annotiert wurden – folglich auch einfacher und in größerem Umfang erhältlich sind –, dürfte die Sicherstellung des erwünschten ethischen Verhaltens zusätzlich erschweren. Ähnlich dem überwachten Lernen spielt auch beim nachgeahmten Lernen die Auswahl der Trainingsdaten bzw. in diesem Fall die Auswahl eines geeigneten Strategie-Vorbilds eine entscheidende Rolle für das angestrebte ethische Verhalten des KI-Systems. Die wahrscheinlich direkteste Möglichkeit, ein quantifizierbares Wertesystem in das Lernen einfließen zu lassen, bietet die Methode des bestärkten Lernens. Über die Belohnungs- und Bestrafungsfunktion kann die Entwicklung des KI-Systems ziemlich direkt in die gewünschte Richtung getrieben werden.

Insgesamt betrachtet wird das Verhalten des KI-Systems sehr stark von der gewählten Lernmethode und den zugehörigen Parameter beeinflusst. Eine Validierung des Verhaltens von KI-Systemen mag daher eher zum gewünschten ethischen Verhalten beitragen als eine Regulierung der Trainingsmethoden.

5 Online-Umsetzung zum Ausprobieren – auch auf Deutsch verfügbar: LINK: http://www.med-ai.com/models/eliza.html.de

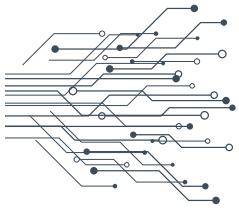
WEN STEUERT DIE KÜNSTLICHE INTELLIGENZ?

Im öffentlichen Diskurs und in der Literatur ist es verbreitet, im Kontext von KI-gesteuerten Maschinen und Fahrzeugen den Begriff „autonom“ zu verwenden, wo eigentlich dem Sinn nach „automatisiert“ gemeint ist. Allerdings ist die synonyme Verwendung dieser beiden Begriffe problematisch:

Der Begriff „automatisiert“ leitet sich von lateinisch *automatus* ab und kann mit „freiwillig“, „spontan“ oder auch „selbstbeweglich“ wiedergegeben werden (14) . Der Begriff „autonom“ dagegen ist ein aus dem Griechischen abgeleitetes, zusammengesetztes Wort. Der Wortbaustein *auto*, abgeleitet vom griechischen Reflexivpronomen *autos* für „sich selbst“, bedeutet so viel wie „selbst“, „eigen“ oder „von sich aus“. Der zweite Wortbaustein *nomos* trägt die Bedeutung von „Gesetz“ oder „Sitte“. „Autonom“ bedeutet als solches daher „eigenes Recht“, „Selbstgesetz“ oder auch „aus sich heraus Gesetz"(15). Gemessen an der Grundbedeutung der beiden Begriffe ergibt sich, dass ihre synonyme Verwendung nicht angebracht ist und eigentlich streng zwischen autonom und automatisch bzw. „selbstgesetzlich“ und „selbstbeweglich“ zu unterscheiden ist.

Um Missverständnissen vorzubeugen, wird daher der Begriff „automatisiert“ in diesem Dokument nur in Bezug auf Maschinen und Fahrzeuge verwendet, während der Begriff „autonom“ ausschließlich für Personen reserviert ist.

Hinzu kommt, dass bei automatisierten Fahrzeugen hinsichtlich des Automatisierungsgrades der Funktion unterschieden wird. Ist von „hochautomatisiert“ die Rede, übernimmt das System die Längs- und Querführung in einem spezifischen Anwendungsfall. Die fahrende Person muss hier das System nicht mehr dauerhaft überwachen. Sie muss jedoch potenziell in der Lage sein die Kontrolle zu übernehmen, da das System Systemgrenzen erkennt und anschließend mit ausreichender Zeitreserve zur Kontrollübernahme auffordert. Bei einem „vollautomatisierten“ Fahrzeug hingegen ist in einem spezifischen Anwendungsfall kein Fahrer mehr erforderlich. Im Anwendungsfall enthalten sind der Straßentyp, der Geschwindigkeitsbereich und die Umfeldbedingungen. Das System kann in diesem spezifischen Anwendungsfall alle Situationen automatisch bewältigen. Der höchste Automatisierungsgrad wird beim „fahrerlosen“ Fahren erreicht. Hier ist vom Start bis zum Ziel keine fahrende Person erforderlich. Die Fahraufgabe wird vollumfänglich bei allen Straßentypen, Geschwindigkeitsbereichen und Umfeldbedingungen vom System übernommen.(16)



” Im vorliegenden Dokument liegt der Fokus auf hochautomatisierten, vollautomatisierten oder fahrerlosen Fahrzeugen, die mittels einer schwachen KI gesteuert werden. Zudem werden Maschinen im industriellen Kontext berücksichtigt, die durch eine schwache KI gesteuert werden. “

ETHIK UND KI – GRUNDZÜGE DER AKTUELLEN DEBATTE

In der Diskussion um den Einsatz und Nutzen von KI haben verschiedene Interessensgruppen, Wirtschaftsverbände, Wissenschaft und Forschung sowie politische Entscheidungsträger auch zu einem ethischen Einsatz von KI Positionen erarbeitet. Die Diskussionen zu Ethik und KI finden auf nationaler, europäischer und internationaler Ebene statt und werden teilweise mit Hilfe von umfangreichen Studien geführt.

In diesem Kapitel werden die Grundzüge der öffentlichen Diskussion zu einem ethischen Einsatz von KI dargestellt. Diese wurden als Grundlage für die Diskussion mit Experten und Expertinnen herangezogen, die im Rahmen des Projektes Normungs- und Standardisierungspotentiale identifiziert. Hier wurden Publikationen mit Veröffentlichungsdatum bis Juli 2019 berücksichtigt. Eine Übersicht befindet sich in Anhang 1.

Politische Initiativen

Die deutsche Bundesregierung schätzt in ihrem Strategiepapier KI als Schlüsseltechnologie ein und sieht sich im internationalen Wettbewerb. In der Tat haben sich verschiedene Länder in den vergangenen Jahren zum Einsatz von Künstlicher Intelligenz positioniert und regulatorische oder strategische Maßnahmen eingeleitet. Dabei gab die US-Administration unter Barack Obama im Jahr 2016 mit dem *Positionspapier Preparing for the Future of Artificial Intelligence (17)* den Startschuss für den internationalen Wettstreit, um KI als Wirtschafts- und Wettbewerbsfaktor zu nutzen.

Weltweit lassen sich vielfältige Entwicklungsdynamiken verzeichnen. Der Einsatz von KI findet im Spannungsfeld zwischen Deregulierung und staatlicher Kontrolle statt. Beide Modelle zielen aber auf einen rapide wachsenden Einsatz von KI in nahezu allen Technikbereichen ab.

Der europäische Ansatz kann deutlich identifiziert werden. Hierbei setzen europäische Länder auf den Schutz des Individuums und wirken beispielsweise regulatorisch auf die Nutzung und Erhebung von personenbezogenen Daten ein (vgl. DSGVO). Auch ethische Fragestellungen beim Einsatz von KI werden in Europa mit politischer Priorität behandelt. (18)

KI-Strategie der Bundesregierung

Neben einer umfangreichen finanziellen Förderung von Forschungs- und Entwicklungsprojekten im Bereich der KI zielt die im November 2018 veröffentlichte KI-Strategie der Bundesregierung darauf ab, gesellschaftlichen Grundwerten und individuellen Grundrechten einen besonderen Stellenwert einzuräumen. Der Einsatz von KI soll dabei stets dem Menschen dienen und ethisch, rechtlich und kulturell verankert werden. Hierbei unterstützen Beratungsgremium, wie die Datenethik-Kommission und die Enquete-Kommission des Bundestages die Bundesregierung mit Empfehlungen.

Ziel sei eine *KI made in Germany*, die dem globalen Wettbewerb standhalten kann und eine verantwortungsvolle und gemeinwohlorientierte Entwicklung und Nutzung von KI voranbringt. Dabei heißt es in der KI-Strategie (2):

„Wir beachten dabei an unserer freiheitlich-demokratischen Grundordnung orientierte ethische und rechtliche Grundsätze im Hinblick auf den gesamten Prozess der Entwicklung und Anwendung Künstlicher Intelligenz. Wir wollen eine europäische Antwort auf datenbasierte Geschäftsmodelle und neue Wege der datenbasierten Wertschöpfung finden, die unserer Wirtschafts-, Werte- und Sozialstruktur entsprechen. Wir wollen die relevanten Akteure – vom Entwickler bis zum Nutzer von KI-Technologie – für die ethischen und rechtlichen Grenzen der Nutzung Künstlicher Intelligenz sensibilisieren; prüfen, ob der Ordnungsrahmen für ein hohes Maß an Rechtssicherheit weiterentwickelt werden muss, und die Beachtung ethischer und rechtlicher Grundsätze im gesamten Prozess der KI-Entwicklung und -Anwendung fördern und fordern.“

KI solle nach Ansicht der Bundesregierung den Bürgerinnen und Bürgern dienen, um Sicherheit, Effizienz und Nachhaltigkeit zu verbessern und soziale und kulturelle Teilhabe, Handlungsfreiheit und Selbstbestimmung zu gewährleisten.

Auch konkrete Anforderungen zu der Beschaffenheit von vertrauenswürdigen bzw. ethischen KI-Systemen definiert das Strategiepapier der Bundesregierung in Teilen. So wird die Erklärbarkeit und Transparenz von KI als ein Schlüssel für das Vertrauen in die KI genannt. Für Nutzer und Nutzerinnen sowie Betroffene sei ein KI-System oftmals weder nachvollziehbar noch transparent, wie es zu Entscheidungen oder Ergebnissen gekommen ist. Entscheidungen müssen sich unter anderem nachvollziehen lassen, damit KI-Systeme als *vertrauenswürdige KI* akzeptiert werden könnten und rechtlichen Anforderungen genügten. Zusätzlich müsse beim Einsatz von KI ein effektiver Schutz gegen Diskriminierung, Manipulation oder sonstigen Missbrauch sichergestellt sein. So würde die DSGVO bereits grundlegende Anforderungen stellen, sobald eine Automatisierung von Entscheidungen mit der Verarbeitung von personenbezogenen Daten zusammenfiele. In einem solchen Fall gäbe es umfassende Informationspflichten beispielsweise ein Recht des oder der Betroffenen für eine menschliche Überprüfung einer automatisierten Entscheidung sowie das Recht zur Einsichtnahme in Entscheidungsprozesse. Ebenso müsse offengelegt werden, welche personenbezogenen Daten bei der Entscheidung berücksichtigt wurden. Grundsätzlich ist erforderlich, dass bei der Entwicklung, Programmierung, Einführung und der Nutzung von KI-Systemen (unter Einbeziehung der Trainings- und Anwendungsdaten) **Transparenz, Nachvollziehbarkeit, Diskriminierungsfreiheit** und **Überprüfbarkeit** der KI-Systeme gewährleistet seien. Diese Attribute werden im Kapitel „Welche neuen Normen und Standards werden zukünftig benötigt?" auf Normungs- und Standardisierungspotentiale hin überprüft.

Eine tragende Rolle bei der Umsetzung der Ziele der KI-Strategie spricht die Bundesregierung Normen und Standards zu und sieht dabei DIN und DKE bei der Entwicklung nationaler, europäischer und internationaler Normen und Standards in zentraler Funktion. Wichtige Fragen seien, laut KI-Strategie, dabei die Festlegung von Begriffen und Klassifizierungen von KI (Dimensionen der Selbständigkeit, Selbständigkeit des Lernens, mit KI verbundene Risiken) sowie Ethikaspekten in der KI-Normung („ethics by design"). Auch die Überprüfung bestehender Normen und Standards auf *KI-Tauglichkeit* sei dabei zu berücksichtigen.[6] (2)

Unabhängige Hochrangige Expertengruppe für Künstliche Intelligenz (HLEG AI)

Im Juni 2018 veröffentliche die von der Europäischen Kommission eingesetzte HLEG ihre Ethik-Leitlinien für eine vertrauenswürdige KI. Ziel der Leitlinien ist die Förderung einer *vertrauenswürdigen KI*, die sich während des gesamten Lebenszyklus durch drei Komponenten auszeichnen soll: (19)

- **a.** Sie sollte rechtmäßig sein und somit alle anwendbaren Gesetze und Bestimmungen einhalten.
- **b.** Sie sollte ethisch sein und somit ethische Werte und Grundsätze einhalten.
- **c.** Sie sollte robust sein, und zwar in technischer als auch sozialer Hinsicht.

Die HLEG bietet somit eine Orientierung für die Förderung und Sicherung einer ethischen und robusten KI und leistet Hilfestellung für die Umsetzung in soziotechnischen Systemen. Die 52-köpfige Gruppe von Experten und Expertinnen meint dabei, dass der Einsatz von KI das Potential habe, die Gesellschaft nachhaltig zu verändern:

„Die KI ist kein Selbstzweck, sondern ein vielversprechendes Mittel, um das menschliche Gedeihen und somit das Wohlbefinden von Individuum und Gesellschaft und das Gemeinwohl zu steigern sowie zur Förderung von Fortschritt und Innovation beizutragen." (19)

6 Die KI-Strategie der Bundesregierung ist unter diesem Link abrufbar: https://www.bmbf.de/files/Nationale_KI-Strategie.pdf

Dabei liegen einer Anwendung von KI stets Grund- und Menschenrechte zugrunde, die durch EU-Verträge und die EU-Grundrechtecharta rechtsverbindlich festgelegt seien. Erwähnt werden dabei vor allem die Achtung der **Menschenwürde**, die **Freiheit des Einzelnen**, die Achtung von **Demokratie** und **Rechtsstaatlichkeit** sowie **Gleichheit, Nichtdiskriminierung** und **Solidarität**.

KI-Systeme sollten nach der HLEG das Wohl des einzelnen Menschen und das Gemeinwohl fördern. Als „ethische Imperative" werden vier Grundsätze formuliert, um die Entwicklung, Einführung und Nutzung von KI-Systemen auf vertrauenswürdige Art und Weise zu gewährleisten. Alle KI-Akteure sollten stets bestrebt sein, sie zu befolgen.

a. Achtung der menschlichen Autonomie

„Wenn Menschen mit KI-Systemen interagieren, müssen sie in der Lage sein, die Selbstbestimmung über die eigene Person in vollem Umfang und wirksam auszuüben […]. KI-Systeme sollten Menschen nicht auf ungerechtfertigte Weise unterordnen, nötigen, täuschen, manipulieren, konditionieren oder in eine Gruppe drängen. KI-Systeme sollten vielmehr dazu dienen, die kognitiven, sozialen und kulturellen Fähigkeiten des Menschen zu stärken, ergänzen und fördern." (19)

b. Schadensverhütung

„KI-Systeme und die Umgebungen, in denen sie operieren, müssen sicher und geschützt sein. Sie müssen technisch robust sein und es muss gewährleistet sein, dass sie nicht für Missbrauch anfällig sind." (19)

c. Fairness

„Die substanzielle Dimension impliziert eine Verpflichtung zur Gewährleistung einer gleichen und gerechten Verteilung von Vorteilen und Kosten und die Gewährleistung, dass Personen und Gruppen vor unfairer Verzerrung, Diskriminierung und Stigmatisierung geschützt werden. […] Zur verfahrenstechnischen Dimension der Fairness gehört die Möglichkeit, sich gegen Entscheidungen der KI-Systeme und der sie betreibenden Menschen zu wehren und einen wirksamen Rechtsbehelf einzulegen." (19)

d. Erklärbarkeit

„Erklärbarkeit ist unabdingbar, wenn beim Benutzer dauerhaftes Vertrauen in KI-Systeme entstehen soll. Das bedeutet, dass Prozesse transparent sein müssen, dass die Fähigkeiten und der Zweck von KI-Systemen offen zu kommunizieren sind und dass Entscheidungen – im größtmöglichen Umfang – den direkt und indirekt davon betroffenen Personen erklärbar sein müssen. […] Eine Erklärung, warum ein Modell ein bestimmtes Ergebnis oder eine bestimmte Entscheidung erzeugt hat (und welche Kombination aus Eingabefaktoren dazu geführt hat) ist nicht immer möglich. Diese Fälle werden als „Blackbox"-Algorithmen bezeichnet und erfordern besondere Beachtung. Unter diesen Umständen sind möglicherweise andere Erklärbarkeitsmaßnahmen notwendig (z. B. Rückverfolgbarkeit, Nachprüfbarkeit und transparente Kommunikation über die Fähigkeiten des Systems), solange das System als Ganzes Grundrechte achtet." (19)

Die HLEG formuliert im weiteren Verlauf Anforderungen an eine vertrauenswürdige KI. Diese Anforderungen werden zum Großteil in Kapitel „Welche neuen Normen und Standards werden zukünftig benötigt?" aufgegriffen, um das Normungs- und Standardisierungspotential zu analysieren. Folgende Grafik gibt einen Überblick über die Beziehung der sieben Anforderungen, die sich gegenseitig unterstützen und während des gesamten Lebenszyklus eines KI-Systems umgesetzt und bewertet werden sollen:

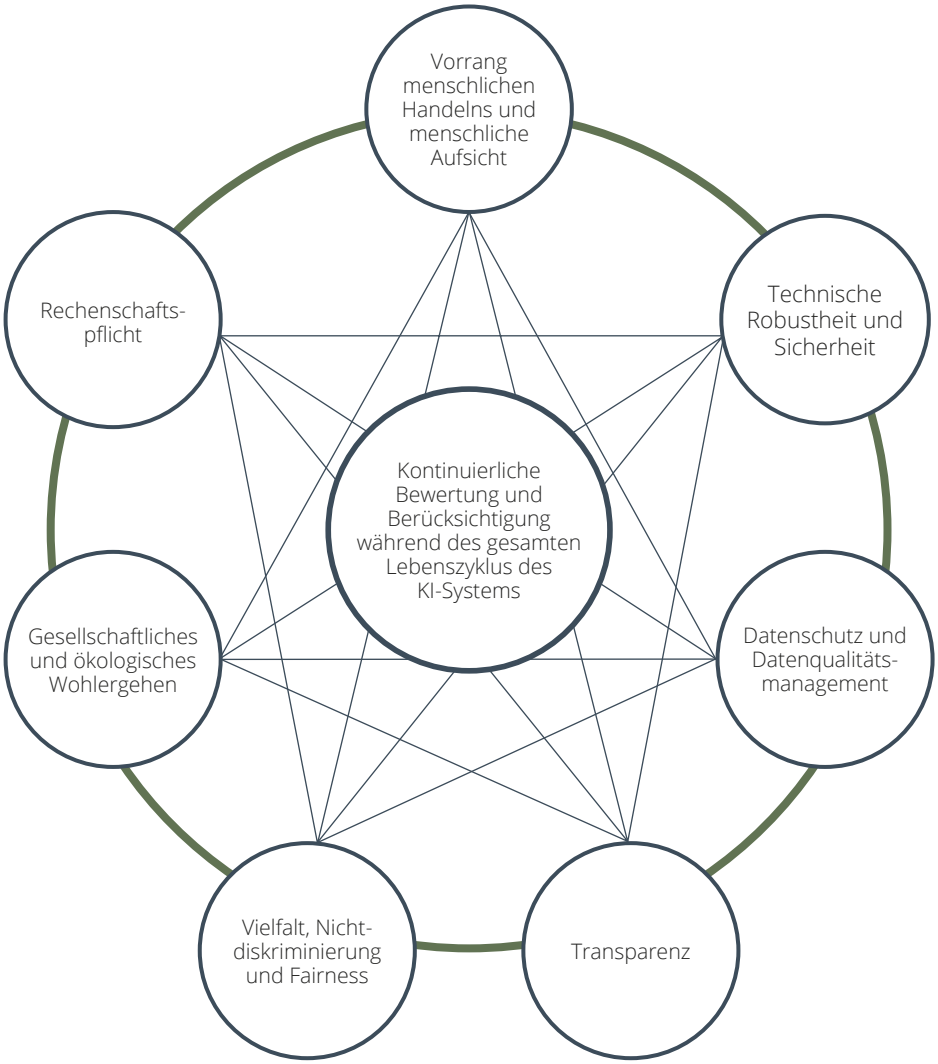


Abbildung 2: Werte, die von der HLEG identifiziert wurden

Quelle: ETHIK-LEITLINIEN FÜR EINE VERTRAUENSWÜRDIGE KI, HLEG, © European Union, 1995-2019 (CC BY 4.0)

Die HLEG betrachtet die dargestellten Ergebnisse als einen dynamischen Prozess, den es regelmäßig zu prüfen und aktualisieren gilt. Sie sollen als Ausgangspunkt für Diskussionen über eine vertrauenswürdige KI für Europa gelten und ersetzen keine politischen Entscheidungen oder Regulierungen. Über Europa hinaus – so der Wunsch der Experten und Expertinnen – sollen diese Leitlinien auch die Forschung, Reflexion und Diskussion über einen ethischen Rahmen für KI-Systeme auf weltweiter Ebene fördern.[7]

Internationale Organisation für Zusammenarbeit und Entwicklung (OECD)

Auf die EU-Leitlinien gestützt, hat auch die Internationale Organisation für Zusammenarbeit und Entwicklung (OECD) – eine internationale Organisation mit 36 Mitgliedstaaten, überwiegend Mitgliedstaaten aus Europa und Nordamerika – Grundsätze für KI formuliert. So will die OECD eine

7 Weitere Informationen zu den Leitlinien der HLEG AI sind unter diesem Link abrufbar: https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

innovative und vertrauenswürdige KI fördern, die Menschenrechte und demokratische Werte achte. Insgesamt fünf Empfehlungen hat die Sachverständigengruppe für KI formuliert (20):

a. KI sollte für die Menschen und den Planeten Nutzen bringen, indem sie ein inklusives Wachstum, eine nachhaltige Entwicklung und die Lebensqualität fördert.

b. KI-Systeme sollten so konzipiert werden, dass sie das Prinzip der **Rechtsstaatlichkeit**, die **Menschenrechte, demokratische Werte** und die **Vielfalt** achten, und sie sollten angemessene Schutzmaßnahmen – z.B. bei Bedarf die Möglichkeit menschlichen Eingreifens – vorsehen. Das Ziel ist dabei eine faire und gerechte Gesellschaft.

c. KI-Systeme sollten **Transparenz** und **verantwortungsvolle Offenlegung** gewährleisten um sicherzustellen, dass die Menschen KI-basierte Ergebnisse verstehen und hinterfragen können.

d. KI-Systeme müssen über ihren gesamten Lebenszyklus **robust** und **sicher** funktionieren, und potenzielle Risiken sollten kontinuierlich beurteilt und kontrolliert werden.

e. Die Organisationen und Personen, die KI-Systeme entwickeln, einführen oder betreiben, sollten für den einwandfreien Betrieb gemäß den oben aufgeführten Grundsätzen rechenschaftspflichtig sein.

Den Regierungen der Mitgliedsländer empfiehlt das OECD-Dokument, öffentliche und private Investitionen in Forschung und Entwicklung zu erleichtern, um Innovationen in vertrauenswürdige KI zu fördern sowie die Schaffung eines Politikumfelds, das den Weg für die Einführung vertrauenswürdiger KI-Systeme bereitet. Grundsätzlich wäre eine grenz- und branchenübergreifende Zusammenarbeit notwendig, um bei der verantwortungsvollen Steuerung vertrauenswürdiger KI Fortschritte zu erzielen.[8]

Wirtschaftsinitiativen

Auf nationaler, europäischer und internationaler Ebene gibt es zahlreiche Initiativen von Verbänden, Industrie und Wirtschaft die sich mit ethischen Fragen beim Einsatz von KI auseinandersetzen. Dem Einsatz von KI wird ein großes wirtschaftliches Potential zugeordnet, was im Rahmen eines globalen Wettbewerbs wahrgenommen werden muss. Gleichzeitig wird erkannt, dass es eine politische und gesellschaftliche Diskussion gibt, die sich mit den Grenzen des Einsatzes von KI auseinandersetzt. Dies könne den Einsatz und die Akzeptanz von KI beeinflussen, so dass die Entwicklung ethischer Prinzipien mit der weiteren Entwicklung der KI einhergehen müsse.

Grundsätzlich wird die **Autonomie und Entscheidungsfreiheit der Nutzer** als ein hohes Gut bewertet. Hierzu äußerte sich beispielsweise der Bundesverband Digitale Wirtschaft (BVDW) und forderte, den Mehrwert eines KI-Einsatzes immer so offen und klar an den Nutzer oder die Nutzerin zu kommunizieren, dass beim Nutzer oder der Nutzerin Vertrauen entsteht. Dabei solle es jedoch nicht um die Offenlegung von Programmierungen und die Funktionsweise von Algorithmen gehen – diese sollten weiterhin im Rahmen des Geschäftsgeheimnisses beim Unternehmen bleiben. Vielmehr sollte darüber informiert werden, welche Daten zugrunde liegen und welche Ziele mit dem Algorithmus verfolgt werden. Dies sollte in einfacher Weise erfolgen, die den Nutzer oder die Nutzerin bzw. den Konsumenten oder die Konsumentin nicht überfordere. (21)

Darüber hinaus wird oftmals betont, dass die Bewertung ethischer Fragen im Umgang mit KI stets unter Betrachtung des Anwendungsfalls erfolgen solle, da die spezifischen ethischen Herausforderungen von KI beträchtlich variieren könnten. So stellt beispielsweise der KI-Expertenrat[9] – ein Gremium aus führenden Industrievertretern – fest, dass eine KI, die für eine vorausschauende Wartung in der Industrie eingesetzt werde, andere ethische Fragen aufwerfe als diejenige, die beim automatisierten Fahren Anwendung finde. Somit sei eine Einzelfallbetrachtung in vielen

8 Die OECD-Grundsätze für KI sind unter diesem Link abrufbar: https://www.oecd.org/going-digital/ai/principles/
9 Informationen zum KI Expertenrat finden sich unter folgendem Link: https://www.microsoft.com/de-de/berlin/ki-expertenrat.aspx

Fällen angebracht und zielführend. Diese sollten die entwickelnden und anwendenden Unternehmen eigenständig und selbstverantwortlich bewerten. Zudem sollte der **Einsatz von KI stets kenntlich gemacht werden**, insbesondere im Rahmen der Mensch-Maschine-Interaktion, so dass Missverständnisse und /oder eine Täuschung nicht entstünden.

Im Bereich der **Nachvollziehbarkeit** von KI-Entscheidungen müssten Entscheidungenkriterien transparent gemacht werden. Nur so könne ein Vertrauen aufgebaut werden, da KI unmittelbare Auswirkungen auf das soziale Leben und die wirtschaftlichen Chancen von Individuen haben könne. Gleichzeitig wird die besondere Natur von KI-Systemen anerkannt, die eine vollumfängliche Nachvollziehbarkeit von Entscheidungen technisch schwierig mache, insbesondere bei adaptiven Systemen. Ziel müsse es dabei sein, eine solche Nachvollziehbarkeit zumindest näherungsweise zu erreichen, um Black-Box-Systeme zu verhindern.

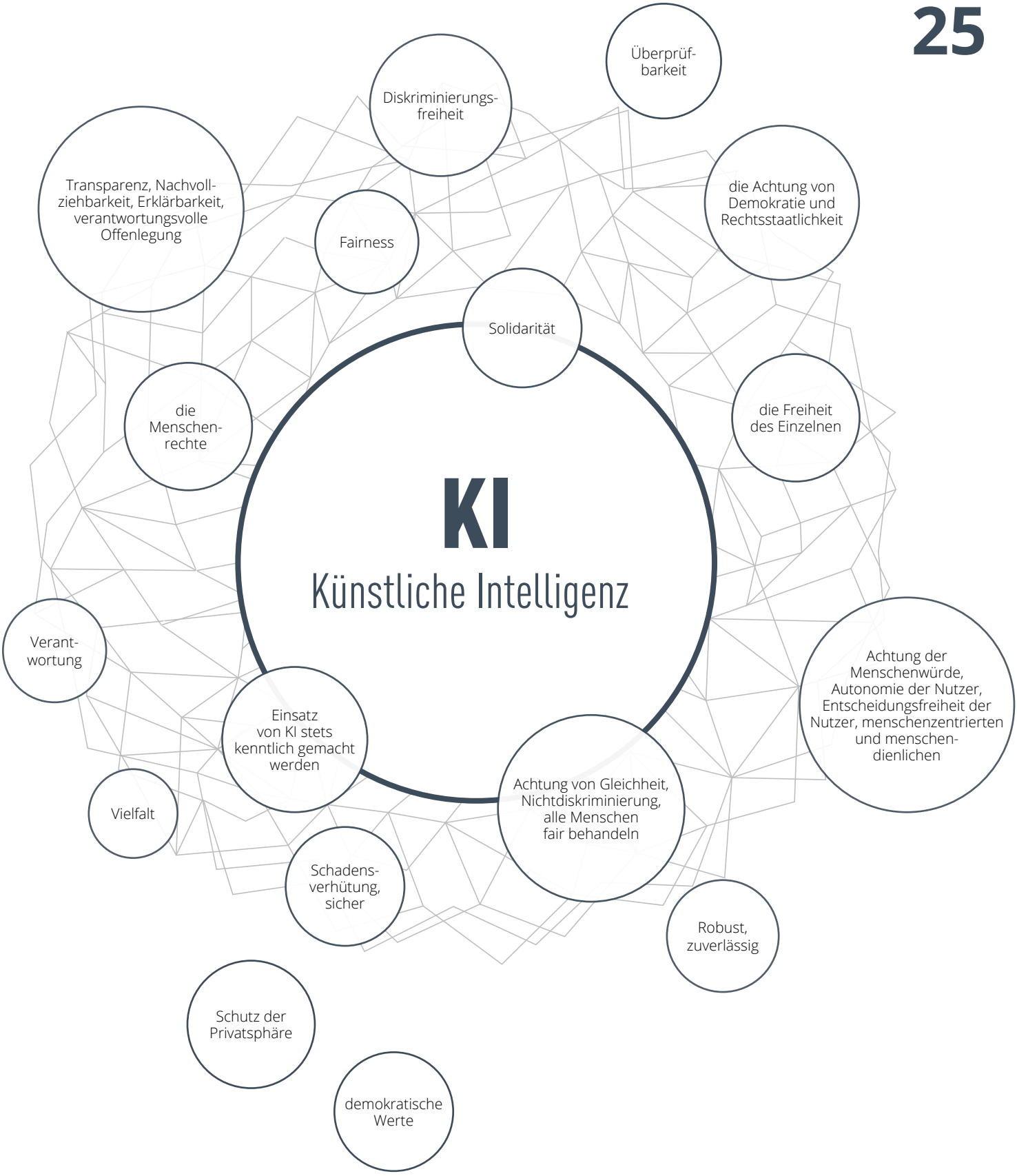
Letztendlich sei jedoch auch die Datenbasis zur Vorbeugung einer systematischen **Diskriminierung** ausschlaggebend. Eine breite und diverse Datenbasis sei elementar für eine korrekte Funktion von KI sowie für das Nutzervertrauen.

Einen ersten Vorstoß seitens einer Zertifizierung eines vertrauenswürdigen KI-Systems hat der KI-Bundesverband mit seinem KI-Gütesiegel getätigt. Nach Angaben des Verbandes verfolge das KI-Gütesiegel das Ziel einen **menschenzentrierten** und **menschen-dienlichen** Einsatz von KI zu sichern. Durch das Einhalten einer übergeordneten Werte- und Prozessverständnisses stelle das Gütesiegel eine ethisch verträgliche Service- und Produktentwicklung sicher. Im Zentrum stünden die Gütekriterien Ethik, Unvoreingenommenheit, Transparenz sowie Sicherheit und Datenschutz. Bislang umfasst das Gütesiegel eine Selbstverpflichtungserklärung. (22)

Auch führende globale Technologiekonzerne äußern sich in einigen Fällen zum Einsatz von KI. Globale Konzerne wie Microsoft, Google, IBM oder SAP haben ethische Prinzipien definiert, die Vertrauen in KI schaffen sollen und an denen sie sich orientieren. Grundsätzlich soll der Einsatz von KI **dem Menschen dienen**, sobald sie in verschiedenen Bereichen wie Gesundheit, Sicherheit, Energie, Herstellung oder Unterhaltung eingesetzt wird. Dementsprechend soll KI alle Menschen **fair** behandeln, was auch eine besondere Datenqualität voraussetzt. KI-Entwickler und Entwicklerinnen sollen auf *unfair bias* wie Nationalität, Geschlecht, Einkommen, sexuelle Orientierung oder politische Einstellung sensibilisiert werden, und die Teams aus Entwickler und Entwicklerinnen sollten selbst die Vielfalt der Welt wiederspiegeln. Zudem sollte KI **zuverlässig** und **sicher** sein, auch in unerwarteten Situationen. Die Sicherheit der KI ist entscheidend für ihre Akzeptanz und muss dauerhaft durch strenge Test und Verfahren überwacht werden. Neben dem **Schutz der Privatsphäre**, also der Frage, wie welche Nutzerdaten erhoben und genutzt werden können, widmen sich die Technologiekonzerne auch der **Transparenz** der KI-Systeme. Hierbei müsse vermieden werden, dass KI anonym Entscheidungen über den Menschen treffe, die nicht überprüfbar seien. Dies bedeute, dass Ergebnisse von KI-Analysen aufbereitet werden müssten, und dass Menschen sie nachvollziehen und bewerten können sollten. Eine ethische KI müsse den Menschen in die Lage versetzen, Fehlentscheidungen und Vorurteile zu erkennen und diese zu melden und/oder zu korrigieren. Auch zur Frage der **Verantwortung** äußert sich die Industrie und stellt fest, dass wie bei anderen Technologien die Menschen für den Einsatz von KI verantwortlich seien. Es müsse verhindert werden, dass es überhaupt zu Rechtsverletzungen komme. (25), (27), (28), (29)

Aus den genannten Debatten, die vor allem durch Industrie, Politik und Nichtregierungsorganisationen geführt werden, lassen sich einer ethischen bzw. vertrauenswürdigen KI Anforderungen und Eigenschaften zuschreiben. Hierbei ist die Forderung nach einer erklärbaren KI deutlich hervorzuheben, sowie Aspekte der IT-Sicherheit, des Datenschutzes sowie die Mensch-Maschine Schnittstelle. Diese und weitere Aspekte erschaffen nach der dargestellten Diskussion einen ethischen Anspruch beim KI-Einsatz, der im Rahmen des vorliegenden Projektes eine Grundlage für die Analyse der Normung- und Standardisierungsbedarfe bietet.

Abbildung 3: Attribute für ethisches Verhalten einer KI basierend auf der öffentlichen Debatte bis Juli 2020



AKTUELLE NORMUNGS- UND STANDARDISIERUNGSAKTIVITÄTEN

Wer arbeitet bereits am Thema künstliche Intelligenz und welche Gruppen betrachten dabei auch ethische Aspekte und automatisiertes Fahren?

Welche Ansätze verfolgt zum Beispiel die IEEE Global Initiative in Ihrer ersten Ausgabe „Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems“?

Und wie sehen die aktuellen Aktivitäten diesbezüglich bei DIN und DKE, CEN/CENELEC und ISO/IEC aus?

Die Antworten auf diese Fragen inklusive einer Auflistung der momentan laufenden Projekte des Gremiums ISO/IEC JTC 1/SC 42 *Artificial Intelligence* sind in diesem Kapitel zu finden.

Darüber hinaus wird auch die Frage adressiert, was Normung und Standardisierung in diesem Kontext leisten kann.

WAS KANN NORMUNG UND STANDARDISIERUNG LEISTEN?

Technische Normen und Standards entstehen nach festgesetzten Regeln. Hierbei ist der Konsens grundsätzlich ein wichtiges Prinzip im Normungs- und Standardisierungsprozess. Das bedeutet, die Expertinnen und Experten verständigen sich unter Berücksichtigung des Standes der Technik auf eine gemeinsame Version der Inhalte, die versucht, alle Interessen der Beteiligten zu berücksichtigen und Gegenargumente auszuräumen. Dieses Prinzip, sowie andere wichtige Grundsätze der Normung (u.a. Freiwilligkeit, Öffentlichkeit, Kohärenz), können im KI-Bereich gut Anwendung finden und stützen die nachhaltige Entwicklung der Technologie. Normen und Standards dienen der Vereinheitlichung technischer Anforderungen, der Vereinheitlichung von Prozessen und Terminologien sowie der Sicherung von Qualität. Technische Normen und Standards sollten aber nicht mit kulturell verankerten gesellschaftlichen Standards verwechselt werden und unterstützen daher nach den Grundsätzen der Normung nicht die Ansichten einer bestimmten Gruppe. Dies soll insbesondere auch bei der Diskussion um ethische Anforderungen an den Umgang mit Künstlicher Intelligenz berücksichtigt werden.

