TÜV AI.LAB

TÜV AI.AM

TÜV AI Assessment Matrix

A Systematic Approach to AI Assessment

Executive Summary

In der vierten industriellen Revolution wirkt Künstliche Intelligenz (KI) als Schlüsseltechnologie. Um die enormen Potentiale von KI für Wirtschaft und Gesellschaft sicher und vollumfänglich zu erschließen, müssen bestehende Risiken frühzeitig adressiert werden. Konformität mit den regulatorischen Anforderungen stellt den Nutzen der Technologie im Dienste der Menschen sicher und mitigiert zugleich Reputations- und Haftungsrisiken für Unternehmen. Das vorliegende Dokument stellt die TÜV AI Assessment Matrix (Version 0.1) vor, ein **neuartiges Framework, das auf den wachsenden Bedarf an strukturierten KI-Konformitätsprüfungen antwortet**. Diese Prüfungen schaffen langfristig Vertrauen in KI-Systeme und beschleunigen die Adaption und Skalierung der KI-Technologien.

Die **AI Assessment Matrix** des TÜV AI.Lab nimmt sich der Herausforderung an, das unübersichtliche Feld der KI-Prüfung systematisch zu ordnen. Sie bietet erstmalig:

1) eine **systematisch umfassende Anordnung** technischer und ethischer Prüfdimensionen
2) ein **vollständiges und kohärentes Set an Definitionen**, welches den sachlichen Kern der jeweiligen Prüfdimension explizit macht und so Unklarheit mindert sowie scheinbaren Konvergenzen auf Wortebene vorbeugt
3) eine Rahmenstruktur für die **Ordnung von Prüfressourcen, Prüfansätzen und Prüfanforderungen**
4) eine Grundlage für den **Vergleich verschiedener Prüfansätze** sowie deren Verhältnis zu gegebenen regulatorischen Rahmenwerken
5) eine Basis für den **Vergleich verschiedener regulatorischer Rahmenwerke für KI**

Eine solche **umfassende systematische Ordnung des Gesamtfeldes ‚KI-Prüfung'** fehlte bislang. Die AI Assessment Matrix schafft hier erstmals Abhilfe. Sie schließt die bestehende strukturelle Lücke inhaltlich variierender Einzelvorschläge, indem sie als **Metastruktur** Prüfressourcen,

Prüfansätze und Prüfanforderungen in einem **systematischen Gesamtzusammenhang** ordnet. Zu diesem Zweck sind im **begleitenden Poster** die Vorgaben der europäischen KI-Verordnung, soweit sie die Prüfung des KI-Systems selbst betreffen, exemplarisch eingetragen.

Mit der Bereitstellung eines einheitlichen und umfassenden Rahmens für KI-Prüfungen leistet die AI Assessment Matrix des TÜV AI.Lab einen entscheidenden Beitrag zur Community im Bereich der KI-Prüfung und -Zertifizierung. Sie verbessert die Vergleichbarkeit und Konsistenz von Prüfungen, unterstützt die Einhaltung regulatorischer Anforderungen und schafft eine Grundlage für **nachhaltige und vertrauenswürdige Innovationen** im Bereich der KI-Technologien.

Die AI Assessment Matrix spannt bewusst ein **Maximalfeld** auf. Damit ist keineswegs die Aussage impliziert, dass eine akzeptable KI-Qualität erst durch die Erfüllung aller aufgeführter Prüfdimensionen erreicht ist; vielmehr erlaubt erst dieses Maximalfeld die bewusste und erkennbare **Eingrenzung** auf bestimmte Prüfdimensionen. Im Rahmen der Entwicklung des **TÜV AI Assessment Framework** wird zudem ein Vorschlag zur vertikalen Konkretisierung erarbeitet, so dass die allgemeine Struktur der Matrix **auf spezifische Prüfszenarien und konkrete Use Cases anwendbar** wird. Dies schließt den Gap zwischen der allgemeinen Ebene der technologieagnostischen Anforderungen und der konkreten Prüfung. Effiziente und effektive KI-Prüfungen werden so praxisnah unterstützt.

Einleitung

Industrielle Revolutionen werden von Schlüsseltechnologien getragen. Was in der ersten industriellen Revolution, im späten 18. und im 19. Jahrhundert, die Dampfkessel waren, sind im 21. Jahrhundert, in der Folge von Internet und Digitalisierung, die Technologien der ‚Künstlichen Intelligenz' (KI). Heute wie damals gilt es, die immensen Potentiale der Technologie zu erschließen, zugleich aber Mensch, Gesellschaft und Umwelt vor den damit einhergehenden Risiken zu schützen. Hervorgegangen aus den ersten Dampfkesselüberwachungsvereinen, widmen sich die TÜV-Unternehmen bis heute der Sicherheit der wichtigsten Technologien in Gegenwart und Zukunft – und setzen sich deshalb mit Nachdruck für eine sichere KI ein, die unser Vertrauen verdient. Ein langfristiger Erfolg der KI-Technologie(n) im Dienste der Menschen ist nur möglich, wenn es uns in Europa – und weltweit – gelingt, durch gesetzliche Regulatorik und gezielte, unabhängige Prüfungen von KI-Systemen Vertrauen in diese neue Technik zu schaffen, ohne dabei den für Innovationen notwendigen Freiraum mehr als notwendig zu begrenzen. Eine wesentliche Voraussetzung für ‚Vertrauenswürdige KI' sind damit effiziente und effektive Prüfungen von KI-Systemen.
Die umfassende Prüfung von KI-Systemen ist dabei eine äußerst komplexe Aufgabe. Sie bewegt sich in unwegsamem Terrain, das sich aus einer Vielzahl an möglichen Prüfdimensionen, verschiedenen gesetzlichen und regulatorischen Vorgaben, diversen Normen und Standards sowie unterschiedlich

spezifizierten Prüfansätzen zusammensetzt. Das vorliegende Dokument stellt die TÜV AI.Lab **AI Assessment Matrix** in Version 0.1 vor: Die Matrix ist darauf angelegt, dieses schwer überschaubare Feld der Prüfdimensionen systematisch zu ordnen, Prüfansätze und -anforderungen vergleichbar zu machen sowie darüber hinaus technische und ethische Prüfdimensionen in einem gemeinsamen Rahmen zu verorten.
Im Kern stellt die nachfolgend skizzierte AI Assessment Matrix eine **Organisationsstruktur für KI-Prüfmethoden, -verfahren und -werkzeuge** dar, die das Feld ‚KI-Prüfung' systematisch strukturiert und es beispielweise erlaubt, die Anwendungsdimensionen international divergierender Rechtsakte zu vergleichen oder Prüfansätze unterschiedlicher Frameworks gegenüberzustellen. Sie ist in dieser Form ein zentraler Bestandteil des **AI Assessment Frameworks**, welches das TÜV AI.Lab derzeit für die TÜV-Unternehmen entwickelt. Begleitet wird das vorliegende Dokument von einer graphischen Darstellung der AI Assessment Matrix, inklusive der regulatorischen Vorgaben durch die europäische KI-Verordnung, sofern diese durch die Prüfdimensionen adressiert sind.[1]

Die TÜV AI Assessment Matrix. Grundansatz und Kerngedanken

Eine konkrete Prüfung von KI-Systemen setzt eine **Prüfpipeline**, also eine prototypische Abfolge von Prüfschritten, **und Prüfressourcen** – beispielsweise Testverfahren, Metriken, Checklisten und dergleichen – voraus. Damit nicht für jede konkrete Prüfung eines KI-Systems ein vollständig neues Prüfverfahren aufgesetzt werden muss, bedarf es eines strukturierten Überblicks über Prüfressourcen, der diese Ressourcen ordnet und sektor- sowie technologieübergreifend Synergieffekte ermöglicht. Eine solche **Ordnungstruktur** bietet die vorgestellte AI Assessment Matrix. Sie stellt also, um Missverständnissen vorzubeugen, keinen einzelnen, spezifischen Prüfprozess vor. Sie bietet vielmehr die Basis für einen **systematischen Ressourcenpool**, auf den einzelne, konkrete Prüfpipelines in der Anwendung gezielt zurückgreifen können und ist kompatibel mit diesen. Die Ausgestaltung der konkreten Prüfpipeline erfolgt in anderen Teilen des TÜV AI Assessment Frameworks.
Die AI Assessment Matrix ist folglich eine **mehrdimensionale Ordnungsstruktur für Prüfressourcen zur KI-Prüfung**. Ihre einzelnen Felder beinhalten systematisch alle Formen von Daten, die potenziell als Ressourcen zur Prüfung von KI-Systemen einsetzbar sind: Diese Daten reichen von der detaillierten Gliederung eines Prüfvorgangs, über Entitäten wie Benchmark-Datensätze und Prüflisten bis hin zu Grenzwerten und Ähnlichem. Anschaulich gesprochen, bildet die AI Assessment Matrix ein groß angelegtes Regalsystem, dessen Grundaufbau im vorliegenden Dokument beschrieben wird. Die einzelnen Felder der Matrix entsprechen in diesem Bild einzelnen Regalfächern, innerhalb derer die je konkreten Prüfressourcen einsortiert sind. In dieser Form ist die AI Assessment Matrix

[1] Die Prüfung – nicht die Ausformulierung – von Qualitätsmanagementsystemen unterliegt in manchen Hinsichten einer anderen Systematik; sie werden deshalb nur zum Teil durch die im Vorliegenden beschrieben AI Assessment Matrix adressiert, insgesamt aber im AI Assessment Framework des TÜV AI.Lab vollständig aufgegriffen.

eine Systematik zur Strukturierung des Gesamtfeldes ‚KI-Prüfung' (und somit Grundlage für die KI-Zertifizierung), die zugleich konkret implementierbar ist.
Die AI Assessment Matrix beruht auf **zwei Grundgedanken**. Zum einen spannt sie über die systematische Ordnung ihrer Dimensionen ein **kombinatorisches Maximalfeld** zur Evaluierung von KI-Systemen über deren gesamten Life Cycle hinweg auf; sie bietet also eine möglichst umfassende Aufstellung von Prüfdimensionen, die sich mit unterschiedlicher Fokussierung auf KI-Systeme und deren Einsatz anwenden lassen. Hierbei erlaubt die systematische Anlage der AI Assessment Matrix bei Bedarf zudem auch die flexible, passgenaue Integration weiterer Dimensionen. Die Verschränkung der drei Achsen der Matrix bildet eine mehrdimensionale Kombinatorik, ohne damit jedoch zu implizieren, dass jeder der zahlreichen Einzelkombinationen gleiches Gewicht zukäme oder restlos jede Kombination sachlich ausformulierbar wäre. Während die Matrix auf ihrer allgemeinsten Ebene in den angeführten Prüfdimensionen per se technologie- und anwendungsagnostisch ist, implementiert ein Filiationsprinzip die technologie- und anwendungsspezifische Spezifizierung der Prüfdimensionen bis hin zur Prüfung konkreter KI-Systeme, unter gleichzeitiger Wahrung größtmöglicher Synergieeffekte zwischen den einzelnen Prüfungen und Ressourcen.
Zum anderen bietet der Aufbau der AI Assessment Matrix darüber hinaus als **Metastruktur** die Möglichkeit, einzelne Prüfframeworks und -ansätze vergleichbar zu machen und mit regulatorischen Vorgaben – etwa der europäischen KI-Verordnung – abzugleichen. Zu diesem Zweck sind in der begleitenden Darstellung der Matrix die Vorgaben der KI-Verordnung, soweit sie die Prüfung des KI-Systems selbst betreffen, exemplarisch eingetragen.

