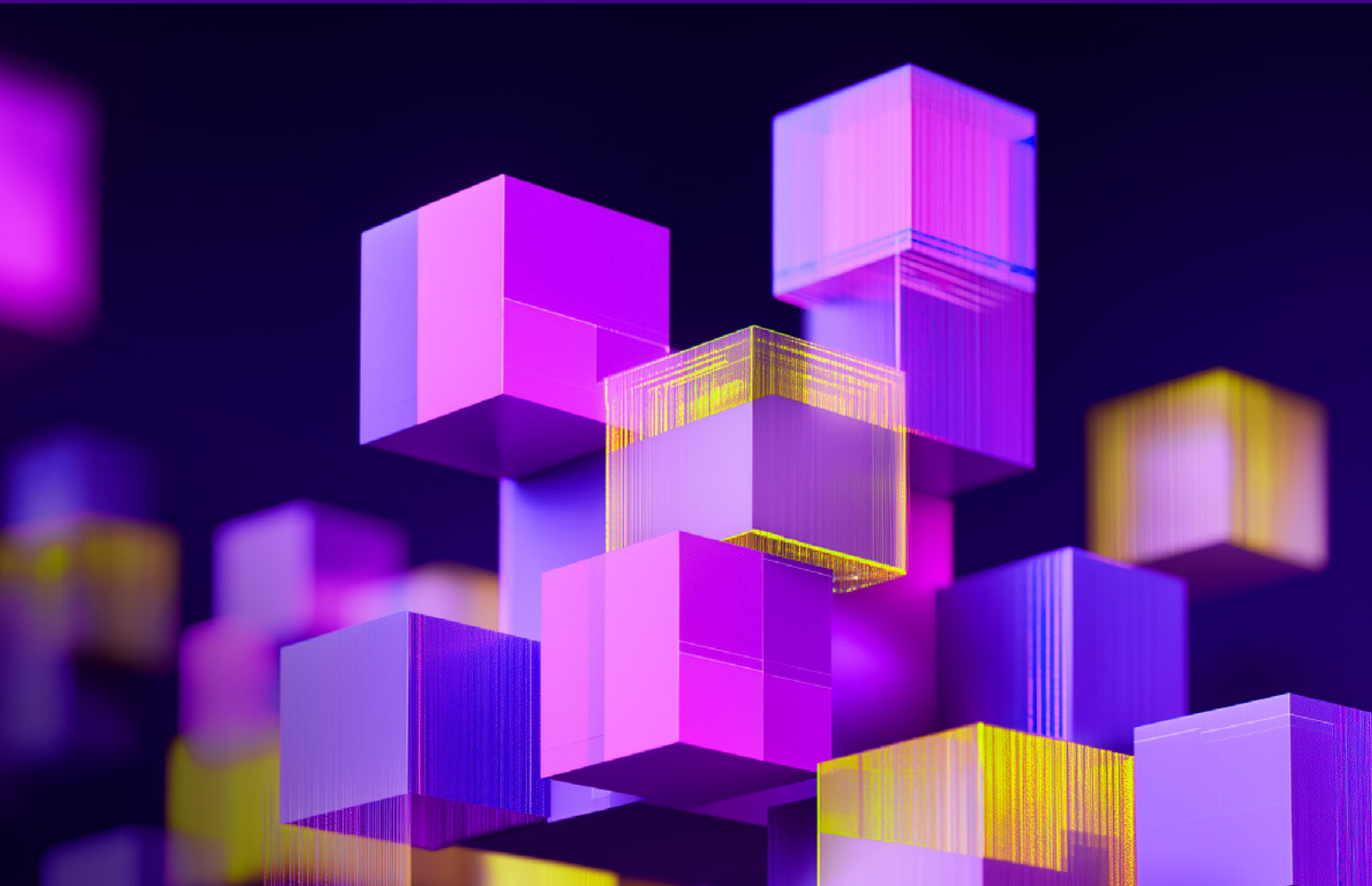


MISSION KI

November 2025

Qualitätsstandard für Niedrigrisiko-KI



Hinweis

Teile der im MISSION KI Qualitätsstandard enthaltenen Indikatoren basieren auf der „Specification for the assessment of the trustworthiness of AI systems“, die gemeinsam von IEEE, VDE, Positive AI und IRT SystemX entwickelt wurde.



Vorwort

Die Förderung vertrauenswürdiger Künstlicher Intelligenz (KI) ist ein zentraler Faktor für die nachhaltige Entwicklung und Akzeptanz von KI-Technologien. Vor diesem Hintergrund wurde **MISSION KI** – Nationale Initiative für Künstliche Intelligenz und Datenökonomie ins Leben gerufen. Die Initiative wird vom Bundesministerium für Digitalisierung und Staatsmodernisierung (BMDS) gefördert und von acatech – Deutsche Akademie der Technikwissenschaften umgesetzt. Ziel von **MISSION KI** ist es, Deutschlands digitale Wettbewerbsfähigkeit zu stärken. **MISSION KI** ist ein Hebelprojekt der Digitalstrategie der Bundesregierung.

Ein zentrales Element dieser Initiative ist die Entwicklung eines freiwilligen Qualitätsstandards für Niedrigrisiko-KI¹, der einen transparenten, standardisierten und vergleichbaren Nachweis der Qualität von KI-Systemen ermöglicht. Der Standard legt die Rahmenbedingungen und das Prüfverfahren für den Qualitätsnachweis fest und erläutert die Zusammenhänge zwischen den zentralen Komponenten der Prüfung.

Die Erarbeitung des **MISSION KI** Qualitätsstandards für Künstliche Intelligenz erfolgte in enger Zusammenarbeit einer Projektgemeinschaft bestehend aus PwC Deutschland (Leitung), dem AI Quality & Testing Hub, dem Fraunhofer IAIS, dem TÜV AI.Lab und dem VDE.

¹ Für eine bessere Lesbarkeit werden im vorliegenden Dokument die Begriffe MISSION KI Qualitätsstandard bzw. Qualitätsstandard als Kurzversionen zu MISSION KI Qualitätsstandard für Niedrigrisiko-KI mit synonyme Bedeutung verwendet.

Inhalt

Vorwort	3
Abbildungsverzeichnis	5
Einleitung	6
Einführung	7
1. Anwendungsbereich des Prüfverfahrens	10
1.1 KI-System als Prüfgegenstand	10
1.2 Qualitätsdimension als Strukturelement der Prüfung	12
1.3 Prüfaussage	13
1.4 Gültigkeit	13
2. Prüftiefe	15
3. Durchführung der Prüfung	17
3.1 Anwendungsfallbeschreibung	17
3.2 Schutzbedarfsanalyse	18
3.3 Prüfanforderungen	19
3.4 Bereitstellen von Evidenzen	20
3.5 Validierung der Evidenzen	22
3.6 Gesamtbewertung	23
3.7 Prüfbericht	23
3.8 Überwachen der Gültigkeit	24
4. Anhänge	25
4.1 Glossar	25
4.2 Anwendungsfallbeschreibungsvorlage	31
4.3 Schutzbedarfsanalyse	33
4.4 Prüfkatalog	39
4.5 Prozessablauf der Gesamtbewertung	40
4.6 Prüfmethodensammlung	44
4.7 Prüfberichtsvorlage	45
Autorenverzeichnis	60

Abbildungsverzeichnis

Abbildung 1 Abstrahierter Aufbau und Abgrenzung eines KI-Systems. Quelle: Fraunhofer IAIS Prüfkatalog (2021)	11
Abbildung 2 Übersicht der Prüfschritte	17
Abbildung 3 Adaptierte VCIO-Systematik der Prüfanforderungen	19
Abbildung 4 Übersicht der Qualitätsdimensionen und Kriterien	20
Abbildung 5 Schutzbedarfe als implizite Mindeststufen in der Gesamtbewertung, die den Bestätigungsgrad des Mindeststandards für ein KI-System im konkreten Anwendungsfall festlegen.	23
Abbildung 6 Prozessablauf der Gesamtbewertung	40
Abbildung 7 Darstellung des grundlegenden Ansatzes zur Aggregation der Einstufung von Indikatoren durch Observablen auf die Kriterienebene.	41
Abbildung 8 Beispielhafte Darstellung der Durchschnittsbildung auf Basis der Indikatoreneinstufung.	41
Abbildung 9 Detaillierte Beschreibung der Bewertungsaggregation von Indikatoren- auf Kriterienebene.	42
Abbildung 10 Einstufung der Kriterien anhand des gerundeten numerischen Wertes.	42
Abbildung 11 Schutzbedarfe als implizite Mindeststufen in der Gesamtbewertung, die die Prüftiefe des Mindeststandards für ein KI-System in mit konkreten Anwendungsfall festlegen.	43

Einleitung

Der **MISSION KI** Qualitätsstandard richtet sich an Anbieter von KI-Systemen, insbesondere Start-ups sowie kleine und mittlere Unternehmen (KMU), deren Anwendungen unterhalb der Hochrisiko-Schwelle der Europäischen KI-Verordnung liegen. Diese Systeme prägen eine Vielzahl wirtschaftlich und gesellschaftlich relevanter Prozesse, von der Produktion über den Kundenservice bis zur Verwaltung. Für sie besteht ein besonderer Bedarf an praxisnahen und überprüfbaren Kriterien, um Qualität und Vertrauenswürdigkeit nachvollziehbar zu belegen.

Der Standard bietet dafür ein strukturiertes Vorgehen zur internen Bewertung, Dokumentation und externen Kommunikation von Qualitätsmaßnahmen. Er unterstützt Organisationen dabei,

- ihre Prozesse und Verantwortlichkeiten im Umgang mit KI zu systematisieren,
- Transparenz gegenüber Kunden, Partnern und Aufsichtsbehörden zu schaffen,
- und sich auf künftige regulatorische Anforderungen vorzubereiten.

Neben KI-Anbietern profitieren auch Kunden, Investoren und öffentliche Auftraggeber von der Anwendung des Standards: Einheitliche Bewertungskriterien schaffen Vergleichbarkeit, fördern Vertrauen und erleichtern fundierte Entscheidungen in Beschaffungs- und Bewertungsprozessen.

Der **MISSION KI** Qualitätsstandard trägt dazu bei, Qualität als Wettbewerbsvorteil zu etablieren und die breite Akzeptanz vertrauenswürdiger KI-Anwendungen in Wirtschaft und Verwaltung zu stärken.

Einführung

Zweck und Verwendung

Dieses Dokument beschreibt den **MISSION KI** Qualitätsstandard für Niedrigrisiko-KI und legt die Kriterien, das Prüfverfahren sowie die Anforderungen an Evidenzen und Validierung fest. Der Standard dient Organisationen als Grundlage für Selbstprüfungen ihrer KI-Systeme und kann optional durch eine Validierung externer Prüfstellen ergänzt werden.

Aufbau des Dokuments

Kapitel 1 definiert Anwendungsbereich und Rahmenbedingungen der Prüfung. Es präzisiert, welche Systemteile gemäß Zweckbestimmung einzubeziehen sind, erläutert die Strukturelemente (Qualitätsdimensionen, Kriterien, Indikatoren, Observablen) und die daraus resultierende Prüfaussage.

Kapitel 2 erläutert das Konzept der Prüftiefe und beschreibt dessen Festlegung im Verfahren – einschließlich Anforderungen an berechnete/validierende Personen.

Kapitel 3 beschreibt die Durchführung der Prüfung: Anwendungsfallbeschreibung, Schutzbedarfsanalyse, Einstufung nach Prüfkatalog, Evidenzen/Tests, Validierung, Gesamtbewertung, Prüfbericht und Überwachung der Gültigkeit.

Die zur Umsetzung erforderlichen Details und Vorlagen (u. a. Anwendungsfallbeschreibung, Schutzbedarfsanalyse, Prüfkatalog, Prüfmethodensammlung, Glossar) befinden sich in den Anhängen.

Leitprinzipien des Standards

Der Qualitätsstandard folgt sechs Leitprinzipien, die nachfolgend erläutert und in Tabelle 1 definiert sind. Mit Inkrafttreten der Europäischen KI-Verordnung (KI-VO) unterliegen viele Anbieter bereits einer Konformitätsbewertung. Eine Prüfung nach dem **MISSION KI** Qualitätsstandard erfolgt demgegenüber freiwillig und primär aus wirtschaftlichem Interesse.

Ergänzend zur KI-VO richtet sich der Standard einerseits an Organisationen im nicht regulierten Bereich, andererseits an solche, die bereits regulatorisch betroffen sind oder es durch die Weiterentwicklung ihrer Systeme werden könnten und hierfür einen praxisnahen Ansatz zur Vorbereitung benötigen. Auch Organisationen, die freiwillige Vertrauenswürdigkeitsanforderungen erfüllen wollen, profitieren von der Anwendung.

Um Anforderungen aus Regulierung und Normenumfeld zu adressieren, ist der Qualitätsstandard anschlussfähig, d. h. kompatibel und widerspruchsfrei zur KI-VO sowie zu weiteren relevanten Richtlinien. Für die wirksame Anwendung im freiwilligen Bereich muss die Durchführung effizient sein – zeitlich, personell und finanziell überschaubar – und in einem angemessenen Verhältnis von Aufwand zu Nutzen stehen. Daher ist der Standard grundsätzlich als Selbstprüfung ausgelegt. Ein belastbares Qualitätsversprechen setzt belastbare Prüfergebnisse voraus: durch Nachweisbarkeit, Reproduzierbarkeit und Objektivität. Der gleichzeitige Anspruch auf Effizienz, Belastbarkeit und Vergleichbarkeit erfordert ein hohes Maß an Systematik.

Schließlich muss der Standard zugänglich sein, sodass Ergebnisse kommunizierbar, verständlich und nachprüfbar sind – und die Einhaltung des Qualitätsstandards zu einem Wettbewerbsvorteil für geprüfte Unternehmen wird.

Leitprinzipien des MISSION KI Qualitätsstandards

Leitprinzipien des MISSION KI Qualitätsstandards

Anschlussfähigkeit	<ul style="list-style-type: none"> • Es gibt eine Schnittmenge zwischen Qualitätsstandard und KI-VO bzw. sektoraler Regulierung wie der Medical Device Regulation (MDR) oder anderer Vorschriften/Standards. • Die Anforderungen des Qualitätsstandards sind kompatibel mit den Hochrisikoranforderungen aus Abschnitt 2 KI-VO. • Der Qualitätsstandard darf die Anforderungen der Vorschriften und Standards auch „übertreffen“.
Belastbarkeit	<ul style="list-style-type: none"> • Der Qualitätsstandard soll einen Mindeststandard darstellen, der ein angemessenes Maß an Qualität glaubwürdig signalisiert, dementsprechend sollen die Ergebnisse der Prüfung möglichst belastbar im Sinne einer entsprechenden Absicherung sein.
Effizienz	<ul style="list-style-type: none"> • Eine Durchführung der Prüfung nach dem Qualitätsstandard muss in Bezug auf die zeitlichen, personellen und finanziellen Aufwendungen überschaubar sein. • Eine größere Prüftiefe muss im guten Verhältnis zu den gelieferten Mehrwerten liegen. • Ein hohes Qualitätslevel soll aufrechterhalten werden.
Freiwilligkeit	<ul style="list-style-type: none"> • Eine Prüfung nach dem Qualitätsstandard erfolgt primär aus wirtschaftlichen Interessen ohne gesetzliche Vorschriften (z. B. verständliches, glaubwürdiges Signalisieren von Qualität). • Die Prüfanforderungen bieten auch der fordernden Zielgruppe konkrete Mehrwerte (d. h. bedarfsorientierte Interpretation von Qualität). • Die Prüfanforderungen bieten der liefernden Zielgruppe weitere konkrete Mehrwerte (z. B. teilweise Compliance mit KI-VO, Prozessverbesserung...).
Vergleichbarkeit	<ul style="list-style-type: none"> • Prüfungen auf Basis des Qualitätsstandards sollen untereinander vergleichbar sein. • Die Tests sollten einheitlich und replizierbar sein. • Das Prüfverfahren soll möglichst objektiv sein.
Zugänglichkeit	<ul style="list-style-type: none"> • Die Prüfaussage ist für alle Stakeholder verständlich. • Die Ergebnisse sind leicht kommunizierbar. • Die Prüfanforderungen sind konkret, greifbar, nachprüfbar und möglichst quantifizierbar.

Tabelle 1: Leitprinzipien des MISSION KI Qualitätsstandards

Grundlagen des Prüfverfahrens

Der Qualitätsstandard wurde unter Berücksichtigung der oben genannten Leitprinzipien entwickelt. Gleichzeitig existieren bereits etablierte Vorarbeiten wie die VDE SPEC 90012², der Fraunhofer IAIS KI-Prüfkatalog³ und die Joint Specification V1.0 for the Assessment of the Trustworthiness of AI Systems⁴, welche erprobte KI-Prüfansätze beschreiben und bei der Entwicklung des **MISSION KI Qualitätsstandards** berücksichtigt wurden.

2 VCIO based description of systems for AI trustworthiness, VDE SPEC 90012 V1.0 (en) characterisation, 2022. URL: <https://www.vde.com/resource/blob/2242194/a24b13db01773747e6b7bba4ce20ea60/vcio-based-description-of-systems-for-ai-trustworthiness-characterisationvde-spec-90012-v1-0--en--data.pdf> (Zugriff: 21.06.2025).

3 Poretschkin, M., et al., KI-Prüfkatalog: Ein Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz, Fraunhofer IAIS, 2021. URL: https://www.iais.fraunhofer.de/content/dam/iais/publikationen/studien-und-whitepaper/2021/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf (Zugriff: 21.06.2025).

4 Joint Specification V1.0 for the Assessment of the Trustworthiness of AI Systems, 2024 (Zugriff: 06.11.2025). Die Spezifikation ist unter trustalliance.ai nach vorheriger Anmeldung kostenfrei verfügbar.

Die im April 2022 vorgestellte Spezifikation VDE SPEC 90012 dient dazu, KI-spezifische Risiken eines gegebenen KI-Systems im Hinblick auf ethische, relevante Eigenschaften zu beurteilen. Der Katalog nutzt das VCIO-Modell, um die Erfüllung von vordefinierten Werten zu beurteilen und mit einem KI-Vertrauenslabel von „A“ am besten bis „D“ am schlechtesten zu visualisieren. Auch der Fraunhofer IAIS KI-Prüfkatalog (April 2021) zielt darauf ab, die Risiken einer KI-Anwendung im Rahmen von sechs Dimensionen der Vertrauenswürdigkeit zu überprüfen. Dabei wird ein risiko-basierter und Anwendungskontext-abhängiger Ansatz verfolgt, der mitunter konkrete technische Tests für die Mitigation von Risiken vorschlägt.

Das im **MISSION KI** Qualitätsstandard definierte Prüfverfahren konsolidiert die Stärken dieser drei Ansätze zu einem integrierten Vorgehensmodell. Im Sinne der Zugänglichkeit und Verständlichkeit werden aus der VDE SPEC 90012 die Prüfsystematik und die Grundlagen für die Bewertung nach VCIO-Modell übernommen.

Um eine Bewertung in Abhängigkeit von den Schutzbedarfen eines KI-Systems zu ermöglichen, wird das VCIO-Modell durch die Berücksichtigung der Zweckbestimmung eines KI-System erweitert. Dazu wird dem VCIO-Ansatz eine Schutzbedarfsanalyse (vgl. 3.2) ähnlich der des KI-Prüfkatalogs vorgeschaltet. Zusätzlich werden die Observablen (unterste Stufe des VCIO-Ansatzes) durch das Einholen von anwendungsbezogenen Evidenzen und technischen Tests (vgl. 3.4) erweitert, um zusätzlich die Belastbarkeit des Prüfergebnisses zu erhöhen.

Zur Konkretisierung der Werte (welche im Rahmen dieses Prüfstandards als Qualitätsdimensionen bezeichnet werden, vgl. 1.2), Kriterien, Indikatoren und Observablen werden die Inhalte aus VDE SPEC 90012 und KI-Prüfkatalog zusammengeführt.

1. Anwendungsbereich des Prüfverfahrens

Das Prüfverfahren ist auf KI-Systeme anwendbar, denen ein bestimmter Zweck zugewiesen ist (siehe 1.1). Der Betrachtungsumfang entspricht einer produktorientierten Prozessprüfung. Diese erhebt, welche konkreten Qualitätsmaßnahmen entlang der Entwicklung des KI-Systems ergriffen und welche Vorkehrungen etwa für die Qualitätssicherung während des Betriebs getroffen wurden (siehe 1.2).

Damit fokussiert sich der Anwendungsbereich typischerweise auf die Anbieterseite des KI-Systems („Anbieter“ bzw. engl. „Provider“ im Sinne der KI-VO). Das Prüfverfahren kann auch nachgelagert beim Betreiber (engl. „Deployer“) durchgeführt werden, sofern die erforderlichen Informationen bzw. Dokumentationen vorliegen.⁵

Als Prüfling wird stets die Organisation verstanden, die die Prüfung durchführt. Im Ergebnis der Prüfung wird das erreichte Level an ergriffenen Qualitäts- und Absicherungsmaßnahmen den Schutzbedarfen gegenübergestellt, die sich aus der Zweckbestimmung des KI-Systems ergeben (siehe 1.3). Die Gültigkeit der Prüfergebnisse ist dabei fest an die Zweckbestimmung des KI-Systems sowie die als Grundlage der Prüfung spezifizierte technische Umsetzung gebunden (siehe 1.4). Die genannten Aspekte des Anwendungsbereichs werden in den folgenden Abschnitten näher erläutert.

1.1 KI-System als Prüfgegenstand

Als Prüfgegenstand sind KI-Systeme zulässig, wobei KI-Systeme im Sinne der Europäischen KI-Verordnung gemeint sind (siehe Glossar). Für die Anwendbarkeit des Prüfverfahrens müssen beide Charakteristika eines KI-Systems erfüllt sein: die Implementierung eines KI-Verfahrens sowie eine eindeutig festgelegte Zweckbestimmung.

1.1.1 KI-System

Ein KI-System ist „ein maschinengestütztes System, das für einen in unterschiedlichem Grad autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können“ (Artikel 3, Absatz 1, KI-VO).

Im Rahmen des vorliegenden Qualitätsstandards wird ein „KI-System“ in technischer Hinsicht als ein funktionaler Zusammenschluss einer oder mehrerer KI-Komponente(n) und Nicht-KI-Komponenten im Hinblick auf eine spezifische Zweckbestimmung und einen konkreten Anwendungskontext verstanden. Ein zentrales Merkmal von KI ist das Ableiten bzw. Lernen aus Dateneingaben. Dieses kann durch eine oder mehrere KI-Komponenten realisiert werden, die beispielsweise auf maschinellen Lernverfahren (ML) basieren.

Ein KI-System kann entweder alleinstehen und direkt mit Nutzern über eine Schnittstelle interagieren (z. B. ein Übersetzungsmodell als Webapplikation), oder aber in größere Produkte bzw. Systeme integriert sein (z. B. Steuerungs- oder Überwachungselemente in Maschinen).

⁵ Hierbei ist anzumerken, dass Unternehmen, die ein extern bereitgestelltes KI-System einsetzen, etwa bereits durch die Zuweisung oder Änderung der Zweckbestimmung des KI-Systems sowie durch das Vornehmen signifikanter technischer Änderungen die Rolle des Anbieters gemäß der Europäischen KI-Verordnung erhalten. Siehe hierzu etwa Erwägungsgrund 84, KI-Verordnung.

Für die Durchführung und Gültigkeit der Prüfung ist es wesentlich, dass das KI-System als Prüfgegenstand klar definiert und, falls zutreffend, von anderen Systemen abgegrenzt ist.⁶ Dabei ist auf Konsistenz mit der Zweckbestimmung des KI-Systems (siehe 1.1.2) zu achten. Alle Komponenten, die zur Realisierung des bestimmten Zwecks erforderlich sind, sollten als Teil des KI-Systems verstanden werden und über spezifizierte Schnittstellen von dem umgebenden System abgegrenzt werden. Eine Hilfestellung zur technischen Abgrenzung bietet etwa Kapitel 2 des Fraunhofer IAIS Prüfkatalogs. Aus diesem ist Abbildung 1 entnommen, die ein KI-System mit einer ML-Komponente vereinfacht darstellt.

Ein KI-System zur Maschinenüberwachung könnte beispielsweise über Schnittstellen zu Sensoren und ein Steuerungselement verfügen. Die Sensoren messen die Maschinendaten und das Steuerungselement erhält als weitere Eingabe die Prognose des KI-Systems und leitet darauf basierend ggf. Aktionen wie das Abschalten der Maschine ein.

Alle Komponenten, die zur Erzeugung der Prognose basierend auf den Sensordaten benötigt werden (z. B. Datenvorverarbeitung, ML-Modell), könnten in diesem Fall zum KI-System gezählt werden. Alle weiteren Komponenten (z. B. ein Expertensystem, das auf die Prognosen reagiert, oder die Sensoren) werden, davon abgegrenzt, als weitere Systemumgebung definiert, die nicht Gegenstand der Prüfung ist.

Ferner ist für die Durchführung der Prüfung erforderlich, dass das KI-System gemäß der beschriebenen Spezifikation und Abgrenzung (siehe auch 3.1) in einem hinreichend hohen Reifegrad implementiert ist. Das Prüfverfahren ist nicht auf Prototypen anwendbar. Der Reifegrad des KI-Systems ist im Einzelfall durch den Prüfling zu beurteilen.