NORMEN, STANDARDS, KONSORTIALSTANDARDS ODER AKTIVITÄTEN IM BEREICH DER TECHNISCHEN REGELSETZUNG

Normung und Standardisierung begleitet innovative KI-Entwicklungen und bereitet die Grundlage für verlässliche KI-Anwendungen. Aktivitäten gibt es mittlerweile auf internationaler Ebene bei der *Internationalen Organisation für Normung* (ISO) und der *Internationalen Elektrotechnischen Kommission* (IEC), auf europäischer Ebene beim *Europäischen Kommitee für Normung* (CEN) und dem *Europäischen Kommitee für Eletrotechnische Normung* (CENELEC) sowie auf nationaler Ebene bei DIN und DKE. Darüber hinaus gibt es einige Aktivitäten innerhalb des *Institute of Electrical and Electronics Engineers* (IEEE) und der *Internationalen Fernmeldeunion* (ITU).

Auch im Bereich der **ethischen Anwendung von KI** wurden bereits einige Normungs- und Standardisierungsaktivitäten angestoßen. Diese Vorhaben sind zum Großteil in der Planungs- und Orientierungsphase und sollen zunächst das Normungs- und Standardisierungspotential im Bereich Ethik und KI eruieren.

Gesellschaftliche Bedenken und Ethikaspekte von KI sind aktuell ein zentrales Thema in der KI-Normung und Standardisierung, denn Bedenken können aus technischer Sicht durch Normen verringert werden. Beispielsweise können Normen Leitlinien zur Risikominderung liefern, die bei der Nutzung von KI potenzielle Auswirkungen auf die Gesellschaft haben. Normen und Standards können auch Best Practices für die Entwicklung und das Training von KI-Systemen bereitstellen, um Verzerrungen in Algorithmen oder Trainingsdatensätzen zu minimieren. Im Gemeinschaftskommittee von ISO und IEC, dem **ISO/IEC Joint Technical Committee 1** (JTC 1), werden Normen und Standards zu IT-Themen erarbeitet. Eine Untergruppe dieses Gemeinschaftsausschusses befasst sich seit April 2018 mit dem Thema *Artificial Intelligence*. Die Arbeitsgruppe *Trustworthiness* (WG3) dieses Gremium (ISO/IEC JTC 1/SC 42 *Artificial Intelligence*) befasst sich mit der Frage: Wie kann Vertrauenswürdigkeit von KI Systemen bzw. Organisationen, die KI zur Verfügung stellen oder nutzen, hergestellt werden?

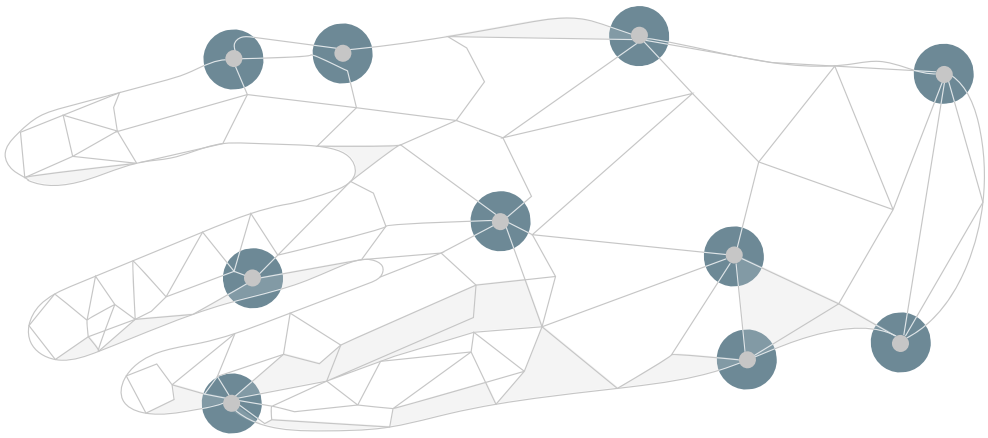
Vertrauenswürdigkeit ist ein notwendiger Aspekt, um unter anderem eine breite Markteinführung von KI erfolgreich zu erreichen. Das Subcommittee 42 Working Group 3 (SC42 WG3) hat sich darauf konzentriert, ein breites Spektrum von Themen im Zusammenhang mit **Sicherheit, Schutz, Privatsphäre, Robustheit, Elastizität, Zuverlässigkeit, Transparenz, Kontrollierbarkeit** usw. im Kontext von KI-Anwendungen und -Systemen zu untersuchen und zu normen. Dies ist die Liste der aktuellen Projekte:

- ISO/IEC 23894 Information Technology — Artificial Intelligence — Risk Management
- ISO/IEC TR 24027 Information Technology — Artificial Intelligence (AI) — Bias in AI systems and AI aided decision making

- ISO/IEC TR 24028 Information Technology — Artificial Intelligence (AI) — Overview of trustworthiness in Artificial Intelligence
- ISO/IEC TR 24029-1 Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview
- ISO/IEC 24029-2 Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Formal methods methodology
- ISO/IEC TR 24368 Information technology — Artificial intelligence — Overview of ethical and societal concerns
- ISO/IEC TR Artificial intelligence — Functional Safety and AI systems

KI kann von verschiedenen Branchen und Anwendungsbereichen verwendet werden, was dazu führen kann, dass kontextspezifische KI-Normen entwickelt werden müssen. Die Entwicklung solcher Normen erfordert Erfahrung und Wissen mit bzw. über KI und deren Anwendungsbereich. Governance Implikationen von KI ist derzeit das erste gemeinsame Arbeitsprojekt. Die Arbeitsgruppe wurde zwischen ISO/IEC JTC 1/SC 40 *Governance of IT* und ISO/IEC JTC 1/SC 42 *Artificial Intelligence* gegründet. Governance als unternehmensweite Disziplin und Verantwortung wird im Allgemeinen als stabiler angesehen als die Systeme, die ihrer Ausrichtung, Verantwortlichkeit und Aufsicht unterliegen. Technologien und die damit verbundenen Prozesse verändern und entwickeln sich ständig weiter.

KI ist nicht inhärent gut oder schlecht; sie ist weder ethisch noch unethisch. Als Artefakt ist sie jedoch ethisch nicht neutral, sowohl auf der Ebene des Designs als auch der Nutzung oder der Innovation. Ihre axiologische Bedeutung (das heißt ethisch oder unethisch) bzw. ihre axiologischen Eigenschaften (das heißt gut oder schlecht) hängen von den Werten der beteiligten Personen ab. Die Gesamtheit der Normen ist heterogen, weshalb nicht alle Normen von der gleichen Art sind (z.B. sind einige Normen ethisch, andere dagegen nicht). Zudem wird die Teilmenge der Normen, die als Grundlage für die Entwicklung der KI ausgewählt wird (sei es in der Entwurfsphase oder in späteren Phasen), stets die eigenen axiologischen Verpflichtungen widerspiegeln, nicht zuletzt auch die eigenen ethischen Werte, auch wenn nicht alle ausgewählten Normen ethischer Natur sind. Als solche sollten ihr Entwurf und ihre Nutzung auf der Ethik basieren, die eine Organisation für sich und ihre Arbeit kritisch definiert. Zusätzlich wird ISO/IEC AWI 38507 Richtlinien für Mitglieder der Leitungsgremien von Organisationen jeder Art – öffentlich, privat, gewinnorientiert, gemeinnützig etc. – und jeder Größe über die effektive, effiziente und akzeptable Nutzung KI in ihren Unternehmen bereitstellen.



Nachfolgend ist eine Übersicht über alle Projekte des ISO/IEC JTC 1/SC 42 *Artificial Intelligence* gegeben.

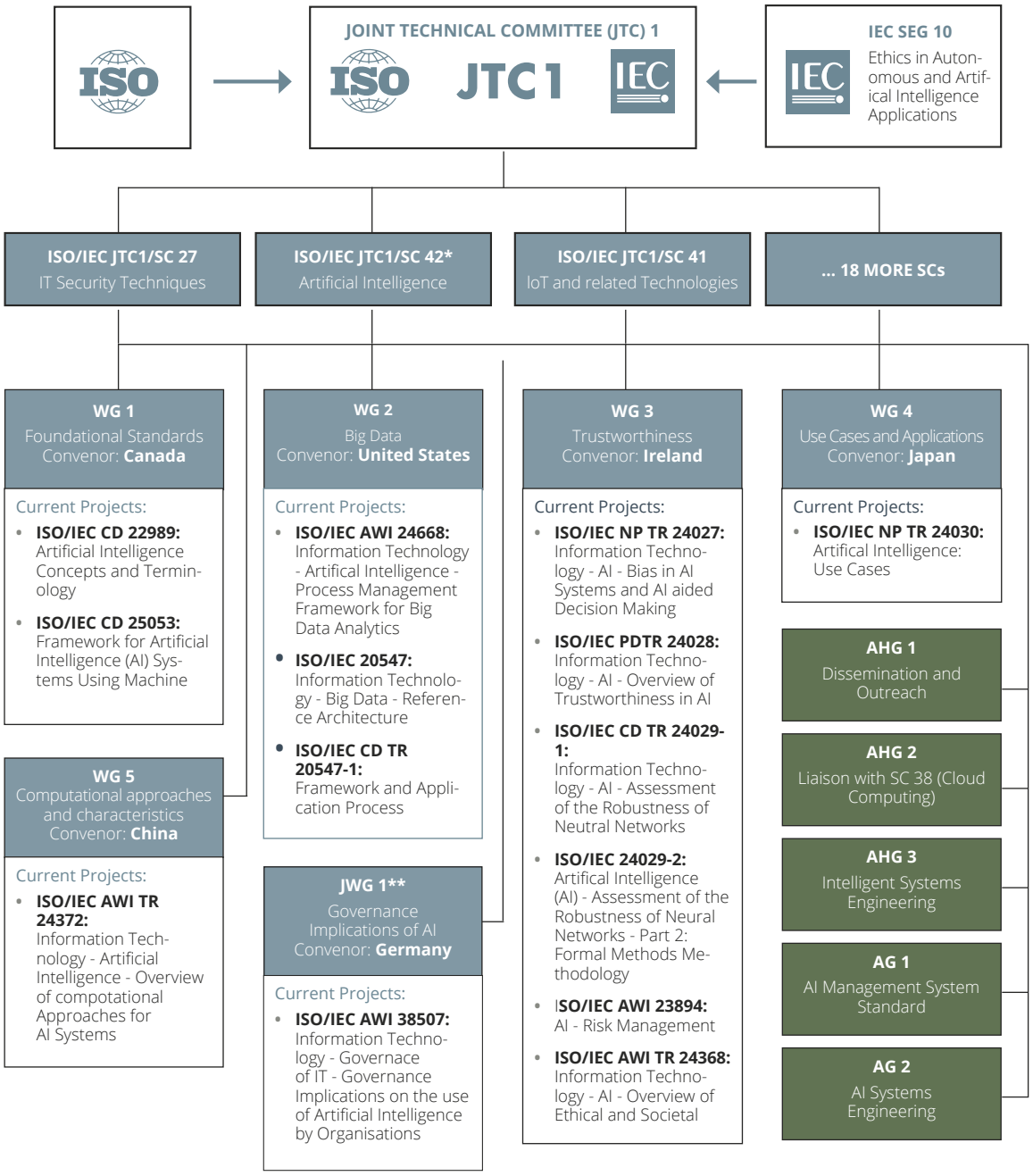


Abbildung 4: Gemeinschaftkommitee von ISO und IEC

* 30 partticipating & 14 observing Members
** (with ISO/IECJTC1/SC 40)

Explantatory notes:
- JTC: Joint Technical Committee
- SC: Subcommittee
- WG: Working Group
- JWG: Joint Working Group
- ADH: Ad Hoc Group
- AG: Advisory Group

Im März 2019 wurde eine *Advisory Group* im Gemeinschaftgreium ISO/IEC JTC 1 gegründet, die sich mit dem Thema *automatisierte und datenreiche Fahrzeuge* beschäftigt: JTC 1/AG 6: *Autonomous and Data Rich Vehicles*. Die Gruppe soll eine Bewertung des aktuellen Stands der Normungsaktivitäten zum Thema Fahrzeugdaten erarbeiten, die relevant sind für automatisierte und datenreiche Fahrzeuge innerhalb von ISO und IEC, in anderen Standardisierungsorganisationen und Industriekonsortien sowie in Rechts- und Regulierungsbehörden. Die Gruppe wird keine Normen oder Standards erarbeiten, sondern Empfehlungen und Berichte an ISO/IEC JTC 1 geben. In einem dieser Berichte fassten sie erste Ergebnisse zur Beurteilung der Nutzung von automatisierten Fahrzeugen zusammen: Vollständig automatisierte (Level 5) Fahrzeuge sind bereits in Sektoren wie Bergbau und Landwirtschaft im Einsatz. Auf offener Straße beginnen in einigen Märkten quasi autonome Fahrzeuge der Stufe 3 aufzutauchen. In der kommerziellen Luftfahrt gab es viele Jahre lang Level-3-Autonomie mit menschlichen Piloten, die als Notfall-Backup fungierten, aber ansonsten mit Flugzeugen, die meist mit Autopilot fliegen.

Auf internationaler Ebene wurde bei IEC Anfang 2019 die **SEG10** *Ethics in Autonomous and Artificial Intelligence Applications* gegründet. Das Gremium soll über einen Zeitraum von zwei Jahren evaluieren, in welchen Normungs- und Standardisierungsbereichen innerhalb der technischen Arbeit des IEC ethische Aspekte beim KI-Einsatz berücksichtigt werden müssen. Dabei sollen zunächst Leitlinien definiert werden, anhand derer technische Gremien (TCs) ethische Aspekte in Bezug zu automatisierten und/oder KI-Anwendungen identifizieren können. Oberstes Ziel ist die Formulierung von Normungs- und Standardisierungspotentialen, die mittelfristig in eine technische Arbeit übergehen können. Deutschland und China teilen sich den Vorsitz. Eine enge Kooperation mit JTC 1/SC 42 ist avisiert.

Auf europäischer Ebene wurde bei CEN-CENELEC im Jahr 2019 eine **Focus Group on Artificial Intelligence** gegründet, die eine Roadmap zu KI erarbeiten soll. Die Fokusgruppe unterstützt CEN-CENELEC bei der Untersuchung der Notwendigkeit einer europäischen Normung für KI innerhalb von CEN-CENELEC, unter Berücksichtigung der Leitlinien der *High Level Expert Group on Artificial Intelligence* (19) eingesetzt durch die Europäische Kommission und der COM(2018) 237[10] zum Thema Artifical Intelligence for Europe. Die Fokusgruppe soll eine gemeinsame Vision hinsichtlich der europäischen KI Normung erarbeiten und eine Übersicht über laufende Normungs- und Standardisierungsaktivitäten im Umfeld von KI erarbeiten. Hierbei soll unter anderem untersucht werden, in welchen technischen Gremien (TCs) ein KI-Bezug besteht und inwiefern hier Abstimmungsbedarfe existieren. Die *Focus Group* soll als Anlaufstelle für europäische TCs und für die Ambitionen der Europäischen Kommission zur Normung und Standardisierung von KI dienen. Die Gruppe sollte Empfehlungen ausarbeiten, wie die KI-Ethik im europäischen Kontext am besten angegangen werden kann[11].

Das nationale Spiegelgremium der internationalen Normung bei ISO/IEC Joint Technical Committee 1 Sub Committee 42 *Artificial Intelligence* und der europäischen CEN-CENELEC *Focus Group on Artificial Intelligence* ist der Arbeitsausschuss *Künstliche Intelligenz* im Normenausschuss Informationstechnik und Anwendungen (NIA)[12] bei DIN. Als Spiegelgremium erarbeiten 46 Expertinnen und Experten aus Forschung, Wissenschaft, Wirtschaft und der öffentlichen Hand die deutsche Position zu europäischen und internationalen Projekten und Themen und bringen diese in den jeweiligen Gremien ein. Der Arbeitsausschuss erarbeitet Vorschläge für Normungs- und Standardisierungsthemen auf all diesen Ebenen. Auf Initiative des Arbeitsausschusses *Künstliche Intelligenz* bei DIN wurde unter anderem das ISO/IEC-Normungsprojekt zum Thema Risikomanagement gestartet und aktuell eine Diskussion zum Thema Datenqualität im JTC 1/SC 42 initiiert, die im April 2020 zum Start der Abstimmung über eine Normenreihe zum Thema *Artificial intelligence — Data quality for analytics and machine learning (ML)* geführt hat. Des Weiteren übernehmen Expertinnen und Experten des Ausschusses aktive Rollen in der internationalen Normung.

10 COM (2018) 237: European Commission: https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF
11 CEN-CENELEC, CEN-CENELEC Focus Group on AI Terms of Reference: https://www.cencenelec.eu/news/articles/Pages/AR-2019-001.aspx
12 Informationen zum NIA finden sich hier https://www.din.de/de/mitwirken/normenausschuesse/nia

Das ISO Technical Committee (TC) 204 *Intelligent transport systems*, das Normung von Informations-, Kommunikations- und Steuerungssystemen im Bereich des städtischen und ländlichen Landverkehrs, einschließlich intermodaler und multimodaler Aspekte, Reiseinformationen, Verkehrsmanagement, öffentlicher Verkehr, gewerblicher Verkehr, Notfalldienste und kommerzielle Dienste im Bereich der intelligenten Verkehrssysteme (ITS) behandelt, hat eine Advisory Group zum Thema Big Data und Künstliche Intelligenz gegründet. Die Advisory Group wird Empfehlungen und einen Bericht zum Thema Big Data und Künstliche Intelligenz im Bereich des Anwendungsbereichs des ISO/TC 204 geben.