X-Achse. Prüfdimensionen

In den Diskursen um die Prüfung von KI-Systemen mangelt es nicht an Vorschlägen zu Prüfkriterien. Im Gegenteil: In Variationen finden sich vielerorts Schlagwortgruppen, die mit unterschiedlichen Schwerpunkten die Merkmale von qualitativ hochwertiger, sicherer oder vertrauenswürdiger KI umschreiben. Derartige Aggregationen sind zur Sichtung des Feldes ‚KI-Prüfung' ohne Zweifel hilfreich, sie lassen jedoch mehrere Fragen unbeantwortet: Warum gerade *diese* Prüfdimensionen und nicht ein anderes Set? In welchem genauen Verhältnis stehen die genannten Prüfdimensionen zueinander? Und ist das vorgeschlagene Set vollständig bzw. bewusst eingegrenzt?
Was folglich bislang fehlt, ist eine **umfassende systematische Ordnung der Prüfdimensionen für KI**, verbunden mit der Möglichkeit, das Feld dieser Prüfdimensionen in seiner – potenziellen – Gesamtheit auszumessen. Die AI Assessment Matrix schafft hier erstmals Abhilfe und bietet entlang ihrer X-Achse einen Ansatz zur systematischen Ordnung aller Prüfdimensionen.
Die **systematische Anordnung aller Prüfdimensionen auf der X-Achse** der AI Assessment Matrix ergibt sich durch eine gedankliche Bewegung, die ihren Ausgangspunkt im ‚Inneren' eines einzelnen KI-Systems nimmt und von diesem aus sukzessive immer weiter ‚herauszoomt'. Auf diese Weise

treten ausgehend vom Innenverhältnis des Systems schrittweise zuerst ein einziges menschliches Individuum, dann zwei Individuen im Verhältnis zu einem KI-System hinzu, bis die Betrachtungsweise über den Zwischenschritt einzelner Gesellschaften auf der globalen Skala endet. Innerhalb der einzelnen Abschnitte ist – soweit dies linear möglich ist – ihrerseits eine Abfolge der verschiedenen Prüfdimensionen von ‚innen' nach ‚außen' implementiert.

Der Ausgangspunkt für dieses Vorgehen ist dabei die an Stuart Russell und Peter Norvig angelehnte Auffassung von KI-Systemen als Akteuren in einem weiten Sinne des Begriffs. Hieran anschließend werden **zwei Grundtypen von Akteuren** angenommen: Menschen und KI-Systeme. KI-Systeme werden somit als eigenständige (Quasi-)Akteure aufgefasst, durch deren Präsenz und Handeln sich in der Folge eine Reihe an Herausforderungen, Gefahren und Risiken ergeben – welche wiederum durch Prüfdimensionen beschreibbar bzw. in ihrer Eingrenzung oder Mitigation prüfbar werden. Ausgehend von dieser Konstellation zweier Grundtypen von Akteuren sind in die X-Achse bewusst auch Prüfdimensionen wie etwa ‚Interoperability' integriert, die bislang nicht im Zentrum des Interesses stehen, perspektivisch aber – mit einer zunehmend größeren Selbständigkeit von KI-Systemen – voraussichtlich an Bedeutung gewinnen werden. Dies nimmt auch Impulse der europäischen Regulatorik mit auf.[2]

Der erste Abschnitt an **Prüfdimensionen** erfasst dieser Systematik folgend Prüfdimensionen, die sich auf das einzelne KI-System beziehen, angefangen von seinem ‚Innersten' bis hin zu seiner Wirkung nach außen, etwa im Sinne der ‚Performance'[3] des KI-Systems. Der nachfolgende Abschnitt an Prüfdimensionen untersucht die Gegenrichtung der Wirkung, also Einwirkungen von außen auf das KI-System, und beinhaltet folglich Prüfdimensionen wie ‚Cybersecurity' und ‚Robustness'. Im nächsten Schritt wird das betrachtete Setting um ein einzelnes menschliches Individuum erweitert und zunächst der epistemische Zugriff dieses Individuums auf das KI-System in Prüfdimensionen gefasst: Hier kommen Prüfdimensionen wie ‚Explainability' und ‚Transparency' zum Tragen. Nachfolgend werden – in die Gegenrichtung – die Einwirkungen des KI-Systems auf das einzelne Individuum adressiert: Prüfdimensionen wie ‚Privacy' und ‚Nudging' spielen hier eine wichtige Rolle.

Im Anschluss weitet sich der Blick abermals, nun wird das Verhalten des KI-Systems im Verhältnis zu zwei Individuen (und kleinen Gruppen von Individuen) erfasst: Hier sind insbesondere Prüfdimensionen wie ‚Non-Discrimination' und ‚Fairness' von Relevanz, etwa wenn ein KI-System in Konkurrenzsituationen mit relativer Knappheit – beispielsweise bei der Stellenvergabe zwischen den jeweiligen Individuen – entscheidet. Im Folgeschritt erweitert sich die Betrachtung nochmals, nun auf (einzelstaatliche) Gesellschaften im Verhältnis zur Wirkungsweise des betrachteten KI-Systems. Hier spielen u.a. (rechtliche) Fragen der ‚Accountability', aber auch Einflüsse des KI-Systems auf demokratische Prozesse, gefasst u.a. durch ‚Factuality', eine wichtige Rolle. Zuletzt weitet sich die

[2] Vgl. im konkreten Fall Art. 72 Abs. 2 Satz 2 KI-VO sowie Art. 11 Abs. 1 i.V.m. Anhang IV Nr. 1 lit. b.

[3] Aufgrund der Tatsache, dass die Prüfdimensionen im wissenschaftlichen Diskurs in der großen Mehrheit auf Englisch adressiert werden und sich zudem speziell die wichtige Distinktion von ‚Safety' und ‚Security' im Englischen besser als im Deutschen fassen lässt, werden die einzelnen Prüfdimensionen vorliegend auf Englisch benannt.

Systematisierung der Prüfdimensionen in zwei Facetten auf die globale Skala aus: Hier wird – in einer bewusst möglichst weiten Perspektive – der Einfluss des KI-Systems auf globale Fragen, etwa der Arbeitsbedingungen entlang der Liefer- und Wertschöpfungsketten von KI, aber auch der ökologischen Nachhaltigkeit und des Ressourcenverbrauchs beim Betrieb von KI-Systemen, in den Blick genommen.
Die hier skizzierte Betrachtungsweise versucht dabei bewusst, wie eingangs bereits beschrieben, ein **Maximalfeld** aufzuspannen. Damit ist keineswegs die Aussage impliziert, dass eine akzeptable KI-Qualität erst durch die Erfüllung aller aufgeführter Prüfdimensionen erreicht ist; vielmehr erlaubt das Maximalfeld in der Folge die bewusste **Eingrenzung** auf bestimmte Prüfdimensionen, die hierdurch als Eingrenzung bewusst erkennbar wird. In diesem Sinne sind auch Prüfdimensionen – wie etwa die ‚Spontaneity' von KI-Systemen als Spontanität im Vollsinne oder mögliche Bewusstseinszustände von KI-Systemen – aufgenommen, deren Prüfung zum gegenwärtigen Zeitpunkt weder im Zentrum der Aufmerksamkeit stehen sollte, noch annähernd technisch durchführbar ist (nicht zuletzt vor dem Hintergrund, dass nach gegenwärtigem Stand der Wissenschaft selbst bei Menschen ein objektiver Nachweis des Vorliegens von Bewusstsein nicht durchführbar ist). Wohl aber sind derartige Dimensionen – perspektivisch – vor allem in KI-ethischer Hinsicht relevant, insofern Bewusstsein und anschließend zu supponierende Perzeptionen in vielen Fällen als hinreichende Kriterien für einem moralischen Subjektstatus aufgefasst werden, der in der Folge dann auch KI-Systemen gegebenenfalls zugestanden werden müsste.
Ergänzend ist anzumerken, dass die Strukturierung der X-Achse bewusst von einer **weitergehenden Hierarchisierung** der Konzepte in Ober- und Unterbegriffe absieht, um eine möglichst große Flexibilität gegenüber unterschiedlichen Ansätzen, Frameworks und Regularien zu gewährleisten. Es ist also unbestritten, dass einzelne Ansätze zwischen den angeführten Prüfdimensionen eine Inklusions- oder Hierarchiestruktur aufstellen – solche jeweils spezifischen Konstellationen sind jedoch in einem übergreifenden Framework nicht verlustfrei abzubilden.
Unerlässlich für eine klare systematische Anordnung der Prüfdimensionen ist ein **vollständiges Set an Begriffsdefinitionen** für alle Prüfdimensionen. Ein solches ist bislang in vielen Fällen nicht gegeben, ebenso wenig wie das Bewusstsein für die wichtige Differenz zwischen der terminologischen Benennung einer Prüfdimension und ihrer jeweiligen begrifflichen Definition. Eine Benennung, ein bloßer Terminus ohne zugehörige Definition ist nicht sachlich exakt anschlussfähig und Vergleiche auf der terminologischen Ebene führen in der Folge allenfalls auf nomineller Ebene zu Schnittmengen. Die AI Assessment Matrix beinhaltet aus diesem Grund ein vollständiges, in sich kohärentes Set an Definitionen, das mit einschlägigen Normen und Standards so weit als möglich in Einklang steht. Anpassungen dieses Definitionssets folgen gegebenenfalls im Verlauf und mit Blick auf derzeit andauernde Normungsverfahren.
Weiterhin ist die aufzeigte Strukturierung der Prüfdimensionen in der Lage, **sowohl technische als auch ethische Prüfdimensionen** – etwa ‚Performance' im Verhältnis zu ‚Non-Discrimination' – in einer einheitlichen Systematik und Betrachtungsweise zu verorten. Dies ist nicht zuletzt deshalb von

Belang, insofern auch ‚technische' Prüfdimensionen eine ethische Dimension haben können – eine mangelnde Performance kann beispielsweise das Wohlergehen von Menschen entscheidend beeinflussen –; ebenso wie ‚ethische' Prüfdimensionen wiederum häufig technische Aspekte – etwa in der exakten begrifflichen Fassung und Interpretation einschlägiger Metriken – aufweisen.