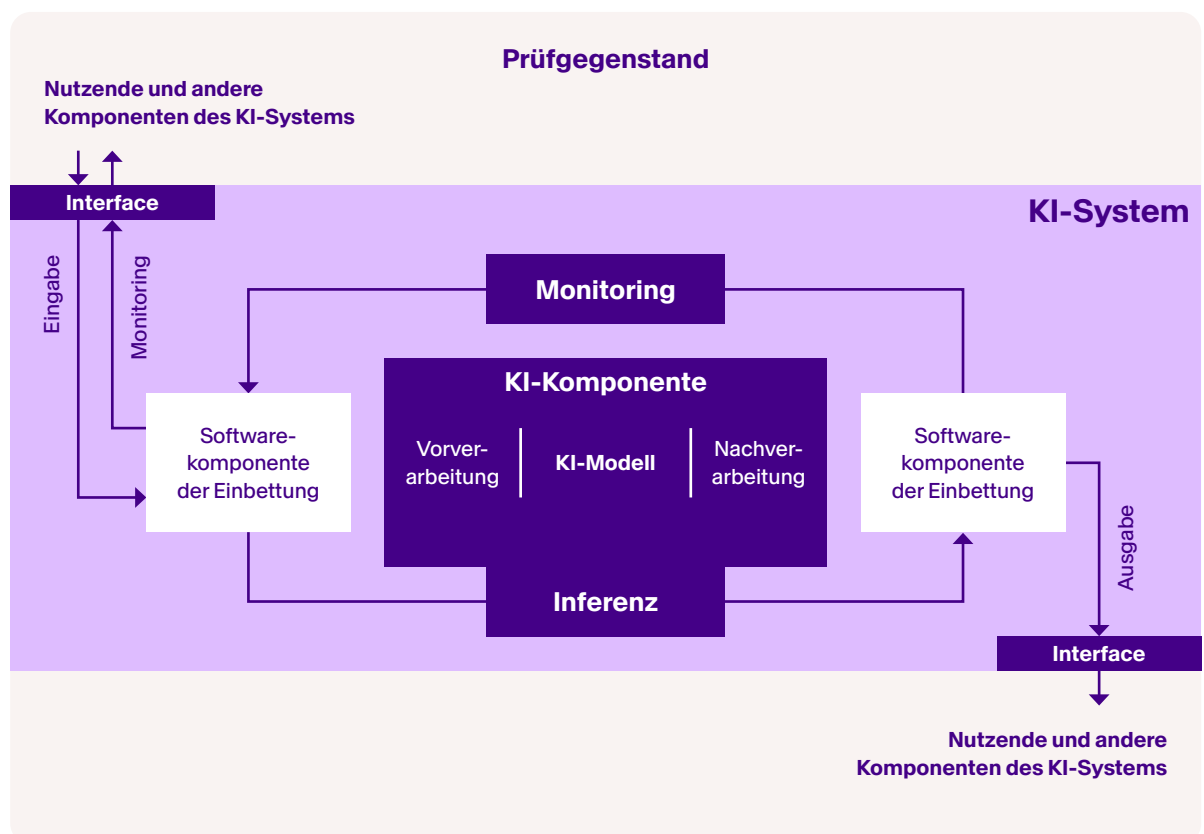


Abbildung 1: Abstrahierter Aufbau und Abgrenzung eines KI-Systems. Quelle: Fraunhofer IAIS Prüfkatalog (2021)

⁶ Der Aufbau des KI-Systems und die Abgrenzung des KI-Systems von anderen Systemen (z. B. durch Schnittstellen) ist zu Beginn der Prüfung unter anderem als Architekturdiagramm zu spezifizieren, siehe Abschnitt 4.1. Falls ein KI-System mehrere KI-Komponenten umfasst, müssen einzelne Prüfanforderungen ggf. je Komponente wiederholt angewendet werden.

1.1.2 Zweckbestimmung

Die Zweckbestimmung des KI-Systems ist „die Verwendung, für die ein KI-System laut Anbieter bestimmt ist, einschließlich der besonderen Umstände und Bedingungen für die Verwendung, entsprechend den vom Anbieter bereitgestellten Informationen in den Betriebsanleitungen, im Werbe- oder Verkaufsmaterial und in diesbezüglichen Erklärungen sowie in der technischen Dokumentation“ (siehe Artikel 3, Absatz 12, KI-Verordnung).

Der **MISSION KI** Qualitätsstandard legt den Fokus auf KI-Anbieter und die Bewertung eines KI-Systems auf produktbezogene Prozesse und Maßnahmen bis zum Zeitpunkt der Übergabe an den Betreiber. Typischerweise ist es deshalb nicht erforderlich (oder möglich, z. B. wenn Anbieter und Betreiber nicht identisch sind), dass die Zweckbestimmung im Detail auf die (spätere) Betriebsumgebung oder die Produktionsdaten eingeht.

Der Zweck sollte jedoch im Rahmen der Anwendungsfallbeschreibung (siehe 3.1) so präzise eingegrenzt werden, dass eine ausreichende Qualitätssicherung seitens des Anbieters in Bezug auf diesen Zweck argumentiert werden kann.

1.1.3 Drittkomponenten

Das Prüfverfahren schließt KI-Systeme, die Drittkomponenten einbinden oder gar darauf basieren nicht grundsätzlich aus. Beispielsweise liegen sogenannte „KI-Modelle mit allgemeinem Verwendungszweck“ („general-purpose AI“, z. B. große Sprachmodelle) im Anwendungsbereich des Prüfverfahrens, wenn diesen ein bestimmter Zweck als Grundlage für die Prüfung zugewiesen wird (siehe 1.1.2).

Die Einbindung externer Komponenten wird im Prüfkatalog explizit aus Perspektive der Supply Chain-Zuverlässigkeit und Cybersicherheit berücksichtigt. Bei einer Zusammenarbeit mit Drittanbietern ist es für den Erfolg der Prüfung jedoch zentral, dass relevante Daten und Dokumentationen eingeholt werden.

Generell gilt, dass im Fall mangelnder Nachweise oder Validierbarkeit eine hohe bzw. gute Qualitätseinstufung des KI-Systems im Ergebnis nicht möglich ist (siehe Kapitel 2).

1.2 Qualitätsdimension als Strukturelement der Prüfung

Im Rahmen des Prüfverfahrens wird ein KI-System entlang von sechs Werten erfasst und analysiert, die im Kontext des **MISSION KI** Qualitätsstandard als Qualitätsdimensionen bezeichnet werden:

- Datenqualität, -schutz- und -Governance,
- Nicht-Diskriminierung,
- Transparenz,
- Menschliche Aufsicht und Kontrolle,
- Verlässlichkeit,
- KI-spezifische Cybersicherheit.

Diese Qualitätsdimensionen bilden die Grundlage des **Prüfkatalogs** der **MISSION KI**, der die Gesamtheit der Prüfanforderungen umfasst. Ihren Ursprung haben die Qualitätsdimensionen in den „Ethik-Leitlinien für eine vertrauenswürdige KI“, entwickelt durch die von der Europäischen Kommission einberufenen hochrangigen Expertengruppe (HEG-KI⁷). Dieses Dokument zählt eine Reihe zentraler Prinzipien auf, die zur Bewertung von Vertrauenswürdigkeit von KI-Systemen dienen sollen. Sie bilden auch die Grundlage des Vertrauenswürdigkeitsbegriffs der KI-Verordnung der EU und werden entsprechend in Erwägungsgrund 27 als Grundlage für die Entwicklung von vertrauenswürdigen KI-Systemen genannt.

⁷ Ethics Guidelines for Trustworthy AI, High-level Expert Group on AI, 2019. URL: [https://digital-strategy.ec.europa.eu/en](https://digital-strategy.ec.europa.eu/en/Zugriff: 27.10.2025) (Zugriff: 27.10.2025).

Im Sinne der Anschlussfähigkeit an die europäische KI-Regulierung, ist es daher naheliegend, dass auch der **MISSION KI** Qualitätsstandard diese Prinzipien aufgreift und integriert. Im europäischen und deutschen Raum sind inzwischen eine Reihe von (Kriterien-)Katalogen und Standards zur KI-Vertrauenswürdigkeit veröffentlicht worden, die auf der HEG-KI basieren. Die Qualitätsdimensionen der **MISSION KI** beruhen dabei konkret auf der VDE SPEC 90012 und dem KI-Prüfkatalog des Fraunhofer IAIS.

Die sechs Qualitätsdimensionen bilden das strukturgebende Element für die beiden wichtigsten Bestandteile der Prüfung: die Schutzbedarfsanalyse (siehe 3.2) und die Prüfanforderungen des Prüfkatalogs, die der VCIO-Einstufung aus der VDE SPEC 90012 folgen (siehe 3.3). Diese Struktur unterteilt die Qualitätsdimensionen weiter in Kriterien, Indikatoren und Observablen (siehe Abbildung 3 und Beschreibungen in 3.3).⁸

Die Schutzbedarfsanalyse identifiziert potenzielle Schutzbedarfe eines KI-Systems auf Grundlage seiner Zweckbestimmung, während die Einstufung bewertet, in welchem Umfang entsprechende Qualitätsmaßnahmen und technische Vorkehrungen umgesetzt wurden. Die Ergebnisse der Schutzbedarfsanalyse und der Einstufung in Prüfanforderungen sind separat je Qualitätsdimension und der dazugehörigen Kriterien strukturiert.

1.3 Prüfaussage

Die Gesamtaussage der Prüfung basiert hauptsächlich auf zwei zentralen Abschnitten. Die Schutzbedarfsanalyse (siehe 3.2) bestimmt für jede Qualitätsdimension den erforderlichen Absicherungsgrad, der aus potenziellen Schadenshöhen abgeleitet wird. Die Einstufung entlang der Prüfanforderungen (siehe 3.3) zeigt durch eine Prüfung anhand von Evidenzen, in welchem Umfang Maßnahmen nach dem Stand der Technik umgesetzt und dokumentiert wurden.

Beide Ergebnisse werden auf Ebene der Kriterien gegenübergestellt (siehe 3.6), um zu beurteilen, ob die getroffenen und dokumentierten Qualitätsmaßnahmen den ermittelten Schutzbedarfen entsprechen. Das Bestehen bescheinigt, dass aus der Dokumentation der Evidenzen die Angemessenheit der umgesetzten Maßnahmen des Prüfkatalogs (d. h. Prüfanforderungen) entsprechend den Schutzbedarfen hervorgeht.

Da es sich bei der Qualitätsprüfung um eine Plausibilitätsprüfung ohne eine vollumfängliche Risikoanalyse handelt, trifft das Bestehen keine Aussage über die bestehenden Residualrisiken des KI-Systems.

1.4 Gültigkeit

Das Prüfverfahren beschreibt eine zeitpunktbezogene Prüfung. Relevant für die Prüfung und die Bewertung sind vor allem die Zweckbestimmung sowie die technische Umsetzung des KI-Systems zum Zeitpunkt der Prüfung. Diese werden in der Anwendungsfallbeschreibung festgehalten (siehe 3.1).

Das auf Grundlage des Prüfverfahrens und des Prüfkatalogs erstellte Prüfergebnis gilt ausschließlich für das zu Beginn der Prüfung spezifizierte System. Es verliert seine Gültigkeit, sobald der ursprüngliche Anwendungszweck oder die technische Umsetzung wesentlich verändert werden und diese Änderungen die Anforderungen des Qualitätsstandards beeinflussen.

⁸ Der Begriff VCIO steht für eine Einteilung in Values-Criteria-Indicators-Observables. Diese Struktur wurde übernommen, der Begriff „Value“ wurde ersetzt durch „Qualitätsdimension“.

Bei KI-Systemen, die während des Betriebs (z. B. inkrementell) weiterlernen, werden zudem die Methodik und die Bedingungen des weiteren Trainings (z. B. Anforderungen an die Trainingsdaten, Logging von Loss-Werten) in der Anwendungsfallbeschreibung erfasst als auch in den Prüfanforderungen berücksichtigt.⁹ Auch hiervon hängt das Prüfergebnis ab und würde bei signifikanten Abweichungen seine Gültigkeit verlieren.

Gleichzeitig gibt es auch bei KI-Systemen, die während des Betriebs nicht weiterlernen, oftmals externe Dynamiken aufgrund von Data Drifts oder Concept Drifts in der bestimmten Anwendung. Daher muss die Gültigkeit der Prüfergebnisse beschränkt sein.

⁹ Dies gilt für die Prüfanforderungen nur in dem Rahmen, in dem kontinuierliche Prüfungs- und Monitoringaspekte im Umfang liegen. Momentan fokussiert sich der Prüfkatalog auf eine zeitpunktbezogene Prüfung für den Anbieter eines KI-Systems. Aspekte der Anwendung und kontinuierlichen Umsetzung werden in den Anforderungen nur so weit wie möglich für einen Anbieter im Detail berücksichtigt.

2. Prüftiefe

Die Prüftiefe steuert maßgeblich die Durchführung des Prüfverfahrens und den Aufwand, mit dem die Prüfung durchgeführt wird. Sie bestimmt den Grad der Absicherung und Prüfung der Plausibilität von Prüfaussagen und definiert die Rollen, die zur Validierung der Ergebnisse im Prüfverfahren gebraucht werden.

Die Prüftiefe ergibt sich aus drei zentralen Eigenschaften der Prüfung:

- 1) Dem Detailgrad an Maßnahmen, die je nach Mindestanforderung aus dem Prüfkatalog erfüllt werden müssen. Diese Mindestanforderung wird durch den Schutzbedarf festgelegt und führt zu einer klar festgelegten, mindestens notwendigen Stufe in den Anforderungen. Eine höhere Prüftiefe geht hier u. a. mit umfangreicheren Anforderungen zu Daten, Modellen, Prozessen, Bewertungen, Tests oder Organisationsstrukturen einher.
- 2) Den bereitzustellenden Evidenzen in Form von Dokumentationen und durchgeführten technischen Maßnahmen wie Tests, die in höheren Prüftiefen auch reproduzierbar sein müssen.
- 3) Dem Grad an Validierung der Prüfaussagen durch einen oder mehrere Prüfer.

Die genaue Zusammenstellung dieser Eigenschaften hängt für die einzelnen Prüfanforderungen vom gewählten Schutzbedarf ab. Diese Zusammenhänge sind in Tabelle 2 zusammengestellt.

Eine größere Prüftiefe ist in der Regel mit einem erhöhten Prüfaufwand verbunden, der bis zur Einbindung unabhängiger Prüfer auf höheren Validierungsstufen reichen kann. Unabhängige Prüfer meint im Rahmen der Selbstprüfung, dass nicht an der Entwicklung beteiligte Experten aus der zu prüfenden Organisation herangezogen werden.

Darüber hinaus ist die freiwillige Einbeziehung externer Prüfstellen ausdrücklich möglich. Deren Beteiligung kann insbesondere bei hohen Prüftiefen und erhöhten Schutzbedarfen den Grad der Absicherung weiter stärken.

Eine Erhöhung der Anforderungen des Prüfkatalogs über die höchste Stufe hinaus ist nicht explizit vorgesehen, aber ebenfalls problemlos freiwillig möglich. Betreiber-spezifische Nachweise rund um Betriebsumgebung und Monitoring können im Prüfkatalog (falls gewünscht) ergänzt werden.

Der Grundansatz des **MISSION KI** Qualitätsstandards macht die Faktoren Belastbarkeit und Effizienz entscheidend für die Definition der Prüftiefe. Der Umfang an Maßnahmen und Evidenzen, sowie die Notwendigkeit der Validierung ist an die Schutzbedarfsanalyse gekoppelt, sodass in den Teilen der Prüfung mit dem höchsten Schutzbedarf auch die höchste Qualitätsabsicherung der Prüfaussage sichergestellt wird.

In Bereichen mit geringerem Schutzbedarf wird der Aufwand der Nachweiserbringung gezielt reduziert. Selbst in der höchsten Prüftiefe gehen die geforderten Maßnahmen nicht über bewährte Best-Practices hinaus. Ein hoher Grad an Belastbarkeit der Prüfaussagen wird stattdessen durch Validierung nach dem Vier-Augen-Prinzip und durch Reproduzierbarkeit der Ergebnisse gewährleistet.

Ableitung der Prüftiefe				
	Schutzbedarf			
	–	gering	moderat	hoch
Mindestanforderung aus dem Prüfkatalog	Es müssen keine Maßnahmen umgesetzt werden. (Observablenstufe D)	Einfache Maßnahmen müssen umgesetzt werden. (Observablenstufe C)	Fortgeschrittene Maßnahmen müssen umgesetzt werden. (Observablenstufe B)	Best-Practice Maßnahmen müssen umgesetzt werden. (Observablenstufe A)
Bereitstellung von Evidenzen	–	Erstellen und Nachhalten von einfachen Nachweisen (z. B. Dokumentation) über die durchgeführten Maßnahmen; Erstellen und Nachhalten von technischen Nachweisen über die durchgeführten Tests und die dazugehörigen Ergebnisse (z. B. Systemauszüge, Dashboards).	Erstellen und Nachhalten von ausführlichen Nachweisen (z. B. Dokumentation) über die durchgeführten Maßnahmen; Erstellen und Nachhalten von technischen Nachweisen über die durchgeführten Tests und die dazugehörigen Ergebnisse mit Metadaten (z. B. Systemauszüge, Dashboards, Logs zu Modellversionen und Datensätzen).	Erstellen und Nachhalten von ausführlichen Nachweisen (z. B. Dokumentation) über die durchgeführten Maßnahmen; Erstellen und Nachhalten von reproduzierbaren technischen Nachweisen über die durchgeführten Tests und die dazugehörigen Ergebnisse mit Metadaten (z. B. Systemauszüge, Dashboards, Logs zu Modellversionen und Datensätzen); Reproduktion der Ergebnisse; Nachweise über die Durchführung der internen Prüfung (z. B. über „Audit Trail“).
Validierung	–	Validierung durch berechtigte Person des verantwortlichen Teams selbst (z. B. Product oder Technical Owner); Schutzbedarfe und Einstufungen können durch Personen ermittelt und bestätigt werden, die an der Entwicklung beteiligt waren; Erforderliche Qualifikation: Erfahrung in Entwicklung/ Betrieb von KI-Systemen oder vergleichbare Qualifikation.	Validierung durch berechtigte Person, die nicht an der Entwicklung des KI-Systems beteiligt war (z. B. Product oder Technical Owner eines anderen Teams); Schutzbedarfe und Einstufungen können durch Personen ermittelt werden, die an der Entwicklung beteiligt waren; Erforderliche Qualifikation: Erfahrung in Entwicklung/ Betrieb von KI-Systemen oder vergleichbare Qualifikation.	Zusätzliche Validierung durch berechtigte Personen einer hierarchisch unabhängigen Stelle (z. B.: Interne Revision); Schutzbedarfe und Einstufungen können durch Personen ermittelt werden, die an der Entwicklung beteiligt waren; Validierungen werden nach dem 4-Augen Prinzip durchgeführt, d. h. von zwei Personen, die nicht an der Entwicklung des KI-Systems beteiligt waren; Erforderliche Qualifikation: Es müssen mindestens einmal Erfahrung in Systeme oder vergleichbare Qualifikation und einmal Audit-/Prüferfahrung vorhanden sein.

Tabelle 2: Ableitung der Prüftiefe basierend auf den festgestellten Schutzbedarfen

3. Durchführung der Prüfung

Die folgenden Abschnitte beschreiben im Detail die Durchführung der Prüfung (siehe Abbildung 2).

Im ersten Schritt fertigt die zu prüfende Organisation eine Anwendungsfallbeschreibung an (siehe 3.1), welche den Prüfgegenstand und damit den spezifischen Geltungsbereich der Prüfung beschreibt. Auf dieser Basis wird im zweiten Schritt eine Schutzbedarfsanalyse durchgeführt, um einerseits besonders kritische Qualitätsdimensionen für die Absicherung des KI-Systems und andererseits ggf. für das KI-System „nicht anwendbare“ Kriterien zu identifizieren (siehe 3.2). Für alle relevanten Qualitätsdimensionen wird im dritten Schritt die Einstufung des KI-Systems gemäß der Prüfanforderungen vorgenommen (siehe 3.3). Hierzu werden im vierten Schritt außerdem Kontrollen bzw. Evidenzen für die Einstufung dokumentiert, insbesondere mithilfe von technischen Prüfmethoden (siehe 3.4). Im fünften Schritt wird die Einstufung anhand der Evidenzen in der erforderlichen Prüftiefe validiert (siehe 3.5). Schließlich wird die Einstufung der Schutzbedarfsanalyse als Gesamtergebnis systematisch gegenübergestellt (siehe 3.6). Die Durchführung der Prüfung wird mit der Ergebnisdokumentation abgeschlossen (siehe 3.7), wobei die Gültigkeit der Prüfergebnisse ggf. anschließend überwacht oder verlängert werden kann (siehe 3.8).

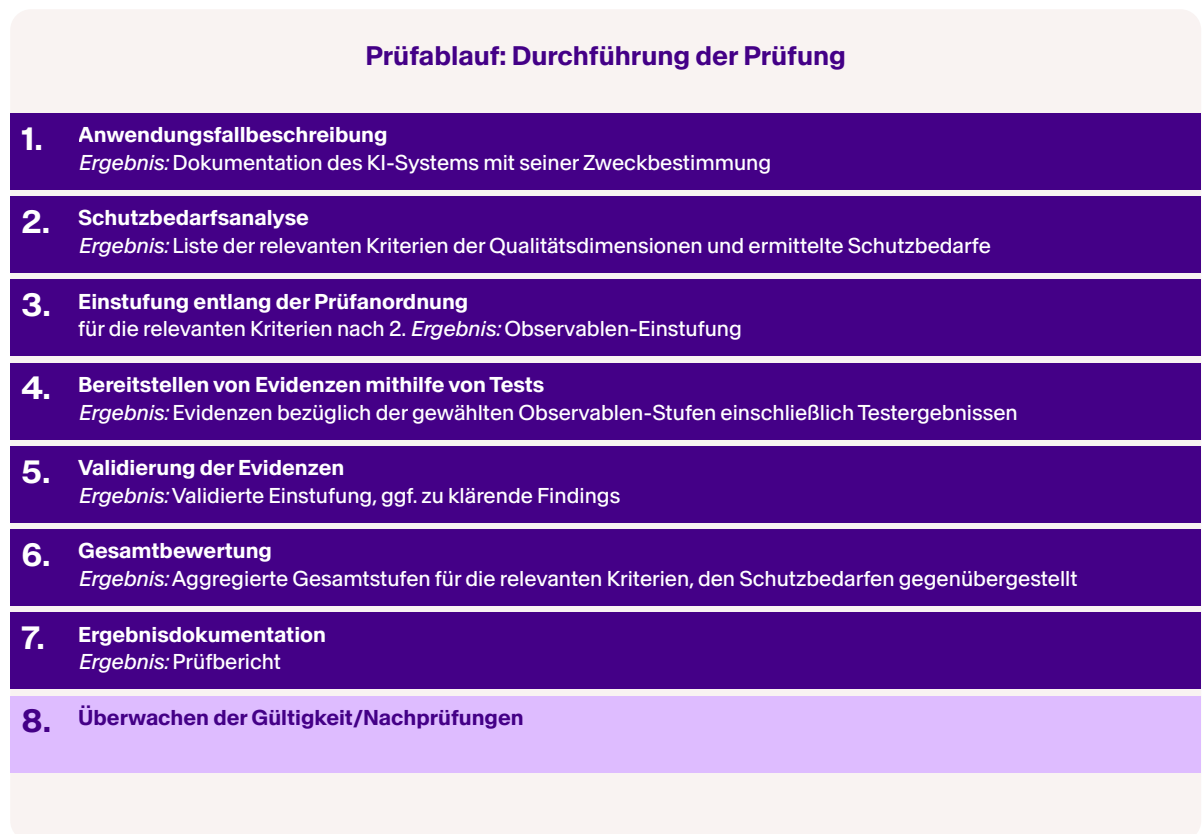


Abbildung 2: Übersicht der Prüfschritte.

3.1 Anwendungsfallbeschreibung

Die Anwendungsfallbeschreibung dient der initialen Spezifikation des Prüfgegenstandes, insbesondere seiner Zweckbestimmung und der Definition des Anwendungsbereichs des KI-Systems. Sie ermöglicht (internen) Prüfern, die zur Validierung herangezogen werden, aber mit dem KI-System bisher keine Berührungspunkte hatten, eine Ersteinschätzung des Prüfgegenstandes.

Die beigegefügte Vorlage zur **Anwendungsfallbeschreibung** (siehe Anhang) unterstützt bei der Aufbereitung der für die Selbstprüfung relevanten Informationen.

Eine Aussage über die Vertrauenswürdigkeit eines KI-Systems ist stets an dessen Zweckbestimmung und Anwendungsbereich gebunden. Schutzbedarfe können sich zwischen unterschiedlichen Anwendungsbereichen für dasselbe KI-System hochgradig unterscheiden.

Die Schutzbedarfsanalyse und die Einstufung der Kriterien sind daher im Kontext des in der Anwendungsfallbeschreibung festgelegten Einsatzkontextes zu verstehen. Deshalb bildet der Einsatzkontext auch ein grundlegendes Element des Prüfberichts und der Prüfaussage.

Neben dem Einsatzkontext, den Grenzen des Eingabebereichs und dem Ausgabeformat werden weitere relevante Kontextinformationen abgefragt. Das schließt potenzielle Veränderungen des KI-Systems im Betrieb und – damit auch nach dem Zeitpunkt der Prüfung – sowie die Art des Betriebs (lokal oder in der Cloud) mit ein. Ebenso wird spezifiziert, ob und in welcher Form Menschen am Betrieb und an der Aufsicht des KI-Systems beteiligt werden sollen.

Die gesammelten Informationen geben in der **Prüfberichtsvorlage** (siehe Anhang) den erforderlichen Kontext für die Prüfaussage, wobei die Abfrage sensibler Daten vermieden wird. So können Unternehmen die Nutzungsbedingungen ihres KI-Systems, auf die sich die Prüfung bezieht, im Prüfbericht transparent und nachvollziehbar darstellen.