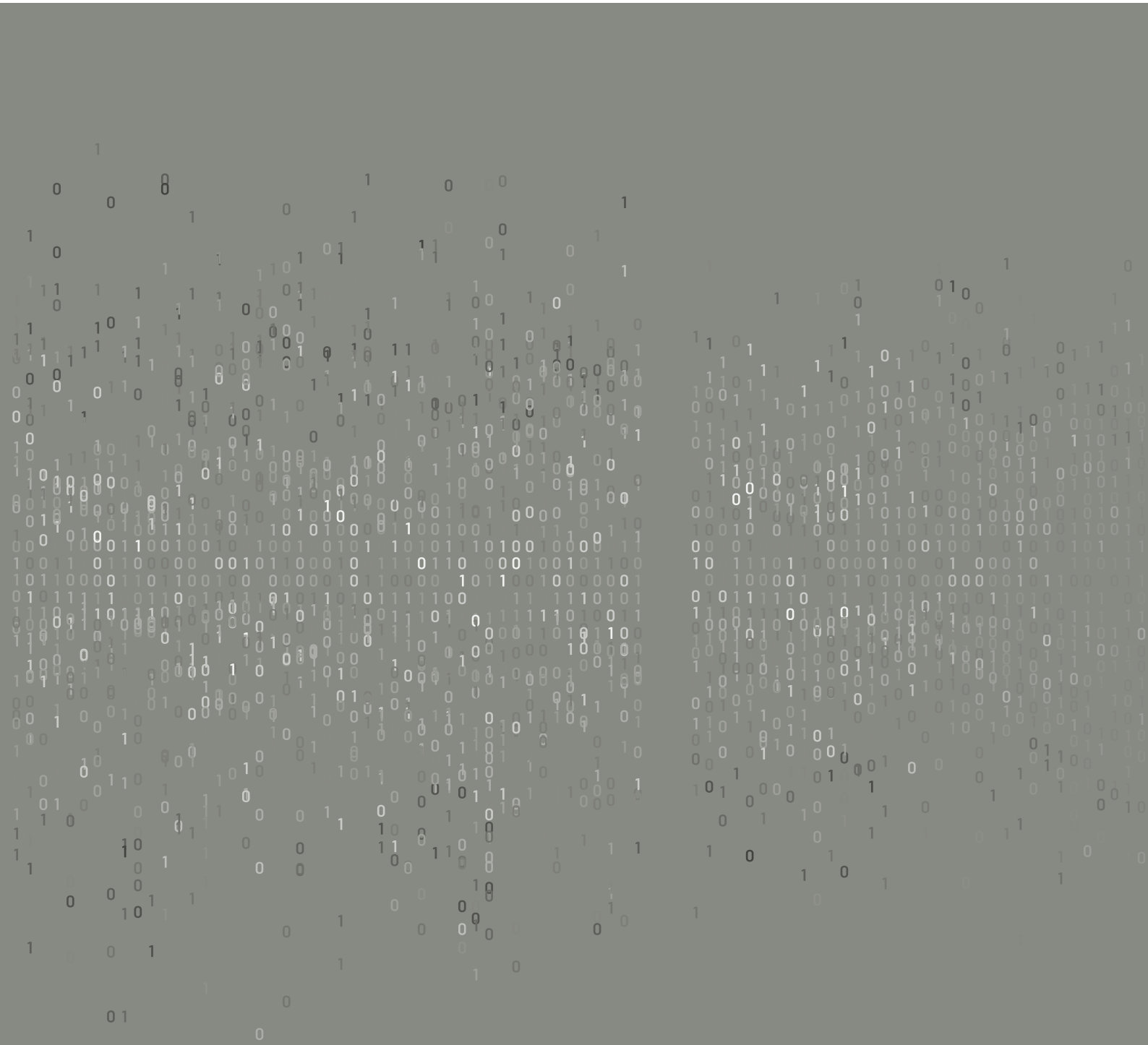
Im 2. Halbjahr 2019 wurde ein Normungsprojekt im ISO/TC 241 *Road traffic safety management systems* gestartet, das Richtlinien für Hersteller von automatisierten Fahrzeugen der Stufe 5 enthalten wird, wie sie von der internationalen Society of Automotive Engineers (SAE) 2014 definiert wurden: ISO/AWI 39003 *Road Traffic Safety (RTS) — Guidance on safety ethical considerations for autonomous vehicles*. Die Richtlinien werden die ethischen Erwägungen und Prioritäten definieren, die ein automatisiertes Fahrzeug der Stufe 5 haben sollte, wenn es wichtige Entscheidungen zum sicheren Fahren trifft. Es ist beabsichtigt, dass die Hersteller solcher Fahrzeuge ihre Fahrzeuge selbst als mit dieser Norm übereinstimmend zertifizieren, bevor sie das Fahrzeugmodell auf den Markt bringen.

Auf nationaler Ebene wurde 2019 die DIN SAE SPEC 91381:2019-06 *Begriffe und Definitionen in Bezug auf die Prüfung automatisierter Fahrzeugtechnologien* veröffentlicht. Die Spezifikation definiert Begriffe für den Bereich der automatisierten Fahrzeugtechnologien, insbesondere Begriffe zu Simulationen und Testumgebungen dieser Technologien. Sie soll als Werkzeug für die nächsten Entwicklungsschritte und Forschungsaktivitäten sowie zur besseren Kommunikation internationaler Partner dienen. Begriffe zu den Levels der Automatisierung und Fahrzeugparametern werden in dieser Spezifikation nicht definiert.

Eine bereits ausführliche Ausarbeitung zur ethischen Betrachtung automatisierter und intelligenter Systeme hat IEEE Global Initiative mit ihrer ersten Edition von *Ethically Aligned Design – A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*[13] veröffentlicht. Hierbei geht das internationale und interdisziplinäre Team von Autoren und Autorinnen davon aus, dass KI nur die volle Wirkung erreiche, wenn sie auf Werte und ethischen Prinzipien aufbaue. Als ethisch wird hierbei auch **soziale Fairness, ökologische Nachhaltigkeit** und die **menschliche Selbstbestimmung** verstanden. Bestandteil der formulierten Prinzipien ist die Beachtung der **Menschenrechte** und der **menschlichen Gesundheit**. Gleichzeitig sollten Nutzer stets über die **Nutzung von Daten** informiert sein und darüber entscheiden können. Dies gehe wiederum mit dem Grundsatz der **Transparenz** einher, die auch eine **Darlegung der Entscheidungswege** fordert. Diese technische Zuverlässigkeit müsse stetig überwacht werden und durch Validierung und Verifikation mittelfristig zertifizierbar gemacht werden. Konkrete Projekte zur Erarbeitung von IEEE Standards werden derzeit in der IEEE P7000-Reihe[14] umgesetzt, mit Schwerpunkten in den Bereichen Interoperabilität, Funktionalität und Sicherheit.

13 Mehr Informationen zu den Aktivitäten von IEEE sind hier zu finden: https://ethicsinaction.ieee.org/
14 Mehr Informationen zu der IEEE P7000-Reihe finden sich hier: https://ethicsstandards.org/p7000/

WELCHE NEUEN NORMEN UND STANDARDS WERDEN ZUKÜNFTIG BENÖTIGT?



Worin besteht das zukünftige Normungspotential für einen Einsatz von KI unter Berücksichtigung ethischer Aspekte?

Wo liegen die Bedarfe?

35

ZEHN ATTRIBUTE FÜR ETHISCHES VERHALTEN EINER KI

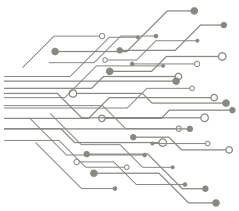
Aus der Analyse der in Anhang 1 aufgeführten Ethik-Richtlinien sowie der existierenden Normungs- und Standardisierungsaktivitäten ergeben sich die folgenden Attribute, die eine KI erfüllen soll, um sich ethisch richtig zu verhalten. Sie sind hier bereits nach Ähnlichkeit gruppiert und Dopplungen durch verschiedene Dokumente sind nicht dargestellt.

- Transparenz, Nachvollziehbarkeit, Erklärbarkeit, verantwortungsvolle Offenlegung
- Diskriminierungsfreiheit
- Überprüfbarkeit
- Achtung der Menschenwürde, Autonomie der Nutzer, Entscheidungsfreiheit der Nutzer, menschenzentrierter und -dienlicher Einsatzzweck
- Freiheit des Einzelnen
- Achtung von Demokratie und Rechtsstaatlichkeit
- Achtung von Gleichheit, Nichtdiskriminierung, alle Menschen fair behandeln
- Solidarität
- Schadensverhütung, sicher
- Fairness
- Menschenrechte
- demokratische Werte
- Vielfalt
- Robust, zuverlässig
- Einsatz von KI stets kenntlich gemacht werden
- Schutz der Privatsphäre
- Verantwortung

Die folgenden Attribute können mittels Normung und Standardisierung nicht gesondert abgebildet werden, da hier Anforderungen zunächst durch den Gesetzgeber festgelegt werden sollten. Technische Normen und Standards können gegebenenfalls anschließend bei der Konkretisierung unterstützen.

- Freiheit des Einzelnen
- Achtung von Demokratie und Rechtsstaatlichkeit
- Solidarität
- Menschenrechte
- demokratische Werte
- Verantwortung
- Diskriminierungsfreiheit, Vielfalt, Fairness

Aus diesen Einschränkungen resultieren zehn Attribute, die zunächst auf eine mögliche technische Umsetzung und anschließend auf die Chance, sie zu vereinheitlichen, untersucht wurden:



- Autonomie des Menschen
- Datenschutz
- Erklärbarkeit
- Reproduzierbarkeit
- Robustheit
- Rückverfolgbarkeit
- Sicherheit
- Transparente Kommunikation ob Mensch oder KI
- Verständlichkeit – effiziente, zuverlässige und sichere Kommunikation zwischen Mensch und Maschine
- Überprüfbarkeit

Alle zehn Attribute können durch technische Anforderungen adressiert werden. Die nachfolgende Analyse dieser Attribute umfasst zusätzlich eine Bewertung, ob die Festlegung einer technischen Anforderung durch Normen und Standards opportun ist. Grundlage dafür sind verschiedene Expertenworkshops zu denen DIN und DKE zur Erstellung des Whitepapers öffentlich eingeladen haben.

AUTONOMIE DES MENSCHEN

Der Begriff Autonomie bezeichnet im Allgemeinen eine Form der Selbstständigkeit – je nach Bezug (Politik, Verwaltung, Philosophie, usw.) mag die Bedeutung des Begriffes eine Färbung bzw. Schärfung erfahren. Im ethischen Kontext bezeichnet Autonomie die Selbstbestimmung einer Entität, insbesondere die des Menschen. Ein Mensch kann im ethischen Sinne als autonom bezeichnet werden, wenn die Bestimmungsgründe seines Handelns nicht „außen", sondern „innen" verortet sind, wenn also sein Handeln ausschließlich den Gesetzmäßigkeiten folgt, die seinem Gewissen entspringen (23).

Im vorliegenden Dokument wird unter Autonomie des Menschen insbesondere auch die Hoheit des Menschen über Entscheidungen einer KI verstanden – begründet durch die Definition der Ethik-Kommission eingesetzt durch den Bundesminister für Verkehr und digitale Infrastruktur. Diese schreibt in ihrem Bericht vom Juni 2017, Autonomie des Benutzers diene der *„[...] Steigerung von Mobilitätschancen und [der] Ermöglichung weiterer Vorteile [des Menschen]. Die technische Entwicklung gehorcht dem Prinzip der Privatautonomie im Sinne eigenverantwortlicher Handlungsfreiheit."* (16)

Ein ähnliches Begriffsverständnis verwendet die HLEG eingesetzt durch die Europäische Kommission in ihren Ethics Guidelines for Trustworthy AI:

„KI-Systeme sollten die menschliche Autonomie und Entscheidungsfindung, wie es der Grundsatz der Achtung der menschlichen Autonomie vorsieht, unterstützen. Dies erfordert, dass KI-Systeme sowohl einer demokratischen, florierenden und gerechten Gesellschaft dienen, indem sie das menschliche Handeln und die Wahrung der Grundrechte fördern, als auch die menschliche Aufsicht ermöglichen." (19)

Ansätze zur Standardisierung

Die Standardisierbarkeit von einer, die Autonomie des Menschen achtenden KI ist dem Grundsatz nach möglich. Dabei muss nicht zwangsweise ein Bezug zu einer bestimmten Anwendung der KI hergestellt sein.

Im Einzelfall bzw. wenn der Bezug zu einer bestimmten Anwendung der KI hergestellt wird, mögen Fragen darüber entstehen, wie genau der Umstand adressiert werden soll, dass die Grenze zwischen Assistenz und Manipulation des Menschen durch KI fließend ist. So mag es einerseits zu den vorgesehenen Aufgaben eines KI-gestützten Assistenzsystems gehören, einem Menschen Vorschläge zu unterbreiten, eine Vorauswahl für ihn zu treffen oder ihm Aufgaben bzw. Entscheidungen teilweise oder vollständig abzunehmen. Andererseits greift das Assistenzsystem dadurch zunächst in die Autonomie des Menschen ein.

Darüber hinaus mag die Autonomie des Menschen jedoch auch gestärkt werden, indem dieser durch das Assistenzsystem um gewisse Aufgaben entlastet wird und sich so anderen (bzw. „wichtigeren") Aufgaben zuwenden kann, bzw. auch äußerst zeitkritische oder komplexe Aufgaben bearbeitet werden können, die das Vermögen eines Menschen höchstwahrscheinlich übersteigen würden (z.B. die Auslösung eines Airbags).

Überprüfbarkeit der KI auf Achtung der Autonomie des Menschen: Im Sinne der Achtung der Autonomie des Menschen ist es grundsätzlich erwünscht, dass eine, von KI getroffene Entscheidung durch eine, von einem Menschen getroffene Entscheidung überstimmt werden kann – hierzu mag das Vorbild des „Not-Aus-Knopfes" aus dem Bereich der Funktionalen Sicherheit dienen. Im Einzelfall mag es angebracht sein, die Entscheidungshoheit dem Menschen zu entziehen oder diese auf einen Personenkreis zu beschränken. Diese Frage ist jedoch tendenziell gesellschaftspolitisch zu lösen.

Ein Aspekt der Verwendung von KI ist die Übergabe der Entscheidungsgewalt vom Menschen auf das KI-System und umgekehrt. Für manche KI-Anwendungen, wie beispielsweise dem automatisierten Fahren, erscheint es notwendig, die Frage zu adressieren, wie das KI-System darauf reagieren soll, wenn der Mensch (als Ausdruck seiner Autonomie) die Übernahme der Entscheidungsgewalt verweigert.

Ein weiterer Aspekt ist die Frage nach dem Umgang mit menschlichen Fehlern seitens des KI-Systems. Hier sollte bereits zum Zeitpunkt des Entwurfs einer KI die Frage beantwortet werden, wie das KI-System später auf menschlichen Fehlentscheidungen reagieren sollte, insbesondere, wenn das KI-System die (definitive) Fehlentscheidung des Menschen erkennt und diese rechtzeitig korrigieren könnte. Als Grundsatz mag hier gelten, dass einer KI umso eher Entscheidungshoheit gewährt werden kann, je eher sie dem Menschen in der Entscheidungsfindung nachweislich überlegen ist. Dies trifft insbesondere auf Anwendungen zu, die ausgesprochen zeitkritische und/oder objektive Entscheidungen verlangen und die Entscheidung des Menschen eventuell zu spät getroffen würde oder zu sehr von äußeren Einflüssen geprägt wäre.

DATENSCHUTZ

Die Europäische HLEG fordert, dass

„KI-Systeme während des gesamten Lebenszyklus eines Systems die Einhaltung des Datenschutz gewährleisten müssen. Dies umfasst die anfänglich vom Benutzer bereitgestellten Informationen sowie die Informationen, die über den Benutzer im Verlauf seiner Interaktion mit dem System generiert wurden (z. B. Ausgaben, die das KI-System für bestimmte Benutzer generiert hat oder wie Benutzer auf bestimmte Empfehlungen geantwortet haben). Durch digitale Aufzeichnungen des menschlichen Verhaltens können KI-Systeme nicht nur auf die Vorlieben von Personen schließen, sondern auch auf deren sexuelle Orientierung, Alter, Geschlecht, religiöse oder politische Ansichten. Damit Einzelpersonen dem Datenerhebungsprozess vertrauen können, muss sichergestellt werden, dass die über sie erhobenen Daten nicht dazu verwendet werden, sie rechtswidrig oder in unfairer Weise zu diskriminieren." (19)

Auch in der ISO/IEC TR 24028 — Information technology — Artificial intelligence (AI) — Overview of trustworthiness in artificial intelligence wird angemerkt, dass

„der Missbrauch oder die Offenlegung bestimmter Daten, insbesondere personenbezogener und sensibler Daten, schädliche Auswirkungen auf die betroffenen Personen haben [kann].“ (24)

Ansätze zur Standardisierung

Die DSGVO unterscheidet zwischen Entscheidungen von Menschen und Entscheidungen von Maschinen, wobei sie vollautomatisierte Entscheidungsfindungen verbietet. Es ist wichtig, dass immer ein „human in the loop“, also ein Mensch beteiligt ist. Beim Datenschutz in Deutschland und Europa ist die aktuelle Gesetzeslage anwendbar. Eine Ausarbeitung und technische Umsetzung dieser Gesetze kann und wird durch Normen nur ergänzt werden und hilft so bei deren Implementieren (z.B. DIN 66398).

ISO/IEC 38505-1:2017: „Information technology – Governance of IT – Part 1: Application of ISO/IEC 38500 to the governance of data“ legt Grundsätze für die wirksame, effiziente und akzeptable Verwendung von Daten fest und erklärt, dass unangemessene Verwaltung von Daten Vertrauensverlust für Organisationen bedeuten kann. Das Dokument listet unter den Verantwortungsbereichen von Leitungsorganen von Organisationen, dass diese sicherstellen müssen, dass es ein klares Verständnis darüber gibt, welche Daten von der Organisation genutzt werden und für welchen Zweck diese genutzt werden. Es bedarf eines wirksamen Managementsystems, um sicherzustellen, dass Verpflichtungen wie Datenschutz, Privatsphäre und Achtung des geistigen Eigentums erfüllt werden können.

ERKLÄRBARKEIT

Der Begriff Erklärbarkeit zielt im Kontext einer vertrauenswürdigen KI darauf ab, den Prozess der Entscheidungsfindung einer KI – die Verbindung(en) zwischen einem Eingangsdatensatz und dem resultierenden Endergebnis – für Menschen verständlich offen zu legen. Die Erklärbarkeit von KI-Systemen wird auch in der KI-Strategie der Bundesregierung als eine bedeutsame Komponente einer vertrauenswürdigen KI betrachtet, um diskriminierendem Verhalten vorzubeugen bzw. dieses sichtbar zu machen. Denn obwohl KI-Systeme von Menschen programmiert werden, können sie durch die schnelle Verarbeitung von großen Datenmengen Ergebnisse produzieren, deren Herleitung nicht einfach erklärbar sein kann. Es ist somit eine zentrale Anforderung einer ethischen KI, durch transparente Prozesse eine Erklärbarkeit der Performance zu erreichen.

In der Forschung und Diskussion über den KI-Einsatz hat sich der Begriff *Erklärbare Künstliche Intelligenz* (engl. ‚Explainable Artificial Intelligence‘, XAI) etabliert. Das Konzept der XAI soll nachvollziehbar machen, wie dynamische und nicht-linear funktionierende Systeme zu Ergebnissen kommen. Ohne ein solches Konzept wären KI-Systeme Black-Box-Systeme bei denen der Anwender keine Kontrollmöglichkeiten hat zu verstehen, wie Entscheidungen zustande gekommen sind. Ziel ist es, erklärbare Modelle zu produzieren, die gleichzeitig eine hohe Lernleistung beibehalten.