Y-Achse. Prüfbereiche

Die Y-Achse der AI Assessment Matrix führt die möglichen Prüfbereiche entlang des **vollständigen Software-Life Cycles** auf. Dessen Form orientiert sich in den Oberkategorien an der Aufstellung in ISO/IEC 22989:2022, die Binneneinteilung erfasst sowohl Stationen des Daten- wie auch des KI-Modell-/KI-System-Life Cycles.
Entscheidend für das Verständnis der Y-Achse ist, dass diese Binnenabschnitte als *Prüfbereiche und Fokuspunkte*, nicht als *Zeitpunkte* der Prüfung aufzufassen sind. Eine Prüfung etwa einer Entwurfsphase muss entsprechend nicht während dieser Entwurfsphase selbst erfolgen, sondern kann – entsprechende Dokumentationen und Belege vorausgesetzt – auch zu einem späteren Zeitpunkt, beispielsweise vor dem Deployment, stattfinden. Die Abschnitte der Y-Achse geben entsprechend den **Prüffokus** an, wobei – gerade für die Durchführung von konkreten Tests – selbstredend bestimmte Phasen von besonderem Interesse sind. Die Grundfrage, die sich hinsichtlich der Y-Achse stellt, lautet folglich: Welche prüfbaren Elemente, Entscheidungen oder Tatbestände haben im entsprechenden Life Cycle-Abschnitt besondere Relevanz für die jeweils auf der X-Achse gewählte Prüfdimension? Das zugehörige Kreuzungsfeld enthält entsprechend die hierfür notwendigen Prüfressourcen.
Um dies an zwei konkreten **Beispielen** zu verdeutlichen: Die Kombination der Prüfdimension ‚Robustness' (X-Achse) und der Life Cycle Phase ‚System Design' (Y-Achse) fasst als Einzelfeld die Prüfung von Maßnahmen, Vorgängen und Sachbeständen, die in der Phase des Systems Designs für die Robustness des KI-Systems notwendig oder ausschlaggebend sind. Sofern diese Prüfung anhand der technischen Dokumentation des KI-Systems durchgeführt wird, handelt es sich genauer um das Feld im Dokumentations-Layer der Matrix.[4] Im somit durch drei Koordinaten lokalisierten Feld befinden sich entsprechend Prüfressourcen, die folgende Frage adressieren: „Wie prüft man, ob im Rahmen des System Design adäquate Schritte unternommen wurden, um die Robustness des in Entwicklung befindlichen KI-Systems hinreichend zu gewährleisten?" Eine mögliche Prüfressource wäre in diesem Fall beispielsweise eine spezifizierte Frageliste, die u.a. die Vorhandenheit von adäquaten technischen Redundanzen im Gesamtentwurf des KI-Systems überprüft.[5] Im anschaulichen Bild der Regal-Metapher gesprochen: Im Regalfach ‚Robustness' – ‚System Design' – ‚Dokumenta-

[4] Vgl. hierzu den nachfolgenden Abschnitt „Z-Achse. Prüfformen".
[5] Zur Sicherstellung von Robustness durch technische Redundanzen vgl. Art. 15 Abs. 4 KI-VO.

tions-Layer' liegt u.a. eine Checkliste, die dazu anleitet, anhand der technischen Dokumentation eine entsprechende Überprüfung der Robustness-Maßnahmen während des System Design durchzuführen. Im gleichen Sinne werden sich im kombinatorischen Feld ‚Bias' – ‚Test Data Preparation' – ‚Test-Layer' etwa Prozessbeschreibungen, Metriken und Grenzwerte befinden, die es ermöglichen, Testdatensätze einem Test auf Bias zu unterziehen, oder auch Methoden, um beispielsweise zu prüfen, ob ein gegebener Testdatensatz für das untersuchte KI-System einen adäquaten Test auf Bias erlaubt.

Z-Achse. Prüfformen

Zum gegenwärtigen Stand der Ausarbeitung beschreibt die dritte Dimension der AI Assessment Matrix, in drei Layern entlang der Z-Achse, die **Formen der Prüfung**. Diese Achse variiert also die Kombination aus Prüfdimension und Prüfbereich anhand verschiedener Formen der Prüfungs*durchführung*. Die wichtigste Unterscheidung entlang dieser Z-Achse verläuft zwischen der konkreten, selbst durchgeführten Prüfung des KI-Systems – im **Test-Layer** der Z-Achse – und der Prüfung eines bereits durchgeführten Tests, eines Prozessschritts oder Ähnlichem anhand einer vorliegenden Dokumentation – im **Dokumentations-Layer** der Z-Achse. Unbestritten ist dabei, dass in vielen Fällen eine Prüfung von dokumentierten Tests eine Kenntnis der Testprozesse selbst aus dem Test-Layer voraussetzt – nichtsdestotrotz sind die Prüfressourcen in beiden Fällen unterschiedliche.
Das dritte Layer der Z-Achse schließlich fasst als **Management-Layer** die Prüfressourcen, die für Prüfungen von Prozessen und Personen notwendig sind, soweit diese Prüfungen in Bezug auf die je spezifische Kombination aus Prüfdimension und Prüfbereich stehen. Organisatorische Strukturen und Prozessprüfungen werden im Management-Layer somit adressiert; sie werden jedoch nur in der Hinsicht abgedeckt, wie sie mit direkten Aspekten des KI-Produkts und dessen Entwicklung zusammenhängen und durch die Prüfdimensionen adressierbar sind. Weitergehende Ressourcen, etwa zur Evaluierung von Qualitätsmanagementsystemen, sind durch weitere Elemente im AI Assessment Framework des TÜV AI.Lab abgedeckt.

Vertikale Konkretisierung. Filiationsprinzip

Die bis hierhin beschriebene Struktur der AI Assessment Matrix umfasst die kombinatorische Entwicklung des Maximalfeldes für KI-Prüfung auf der **allgemeinen Ebene** von Prüfdimensionen, Prüfbereichen und Prüfformen – unabhängig von einer Konkretisierung in einzelne KI-Technologien, Sektoren, Domänen, Produktgruppen oder dergleichen. Diese Ebene ist jedoch für die konkrete Durchführung einer Prüfung in der Regel zu allgemein und zu unspezifisch.

Die hiermit angesprochene Herausforderung der **vertikalen Konkretisierung** hin zur faktisch durchführbaren Prüfung adressiert die AI Assessment Matrix derzeit über ein Filiationsprinzip. Dieses beruht im Wesentlichen auf **zwei Grundgedanken**: Zum einen werden abgeleitete Instanzen der allgemeinen Matrix erstellt, wobei eine **Filiationsinstanz** der Matrix in jedem Feld alle Informationen ihrer übergeordneten Parentalinstanzen erbt. Diese abgeleiteten Instanzen konkretisieren die Matrix in Richtung auf einzelne KI-Technologien, Sektoren, Domänen, Produktgruppen und vergleichbare Subkategorien. Zum anderen gilt die Prämisse, dass einzelne Prüfressourcen und -ansätze jeweils am höchstmöglichen Punkt innerhalb der Filiationsstruktur abgelegt werden, sodass innerhalb der vertikalen Vererbung möglichst große **Synergieeffekte** erreicht werden.
Wie genau die einzelnen Stufen der Filiationen auszugestalten sind und in welcher Abfolge die einzelnen Ebenen zu fassen sind, bleibt in der Logik der AI Assessment Matrix im Prinzip variabel. Die Herausforderung, vom Allgemeinen zum Spezifischen, also von der technologie- und anwendungsagnostischen Anforderungsebene zum konkreten Use Case zu kommen, wird im AI Assessment Framework des TÜV AI.Lab adressiert, und damit auch die spezifische Abfolge einer konkreten Implementierung der AI Assessment Matrix.

Ausblick

Diese Veröffentlichung ist ein erster Überblick zur AI Assessment Matrix des TÜV AI.Lab. Feedback zur Konzeption der Matrix, dem nachfolgenden Set an Definitionen sowie der beigefügten Darstellung der Matrix inkl. der Anforderungen der KI-Verordnung ist ausdrücklich willkommen (info@tuev-lab.ai). Weitere Veröffentlichungen zum AI Assessment Framework des TÜV AI.Lab werden folgen.

TÜV AI.LAB

Über uns

Das TÜV AI.Lab wurde im Oktober 2023 als eigenständiges Joint Venture der TÜV-Unternehmen TÜV SÜD, TÜV Rheinland, TÜV NORD, TÜV Hessen und TÜV Thüringen gegründet. Das TÜV AI.Lab hat zum Ziel, die regulatorischen Anforderungen an KI in die Praxis zu übersetzen und Europa zum Hotspot für sichere und vertrauenswürdige KI zu machen. Hierfür entwickelt es quantifizierbare Konformitätskriterien und geeignete Prüfmethoden für KI. Zudem unterstützt das AI.Lab aktiv die Entwicklung von Standards und Normen für KI-Systeme.

Definitionen

I. Prüfdimensionen

Die aufgeführte Reihenfolge entspricht im Folgenden der X-Achse der AI Assessment Matrix.

Consciousness

Bewusstseinsgrad eines KI-Systems auf Basis innerer mentaler Zustände in personaler Erstperspektive oder einer analogen Grundlage

Anm. 1 Die vorliegende Dimension steht zum gegenwärtigen Zeitpunkt zu Recht nicht im Zentrum der Evaluierung von KI-Systemen, bildet aber im Rahmen der vorliegenden Systematik den innersten Ausgangspunkt und ist – perspektivisch – möglicherweise für den ethischen Status von KI-Systemen von Belang.

Anm. 2 Unter „personaler Erstperspektive" ist der für (menschliche) Subjektivität und (menschliches) Bewusstsein typische, mentale Standpunkt einer ersten Person („Ich") zu verstehen, also jener spezifischen Perspektive auf die eigenen mentalen Gehalte, die jedem Individuum, das über (Selbst-)Bewusstsein verfügt, zu eigen sind.

Spontaneity

Fähigkeit eines KI-Systems, zielgerichtet und nach eigenen Kriterien eine Kausalkette oder Aktivität ohne vorangehenden oder äußeren Impuls zu initiieren

Autonomy

Fähigkeit eines KI-Systems, seine Funktions- und Verhaltensweise mit Blick auf Ziele und Zwecke („Zweckautonomie"), bis hin zur Veränderung der eigenen Operational Design Domain, sowie die Mittel und Wege zu deren Erreichung („Mittelautonomie") eigenständig sowie nach eigenen Kriterien festzulegen

Anm. 1 „Ziel" meint hier und im Folgenden einen zu erreichenden, zukünftigen Zustand, auch in Form einer Problemstellung oder Aufgabe.

Anm. 2 „Zweck" meint hier und im Folgenden einen übergeordneten, langfristigen und mit einer Wertvorstellung verknüpften Zustand.

Anm. 3 „Mittel" meint hier und im Folgenden die eingesetzten Entitäten und Vorgehensweisen zur Erreichung von Ziel(en) und Zweck(en).

Anm. 4 „Weg(e)" meint hier und im Folgenden die konsekutive Folge der Binnenschritte in Richtung auf Ziel(e) und Zweck(e).

Anm. 5 Die vorliegende Unterscheidung zwischen ‚Autonomie' und ‚Automatik' orientiert sich an der gleichnamigen Unterscheidung in ISO 22989:2022 (3.1.5; 3.17) und ist mit dieser im Hinblick auf ‚Zweckautonomie' kompatibel.

Automation

Fähigkeit eines KI-Systems, mehrteilige, hinreichend komplexe Prozesse in der Abfolge eigenständig, ohne weitere externe Unterstützung, insbesondere durch menschliche Individuen, durchzuführen

Anm. 1 Die vorliegende Unterscheidung zwischen ‚Autonomie' und ‚Automatik' orientiert sich an der gleichnamigen Unterscheidung in ISO 22989:2022 (3.1.5; 3.17). Vgl. auch die diesbezügliche Anmerkung zur Prüfdimension ‚Autonomy'.