3.2 Schutzbedarfsanalyse

Die **Schutzbedarfsanalyse** (siehe Anhang) dient der Vorfilterung der für den jeweiligen Anwendungsfall relevanten Kriterien. Hintergrund ist die Annahme, dass je nach Aufgabe und Einsatzkontext des KI-Systems nicht jedes Kriterium einer Qualitätsdimension gleichermaßen relevant für die Bewertung der Qualität ist.

Die Schutzbedarfsanalyse bietet so eine effiziente Methode, gezielt relevante Qualitätsdimensionen auf Ebene der Kriterien zu identifizieren und diese nach Höhe ihres Schutzbedarfs zu kategorisieren. Gleichzeitig ermöglicht die Schutzbedarfsanalyse einen praktikablen Weg, um eine Anschlussfähigkeit an die europäische KI-Verordnung zu erreichen, die bei „Hochrisiko“-KI-Systemen von einer Zweckbestimmung ausgeht.

Die Untersuchung der Höhe des Schutzbedarfs erfolgt separat je Qualitätsdimension auf Ebene der Kriterien. Sie richtet sich nach den potenziellen Schäden, die bei Nichteinhaltung der Anforderungen im Kontext der einzelnen Qualitätsdimensionen auf Ebene der Kriterien realisiert werden könnten. Konkret werden die möglichen Schadensszenarien („Worst-Case“) aus den in Art. 1 Abs. 1 KI-Verordnung gelisteten Schutzzielen abgeleitet. Der jeweilige Schutzbedarf wird in drei Kategorien eingeteilt – gering, moderat, hoch; alternativ können Kriterien einer Qualitätsdimension auch als „nicht anwendbar“ bewertet werden. Detailliertere Erläuterungen zur Analyse und dem Bewertungsschema werden in Anhang 4.3 ausgeführt. Basierend auf den Einschätzungen zur Relevanz der Qualitätsdimensionen und ihrer zugehörigen Kriterien (durch Ausschluss oder Bestimmung der Schutzbedarfe) kann die weitere Prüfung zielgerichtet und effizient durchgeführt werden.

Für „nicht anwendbare“ Kriterien oder – im Falle, dass alle Kriterien einer Qualitätsdimension „nicht anwendbar“ sind – „nicht anwendbare“ Qualitätsdimensionen ist prinzipiell keine tiefergehende Analyse in Form der Einstufung (siehe 3.3) vorgesehen. Auf ausdrücklichen Wunsch des Anbieters kann jedoch von diesem Prinzip abgewichen und auch für „nicht anwendbare“ Kriterien oder Qualitätsdimensionen eine Einstufung durchgeführt werden. Die festgestellten Schutzbedarfe wiederum dienen implizit als Mindeststufe in der Gesamtbewertung (vgl. Kapitel 2).

3.3 Prüfanforderungen

Ein Kernbestandteil des Prüfverfahrens ist die Bestimmung – „Einstufung“ – der Qualität eines KI-Systems. Um eine systematische und effiziente Einstufung zu erlauben, sind die Anforderungen des Qualitätsstandards hierarchisch in einer an die VCIO-Systematik der VDE SPEC 90012 angelehnten Struktur organisiert (siehe Abbildung 3).

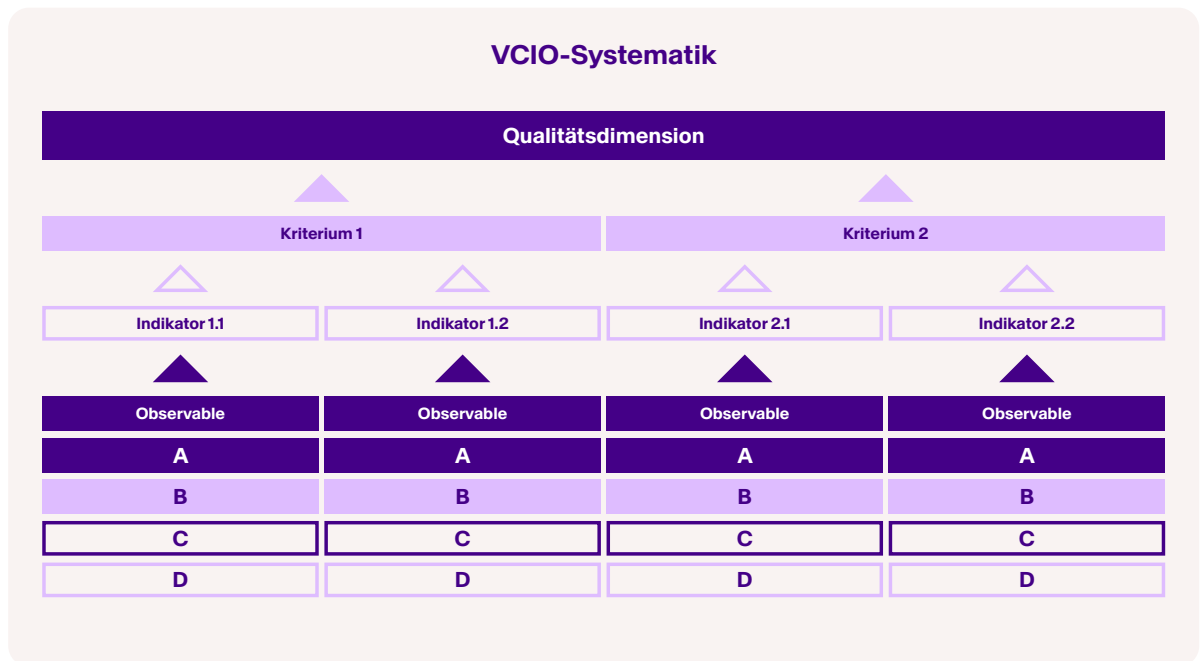


Abbildung 3: Adaptierte VCIO-Systematik der Prüfanforderungen

Der angehängte Prüfkatalog enthält die Gesamtheit der Qualitätsdimensionen, Kriterien, Indikatoren und Observablen. Im Rahmen des **MISSION KI** Qualitätsstandards haben die Begriffe aus der VCIO-Systematik folgende Bedeutung:

Qualitätsdimension: Sie bilden die Grundlage des Qualitätsbegriffs im **MISSION KI** Qualitätsstandard und definieren die übergeordneten Qualitätsziele, die ein KI-System erfüllen soll.

Kriterien: Qualitätsdimensionen werden in Kriterien (siehe Abbildung 4) unterteilt, um Qualitätsziele weiter einzugrenzen, wie zum Beispiel „Verlässlichkeit“ in „Leistungsfähigkeit und Robustheit“ und „Rückfallpläne und funktionale Sicherheit“. Die Schutzbedarfsanalyse wird auf dieser Ebene den Prüfanforderungen gegenübergestellt, da Qualitätsdimensionen sehr weitreichend sind, die nachgelagerten Indikatoren dagegen zu granular.

Indikatoren: Jedes Kriterium wird durch Indikatoren spezifiziert, die für die Qualitätssicherung des KI-Systems entscheidend sind. Indikatoren sind in drei Typen erforderlicher Handlungen unterteilt:

- **Typ 1 (Analyse):** Fokus auf Risikoanalysen und Zweckbestimmung des KI-Systems.
- **Typ 2 (Maßnahmen):** Unterteilt in organisatorische (z. B. Governance oder Schulungen) und technische Maßnahmen (z. B. Datenaufbereitung oder Tests), die zur Mitigation der festgestellten Schutzbedarfe/Risiken umgesetzt werden müssen.
- **Typ 3 (Bewertung):** Erfordert eine Begründung für und Abnahme von verbleibenden Residualrisiken nach Durchführung der Maßnahmen.

Manche Indikatoren können für mehrere Kriterien relevant sein, was durch eine Referenz im Prüfkatalog hervorgehoben ist (siehe Prüfkatalog auf Ebene der Observablen).



Abbildung 4: Qualitätsdimensionen und Kriterien

Observablen: Diese sind die feststellbaren/messbaren Merkmale, die die Qualität eines Systems bestimmen und in Stufen von A (am besten) bis Stufe D (überhaupt nicht erfüllt) eingeteilt sind. Jede Observable zeigt den Erfüllungsgrad eines Indikators an. Die Abstufung der Observablen erfolgt über die Komplexität und den Umfang der zu erfüllenden Anforderungen.

Die Schutzbedarfsanalyse bestimmt die Anwendbarkeit der Kriterien und damit auch der dazugehörigen Indikatoren für ein konkretes KI-System. Die Einstufung muss daher nur für anwendbare Indikatoren stattfinden. Durch die gleiche Anzahl von Stufen bei Schutzbedarfen und Observablen, können sie zur abschließenden Bewertung eines KI-Systems gegenübergestellt werden (siehe 3.6).

„Einstufung“ im Rahmen der Prüfungsdurchführung bedeutet, dass je Indikator diejenige Observablen-Stufe auswählt, die auf das spezifische KI-System zutrifft. Die Einstufung und damit die Umsetzung der geforderten Analysen, Maßnahmen und Bewertungen muss über das Erbringen von Evidenzen nachweisbar sein (siehe 3.4). Abschließend kann eine Gesamtbewertung des Systems über eine Aggregation vorgenommen werden (siehe 3.6).

3.4 Bereitstellen von Evidenzen

Als Begründung bzw. Nachweis für die Korrektheit der Einstufung des KI-Systems müssen entsprechende Kontrollen und Evidenzen durch den Prüfling systematisch erfasst und bereitgestellt werden. Die genauen Anforderungen an den Umfang der Evidenzen ergeben sich aus den jeweiligen Observablen und sind dem Prüfkatalog zu entnehmen. Dabei ist sicherzustellen, dass alle in den Observablen genannten Punkte vollständig erfüllt und durch geeignete Evidenzen belegt sind.

Einzelne Nachweise können dabei mehrere Indikatoren auch über verschiedene Qualitätsdimensionen hinweg abdecken. Doppelte Nachweiserbringung ist nicht erforderlich. Bereits vorhandene Nachweise und Zertifizierungen (z. B. DSGVO-Elemente zum Schutz personenbezogener Daten, ISO 27001 oder IT-Grundschutz) werden als Evidenzen anerkannt.

Liste der Observablen und dazugehörigen Evidenzarten

Analysen und Bewertungen	
Unterkategorie	Evidenzarten
(Zweck-)Definition	schriftl. Dokumentation – Beschreibung, Begründung
Risiko	schriftl. Dokumentation – Analyse, Testergebnis – Ergebnisdokumentation, Begründung
Metriken & Schwellenwerte	schriftl. Dokumentation – Analyse, Testergebnis – Ergebnisdokumentation
Bewertungen	schriftl. Dokumentation – Beschreibung, Begründung
Organisatorische Maßnahmen	
Unterkategorie	Evidenzarten
Governance	schriftl. Dokumentation – Guideline/Policy, Prozessdokumentation, Plan, Beschreibung
Systemnahe Prozesse	schriftl. Dokumentation – Prozessdokumentation, Plan, Beschreibung, nicht-schriftl. Dokumentation – Screenshots, Computerscript/Code
Benutzerinstruktionen	schriftl. Dokumentation – Prozessdokumentation, Plan, Beschreibung
Schulungsmaterialien	schriftl. Dokumentation – Plan, Beschreibung
Technische Maßnahmen	
Unterkategorie	Evidenzarten
Daten	Daten – Stichprobe, Daten- Datendokumentation, Daten – Voller Datenzugang, nicht-schriftl. Dokumentation – Screenshots, nicht-schriftl. Dokumentation – Computerscript/Code
Tests	Testergebnis – Testdurchführung, Testergebnis – Ergebnisdokumentation, Testergebnis – Screenshot
Modelle	Modellmerkmale – Modelldokumentation, Modellmerkmale – Modellzugang
Betrieb	schriftl. Dokumentation – Prozessdokumentation, Plan, Beschreibung, nicht-schriftl. Dokumentation – Logs, nicht-schriftl. Dokumentation – Screenshots, nicht-schriftl. Dokumentation – Computerscript/Code

Tabelle 3: Liste der Observablen und dazugehörigen Evidenzarten

Im Rahmen der Prüfung kommen unterschiedliche Arten von Evidenzen zum Einsatz. Dazu wird eine Liste an Evidenzarten (Tabelle 3) vorgeschlagen, die anzeigt, wie die Observablen umgesetzt werden können.

Unter Indikatoren der Kategorie „Analysen und Bewertungen“ sind Evidenzen und den ihnen zugeordneten Observablen überwiegend in Form von Dokumentationen, Beschreibungen des KI-Systems und seiner Komponenten (z.B. System- oder Architekturdiagramme), Protokollierungen

oder anderen Prozessartefakten gefragt. Ist eine Dokumentation für verschiedene Kriterien bzw. Indikatoren relevant, muss diese nicht mehrfach bereitgestellt werden, sondern kann bei Wiederholung referenziert werden.¹⁰

Die Kategorie „Organisatorische Maßnahmen“ bewertet mit den entsprechenden Indikatoren und nachgelagerte Observablen inwiefern in einer Organisation Maßnahmen getroffen wurden, die für die qualitätsgemäße Entwicklung und Ausführung eines KI-Systems nötig sind. In dieser Kategorie werden beispielsweise Prozesse, Schulungen oder Guidelines abgefragt.

In der dritten Kategorie werden „Technische Maßnahmen“ beschrieben. Für einige Indikatoren und ihre Observablen vom Typ „Maßnahmen“ sind Evidenzen zu spezifischen Systemeigenschaften erforderlich, die beispielsweise durch die Evaluierung geeigneter Metriken, Tests und Analysen des KI-Systems und seiner Komponenten erzeugt werden können.

Die so bereitgestellten, vorwiegend prozessualen Evidenzen schaffen Transparenz und können, wenn sie mehrere Kriterien abdecken, durch Verweise anstelle von Duplikaten genutzt werden. Reichen diese Nachweise für die Bewertung technischer Systemeigenschaften nicht aus, folgt eine vertiefte, messbasierte Prüfung, deren Ergebnisse nachvollziehbar und wiederholbar dokumentiert werden. Wo Prüfanforderungen technische Aspekte adressieren, sollen Prüflinge repräsentative, reproduzierbare Testergebnisse beilegen (z. B. Datensatz-/Split-Beschreibung, Testkonfiguration, Versionen, Seed/Run-Informationen). Das Prüfverfahren gibt dafür keine konkreten Metriken verpflichtend vor, da in Abhängigkeit des KI-Systems und des Anwendungskontexts verschiedene Metriken relevant sein können. Der Qualitätsstandard beinhaltet eine **Prüfmethodensammlung**, die als Hilfestellung dient (siehe Anhang). Die technische Prüfmethodensammlung fokussiert sich auf Verfahren, die KI-Systemeigenschaften substantzieren, indem sie spezifische Eigenschaften von Daten oder KI-Komponenten auf Testausgaben abbilden. Jede technische Prüfmethode in der Sammlung ist über Indikatoren-IDs mit dem angehängten **Prüfkatalog** verknüpft. Prüflinge erhalten so eine Übersicht etablierter Verfahren, die zur Evidenzerzeugung bezüglich eines Indikators prinzipiell geeignet sind. Natürlich müssen spezifische technische Prüfmethode immer noch basierend auf einer Analyse des zu prüfenden KI-Systems, insbesondere dessen Aufgabenstellung und Anwendungsbereich, ausgewählt werden (siehe z. B. VE1.2 und VE1.3 im **Prüfkatalog**). Deshalb ist es nicht möglich, eine Vorgabe zur Nutzung bestimmter technischer Prüfmethode zu treffen.

3.5 Validierung der Evidenzen

Die durch den Prüfling vorgenommene Einstufung wird im nächsten Schritt validiert. Dies umfasst die Bestätigung des Vorhandenseins und der Plausibilität der Maßnahmen, die durch die Evidenzen nachgewiesen werden sollen. Die Validierung erfolgt in Bezug auf die gewählten Observablen-Stufen und unter Berücksichtigung der bereitgestellten Evidenzen (siehe 3.4).

Wie in Kapitel 2 beschrieben, ergibt sich der erforderliche Grad der Validierung je Indikator direkt aus der zu validierenden Observablen-Einstufung.

Evidenzen zu technischen Maßnahmen müssen, wo anwendbar und je nach Prüftiefe, reproduzierbare Tests unter identischer Konfiguration (z. B. Daten/Splits, Seeds, Hyperparameter) beinhalten. Die **Prüfmethodensammlung** dient der methodischen Absicherung. Sie kann zusätzlich durch unabhängige Prüfer für eine Validierung von systematisch erfassten Evidenzen (z. B. erwarteter Wert, akzeptabler Bereich) genutzt werden.

¹⁰ Beispiel: Für die Anforderung „Wurden bzgl. Fairness Zielgruppen definiert?“ kann ein Artefakt die „Dokumentation Fairnessanalyse v2.pdf“ sein, wobei sich die Verknüpfung über „Tabelle 2: Potenziell benachteiligte Gruppen“ ergibt. Außerdem können die Indikatoren bezüglich der Risikoanalyse alle auf ein gemeinsames Risikoanalysen-Dokument verweisen.

3.6 Gesamtbewertung

Für den Abschluss einer vollständigen Prüfung erfolgt nach der Observablen-Einstufung eine Gesamtbewertung des KI-Systems. Hierzu wird diese Einstufung auf Ebene der Kriterien aggregiert und systematisch den entsprechenden Schutzbedarfen gegenübergestellt (siehe Abbildung 5).

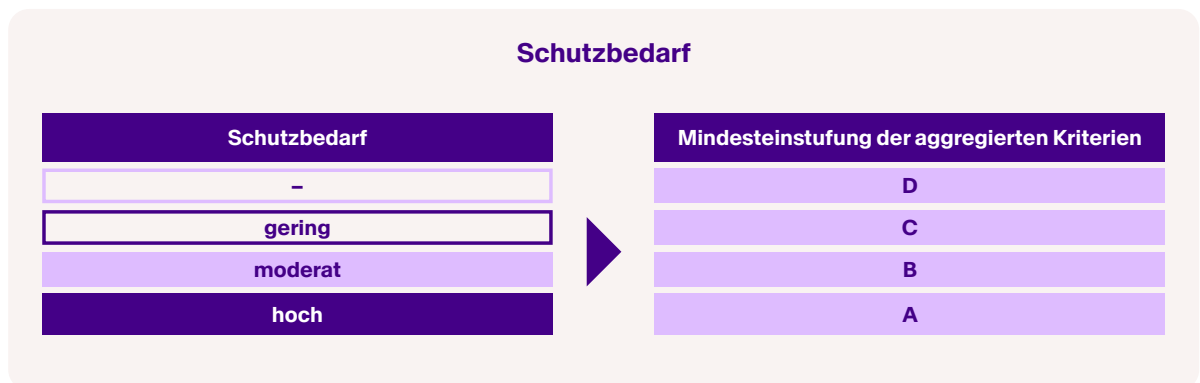


Abbildung 5: Schutzbedarfe als implizite Mindeststufen in der Gesamtbewertung, die den Bestätigungsgrad des Mindeststandards für ein KI-System im konkreten Anwendungsfall festlegen.

Die Gesamtbewertung überprüft so, ob die jeweils für das KI-System ermittelte und auf Kriterien aggregierte Einstufung den Schutzbedarfen für die jeweiligen Kriterien gerecht wird. Die auf Basis der Schutzbedarfe festgesetzten, mindestens erforderlichen Niveaus müssen erreicht oder übertroffen werden.

In der anschließenden Gegenüberstellung der aggregierten Einstufung mit der Schutzbedarfsanalyse dienen die unterschiedlichen Stufen der Schutzbedarfe (hoch, moderat, gering) als erforderliches Mindestniveau. Dieses Niveau muss je (anwendbarem) Kriterium in der Einstufung erreicht werden, um die Prüfung zu bestehen. Beispielsweise würde ein hoher Schutzbedarf erfordern, dass die aggregierte Kriterien-Stufe einem „A“ entspricht. Bei einem moderaten Schutzbedarf muss mindestens die aggregierte Kriterien-Stufe „B“ oder höher erreicht werden.

Wird in allen Kriterien das Mindestniveau erreicht, ist die Qualitätsprüfung insgesamt bestanden. Wird das Niveau nicht in allen Kriterien erfüllt, ist die Qualitätsprüfung nicht bestanden. Wird das Mindestniveau in einzelnen Kriterien übertroffen, liegt eine Übererfüllung vor.

Eine Anleitung zum genauen **Prozessablauf der Gesamtbewertung** ist dem Qualitätsstandard angehängt.

3.7 Prüfbericht

Nach Abschluss der Prüfung wird das Ergebnis in einem Prüfbericht zusammengefasst. Der Prüfbericht beinhaltet eine formale Erklärung über die Richtigkeit der getätigten Aussagen und spiegelt die wichtigsten Aspekte der Prüfung übersichtlich wider:

- Informationen zum KI-System, die u. a. eine Beschreibung des Anwendungskontextes, der genutzten Daten sowie der (KI-)Komponenten und deren Zusammenwirken umfassen.
- Ermittelte Schutzbedarfe nach Kriterien aufgeschlüsselt und mit Erläuterung versehen, falls das Kriterium als „nicht anwendbar“ bewertet wurde.
- Jeweils erreichte Stufe der Prüfanforderungen, die dem Schutzbedarf auf Ebene der Kriterien gegenübergestellt werden.
- Ausgewählte Qualitätsmaßnahmen und Vorkehrungen zur Absicherung des KI-Systems, die die Einstufung untermauern.
- Auskunft über die Prüftiefe, d. h. die technischen Tests und welcher Grad der Validierungen durchgeführt wurde.

- Erläuterungen und Kommentare, die die Interpretation der Ergebnisse erleichtern oder im Kontext nachvollziehbar machen.
- Erklärung darüber welche Kriterien erfüllt, nicht erfüllt oder übertroffen wurden.
- Für die Prüfung verantwortliche Personen.

Die Darstellung der Ergebnisse muss der Prüfberichtsvorlage (siehe Anhang) entsprechen. Der Prüfbericht ist nur mit der Unterschrift der für die Prüfung und die potenziellen Validierungen verantwortlichen Personen gültig. Die damit einhergehende Erklärung über die Richtigkeit der getätigten Aussagen ist im Rahmen des **MISSION KI** Qualitätsstandards verpflichtend.

3.8 Überwachen der Gültigkeit

Es muss durch konkrete Maßnahmen und Bestimmung sichergestellt sein, dass die Gültigkeit des Ergebnisses der Prüfung (siehe Gültigkeit 1.4) erhalten bleibt.

Die Prüfaussage ist grundsätzlich nur gültig für die eindeutig bestimmte Version des KI-Systems, die zum Zeitpunkt der Prüfung vorlag. Dies soll aus Dokumentation und klarer Versionierung im Kontext des Prüfberichts hervorgehen.

Die Prüfaussage verliert ihre Gültigkeit so bald:

- Sich der ursprüngliche Anwendungszweck ändert.
- Externe Bedingungen (z. B. Concept oder Data Drift) die Prüfaussage verändern würden.
- Die technische Umsetzung des KI-Systems (Software oder Hardware) signifikant verändert wird.

Bei den folgenden, nicht signifikanten Änderungen bleibt die Gültigkeit bestehen:

- Gleichwertige Neuverteilung von Verantwortlichkeiten.
- Änderungen an Hardware- oder Softwarekomponenten (zum Beispiel neue Generation von GPUs, neue Version einer Softwarebibliothek etc.), die nachweislich keinen signifikanten Einfluss auf die geprüften Qualitätsdimensionen hatten (siehe Monitoring).

4. Anhänge

4.1 Glossar

4.1.1 Oberbegriffe

Qualitätsdimension

Eine Qualitätsdimension bezeichnet eine erstrebenswerte, hochstufige, allgemeine Eigenschaft eines →KI-Systems, deren Vorliegen mittelbar – d. h. über Konkretisierungen und Operationalisierungen (→Kriterium, →Indikator, →Observable) – getestet werden kann und die im Verbund mit anderen gleichrangigen Eigenschaften bei Vorliegen eines korrespondierenden Schutzbedarfs die Qualität dieses →KI-Systems insgesamt definiert. Das Modell aus Qualitätsdimension, →Kriterium, →Indikator, →Observable wird gemeinsam aus der Literatur als „VCIO“ („Value-Criterium-Indicator-Observable“) referenziert. Der dort verwendete Begriff „Wert“ (Value) wird innerhalb von **MISSION KI** weitgehend bedeutungsgleich durch „Qualitätsdimension“ ersetzt.

Kriterium

Um zu bestimmen, ob ein →KI-System einzelne →Qualitätsdimensionen erfüllt, werden Kriterien spezifiziert. Kriterien stellen somit eine Konkretisierung einer Qualitätsdimension in Richtung Operationalisierung und beobachtbarer spezifischer Sachverhalte sowie auf konkrete Risiken/ Schutzbedarfe hin dar.

Indikator

Normalerweise ist es nicht möglich, direkt zu messen, ob einzelne →Kriterien erfüllt werden. Um dies zu prüfen, werden die →Kriterien weiter in Indikatoren aufgeschlüsselt. Indikatoren ergeben sich aus Bedingungen, die zur Erfüllung übergeordneter →Kriterien auf einer abstrakten Ebene beitragen. Somit liefern Indikatoren Informationen zu spezifischen Eigenschaften eines →KI-Systems, die für die qualitative Erfüllung/Nichterfüllung oder den Erfüllungsgrad eines Kriteriums entscheidend sind (VDE SPEC 90012: 2.28). Maximalwertindikatoren sind Indikatoren, denen im Rahmen der Bewertung eine gesonderte Rolle zukommt. Aufgrund ihrer Kritikalität für den ermittelten Schutzbedarf stellt ihre Einstufung für den Durchschnitt einen bindenden Maximalwert dar.