Die Erklärbarkeit der Performance von KI-Systemen wird unter anderem durch die HLEG gefordert und definiert:

„[...] processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected.“ (19)

Darüber hinaus fordert die HLEG, dass Entscheidungen (wenn möglich) für diejenigen, die von dem Einsatz direkt oder indirekt betroffen sind, erklärbar sein sollen. Auch Konzerne wie IBM schätzen die Erklärbarkeit von KI-Systemen und deren Entscheidungen als zentrales Vertrauensvehikel ein und konstatieren:

” Explainability is key for users interacting with AI to understand the AI's conclusions and recommendations. Your users should always be aware that they are interacting with an AI. “ (25)

Als nationaler Standard geht DIN SPEC 92001-1 auf die Erklärbarkeit bzw. Verständlichkeit von KI Systemen ein. Hierbei sollen alle möglichen Personen- und Interessensgruppen (‚Stakeholder') die Entscheidungen eines KI-Systems verstehen:

” Comprehensibility represents the degree to which a stakeholder with defined needs can understand the causes of an AI module's output. The causes include the reason for a specific output, i.e. the input leading on to it, and the whole process of decision-making. This means that the AI component is transparent and explainable. Furthermore, a qualitative understanding between the input variables and the response is provided with respect to the stakeholder's level of expertise and need for comprehension. “ (26)

Ein KI-System kann grundsätzlich die Anforderung der Erklärbarkeit erfüllen. Dabei kann über *process mining* oder *event logs* eine technische Erklärbarkeit realisiert werden, die in Retrospektive zu einem definierten Zeitpunkt die zugrunde liegende Datenbasis sowie die Entscheidungsherleitung nachempfinden können. Eine vorausschauende Antizipation von KI-Entscheidungen ist im Gegensatz dazu im Zweifelsfall nicht gleichermaßen möglich.

Ansätze zur Standardisierung

Die Erklärbarkeit von KI-Systemen lässt sich grundsätzlich durch Normen und Standards technisch regeln. Da KI-Systeme veränderlich sein können, beispielsweise, weil das KI-System laufend dazulernt und sein Verhalten entsprechend anpasst, mag ein KI-System trotz gleicher Entscheidungsgrundlage zu unterschiedlichen Zeitpunkten zu unterschiedlichen Ergebnissen kommen. In Konsequenz ist Erklärbarkeit zwar retrospektiv in der Regel möglich, während dies prospektiv nicht zwangsweise der Fall sein muss. Eine allgemeine Forderung nach Erklärbarkeit zu jedem beliebigen Zeitpunkt könnte sich daher im Einzelfall als unerfüllbar erweisen.

Zusätzlich sollte die Art der KI, die zum Einsatz kommt, bei der Analyse des Normungs- und Standardisierungsbedarfs berücksichtigt werden. Hier seien unterschiedliche Anforderungen zu berücksichtigen, beispielsweise für *supervised learning*-Systeme oder adaptive Systeme.

Konkrete Vorhaben im Bereich Normung und Standardisierung könnten die Definition und Ausgestaltung von Anforderungen der technischen Erklärbarkeit durch *process mining* oder *event logs* sein. Hierbei könnte beispielsweise die Häufigkeit solcher Mechanismen definiert werden.

Ein mögliches Normungs- und Standardisierungspotenzial liegt darin, KI-Modelle und -Systeme auf Erklärbarkeit hin zu überprüfen und somit auf ihre Konformität hin zu bewerten. Hierbei sollte stets die anwendungsbezogene Notwendigkeit in den Prozess mit einbezogen werden. In vielen Fällen ist die Erklärbarkeit von KI-Entscheidungsmustern relevant und dringend geboten (z.B. beim vollautomatisiertes Fahren). In manchen Fällen ist die Relevanz möglicherweise nicht eindeutig, da weniger der KI-Entscheidungsprozess von Interesse ist, sondern beispielsweise die Datengrundlage. Somit wird es als wichtig erachtet, zur Überprüfung der Erklärbarkeit auf die Datensätze bzw. Modelle, mit denen die KI trainiert wurde, sowie gegebenenfalls den Zeitpunkt einer bestimmten Entscheidung durch das KI-System zurückgreifen zu können. Andernfalls mag die Entscheidung einer KI nicht detailliert nachvollzogen werden können. Der Grad der Erklärbarkeit und die Notwendigkeit dessen können sich aber dynamisch darstellen und von den Umständen abhängen. Somit werden hier die kontextbezogene Normung und Standardisierung präferiert.

Der Einsatz einer KI soll vom Menschen als vertrauenswürdig und sinnvolle Unterstützung wahrgenommen werden. Um dieses Ziel zu erreichen, wird eine ausreichende Erklärbarkeit der KI als wichtiger Faktor gesehen. Da es vorwiegend in den Händen des Entwicklers oder der Entwicklerin eines KI-Systems liegt, wie genau die Interaktion zwischen Mensch und Maschine stattfinden wird, kommt ihm oder ihr hierbei eine tragende Rolle zu, die Entscheidungsprozesse des KI-Systems entsprechend einsehbar bzw. erklärbar zu gestalten und so einer Black-Box-Problematik entgegenzuwirken. Ziel muss es sein, KI-Systeme als künstlich-intelligenten Partner zu verstehen, denen Menschen angemessen vertrauen können.

REPRODUZIERBARKEIT

Die HLEG definiert Reproduzierbarkeit wie folgt:

„Ein KI-Experiment zeigt das gleiche Verhalten, wenn es unter den gleichen Bedingungen wiederholt wird.“ (19)

Ergebnisse von KI-Systemen, die auf Machine Learning Technologien beruhen und in der Anwendung weiterlernen, sind nur schwierig bis gar nicht reproduzierbar. Um ein Machine Learning System, das in der Anwendung weiterlernt ‚zu reproduzieren, kann versucht werden, dies anhand eines Ähnlichkeitsmaß zu messen, da situationsbedingt Abweichungen wahrscheinlich und eine vollständige Reproduktion des Systems unwahrscheinlich ist.

Ansätze zur Standardisierung

Aufgrund der Divergenz der Reproduzierbarkeit der Ergebnisse verschiedener KI-Technologien ist es wichtig, zu definieren, welche Anforderungen KI-Systeme mit bestimmten Anwendungscharakteristiken haben müssen und welche KI-Systeme eventuell für bestimmte sensible Anwendungsfelder nicht empfehlenswert sind.

Eine Prüfnorm könnte die Reproduzierbarkeit von Ergebnissen eines KI-Systems anhand definierter Merkmale und Prozesse messen und eine Klassifizierung vornehmen.

ROBUSTHEIT

Die Robustheit eines Systems beschreibt die Eigenschaft, unter allen Bedingungen noch angemessen zuverlässig und sicher zu funktionieren. Hierbei spielt das Konzept der Fehlertoleranz eine Rolle, das Abweichungen der Normalfunktion zulässt, ohne das die Funktionalität des Systems in relevantem Maße beeinträchtigt wird. Im Softwarebereich wird Robustheit als Qualitätskriterium verstanden, obwohl eine Hundertprozentige Robustheit nicht erreichbar ist. Beim KI-Einsatz werden Anforderungen an das System gestellt, da es auf sich verändernde Umwelteinflüsse reagieren soll. Eine hohe Robustheit ist notwendig, um ein Vertrauen in KI-Modelle und –Systeme herzustellen.

Die Robustheit von KI-Systemen wird unter anderem durch die HLEG gefordert und definiert:

„Die Robustheit eines KI-Systems umfasst sowohl seine technische Robustheit (Angemessenheit für einen bestimmten Kontext, wie z. B. den Anwendungsbereich oder die Phase des Lebenszyklus) als auch seine soziale Robustheit (Gewährleistung einer angemessenen Berücksichtigung des Kontexts und der Umgebung, in denen das System eingesetzt wird). Das ist entscheidend, um sicherzustellen, dass auch bei guten Absichten keine unbeabsichtigten Schäden auftreten können. Robustheit ist das dritte der drei Schlüsselelemente, die für die Erreichung einer vertrauenswürdigen KI erforderlich sind.“ (19)

Im Draft ISO/IEC TR 24029-1 wird Robustheit als die Fähigkeit eines KI-Systems, sein Leistungsniveau unter allen Bedingungen aufrechtzuerhalten, definiert.

Ansätze zur Standardisierung

Die Definition von Robustheit von KI-Systemen lässt sich durch Normen und Standards technisch regeln. Da Robustheit auch als Qualitätskriterium angesehen wird, müssen Robustheitsanforderungen definiert werden, welche an den Anwendungsbereich des KI-Systems und den direkten Einfluss auf den Menschen angepasst sein sollte.

Möglich wäre die Entwicklung einer Definition unterschiedlicher Level von Robustheit, um speziell auf die Anwendungsbereiche und Bedürfnisse einzugehen. Ein Ansatz, unterschiedliche Level von Robustheit zu definieren, kann anhand der Regelungen zur Sicherheitsanforderungsstufe (Sicherheits-Integritätslevel; SIL) aus dem Bereich der Funktionalen Sicherheit identifiziert wer-

den. Hierbei werden Sicherheitsfunktionen auf deren Zuverlässigkeit bewertet und verschiedenen Ebenen (‚Level') zugeordnet und für die jeweilige Sicherheitsfunktion festgelegt. Dabei stellt die Sicherheitsanforderungsstufe ein Maß für die Zuverlässigkeit des Systems in Abhängigkeit der Gefährdung dar.

Die Akzeptanz in die Funktion von KI-Systemen und –Modellen wird maßgeblich von deren Robustheit bestimmt. Insbesondere wenn Menschen mit den Systemen in Kontakt kommen, muss die sichere Funktionalität gewährleistet sein. Der Kontakt mit dem Mensch stellt die höchsten Sicherheitsanforderungen an ein KI-System, da dabei im Zweifel die menschliche Integrität beeinflusst werden kann. Eine hohe Robustheit eines KI-Systems bedeutet somit, dass der Mensch jederzeit in Prozesse eingreifen kann, bzw. diese stoppen kann.

Um die Robustheit eines KI-Systems zu testen gibt es verschiedene Methoden: statistische Methoden, formale Methoden und empirische Methoden. Da KI-Systeme einige Besonderheiten, insbesondere beim Testen, zeigen, müssen die vorhandenen Methoden dahingehend analysiert und bewertet werden. Ein einheitliches Verfahren und eine Klassifizierung oder Bewertung kann eine größere Vergleichbarkeit zwischen verschiedenen KI-Systemen schaffen und das Vertrauen in KI-Anwendungen steigern (Prüfnorm). Zusätzlich müssten Test und Testdesigns definiert werden, um KI-Modellen und –Systemen Eigenschaften zuzusichern.

RÜCKVERFOLGBARKEIT

Die HLEG versteht unter dem Begriff Rückverfolgbarkeit folgendes:

„ Die Rückverfolgbarkeit eines KI-Systems bezieht sich auf die Möglichkeit, die Daten-, Entwicklungs- und Bereitstellungsprozesse des Systems nachzuvollziehen, in der Regel anhand von dokumentierten Prozessaufzeichnungen. “ (19)

Ansätze zur Standardisierung

Die Rückverfolgbarkeit von KI-Systemen lässt sich grundsätzlich durch Normen und Standards technisch regeln. Für jedes System sollten Informationen des System Engineerings, der Systemdefinition und des Betriebskonzepts dargelegt werden. Diese Informationen sollten an die Zielgruppe (Entwickler und Entwicklerinnen, Anwender und Anwenderinnen, etc.) in ihrer Komplexität und Detailtiefe angepasst werden.

Anhand einer Prüfnorm sollten unter anderem die folgenden Parameter rückverfolgt werden können:

- Datensätze
- Prozessbeschreibung
- Feldversuche
- Anwendungsfelder (use cases)
- Sicherheitsrelevante Aspekte

Darauf basierend kann sich eine Konformitätsbewertung des KI-Systems anschließen.

Ein anderer Aspekt, der durch Normung konkretisiert werden kann, ist die Definition und Ausgestaltung von Anforderungen der technischen Erklärbarkeit durch *process mining* oder *event logs*. Hierbei könnte beispielsweise die Häufigkeit solcher Mechanismen definiert werden. Auch könnten Aspekte der Prüfung konkret in Normen und Standards geregelt werden (Prozesse, Anforderungen, Prüfkompetenzen).

SICHERHEIT

Sicherheit ist ein Überbegriff, der bei einer Vielzahl von Systemanforderungen vorgegeben wird. Der deutsche Begriff wird dabei der Unterscheidung zwischen „safety" und „security" nicht gerecht. Dabei ist diese Trennung äußerst relevant, da sie unterschiedliche Anforderungen definiert. *Security* bedeutet Sicherheit im Sinne von Schutz vor Angriffen, wohingegen *Safety* den Schutz vor Gefährdungen meint, die von Systemen selbst ausgehen. Wie für alle informationstechnischen Systeme sind sowohl *Safety* als auch *Security* beim KI-Einsatz notwendig und müssen in der Systemkonzeption stets bedacht werden. Gefährdungen können sowohl durch einen Funktionsausfall oder -änderung definiert sein, die durch das System selbst (‚Systemfehler') oder durch unbefugten Einfluss von außen entstehen können. Dies gilt insbesondere für soziotechnische Systeme, die z.T. durch KI-Systeme abgebildet werden.

Die HLEG definiert die *Safety*-Anforderungen ausgehend vom Systemeinsatz wie folgt:

„The level of safety measures required depends on the magnitude of the risk posed by an AI system, which in turn depends on the system's capabilities. Where it can be foreseen that the development process or the system itself will pose particularly high risks, it is crucial for safety measures to be developed and tested proactively." (19)

Ansätze zur Standardisierung

Im Bereich der IT-Sicherheit existieren bereits eine Reihe etablierter (v.a. internationaler) Normen und Standards. Diese sind nicht explizit auf KI-Systeme ausgerichtet bieten aber einen hilfreichen Rahmen für den sicheren KI-Einsatz. Ziel muss es sein, bestehende Normen und Standards, die für den Schutz von KI-Anwendungen relevant sind, zusammenzuführen und auf die KI-Einsatzbereiche zu prüfen. Darüber hinaus sind neue Normen und Standards zu erarbeiten, die sich aus den bereits aufgeführten Standardisierungspotentialen ergeben. Hierbei muss beachtet werden, dass, wie bei bestehenden IT-Security Normen, eine horizontale Betrachtung des Themas zuerst erfolgen muss, so dass sich die vertikalen Anwendungsgebiete der Sicherheit von KI-Systemen anhand dieser Grundlage orientieren können. Dadurch kann eine einheitliche Vorgehensweise gefördert werden.

Eine standardisierbare Möglichkeit die Sicherheit eines KI-Systems zu stärken besteht beispielsweise darin, das KI-System mit einer deterministischen Schleife zu umgeben. Ferner sollte nicht nur der Schutz von KI-Systemen, sondern auch der Schutz mit Hilfe von KI-Systemen betrachtet werden.

TRANSPARENTE KOMMUNIKATION OB MENSCH ODER KI

Die HLEG fordert Transparenz in der Kommunikation mit einer KI:

„KI-Systeme dürfen gegenüber den Nutzern nicht als Menschen auftreten. Menschen haben das Recht, darüber informiert zu werden, dass sie mit einem KI-System interagieren. Dies bedeutet, dass KI-Systeme als solche erkennbar sein müssen. Zur Gewährleistung der Einhaltung der Grundrechte sollte darüber hinaus bei Bedarf die Möglichkeit bestehen, sich gegen diese Interaktion und zugunsten einer zwischenmenschlichen Interaktion zu entscheiden. Darüber hinaus sollten die Fähigkeiten und Einschränkungen des KI-Systems den Anwendern der KI und den Endnutzern in einer dem jeweiligen Anwendungsfall angemessenen Weise mitgeteilt werden. Das könnte Informationen zur Präzision des KI-Systems sowie seiner Grenzen umfassen." (19)

Ansätze zur Standardisierung

Die Entscheidung, ob sich KI-Systeme als solche zu erkennen geben müssen, ist eine Entscheidung, die auf regulatorischer Ebene getroffen werden sollte. Eine konkrete Ausarbeitung und die technische Umsetzung einer solchen Vorschrift kann durch Normen ergänzt und konkretisiert

werden. Aufgrund der verschiedenen Einsatzmöglichkeiten und -situationen, ist es wichtig zu definieren, wie sich das KI-System als solches zu erkennen gibt (einmalige oder wiederholte Information, schriftlich, per Lautsignal, etc.). Diese Details können normativ erarbeitet werden.

VERSTÄNDLICHKEIT – EFFIZIENTE, ZUVERLÄSSIGE UND SICHERE KOMMUNIKATION ZWISCHEN MENSCH & MASCHINE

Die Ethik-Kommission eingesetzt durch den Bundesminister für Verkehr und digitale Infrastruktur schreibt in Abschnitt drei „Ethische Regeln für den automatisierten und vernetzten Fahrzeugverkehr" in ihren Bericht vom Juni 2017:

„ Software und Technik hochautomatisierter Fahrzeuge müssen so ausgelegt werden, dass die Notwendigkeit einer abrupten Übergabe der Kontrolle an den Fahrer („Notstand") praktisch ausgeschlossen ist. Um eine effiziente, zuverlässige und sichere Kommunikation zwischen Mensch und Maschine zu ermöglichen und Überforderung zu vermeiden, müssen sich die Systeme stärker dem Kommunikationsverhalten des Menschen anpassen und nicht umgekehrt erhöhte Anpassungsleistungen dem Menschen abverlangt werden. " (16)

In diesem Sinne regelt der neu geschaffene § 1a Abs. 2 Nr. 5 StVG, dass Kraftfahrzeuge mit hoch- oder vollautomatisierter Fahrfunktion über eine technische Ausrüstung verfügen müssen, die dem Fahrzeugführer oder der Fahrzeugführerin das Erfordernis der eigenhändigen Fahrzeugsteuerung mit ausreichender Zeitreserve vor der Abgabe der Fahrzeugsteuerung an den Fahrzeugführer oder die Fahrzeugführerin optisch, akustisch, taktil oder sonst wahrnehmbar kommunizieren.