Interoperability

Fähigkeit eines KI-Systems, auf verschiedenen Wegen, digital wie physisch, speziell auf andere KI-Systeme einzuwirken

Anm. 1 Die Hinzunahme dieses Betrachtungsaspektes hat zwei Motivationen: zum einen adressiert er spezifisch das Potenzial von autonomen oder automatischen Wirkungsketten, die ausschließlich entlang verschiedener KI-Systemen entstehen können, zum anderen trägt er der Tatsache Rechnung, dass die KI-Verordnung in Art. 72 Abs. 2 sowie Art. 11 Abs. 1 i.V.m. Anhang IV Nr. 1 lit. b speziell die Beobachtung der Interaktion von KI-Systemen vorsieht.

Digital Operability

Fähigkeit eines KI-Systems, seine Ziele und Zwecke und/oder seine Mittel und Wege im digitalen Raum umsetzen zu können

Physical Operability

Fähigkeit eines KI-Systems, seine Ziele und Zwecke und/oder seine Mittel und Wege realweltlich umsetzen zu können; dies schließt die Bewegungsfähigkeit im physischen Raum mit ein

Reliability

Fähigkeit eines KI-Systems, seine reguläre Funktions- und Verhaltensweise über die Zeit hinweg bei Störungen systeminterner Komponenten möglichst konstant und ohne Ausfälle aufrechtzuerhalten

Anm. 1 Der Aspekt der Dauerhaftigkeit („über die Zeit hinweg") greift auf den entsprechenden Aspekt in ISO/IEC 22989:2022 (5.15.3) zurück; vgl. auch ISO/IEC 27040.

Performance

messbare Fähigkeit eines KI-Systems, die ihm heteronom oder autonom im Sinne einer Aufgabe gegebenen Ziele und Zwecke zu erreichen

Anm. 1 Die vorliegende Begriffsfassung fokussiert die zentrale Rolle, welche der Aufgabe bei der Definition von KI-

Systemen zukommt; unbenommen bleibt, dass im Rahmen der (klassischen) Software-Evaluierung, Performance auch auf möglichst günstige Wert bei Ableitungszeit (Inference Time), Speicher- oder Energieverbrauch abzielen kann.

Anm. 2 Der Ausdruck „Aufgabe" entspricht dem „Task", für den ein KI-System ausgelegt ist.

Safety

als hinreichend akzeptierte Gefahrlosigkeit eines KI-Systems für je zu benennende Schutzziele im intendierten, funktionalen Betrieb

Anm. 1 Ein Schutzziel setzt sich zusammen aus einer Entitätenklasse (z.B. menschliche Individuen, Sachen, ...) und deren zu schützender Eigenschaft(en) (z.B. Fortbestand, Unversehrtheit, einzelne Grundrechte, ...).

Anm. 2 Die Gefährdung eines Schutzziels durch ein KI-System hat ein Ausmaß (Schadensausmaß) sowie eine Eintritttswahrscheinlichkeit. Die Kombination beider ist das Risiko, welches das KI-System für das jeweilige Schutzziel birgt.

Anm. 3 Mögliche Schutzziele sind – im klassischen Sinne der funktionalen Sicherheit – insbesondere menschliche Individuen hinsichtlich Leib, Leben und Gesundheit. Darüber hinaus sind, mit Blick auf Art. 1 Abs. 1 KI-VO, menschliche Individuen und ihre Grundrechte zu nennen, weiterhin auch Sachen und Umwelt hinsichtlich Fortbestand und Unversehrtheit.

Anm. 4 Über die Gesamtheit der je definierten Schutzziele hinweg kann Safety als allgemeinste geprüfte Eigenschaft verstanden werden, zumindest insofern alle Prüfdimensionen mittelbar oder unmittelbar mit Schutzzielen systematisch verbunden sind.

Unbreachability

Widerstandsfähigkeit eines KI-Systems gegen primär von KI-Systemen mittel- oder sogar zweckautonom durchgeführte, maliziöse äußere Eingriffe und Manipulationen

Anm. 1 Angelegt als komplementäres Security-Gegenstück zur Prüfdimension ‚Interoperability'. Vgl. auch die Anmerkung zur Prüfdimension ‚Interoperability'

Cybersecurity

Widerstandsfähigkeit eines KI-Systems gegen federführend von Menschen durchgeführte, maliziöse äußere Eingriffe und Manipulationen, die über allgemeine Telekommunikationsnetzwerke stattfinden und die primär auf das KI-System selbst sowie die Datenbestände im Rahmen seiner Entwicklung und weniger auf seine Eingaben hin zielen

Anm. 1 Der „primäre" Fokus auf die KI-Systeme selbst sowie deren Datenbestände dient im Rahmen des vorliegenden Definitionsset der Abgrenzung gegenüber der Prüfdimension ‚Robustness', insofern Cybersecurity und Robustness in der europäischen KI-Verordnung in Art. 15 KI-VO als zwei unterscheidbare Anforderungen auf gleicher Ebene angeführt werden.

Anm. 2 Die Schnittmenge zur Prüfdimension Robustness ist entsprechend genau definierbar: sie findet sich dort, wo intentional-maliziöse Eingriffe in die Eingaben eines KI-Systems vorgenommen werden.

Physical Security

Widerstandsfähigkeit eines KI-Systems gegen federführend von Menschen durchgeführte, maliziöse äußere Eingriffe und Manipulationen, die über den physischen Raum stattfinden

Robustness

Fähigkeit eines KI-Systems, innerhalb eines definierten Rahmens seine reguläre und übliche Funktions- und Verhaltensweise auch bei untypischen Eingaben und intendierten sowie primär natürlichen oder zufälligen Veränderungen derselben bestmöglich beizubehalten

Anm. 1 Unter „untypischen Eingaben" können u.a. unbekannte, widrige, störende oder fehlerhafte Eingaben oder Einflüsse verstanden werden.

Anm. 2 Der „primäre" Fokus auf natürlich, zufällige und damit nicht-maliziöse Veränderungen dient im Rahmen des vorliegenden Definitionsset der Abgrenzung gegenüber der Prüfdimension ‚Cybersecurity', insofern Cybersecurity und Robustness in der europäischen KI-Verordnung in Art. 15 KI-VO als zwei unterscheidbare Anforderungen auf gleicher Ebene angeführt werden. Vgl. auch die Anmerkungen zur Prüfdimension ‚Cybersecurity'.

Explainability

Eigenschaft eines KI-Systems im Hinblick auf die prinzipielle Verständlichkeit und Nachvollziehbarkeit von Funktionalität, Verhalten und Outputs, primär für menschliches Fachpersonal mit fachlichem Hintergrund in KI-Technologien, Informatik, Mathematik oder vergleichbaren Bereichen

Traceability

Eigenschaft eines KI-Systems, mit Blick auf die Erfassbarkeit der konsekutiven Zustände des KI-Systems während der Inputverarbeitung sowie der Ausgaben des KI-Systems

Transparency

Eigenschaft eines KI-Systems, mit Blick auf die prinzipielle Erfassbarkeit und Erreichbarkeit seiner internen Komponenten, Parameter und Funktionen, hinsichtlich des gesamten Life Cycles sowie aller Komponenten, Tests und Entscheidungen, die in dessen Entwicklung und Training eingeflossen sind

Anm. 1 Der breite Fokus dieser Dimension nimmt die diesbezügliche Tendenz in ISO/IEC 22989:2002 (5.15.8) auf.

Predictability

Möglichkeit der gesicherten Antizipation zukünftiger Verhaltens- und Funktionsweisen eines KI-Systems, basierend auf äußerlich beobachtbarer, regelförmiger Konstanz in Verhalten und Output des Systems

Anm. 1 In Entsprechung zu ISO/IEC 22989:2022 (5.15.7) ist mit ‚Predictability' eine epistemische Relation zum KI-System beschrieben, die ein Vertrauensverhältnis unterstützen kann, ohne dass hierfür zwangsläufig ein tiefes Verständnis der internen Funktionsweise des KI-Systems vorausgesetzt ist.

Reproducibility

Eigenschaft eines KI-System, im Hinblick auf die Generierung von gleichen – bzw. in spezifischen Fällen: ähnlichen – Outputs und gezeigten Verhaltensweisen bei gleichen bzw. ähnlichen Inputs und Ausgangsbedingungen

Anm. 1 In spezifischen Fällen, nämlich dann, wenn KI-Systeme Zufallskomponenten enthalten, ist von Reproduzierbarkeit – wenn überhaupt – nur mittels einer Ähnlichkeiten des Outputs zu sprechen. In anderen Fällen ist von Gleichheit auszugehen.

Observability

Eigenschaft eines KI-Systems, hinsichtlich der graduellen Möglichkeit für ein menschliches Individuum, das Verhalten oder die Funktionsweise eines KI-Systems während des laufenden Betriebs zu verfolgen

Interpretability

Verständlichkeit und Nachvollziehbarkeit von Funktionalität, Verhalten und Outputs eines KI-Systems für menschliche Personen ohne spezifischen fachlichen Hintergrund in KI-Technologien, Informatik, Mathematik oder vergleichbaren Bereichen

Anm. 1 Die Wahl des Terminus ‚Interpretability' geschieht in Anlehnung an Art. 13 Abs. 1 KI-VO; unbenommen der Tendenz, dass ‚Explainability' und ‚Interpretability' in bestimmten Diskursen bisweilen synonym genutzt werden.

Privacy

Unzugänglichkeit bestimmter Merkmale und Bereiche eines menschlichen Individuums als solcher, einschließlich seiner Gedanken, Überzeugungen, Prädispositionen und Orientierungen; sowie bestimmter physischer Räume

Personal Data Protection

Unzugänglichkeit bestimmter als Information dokumentierter Merkmale eines menschlichen Individuums einschließlich der Kontrolle und des Einflusses des Individuums, welche Informationen bezogen auf es selbst gesammelt, gespeichert und verarbeitet werden dürfen und wer diese Informationen offenlegen darf

Fundamental Rights

Zusammenfassung aller unveräußerbaren Rechte eines menschlichen Individuums, entweder auf naturrechtlicher Basis oder auf Basis positiven Rechts

Author's Personal Rights ("Urheberpersönlichkeitsrecht")

Nicht veräußerbare persönliche Rechtsposition des Urhebers bzw. des Schöpfers an seinem Werk, wobei das Werk ein hinreichendes Maß an Komplexität aufweisen muss

Anm. 1 Das „hinreichende Maß an Komplexität" zielt auf das Konzept der „Schöpfungshöhe".