Observable

Um die Erfüllung der →Indikatoren zu bewerten, werden Observablen definiert, die in Stufen anzeigen, inwieweit ein →Indikator erfüllt wird. Die Stufen der Observablen sind hierbei vordefiniert und zeigen das Zielerreichungslevel an. Eine Observable ist somit eine messbare Größe, die verwendet wird, um den Zustand oder die Eigenschaften eines Systems anhand eines Indikators zu bestimmen (VDE SPEC 90012: 2.36).

Beispiele: Güte der Datensatzdokumentation oder der Risikovermeidung.

Test

Versuch zur Feststellung bestimmter Eigenschaften, Leistungen oder von Vergleichbarem.

4.1.2 Einzelne Qualitätsdimensionen

(KI-spezifische) Cybersicherheit

Widerstandsfähigkeit von →KI-Modellen, →KI-Komponenten und →KI-Systemen gegen KI-spezifische, maliziöse äußere Eingriffe und Manipulationen, die über allgemeine Telekommunikationsnetzwerke stattfinden.

Daten-Governance

Behandlung, Bearbeitung und Absicherung der Daten, die innerhalb des Lebenszyklus eines →KI-Systems genutzt werden mit dem Ziel, dass die Daten hohen Qualitäts- und Integritätsstandards genügen und im Einklang mit geltenden Vorschriften zum Schutz der Privatsphäre und dem →Datenschutz verwendet werden (vgl. ErwGr. 27 KI-VO).

Datenqualität (engl. Data Quality)

Eigenschaft der Trainings- Validierungs- und Testdaten eines →KI-Systems, hinsichtlich ihrer faktischen Korrektheit, Vollständigkeit und Freiheit von ungerechtfertigter Verzerrung (Bias).

Datenschutz

Bewahrung und Unzugänglichkeit bestimmter als Information dokumentierter personenbezogener Daten.

Menschliche Aufsicht und Kontrolle (engl. Human Oversight and Control)

Eigenschaft eines →KI-Systems einschließlich seines Embeddings im Anwendungskontext, hinsichtlich der Möglichkeit für ein – fachlich entsprechend kompetentes – menschliches Individuum, das Verhalten und/oder die Funktionsweise dieses →KI-Systems grundsätzlich sowie während des laufenden Betriebs adäquat zu beobachten und zu ändern sowie ggfs. zu beenden.

Nicht-Diskriminierung (engl. Non-Discrimination)

Merkmal eines durch ein →KI-System durchgeführten, offenen Prozesses, wenn im Verlauf dieses Prozesses mehrere menschliche Individuen im Vergleich zueinander behandelt werden und dieser Prozess in juristischer Hinsicht frei von Schlechterbehandlung eines menschlichen Individuums aufgrund einer gesetzlich geschützten Eigenschaft ist.

Transparenz (engl. Transparency)

Eigenschaft eines →KI-Systems, das →erklärbar und →interpretierbar ist. Im Rahmen des vorliegenden Qualitätsstandards beinhaltet „Transparenz“ zudem eine Dokumentation der Eigenschaften des →KI-Systems.

Verlässlichkeit

Eigenschaft eines →KI-Systems, das eine hinreichende →Leistung, hinreichende →Robustheit aufweist und ein hinreichendes →Monitoring erlaubt.

4.1.3 Einzelne Kriterien

Erklärbarkeit (engl. Explainability)

Eigenschaft eines →KI-Systems im Hinblick auf die prinzipielle Verständlichkeit, und Nachvollziehbarkeit von Funktionalität, Verhalten und Output, für menschliches Fachpersonal, aber auch betroffene Personen und →Nutzer. Erklärbarkeit wird häufig als eine lokal und vom Systemdesign unabhängig („post-hoc“) gemessene Eigenschaft eines →KI-Modells oder →KI-Systems verstanden.

Interpretierbarkeit (engl. Interpretability)

Eigenschaft eines →KI-Modells, grundsätzlich in seinen Modellparametern, -gewichten oder anderen (mathematischen) Eigenschaften möglichst direkt nachvollziehbar und für Fachpersonal direkt verständlich zu sein. Interpretierbarkeit ist häufig explizit gegeben als Teil des Modellarchitekturdesigns, im Kontrast zur expliziten Wahl undurchsichtiger „black box“-Modelle (z. B. die Wahl eines interpretierbaren Entscheidungsbaums statt eines Neuralen Netzwerks für eine Klassifizierungsaufgabe).

Leistungsfähigkeit (engl. Performance)

Eigenschaft eines →KI-Systems, hinsichtlich der Fähigkeit, seine vorgegebenen Ziele und Zwecke möglichst vollständig zu erreichen.

Monitoring

Vorgehen, bei dem während des Betriebs eines →KI-Systems Abweichungen zwischen beobachtbaren Istzuständen und den angestrebten Sollzuständen detektiert werden.

Robustheit (engl. Robustness)

Fähigkeit eines →KI-Systems, seine reguläre und übliche Verhaltens- und Funktionsweise auch bei nicht-maliziösen, widrigen, störenden oder fehlerhaften Eingaben oder Einflüssen von außen bestmöglich beizubehalten.

Rückverfolgbarkeit (engl. Traceability)

Eigenschaft eines →KI-Systems im Hinblick auf die Erfassbarkeit der konsekutiven Folge aller Entscheidungen, die entlang des gesamten Lebenszyklus in ein →KI-System eingehen oder eingegangen sind.

4.1.4 Horizontale Konzepte**Dokumentation (engl. Documentation)**

Systematische Erfassung, Sammlung, Aufbewahrung und Bereitstellung von Informationen verschiedener Art, welche gesetzlichen, internen oder externen Vorgaben entsprechen.

Fairness

Merkmal eines durch ein →KI-System durchgeführten, offenen Prozesses, wenn im Verlauf dieses Prozesses mehrere menschliche Individuen im Vergleich zueinander behandelt werden und dieser Prozess in juristischer Hinsicht frei von Schlechterbehandlung eines menschlichen Individuums aufgrund einer gesetzlich geschützten Eigenschaft ist sowie zudem den Gerechtigkeitsvorstellungen von zu benennenden Individuen entspricht.

Safety

Eigenschaft eines →KI-Systems, hinsichtlich der Gefahrlosigkeit des Systems für menschliche Individuen hinsichtlich der Risiken für Leib, Leben und Gesundheit sowie für Sachen hinsichtlich Beschädigung im intendierten, funktionalen Regelbetrieb.

Security

Widerstandsfähigkeit eines →KI-Systems gegen maliziöse äußere Eingriffe und Manipulationen.

4.1.5 Weitere Begriffe**Anschlussfähigkeit**

Eigenschaft des →Qualitätsstandards in Abhängigkeit von Vereinbarkeit und Widerspruchsfreiheit mit anderen KI-Regularien wie der europäischen KI-Verordnung.

Anwendungsbereich

Die Gesamtheit der möglichen Eingabedaten, die für ein →KI-System relevant sind. Er umfasst die Kontexte, in denen ein →KI-System angewendet oder genutzt werden kann. In bestimmten Zusammenhängen kann der Anwendungsbereich synonym mit dem Begriff „Operational Design Domain“ (ODD) verwendet werden, um die spezifischen Bedingungen und Parameter zu definieren, unter denen ein →KI-System effektiv arbeitet.

Belastbarkeit

Eigenschaft des Qualitätsstandards in Abhängigkeit von →Prüftiefe und Objektivität des Testens, wobei der Bestätigungsgrad, externe Prüfungen sowie vergleichbare Faktoren diese entsprechend erhöhen.

Betroffene Personen

Natürliche Personen, die durch ein →KI-System beeinflusst oder betroffen werden, ohne notwendigerweise direkt oder aktiv mit dem System zu interagieren. Der Unterschied zu →Nutzern liegt in der Art der Interaktion: Während →Nutzer das →KI-System direkt und aktiv bedienen, beispielsweise als Endnutzer oder Benutzer, stehen betroffene Personen in einem eher indirekten und passiven Verhältnis zum →KI-System, können jedoch durch dessen Entscheidungen oder Funktionen beeinflusst werden.

Effizienz

Eigenschaft des Qualitätsstandards in Abhängigkeit von zeitlichen, personellen und finanziellen Aufwendungen bei gleichzeitiger Sicherstellung eines hohen Qualitätslevels.

KI-Anbieter (engl. AI Provider)

Ein „Anbieter“ ist eine natürliche oder juristische Person, Behörde, Einrichtung oder sonstige Stelle, die ein →KI-System oder ein →KI-Modell mit allgemeinem Verwendungszweck entwickelt oder entwickeln lässt und es unter ihrem eigenen Namen oder ihrer Handelsmarke in Verkehr bringt oder das →KI-System unter ihrem eigenen Namen oder ihrer Handelsmarke in Betrieb nimmt, sei es entgeltlich oder unentgeltlich. (vgl. Art. 3 Abs. 3 KI-VO)

KI-Betreiber (engl. AI Deployer)

Ein „Betreiber“ ist eine natürliche oder juristische Person, Behörde, Einrichtung oder sonstige Stelle, die ein →KI-System in eigener Verantwortung verwendet, es sei denn das →KI-System wird im Rahmen einer persönlichen und nicht beruflichen Tätigkeit verwendet. (vgl. Art. 3 Abs. 4 KI-VO)

KI-Komponente

Eine „KI-Komponente“ umfasst ein implementiertes →KI-Modell; ggf. gemeinsam mit den Methoden, die sich unmittelbar auf die Vor- oder Nachverarbeitung der Ein-/Ausgaben dieses Modells beziehen sowie deren Schnittstellen.

Beispiel: ein KI-Bilderkennungsmodell inkl. Methoden zur Vorverarbeitung der Bilder, das als Input rohe Bild-/Videodaten nimmt und als Output eine Aussage darüber trifft, ob ein Mensch auf dem Bild zu sehen ist.

KI-Lebenszyklus (engl. AI life cycle)

Entwicklung eines Systems, eines Produkts, einer Dienstleistung, eines Projekts oder einer anderen vom Menschen geschaffenen Einheit – welche KI nutzt – von der Konzeption bis zur Stilllegung. (in Anlehnung an und Fortführung von ISO/IEC 22989:2022)

KI-Modell

Ein „KI-Modell“ umfasst ausschließlich die funktionalen, KI-spezifischen Parameter, ggf. die Gewichte und Biases (inferentiellen Input-Output-Mappings) sowie die Architektur; nicht erfasst ist hierbei die weiterführende Implementierung und Einbindung, diese wird erst vom Begriff der →KI-Komponente miterfasst.

Beispiel: Ein neuronales Netz zur Bildverarbeitung, das Pixeln zugeordnete Zahlenwerte als Eingabe bekommt und als Ausgabe die Wahrscheinlichkeit ausgibt, dass auf dem Bild ein Mensch zu sehen ist.

KI-System

Ein „KI-System“ ist ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können. (vgl. Art. 3 Abs. 1 KI-VO) Im Rahmen des vorliegenden Qualitätsstandards wird ein KI-System dabei in technischer Hinsicht als ein funktionaler Zusammenschluss von einer oder mehreren →KI-Komponente(n) und Nicht-KI-Komponenten im Hinblick auf eine spezifische Zweckbestimmung und einen konkreten Anwendungskontext

verstanden. Mehrere KI-Systeme können zu einem größeren KI-System zusammengeschlossen werden. Komponenten, mit denen das KI-System interagieren kann, die aber nicht zwingend notwendig für das Funktionieren des KI-Systems oder beliebig austauschbar sind, werden nicht als Teil des KI-Systems erachtet.

Netzwerk

Zusammenschluss von mindestens zwei Computern oder anderen elektronischen Geräten, der den Austausch von Daten und die Nutzung gemeinsamer Ressourcen ermöglicht.

Niederschwelligkeit

Angebot oder Dienst, gekennzeichnet durch einen geringen erforderlichen Aufwand (z. B. kurze Prüfdauer) bei zeitgleich hohem Nutzen (hohe Qualität sichergestellt).

Nutzer

Natürliche Personen, die direkt und aktiv mit einem →KI-System interagieren, entweder als Endnutzer, die das →KI-System für persönliche oder geschäftliche Zwecke verwenden, oder als Benutzer, die das →KI-System im professionellen Kontext anwenden. Der Begriff umfasst sowohl diejenigen, die das →KI-System bedienen, als auch jene, die es für die Erreichung spezifischer Ziele nutzen. Je nach Kontext kann dies die Verwendung von →KI-Systemen in alltäglichen Anwendungen oder in spezialisierten, beruflichen Szenarien beinhalten.

Proprietäre Daten

Daten, welche eigentumsrechtlich zugeordnet sind, einschließlich lizenzgebundener und urheberrechtlich geschützter Daten.

Prüfer

Sind unabhängige interne oder externe natürliche Personen, welche zu keinem Zeitpunkt in die Entwicklung des →KI-Systems oder damit verbundene Strukturen involviert sein dürfen. Als natürliche Personen handeln sie in Vertretung juristischer Personen, beispielsweise als interne Prüfer des →KI-Anbieters bzw. →KI-Betreibers oder als externe →Prüfer einer unabhängigen Partei.

Prüfling (zu prüfende Organisation)

Der →KI-Anbieter bzw. →KI-Betreiber des zu prüfenden →KI-Systems wird als Prüfling bezeichnet. Der Prüfling wird durch eine oder mehrere natürliche Personen vertreten.

Prüfmethode

Methodisches Vorgehen zur Erhebung einzelner oder mehrerer inhaltlich zusammenhängender Prüfevidenzen und ihrer Bewertung im Kontext des Prüfverfahrens. Eine Prüfmethode kann sowohl manuelle als auch automatisierte Anteile enthalten. Unter den Begriff der technischen Prüfmethoden fallen solche, bei denen sich die Erhebung der Prüfevidenzen wesentlich auf technische Verfahren und/oder Hilfsmittel stützt.

Prüftiefe

Die Prüftiefe legt fest, mit welchem Aufwand die Prüfung durchgeführt wird, bestimmt den Grad der Absicherung und Plausibilität von Prüfaussagen und die legt die Rollen fest, die zur Validierung der Ergebnisse im Prüfverfahren gebraucht werden. Die Prüftiefe setzt sich aus drei zentralen Eigenschaften der Prüfung zusammen: 1) Detailgrad der Maßnahmen je nach Prüfanforderungen aus dem Prüfkatalog 2) Den bereitzustellenden Evidenzen 3) Dem Grad an Validierung.

Prüfverfahren

Vorgehen im Rahmen der Prüfung, welches die Bestimmung der →Prüftiefe, der Methodik zum Einholen von Evidenzen, der Vergleichbarkeit sowie der Bewertung der →Kriterien umfasst.

Qualitätsstandard

Ein Standard ist ein Dokument, das ein Verfahren festlegt, sodass objektiv prüfbar ist, ob ein Prüfgegenstand konform mit den im Standard festgelegten Kriterien ist. Ein Qualitätsstandard ist ein Standard, dessen Kriterien Rückschlüsse über die Qualität des Prüfgegenstandes erlauben.

Vergleichbarkeit

Eigenschaft des → Qualitätsstandards in Abhängigkeit von Einheitlichkeit der Prüfungen und Replizierbarkeit der Tests mit dem Ziel der Objektivierbarkeit.

Zugänglichkeit

Eigenschaft von entweder einer Prüfaussage, welche Auskunft über ihre Verständlichkeit gibt oder von einem Prüfansatz, welcher das Ausmaß der Anwendbarkeit auf bestimmte Zielgruppen beschreibt.

4.2 Anwendungsfallbeschreibungsvorlage

Allgemeine Informationen	
Fragestellungen	Angaben
1 Name des Anwendungsfalls	
2 Beschreiben Sie den Anwendungsfall. (maximal 300 Wörter)	
3 Welche Aufgabenstellung wird durch das KI-System gelöst? (maximal 300 Wörter)	
4 Was ist der Eingabebereich des KI-Systems, d. h. in welchen Situationen/bei welchen Eingaben soll das KI-System funktionieren ¹¹ ?	
5 Wo liegen die Grenzen des Eingabebereichs des KI-Systems, d. h. in welchen Situationen/bei welchen Eingaben funktioniert das KI-System nur eingeschränkt oder gar nicht mehr?	
6 Was ist der Ausgabebereich des KI-Systems?	
7a Für welche Nutzergruppe ist der Einsatz des KI-Systems denkbar? Gibt es Voraussetzungen bzgl. Ausbildung oder Wissen dieser Nutzergruppe?	
7b Welche Gruppen von Personen sind möglicherweise unmittelbar durch die Ausgaben des KI-Systems (positiv und/oder negativ) betroffen ¹² ?	
8a Ist vorgesehen, dass Menschen am Betrieb des KI-Systems beteiligt sind?	<input type="checkbox"/> Ja <input type="checkbox"/> Nein
8b Sind Menschen an der Aufsicht des KI-Systems beteiligt?	<input type="checkbox"/> Ja <input type="checkbox"/> Nein
8c Falls ja, müssen sie aktiv eingreifen, um Einfluss zu nehmen, oder ist die Prüfung durch einen Menschen notwendiger Teil des Prozesses?	<input type="checkbox"/> Ja <input type="checkbox"/> Nein
9 Gibt es besondere regulatorische Anforderungen bezogen auf den Einsatzkontext ¹³ ?	

11 Der Eingabebereich kann beispielsweise durch Rahmenbedingungen oder technische Voraussetzungen eingeschränkt sein. Im Kontext autonomen Fahrens spricht man hier auch von einer Operational Design Domain (ODD). Ein Beispiel wäre eine Objekterkennung, die nur unter bestimmten Lichtbedingungen funktioniert.

12 Damit sind die Personengruppen gemeint, auf die sich unentdeckte fehlerhafte Ausgaben des KI-Systems auswirken, bspw. Bewerber bei einem Bewerbungstool oder Leser bei automatisch generierten Zeitungsartikeln.

13 Beispielsweise Anforderungen folgend aus der Einstufung nach KI-Verordnung oder Anforderungen aus sektoraler Regulierung. Nachweise über erfüllte Anforderungen können im weiteren Verlauf der Prüfung als Evidenzen dienen.

Allgemeine Informationen

Fragestellungen	Angaben
10 Betreiben Sie Ihre KI-Modelle on-premise, lokal oder in der Cloud?	<input type="checkbox"/> Cloud <input type="checkbox"/> on-premise <input type="checkbox"/> hybrid <input type="checkbox"/> hybrid
11 Wird mindestens ein Modell im Betrieb weitertrainiert?	<input type="checkbox"/> Ja, das Modell trainiert kontinuierlich weiter <input type="checkbox"/> Ja, aber weiteres Training muss manuell initialisiert werden <input type="checkbox"/> Nein
12 Architektur	<input type="checkbox"/> Large Language Model <input type="checkbox"/> Entscheidungsbaum <input type="checkbox"/> Convolutional NN <input type="checkbox"/> Genetischer Algorithmus <input type="checkbox"/> Random forest <input type="checkbox"/> Regression <input type="checkbox"/> SVM <input type="checkbox"/> Clustering <input type="checkbox"/> Anderes

4.3 Schutzbedarfsanalyse

Allgemein ist die Schutzbedarfsanalyse eine Untersuchung der Schutzbedarfe bezüglich der einzelnen Qualitätsdimensionen und ihrer dazugehörigen Kriterien. Sie dient der Vorfilterung der für den jeweiligen Anwendungsfall relevanten Qualitätsdimensionen auf Kriterienebene und der Bestimmung eines zu erreichenden Sollwertes für die nachfolgende Prüfung.

Der Hintergrund ist, dass je nach Aufgabe und Einsatzgebiet eines KI-Systems nicht alle Qualitätsmerkmale gleich wichtig sind. Auch innerhalb eines Merkmals können einzelne Punkte unterschiedlich stark ins Gewicht fallen, wenn man die Qualität des KI-Systems bewertet.

So ermöglicht die Schutzbedarfsanalyse effizient, gezielt relevante Qualitätsdimensionen auf Kriterienebene zu identifizieren und die Höhe des ihnen korrespondierenden Schutzbedarfes zu bestimmen.

Zugleich ist die Schutzbedarfsanalyse ein praktikabler Weg, eine Anschlussfähigkeit an die europäische KI-Verordnung (KI-VO) zu erreichen. Diese geht v. a. bei Hochrisiko-KI-Systemen, aber auch allgemein für KI-Systeme von deren jeweiliger Zweckbestimmung aus und betrachtet die Kombination des KI-Systems mit der jeweiligen Aufgabe und dem Anwendungskontext. Gegenstand der Schutzbedarfsanalyse ist also das KI-System in seinem intendierten Anwendungskontext („Use Case“).

Ergänzend zu den Angaben im Hauptteil dieses Dokuments werden im vorliegenden Anhang folgende Punkte näher erläutert: zunächst werden einige relevante Definitionen für die interne Systematik der Schutzbedarfsanalyse angeführt (4.3.1) und die Abwägung zwischen Risikobewertung und Schutzbedarfsanalyse erörtert (4.3.2). Nach einer systematischen Detailbeschreibung der Schutzbedarfsanalyse (4.3.3) und den Schutzzielen (4.3.4) wird abschließend die Anwendung und Durchführung (4.3.5) des Frameworks erläutert.

4.3.1 Definitionen

Für die Beschreibung der sachlichen Konstellation der Schutzbedarfsanalyse seien für die internen Belange dieses Dokuments folgende Begriffe vorab definiert:

- **Gefahr** erfasst eine Sachlage, bei der die Möglichkeit eines Schadenseintritts besteht.
- **Risiko** beschreibt, entsprechend der gängigen Nutzung, das Produkt aus der Wahrscheinlichkeit des Schadenseintritts und dem Ausmaß des zu erwartenden Schadens.
- **Schutz** bezeichnet die Abwehr, Mitigation oder Verhinderung eines Schadens oder einer Gefährdung.
- **Schutzbedarf** ist das angemessene oder notwendige Maß an Schutz für ein Schutzziel und orientiert sich an dem Ausmaß des bei Verletzungen jeweils drohenden Schadens.
- **Schutzziel** meint die jeweilige Eigenschaft aufseiten der betroffenen Entitäten mit Blick auf die Regulierung oder Prüfung. Schutzziele im Sinne dieser Schutzbedarfsanalyse sind die im Art. 1 Abs. 1 KI-VO genannten Werte (Gesundheit, Sicherheit und die in der EU-Charta verankerten Grundrechte, einschließlich Demokratie, Rechtsstaatlichkeit und Umweltschutz).

4.3.2 Risikobewertung vs. Schutzbedarfsanalyse

Nach reiflicher Überlegung und eingehenden Diskussionen im Rahmen des Projekts wird die Schutzbedarfsanalyse anstelle einer Risikobewertung eingesetzt. Diese Modifikation beruht auf mehreren Argumenten.

Zuerst ist hier die Schwierigkeit des Risikobegriffs im Kontext von KI zu nennen. Basierend auf der zuvor beschriebenen gängigen Definition umfasst der Risikobegriff die Berücksichtigung von Wahrscheinlichkeiten. Dazu zählen sowohl die Eintrittswahrscheinlichkeit eines Schadens als auch die Ausfallwahrscheinlichkeiten der zu prüfenden Entität. Bislang liegen im KI-Bereich aber noch keine Erfahrungswerte im erforderlichen Maße und in der erforderlichen Breite vor. So ist aktuell eine Ermittlung der jeweiligen Wahrscheinlichkeiten, insbesondere mit Blick auf Skalierbarkeit, nur unter größten Schwierigkeiten umsetzbar.

Weiterhin geht eine Risikobewertung aufgrund der kombinierten Betrachtung der Schadenswahrscheinlichkeit und des Schadensausmaßes mit einer deutlich aufwendigeren Prüfung einher. Dagegen ist eine Schutzbedarfsanalyse, wie im Folgenden beschrieben, deutlich effizienter durchführbar.

Eine Schutzbedarfsanalyse erlaubt eine Vorfilterung der für das KI-System relevanten Qualitätsdimensionen auf Kriterienebene, indem die Fragen der Schutzbedarfsanalyse die Feststellung ermöglichen, welche der Qualitätsdimensionen und ihrer darunter gruppierten Kriterien für das untersuchte KI-System konkret anwendbar sind. Dies erhöht im entscheidenden Maße die Effizienz der Prüfungsdurchführung und folglich auch die Attraktivität des Prüfprodukts.

Aus diesen Gründen wird statt einer Risikobewertung eine Schutzbedarfsanalyse durchgeführt. Diese reduziert den Prüfaufwand und adressiert zeitgleich die jeweiligen Schutzbedarfe je Qualitätsdimension auf Kriterienebene. Der grundlegende Leitsatz für die Schutzbedarfsfeststellung ist eine Aufnahme an Fragen zu Gefahren und Gefährdungsfaktoren, wenn diese tatsächlich einen Einfluss auf die Höhe und Art des Schadens haben. Im Rahmen der Schutzbedarfsanalyse werden lediglich sogenannte „first order effects“ betrachtet. Dies ist darin begründet, dass eine weitergehende Erfassung von Folgeeffekten in Anbetracht der sich daraus ergebenden Möglichkeiten nicht umfassend auf effiziente Weise abdeckbar ist.

4.3.3 Systematische Beschreibung der Schutzbedarfsanalyse

Für das systematische Verständnis der Schutzbedarfsanalyse ist folgende Gesamtkonstellation zu betrachten: Ein KI-System hat über seine spezifische Verfasstheit und seine Zweckbestimmung die Möglichkeit, auf andere Entitäten (Menschen, Tiere, Dinge, die Gesellschaft, die Umwelt im Ganzen) einzuwirken. Diese Entitäten haben Eigenschaften, denen eine intrinsische Werthaftigkeit zugesprochen wird. Das sind etwa Grundwerte wie Leben, Gesundheit, Unversehrtheit sowie Fortbestand und Schadfreiheit aber auch die über Grundrechte formulierten Werte, beispielsweise die Gleichberechtigung der Geschlechter.