Ansätze zur Standardisierung

Die Anforderungen an das Kommunikationsverhalten des KI-Systems können normativ festgelegt werden. Normativ behandelt werden könnten beispielsweise Zeitpunkt und Nachdruck der Übergabeaufforderung durch das KI-System – beide sollten sich nach dem Automatisierungsgrad der Anwendung und der Dringlichkeit des menschlichen Eingreifens richten. Auch das Thema einer ausreichenden Zeitreserve zwischen einer „rechtzeitigen" Übergabeaufforderung und der tatsächlichen Übergabe kann bzw. sollte behandelt werden. Beispielsweise mag im Rahmen der vollautomatisierten Fortbewegung eine Zeitreserve von 40 s angemessen erscheinen, damit sich der Fahrer wieder umfassend in das Fahrgeschehen und -umfeld einarbeiten kann. In Einzelfällen mögen auch kürzere Übergabefrist angemessen sein (z.B. im Parkhaus). Ein Ausdefinieren aller denkbaren Einzelfälle erscheint dabei nicht zielführend, wohl aber die Schaffung von Mindeststandards zu diesem Thema.

Darüber hinaus ist die Mensch-Maschine-Schnittstelle bei einer vertrauenswürdigen KI für die Kommunikation überaus wichtig. Falls es zum direkten Kontakt zwischen Mensch und Maschine kommt, sollte die Schnittstelle die Möglichkeit bieten, anwenderspezifisch personalisiert zu werden, um das Wohlbefinden des Benutzers oder der Benutzerin und sein bzw. ihr Vertrauen in die KI zu stärken.

ÜBERPRÜFBARKEIT

Ein System ist überprüfbar, wenn eine Bestätigung durch objektive Nachweise, dass die festgelegten Anforderungen erfüllt sind, möglich ist (ISO/IEC TR 29110-1:2016). Hierbei kann getestet werden, ob ein KI-Modul hinreichend gut gemäß seiner Spezifikation aufgebaut ist. In der Entwicklungs-, Bereitstellungs- und Betriebsphase eines KI-Modullebenszyklus werden für KI-Systeme Anforderungen oftmals in impliziter Form als Lerndaten in eine Architektur umgesetzt und anschließend in Code umgesetzt. Eine Verifizierung kann sicherstellen, dass der resultierende Code in Bezug auf ein definiertes Testdatenset korrekt ist. Eine explizite, quantisierbare Anforderungsspezifikation, menschliche Plausibilisierung und Überprüfung definierter Datenqualitätsmerkmale sind wichtige Mittel zur Systemverifikation (DIN SPEC 92001-1: 2019). Bei der Überprüfbarkeit

einer KI Technologie ist zu beachten, dass die Anforderungen an deren Überprüfbarkeit je nach Anwendung und Einbettung der Technologie variieren kann, sie ist kontextabhängig. Die Modellierung sollte dem sozialen Prozess angepasst sein. Unter anderem kann die Einhaltung von Zielvorgaben und Merkmalen (deren Qualität eingeschlossen), sowie Datenqualität und Eigenschaftsvektoren überprüft werden.

Ein Qualitätsmaß kann zum Beispiel die Erfüllung der Zielvorgaben zu Fairness sein. Je nach KI Technologie ist es erforderlich, ein Black-Box-Verfahren zur Überprüfung zu nutzen (z.B. künstliche neuronale Netze), hierbei werden bestimmte Eigenschaften dem System vorgegeben und die Eigenschaften des Ergebnisses anhand von ausgewählten Stichproben bewertet.

Die HLEG spricht in diesem Zusammenhang von dem Begriff „Nachprüfbarkeit".

„Nachprüfbarkeit bezieht sich auf die Fähigkeit eines KI-Systems, einer Bewertung der Algorithmen, Daten und der Verfahren zum Entwurf des Systems unterzogen zu werden. Nachprüfbarkeit ist eine der sieben Anforderungen, die ein vertrauenswürdiges KI-System erfüllen sollte. Dies bedeutet nicht unbedingt, dass Informationen über Geschäftsmodelle und geistiges Eigentum im Zusammenhang mit dem KI-System offengelegt werden müssen. Mit der Sicherstellung von Verfahren zur Rückverfolgbarkeit und Berichterstattung bereits in der frühen Entwurfsphase des KI-Systems kann ein Beitrag zur Nachprüfbarkeit des Systems geleistet werden." (19)

Im ISO/IEC DTR 24028:2019 wird zudem der englische Begriff „Controllability" verwendet.

„Die Kontrollierbarkeit kann erreicht werden, indem zuverlässige Mechanismen bereitgestellt werden, mit denen ein Operator die Kontrolle über das KI-System übernehmen kann. Um diese Kontrollierbarkeit zu erreichen, muss als erstes beantwortet werden, wer inwieweit Kontrolle über wessen KI-Systeme erhält, an denen mehrere Stakeholder beteiligt sind, z. der Dienstleister oder Produktanbieter, der Anbieter der konstituierenden KI, der Benutzer oder ein Akteur der Aufsichtsbehörde."

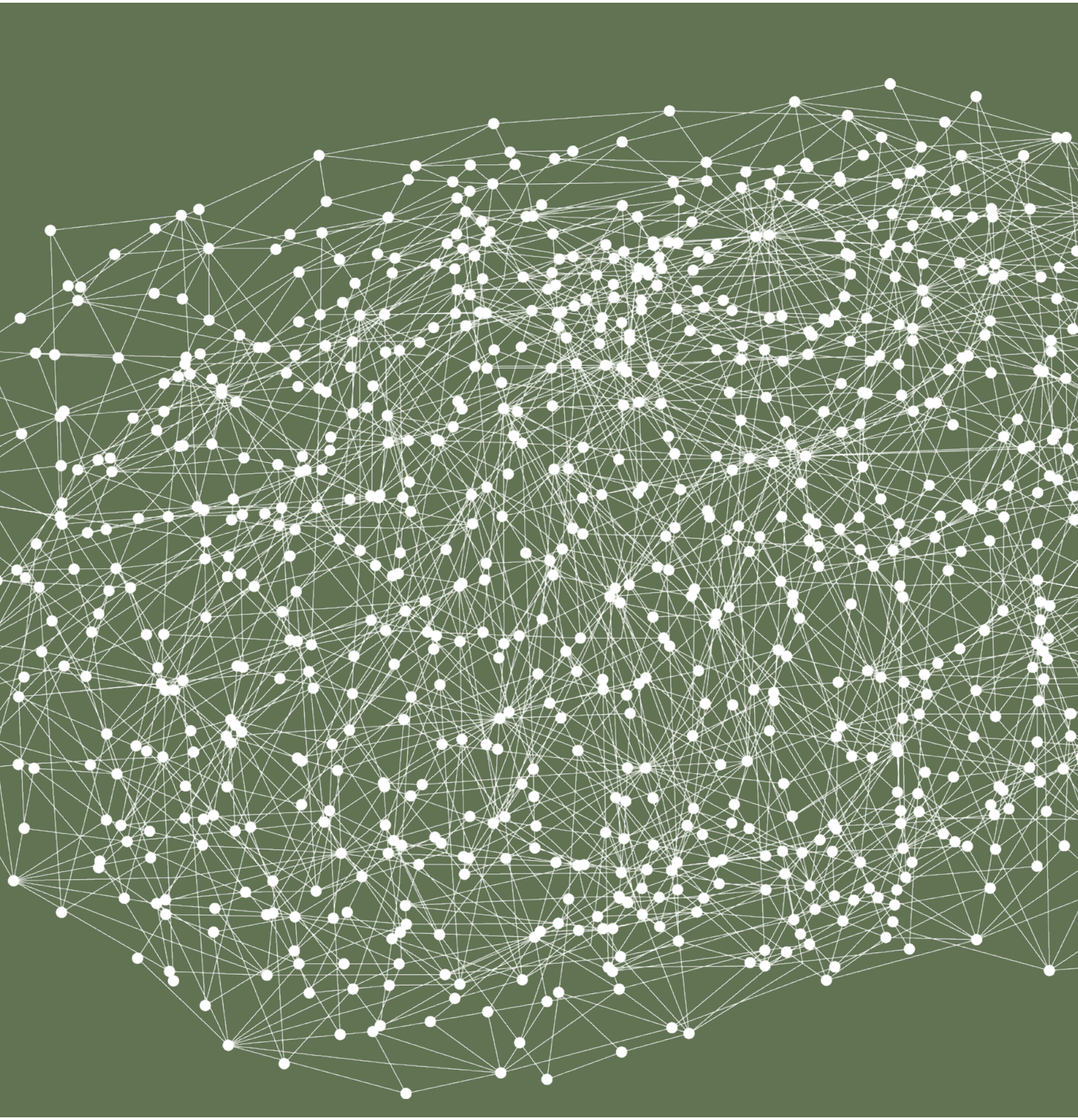
Ansätze zur Standardisierung

KI-Systeme sind per Design oft weniger deterministisch als herkömmliche Softwaresysteme und selten vollständig erklärbar. Die Software eines KI-Systems umfasst sowohl KI- als auch Nicht-KI-Komponenten. Damit ein KI-System ordnungsgemäß funktioniert, müssen alle seine Komponenten den anerkannten Soft- und Hardwarepraktiken folgen (einschließlich Unit- und Funktionstests), während seine KI-Komponenten (diejenigen, die Logik mit KI-Technologien implementieren) eine modifizierte Version dieser Praktiken verwenden würden.

Je nach Einsatzgebiet und Kombinierung des KI-Systems werden unterschiedliche Anforderungen an die Überprüfbarkeit einer Technologie gestellt. Diese Anforderungen müssen festgelegt werden. Eine Möglichkeit dies umzusetzen ist eine Klassifizierung anhand einer Norm bzw. eines Standards. Eine andere Option ist eine Regulierung des Gesetzgebers.

Es ist notwendig, dass Funktionstests bei KI-Systemen gegebenenfalls mit Unsicherheiten umgehen können. Es stellt eine Herausforderung dar, die Anforderungen an nicht-deterministische Softwarekomponenten unter Verwendung bestehender Standards und Praktiken zu spezifizieren und/oder zu testen. Es kann zusätzlich schwierig sein, festzustellen, ob ein einzelner Test seine Erfolgskriterien erfüllt hat. (24) Normen und Standards, die Techniken zur Verifikation und Validierung von KI-Systemen beschreiben, können diese Lücke füllen (z.B. Prüfnormen und Prozessnormen). Hierfür sollte auf den bestehenden Normen aufgebaut werden und diese mit spezifischen KI-Anforderungen ergänzt werden, wie in der aktuell entstehenden Risikomanagement Norm zu KI: ISO/IEC 23894 *Information technology - Artificial intelligence - Risk management*.

Je höher die Automatisierung der KI Technologie ist, desto größer muss das Level der Überprüfbarkeit sein. Eine Organisation, die ein KI-System entwickeln oder anwenden möchte, sollte prüfen, welche Risiken im Allgemeinen nicht akzeptabel sind und anhand dieser Prüfung Anfor-



HANDLUNGSEMPFEHLUNGEN
für zukünftige Normungs- und Standardisierungsaktivitäten

47

Wie können die Anforderungen aus den Ethik-Richtlinien in Normen und Standards abgebildet werden?

Auf einige Anforderungen können Normen und Standards Antworten liefern und bei anderen können mögliche zukünftige Normen und Standards unterstützen.

In diesem Kapitel werden daher die oben dargestellten Vorschläge für zukünftige Normungs- und Standardisierungsaktivitäten zu den Forderungen aus den Ethik-Richtlinien in Bezug gesetzt und als Handlungsempfehlungen formuliert.

Über ihre Umsetzung entscheiden die entsprechenden Normungsgremien.

Die Ethikkommissionen der europäischen und deutschen Regierung, Bundesverbände und Unternehmen fordern den ethisch wertvollen Einsatz von Künstlicher Intelligenz. Im Kapitel „Was zeichnet eine Künstliche Intelligenz aus?“ wurden die verschiedenen Positionen und Akteure vorgestellt. Aus ihren Ansichten und den Diskussionen von drei Expertenworkshops ergaben sich für das Projekt zehn Attribute, die zur Wahrung ethisch wertvollen Verhaltens eines automatisierten Fahrzeuges oder einer automatisierten Maschine, welche durch eine KI gesteuert wird, beitragen. Welche dieser Attribute durch Normen oder Standards definiert und/ oder gefestigt werden können, wurde im Kapitel „Zehn Attribute für ethisches Verhalten einer KI“ aufgezeigt.

Die im Kapitel „Welche neuen Normen und Standards werden zukünftig benötigt?“ vorgestellten Vorschläge für zukünftige Normungs- und Standardisierungsaktivitäten liefern Möglichkeiten, den in den unterschiedlichen Ethik-Richtlinien (s. Anhang 1) geforderten Attributen einer ethisch wertvollen KI mittels Normen und Standards zu begegnen. Diese Bedarfe sind nun hier ohne Priorisierung in acht Handlungsempfehlungen formuliert.

I. Es wird empfohlen, dass sich die entsprechenden Normungsgremien mit dem Aufstellen von Prüfkriterien, mit denen eine KI auf die Achtung der Autonomie des Menschen überprüft werden kann, befassen. Dabei sei die Forderung der Ethikkommission automatisiertes Fahren des BMVI (16) nach einer Steigerung der Mobilitätschancen des Menschen zu berücksichtigen sowie die Anforderung der HLEG, „dass KI-Systeme sowohl einer demokratischen, florierenden und gerechten Gesellschaft dienen, indem sie das menschliche Handeln und die Wahrung der Grundrechte fördern, als auch die menschliche Aufsicht ermöglichen.“ (19) Wann die Entscheidungshoheit dem Menschen entzogen werden sollte und der KI überlassen wird, ist gesellschaftspolitisch zu lösen.

II. Der Datenschutz ist durch die aktuelle Gesetzeslage geregelt. Die Ausarbeitung und technische Umsetzung dieser Gesetze in Bezug auf eine KI kann durch Normen nur ergänzt werden (z.B. DIN 66398).

III. Es wird empfohlen, dass sich die entsprechenden Normungsgremien mit dem Aufstellen von Prüfkriterien in Bezug auf die Erklärbarkeit einer KI befassen. Um eine umfängliche Überprüfung der Erklärbarkeit zu ermöglichen wird empfohlen, die jeweils relevanten Trainingsdatensätze und -modelle eines KI-Systems sowie den Zeitstempel einer zu Überprüfenden Entscheidung vorzuhalten.

IV. Der Grad der Erklärbarkeit erscheint auf Grundlage der vorliegenden Analyse kontextbezogen. Entsprechend sollte auch die Normung und Standardisierung den Kontext der KI-Anwendung berücksichtigen. Zudem wird eine Berücksichtigung der Art der KI empfohlen. Diese Normen und Standards sollen u.a. dem übergeordneten Ziel dienen, den Menschen zu ermächtigen, informierte Entscheidungen in Bezug auf KI-Systeme zu treffen, und damit ein dauerhaftes Vertrauen in diese Technologie aufzubauen.

V. Die Ethikkommission automatisiertes Fahren des BMVI fordert: „Technik muss nach ihrem jeweiligen Stand so ausgelegt sein, dass kritische Situationen gar nicht erst entstehen […]“. Zudem fordert die HLEG: „Ein System ist dann zuverlässig, wenn es mit einer Reihe von Eingaben und in verschiedenen Situationen einwandfrei funktioniert. Dies ist erforderlich, um ein KI-System zu überprüfen und unerwünschte Schäden zu vermeiden. Wiederholbarkeit beschreibt, ob ein KI-Experiment das gleiche Verhalten aufweist, wenn es unter gleichen Bedingungen wiederholt wird.“ Weil KI-Systeme, je nach eingesetzter Technologie, nicht unbedingt reproduzierbare Ergebnisse liefern, wird empfohlen, eine Prüfnorm mit dem Ziel der Klassifizierung von KI-Systemen zu erstellen. Dabei soll definiert werden, welche Anforderungen KI-Systeme mit bestimmten Anwendungscharakteristiken haben müssen und welche KI-Systeme eventuell aufgrund ihres niedrigen Levels an Reproduzierbarkeit für bestimmte sensible Anwendungsfelder nicht empfehlenswert sind.

VI. Von der Ethikkommission des BMVI sowie der HLEG wird ein robustes KI-System gefordert, mit dem übergeordneten Ziel, das Vertrauen in die Technologie zu stärken. Basierend auf den Projektergebnissen wird empfohlen, dass sich Normung und Standardisierung ausschließlich auf technische Robustheit konzentrieren. Hierbei können Normen oder Standards entwickelt werden, die den unterschiedliche Level von Robustheit – analog zu den Regelungen der Sicherheitsanforderungsstufen SIL – definieren. Andererseits wird empfohlen, Robustheit als Qualitätskriterium anzusehen und dadurch Anforderungen durch eine Norm oder einen Standard zu definieren.

VII. Es wird empfohlen, bereits bestehende Normen und Standards, die für den Schutz von KI-Anwendungen relevant sind, zusammenzuführen und auf die KI-Einsatzbereiche zu prüfen, um der geforderten Sicherheit eines KI-Systems gerecht zu werden.

VIII. Es wird empfohlen, die Entscheidung, ob sich KI-Systeme als solche zu erkennen geben müssen, auf regulatorischer Ebene zu treffen. Eine konkrete Ausarbeitung und die technische Umsetzung einer solchen Vorschrift kann durch Normen und Standards ergänzt und konkretisiert werden.

Eine gestiegene Rechenleistung, günstiger Speicherplatz und eine fortschreitende Digitalisierung von Gesellschaft und Industrie treiben die Entwicklung und den Einsatz von KI weiterhin stark an. Das Potential von KI wird dabei als hoch eingeschätzt. So betrachtet die Bundesregierung KI als Schlüsseltechnologie, die für die wirtschaftliche Leistung des Landes von Bedeutung ist. Gleichzeitig wirft der Einsatz von KI eine Reihe von ethischen Fragen auf, die in einem politischen und gesellschaftlichen Prozess derzeit und zukünftig bearbeitet werden.

Das Whitepaper zeigt auf, welche zentrale Rolle die Normung und Standardisierung bei der Entwicklung und dem Einsatz von KI spielen kann.

Es wird der aktuelle Stand der Diskussion zu Ethik und KI im Bereich automatisierter Fahrzeuge und Maschinen dargestellt und Normungs- und Standardisierungsbedarfe werden aufgezeigt. Die zentrale Leitfrage, ob Normung und Standardisierung hierbei eine bedeutsame Rolle spielen kann, wird positiv beantwortet und durch die genannten Handlungsempfehlungen belegt.