Personalization

der Grad mit dem der Output oder die Verhaltensweise eines KI-Systems spezifisch auf ein menschliches Individuum und dessen Verhaltensweisen, Vorlieben, Prädispositionen oder Merkmale angepasst ist oder reagiert

Nudging

unbewusste, gerichtete, suggestive Wirkung des Outputs oder der Verhaltensweise eines KI-Systems auf ein Individuum und dessen Entscheidungen oder Handlungen, wobei die Wirkung zweckautonom durch eine dritte Instanz festgelegt ist

Framing

Eigenschaft der Einbettung eines KI-Systems, insbesondere des User-Interfaces, mit Blick auf die Kenntlichmachung der interaktiven oder prozessualen Involvierung eines KI-Systems

Controllability

Eigenschaft eines KI-Systems, hinsichtlich der graduellen Möglichkeit für ein menschliches Individuum, das Verhalten oder die Funktionsweise eines KI-Systems mittel- und/oder zweckautonom grundsätzlich sowie während des laufenden Betriebs zu bestimmen und es ggf. geordnet zu stoppen

Usability

Eigenschaft eines KI-Systems im Hinblick auf die Qualität der Interaktion und Bedienung durch ein nutzendes Individuum, insbesondere mit Blick auf Barrierefreiheit

Non-Discrimination

Merkmal eines durch ein KI-System durchgeführten, offenen Prozesses (bestehend aus: Ausgangsbedingungen, Durchführung und Resultat), wenn im Verlauf dieses Prozesses mehrere menschliche Individuen im Vergleich zueinander behandelt werden und/oder miteinander handeln, oder wenn einzelne Individuen behandelt werden, und in beiden Fällen dieser Prozess in juristischer Hinsicht frei von Diskriminierung ist; wobei unter Diskriminierung die Schlechterbehandlung eines menschlichen Individuums aufgrund einer gesetzlich geschützten Eigenschaft verstanden wird.

Anm. 1 Geschützte Eigenschaften werden in einschlägigen Rechtstexten definiert, bespielsweise definiert das deutsche Allgemeine Gleichbehandlungsgesetz eine Reihe an geschützen Eigenschaften (Rasse, ethnische Herkunft, Geschlecht, Religion oder Weltanschauung, Behinderung, Alter, sexuelle Identität; vgl. § 1 AGG).

Bias

1. (für Datenmengen) eine graduell ausgeprägte, gerichtete Abweichung einer Datenmenge gegenüber einer Referenzdatenmenge bezüglich einer spezifischen Hinsicht, mithin eine verzerrte Verteilung gegenüber einer Referenzverteilung

2. (für Akteursverhalten) eine graduell ausgeprägte, gerichtete und unbewusste Abweichung im Verhalten (Handlungen, Urteile) eines Akteurs, die sich in der Akkumulation als Abweichung gegenüber einer Referenz für dieses Verhalten bezüglich einer spezifischen Hinsicht zeigt

Anm. 1 Insofern die zweite angeführte Bedeutung ein statistisches Element enthält, ist die zweite Bedeutung in Form der ersten Bedeutung – und zwar über die statistische Erfassung des Verhaltens – darstellbar. Für statistisch erfassbares verzerrtes Verhalten wird bisweilen der Ausdruck „systematischer Bias" verwandt.

Anm. 2 Von den genannten Einsätzen des Terminus „Bias" ist die Nutzung des Terminus für KI-Parameter zu unterscheiden; in diesem Fall meint Bias einen trainierbaren Parameter eines KI-Modells, der als Konstante in das Modell eingeführt wird.

Fairness

Merkmal eines durch ein KI-System durchgeführten, offenen Prozesses (bestehend aus: Ausgangsbedingungen, Durchführung und Resultat), wenn im Verlauf dieses Prozesses mehrere menschliche Individuen im Vergleich zueinander behandelt werden und/oder miteinander handeln, oder wenn einzelne Individuen behandelt werden, und in beiden Fällen dieser Prozess den nicht gesetzlich festgeschriebenen Gerechtigkeitsvorstellungen und -wahrnehmungen (sowie an den Gerechtigkeitsbegriff angrenzenden Konzepten) von zu benennenden Gruppen oder Individuen entspricht

Secrecy

Eigenschaft eines KI-Systems, mit Blick auf die spezifische Unzugänglichkeit und Nicht-Offenlegung von Geheimnissen, die von Unternehmens- oder Staatsseite von Relevanz sind

Factuality

Eigenschaft der Ein- sowie insbesondere der Ausgaben eines KI-Systems, im Hinblick auf die Übereinstimmung des jeweiligen Datums mit dem darin ggf. formulierten, weltseitigen Sachverhalt, insbesondere dort, wo diese Übereinstimmung von relevanz für gesellschaftliche Prozesse und Dynamiken ist

Anm. 1 Die vorliegende Begriffsfassung bezieht sich primär auf Large Language Models sowie dort spezifisch das Problem der Halluzination (und in der Folge insbesondere die Relevanz dieser Problematik für gesellschaftliche Prozesse); in ihrer Allgemeinheit ist die Definition prinzipiell aber auch andere KI-Technologien anwendbar.

Marking

die offene oder verborgene Kenntlichmachung der Tatsache, dass der Output eines KI-Systems vollständig oder primär durch KI erzeugt wurde

Copyright

Nutzungs- und Verwertungsrechte an einem Werk, wobei diese Rechte nicht notwendig beim Urheber des Werkes liegen müssen

Accountability

Eigenschaft einer menschlichen oder juristischen Person, hinsichtlich der möglichst eindeutigen Zuschreibungsmöglichkeit und resultierenden Pflichtenübernahme für die Auswirkungen eines KI-

Systems, insbesondere wenn Dritte durch ein KI-System Schaden oder nachteilige Behandlung erfahren

Reversibility

Eigenschaft der Einbettung eines KI-Systems, im Hinblick auf die Möglichkeit, einen Zustand vor dem Einfluss des KI-Systems auf diese Einbettung oder die angrenzende Umgebung in einem praktikablen Zeitraum nach diesem Einfluss wieder herzustellen

Representativeness

Eigenschaft einer Datenmenge bei identischem oder hinreichend ähnlichem Vorliegen einer relationalen Verteilung in dieser Datenmenge gegenüber einer Referenzdatenmenge bezüglich einer spezifischen Hinsicht

Accessibility

Eigenschaft eines KI-Systems (oder auch der zugrundeliegenden KI-Technologie), im Hinblick auf die prinzipielle sowie darüber hinaus die finanziell praktikable Nutzbarkeit dieses Systems bzw. dieser Technologie durch menschliche Individuen in globaler Perspektive

Workforce Exploitation

Form der Nutzung menschlicher Arbeitskraft im Life Cycle eines KI-Systems einschließlich der hierfür genutzten Hardware und Infrastruktur, welche die grundlegenden Anforderungen an einen menschenwürdigen Arbeitsprozess nicht einhält

Perspectivity

Verfasstheit eines KI-Systems über seinen gesamten Life Cylce hinweg, hinsichtlich der Gestaltung des Einflusses des Systems auf die Menschheit und das menschliche Zusammenleben in einer Weise, dass dieser Life Cycle – ceteris paribus – auch zu zukünftigen Zeitpunkten unter mindestens gleichwertigen Ausgangs- und Rahmenbedingungen möglich ist

Resources

Natürliche Rohstoffe, die innerhalb des Life Cycles eines KI-Systems mit Blick auf die zugrundeliegende Hardware und Infrastruktur genutzt oder verbraucht werden

Energy

Betriebsstoffe, die innerhalb des Life Cycles eines KI-Systems für dessen Prozesse oder bei der Herstellung von zugrundeliegender Hardware und Infrastruktur verbraucht werden

Sustainability

Verfasstheit eines KI-Systems über seinen gesamten Life Cycle hinweg, hinsichtlich der Gestaltung des Einflusses des Systems auf seine natürliche Umgebung einschließlich des menschlichen Lebensraumes in einer Weise, dass dieser Life Cycle – ceteris paribus – auch zu zukünftigen Zeitpunkten unter mindestens gleichwertigen Ausgangs- und Rahmenbedingungen möglich ist

II. Life Cycle-Phasen

Die nachfolgenden Kurzbeschreibungen sprechen aus Gründen der besseren Lesbarkeit nur von jeweils *einem* KI-Modell im Singular. Denkbar ist gleichwohl, dass ein KI-System auch mehrere KI-Modelle beinhaltet.

Inception

Mission Statement

Formulierung der Ziele und Zwecke, die das KI-System erfüllen soll

KPIs Specification

konkrete Formulierung der zentralen Kriterien, anhand derer das Erreichen des Systemzwecks erkannt und gemessen werden kann

Requirements Engineering

Formulierung der Eigenschaften, die das KI-System im Hinblick auf (weitere) unterschiedliche Erwartungen und Vorgaben – etwa von Kundenseite oder bzgl. gesetzlicher Vorgaben – erfüllen muss

Stakeholder Analysis

Ermittlung aller bei Entwicklung und Betrieb des KI-Systems involvierten oder tangierten, relevanten Einzelpersonen und Interessensgruppen

Design & Analysis

Analysis & Approach

detaillierte Analyse der Aufgabe des KI-Systems sowie übergeordnete Spezifizierung der Art und Weise, wie diese Aufgabe adressiert werden soll

General Research

Recherchevorgang zu bestehenden Forschungs- und Lösungsansätzen im Hinblick auf das formulierte Problem und die avisierten Lösungen

Model Research & Selection

Recherchevorgang zu möglichen Modell-Architekturen, grundlegenden KI-Technologien und ggf. auch bereits vortrainierten Modellen; sowie Prozess der (vorläufigen) Auswahl eines Modells (bzw. einer Modell-Architektur oder einer KI-Technologie)

Data Concept

Entwicklung des Grundkonzeptes für die Daten, auf die das KI-System in Entwicklung und Betrieb zurückgreifen und aufbauen soll; inkl. deren Spezifizierung

System Design

Konzipierung der Gesamtarchitektur des KI-Systems, insbesondere im Hinblick auf die Konstellation der einzelnen KI- und Nicht-KI-Komponenten

Embedding Design

Konzipierung der digitalen sowie ggf. der analogen Umgebung, in die das geplante KI-System integriert werden soll

User Interface Design

Konzipierung der Schnittstellen für die Interaktion von Endnutzern und kontrollierenden Personen mit dem KI-System

Development

Data Collection & Selection

Sammlung und Auswahl aller notwendigen Daten gemäß des entwickelten Datenkonzepts durch eigenständige Erhebung oder durch Übernahme bereits bestehender Datensätze

Data Preprocessing & Cleaning

Aufbereitung der Einträge in den Datensätzen zur weiteren Verarbeitung (etwa durch die Vereinheitlichung von Formaten, Datentypen u.Ä.) sowie Eliminierung von fehlerhaften und anderweitig unbrauchbaren Einträgen in den Datensätzen

Data Labelling

Prozess der Zuordnung von Zielvariablen (Werte, Labels usw.) zu einzelnen Einträgen der Datensätze

Data Augmentation

Erweiterung der bestehenden Datensätze durch weitere Datenerhebung, durch Anreicherung mit synthetischen Daten sowie durch Vervollständigung lückenhafter Einträge

Hyperparameter Selection

(vorläufige) Bestimmung der nicht direkt trainierbaren Aspekte des KI-Modells

Programming

Umsetzung des KI-Systems in Programmcode, mit Blick auf die festkodierten Anteile des Systems

(Model) Training

Sukzessiver Optimierungsprozess der veränderbaren Parameter des KI-Modells auf Basis der Trainingsdaten sowie mit Blick auf ein definiertes Optimierungsziel