Je nach seiner Verfasstheit stellt ein KI-System in seiner Grundverfassung, vor allem aber durch seinen intendierten Verwendungszweck, in je spezifischen Hinsichten potenziell eine Gefährdung für diese Eigenschaften der genannten Entitäten und somit ein Schadenspotential dar. In der intendierten Abwendung dieses Schadens werden die genannten Grundwerte entsprechend zu Schutzzielen. Für den Einsatz des KI-Systems ergibt sich – je nach Ausmaß der Gefährdung und des Schadenspotentials – ein Schutzbedarf in bestimmter Höhe. Hierbei greift eine Schutzbedarfsanalyse, wie in Abschnitt 03 erläutert, nur auf die Höhe und das Ausmaß des Schadens, nicht aber auf exakt quantifizierte Eintrittswahrscheinlichkeiten zurück.

Aus dem bestehenden Gefahren- und Schadenspotenzial sowie den gewählten Schutzzielen ergeben sich die Eigenschaften, die das untersuchte KI-System aufweisen muss, um diese Gefährdungen zu minimieren. Diese Eigenschaften des KI-Systems werden durch die entsprechenden Prüfungen der verschiedenen Qualitätsdimensionen (wie VE: Verlässlichkeit, CY: KI-spezifische Cybersicherheit usw.) sichergestellt. Soweit der Schutzbedarf je Qualitätsdimension und Kriterium graduierbar ist, ist auch die Ausprägtheit dieser Qualitätsdimensionen aufseiten des KI-Systems unterschiedlich stark notwendig.

Entsprechend dieser Gesamtkonstellation ermittelt die Schutzbedarfsanalyse folglich für jede einzelne Qualitätsdimension auf Kriterienebene den Schutzbedarf, je mit Blick auf die gewählten Schutzziele. Sie stellt also mittels gezielter Fragen fest, wie hoch der Schutzbedarf für jede einzelne Qualitätsdimension auf Kriterienebene zu veranschlagen ist. Dabei berücksichtigt sie das konkrete KI-System in seinem Gebrauchskontext – etwa das Kriterium „Rückfallpläne und funktionale Sicherheit“ der Qualitätsdimension VE (Verlässlichkeit). Diese Feststellung bildet als Soll-Zustand einen Mindestwert, der in den nachfolgenden Prüfungen für diese Qualitätsdimension auf Kriterienebene jeweils zu erreichen ist.

Für eine möglichst effiziente Durchführung der Schutzbedarfsanalyse (und der nachfolgenden Prüfung) wurden zwei weitere Entscheidungen getroffen: erstens hinsichtlich der Anwendbarkeit der Qualitätsdimensionen auf Kriterienebene und zweitens hinsichtlich der Graduierung des Schutzbedarfs.

Im Rahmen der ersten Entscheidung wird vor der detaillierten Ermittlung des Schutzbedarfs geprüft, ob die Kriterien einer Qualitätsdimension jeweils sinnvoll auf das zu überprüfende KI-System anwendbar sind. So ist beispielsweise das Kriterium „Vermeidung von ungerechtfertigter Diskriminierung und Verzerrung“ der Qualitätsdimension ND (Nicht-Diskriminierung) möglicherweise nur eingeschränkt auf ein technisches System für vorausschauende Wartung anwendbar. Nur wenn diese Anwendbarkeit grundsätzlich gegeben ist, wird der Schutzbedarf für das fragliche Kriterium einer Qualitätsdimension im Detail ermittelt. Das soll sicherstellen, dass – sowohl bei der Schutzbedarfsanalyse selbst als auch bei der Durchführung der Prüfungen – nur solche Qualitätsdimensionen auf Kriterienebene geprüft werden, die überhaupt sinnvoll auf das konkrete KI-System anwendbar sind. Zudem soll gewährleistet werden, dass diese Qualitätsdimensionen mit dem gegebenen Schutzbedarf korrespondieren. Ausnahmen sind bei diesem Vorgehen die Kriterien „Leistungsfähigkeit und Robustheit“ und „Rückfallpläne und funktionale Sicherheit“ der Qualitätsdimension VE (Verlässlichkeit), das Kriterium „Datenqualität“ der Qualitätsdimension DA (Datenqualität, -schutz und -Governance) und das Kriterium „Rückverfolgbarkeit & Dokumentation“ der Qualitätsdimension TR (Transparenz) – sie werden als so grundlegend für die Qualität von KI-Systemen erachtet, dass sie immer geprüft werden.

Bezüglich der Graduierung des Schutzbedarfs wurde entschieden, im Sinne der Effizienz einen Mittelweg zwischen einer bloß binären Unterscheidung (Schutzbedarf: ja/nein) und einer allzu starken Granularität zu wählen. Im Sinne dieser Entscheidung wird der Schutzbedarf in drei Stufen („gering“, „moderat“, „hoch“) festgestellt. Sind Fragen zum Schutzbedarf nur mit „Ja“ oder „Nein“ zu beantworten, wird bei der Antwort „Ja“ ein hoher Schutzbedarf angenommen. Eine Ausnahme bildet das Schutzziel „Gesundheit“. Hier wird mit „Ja“ ein moderater Schutzbedarf angenommen, um eine Differenzierung zum Schutzziel „Leib und Leben“ vorzunehmen.

Nach der Beantwortung der Einzelfragen wird der Schutzbedarf auf Ebene der einzelnen Kriterien einer Qualitätsdimension nach einem spezifischen Verfahren aggregiert. Damit wird, wie angeführt, der Mindestwert formuliert, den es für ein erfolgreiches Bestehen der Prüfung für jedes einzelne Kriterium einer Qualitätsdimension zu erreichen gilt. Das genaue Verfahren zur Aggregation ist im Prozessablauf der Gesamtbewertung (siehe 4.5) beschrieben.

Im Ganzen betrachtet ermittelt die Schutzbedarfsanalyse durch ein effizientes Frageformat den Schutzbedarf für jede einzelne zu prüfenden Qualitätsdimension auf Kriterienebene und mit Blick auf die nachfolgend angeführten Schutzziele. Der ermittelte Schutzbedarf bildet den Soll-Zustand für die nachfolgende Prüfung des KI-Systems auf Ebene der einzelnen Kriterien je Qualitätsdimension.

4.3.4 Schutzziele

Die Schutzziele ergeben sich – im Sinne der Anschlussfähigkeit an die KI-VO – aus Art. 1 Abs. 1 (KI-VO). Dieser Artikel listet Gesundheit, Sicherheit und die in der EU-Charta verankerten Grundrechte, einschließlich Demokratie, Rechtsstaatlichkeit und Umweltschutz. Die KI-VO fordert für diese Schutzziele die Gewährleistung eines hohen Schutzniveaus. Um den Prüfaufwand zugänglich und niederschwellig zu halten, werden die Schutzziele zur Vermeidung einer unnötigen Detailtiefe in handhabbare Kategorien eingeteilt und erfasst. Diese Kategorien umfassen Leib und Leben, Gesundheit, Grundrechte, Eigentum und Sachen, Umwelt, Schutz personenbezogener Daten, Menschenwürde und Nichtdiskriminierung.

4.3.5 Durchführung der Schutzbedarfsanalyse

Für die Durchführung der Schutzbedarfsanalyse wird das Template in 4.3.6 verwendet. Die im Template enthaltene Fragenliste wird für jedes Kriterium und jede Qualitätsdimension durchgearbeitet, um zu ermitteln, ob das KI-System potenzielle Gefährdungen für die jeweiligen Schutzziele aufweist. Grundlage für die Beantwortung ist die Anwendungsfallbeschreibung. Bei Bedarf können weitere Informationen für einzelne Fragen der Schutzbedarfsanalyse beim Prüfling eingeholt werden.

Jeder Fragesatz pro Qualitätsdimension wird durch Applikationsfragen (AF) eingeleitet. Diese prüfen grundsätzlich, welche Kriterien je Qualitätsdimension überhaupt sinnvoll für das gegebene KI-System im Hinblick auf den Schutzbedarf anwendbar sind. Kriterien können entweder keine (sofern eine der Ausnahmen greift und das Kriterium immer anwendbar ist, vgl. 04), eine oder mehrere AF haben. Hat ein Kriterium mehr als eine AF, gilt es dann als „anwendbar“, wenn mindestens eine der AF mit „Ja“ beantwortet wurde. Die generelle Anwendbarkeit hat keinen Einfluss auf die Höhe des Schutzbedarfs. Ist ein Kriterium basierend auf diesen AF auf das gegebene KI-System nicht anwendbar, ist die Höhe des Schutzbedarfs gemäß den Ausführungen in Abschnitt 04 nicht sinnvoll zu bestimmen und die Beantwortung der weiteren Fragen zu diesem Kriterium entfällt. Dies steigert die Effizienz der Prüfung zweifach. Einerseits reduziert sich der Prüfaufwand, da keine weiteren Fragen das Kriterium betreffend beantwortet werden müssen. Andererseits kann das entsprechende Kriterium bei der weiteren Prüfung des KI-Systems außer Betracht bleiben.

Im Anschluss an die AF werden sogenannte Grundfragen (GF) gestellt. Antworten auf diese Fragen geben eine zwingende Mindesthöhe für die Höhe des Schutzbedarfs vor. Bei mehreren GF bestimmt der höchste Schutzbedarf den Schutzbedarf des gesamten Kriteriums. Den GF kommt folglich auf diese Weise besonderes Gewicht zu: sie betreffen Schutzbedarfe, die – etwa im Falle von Personenschäden – nicht durch andere, geringere Schutzbedarfe ‚ausgeglichen‘ oder gemindert werden können.

Als dritter Fragentyp folgen Erweiterungsfragen (EF). Die Antworten auf diese Fragen bilden pro Kriterium je Qualitätsdimension den Durchschnitt. Anders als die GF ermitteln die EF einen „durchschnittlichen Gesamteindruck“ des Schutzbedarfs für Schutzziele, deren Relevanz als relativ gering, aber dennoch nicht irrelevant im Vergleich zu den über die GF abgefragten Schutzbedarfen eingeschätzt wird. Bei < 0.5 in der ersten Dezimale wird im Rahmen der Durchschnittsbildung abgerundet, bei ≥ 0.5 wird aufgerundet.

Je Frage wird eine Spalte mit der jeweiligen Schutzzielkategorie aufgeführt. Die Selektion der Schutzzielkategorien richtet sich dabei nach den potenziellen Schäden, die bei einer Verletzung der Anforderungen an das Kriterium einer Qualitätsdimension auftreten können. Lässt sich eine Frage keiner klaren Schutzzielkategorie zuordnen, wird sie unter der Schutzzielkategorie „allgemein“ gelistet. Einzelne Fragen sind für die Einstufung mehrerer Kriterien oder Qualitätsdimensionen relevant und kommen deswegen mehrfach vor.

Für die adäquate Beantwortung der Fragen ist jeweils das KI-System als solches zu betrachten, also sowohl mit Blick auf seinen intendierten Verwendungszweck als auch auf seinen Anwendungskontext, aber noch ohne Berücksichtigung von implementierten Maßnahmen oder Ähnlichem.

Entsprechend wird der Schutzbedarf mit Blick auf den Verwendungszweck als solchen bestimmt. Eine Abfrage des Vorliegens ggf. notwendiger Mitigierungs-, Sicherheits- oder Schutzmaßnahmen wird im Rahmen der nachfolgenden VCIO-Prüfung abgefragt.

Die Antworten auf die Einzelfragen werden nach einem spezifischen Verfahren zum Gesamtschutzbedarf eines Kriteriums je Qualitätsdimension aggregiert. Dieses Aggregationsverfahren ist im Sinne der Effizienz eng mit der Durchführung der Schutzbedarfsanalyse verzahnt. Die Aggregation des Schutzbedarfs gestaltet sich dabei, in Verbindung mit der Durchführung der Schutzbedarfsanalyse, pro Kriterium je Qualitätsdimension wie folgt:

Schritt 1: Beantwortung der AF. Ergeben diese durchgängig, dass keine Anwendbarkeit gegeben ist, kann die Schutzbedarfsanalyse für dieses Kriterium mit Schritt 7 der vorliegenden Beschreibung beendet werden. Gleiches gilt für die Ausnahmen (vgl. 04), welche Kriterien betreffen, die immer anwendbar sind und für welche es folglich keine AF gibt. Ist Anwendbarkeit gegeben, ist mit Schritt 2 fortzufahren.

Schritt 2: Beantwortung der GF.

Schritt 3: Bestimmung des Maximalwertes für die GF. Falls eine der Grundfragen mit „3“ bzw. „Hoch“ beantwortet ist, ist der Schutzbedarf für dieses Kriterium insgesamt auf „Hoch“ zu setzen und die Schutzbedarfsanalyse – im Sinne der Effizienz – für dieses Kriterium nach der entsprechenden Frage mit Schritt 7 der vorliegenden Beschreibung zu beenden. (Ist ein Gesamteindruck des Schutzbedarfs gewünscht, kann die Analyse auch weiter durchgeführt werden).

Schritt 4: Beantwortung der EF.

Schritt 5: Ermittlung des Durchschnittswertes für die EF gemäß den obenstehenden Bestimmungen zu den EF.

Schritt 6: Gilt: **Niveau aus GF \geq Durchschnitt aus EF**, so ist der Schutzbedarf für dieses Kriterium gleich dem höchsten Schutzbedarf aus den GF.

Schritt 7: Gilt: **Niveau aus GF $<$ Durchschnitt aus EF**, so ist der Schutzbedarf für dieses Kriterium gleich dem Durchschnitt aus allen GF und EF, wobei abermals auf die obigen Bestimmungen zur Durchschnittsbildung zurückzugreifen ist. Gibt es keine GF, entspricht der Durchschnitt aus EF dem Schutzbedarf für das Kriterium.

Schritt 8: Damit ist der Schutzbedarf für das untersuchte Kriterium bestimmt. Er ist oben im Reiter für die untersuchte Qualitätsdimension des Kriteriums im dafür vorgesehenen Feld zu vermerken.

Schritt 9: Nachfolgend ist mit dem nächsten Kriterium der Qualitätsdimension oder (wenn alle Kriterien der Qualitätsdimension bereits bearbeitet wurden) im nächsten Reiter mit den Kriterien der nächsten Qualitätsdimension **fortzufahren**.

4.3.6 Template Schutzbedarfsanalyse

Schutzbedarfsanalyse



Überblick der Reiter

Reitername	Inhalt
DA	Fragenkatalog zur Qualitätsdimension Datenqualität, -schutz und -Governance
ND	Fragenkatalog zur Qualitätsdimension Nicht-Diskriminierung
TR	Fragenkatalog zur Qualitätsdimension Transparenz
MA	Fragenkatalog zur Qualitätsdimension Menschliche Aufsicht und Kontrolle
VE	Fragenkatalog zur Qualitätsdimension Verlässlichkeit
CY	Fragenkatalog zur Qualitätsdimension KI-spezifische Cybersicherheit

Dieses Dokument ist Teil des MISSION KI Qualitätsstandards. ©acatech – Deutsche Akademie der Technikwissenschaften e.V.
Dieses Werk ist lizenziert unter der Creative Commons Lizenz Namensnennung – Keine Bearbeitungen 4.0 International (CC BY-ND 4.0).
<https://creativecommons.org/licenses/by-nd/4.0/deed.de>

Schutzbedarfsanalyse

Qualitätsdimension: Datenqualität, -schutz und -Governance (DA)

Gesamteinschätzung je Kriterium

Schutzbedarf Kriterium „Datenqualität“:			1 / 2 / 3	immer anwendbar
Schutzbedarf Kriterium „Schutz personenbezogener Daten“:			1 / 2 / 3 / nicht anwendbar	
Schutzbedarf Kriterium „Schutz proprietärer Daten“:			1 / 2 / 3 / nicht anwendbar	

Frageindex	Frage	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
	Applikationsfragen	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
DA-Z12	Verarbeitet oder erzeugt das KI-System personenbezogene Daten oder wurden diese im Rahmen des Trainings des KI-Modells bzw. der KI-Modelle verwendet? ja = 3 nein = 1	allgemein	Schutz personenbezogener Daten			"Verarbeiten" meint explizit keine Schritte um eben diese Daten für weitere Prozesse zu entfernen, auszuschließen oder zu anonymisieren, sofern diese nicht gespeichert werden
DA-Z13	Verarbeitet oder erzeugt das KI-System proprietäre Daten oder wurden diese im Rahmen des Trainings des KI-Modells bzw. der KI-Modelle verwendet? ja = 3 nein = 1	allgemein	Schutz proprietärer Daten			"Verarbeiten" meint explizit keine Schritte um eben diese Daten für weitere Prozesse zu entfernen, auszuschließen oder zu anonymisieren, sofern diese nicht gespeichert werden
	Grundfragen	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
DA-Z16	Wie hoch ist der Schutzbedarf bezüglich des Kriteriums „Leistungsfähigkeit und Robustheit“ (VE)? gering = 1 moderat = 2 hoch = 3	allgemein	Datenqualität			Automatik einfügbar zur Übernahme der Einschätzung des Schutzbedarfs
DA-Z17	Wie hoch ist der Schutzbedarf bezüglich des Kriteriums „Schutz personenbezogener Daten“ (DA)? gering = 1 moderat = 2 hoch = 3	allgemein	Datenqualität			Automatik einfügbar zur Übernahme der Einschätzung des Schutzbedarfs
DA-Z18	Wie hoch ist der Schutzbedarf bezüglich des Kriteriums „Vermeidung von ungerechtfertigter Diskriminierung und Verzerrung“ (ND)? gering = 1 moderat = 2 hoch = 3	allgemein	Datenqualität			Automatik einfügbar zur Übernahme der Einschätzung des Schutzbedarfs
DA-Z19	Verarbeitet das KI-System personenbezogene Daten angelehnt an Art. 9 Abs. 1 DSGVO? ja = 3 nein = 1	allgemein	Schutz personenbezogener Daten			Personenbezogene Daten angelehnt an Art. 9 Abs. 1 DSGVO sind Daten, aus denen die rassische oder ethnische Herkunft, politische Meinungen, religiöse oder weltanschauliche Über-zeugungen oder die Gewerkschaftszugehörigkeit einer natürlichen Person hervorgehen.
DA-Z20	Verarbeitet das KI-System genetische Daten, biometrische Daten zur eindeutigen Identifizierung einer natürlichen Person, Gesundheitsdaten oder Daten zum Sexualleben oder der sexuellen Orientierung einer natürlichen Person? ja = 3 nein = 1	allgemein	Schutz personenbezogener Daten			
DA-Z21	Können mangelhafter Schutz oder mangelhaftes Management der von dem KI-System genutzten oder erzeugten Daten – im realistischen worst-case – tödliche Personenschäden verursachen? ja = 3 nein = 1	Leib und Leben	Schutz personenbezogener Daten			Zu denken ist beispielsweise politische Verfolgung: in einem solchen Falle kann mangelhafter Schutz sensibler Daten durch den Zugriff entsprechender Stellen Lebensgefahr für die jeweilige Person bedeuten.
DA-Z22	Können mangelhafter Schutz oder mangelhaftes Management der von dem KI-System genutzten oder erzeugten Daten – im realistischen worst-case – Gesundheitsschäden (körperliche Gesundheit) bei natürlichen Personen verursachen? ja = 2 nein = 1	Gesundheit	Schutz personenbezogener Daten			Zu denken ist beispielweise an Angriffe auf Politiker, auch in diesem Fall vor dem Hintergrund eines mangelhaften Daten-schutzes.
DA-Z23	Kann das KI-System die Überwachung von natürlichen Personen erleichtern bzw. stärken oder dient sie unmittelbar der Überwachung von natürlichen Personen? ja = 3 nein = 1	allgemein	Schutz personenbezogener Daten			

Frageindex	Frage	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
	Erweiterungsfragen	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
DA-Z26	Kann das KI-System prinzipiell, angesichts der Aufgabenstellung und Datengrundlage, personenbezogene Daten erzeugen? ja = 3 nein = 1	allgemein	Schutz personenbezogener Daten			
DA-Z27	Können die Ergebnisse des KI-Systems prinzipiell verwendet werden, um einen Personenbezug in den Daten herzustellen? ja = 3 nein = 1	allgemein	Schutz personenbezogener Daten			
DA-Z28	Liegt für das KI-System ein Data Protection Impact Assessment (DPIA) mit hoher bzw. kritischer Einstufung im Ergebnis vor? ja = 3 nein = 1	allgemein	Schutz personenbezogener Daten			
DA-Z29	Wie häufig wird das KI-System während des Betriebs weitertrainiert oder durch neue Versionen aktualisiert? nie = 1 hin und wieder = 2 oft = 3	allgemein	Datenqualität			Beispielhafte Grenzen: nie = kein einziges mal hin und wieder = monatlich oder seltener oft = häufiger als monatlich
DA-Z30	Werden personenbezogenen Daten anonymisiert, bevor sie für das KI-System verwendet werden? alle und immer = 1 mit leichten bis moderaten Einschränkungen = 2 mit großen oder größten Einschränkungen = 3	Schutz personenbezogener Daten	Schutz personenbezogener Daten			Beispielhafte Grenzen: alle und immer = Anonymisierung (alle Daten sind vollständig anonymisiert) mit leichten bis moderaten Einschränkungen = Pseudonymisierung (Daten sind pseudonymisiert, jedoch besteht ein Risiko der Re-Identifizierung) mit großen oder größten Einschränkungen = Daten werden weder anonymisiert noch pseudonymisiert oder der Prozentsatz liegt unter 50 %
DA-Z31	Wie hoch kann der finanzielle Schaden für das Unternehmen maximal sein, falls geistiges Eigentum nicht ausreichend geschützt wird? Der potentiell finanzielle Schaden dient hier als stellvertretende Bemessung des Schadens für den oder die Betroffenen. kein bis niedriger finanzieller Schaden = 1 moderater finanzieller Schaden = 2 schwerwiegender oder katastrophaler finanzieller Schaden = 3	Eigentum und Sachen	Schutz proprietärer Daten			Beispielhafte Grenzen: kein bis niedriger finanzieller Schaden = Schaden in Höhe von bis zu 1 % des Vorjahresumsatzes oder unter 1.000€ moderater finanzieller Schaden = Schaden in Höhe von 1 % bis 5 % des Vorjahresumsatzes oder zwischen 1.000€ bis 50.000€ schwerwiegender oder katastrophaler finanzieller Schaden = Schaden in Höhe von über 5 % des Vorjahresumsatzes oder über 50.000€
DA-Z32	Wie hoch kann der finanzielle Schaden für das Unternehmen maximal sein, falls personenbezogene Daten nicht ausreichend geschützt werden (kumulierter Schaden)? Der potentiell finanzielle Schaden dient hier als stellvertretende Bemessung des Schadens für den oder die Betroffenen. kein bis niedriger finanzieller Schaden = 1 moderater finanzieller Schaden = 2 schwerwiegender oder katastrophaler finanzieller Schaden = 3	Eigentum und Sachen	Schutz personenbezogener Daten			Beispielhafte Grenzen: kein bis niedriger finanzieller Schaden = Schaden in Höhe von bis zu 1 % des Vorjahresumsatzes oder unter 1.000€ moderater finanzieller Schaden = Schaden in Höhe von 1 % bis 5 % des Vorjahresumsatzes oder zwischen 1.000€ bis 50.000€ schwerwiegender oder katastrophaler finanzieller Schaden = Schaden in Höhe von über 5 % des Vorjahresumsatzes oder über 50.000€

Schutzbedarfsanalyse

Qualitätsdimension: Nicht-Diskriminierung (ND)

Gesamteinschätzung je Kriterium

Schutzbedarf Kriterium „Vermeidung von ungerechtfertigter Diskriminierung und Verzerrung“:		1 / 2 / 3 / nicht anwendbar
--	--	-----------------------------