LITERATURVERZEICHNIS

1. **S.J. Russell, P. Norvig.** *Artificial Intelligence. s.l. :* Pearson Education Inc, 2010.
2. **Bundesregierung.** *Strategie Künstliche Intelligenz der Bundesregierung.* 2018.
3. **Shane, Janelle.** *The danger of AI is weirder than you think. s.l. :* TED2019, 2019.
4. **Dastin, Jeffrey.** *Amazon scraps secret AI recruiting tool that showed bias against women.* s.l. : Reuters, 2018.
5. **Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner.** *Machine Bias.* s.l. : ProPublica, 2016.
6. **Lenk, Hans.** *Technik und Ethik. Stuttgart :* Reclam Verlag, 1993.
7. **Haugeland, J.** *Artificial Intelligence: The Very Idea. s.l. :* MIT Press, 1985.
8. **Turing, A.** *Computing machinery and intelligence. s.l. :* Mind, 1950.
9. **D. Poole, A. K. Mackworth, R. Goebel.** *Computational intelligence: A logic approach.* Oxford : University Press, 1998.
10. **Hochrangige Expertengruppe für Künstliche Intellig.** *Eine Definition der KI: Wichtigste Fähigkeiten und Wissenschaftsgebiete.* Brüssel : Europäische Kommission, 2018.
11. **Nilson, N. J.** *The Quest for Artificial Intelligence.* Cambridge : University Press, 2010.
12. **Google.** *Google AI Blog. An AI System for Accomplishing Real-World Tasks Over the Phone.* [Online] 08. Mai 2018. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html.
13. *Google und die Frau am Herd.* **Wolfangel, Eva.** 29, s.l. : Zeit, 2017.
14. **Whitaker, William.** *William Whitaker's Words.* [Online] University of Notre Dame, South Bend, IN, 2010. [Zitat vom: 27. Mai 2020.] http://archives.nd.edu/cgi-bin/wordz.pl?keyword=automatus.
15. **Harper, Douglas.** *Online Etymology Dictionary.* [Online] MaoningTech, 2018. [Zitat vom: 27. Mai 2020.] https://www.etymonline.com/word/autonomous; https://www.etymonline.com/word/auto-?ref=etymonline_crossreference.
16. **Ethik-Kommission Automat. und Vernetztes Fahren.** *Bericht.* s.l. : Bundesministerium für Verkehr und digitale Infrastruktur, 2017.
17. **Council, National Science and Technology.** *Preparing for the Future of Artificial Intelligence. Washington :* U.S Government, 2016.
18. **O. J. Groth, M. Nitzberg, D. Zehr.** *Vergleich nationaler Strategien zur Förderung von Künstlicher Intelligenz, Teil 1.* Sankt Augustin/Berlin : Konrad Adenauer Stiftung, 2018.
19. **Hochrangige Expertengruppe für Künstliche Intellig.** *Ethik-Leitlinien für eine vertrauenswürdige KI.* Brüssel : Europäische Kommission, 2018.
20. **OECD, Organisation für wirtschaftliche Zusammenarbeit und Entwicklung.** *Artificial Intelligence .* OECD Principles on AI. [Online] https://www.oecd.org/going-digital/ai/principles/.
21. **BVDW, Bundesverband Digitale Wirtschaft.** *Acht Leitlinien für Künstliche Intelligenz.* 2019.
22. **KI Bundesverband e.V.** *KI Gütesiegel. Berlin :* s.n., 2019.
23. *Lexikon der Psychologie .* Heidelberg : Spektrum Akademischer Verlag, 2000.
24. **ISO/IEC JTC 1/SC 42 Artificial intelligence.** ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence.
25. **IMB Corp.** *Everyday Ethics for Artificial Intelligence.* 2019.
26. DIN SPEC 92001-1 *Künstliche Intelligenz - Life Cycle Prozesse und Qualitätsanforderungen - Teil 1:* Qualitäts-Meta-Modell. 2019.
27. **Google.** *Artificial Intelligence at Google: Our Principles.* https://ai.google/principles
28. **Microsoft Corporation.** *The Future Computed. Die gesellschaftliche Bedeutung von Künstlicher Intelligenz* (KI). Washington, 2018.
29. **SAP.** *SAP's Guiding Principles for Artificial Intelligence.* 2018 https://www.sap.com/products/intelligent-technologies/artificial-intelligence/ai-ethics.html?pdf-asset=940c6047-1c7d-0010-87a3-c30de2ffd8ff&page=1.

BEGRIFFSGLOSSAR

BEGRIFF	DEFINITION	QUELLE
Automatisiertes Fahren	Automatisierte Systeme können die Fahrzeugführung in speziellen Situationen, für einen begrenzten Zeitraum komplett übernehmen. Solche Systeme können beispielsweise auf Autobahnen bis zu einer bestimmten oberen Geschwindigkeitsgrenze automatisch den gewünschten Abstand zum Vorderfahrzeug einhalten, dabei gleichzeitig die Spurhaltung kontrollieren und zukünftig auch die Spur wechseln. Weiter entwickelte Systeme, die keine Fahrerin oder keinen Fahrer mehr zu Fahrzeugsteuerung benötigen, werden als autonome (fahrerlose) Systeme bezeichnet.	Bundesministerium für Verkehr und digitale Infrastruktur (BMVI). *„Automatisiertes und vernetztes Fahren“* **VERFÜGBAR ONLINE:**
Automatisierung	Automatisierung ist das Ergebnis des Automatisierens, d.h. des Einsatzes von Automaten (DIN IEC 60050-351). Automaten sind hierbei künstliche Systeme, die selbsttätig ein Programm befolgen und dabei aufgrund des Programms Entscheidungen zur Steuerung und ggf. Regelung von Prozessen treffen (z.B. flexibles FFS). Die Entscheidungen des Systems beruhen auf der Verknüpfung von Eingaben mit den jeweiligen Zuständen eines Systems und haben Aufgaben zur Folge (DIN IEC 60050-351). Automatisch ablaufende Prozesse vollziehen sich vielfach nach dem Regelkreisprinzip - also unter zielorientierter Prozessbeeinflussung durch die Rückkoppelung von Kontrollergebnissen. Je nach Umfang der Übernahme von Steuerungs- und Regelungsaufgaben durch die Maschine wird von Teil- oder Vollautomatisierung gesprochen.	Prof. Dr. Kai-Ingo Voigt. *„Gabler Wirtschaftslexikon-Automatisierung“*, Springer Gabler **VERFÜGBAR ONLINE:**
Automatisierungsgrad /-stufe	Der Anteil, den die automatisierten Funktionen an der Gesamtfunktion eines Produktionssystems haben, wird als Automatisierungsgrad bezeichnet (DIN IEC 60050-351). Bei einer schrittweisen Erhöhung des Einsatzes von Automaten im Produktionsprozess kann von Automatisierungsstufen gesprochen werden.	Prof. Dr. Kai-Ingo Voigt. *„Gabler Wirtschaftslexikon-Automatisierung“*, Springer Gabler **VERFÜGBAR ONLINE:**
Ethik	Ethik ist eine wissenschaftliche Disziplin und ein Teilgebiet der Philosophie. Im Allgemeinen geht es um Fragen wie „Was ist eine gute Tat?“, „Welchen Wert hat das menschliche Leben?“, „Was ist Gerechtigkeit?“ oder „Was ist ein gutes Leben?“. In der wissenschaftlichen Ethik gibt es vier Hauptforschungsgebiete: i) Metaethik: Sie bezieht sich vor allem auf die Bedeutung und den Bezug normativer Sätze und die Frage, wie deren Wahrheitswerte (falls vorhanden) bestimmt werden können. ii) Normative Ethik: Sie beschäftigt sich mit praktischen Mitteln zur Bestimmung einer moralischen Handlungsweise durch Überprüfung der Normen für richtiges und falsches Handeln und Zuweisung eines Wertes zu bestimmten Handlungen. iii) Deskriptive Ethik: Gegenstand ist die empirische Untersuchung des moralischen Verhaltens und der Überzeugungen der Menschen. iv) Angewandte Ethik: Gegenstand ist das Handeln, zu dem die Menschen verpflichtet sind (oder das ihnen erlaubt ist) in einer bestimmten (oft historisch neuen) Situation oder einem bestimmten Kontext von (oft historisch ungekannten) Handlungsmöglichkeiten. Die angewandte Ethik beschäftigt sich mit realen Situationen, in denen Entscheidungen unter Zeitdruck und oftmals mit begrenzter Rationalität getroffen werden müssen. Die KI-Ethik wird im Allgemeinen als ein Beispiel für angewandte Ethik betrachtet. Sie konzentriert sich auf die normativen Fragen, die sich aus dem Entwurf, der Entwicklung, der Umsetzung und Verwendung von KI ergeben.	Hochrangige Expertengruppe für KI (HEG-KI). (2019) *“Ethik-Leitlinien für eine vertrauenswürdige KI”*, Europäischen Kommission **VERFÜGBAR ONLINE:**

BEGRIFF	DEFINITION	QUELLE
Fahrzeug	u. a. mit Rädern, Kufen oder Tragflächen ausgerüstete Konstruktion mit Eigen- oder Fremdantrieb zur Beförderung von Personen und Lasten	Duden **VERFÜGBAR ONLINE:**
Längsführung	bspw. Spurhalten	Bundesministerium für Verkehr und digitale Infrastruktur (BMVI). *„Automatisiertes und vernetztes Fahren"* **VERFÜGBAR ONLINE:**
Maschine	eine mit einem anderen Antriebssystem als der unmittelbar eingesetzten menschlichen oder tierischen Kraft ausgestattete oder dafür vorgesehene Gesamtheit miteinander verbundener Teile oder Vorrichtungen, von denen mindestens eines bzw. eine beweglich ist und die für eine bestimmte Anwendung zusammengefügt sind	Europäische Union. (2006) „Richtlinie 2006/42/EG des europäischen Parlaments **VERFÜGBAR ONLINE:** und des Rates vom 17. Mai 2006 über Maschinen und zur Änderung der Richtlinie 95/16/EG (Neufassung)" **VERFÜGBAR ONLINE:**
Moral	Der Begriff „Moral" bezieht sich auf konkrete, faktische Verhaltensmuster, Bräuche und Sitten, die bei bestimmten Kulturen, Gruppen oder Einzelpersonen zu einem bestimmten Zeitpunkt beobachtet werden können. Der Begriff „ethisch" bezieht sich auf eine Bewertung solcher konkreten Handlungen und Verhaltensweisen aus einer systematischen, wissenschaftlichen Perspektive.	Hochrangige Expertengruppe für KI (HEG-KI). (2019) *"Ethik-Leitlinien für eine vertrauenswürdige KI"*, Europäischen Kommission **VERFÜGBAR ONLINE:**
Querführung	bspw. Bremsen	Bundesministerium für Verkehr und digitale Infrastruktur (BMVI). *„Automatisiertes und vernetztes Fahren"* **VERFÜGBAR ONLINE:**

ANHANG 1 – Übersicht über Veröffentlichungen zum Thema der ethischen Anwendung von KI

Hier ist eine Übersicht von Ethik-Richtlinien für KI gegeben. Sie diente neben den existierenden Normungsdokumenten (siehe Anhang 2) als Grundlage für die Recherche von Attributen, für eine vertrauenswürdige KI. Zum Zeitpunkt der Veröffentlichung gibt es bereits einige weitere relevante Dokumente in diesem Themenfeld. Da die Ergebnisse der Recherche allerdings in die Diskussion mit Expertinnen und Experten eingeflossen sind, konnten nur Dokumente, die bis Mitte 2019 veröffentlicht waren berücksichtigt werden. Die Vollständigkeit der Liste kann nicht gewährleistet werden. Es wurde jedoch angestrebt Dokument aus nationalen, europäischen und internationalen Gremien, Organisationen und Unternehmen zu berücksichtigen.

- **AI4People. AI4People's Ethical Framework for a Good AI Society:** *Opportunities, Risks, Principles, and Recommendations,* 2018.
- **Bundesregierung.** *Strategie Künstliche Intelligenz der Bundesregierung*. 2018
- **BVDW, Bundesverband Digitale Wirtschaft.** *Acht Leitlinien für Künstliche Intelligenz.* 2019.
- **Ethik-Kommission Automat. und Vernetztes Fahren.** *Bericht. s.l. : Bundesministerium für Verkehr und digitale Infrastruktur,* 2017.
- **Europäische Gruppe für Ethik der Naturwissenschaften und der Neuen Technologien.** *Erklärung zu künstlicher Intelligenz, Robotik und „autonomen" Systemes.* Brüssel: Europäische Kommission, 2018.
- **Google.** *Artificial Intelligence at Google: Our Principles.* https://ai.google/principles
- **Hochrangige Expertengruppe für Künstliche Intellig.** *Ethik-Leitlinien für eine vertrauenswürdige KI. Brüssel :* Europäische Kommission, 2018.
- **IBM Corp.** *Everyday Ethics for Artificial Intelligence.* 2019.
- **KI Bundesverband e.V.** *KI Gütesiegel.* Berlin : s.n., 2019.
- **Microsoft Corporation.** *The Future Computed. Die gesellschaftliche Bedeutung von Künstlicher Intelligenz (KI).* Washington, 2018.
- **OECD, Organisation für wirtschaftliche Zusammenarbeit und Entwicklung**. A*rtificial Intelligence . OECD Principles on AI.* [Online] https://www.oecd.org/going-digital/ai/principles/.
- **SAP.** *SAP's Guiding Principles for Artificial Intelligence.* 2018 https://www.sap.com/products/intelligent-technologies/artificial-intelligence/ai-ethics.html?pdf-asset=940c6047-1c7d-0010-87a3-c30de2ffd8ff&page=1.
- **Verband der Automobilindustrie (VDA).** *Automatisierung – Von Fahrerassistenzsystemen zum automatisierten Fahren,* 2015.

ANHANG 2 – Übersicht über Normungs- und Standardisierungsaktivitäten

Hier sind Normungs- und Standardisierungsdokumente aufgelistet, die zum Thema Künstliche Intelligenz derzeit in der Erarbeitung oder bereits erschienen sind. Nicht alle Dokumente fokussieren sich auf das ethische Verhalten von KI. Auf diese Dokumente wird gesondert im Kapitel „Normen, Standards, Konsortialstandards oder Aktivitäten im Bereich der technischen Regelsetzung" eingegangen.

Tabelle 2 Übersicht über aktuelle Normungs- und Standardisierungsaktivitäten im Bereich KI

TITEL	ANWENDUNGSBEREICH	AUTOR*INNEN	ERSCHIENEN
ISO/IEC 22989 Artificial intelligence — Concepts and terminology	This document establishes terminology for Artificial Intelligence (AI) and describes concepts in the field of AI. This document can be used in the development of other standards and in support of communications among diverse, interested parties/stakeholders. This document is applicable to all types of organizations (e.g., commercial enterprises, government agencies, not-for-profit organizations).	ISO/IEC JTC 1/ SC 42 „Artificial Intelligence"	In Erarbeitung
ISO/IEC 23894 Information Technology — Artificial Intelligence — Risk Management	This document provides guidelines on managing risk faced by organizations during the development and application of Artificial Intelligence (AI) techniques and systems, to assist organizations to integrating risk management for AI into significant activities and functions. It moreover describes processed for the effective implementation and integration of AI risk management. The application of these guidelines can be customized to any organization and its context. This document uses the guidelines described in the International Standard ISO 31000 (Risk management – Guidelines) and in addition provides additional guidance that arises by the application of AI to existing processes in any organization or when an organization provides an AI system for use by others	ISO/IEC JTC 1/ SC 42 „Artificial Intelligence"	In Erarbeitung
ISO/IEC AWI 38507 Information Technology — Governance of IT — Governance implications of the use of Artificial Intelligence by organizations	This document provides guidance for members of the governing bodies of organizations on the effective, efficient, and acceptable uses of artificial intelligence within their organizations. This document also provides guidance to a wider community, including: ● executive managers; ● external businesses or technical specialists, such as legal or accounting specialists, retail or industrial associations, or professional bodies; ● public authorities and policy-makers; ● internal and external service providers (including consultants); ● auditors. This document is applicable to the governance of current and future uses of artificial intelligence as well as the implications of such use for the organization itself.	ISO/IEC JTC 1/ SC 42 „Artificial Intelligence"	In Erarbeitung

TITEL	ANWENDUNGSBEREICH	AUTOR*INNEN	ERSCHIENEN
	This document is applicable to all organizations, including public and private companies, government entities, and not-for-profit organizations. This document is applicable to organizations of all sizes from the smallest to the largest, regardless of the extent of their dependence on data or information technologies.		
ISO/IEC TR 24027 Information Technology — Artificial Intelligence (AI) — Bias in AI systems and AI aided decision making	This document addresses bias in relation to AI systems, especially with regards to AI-aided decision making. Measurement techniques and methods for assessing bias are described, with the aim to address bias related vulnerabilities, and mitigation thereof. All AI system lifecycle phases are in scope, including but not limited to data collection, training, continual learning, design, testing, evaluation, and use.	ISO/IEC JTC 1/ SC 42 „Artificial Intelligence“	In Erarbeitung
ISO/IEC TR 24028 Information Technology — Artificial Intelligence (AI) — Overview of trustworthiness in Artificial Intelligence	This document surveys topics related to trustworthiness in AI systems, including the following: • approaches to establish trust in AI systems through transparency, explainability, controllability,etc.; • engineering pitfalls and typical associated threats and risks to AI systems, along with possible mitigation techniques and methods; and • approaches to assess and achieve availability, resiliency, reliability, accuracy, safety, security, and privacy of AI systems. Out of scope is the following: • specification of levels of trustworthiness for AI systems.	ISO/IEC JTC 1/ SC 42 „Artificial Intelligence“	2020
ISO/IEC TR 24029-1 Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview	The present document provides background about the existing methods to assess the robustness of neural networks.	ISO/IEC JTC 1/ SC 42 „Artificial Intelligence“	2020
ISO/IEC 24029-2 Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Formal methods methodology	This document provides methodology on the use of formal methods to assess robustness properties of neural networks. The document focuses on how to manage and put in place formal methods to prove robustness properties.	ISO/IEC JTC 1/ SC 42 „Artificial Intelligence“	In Erarbeitung
ISO/IEC TR 24368 Information technology — Artificial intelligence — Overview of ethical and societal concerns	This document provides a high-level overview of the programme of work in SC 42 in the area of ethics and societal concerns relative to Artificial Intelligence (AI) systems and applications. This document provides information in relation to principles, processes and methods in this area. This document is intended for technologists, regulators, interest groups, and society at large. This document is not intended to advocate for any specific set of values (value systems).	ISO/IEC JTC 1/ SC 42 „Artificial Intelligence“	In Erarbeitung
April 2020 genehmigt: **ISO/IEC TR** Artificial intelligence — Functional Safety and AI systems	The document describes the properties, related risk factors, available methods and processes relating to: l Use of AI inside a safety related function to realise the functionality l Use of non-AI safety related functions to ensure safety for an AI controlled equipment l Use of AI systems to design and develop safety related functions.	ISO/IEC JTC 1/ SC 42 „Artificial Intelligence“	In Erarbeitung