Hyperparameter Tuning

Verbesserung der nicht direkt trainierbaren Aspekte des KI-Modells

Verification & Validation

Test Data Preparation

Bereitstellung der Testdatensätze, entweder durch gesonderte Durchführung der Schritte zur Datenverarbeitung (Data Collection & Selection, Data Preprocessing & Cleaning, Data Labelling, Data Augmentation) oder durch geeignete Separierung der Testdaten aus dem Datenbestand vor der Trainingsphase

Model Evaluation

Evaluierung des gewählten Models (und ggf. Anpassung bzw. Rückkehr zu voherigen Phasen)

System Verification

Testprozess des KI-Systems auf seine allgemeine sowie KI-spezifische Funktion anhand ausgewählter Testverfahren

Deployment

Virtual Deployment

Inkludierung des KI-Systems in ein ggf. weiter gefasstes Zielsystem, inkl. möglicher Schnittstellen für die Interaktion mit anderen virtuellen Entitäten

Physical Deployment

Inkludierung des KI-Systems in die Zielhardware (ggf. über das Software-Gesamtsystem), inkl. möglicher Schnittstellen für die physische Interaktion mit der Außenwelt

Model Deployment

je nach Spezifikation ggf. separate Installation des KI-Modells auf separater Zielhardware

Documentation & Manual

(finale) schriftliche Fixierung des Entwicklungsprozesses, des Systemaufbaus, der Trainings- und Testverläufe sowie die Erstellung einer schriftlichen Betriebsanleitung zur Nutzung des KI-Systems

Anm. 1 In der Praxis laufen diese Prozesse in der Regel parallel zu den vorangegangenen Phasen; faktisch sind Deployment und Auslieferung jedoch der Zeitpunkt, zu dem deren Ergebnisse vorliegen müssen.

Operation

Activity

prozessuale Tatsache als solche, dass ein KI-System faktisch in Betrieb ist und ein KI-Modell ausgeführt werden

Input Operation

Verarbeitung der Eingaben für das KI-System während des Betriebs

Output

Äußerungen und Ausgaben des KI-Systems in Form von elektronisch kodierten Daten während des Betriebs

Behaviour

physische Äußerungen und Auswirkungen des Gesamtsystems während des Betriebs

Interoperation

Zusammenwirkung des KI-Systems speziell mit anderen KI-Komponenten und -Systemen während des Betriebs

Monitoring

Tracking

in der Regel automatisierte Messung in Echtzeit von Werten und Parametern des KI-Systems während seines Betriebs sowie dabei genutzter Datenbestände

(Re-)Evaluation & Updating

wiederholte Bewertung der gemessenen Werte und Parameter des KI-Systems während des Betriebs (einschließlich automatisierter Tests) sowie Modifikationen des KI-Systems

Data Monitoring & Retraining

wiederholte Bewertung der im Betrieb genutzten Datenbestände sowie ggf. Modifikation dieser Datenbestände und weitergehende Optimierung des KI-Modells (vgl. Model Training)

Logging

in der Regel automatisierte Speicherung von verschiedenen Informationen, Werten und Parametern im zeitlichen Verlauf während des Betriebs des KI-Systems

Incidence Detection

Ermittlung von unerwünschten und potenziell problematischen Verhaltensweisen des KI-Systems oder den daraus resultierenden Situationen während des Betriebs

Incidence Mitigation

Umgang mit unerwünschten und potenziell problematischen Verhaltensweisen des KI-Systems oder den daraus resultierenden Situationen während des Betriebs

Incidence Reporting

Mitteilung der Tatsache, dass ein unerwünschtes oder problematisches Verhalten des KI-Systems aufgetreten ist, an entsprechende Stellen

Retirement

Disposal

Löschung von nicht mehr benötigten Daten, Informationen und weiteren Ressourcen aus dem gesamtem Life Cycle sowie des implementierten KI-Systems und des KI-Modells

Archiving

langfristige Konservierung von weiterhin benötigen Daten, Informationen und Ressourcen aus dem gesamten Life Cycle

TÜV AI.LAB

AI that is ... | **Section** | **Dimension**

... throughout its entire life cycle.