Frageindex	Frage	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Erläuterungen und Beispiele
Applikationsfragen		Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
ND-Z10	<p>Betreffen die Ergebnisse oder das Verhalten des KI-Systems natürliche Personen auf unterschiedliche Art und Weise und ist diese Ungleichbehandlung mit der Ausprägung geschützter Merkmale (Rasse, ethnische Herkunft, Geschlecht, Religion oder Weltanschauung, Behinderung, Alter, sexuelle Identität) bei diesen Personen verbunden?</p> <p>Im Falle einer Selbstbewertung ist eine Erläuterung in der Kommentarspalte (wie genau sind natürliche Personen betroffen?) zwingend notwendig.</p> <p>ja = 3 nein = 1</p>	allgemein	Vermeidung von ungerechtfertigter Diskriminierung und Verzerrung			<p>Hier sind betreffene Personen natürliche Personen, die durch ein KI-System beeinflusst oder betroffen werden, ohne notwendigerweise direkt oder aktiv mit dem System zu interagieren. Der Unterschied zu Nutzern liegt in der Art der Interaktion: Während Nutzer das KI-System direkt und aktiv bedienen, beispielsweise als Endnutzer oder Benutzer, stehen betroffene Personen in einem eher indirekten und passiven Verhältnis zum KI-System, können jedoch durch dessen Entscheidungen oder Funktionen beeinflusst werden (vgl. Glossar).</p> <p>Geschützte Eigenschaften sind (in Anlehnung an das AGG): Rasse, ethnische Herkunft, Geschlecht, Religion oder Weltanschauung, Behinderung, Alter, sexuelle Identität.</p>
Grundfragen		Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
ND-Z14	<p>Benutzt das KI-System Daten, die Aufschluss über die Zugehörigkeit einer natürlichen Person zu einer Personengruppe mit geschützten Eigenschaften geben?</p> <p>ja = 3 nein = 1</p>	allgemein	Vermeidung von ungerechtfertigter Diskriminierung und Verzerrung			<p>Geschützte Eigenschaften sind (in Anlehnung an das AGG): Rasse, ethnische Herkunft, Geschlecht, Religion oder Weltanschauung, Behinderung, Alter, sexuelle Identität.</p>
ND-Z15	<p>Regelt das KI-System den Zugang zu essenziell die Persönlichkeit betreffenden Diensten oder Aktivitäten?</p> <p>ja = 3 nein = 1</p>	allgemein	Vermeidung von ungerechtfertigter Diskriminierung und Verzerrung			
ND-Z16	<p>Ist das KI-System in Entscheidungsprozesse eingebunden, welche Persönlichkeitsrechte weitreichend beeinflussen?</p> <p>ja = 3 nein = 1</p>	allgemein	Vermeidung von ungerechtfertigter Diskriminierung und Verzerrung			
Erweiterungsfragen		Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
ND-Z19	<p>Könnte ein Missbrauch (z.B. Überwachung von Personen) des KI-Systems Gruppen, die sich anhand geschützter Eigenschaften definieren, in besonders starkem Ausmaß schaden?</p> <p>ja = 3 nein = 1</p>	allgemein	Vermeidung von ungerechtfertigter Diskriminierung und Verzerrung			<p>Geschützte Eigenschaften sind (in Anlehnung an das AGG): Rasse, ethnische Herkunft, Geschlecht, Religion oder Weltanschauung, Behinderung, Alter, sexuelle Identität.</p>
ND-Z20	<p>Wie viele Nutzer und (durch die Nutzung) betroffene Personen hat das KI-System innerhalb eines bestimmten Zeitraums?</p> <p>niedrige Anzahl an Nutzer*innen/betroffenen Personen = 1 moderate Anzahl an Nutzer*innen/betroffenen Personen = 2 hohe Anzahl an Nutzer*innen/betroffenen Personen = 3</p>	allgemein	Vermeidung von ungerechtfertigter Diskriminierung und Verzerrung			<p>Beispielhafte Grenzen: niedrige Anzahl an Nutzern/betroffenen Personen im Monat = 0-1.000 moderate Anzahl an Nutzern/betroffenen Personen im Monat = 1.001-10.000 hohe Anzahl an Nutzern/betroffenen Personen im Monat = >10.000</p> <p>Hier sind betroffene Personen natürliche Personen, die durch ein KI-System beeinflusst oder betroffen werden, ohne notwendigerweise direkt oder aktiv mit dem System zu interagieren. Der Unterschied zu Nutzer*innen liegt in der Art der Interaktion: Während Nutzer*innen das KI-System direkt und aktiv bedienen, beispielsweise als Endnutzer*innen oder Benutzer*innen, stehen betroffene Personen in einem eher indirekten und passiven Verhältnis zum KI-System, können jedoch durch dessen Entscheidungen oder Funktionen beeinflusst werden (vgl. Glossar).</p> <p>Hier meint Nutzer natürliche Personen, die direkt und aktiv mit einem KI-System interagieren, entweder als Endnutzer*innen, die das KI-System für persönliche oder geschäftliche Zwecke verwenden, oder als Benutzer*innen, die das KI-System im professionellen Kontext anwenden. Der Begriff umfasst sowohl diejenigen, die das KI-System bedienen, als auch jene, die es für die Erreichung spezifischer Ziele nutzen. Je nach Kontext kann dies die Verwendung von KI-Systemen in alltäglichen Anwendungen oder in spezialisierten, beruflichen Szenarien beinhalten (vgl. Glossar).</p>

Frageindex	Frage	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Erläuterungen und Beispiele
ND-Z22	Hat das KI-System hinsichtlich seines Anwendungsbereiches das Potenzial, gesetzlich verbotene Schlechterbehandlung abzubilden, oder besteht die begründete Annahme, dass es derartige Ungleichbehandlungen verstärkt? ja = 3 nein = 1	Nichtdiskriminierung	Vermeidung von ungerechtfertigter Diskriminierung und Verzerrung			
ND-Z23	Hat das KI-System aufgrund seines Anwendungsbereichs oder seiner Funktionsweise das Potenzial, Ausgaben zu produzieren, die möglicherweise gesetzlich verboten sind (z.B. Beleidigungen), oder lässt sich mit hinreichender Plausibilität annehmen, dass die zugrundeliegende (Trainings-/Test-)Datenbasis derartige Instanzen oder Inhalte enthält? ja = 3 nein = 1	Menschenwürde	Vermeidung von ungerechtfertigter Diskriminierung und Verzerrung			

Schutzbedarfsanalyse

Qualitätsdimension: Transparenz (TR)

Gesamteinschätzung je Kriterium

Schutzbedarf Kriterium „Rückverfolgbarkeit & Dokumentation“:			1 / 2 / 3	immer anwendbar
Schutzbedarf Kriterium „Erklärbarkeit & Interpretierbarkeit“:			1 / 2 / 3 / nicht anwendbar	

Frageindex	Frage	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
	Applikationsfragen	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
TR-Z11	Reduziert sich die Sicherheit (safety) des Systems, wenn eine Einschränkung von Erklärbarkeit und Interpretierbarkeit des Systems gegeben ist? Hinweis: Hier geht es darum die "Funktionsweise" des Systems zu verstehen und nicht um Rückverfolgbarkeit und Dokumentation. ja = 3 nein = 1	allgemein	Erklärbarkeit & Interpretierbarkeit			Als geringes Maß können Auswirkungen verstanden werden, die Nutzer und/oder Betroffene vermutlich ohne Beschwerde hinnehmen würden.
TR-Z12	Wirkt sich das Verhalten oder Ergebnis des KI-Systems substanziell auf die Handlungsfreiheit von natürlichen Personen oder auf persönliche Rechte aus? ja = 3 nein = 1	allgemein	Erklärbarkeit & Interpretierbarkeit			Als substanziell können Auswirkungen verstanden werden, die die Handlungsfreiheit oder -autonomie einschränken.
	Grundfragen	Schutzzielkategorie	Kriterium	Einschätzung		Anmerkungen und Beispiele
TR-Z16	Kann eine infolge mangelnder Erklärbarkeit und Interpretierbarkeit verursachte Einschränkung der sicheren und zweckgemäßen Nutzung des KI-Systems – im realistischen worst-case – tödliche Personenschäden verursachen? ja = 3 nein = 1	Leib und Leben	Erklärbarkeit & Interpretierbarkeit			
TR-Z17	Kann eine infolge mangelnder Rückverfolgbarkeit und Dokumentation verursachte Einschränkung der sicheren und zweckgemäßen Nutzung des KI-Systems – im realistischen worst-case – tödliche Personenschäden verursachen? ja = 3 nein = 1	Leib und Leben	Rückverfolgbarkeit & Dokumentation			
TR-Z18	Kann eine infolge mangelnder Erklärbarkeit und Interpretierbarkeit verursachte Einschränkung der sicheren und zweckgemäßen Nutzung des KI-Systems – im realistischen worst-case – Gesundheitsschäden (körperliche Gesundheit) bei natürlichen Personen verursachen? ja = 2 nein = 1	Gesundheit	Erklärbarkeit & Interpretierbarkeit			
TR-Z19	Kann eine infolge mangelnder Rückverfolgbarkeit und Dokumentation verursachte Einschränkung der sicheren und zweckgemäßen Nutzung des KI-Systems – im realistischen worst-case – Gesundheitsschäden (körperliche Gesundheit) bei natürlichen Personen verursachen? ja = 2 nein = 1	Gesundheit	Rückverfolgbarkeit & Dokumentation			
TR-Z20	Regelt das KI-System den Zugang zu essenziell die Persönlichkeit oder die persönlichen Handlungsspielräume betreffenden Diensten und Aktivitäten? ja = 3 nein = 1	allgemein	Rückverfolgbarkeit & Dokumentation			Identisch zu Frage TR-Z21 (anderes Kriterium)
TR-Z21	Regelt das KI-System den Zugang zu essenziell die Persönlichkeit oder die persönlichen Handlungsspielräume betreffenden Diensten und Aktivitäten? ja = 3 nein = 1	allgemein	Erklärbarkeit & Interpretierbarkeit			Identisch zu Frage TR-Z20 (anderes Kriterium)
TR-Z22	Ist das KI-System in Entscheidungsprozesse eingebunden, welche Grundrechte oder demokratische Verfahren betreffen? ja = 3 nein = 1	allgemein	Rückverfolgbarkeit & Dokumentation			Identisch zu Frage TR-Z23 (anderes Kriterium) Beispiel: Überwachung von natürlichen Personen
TR-Z23	Ist das KI-System in Entscheidungsprozesse eingebunden, welche Grundrechte oder demokratische Verfahren betreffen? ja = 3 nein = 1	allgemein	Erklärbarkeit & Interpretierbarkeit			Identisch zu Frage TR-Z22 (anderes Kriterium) Beispiel: Überwachung von natürlichen Personen

Frageindex	Frage	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
TR-Z24	<p>Kann das KI-System – absichtlich oder unabsichtlich – bei fehlender Rückverfolgbarkeit und Dokumentation zu einem anderen Zwecke missbraucht werden, sodass dadurch gesellschaftliche Schäden oder Schäden bzgl. der Grundrechte von natürlichen Personen verursacht werden?</p> <p>ja = 3 nein = 1</p>	allgemein	Rückverfolgbarkeit & Dokumentation			Beispiel: Ein KI-System, welches den Informationsfluss auf Plattformen wie sozialen Netzwerken reguliert, könnte versehentlich die Verbreitung von Fehlinformationen verstärken, indem es Inhalte priorisiert, die polarisierend oder reißerisch sind. Dies könnte zu gesellschaftlicher Spaltung, dem Aufkommen von Extremismus und zur Untergrabung des Vertrauens in öffentliche Institutionen führen.
TR-Z25	<p>In welchem Ausmaß interagiert das KI-System mit Nutzern in einer Art und Weise, die – im realistischen worst-case – ihre Wahrnehmung, Entscheidungen oder Handlungen entscheidend negativ einschränken oder beeinflussen kann?</p> <p>überhaupt nicht oder in geringem Maße = 1 in moderatem Maße = 2 in hohem Maße = 3</p>	allgemein	Rückverfolgbarkeit & Dokumentation			<p>Identisch zu Frage TR-Z26 (anderes Kriterium)</p> <p>Hier meint Nutzer natürliche Personen, die direkt und aktiv mit einem KI-System interagieren, entweder als Endnutzer, die das KI-System für persönliche oder geschäftliche Zwecke verwenden, oder als Benutzer, die das KI-System im professionellen Kontext anwenden. Der Begriff umfasst sowohl diejenigen, die das KI-System bedienen, als auch jene, die es für die Erreichung spezifischer Ziele nutzen. Je nach Kontext kann dies die Verwendung von KI-Systemen in alltäglichen Anwendungen oder in spezialisierten, beruflichen Szenarien beinhalten (vgl. Glossar).</p>
TR-Z26	<p>In welchem Ausmaß interagiert das KI-System mit Nutzern in einer Art und Weise, die – im realistischen worst-case – ihre Wahrnehmung, Entscheidungen oder Handlungen entscheidend negativ einschränken oder beeinflussen kann?</p> <p>überhaupt nicht oder in geringem Maße = 1 in moderatem Maße = 2 in hohem Maße = 3</p>	allgemein	Erklärbarkeit & Interpretierbarkeit			<p>Identisch zu Frage TR-Z25 (anderes Kriterium)</p> <p>Hier meint Nutzer natürliche Personen, die direkt und aktiv mit einem KI-System interagieren, entweder als Endnutzer, die das KI-System für persönliche oder geschäftliche Zwecke verwenden, oder als Benutzer, die das KI-System im professionellen Kontext anwenden. Der Begriff umfasst sowohl diejenigen, die das KI-System bedienen, als auch jene, die es für die Erreichung spezifischer Ziele nutzen. Je nach Kontext kann dies die Verwendung von KI-Systemen in alltäglichen Anwendungen oder in spezialisierten, beruflichen Szenarien beinhalten (vgl. Glossar).</p>
TR-Z27	<p>Wie häufig können Situationen auftreten, in denen die Nutzungssicherheit oder der Nutzen des KI-Systems bei mangelnder Validierbarkeit oder Erklärbarkeit einzelner Ergebnisse eingeschränkt sind?</p> <p>nie = 1 hin und wieder = 2 oft = 3</p>	allgemein	Erklärbarkeit & Interpretierbarkeit			
Erweiterungsfragen		Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
TR-Z30	<p>Ist die Nutzung des KI-Systems für die Nutzer verpflichtend bzw. nicht-freiwillig?</p> <p>ja = 3 nein = 1</p>	allgemein	Erklärbarkeit & Interpretierbarkeit			Hier meint Nutzer natürliche Personen, die direkt und aktiv mit einem KI-System interagieren, entweder als Endnutzer, die das KI-System für persönliche oder geschäftliche Zwecke verwenden, oder als Benutzer, die das KI-System im professionellen Kontext anwenden. Der Begriff umfasst sowohl diejenigen, die das KI-System bedienen, als auch jene, die es für die Erreichung spezifischer Ziele nutzen. Je nach Kontext kann dies die Verwendung von KI-Systemen in alltäglichen Anwendungen oder in spezialisierten, beruflichen Szenarien beinhalten (vgl. Glossar).
TR-Z31	<p>Wie häufig wird das KI-System während des Betriebs weitertrainiert oder durch neue Versionen aktualisiert?</p> <p>nie = 1 hin und wieder = 2 oft = 3</p>	allgemein	Rückverfolgbarkeit & Dokumentation			<p>Identisch zu Frage VE-Z24 (andere Dimension)</p> <p>Beispielhafte Grenzen: nie = kein einziges Mal hin und wieder = monatlich oder seltener oft = häufiger als monatlich</p>
TR-Z32	<p>Wie gut können durch das KI-System verursachte, schadhafte Fehler als solche erkannt werden?</p> <p>immer leicht und eindeutig = 1 mit leichten bis moderaten Mühen = 2 mit großen oder größten Mühen = 3</p>	allgemein	Rückverfolgbarkeit & Dokumentation			
TR-Z33	<p>Kann eine durch mangelnde Rückverfolgbarkeit und Dokumentation verursachte Einschränkung der sicheren und zweckgemäßen Nutzung des KI-Systems – im realistischen worst-case – tödliche Schäden bei Tieren verursachen?</p> <p>ja = 3 nein = 1</p>	Leib und Leben	Rückverfolgbarkeit & Dokumentation			
TR-Z34	<p>Kann eine durch mangelnde Erklärbarkeit und Interpretierbarkeit verursachte Einschränkung der sicheren und zweckgemäßen Nutzung des KI-Systems – im realistischen worst-case – tödliche Schäden bei Tieren verursachen?</p> <p>ja = 3 nein = 1</p>	Leib und Leben	Erklärbarkeit & Interpretierbarkeit			

Frageindex	Frage	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
TR-Z35	<p>Falls Fachkenntnisse über künstliche Intelligenz bei den Nutzern notwendig sind: In welchem Maße sind Fachkenntnisse über künstliche Intelligenz bei Nutzern oder betroffenen Personen des KI-Systems erwartbar?</p> <p>Fachkenntnisse in hohem Maße erwartbar oder nicht notwendig = 1 moderate Fachkenntnisse erwartbar = 2 keine oder geringe Fachkenntnisse erwartbar = 3</p>	allgemein	Erklärbarkeit & Interpretierbarkeit			<p>Hier sind betroffene Personen natürliche Personen, die durch ein KI-System beeinflusst oder betroffen werden, ohne notwendigerweise direkt oder aktiv mit dem System zu interagieren. Der Unterschied zu Nutzern liegt in der Art der Interaktion: Während Nutzer das KI-System direkt und aktiv bedienen, beispielsweise als Endnutzer oder Benutzer, stehen betroffene Personen in einem eher indirekten und passiven Verhältnis zum KI-System, können jedoch durch dessen Entscheidungen oder Funktionen beeinflusst werden (vgl. Glossar).</p> <p>Hier meint Nutzer natürliche Personen, die direkt und aktiv mit einem KI-System interagieren, entweder als Endnutzer, die das KI-System für persönliche oder geschäftliche Zwecke verwenden, oder als Benutzer, die das KI-System im professionellen Kontext anwenden. Der Begriff umfasst sowohl diejenigen, die das KI-System bedienen, als auch jene, die es für die Erreichung spezifischer Ziele nutzen. Je nach Kontext kann dies die Verwendung von KI-Systemen in alltäglichen Anwendungen oder in spezialisierten, beruflichen Szenarien beinhalten (vgl. Glossar).</p>
TR-Z36	<p>Ist es aufgrund der Funktionsweise und Aufgabenstellung des KI-Systems prinzipiell möglich, dass das KI-System die Entscheidungen oder Handlungen von Nutzern oder betroffenen Personen auf eine Art und Weise beeinflusst oder lenkt, die von ihnen nicht hinreichend durchdrungen werden kann?</p> <p>ja = 3 nein = 1</p>	allgemein	Rückverfolgbarkeit & Dokumentation			<p>Identisch zu Frage MA-Z23 (andere Dimension)</p> <p>„Hinreichend durchdrungen“ bedeutet hier, dass Nutzer/ betroffene Personen nicht hinreichend nachvollziehen können, wie ihre Entscheidungen und Handlungen beeinflusst werden.</p> <p>Hier sind betroffene Personen natürliche Personen, die durch ein KI-System beeinflusst oder betroffen werden, ohne notwendigerweise direkt oder aktiv mit dem System zu interagieren. Der Unterschied zu Nutzern liegt in der Art der Interaktion: Während Nutzer das KI-System direkt und aktiv bedienen, beispielsweise als Endnutzer oder Benutzer, stehen betroffene Personen in einem eher indirekten und passiven Verhältnis zum KI-System, können jedoch durch dessen Entscheidungen oder Funktionen beeinflusst werden (vgl. Glossar).</p> <p>Hier meint Nutzer natürliche Personen, die direkt und aktiv mit einem KI-System interagieren, entweder als Endnutzer, die das KI-System für persönliche oder geschäftliche Zwecke verwenden, oder als Benutzer, die das KI-System im professionellen Kontext anwenden. Der Begriff umfasst sowohl diejenigen, die das KI-System bedienen, als auch jene, die es für die Erreichung spezifischer Ziele nutzen. Je nach Kontext kann dies die Verwendung von KI-Systemen in alltäglichen Anwendungen oder in spezialisierten, beruflichen Szenarien beinhalten (vgl. Glossar).</p>
TR-Z37	<p>Wie hoch ist – im realistischen worst-case – der finanzielle Schaden, der entstehen kann, wenn Erklärungen oder Einsichten in die Genese von Ausgaben des KI-Systems fehlen? Der potentiell finanzielle Schaden dient hier als stellvertretende Bemessung des Schadens für den oder die Betroffenen.</p> <p>kein bis niedriger finanzieller Schaden = 1 moderater finanzieller Schaden = 2 schwerwiegender oder katastrophaler finanzieller Schaden = 3</p>	Eigentum und Sachen	Erklärbarkeit & Interpretierbarkeit			<p>Beispielhafte Grenzen: kein bis niedriger finanzieller Schaden = Schaden in Höhe von bis zu 1 % des Vorjahresumsatzes oder unter 1.000€ moderater finanzieller Schaden = Schaden in Höhe von 1 % bis 5 % des Vorjahresumsatzes oder zwischen 1.000€ bis 50.000€ schwerwiegender oder katastrophaler finanzieller Schaden = Schaden in Höhe von über 5 % des Vorjahresumsatzes oder über 50.000€</p>
TR-Z38	<p>Wie hoch ist – im realistischen worst-case – der finanzielle Schaden, der entstehen kann, falls Begründungen oder Rechenschaft über Fehler oder Fehlfunktion des KI-Systems nachträglich nicht vorgelegt werden können? Der potentiell finanzielle Schaden dient hier als stellvertretende Bemessung des Schadens für den oder die Betroffenen.</p> <p>kein bis niedriger finanzieller Schaden = 1 moderater finanzieller Schaden = 2 schwerwiegender oder katastrophaler finanzieller Schaden = 3</p>	Eigentum und Sachen	Rückverfolgbarkeit & Dokumentation			<p>Beispielhafte Grenzen: kein bis niedriger finanzieller Schaden = Schaden in Höhe von bis zu 1 % des Vorjahresumsatzes oder unter 1.000€ moderater finanzieller Schaden = Schaden in Höhe von 1 % bis 5 % des Vorjahresumsatzes oder zwischen 1.000€ bis 50.000€ schwerwiegender oder katastrophaler finanzieller Schaden = Schaden in Höhe von über 5 % des Vorjahresumsatzes oder über 50.000€</p>

Schutzbedarfsanalyse

Qualitätsdimension: Menschliche Aufsicht und Kontrolle (MA)

Gesamteinschätzung je Kriterium

Schutzbedarf Kriterium „Menschliche Handlungsfähigkeit“:			1 / 2 / 3 / nicht anwendbar
Schutzbedarf Kriterium „Menschliche Aufsicht“:			1 / 2 / 3 / nicht anwendbar

Frageindex	Frage	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
Applikationsfragen		Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
MA-Z11	Kann das KI-System prinzipiell ohne menschliche Be(s)tätigung seine Aufgabe vollständig ausführen? ja = 3 nein = 1	AF	Menschliche Aufsicht			Gemeint ist hier mit „ prinzipiell “ die Art und Weise des Einsatzes innerhalb des Verwendungszwecks. Mit „ Aufgabe “ ist die Aufgabe des gesamten KI-Systems gemeint.
MA-Z12	Kann das KI-System eine unmittelbare Auswirkung auf Personen haben? ja = 3 nein = 1	AF	Menschliche Handlungsfähigkeit			Beispiele für Personenbezug: personenbezogene Daten beim Training des KI-Modells bzw. der KI-Modelle, Verarbeitung oder Erzeugung personenbezogener Daten, menschliche Rezipienten des Outputs
Grundfragen		Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
MA-Z15	Wie hoch ist der Schutzbedarf des Kriteriums „Leistungsfähigkeit und Robustheit“ (VE)? gering = 1 moderat = 2 hoch = 3	allgemein	Menschliche Aufsicht			
MA-Z16	Wie hoch ist der Schutzbedarf des Kriteriums „Rückfallpläne und funktionale Sicherheit“ (VE)? gering = 1 moderat = 2 hoch = 3	allgemein	Menschliche Aufsicht			
MA-Z17	Ist die Nutzung des KI-Systems für die Nutzer verpflichtend bzw. nicht-freiwillig? ja = 3 nein = 1	allgemein	Menschliche Handlungsfähigkeit			Hier meint Nutzer natürliche Personen, die direkt und aktiv mit einem KI-System interagieren, entweder als Endnutzer, die das KI-System für persönliche oder geschäftliche Zwecke verwenden, oder als Benutzer, die das KI-System im professionellen Kontext anwenden. Der Begriff umfasst sowohl diejenigen, die das KI-System bedienen, als auch jene, die es für die Erreichung spezifischer Ziele nutzen. Je nach Kontext kann dies die Verwendung von KI-Systemen in alltäglichen Anwendungen oder in spezialisierten, beruflichen Szenarien beinhalten (vgl. Glossar).
MA-Z18	Wie dringlich ist eine schnelle Behebung oder Wiederherstellung bei einem Ausfall oder fehlerhafter Funktion des KI-Systems? keine oder geringe Dringlichkeit = 1 moderate Dringlichkeit = 2 hohe Dringlichkeit= 3	allgemein	Menschliche Aufsicht			Eine schnelle Behebung oder Wiederherstellung ist zum Beispiel dann dringlich, wenn der Schaden durch andauernden Ausfall oder fehlerhafter Funktion kontinuierlich steigt
Erweiterungsfragen		Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
MA-Z21	Wie lassen sich die Ausweichmöglichkeiten für die Nutzer oder die Aufgabe des KI-Systems, falls die gewünschte Funktionalität des KI-Systems nicht zur Verfügung steht, beschreiben? angemessen und schnell verfügbar = 1 mit leichten bis moderate Einschränkungen verfügbar = 2 nur mit großen und größten Einschränkungen verfügbar = 3	allgemein	Menschliche Handlungsfähigkeit			Hier meint Nutzer natürliche Personen, die direkt und aktiv mit einem KI-System interagieren, entweder als Endnutzer, die das KI-System für persönliche oder geschäftliche Zwecke verwenden, oder als Benutzer, die das KI-System im professionellen Kontext anwenden. Der Begriff umfasst sowohl diejenigen, die das KI-System bedienen, als auch jene, die es für die Erreichung spezifischer Ziele nutzen. Je nach Kontext kann dies die Verwendung von KI-Systemen in alltäglichen Anwendungen oder in spezialisierten, beruflichen Szenarien beinhalten (vgl. Glossar).
MA-Z22	In welchem Ausmaß interagiert das KI-System mit Nutzern in einer Art und Weise, die – im realistischen worst-case – die Wahrnehmung, Entscheidungen oder Handlungen der Nutzer entscheidend negativ einschränken oder beeinflussen kann? überhaupt nicht oder in geringem Maße = 1 in moderatem Maße = 2 in hohem Maße = 3	allgemein	Menschliche Handlungsfähigkeit			Hier meint Nutzer natürliche Personen, die direkt und aktiv mit einem KI-System interagieren, entweder als Endnutzer, die das KI-System für persönliche oder geschäftliche Zwecke verwenden, oder als Benutzer, die das KI-System im professionellen Kontext anwenden. Der Begriff umfasst sowohl diejenigen, die das KI-System bedienen, als auch jene, die es für die Erreichung spezifischer Ziele nutzen. Je nach Kontext kann dies die Verwendung von KI-Systemen in alltäglichen Anwendungen oder in spezialisierten, beruflichen Szenarien beinhalten (vgl. Glossar).