TITEL	ANWENDUNGSBEREICH	AUTOR*INNEN	ERSCHIENEN
ISO/AWI 39003 Road Traffic Safety (RTS) — Guidance on safety ethical considerations for autonomous vehicles	This standard will give guidelines for manufacturers of Level 5 Autonomous Vehicles, as defined by the international Society of Automotive Engineers (SAE) in 2014, as part of its „Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems" report. The guidelines will define the ethical considerations and prioritizations that a level 5 Autonomous Vehicle should have when making essential safe driving decisions. It is intended that manufacturers of such vehicles will self-certify their vehicles as being compliant to this standard before they bring the vehicle model to market. However, the panel will take into account the 'neutrality principle' defined in 33.1 of the ISO/IEC Directives, Part 2 when drafting the document.	ISO/TC 241 „Road traffic safety management systems"	In Erarbeitung
DIN SAE SPEC 91381:2019-06 Begriffe und Definitionen in Bezug auf die Prüfung automatisierter Fahrzeugtechnologien	Diese zweisprachige DIN-SAE SPEC definiert Begriffe für den Bereich der automatisierten Fahrzeugtechnologien, insbesondere Begriffe zu Simulationen und Testumgebungen dieser Technologien. Die DIN-SAE SPEC soll als Werkzeug für die nächsten Entwicklungsschritte und Forschungsaktivitäten sowie zur besseren Kommunikation internationaler Partner dienen. Diese DIN-SAE SPEC definiert keine Begriffe zu den Levels der Automatisierung und Fahrzeugparametern. Der Standard soll einen Begriffskatalog zur Vereinheitlichung der Sprache in einem komplexen, interdisziplinären Umfeld darstellen. Diese DIN-SAE SPEC (PAS) richtet sich an Forscher, Entwickler, Softwareentwickler, Teststreckenbetreiber, Prüfeinrichtungen sowie an Hersteller automatisierter Fahrzeuge.	DIN SAE SPEC 91381 Workshop	2019
DIN SPEC 92001-1:2019-04 Künstliche Intelligenz - Life Cycle Prozesse und Qualitätsanforderungen - Teil 1: Qualitäts-Meta-Modell	Das Dokument bietet ein allgemeines Qualitätsmetamodell für Künstliche Intelligenz (KI), das in erster Linie die wichtigsten Aspekte der KI-Qualität beschreibt. Das KI-Qualitätsmetamodell enthält unter anderem die drei wesentlichen Qualitätsmerkmale - Leistung & Funktionalität, Robustheit und Verständlichkeit. Dieses Dokument befasst sich auch mit dem KI-Modul als Teil eines Softwaresystems, seiner Risikobewertung und einem geeigneten Software-Lebenszyklusansatz. Der vorliegende KI-Lebenszyklusansatz stützt sich stark auf die Internationale Norm für System- und Softwareentwicklung ISO/ IEC/IEEE 12207:2017. Ziel dieser DIN SPEC-Reihe ist es, eine sichere und transparente Entwicklung und den Einsatz von KI-Modulen zu ermöglichen. Zu diesem Zweck beschreibt die DIN SPEC eine Reihe von Qualitätsanforderungen, die durch ein KI-Qualitätsmetamodell strukturiert sind. Die DIN SPEC-Reihe gilt für alle Lebenszyklusphasen - Konzeptionierung, Entwicklung, Einsatz, Betrieb und Stilllegung - eines KI-Moduls und berücksichtigt eine Vielzahl unterschiedlicher Lebenszyklusprozesse. Da KI-Technologien für die unterschiedlichsten Aufgaben eingesetzt werden, richtet sich diese DIN SPEC-Reihe an Unternehmen aller Branchen.	DIN SPEC 92001 Workshop	2019

TITEL	ANWENDUNGSBEREICH	AUTOR*INNEN	ERSCHIENEN
DIN SPEC 92001-2 Künstliche Intelligenz - Life Cycle Prozesse und Qualitätsanforderungen - Teil 2: Robustheit	Teil 2 der DIN SPEC 92001-Reihe stellt KI-spezifische Robustheitsanforderungen dar. Diese Qualitätsanforderungen werden mit Hilfe des vorgegebenen KI-Qualitätsmetamodells (DIN SPEC 92001-1) strukturiert.	DIN SPEC 92001 Workshop	In Erarbeitung
DIN 66398 Leitlinie zur Entwicklung eines Löschkonzepts mit Ableitung von Löschfristen für personenbezogene Daten	DIN 66398 gibt Empfehlungen für die Inhalte, den Aufbau und die Zuordnung von Verantwortung in einem Löschkonzept für personenbezogene Daten. Das Dokument beschreibt insbesondere Vorgehensweisen, mit denen auf effiziente Weise Löschfristen und Löschregeln für verschiedene Datenarten bestimmt werden können.	NA 043-01-27-05 AK Identitätsmanagement und Datenschutz-Technologien	2016
ISO/IEC 38505-1:2017: „Information technology – Governance of IT – Part 1: Application of ISO/IEC 38500 to the governance of data	This document provides guiding principles for members of governing bodies of organizations (which can comprise owners, directors, partners, executive managers, or similar) on the effective, efficient, and acceptable use of data within their organizations by • applying the governance principles and model of ISO/IEC 38500 to the governance of data, • assuring stakeholders that, if the principles and practices proposed by this document are followed, they can have confidence in the organization's governance of data, • informing and guiding governing bodies in the use and protection of data in their organization, and • establishing a vocabulary for the governance of data. This document can also provide guidance to a wider community, including: • executive managers, • external businesses or technical specialists, such as legal or accounting specialists, retail or industrial associations, or professional bodies, • internal and external service providers (including consultants), and • auditors. While this document looks at the governance of data and its use within an organization, guidance on the implementation arrangement for the effective governance of IT in general is found in ISO/IEC/TS 38501. The constructs in ISO/IEC/TS 38501 can help to identify internal and external factors relating to the governance of IT and help to define beneficial outcomes and identify evidence of success. This document applies to the governance of the current and future use of data that is created, collected, stored or controlled by IT systems, and impacts the management processes and decisions relating to data. This document defines the governance of data as a subset or domain of the governance of IT, which itself is a subset or domain of organizational, or in the case of a corporation, corporate governance. This document is applicable to all organizations, including public and private companies, government entities, and not-for-profit organizations. This document is applicable to organizations of all sizes from the smallest to the largest, regardless of the extent of their dependence on data.	ISO/IEC JTC 1/SC 40 IT "Service Management and IT Governance"	2017

TITEL	ANWENDUNGSBEREICH	AUTOR*INNEN	ERSCHIENEN
ISO/IEC TR 29110-1:2016	ISO/IEC TR 29110-1:2016 introduces the major concepts required to understand and use the ISO/IEC 29110 series. It introduces the characteristics and requirements of a VSE and clarifies the rationale for VSE-specific profiles, documents, standards and guides. It also introduces process, lifecycle, standardization concepts and defines the organizational terms common to the VSE Profile Set of Documents. It is applicable to a VSE. A VSE is an entity (enterprise, organization, department or project) having up to 25 people. The lifecycle processes described in the ISO/IEC 29110 series, Standardized Profiles and Technical Reports are not intended to preclude nor discourage their use by an entity that is larger than a VSE. It is targeted both at the general audience wishing to understand the series of documents and, more specifically, at users of the ISO/IEC 29110 series. It should be read first when initially exploring VSE Profile documents. While there is no specific prerequisite to read this part of ISO/IEC 29110, it will be helpful to the user in understanding the other parts. The lifecycle processes defined in the set of Standardized Profiles and Technical Reports can be used by a VSE when developing, acquiring and using, as well as when creating and supplying systems having hardware and software elements and software. They can be applied at any level in a systems development, software system's structure and at any stage in the lifecycle. They are not intended to preclude or discourage the use of additional processes that a VSE finds useful.	ISO/IEC JTC 1/ SC 7 „Software and systems engineering“	2016
ISO/TS 17033:2019 Ethical claims and supporting information — Principles and requirements	This document contains principles and requirements for developing and declaring ethical claims and for providing supporting information, where specific standards have not been developed, or to supplement existing standards. This document is intended for use by all types of organizations and is applicable to all types of ethical claims relating to a product, process, service or organization. This document can also be used by those seeking a better understanding of ethical claims and their use. This document can support the development of programmes for aspect-specific and sector-specific ethical claims.	ISO/CASCO „Committee on conformity assessment“	2019
DIN ISO 26000:2011-01 Leitfaden zur gesellschaftlichen Verantwortung	Die Norm ISO 26000 ist als Leitfaden angelegt, der strategische Planung und Umsetzung von gesellschaftlicher Verantwortung im weitesten Sinne erleichtern soll. Der aktuelle Norm-Entwurf weist ausdrücklich darauf hin, dass die ISO 26000 keine Managementsystemnorm (wie zum Beispiel ISO 14001) ist und weder für Zertifizierungszwecke noch für gesetzliche oder vertragliche Anwendungen vorgesehen und auch nicht geeignet ist. Jegliche Angebote zur Zertifizierung oder die Behauptung, gemäß ISO 26000 zertifiziert zu sein, widersprechen der Absicht und dem Zweck dieser Internationalen Norm.	NA 095-04-01 AA „Gesellschaftliche Verantwortung von Organisationen“	2011

TITEL	ANWENDUNGSBEREICH	AUTOR*INNEN	ERSCHIENEN
	Ziel von ISO 26000 ist es vielmehr, denjenigen Organisationen Orientierung und Anleitung zu vermitteln, die sich auf der Höhe der internationalen Diskussion mit den Prinzipien, Praktiken, Kernthemen und Handlungsfeldern gesellschaftlicher Verantwortung ernsthaft auseinandersetzen und ihre Organisation durchgängig und kontinuierlich danach ausrichten wollen. Zentraler Ansatzpunkt der Wahrnehmung gesellschaftlicher Verantwortung nach ISO 26000 sind neben den Führungs- und Steuerungsmechanismen die gelebten Werthaltungen, Denkmuster, Verhaltensweisen und Praktiken der Organisation beziehungsweise der Organisationsmitglieder. Ziel dieser Internationalen Norm ist es, allen Arten von Organisationen, sowohl in der Privatwirtschaft, als auch im öffentlichen oder im gemeinnützigen Sektor von Nutzen zu sein, unabhängig von ihrer Größe und ihren Tätigkeiten in entwickelten oder sich entwickelnden Teilen der Welt. Mit diesem Orientierungsrahmen soll die Norm Organisationen unterstützen, einen Beitrag zur nachhaltigen Entwicklung zu leisten.		
ISO 26262-6:2018 Road vehicles — Functional safety — Part 6: Product development at the software level	ISO 26262 is intended to be applied to safety-related systems that include one or more electrical and/or electronic (E/E) systems and that are installed in series production passenger cars with a maximum gross vehicle mass up to 3 500 kg. ISO 26262 does not address unique E/E systems in special purpose vehicles such as vehicles designed for drivers with disabilities. Systems and their components released for production, or systems and their components already under development prior to the publication date of ISO 26262, are exempted from the scope. For further development or alterations based on systems and their components released for production prior to the publication of ISO 26262, only the modifications will be developed in accordance with ISO 26262. ISO 26262 addresses possible hazards caused by malfunctioning behaviour of E/E safety-related systems, including interaction of these systems. It does not address hazards related to electric shock, fire, smoke, heat, radiation, toxicity, flammability, reactivity, corrosion, release of energy and similar hazards, unless directly caused by malfunctioning behaviour of E/E safety-related systems. ISO 26262 does not address the nominal performance of E/E systems, even if dedicated functional performance standards exist for these systems (e.g. active and passive safety systems, brake systems, Adaptive Cruise Control). ISO 26262-6:2011 specifies the requirements for product development at the software level for automotive applications, including the following: ● requirements for initiation of product development at the software level, ● specification of the software safety requirements, ● software architectural design, ● software unit design and implementation, ● software unit testing, ● software integration and testing, and ● verification of software safety requirements.	ISO/TC 22/SC 32 „Electrical and electronic components and general system aspects“	2018

TITEL	ANWENDUNGSBEREICH	AUTOR*INNEN	ERSCHIENEN
IEEE P7000™ – Model Process for Addressing Ethical Concerns During System Design	outlines an approach for identifying and analyzing potential ethical issues in a system or software program from the onset of the effort. The values-based system design methods addresses ethical considerations at each stage of development to help avoid negative unintended consequences while increasing innovation.	IEEE	In Erarbeitung
IEEE P7001™ – Transparency of Autonomous Systems	provides a Standard for developing autonomous technologies that can assess their own actions and help users understand why a technology makes certain decisions in different situations. The project also offers ways to provide transparency and accountability for a system to help guide and improve it, such as incorporating an event data recorder in a self-driving car or accessing data from a device's sensors.	IEEE	In Erarbeitung
I**EEE P7002™** – Data Privacy Process	specifies how to manage privacy issues for systems or software that collect personal data. It will do so by defining requirements that cover corporate data collection policies and quality assurance. It also includes a use case and data model for organizations developing applications involving personal information. The standard will help designers by providing ways to identify and measure privacy controls in their systems utilizing privacy impact assessments.	IEEE	In Erarbeitung
IEEE P7003™ – Algorithmic Bias Considerations	provides developers of algorithms for autonomous or intelligent systems with protocols to avoid negative bias in their code. Bias could include the use of subjective or incorrect interpretations of data like mistaking correlation with causation. The project offers specific steps to take for eliminating issues of negative bias in the creation of algorithms. The standard will also include benchmarking procedures and criteria for selecting validation data sets, establishing and communicating the application boundaries for which the algorithm has been designed, and guarding against unintended consequences.	IEEE	In Erarbeitung
IEEE P7006™ – Standard on Personal Data AI Agent Working Group	addresses concerns raised about machines making decisions without human input. This standard hopes to educate government and industry on why it is best to put mechanisms into place to enable the design of systems that will mitigate the ethical concerns when AI systems can organize and share personal information on their own. Designed as a tool to allow any individual to essentially create their own personal "terms and conditions" for their data, the AI Agent will provide a technological tool for individuals to manage and control their identity in the digital and virtual world.	IEEE	In Erarbeitung
IEEE P7007™ – Ontological Standard for Ethically driven Robotics and Automation Systems	establishes a set of ontologies with different abstraction levels that contain concepts, definitions and axioms that are necessary to establish ethically driven methodologies for the design of Robots and Automation Systems.	IEEE	In Erarbeitung
IEEE P7008™ – Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems	establishes a delineation of typical nudges (currently in use or that could be created) that contains concepts, functions and benefits necessary to establish and ensure ethically driven methodologies for the design of the robotic, intelligent and autonomous systems that incorporate them. "Nudges" as exhibited by robotic, intelligent or autonomous systems are defined as overt or hidden suggestions or manipulations designed to influence the behavior or emotions of a user.	IEEE	In Erarbeitung

TITEL	ANWENDUNGSBEREICH	AUTOR*INNEN	ERSCHIENEN
IEEE P7009™ – Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems	establishes a practical, technical baseline of specific methodologies and tools for the development, implementation, and use of effective fail-safe mechanisms in autonomous and semi-autonomous systems. The standard includes (but is not limited to): clear procedures for measuring, testing, and certifying a system's ability to fail safely on a scale from weak to strong, and instructions for improvement in the case of unsatisfactory performance. The standard serves as the basis for developers, as well as users and regulators, to design fail-safe mechanisms in a robust, transparent, and accountable manner.	IEEE	In Erarbeitung
IEEE P7010™ – Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems	will establish wellbeing metrics relating to human factors directly affected by intelligent and autonomous systems and establish a baseline for the types of objective and subjective data these systems should analyze and include (in their programming and functioning) to proactively increase human wellbeing.	IEEE	In Erarbeitung
IEEE P7011™ – Standard for the Process of Identifying & Rating the Trust-worthiness of News Sources	will address the negative impacts of the unchecked proliferation of fake news by providing an open system of easy-to-understand ratings. In so doing, it shall assist in the restoration of trust in some purveyors, appropriately discredit other purveyors, provide a disincentive for the publication of fake news, and promote a path of improvement for purveyors wishing to do so. The standards shall target a representative sample set of news stories in order to provide a meaningful and accurate rating scorecard.	IEEE	In Erarbeitung
IEEE P7012™ – Standard for Machine Readable Personal Privacy Terms	will provide individuals with means to proffer their own terms respecting personal privacy, in ways that can be read, acknowledged and agreed to by machines operated by others in the networked world. In a more formal sense, the purpose of the standard is to enable individuals to operate as first parties in agreements with others–mostly companies–operating as second parties. Note that the purpose of this standard is not to address privacy policies, since these are one-sided and need no agreement. (Terms require agreement; privacy policies do not.)	IEEE	In Erarbeitung
IEEE P7014™ – Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems	defines a model for ethical considerations and practices in the design, creation and use of empathic technology, incorporating systems that have the capacity to identify, quantify, respond to, or simulate affective states, such as emotions and cognitive states. This includes coverage of 'affective computing', 'emotion Artificial Intelligence' and related fields.	IEEE	In Erarbeitung
VDE-AR-E 2842-61 – Reihe Entwicklung und Vertrauenswürdigkeit von autonom/kognitiven Systemen	Die VDE-Anwendungsregel wird verschiedene Entwicklungsschritte anhand des Lebenszyklus für KI-Systeme definieren. Hierzu gehören der Entwurf auf Systemebene (u. a. Vertrauenswürdigkeitsattribute), Komponentenebene (u. a. Hardware, Software und KI-Blaupausen zur Anwendung einer KI-Methodik), Integration, Verifikation und Validierung sowie Abnahme und Freigabe. Das Modell umfasst darüber hinaus auch Vorgaben zur Marktbeobachtung und korrigierende sowie schützende Maßnahmen (eng. Corrective And Preventive Action - CAPA).	DKE/AK801.0.8	Teile 1&2 veröffentlicht am 07/2020; weitere Teile 3-7 in Erarbeitung

DIN e. V.

Saatwinkler Damm 42/43
10787 Berlin
Telefon: +49 30 2601-0
E-Mail: presse@din.de
Internet: www.din.de

DKE Deutsche Kommission Elektrotechnik Elektronik Informationstechnik in DIN und VDE

Stresemannallee 15
60596 Frankfurt
Telefon: +49 69 6308-0
Telefax: +49 69 08-9863
E-Mail: standardisierung@vde.com
Internet: www.dke.de

Photonachweise:

LIGHTFIELD STUDIOS/stock.adobe.com (Titelseite/Rückseite), DKosig/istockphoto.com (S.4-5), zinkevych/stock.adobe.com (S.7), fizkes/stock.adobe.com (S.11), agnormark/Fotolia (S.33), weerapat1003/Fotolia (S.33), willyam/stock.adobe.com (S.49)