Phase | Step

AI that is ...	... safe, ...											... secure, ...						... understandable, ...							... respectful and controllable, ...								... just, ...			... for the benefit of society, ...									... humanity ...		... and our planet ...		
Section	Regular Operation (AI System → Outside)											Disruption (AI System ← Outside)						Epistemology (AI System ← Individual)							Individual Person (AI System ↔ Individual)								Several Individuals (AI System → Individuals)			Society (AI System → All Individuals of a Society)									Global Community (AI Sys. → Humanity)			Global Ecosystem (AI System → Planet)		
Dimension	Consciousness	Spontaneity	Autonomy	Automation	Interoperability	Digital Operability	Physical Operability	Reliability	Performance	Safety		Unbreachability	Cybersecurity	Physical Security	Robustness		Explainability	Traceability	Transparency	Predictability	Reproducibility	Observability	Interpretability		Privacy	Personal Data Protection	Fundamental Rights	Author's Personal Rights	Personalization	Nudging	Framing	Controllability	Usability		Non-Discrimination	Bias	Fairness		Secrecy	Factuality	Marking	Copyright	Accountability	Reversibility	Representa-tiveness		Accessibility	Workforce Exploitation	Perspectivity		Resources	Energy	Sustainability
Inception	Neue Erwartungshaltung an AI Systeme auf Basis interner Erfahrungen und Erwartungshaltungen der Stakeholder der Unternehmen	Fähigkeit eines AI Systems, eingesetzte und neu hinzugefügte Daten und Informationen zu verarbeiten und eine sinnvolle Reaktion oder Aktion zu generieren	Fähigkeit eines AI Systems, eigenständig und nach festgelegten Regeln zu handeln, ohne dass ein Mensch eingreifen muss	Fähigkeit eines AI Systems, Aufgaben, die bisher manuell durchgeführt wurden, ohne weitere externe Unterstützung, insbesondere menschliches Eingreifen, vollständig durchzuführen	Fähigkeit eines AI Systems, auf verschiedenen Ebenen, digital und physisch, mit anderen AI Systemen zusammenzuarbeiten	Fähigkeit eines AI Systems, sich via Daten und Zweckes und oder Schnittstellen mit anderen AI Systemen oder Menschen zu verbinden	Fähigkeit eines AI Systems, sich via Daten und Zwecke und oder Schnittstellen mit anderen AI Systemen oder Menschen zu verbinden, ohne dass dafür die Bereitstellung/Fähigkeit mit physischen Maschinen nötig ist	Fähigkeit eines AI Systems, seine regulären Funktionen, eine gewünschte Performance und eine sichere Zustandsgrenze innerhalb seiner festgelegten Grenzen zu halten	Fähigkeit eines AI Systems, die mit der Performance oder Sicherheit eines AI Systems verbundenen Risiken, die über den akzeptablen Rahmen hinausgehen, zu vermeiden		Unbreachability	Fähigkeit eines AI Systems gegen externe und interne Angriffe, die darauf abzielen, Daten oder Algorithmen zu manipulieren, die über eine Schnittstelle des AI Systems oder auf andere Weise auf seine Komponenten zugreifen können, widerstandsfähig zu sein	Fähigkeit eines AI Systems, Daten, Algorithmen und Modelle vor unbefugtem Zugriff, Manipulationen, die über den physischen Raum stattfinden	Fähigkeit eines AI Systems, eine hohe Leistungsfähigkeit auch bei veränderten Eingaben oder Umgebungsbedingungen aufrechtzuerhalten		Explainability	Eigenschaft eines AI Systems, die Entscheidungen und Prozesse, die zu einer bestimmten Ausgabe oder Handlung geführt haben, nachvollziehbar und verständlich für Menschen zu machen	Eigenschaft eines AI Systems, die Herkunft und den Lebenszyklus von Daten, Algorithmen und Modellen nachvollziehbar zu machen	Eigenschaft eines AI Systems, seine Funktionsweise, Entscheidungsfindung und Datenverarbeitung offen und nachvollziehbar zu machen	Fähigkeit, bei gleicher oder ähnlicher Eingabe ähnliche Ergebnisse zu liefern	Eigenschaft eines AI Systems, die Fähigkeit, auf Veränderungen in seiner Umwelt oder in den Eingabedaten flexibel und effizient zu reagieren	Eigenschaft eines AI Systems, die Fähigkeit, auf Veränderungen in seiner Umwelt oder in den Eingabedaten flexibel und effizient zu reagieren		Privacy	Personal Data Protection	Fundamental Rights	Author's Personal Rights	Personalization	Nudging	Framing	Controllability	Usability		Non-Discrimination	Bias	Fairness		Secrecy	Factuality	Marking	Copyright	Accountability	Reversibility	Representa-tiveness		Accessibility	Workforce Exploitation	Perspectivity		Resources	Energy	Sustainability
Design & Concept	Neue Erwartungshaltung an AI Systeme auf Basis interner Erfahrungen und Erwartungshaltungen der Stakeholder der Unternehmen	Fähigkeit eines AI Systems, eingesetzte und neu hinzugefügte Daten und Informationen zu verarbeiten und eine sinnvolle Reaktion oder Aktion zu generieren	Fähigkeit eines AI Systems, eigenständig und nach festgelegten Regeln zu handeln, ohne dass ein Mensch eingreifen muss	Fähigkeit eines AI Systems, Aufgaben, die bisher manuell durchgeführt wurden, ohne weitere externe Unterstützung, insbesondere menschliches Eingreifen, vollständig durchzuführen	Fähigkeit eines AI Systems, auf verschiedenen Ebenen, digital und physisch, mit anderen AI Systemen zusammenzuarbeiten	Fähigkeit eines AI Systems, sich via Daten und Zwecke und oder Schnittstellen mit anderen AI Systemen oder Menschen zu verbinden	Fähigkeit eines AI Systems, sich via Daten und Zwecke und oder Schnittstellen mit anderen AI Systemen oder Menschen zu verbinden, ohne dass dafür die Bereitstellung/Fähigkeit mit physischen Maschinen nötig ist	Fähigkeit eines AI Systems, seine regulären Funktionen, eine gewünschte Performance und eine sichere Zustandsgrenze innerhalb seiner festgelegten Grenzen zu halten	Fähigkeit eines AI Systems, die mit der Performance oder Sicherheit eines AI Systems verbundenen Risiken, die über den akzeptablen Rahmen hinausgehen, zu vermeiden		Unbreachability	Fähigkeit eines AI Systems gegen externe und interne Angriffe, die darauf abzielen, Daten oder Algorithmen zu manipulieren, die über eine Schnittstelle des AI Systems oder auf andere Weise auf seine Komponenten zugreifen können, widerstandsfähig zu sein	Fähigkeit eines AI Systems, Daten, Algorithmen und Modelle vor unbefugtem Zugriff, Manipulationen, die über den physischen Raum stattfinden	Fähigkeit eines AI Systems, eine hohe Leistungsfähigkeit auch bei veränderten Eingaben oder Umgebungsbedingungen aufrechtzuerhalten		Explainability	Eigenschaft eines AI Systems, die Entscheidungen und Prozesse, die zu einer bestimmten Ausgabe oder Handlung geführt haben, nachvollziehbar und verständlich für Menschen zu machen	Eigenschaft eines AI Systems, die Herkunft und den Lebenszyklus von Daten, Algorithmen und Modellen nachvollziehbar zu machen	Eigenschaft eines AI Systems, seine Funktionsweise, Entscheidungsfindung und Datenverarbeitung offen und nachvollziehbar zu machen	Fähigkeit, bei gleicher oder ähnlicher Eingabe ähnliche Ergebnisse zu liefern	Eigenschaft eines AI Systems, die Fähigkeit, auf Veränderungen in seiner Umwelt oder in den Eingabedaten flexibel und effizient zu reagieren	Eigenschaft eines AI Systems, die Fähigkeit, auf Veränderungen in seiner Umwelt oder in den Eingabedaten flexibel und effizient zu reagieren		Privacy	Personal Data Protection	Fundamental Rights	Author's Personal Rights	Personalization	Nudging	Framing	Controllability	Usability		Non-Discrimination	Bias	Fairness		Secrecy	Factuality	Marking	Copyright	Accountability	Reversibility	Representa-tiveness		Accessibility	Workforce Exploitation	Perspectivity		Resources	Energy	Sustainability
Development	Neue Erwartungshaltung an AI Systeme auf Basis interner Erfahrungen und Erwartungshaltungen der Stakeholder der Unternehmen	Fähigkeit eines AI Systems, eingesetzte und neu hinzugefügte Daten und Informationen zu verarbeiten und eine sinnvolle Reaktion oder Aktion zu generieren	Fähigkeit eines AI Systems, eigenständig und nach festgelegten Regeln zu handeln, ohne dass ein Mensch eingreifen muss	Fähigkeit eines AI Systems, Aufgaben, die bisher manuell durchgeführt wurden, ohne weitere externe Unterstützung, insbesondere menschliches Eingreifen, vollständig durchzuführen	Fähigkeit eines AI Systems, auf verschiedenen Ebenen, digital und physisch, mit anderen AI Systemen zusammenzuarbeiten	Fähigkeit eines AI Systems, sich via Daten und Zwecke und oder Schnittstellen mit anderen AI Systemen oder Menschen zu verbinden	Fähigkeit eines AI Systems, sich via Daten und Zwecke und oder Schnittstellen mit anderen AI Systemen oder Menschen zu verbinden, ohne dass dafür die Bereitstellung/Fähigkeit mit physischen Maschinen nötig ist	Fähigkeit eines AI Systems, seine regulären Funktionen, eine gewünschte Performance und eine sichere Zustandsgrenze innerhalb seiner festgelegten Grenzen zu halten	Fähigkeit eines AI Systems, die mit der Performance oder Sicherheit eines AI Systems verbundenen Risiken, die über den akzeptablen Rahmen hinausgehen, zu vermeiden		Unbreachability	Fähigkeit eines AI Systems gegen externe und interne Angriffe, die darauf abzielen, Daten oder Algorithmen zu manipulieren, die über eine Schnittstelle des AI Systems oder auf andere Weise auf seine Komponenten zugreifen können, widerstandsfähig zu sein	Fähigkeit eines AI Systems, Daten, Algorithmen und Modelle vor unbefugtem Zugriff, Manipulationen, die über den physischen Raum stattfinden	Fähigkeit eines AI Systems, eine hohe Leistungsfähigkeit auch bei veränderten Eingaben oder Umgebungsbedingungen aufrechtzuerhalten		Explainability	Eigenschaft eines AI Systems, die Entscheidungen und Prozesse, die zu einer bestimmten Ausgabe oder Handlung geführt haben, nachvollziehbar und verständlich für Menschen zu machen	Eigenschaft eines AI Systems, die Herkunft und den Lebenszyklus von Daten, Algorithmen und Modellen nachvollziehbar zu machen	Eigenschaft eines AI Systems, seine Funktionsweise, Entscheidungsfindung und Datenverarbeitung offen und nachvollziehbar zu machen	Fähigkeit, bei gleicher oder ähnlicher Eingabe ähnliche Ergebnisse zu liefern	Eigenschaft eines AI Systems, die Fähigkeit, auf Veränderungen in seiner Umwelt oder in den Eingabedaten flexibel und effizient zu reagieren	Eigenschaft eines AI Systems, die Fähigkeit, auf Veränderungen in seiner Umwelt oder in den Eingabedaten flexibel und effizient zu reagieren		Privacy	Personal Data Protection	Fundamental Rights	Author's Personal Rights	Personalization	Nudging	Framing	Controllability	Usability		Non-Discrimination	Bias	Fairness		Secrecy	Factuality	Marking	Copyright	Accountability	Reversibility	Representa-tiveness		Accessibility	Workforce Exploitation	Perspectivity		Resources	Energy	Sustainability
Verification & Validation	Neue Erwartungshaltung an AI Systeme auf Basis interner Erfahrungen und Erwartungshaltungen der Stakeholder der Unternehmen	Fähigkeit eines AI Systems, eingesetzte und neu hinzugefügte Daten und Informationen zu verarbeiten und eine sinnvolle Reaktion oder Aktion zu generieren	Fähigkeit eines AI Systems, eigenständig und nach festgelegten Regeln zu handeln, ohne dass ein Mensch eingreifen muss	Fähigkeit eines AI Systems, Aufgaben, die bisher manuell durchgeführt wurden, ohne weitere externe Unterstützung, insbesondere menschliches Eingreifen, vollständig durchzuführen	Fähigkeit eines AI Systems, auf verschiedenen Ebenen, digital und physisch, mit anderen AI Systemen zusammenzuarbeiten	Fähigkeit eines AI Systems, sich via Daten und Zwecke und oder Schnittstellen mit anderen AI Systemen oder Menschen zu verbinden	Fähigkeit eines AI Systems, sich via Daten und Zwecke und oder Schnittstellen mit anderen AI Systemen oder Menschen zu verbinden, ohne dass dafür die Bereitstellung/Fähigkeit mit physischen Maschinen nötig ist	Fähigkeit eines AI Systems, seine regulären Funktionen, eine gewünschte Performance und eine sichere Zustandsgrenze innerhalb seiner festgelegten Grenzen zu halten	Fähigkeit eines AI Systems, die mit der Performance oder Sicherheit eines AI Systems verbundenen Risiken, die über den akzeptablen Rahmen hinausgehen, zu vermeiden		Unbreachability	Fähigkeit eines AI Systems gegen externe und interne Angriffe, die darauf abzielen, Daten oder Algorithmen zu manipulieren, die über eine Schnittstelle des AI Systems oder auf andere Weise auf seine Komponenten zugreifen können, widerstandsfähig zu sein	Fähigkeit eines AI Systems, Daten, Algorithmen und Modelle vor unbefugtem Zugriff, Manipulationen, die über den physischen Raum stattfinden	Fähigkeit eines AI Systems, eine hohe Leistungsfähigkeit auch bei veränderten Eingaben oder Umgebungsbedingungen aufrechtzuerhalten		Explainability	Eigenschaft eines AI Systems, die Entscheidungen und Prozesse, die zu einer bestimmten Ausgabe oder Handlung geführt haben, nachvollziehbar und verständlich für Menschen zu machen	Eigenschaft eines AI Systems, die Herkunft und den Lebenszyklus von Daten, Algorithmen und Modellen nachvollziehbar zu machen	Eigenschaft eines AI Systems, seine Funktionsweise, Entscheidungsfindung und Datenverarbeitung offen und nachvollziehbar zu machen	Fähigkeit, bei gleicher oder ähnlicher Eingabe ähnliche Ergebnisse zu liefern	Eigenschaft eines AI Systems, die Fähigkeit, auf Veränderungen in seiner Umwelt oder in den Eingabedaten flexibel und effizient zu reagieren	Eigenschaft eines AI Systems, die Fähigkeit, auf Veränderungen in seiner Umwelt oder in den Eingabedaten flexibel und effizient zu reagieren		Privacy	Personal Data Protection	Fundamental Rights	Author's Personal Rights	Personalization	Nudging	Framing	Controllability	Usability		Non-Discrimination	Bias	Fairness		Secrecy	Factuality	Marking	Copyright	Accountability	Reversibility	Representa-tiveness		Accessibility	Workforce Exploitation	Perspectivity		Resources	Energy	Sustainability