Frageindex	Frage	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
MA-Z23	<p>Ist es aufgrund der Funktionsweise und Aufgabenstellung des KI-Systems prinzipiell möglich, dass das KI-System die Entscheidungen oder Handlungen von Nutzern oder betroffenen Personen auf eine Art und Weise beeinflusst oder lenkt, die von ihnen nicht hinreichend durchdrungen werden kann?</p> <p>ja = 3 nein = 1</p>	allgemein	Menschliche Handlungsfähigkeit			<p>Identisch zu Frage TR-Z36 (andere Dimension)</p> <p>Hier meint Nutzer natürliche Personen, die direkt und aktiv mit einem KI-System interagieren, entweder als Endnutzer, die das KI-System für persönliche oder geschäftliche Zwecke verwenden, oder als Benutzer, die das KI-System im professionellen Kontext anwenden. Der Begriff umfasst sowohl diejenigen, die das KI-System bedienen, als auch jene, die es für die Erreichung spezifischer Ziele nutzen. Je nach Kontext kann dies die Verwendung von KI-Systemen in alltäglichen Anwendungen oder in spezialisierten, beruflichen Szenarien beinhalten (vgl. Glossar).</p>

Schutzbedarfsanalyse

Qualitätsdimension: KI-spezifische Cybersicherheit (CY)

Gesamteinschätzung je Kriterium

Schutzbedarf Kriterium „Allgemeine KI-spezifische Cybersicherheit“:			1 / 2 / 3	immer anwendbar
Schutzbedarf Kriterium „Widerstandsfähigkeit gegen KI-spezifische Angriffe“:			1 / 2 / 3 / nicht anwendbar	

Frageindex	Frage	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
	Applikationsfragen	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
CY-Z10	Ist das KI-System dazu bestimmt, an ein Netzwerk angeschlossen zu werden und darüber Daten oder Anweisungen zu empfangen? ja = 3 nein = 1	allgemein	Widerstandsfähigkeit gegen KI-spezifische Angriffe			„ Anschluss “ meint hier drahtlose, kabellose Verbindungen. Es sind sowohl interne, als auch externe Netzwerke gemeint.
	Grundfragen	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
CY-Z14	Wie hoch ist die Kritikalität der Daten, die durch das KI-System verarbeitet werden? gering = 1 moderat = 2 hoch = 3	allgemein	„Allgemeine KI-spezifische Cybersicherheit“ und „Widerstandsfähigkeit gegen KI-spezifische Angriffe“			Beispiel im Kontext von Firmendaten (Information Classification Model): Internal Only = 1 Confidential = 2 Restricted = 3 Beispiel im Kontext von personenbezogenen Daten: Allgemeine personenbezogene Daten (z.B. Name, Anschrift, ...) = 1 Personenbezogene Daten mit Verhaltensbezug o.ä. (z.B. Kaufen, Browsen, Bewegung, ...) = 2 Personenbezogene Daten nach Art. 9 DSGVO = 3
CY-Z15	Inwiefern sind die Schäden, die durch den absichtlichen Missbrauch oder die Manipulation des KI-Systems – im realistischen worst-case – verursacht werden können, behebbbar bzw. wiederherstellbar? vollständig behebbbar = 1 teilweise behebbbar = 2 wenig bis nicht behebbbar = 3	allgemein	„Allgemeine KI-spezifische Cybersicherheit“ und „Widerstandsfähigkeit gegen KI-spezifische Angriffe“			„ Missbrauch “ ist hier im weiten Sinne zu verstehen
CY-Z16	Inwiefern sind die Schäden, die – im realistischen worst-case – durch Datenzugriffe von nicht-authorisierten Personen oder durch Datennutzung außerhalb der vertraglichen Bedingungen verursacht werden können, behebbbar bzw. inwiefern sind Zustände vor Zugriff oder Nutzung der Daten wiederherstellbar? vollständig behebbbar = 1 teilweise behebbbar = 2 nicht behebbbar = 3	allgemein	„Allgemeine KI-spezifische Cybersicherheit“ und „Widerstandsfähigkeit gegen KI-spezifische Angriffe“			Hier geht es nur um die direkt Konsequenzen durch unauthorisierte Datenzugriffe und Datennutzungen. Die Konsequenzend durch Missbrauch und Manipulation werden an anderer Stelle abgefragt. Ausgelesene Passwörter können geändert werden, insofern ist der unmittelbare Schaden behebbbar. Die Einsicht in sensitive Informationen kann nicht rückgängig gemacht werden, der unmittelbare Schaden ist nicht behebbar.
CY-Z17	Können absichtlicher Missbrauch oder Manipulation des KI-Systems – im realistischen worst-case – tödliche Personenschäden verursachen? ja = 3 nein = 1	Leib und Leben	„Allgemeine KI-spezifische Cybersicherheit“ und „Widerstandsfähigkeit gegen KI-spezifische Angriffe“			
CY-Z18	Können absichtlicher Missbrauch oder Manipulation des KI-Systems – im realistischen worst-case – Gesundheitsschäden (körperliche Gesundheit) bei natürlichen Personen verursachen? ja = 2 nein = 1	Gesundheit	„Allgemeine KI-spezifische Cybersicherheit“ und „Widerstandsfähigkeit gegen KI-spezifische Angriffe“			
CY-Z19	Können absichtlicher Missbrauch oder Manipulation des KI-Systems gesellschaftlichen Schaden anrichten oder Grundrechtsverletzungen verursachen? ja = 3 nein = 1	allgemein	„Allgemeine KI-spezifische Cybersicherheit“ und „Widerstandsfähigkeit gegen KI-spezifische Angriffe“			
CY-Z20	Können absichtlicher Missbrauch oder Manipulation der KI-Systems besonders starke Auswirkungen auf Gruppen haben, die sich anhand geschützter Eigenschaften definieren? ja = 3 nein = 1	allgemein	„Allgemeine KI-spezifische Cybersicherheit“ und „Widerstandsfähigkeit gegen KI-spezifische Angriffe“			Geschützte Eigenschaften sind (in Anlehnung an das AGG): Rasse, ethnische Herkunft, Geschlecht, Religion oder Weltanschauung, Behinderung, Alter, sexuelle Identität.
	Erweiterungsfragen	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele

Frageindex	Frage	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
CY-Z23	Wie viele schadhafte Ergebnisse oder Ausgaben können – im realistischen worst-case – bei absichtlichem Missbrauch oder bei Manipulation des KI-Systems potenziell durch das KI-System erzeugt werden, bis der sichere Regelbetrieb wiederhergestellt ist? niedrige Anzahl = 1 moderate Anzahl = 2 hohe Anzahl = 3	allgemein	„Allgemeine KI-spezifische Cyber-sicherheit" und „Widerstandsfähigkeit gegen KI-spezifische Angriffe"			Beispielhafte Grenzen: niedrige Anzahl = 0-1.000 moderate Anzahl = 1.001-10.000 hohe Anzahl = >10.000
CY-Z24	Wie gut können schadhafte, durch absichtlichen Missbrauch oder Manipulation verursachte Fehler des KI-Systems als solche erkannt werden? immer leicht und eindeutig = 1 mit leichten bis moderaten Mühen = 2 mit großen oder größten Mühen = 3	allgemein	„Allgemeine KI-spezifische Cyber-sicherheit" und „Widerstandsfähigkeit gegen KI-spezifische Angriffe"			
CY-Z25	Können absichtlicher Missbrauch oder Manipulation des KI-Systems – im realistischen worst-case – tödliche Schäden bei Tieren verursachen? ja = 3 nein = 1	Leib und Leben	„Allgemeine KI-spezifische Cyber-sicherheit" und „Widerstandsfähigkeit gegen KI-spezifische Angriffe"			
CY-Z26	Wie hoch ist – im realistischen worst-case – der finanzielle Schaden, den die Verletzung der Vertraulichkeit oder Integrität der vom KI-System verarbeiteten sensiblen Geschäfts- oder lizenzgebundenen Daten oder des geistigen Eigentums zur Folge haben kann? Der potentiell finanzielle Schaden dient hier als stellvertretende Bemessung des Schadens für den oder die Betroffenen. kein bis niedriger finanzieller Schaden = 1 moderater finanzieller Schaden = 2 schwerwiegender oder katastrophaler finanzieller Schaden = 3	Eigentum und Sachen	„Allgemeine KI-spezifische Cyber-sicherheit" und „Widerstandsfähigkeit gegen KI-spezifische Angriffe"			Beispielhafte Grenzen: kein bis niedriger finanzieller Schaden = Schaden in Höhe von bis zu 1 % des Vorjahresumsatzes oder unter 1.000€ moderater finanzieller Schaden = Schaden in Höhe von 1 % bis 5 % des Vorjahresumsatzes oder zwischen 1.000€ bis 50.000€ schwerwiegender oder katastrophaler finanzieller Schaden = Schaden in Höhe von über 5 % des Vorjahresumsatzes oder über 50.000€
CY-Z27	Wie lassen sich die Auswirkungen von absichtlichem Missbrauch oder Manipulation des KI-Systems – im realistischen worst-case – in Bezug auf die Umwelt (Sustainability) beschreiben? höchstens geringe negative Auswirkungen = 1 moderate negative Auswirkungen = 2 hohe oder katastrophale Auswirkungen = 3	Umwelt	„Allgemeine KI-spezifische Cyber-sicherheit" und „Widerstandsfähigkeit gegen KI-spezifische Angriffe"			Auswirkungen auf die Umwelt können sich auf verschwendete materielle Ressourcen oder Strom beziehen. Beispielhafte Grenzen: höchstens geringe negative Auswirkungen = es werden einzelne unbrauchbare Produkte produziert moderate negative Auswirkungen = es werden einige unbrauchbare Produkte produziert hohe oder katastrophale Auswirkungen = es werden in große Mengen unbrauchbare Produkte produziert

Schutzbedarfsanalyse

Qualitätsdimension: Verlässlichkeit (VE)

Gesamteinschätzung je Kriterium

Schutzbedarf Kriterium „Leistungsfähigkeit und Robustheit“:			1 / 2 / 3	immer anwendbar
Schutzbedarf Kriterium „Rückfallpläne und funktionale Sicherheit “:			1 / 2 / 3	immer anwendbar

Frageindex	Frage	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
	Grundfragen	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
VE-Z11	Inwiefern sind die Schäden, die durch das KI-System – im realistischen worst-case – verursacht werden können, behebbar bzw. wiederherstellbar? vollständig behebbar = 1 teilweise behebbar = 2 nicht behebbar = 3	allgemein	Rückfallpläne und funktionale Sicherheit			
VE-Z12	Kann das KI-System – im realistischen worst-case – unmittelbar tödliche Personenschäden verursachen? ja = 3 nein = 1	Leib und Leben	Leistungsfähigkeit und Robustheit			Identisch zu Frage VE-Z13 (anderes Kriterium)
VE-Z13	Kann das KI-System – im realistischen worst-case – unmittelbar tödliche Personenschäden verursachen? ja = 3 nein = 1	Leib und Leben	Rückfallpläne und funktionale Sicherheit			Identisch zu Frage VE-Z12 (anderes Kriterium)
VE-Z14	Kann das KI-System – im realistischen worst-case – Gesundheitsschäden (körperliche Gesundheit) bei Personen verursachen? ja = 2 nein = 1	Gesundheit	Leistungsfähigkeit und Robustheit			Identisch zu Frage VE-Z15 (anderes Kriterium)
VE-Z15	Kann das KI-System – im realistischen worst-case – Gesundheitsschäden (körperliche Gesundheit) bei natürlichen Personen verursachen? ja = 2 nein = 1	Gesundheit	Rückfallpläne und funktionale Sicherheit			Identisch zu Frage VE-Z14 (anderes Kriterium)
VE-Z16	Regelt das KI-System den Zugang zu essenziell die Persönlichkeit oder die persönlichen Handlungsspielräume betreffenden Diensten und Aktivitäten? ja = 3 nein = 1	allgemein	Leistungsfähigkeit und Robustheit			
VE-Z17	Ist das KI-System in Entscheidungsprozesse eingebunden, welche Grundrechte oder demokratische Verfahren betreffen? ja = 3 nein = 1	allgemein	Leistungsfähigkeit und Robustheit			
	Erweiterungsfragen	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
VE-Z20	Wie gut können durch das KI-System verursachte, schadhafte Ergebnisse oder Ausgaben als solche erkannt werden? immer leicht und eindeutig = 1 mit leichten bis moderaten Mühen = 2 mit großen oder größten Mühen; gar nicht = 3	allgemein	Leistungsfähigkeit und Robustheit			
VE-Z21	Wie viele schadhafte Ergebnisse oder Ausgaben können – im realistischen worst-case – bei fehlerhaftem Zustand oder Verhalten potenziell durch das KI-System erzeugt werden, bis der sichere Regelbetrieb wiederhergestellt ist? niedrige Anzahl = 1 moderate Anzahl = 2 hohe Anzahl = 3	allgemein	Rückfallpläne und funktionale Sicherheit			Beispielhafte Grenzen: niedrige Anzahl = 0–1.000 moderate Anzahl = 1.001–10.000 hohe Anzahl = > 10.000
VE-Z22	Ist uneindeutig, wie die Validität der Ergebnisse des KI-Systems gemessen werden kann? ja = 3 nein = 1	allgemein	Leistungsfähigkeit und Robustheit			Hier meint Nutzer natürliche Personen, die direkt und aktiv mit einem KI-System interagieren, entweder als Endnutzer, die das KI-System für persönliche oder geschäftliche Zwecke verwenden, oder als Benutzer, die das KI-System im professionellen Kontext anwenden. Der Begriff umfasst sowohl diejenigen, die das KI-System bedienen, als auch jene, die es für die Erreichung spezifischer Ziele nutzen. Je nach Kontext kann dies die Verwendung von KI-Systemen in alltäglichen Anwendungen oder in spezialisierten, beruflichen Szenarien beinhalten (vgl. Glossar).

Frageindex	Frage	Schutzzielkategorie	Kriterium	Einschätzung	Optional: Kommentar	Anmerkungen und Beispiele
VE-Z23	Wie lassen sich die Ausweichmöglichkeiten für die Nutzer oder die Aufgabe des KI-Systems beschreiben, für den Fall dass die gewünschte Funktionalität des KI-Systems nicht zur Verfügung steht? als angemessen und schnell verfügbar = 1 als mit leichten bis moderaten Einschränkungen verfügbar = 2 als nur mit großen und größten Einschränkungen verfügbar = 3	allgemein	Rückfallpläne und funktionale Sicherheit			
VE-Z24	Wie häufig wird das KI-System während des Betriebs weitertrainiert oder durch neue Versionen aktualisiert? nie = 1 hin und wieder = 2 oft = 3	allgemein	Leistungsfähigkeit und Robustheit			Identisch zu Frage TR-Z31 (andere Dimension) Beispielhafte Grenzen: nie = kein einziges Mal hin und wieder = monatlich oder seltener oft = häufiger als monatlich
VE-Z25	Kann das KI-System – im realistischen worst-case – tödliche Schäden bei Tieren verursachen? ja = 3 nein = 1	Leib und Leben	Leistungsfähigkeit und Robustheit			Identisch zu Frage VE-Z26 (anderes Kriterium)
VE-Z26	Kann das KI-System – im realistischen worst-case – tödliche Schäden bei Tieren verursachen? ja = 3 nein = 1	Leib und Leben	Rückfallpläne und funktionale Sicherheit			Identisch zu Frage VE-Z25 (anderes Kriterium)
VE-Z27	Wie hoch ist der finanzielle Schaden, der – für alle involvierten Interessengruppen, im realistischen worst-case – durch das KI-System entstehen kann? kein bis niedriger finanzieller Schaden = 1 moderater finanzieller Schaden = 2 schwerwiegender oder katastrophaler finanzieller Schaden = 3	Eigentum und Sachen	Leistungsfähigkeit und Robustheit			Anmerkung: Hier geht es um den unmittelbar messbaren finanziellen Schaden . Beispielhafte Grenzen: kein bis niedriger finanzieller Schaden = Schaden in Höhe von unter 1.000€ moderater finanzieller Schaden = Schaden in Höhe von zwischen 1.000€ bis 50.000€ schwerwiegender oder katastrophaler finanzieller Schaden = Schaden in Höhe von über 50.000€
VE-Z28	Wie hoch ist der finanzielle Schaden, der – für alle involvierten Interessengruppen, im realistischen worst-case – durch das KI-System entstehen kann? kein bis niedriger finanzieller Schaden = 1 moderater finanzieller Schaden = 2 schwerwiegender oder katastrophaler finanzieller Schaden = 3	Eigentum und Sachen	Rückfallpläne und funktionale Sicherheit			Anmerkung: Hier geht es um den unmittelbar messbaren finanziellen Schaden . Beispielhafte Grenzen: kein bis niedriger finanzieller Schaden = Schaden in Höhe von unter 1.000€ moderater finanzieller Schaden = Schaden in Höhe von zwischen 1.000€ bis 50.000€ schwerwiegender oder katastrophaler finanzieller Schaden = Schaden in Höhe von über 50.000€
VE-Z29	Wie lassen sich die Auswirkungen des KI-Systems – im realistischen worst-case – in Bezug auf die Umwelt (Sustainability) beschreiben? höchstens geringe negative Auswirkungen = 1 moderate negative Auswirkungen = 2 hohe oder katastrophale Auswirkungen = 3	Umwelt	Leistungsfähigkeit und Robustheit			Identisch zu Frage VE-Z30 (anderes Kriterium) Auswirkungen auf die Umwelt können sich auf verschwendete materielle Ressourcen oder Strom beziehen. Beispielhafte Grenzen: höchstens geringe negative Auswirkungen = es werden einzelne unbrauchbare Produkte produziert moderate negative Auswirkungen = es werden einige unbrauchbare Produkte produziert hohe oder katastrophale Auswirkungen = es werden in große Mengen unbrauchbare Produkte produziert
VE-Z30	Wie lassen sich die Auswirkungen des KI-Systems – im realistischen worst-case – in Bezug auf die Umwelt (Sustainability) beschreiben? höchstens geringe negative Auswirkungen = 1 moderate negative Auswirkungen = 2 hohe oder katastrophale Auswirkungen = 3	Umwelt	Rückfallpläne und funktionale Sicherheit			Identisch zu Frage VE-Z29 (anderes Kriterium)

Navigation												
Qualitätsdimensionen	Datenqualität, -schutz und -Governance			Nicht-Diskriminierung	Transparenz		Menschliche Aufsicht und Kontrolle		Verlässlichkeit		KI-spezifische Cybersicherheit	
Kriterien	DA1	DA2	DA3	ND1	TR1	TR2	MA1	MA2	VE1	VE2	CY1	CY2
Indikatoren	DA1.1	DA2.1	DA3.1	ND1.1	TR1.1	TR2.1	MA1.1	MA2.1	VE1.1	VE2.1	CY1.1	CY2.1
	DA1.2	DA2.2	DA3.2	ND1.2	TR1.2	TR2.2	MA1.2	MA2.2	VE1.2	VE2.2	CY1.2	CY2.2
	DA1.3	DA2.3	DA3.3	ND1.3	TR1.3	TR2.3	MA1.3	MA2.3	VE1.3	VE2.3	CY1.3	CY2.3
	DA1.4	DA2.4	DA3.4	ND1.4	TR1.4	TR2.4	MA1.4	MA2.4	VE1.4	VE2.4	CY1.4	CY2.4
	DA1.5	DA2.5	DA3.5	ND1.5	TR1.5		MA1.5	MA2.5	VE1.5	VE2.5	CY1.5	
		DA2.6	DA3.6	ND1.6	TR1.6		MA1.6		VE1.6	VE2.6	CY1.6	
		DA2.7		ND1.7	TR1.7		MA1.7		VE1.7	VE2.7		
				ND1.8	TR1.8							
				ND1.9								

Datenqualität, -schutz und -Governance										
DA1										
Datenqualität										
DA1.1	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Die Merkmale der Datensätze müssen dokumentiert werden.	Analyse – Definition Zusammensetzung der Daten – Anzahl der Dateninstanzen / Größe der Daten – Standards für Datenstrukturen/-Formate Prozess der Datenerhebung – Methode der Datenerhebung – Quelle der Daten – Verantwortliche für Datenerhebung Datenverarbeitungsschritte – Erklärung der Features und ihrer möglichen Qualitätsdimensionen und Bereiche der Qualitätsdimensionen – Vorverarbeitung – Labeling – Cleaning Pflege der Daten – Speicherperioden – Aktualisierung der Daten	Analyse – Definition Zusammensetzung der Daten – Anzahl der Dateninstanzen / Größe der Daten – Standards für Datenstrukturen/-Formate Prozess der Datenerhebung – Methode der Datenerhebung – Quelle der Daten Datenverarbeitungsschritte – Erklärung der Features und ihrer möglichen Qualitätsdimensionen und Bereiche der Qualitätsdimensionen – Labeling – Cleaning Pflege des Datensatzes – Speicherperioden – Aktualisierung der Daten	Analyse – Definition Zusammensetzung der Daten – Anzahl der Dateninstanzen / Größe der Daten – Standards für Datenstrukturen/-Formate Prozess der Datenerhebung – Quelle der Daten Datenverarbeitungsschritte – Erklärung der Features und ihrer möglichen Qualitätsdimensionen und Bereiche der Qualitätsdimension Pflege des Datensatzes – Speicherperioden – Aktualisierung der Daten	Die Merkmale der Daten wurden nicht dokumentiert.		Komponente	Maßnahme	Maximalwert	TR1.3	
DA1.2	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Die Risiken, die sich aus fehlender Datenqualität im Kontext des Zweck des KI Systems ergeben müssen analysiert und daraus Datenqualitätsanforderungen abgeleitet werden.	Analyse – Zweck Definition Basierend auf dem Verwendungszweck und Anwendungsbereiches des KI Systems und den möglichen Risiken müssen relevante Datenqualitätscharakteristiken definiert und spezifische Anforderungen an diese gestellt werden Analyse-Risiko Detaillierte Risikoanalyse einschließlich Quantifizierung von Eintrittswahrscheinlichkeiten und Risiken: – Identifikation der möglichen Risiken und ihrer Ursachen – Zuweisung der Risikoverantwortung – Schätzung der Eintrittswahrscheinlichkeit – Schätzung der Aufdeckungswahrscheinlichkeit – Schätzung der Auswirkung – Strukturierte Abstufung und Priorisierung der Risiken Mindestens die folgenden Risikoquellen sind zu berücksichtigen: – Datenqualitätscharakteristiken passen nicht zum Verwendungszweck und Anwendungsbereich – Datenqualitätscharakteristiken sind schlecht erfüllt.	Analyse – Zweck Definition Basierend auf dem Verwendungszweck und Anwendungsbereiches des KI Systems und den möglichen Risiken müssen relevante Datenqualitätscharakteristiken definiert und spezifische Anforderungen an diese gestellt werden Analyse-Risiko Limitierte Risikoanalyse mit Fokus auf den Schutzbedarf ohne Quantifizierung von Wahrscheinlichkeiten: – Identifikation der möglichen Risiken – Zuweisung der Risikoverantwortung – Schätzung der Auswirkung – Strukturierte Abstufung und Priorisierung der Risiken Mindestens die folgenden Risikoquellen sind zu berücksichtigen: – Datenqualitätscharakteristiken passen nicht zum Verwendungszweck und Anwendungsbereich – Datenqualitätscharakteristiken sind schlecht erfüllt.	Analyse – Zweck Definition Basierend auf dem Verwendungszweck und Anwendungsbereiches des KI Systems und den möglichen Risiken müssen relevante Datenqualitätscharakteristiken definiert und spezifische Anforderungen an diese gestellt werden Analyse-Risiko Hauptsächlich qualitative Abschätzung ohne Wahrscheinlichkeiten: – Identifikation der möglichen Gefährdungen – Zuweisung der Risikoverantwortung – Qualitative Schätzung der Auswirkung – Qualitative Abstufung und Priorisierung der Gefährdungen Mindestens die folgenden Risikoquellen sind zu berücksichtigen: – Datenqualitätscharakteristiken passen nicht zum Verwendungszweck und Anwendungsbereich – Datenqualitätscharakteristiken sind schlecht erfüllt.	Die Risiken und Gefährdungen wurden nicht analysiert und es wurden keine Datenqualitätsanforderungen abgeleitet.	ISO/IEC 5259	System/Komponente	Analyse	Maximalwert	TR1.1	
DA1.3	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Die Datensätze müssen dem beabsichtigten Zweck des KI-Systems entsprechen und die daraus abgeleiteten Datenqualitätsanforderungen erfüllen.	Tech. Maßnahmen – Daten Die in DA 1.2 definierten Datenqualitätscharakteristiken müssen unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks getestet und bewertet werden. Dies beinhaltet: – Tests und Metriken zum Nachweis der Datenqualitätscharakteristiken – Mindestens müssen Vollständigkeit, Aktualität und Korrektheit betrachtet werden Die ausgewählten Metriken und Tests sollten möglichst die folgenden Merkmale aufweisen: – wenn durch die Auswahl begründet, sollten die Metriken und Tests sowohl wichtige Basismethoden als auch möglichst fortgeschrittene Methoden (hinsichtlich Komplexität, Informationsgehalt, Implementierungsaufwand, z.B. umfangreiches Prüfwerkzeug) umfassen – die ausgewählten Metriken und Tests sollten wenn möglich durch einen hohen Grad der Automatisierung eine kontinuierliche Evaluation der KI-Modelle erlauben – Festlegung von Schwellenwerten (als Mindestanforderungen an die Funktionalität/Leistung) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks – falls anwendbar: Abstufung der Metriken und Schwellenwerte nach unterschiedlichen Einsatzszenarien mit Bezug zur Definition des Anwendungsbereichs	Tech. Maßnahmen – Daten Die in DA 1.2 definierten Datenqualitätscharakteristiken müssen unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks getestet und bewertet werden. Dies beinhaltet: – Tests und Metriken zum Nachweis der Datenqualitätscharakteristiken – Mindestens müssen Vollständigkeit und Korrektheit betrachtet werden Die ausgewählten Metriken und Tests sollten möglichst die folgenden Merkmale aufweisen: – wenn durch die Auswahl begründet, sollten die Metriken und Tests sowohl wichtige Basismethoden als auch möglichst fortgeschrittene Methoden (hinsichtlich Komplexität, Informationsgehalt, Implementierungsaufwand, z.B. umfangreiches Prüfwerkzeug) umfassen – Festlegung von Schwellenwerten (als Mindestanforderungen an die Funktionalität/Leistung) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks	Tech. Maßnahmen – Daten Die in DA 1.2 definierten Datenqualitätscharakteristiken müssen unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks getestet und bewertet werden. Dies beinhaltet: – Tests und Metriken zum Nachweis der Datenqualitätscharakteristiken Die ausgewählten Metriken und Tests sollten möglichst die folgenden Merkmale aufweisen: – es genügen Basismethoden (hinsichtlich Komplexität, Informationsgehalt, Implementierungsaufwand, z.B. eine simple Metrik) – Festlegung von Schwellenwerten (als Mindestanforderungen an die Funktionalität/Leistung) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks	Es wurde kein Nachweis geführt, dass der Datensatz dem Zweck des KI-Systems und den Datenqualitätsanforderungen entspricht.	Während des Trainings ISO/IEC 5259	Komponente	Maßnahme	Maximalwert	DA1.4	
DA1.4	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Die Daten des KI-Systems müssen in Entwicklung und Betrieb überwacht werden (können).	MA2.3									
						Komponente	Maßnahme	Normal	DA1.3	
DA1.5	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Es muss bewertet werden, ob die ergriffenen Maßnahmen die festgestellten Risiken auf ein annehmbares Maß vermindert haben und die Qualität des KI-Systems den gesetzten Zielvorgaben entspricht.	Bewertung – Zusammenfassende Betrachtung der Effekte der Durchführung der technischen und organisatorischen Maßnahmen – Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen – Ermittlung und Beschreibung des Restrisikos nach Durchführung der technischen und organisatorischen Maßnahmen – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Bewertung – Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Bewertung – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Eine Bewertung wurde nicht durchgeführt.		System	Bewertung	Maximalwert		