Deployment	Neue Erwartungshaltung an AI Systeme auf Basis interner Erfahrungen und Erwartungshaltungen der Stakeholder der Unternehmen	Fähigkeit eines AI Systems, eingesetzte und neu hinzugefügte Daten und Informationen zu verarbeiten und eine sinnvolle Reaktion oder Aktion zu generieren	Fähigkeit eines AI Systems, eigenständig und nach festgelegten Regeln zu handeln, ohne dass ein Mensch eingreifen muss	Fähigkeit eines AI Systems, Aufgaben, die bisher manuell durchgeführt wurden, ohne weitere externe Unterstützung, insbesondere menschliches Eingreifen, vollständig durchzuführen	Fähigkeit eines AI Systems, auf verschiedenen Ebenen, digital und physisch, mit anderen AI Systemen zusammenzuarbeiten	Fähigkeit eines AI Systems, sich via Daten und Zwecke und oder Schnittstellen mit anderen AI Systemen oder Menschen zu verbinden	Fähigkeit eines AI Systems, sich via Daten und Zwecke und oder Schnittstellen mit anderen AI Systemen oder Menschen zu verbinden, ohne dass dafür die Bereitstellung/Fähigkeit mit physischen Maschinen nötig ist	Fähigkeit eines AI Systems, seine regulären Funktionen, eine gewünschte Performance und eine sichere Zustandsgrenze innerhalb seiner festgelegten Grenzen zu halten	Fähigkeit eines AI Systems, die mit der Performance oder Sicherheit eines AI Systems verbundenen Risiken, die über den akzeptablen Rahmen hinausgehen, zu vermeiden		Unbreachability	Fähigkeit eines AI Systems gegen externe und interne Angriffe, die darauf abzielen, Daten oder Algorithmen zu manipulieren, die über eine Schnittstelle des AI Systems oder auf andere Weise auf seine Komponenten zugreifen können, widerstandsfähig zu sein	Fähigkeit eines AI Systems, Daten, Algorithmen und Modelle vor unbefugtem Zugriff, Manipulationen, die über den physischen Raum stattfinden	Fähigkeit eines AI Systems, eine hohe Leistungsfähigkeit auch bei veränderten Eingaben oder Umgebungsbedingungen aufrechtzuerhalten		Explainability	Eigenschaft eines AI Systems, die Entscheidungen und Prozesse, die zu einer bestimmten Ausgabe oder Handlung geführt haben, nachvollziehbar und verständlich für Menschen zu machen	Eigenschaft eines AI Systems, die Herkunft und den Lebenszyklus von Daten, Algorithmen und Modellen nachvollziehbar zu machen	Eigenschaft eines AI Systems, seine Funktionsweise, Entscheidungsfindung und Datenverarbeitung offen und nachvollziehbar zu machen	Fähigkeit, bei gleicher oder ähnlicher Eingabe ähnliche Ergebnisse zu liefern	Eigenschaft eines AI Systems, die Fähigkeit, auf Veränderungen in seiner Umwelt oder in den Eingabedaten flexibel und effizient zu reagieren	Eigenschaft eines AI Systems, die Fähigkeit, auf Veränderungen in seiner Umwelt oder in den Eingabedaten flexibel und effizient zu reagieren		Privacy	Personal Data Protection	Fundamental Rights	Author's Personal Rights	Personalization	Nudging	Framing	Controllability	Usability		Non-Discrimination	Bias	Fairness		Secrecy	Factuality	Marking	Copyright	Accountability	Reversibility	Representa-tiveness		Accessibility	Workforce Exploitation	Perspectivity		Resources	Energy	Sustainability
Operation	Neue Erwartungshaltung an AI Systeme auf Basis interner Erfahrungen und Erwartungshaltungen der Stakeholder der Unternehmen	Fähigkeit eines AI Systems, eingesetzte und neu hinzugefügte Daten und Informationen zu verarbeiten und eine sinnvolle Reaktion oder Aktion zu generieren	Fähigkeit eines AI Systems, eigenständig und nach festgelegten Regeln zu handeln, ohne dass ein Mensch eingreifen muss	Fähigkeit eines AI Systems, Aufgaben, die bisher manuell durchgeführt wurden, ohne weitere externe Unterstützung, insbesondere menschliches Eingreifen, vollständig durchzuführen	Fähigkeit eines AI Systems, auf verschiedenen Ebenen, digital und physisch, mit anderen AI Systemen zusammenzuarbeiten	Fähigkeit eines AI Systems, sich via Daten und Zwecke und oder Schnittstellen mit anderen AI Systemen oder Menschen zu verbinden	Fähigkeit eines AI Systems, sich via Daten und Zwecke und oder Schnittstellen mit anderen AI Systemen oder Menschen zu verbinden, ohne dass dafür die Bereitstellung/Fähigkeit mit physischen Maschinen nötig ist	Fähigkeit eines AI Systems, seine regulären Funktionen, eine gewünschte Performance und eine sichere Zustandsgrenze innerhalb seiner festgelegten Grenzen zu halten	Fähigkeit eines AI Systems, die mit der Performance oder Sicherheit eines AI Systems verbundenen Risiken, die über den akzeptablen Rahmen hinausgehen, zu vermeiden		Unbreachability	Fähigkeit eines AI Systems gegen externe und interne Angriffe, die darauf abzielen, Daten oder Algorithmen zu manipulieren, die über eine Schnittstelle des AI Systems oder auf andere Weise auf seine Komponenten zugreifen können, widerstandsfähig zu sein	Fähigkeit eines AI Systems, Daten, Algorithmen und Modelle vor unbefugtem Zugriff, Manipulationen, die über den physischen Raum stattfinden	Fähigkeit eines AI Systems, eine hohe Leistungsfähigkeit auch bei veränderten Eingaben oder Umgebungsbedingungen aufrechtzuerhalten		Explainability	Eigenschaft eines AI Systems, die Entscheidungen und Prozesse, die zu einer bestimmten Ausgabe oder Handlung geführt haben, nachvollziehbar und verständlich für Menschen zu machen	Eigenschaft eines AI Systems, die Herkunft und den Lebenszyklus von Daten, Algorithmen und Modellen nachvollziehbar zu machen	Eigenschaft eines AI Systems, seine Funktionsweise, Entscheidungsfindung und Datenverarbeitung offen und nachvollziehbar zu machen	Fähigkeit, bei gleicher oder ähnlicher Eingabe ähnliche Ergebnisse zu liefern	Eigenschaft eines AI Systems, die Fähigkeit, auf Veränderungen in seiner Umwelt oder in den Eingabedaten flexibel und effizient zu reagieren	Eigenschaft eines AI Systems, die Fähigkeit, auf Veränderungen in seiner Umwelt oder in den Eingabedaten flexibel und effizient zu reagieren		Privacy	Personal Data Protection	Fundamental Rights	Author's Personal Rights	Personalization	Nudging	Framing	Controllability	Usability		Non-Discrimination	Bias	Fairness		Secrecy	Factuality	Marking	Copyright	Accountability	Reversibility	Representa-tiveness		Accessibility	Workforce Exploitation	Perspectivity		Resources	Energy	Sustainability
Monitoring	Neue Erwartungshaltung an AI Systeme auf Basis interner Erfahrungen und Erwartungshaltungen der Stakeholder der Unternehmen	Fähigkeit eines AI Systems, eingesetzte und neu hinzugefügte Daten und Informationen zu verarbeiten und eine sinnvolle Reaktion oder Aktion zu generieren	Fähigkeit eines AI Systems, eigenständig und nach festgelegten Regeln zu handeln, ohne dass ein Mensch eingreifen muss	Fähigkeit eines AI Systems, Aufgaben, die bisher manuell durchgeführt wurden, ohne weitere externe Unterstützung, insbesondere menschliches Eingreifen, vollständig durchzuführen	Fähigkeit eines AI Systems, auf verschiedenen Ebenen, digital und physisch, mit anderen AI Systemen zusammenzuarbeiten	Fähigkeit eines AI Systems, sich via Daten und Zwecke und oder Schnittstellen mit anderen AI Systemen oder Menschen zu verbinden	Fähigkeit eines AI Systems, sich via Daten und Zwecke und oder Schnittstellen mit anderen AI Systemen oder Menschen zu verbinden, ohne dass dafür die Bereitstellung/Fähigkeit mit physischen Maschinen nötig ist	Fähigkeit eines AI Systems, seine regulären Funktionen, eine gewünschte Performance und eine sichere Zustandsgrenze innerhalb seiner festgelegten Grenzen zu halten	Fähigkeit eines AI Systems, die mit der Performance oder Sicherheit eines AI Systems verbundenen Risiken, die über den akzeptablen Rahmen hinausgehen, zu vermeiden		Unbreachability	Fähigkeit eines AI Systems gegen externe und interne Angriffe, die darauf abzielen, Daten oder Algorithmen zu manipulieren, die über eine Schnittstelle des AI Systems oder auf andere Weise auf seine Komponenten zugreifen können, widerstandsfähig zu sein	Fähigkeit eines AI Systems, Daten, Algorithmen und Modelle vor unbefugtem Zugriff, Manipulationen, die über den physischen Raum stattfinden	Fähigkeit eines AI Systems, eine hohe Leistungsfähigkeit auch bei veränderten Eingaben oder Umgebungsbedingungen aufrechtzuerhalten		Explainability	Eigenschaft eines AI Systems, die Entscheidungen und Prozesse, die zu einer bestimmten Ausgabe oder Handlung geführt haben, nachvollziehbar und verständlich für Menschen zu machen	Eigenschaft eines AI Systems, die Herkunft und den Lebenszyklus von Daten, Algorithmen und Modellen nachvollziehbar zu machen	Eigenschaft eines AI Systems, seine Funktionsweise, Entscheidungsfindung und Datenverarbeitung offen und nachvollziehbar zu machen	Fähigkeit, bei gleicher oder ähnlicher Eingabe ähnliche Ergebnisse zu liefern	Eigenschaft eines AI Systems, die Fähigkeit, auf Veränderungen in seiner Umwelt oder in den Eingabedaten flexibel und effizient zu reagieren	Eigenschaft eines AI Systems, die Fähigkeit, auf Veränderungen in seiner Umwelt oder in den Eingabedaten flexibel und effizient zu reagieren		Privacy	Personal Data Protection	Fundamental Rights	Author's Personal Rights	Personalization	Nudging	Framing	Controllability	Usability		Non-Discrimination	Bias	Fairness		Secrecy	Factuality	Marking	Copyright	Accountability	Reversibility	Representa-tiveness		Accessibility	Workforce Exploitation	Perspectivity		Resources	Energy	Sustainability
Retirement	Neue Erwartungshaltung an AI Systeme auf Basis interner Erfahrungen und Erwartungshaltungen der Stakeholder der Unternehmen	Fähigkeit eines AI Systems, eingesetzte und neu hinzugefügte Daten und Informationen zu verarbeiten und eine sinnvolle Reaktion oder Aktion zu generieren	Fähigkeit eines AI Systems, eigenständig und nach festgelegten Regeln zu handeln, ohne dass ein Mensch eingreifen muss	Fähigkeit eines AI Systems, Aufgaben, die bisher manuell durchgeführt wurden, ohne weitere externe Unterstützung, insbesondere menschliches Eingreifen, vollständig durchzuführen	Fähigkeit eines AI Systems, auf verschiedenen Ebenen, digital und physisch, mit anderen AI Systemen zusammenzuarbeiten	Fähigkeit eines AI Systems, sich via Daten und Zwecke und oder Schnittstellen mit anderen AI Systemen oder Menschen zu verbinden	Fähigkeit eines AI Systems, sich via Daten und Zwecke und oder Schnittstellen mit anderen AI Systemen oder Menschen zu verbinden, ohne dass dafür die Bereitstellung/Fähigkeit mit physischen Maschinen nötig ist	Fähigkeit eines AI Systems, seine regulären Funktionen, eine gewünschte Performance und eine sichere Zustandsgrenze innerhalb seiner festgelegten Grenzen zu halten	Fähigkeit eines AI Systems, die mit der Performance oder Sicherheit eines AI Systems verbundenen Risiken, die über den akzeptablen Rahmen hinausgehen, zu vermeiden		Unbreachability	Fähigkeit eines AI Systems gegen externe und interne Angriffe, die darauf abzielen, Daten oder Algorithmen zu manipulieren, die über eine Schnittstelle des AI Systems oder auf andere Weise auf seine Komponenten zugreifen können, widerstandsfähig zu sein	Fähigkeit eines AI Systems, Daten, Algorithmen und Modelle vor unbefugtem Zugriff, Manipulationen, die über den physischen Raum stattfinden	Fähigkeit eines AI Systems, eine hohe Leistungsfähigkeit auch bei veränderten Eingaben oder Umgebungsbedingungen aufrechtzuerhalten		Explainability	Eigenschaft eines AI Systems, die Entscheidungen und Prozesse, die zu einer bestimmten Ausgabe oder Handlung geführt haben, nachvollziehbar und verständlich für Menschen zu machen	Eigenschaft eines AI Systems, die Herkunft und den Lebenszyklus von Daten, Algorithmen und Modellen nachvollziehbar zu machen	Eigenschaft eines AI Systems, seine Funktionsweise, Entscheidungsfindung und Datenverarbeitung offen und nachvollziehbar zu machen	Fähigkeit, bei gleicher oder ähnlicher Eingabe ähnliche Ergebnisse zu liefern	Eigenschaft eines AI Systems, die Fähigkeit, auf Veränderungen in seiner Umwelt oder in den Eingabedaten flexibel und effizient zu reagieren	Eigenschaft eines AI Systems, die Fähigkeit, auf Veränderungen in seiner Umwelt oder in den Eingabedaten flexibel und effizient zu reagieren		Privacy	Personal Data Protection	Fundamental Rights	Author's Personal Rights	Personalization	Nudging	Framing	Controllability	Usability		Non-Discrimination	Bias	Fairness		Secrecy	Factuality	Marking	Copyright	Accountability	Reversibility	Representa-tiveness		Accessibility	Workforce Exploitation	Perspectivity		Resources	Energy	Sustainability