Datenqualität, -schutz und -Governance										
DA2		Schutz personenbezogener Daten								
DA2.1	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Maßnahme	Gewichtung	Verknüpfung	
Eine Risikoanalyse für den Schutz benutzter personenbezogener Daten im KI-System muss unter Beachtung des Verwendungszwecks durchgeführt werden.	Datenschutzfolgeabschätzung nach DSGVO Art.35 umgesetzt oder Analyse – Risiko Volle Risikoanalyse einschließlich Quantifizierung von Eintrittswahrscheinlichkeiten und Risiken: – Identifikation der möglichen Risiken und ihrer Ursachen – Zuweisung der Risikoverantwortung – Schätzung der Eintrittswahrscheinlichkeit – Schätzung der Aufdeckungswahrscheinlichkeit – Schätzung der Auswirkung – Strukturierte Abstufung und Priorisierung der Risiken Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen: – Datenverarbeitungsprozesse: Erhebung, Speicherung, Verarbeitung und Weitergabe personenbezogener Daten, insbesondere Risiken bei der Umsetzung des Rechts auf Datenübertragbarkeit oder Löschung. – Zweckbindung: Verwendung der Daten über den ursprünglich definierten Zweck hinaus. – Datenminimierung: Erhebung unnötiger oder übermäßiger personenbezogener Daten. – Datenzugriffsrechte und -Speicherung: Unbefugter oder unkontrollierter Zugang zu sensiblen personenbezogenen Daten und Aufbewahrung von Daten über notwendige Fristen hinaus. – Datenweitergabe an Dritte: Unkontrollierte Übertragung personenbezogener Daten an externe Partner oder Dienstleister, Risiken bei der Übermittlung personenbezogener Daten in Länder mit unzureichendem Datenschutz – Missbrauch von Daten: Potenzieller Missbrauch, wie Identitätsdiebstahl oder Diskriminierung durch unsachgemäße Datenverarbeitung. – Anonymisierung und Pseudonymisierung: Unzureichende Anonymisierungs- oder Pseudonymisierungstechniken. – Automatisierte Entscheidungsfindung: Mangelnde Transparenz oder Kontrolle bei automatisierten Entscheidungen, die auf personenbezogenen Daten basieren.	Analyse – Risiko Limitierte Risikoanalyse mit Fokus auf den Schutzbedarf ohne Quantifizierung von Wahrscheinlichkeiten – Identifikation der möglichen Risiken – Zuweisung der Risikoverantwortung – Schätzung der Auswirkung – Strukturierte Abstufung und Priorisierung der Risiken Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen: – Datenverarbeitungsprozesse: Erhebung, Speicherung, Verarbeitung und Weitergabe personenbezogener Daten, insbesondere Risiken bei der Umsetzung des Rechts auf Datenübertragbarkeit oder Löschung. – Zweckbindung: Verwendung der Daten über den ursprünglich definierten Zweck hinaus. – Datenminimierung: Erhebung unnötiger oder übermäßiger personenbezogener Daten. – Datenzugriffsrechte und -Speicherung: Unbefugter oder unkontrollierter Zugang zu sensiblen personenbezogenen Daten und Aufbewahrung von Daten über notwendige Fristen hinaus. – Datenweitergabe an Dritte: Unkontrollierte Übertragung personenbezogener Daten an externe Partner oder Dienstleister, Risiken bei der Übermittlung personenbezogener Daten in Länder mit unzureichendem Datenschutz – Missbrauch von Daten: Potenzieller Missbrauch, wie Identitätsdiebstahl oder Diskriminierung durch unsachgemäße Datenverarbeitung. – Anonymisierung und Pseudonymisierung: Unzureichende Anonymisierungs- oder Pseudonymisierungstechniken. – Automatisierte Entscheidungsfindung: Mangelnde Transparenz oder Kontrolle bei automatisierten Entscheidungen, die auf personenbezogenen Daten basieren.	Analyse – Risiko Hauptsächlich qualitative Abschätzung der Gefährdungen ohne Wahrscheinlichkeiten – Identifikation der möglichen Gefährdungen: – Zuweisung der Risikoverantwortung – Qualitative Schätzung der Auswirkung – Qualitative Abstufung und Priorisierung der Gefährdungen Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen: – Datenverarbeitungsprozesse: Erhebung, Speicherung, Verarbeitung und Weitergabe personenbezogener Daten, insbesondere Risiken bei der Umsetzung des Rechts auf Datenübertragbarkeit oder Löschung. – Datenzugriffsrechte und -Speicherung: Unbefugter oder unkontrollierter Zugang zu sensiblen personenbezogenen Daten und Aufbewahrung von Daten über notwendige Fristen hinaus. – Datenweitergabe an Dritte: Unkontrollierte Übertragung personenbezogener Daten an externe Partner oder Dienstleister, Risiken bei der Übermittlung personenbezogener Daten in Länder mit unzureichendem Datenschutz – Missbrauch von Daten: Potenzieller Missbrauch, wie Identitätsdiebstahl oder Diskriminierung durch unsachgemäße Datenverarbeitung. – Anonymisierung und Pseudonymisierung: Unzureichende Anonymisierungs- oder Pseudonymisierungstechniken.	Es wurde keine Risikoanalyse durchgeführt.	Relevant sind hier die DSGVO und die darin enthaltenen Vorgaben für eine Datenschutzfolgeabschätzung (DSFA), Art.35, DSGVO.	System	Analyse	Minimalwert	DA1.1, DA1.2, TR1.1 (Verwendungszweck)	
DA2.2	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Es müssen nicht-KI-spezifische Maßnahmen ergriffen werden, um unter Beachtung des identifizierten Risikos den Schutz personenbezogener Daten zu gewährleisten.	Maßnahmen entsprechend der Datenschutzfolgeabschätzung nach DSGVO Art.35 umgesetzt oder Orga. Maßnahmen – Governance / Systemnahe Prozesse – Implementierung von Mechanismen zur Datenminimierung, um nur die notwendigen personenbezogenen Daten zu erheben und zu verarbeiten. – Einführung strenger Zugriffskontrollen und von Berechtigungsmanagement, um sicherzustellen, dass nur autorisierte Personen Zugang zu personenbezogenen Daten haben. – Einführung von Mechanismen zum Speicherfristen-Management, um Daten nach Ablauf der festgelegten Fristen sicher zu löschen. – Sicherstellung, dass bei der Weitergabe von Daten an Dritte klare vertragliche Vereinbarungen bestehen – Einführung klarer Einwilligungsprozesse und transparenter Informationsmethoden für betroffene Personen. – Implementierung eines Notfallplans für die Erkennung, Meldung und Behebung von Datenschutzverletzungen innerhalb der gesetzlich vorgeschriebenen Frist. – Sicherstellung, dass bei Übermittlungen personenbezogener Daten in Drittländer geeignete Schutzmaßnahmen (z.B. Standardvertragsklauseln) implementiert sind. Tech. Maßnahmen – Daten – Sicherstellung, dass Datenschutzmaßnahmen bereits in der Entwicklung und Implementierung der Systeme integriert sind ("data-protection-by-design") – Einsatz von Verschlüsselungstechniken zum Schutz gespeicherter Daten – Einsatz von Anonymisierung, Pseudonymisierung und weiterer Technologien wie Differential Privacy, um die Informationen in personenbezogene Daten zu schützen mit Begründung der Notwendigkeit und Wirksamkeit der Maßnahmen – Implementierung von Prozessen, um betroffene Personen über automatisierte Entscheidungen und deren Logik zu informieren (ggfs. Bezug zu KI-spezifischen Maßnahmen zu Erklärbarkeit, siehe TR2) Tech. Maßnahmen – Betrieb – Regelmäßige Prüfungen und Audits der Datenverarbeitungsprozesse und -praktiken, um sicherzustellen, dass die Datenschutzanforderungen entsprechend dem Verwendungszweck eingehalten werden.	Orga. Maßnahmen – Governance / Systemnahe Prozesse – Einführung strenger Zugriffskontrollen und von Berechtigungsmanagement, um sicherzustellen, dass nur autorisierte Personen Zugang zu personenbezogenen Daten haben. – Einführung von Mechanismen zum Speicherfristen-Management, um Daten nach Ablauf der festgelegten Fristen sicher zu löschen. – Sicherstellung, dass bei der Weitergabe von Daten an Dritte klare vertragliche Vereinbarungen bestehen – Einführung klarer Einwilligungsprozesse und transparenter Informationsmethoden für betroffene Personen. – Implementierung eines Notfallplans für die Erkennung, Meldung und Behebung von Datenschutzverletzungen innerhalb der gesetzlich vorgeschriebenen Frist. – Sicherstellung, dass bei Übermittlungen personenbezogener Daten in Drittländer geeignete Schutzmaßnahmen (z.B. Standardvertragsklauseln) implementiert sind. Tech. Maßnahmen – Daten – Sicherstellung, dass Datenschutzmaßnahmen bereits in der Entwicklung und Implementierung der Systeme integriert sind ("data-protection-by-design") – Einsatz von Verschlüsselungstechniken zum Schutz gespeicherter Daten – Einsatz von Anonymisierung, Pseudonymisierung und weiterer Technologien wie Differential Privacy, um die Informationen in personenbezogene Daten zu schützen mit Begründung der Notwendigkeit und Wirksamkeit der Maßnahmen Tech. Maßnahmen – Betrieb – Regelmäßige Prüfungen und Audits der Datenverarbeitungsprozesse und -praktiken, um sicherzustellen, dass die Datenschutzanforderungen entsprechend dem Verwendungszweck eingehalten werden.	Orga. Maßnahmen – Governance / Systemnahe Prozesse – Einführung strenger Zugriffskontrollen und von Berechtigungsmanagement, um sicherzustellen, dass nur autorisierte Personen Zugang zu personenbezogenen Daten haben. – Einführung von Mechanismen zum Speicherfristen-Management, um Daten nach Ablauf der festgelegten Fristen sicher zu löschen. – Sicherstellung, dass bei der Weitergabe von Daten an Dritte klare vertragliche Vereinbarungen bestehen – Implementierung eines Notfallplans für die Erkennung, Meldung und Behebung von Datenschutzverletzungen innerhalb der gesetzlich vorgeschriebenen Frist. – Sicherstellung, dass bei Übermittlungen personenbezogener Daten in Drittländer geeignete Schutzmaßnahmen (z.B. Standardvertragsklauseln) implementiert sind. Tech. Maßnahmen – Daten – Einsatz von Verschlüsselungstechniken zum Schutz gespeicherter Daten – Einsatz von Anonymisierung, Pseudonymisierung und weiterer Technologien wie Differential Privacy, um die Informationen in personenbezogene Daten zu schützen mit Begründung der Notwendigkeit und Wirksamkeit der Maßnahmen	Es wurden keine nicht-KI-spezifischen Maßnahmen zum Schutz personenbezogener Daten ergriffen.	Relevant sind hier die DSGVO und die darin enthaltenen Vorgaben für eine Datenschutzfolgeabschätzung (DSFA), Art.35, DSGVO.	System/Komponente	Maßnahme	Maximalwert	DA1.2, DA1.3, TR2, MA1, CY1.5	
DA2.3	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Es müssen Maßnahmen und Sicherheitskontrollen vorhanden sein, um KI-spezifische Angriffe die während der Vorbereitungs- und Modelltrainingsphase der KI-Modelle stattfinden zu mitigieren.		CY2.3			Aus Plausibilitätsgründen sind nicht zwangsläufig alle Maßnahmen in CY2.3 relevant zum Schutz personenbezogener Daten, z.B., wenn es um das Schutzziel Availability geht.	System/Komponente	Maßnahme	Maximalwert	CY2.2, CY2.3	
DA2.4	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Es müssen Maßnahmen und Sicherheitskontrollen vorhanden sein, um KI-spezifische Angriffe die während des Betriebs auftreten können zu mitigieren.		CY2.2			Aus Plausibilitätsgründen sind nicht zwangsläufig alle Maßnahmen in CY1.6 relevant zum Schutz personenbezogener Daten, z.B., wenn es um das Schutzziel Availability geht.	System/Komponente	Maßnahme	Maximalwert	CY2.2, CY2.3	
DA2.5	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Es müssen Maßnahmen und Sicherheitskontrollen vorhanden sein, um fehlerhafte Nutzung und beabsichtigten Missbrauch des KI-Systems zu verhindern.		CY1.4			Dies sollte vor allem die fehlerhafte Nutzung oder den Missbrauch des Systems abdecken, welche personenbezogene Daten fälschlich verwendet oder offenlegt.	System/Komponente	Maßnahme	Normal		

Datenqualität, -schutz und -Governance											
DA2.6	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung		
Es muss für betroffene Personen und Nutzer möglich sein, ihre mit personenbezogenen Daten einhergehenden Rechte zu Datenverwaltung, -löschung und -benutzung sowie Informationspflichten auch im Betrieb des KI-Systems wahrzunehmen.	<p>Orga. Maßnahme – Benutzerinstruktion</p> <ul style="list-style-type: none">– Bereitstellung einer klaren und verständlichen Erklärung der Datenverarbeitung, der Betroffenenrechte und der Verantwortlichen– Dokumentation und Protokollierung ("Logging") aller Datenverarbeitungsaktivitäten sowie der Ausübung der Betroffenenrechte (siehe TR1.7) <p>Orga. Maßnahme – Governance</p> <ul style="list-style-type: none">– Gewährleistung, dass Nutzer über Änderungen der Datenverarbeitung oder -nutzung zeitnah informiert werden.– Einrichtung klarer Kanäle, um Betroffenen bei der Ausübung ihrer Rechte zu unterstützen (siehe MA1.4)– Implementierung eines Systems zur Einholung und Verwaltung der expliziten Einwilligung zur Datenverarbeitung. <p>Tech. Maßnahme – Betrieb</p> <ul style="list-style-type: none">– Bereitstellung leicht zugänglicher Funktionen für Betroffene, um ihre Rechte auf Auskunft, Berichtigung, Löschung und Datenübertragbarkeit auszuüben einschließlich Möglichkeiten der Datenverarbeitung zu widersprechen oder ihre Einwilligung jederzeit zu widerrufen.– Integration von Mechanismen, die eine automatische Löschung personen-bezogener Daten nach Ablauf der Speicherfrist oder auf Anfrage ermöglichen.	<p>Orga. Maßnahme – Benutzerinstruktion</p> <ul style="list-style-type: none">– Bereitstellung einer klaren und verständlichen Erklärung der Datenverarbeitung, der Betroffenenrechte und der Verantwortlichen– Dokumentation und Protokollierung ("Logging") aller Datenverarbeitungsaktivitäten sowie der Ausübung der Betroffenenrechte (siehe TR1.7) <p>Orga. Maßnahme – Governance</p> <ul style="list-style-type: none">– Einrichtung klarer Kanäle, um Betroffenen bei der Ausübung ihrer Rechte zu unterstützen (siehe MA1.4)– Implementierung eines Systems zur Einholung und Verwaltung der expliziten Einwilligung zur Datenverarbeitung. <p>Tech. Maßnahme – Betrieb</p> <ul style="list-style-type: none">– Integration von Mechanismen, die eine automatische Löschung personen-bezogener Daten nach Ablauf der Speicherfrist oder auf Anfrage ermöglichen.	<p>Orga. Maßnahme – Benutzerinstruktion</p> <ul style="list-style-type: none">– Bereitstellung einer klaren und verständlichen Erklärung der Datenverarbeitung, der Betroffenenrechte und der Verantwortlichen <p>Orga. Maßnahme – Governance</p> <ul style="list-style-type: none">– Einrichtung klarer Kanäle, um Betroffenen bei der Ausübung ihrer Rechte zu unterstützen (siehe MA1.4)– Implementierung eines Systems zur Einholung und Verwaltung der expliziten Einwilligung zur Datenverarbeitung. <p>Tech. Maßnahme – Betrieb</p> <ul style="list-style-type: none">– Integration von Mechanismen, die eine automatische Löschung personen-bezogener Daten nach Ablauf der Speicherfrist oder auf Anfrage ermöglichen.	Möglichkeiten für natürliche Personen ihre Rechte in Bezug auf personenbezogene Daten wahrzunehmen bestehen nicht.		Komponente	Maßnahme	Maximalwert	TR1.7, MA1.4		
DA2.7	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung		
Es muss bewertet werden, ob die ergriffenen Maßnahmen die festgestellten Risiken auf ein annehmbares Maß vermindert haben und die Qualität des KI-Systems den gesetzten Zielvorgaben entspricht.	<p>Bewertung</p> <ul style="list-style-type: none">– Zusammenfassende Betrachtung der Effekte der Durchführung der technischen und organisatorischen Maßnahmen– Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitäts-dimensionen– Ermittlung und Beschreibung des Restrisikos nach Durchführung der technischen und organisatorischen Maßnahmen– Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist– Begründung der Tolerierbarkeit des Restrisikos	<p>Bewertung</p> <ul style="list-style-type: none">– Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen– Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist– Begründung der Tolerierbarkeit des Restrisikos	<p>Bewertung</p> <ul style="list-style-type: none">– Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist– Begründung der Tolerierbarkeit des Restrisikos	Eine Bewertung wurde nicht durchgeführt		System	Bewertung	Maximalwert			
DA3	Schutz proprietärer Daten										
DA3.1	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung		
Eine Risikoanalyse für den Schutz benutzter proprietärer Daten im KI-System muss unter Beachtung des Verwendungszwecks durchgeführt werden.	<p>Analyse – Risiko</p> <p>Volle Risikoanalyse einschließlich Quantifizierung von Eintrittswahrscheinlichkeiten und Risiken:</p> <ul style="list-style-type: none">– Identifikation der möglichen Risiken und ihrer Ursachen– Zuweisung der Risikoverantwortung– Schätzung der Eintrittswahrscheinlichkeit– Schätzung der Aufdeckungswahrscheinlichkeit– Schätzung der Auswirkung– Strukturierte Abstufung und Priorisierung der Risiken <p>Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen:</p> <ul style="list-style-type: none">– Unbefugter Zugriff und mangelnde Kontrolle über den Zugang zu sensiblen Firmendaten durch interne oder externe Akteure.– Risiken von versehentlichem Verlust, Löschung oder Beschädigung proprietärer Daten durch Systemfehler oder menschliches Versagen.– Gefahr durch externe Angriffe, Hacking oder Industriespionage, die auf den Diebstahl, Offenlegung oder Manipulation vertraulicher Firmendaten abzielen– Kontrollverlust über Firmendaten bei deren Speicherung oder Verarbeitung durch externe Dienstleister (z.B. in Cloud-Umgebungen).– Vertrags- oder Vereinbarungsverstöße durch Partner oder Dienstleister hinsichtlich unzulässigen Datenzugriff oder nicht-autorisierte Datenweitergabe– Nichteinhaltung von branchenspezifischen Vorschriften, bezüglich der Nutzung und Weitergabe von Daten.	<p>Analyse – Risiko</p> <p>Limitierte Risikoanalyse mit Fokus auf den Schutzbedarf ohne Quantifizierung von Wahrscheinlichkeiten</p> <ul style="list-style-type: none">– Identifikation der möglichen Risiken– Zuweisung der Risikoverantwortung– Schätzung der Auswirkung– Strukturierte Abstufung und Priorisierung der Risiken <p>Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen:</p> <ul style="list-style-type: none">– Unbefugter Zugriff und mangelnde Kontrolle über den Zugang zu sensiblen Firmendaten durch interne oder externe Akteure.– Risiken von versehentlichem Verlust, Löschung oder Beschädigung proprietärer Daten durch Systemfehler oder menschliches Versagen.– Gefahr durch externe Angriffe, Hacking oder Industriespionage, die auf den Diebstahl, Offenlegung oder Manipulation vertraulicher Firmendaten abzielen– Kontrollverlust über Firmendaten bei deren Speicherung oder Verarbeitung durch externe Dienstleister (z.B. in Cloud-Umgebungen).– Vertrags- oder Vereinbarungsverstöße durch Partner oder Dienstleister hinsichtlich unzulässigen Datenzugriff oder nicht-autorisierte Datenweitergabe– Nichteinhaltung von branchenspezifischen Vorschriften, bezüglich der Nutzung und Weitergabe von Daten.	<p>Analyse – Risiko</p> <p>Hauptsächlich qualitative Abschätzung der Gefährdungen ohne Wahrscheinlichkeiten</p> <ul style="list-style-type: none">– Identifikation der möglichen Gefährdungen:– Zuweisung der Risikoverantwortung– Qualitative Schätzung der Auswirkung– Qualitative Abstufung und Priorisierung der Gefährdungen <p>Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen:</p> <ul style="list-style-type: none">– Unbefugter Zugriff und mangelnde Kontrolle über den Zugang zu sensiblen Firmendaten durch interne oder externe Akteure.– Risiken von versehentlichem Verlust, Löschung oder Beschädigung proprietärer Daten durch Systemfehler oder menschliches Versagen.– Gefahr durch externe Angriffe, Hacking oder Industriespionage, die auf den Diebstahl, Offenlegung oder Manipulation vertraulicher Firmendaten abzielen– Kontrollverlust über Firmendaten bei deren Speicherung oder Verarbeitung durch externe Dienstleister (z.B. in Cloud-Umgebungen).– Vertrags- oder Vereinbarungsverstöße durch Partner oder Dienstleister hinsichtlich unzulässigen Datenzugriff oder nicht-autorisierte Datenweitergabe– Nichteinhaltung von branchenspezifischen Vorschriften, bezüglich der Nutzung und Weitergabe von Daten.	Es wurde keine Risiko- oder Gefährdungsanalyse durchgeführt.	Relevant sind hier die DSGVO und die darin enthaltenen Vorgaben für eine Datenschutzfolgeabschätzung (DSFA), Art.35, DSGVO	System	Analyse	Maximalwert	DA1.1, DA1.2, TR1.1 (Verwendungszweck)		
DA3.2	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung		
Es müssen nicht-KI-spezifische Maßnahmen ergriffen werden um unter Beachtung des identifizierten Risikos den Schutz proprietärer Daten zu gewährleisten.	<p>Orga. Maßnahmen – Governance / Systemnahe Prozesse</p> <ul style="list-style-type: none">– Einführung von Mechanismen zum Speicherfristen-Management, um Daten nach Ablauf der festgelegten Fristen sicher zu löschen.– Klare vertragliche Vereinbarungen mit externen Dienstleistern über den Zugriff, die Nutzung und den Schutz proprietärer Daten– Gewährleistung, dass Unternehmen jederzeit die Kontrolle über ihre Daten behalten, insbesondere bei der Nutzung externer Cloud- oder Speicherdienste.– Implementierung eines Notfallplans für die Erkennung, Meldung und Behebung von Datenschutzverletzungen <p>Tech. Maßnahmen – Daten</p> <ul style="list-style-type: none">– Einführung strenger Zugriffskontrollen und von Berechtigungsmanagement, um sicherzustellen, dass nur autorisierte Personen Zugang zu proprietären Daten haben.– Sicherstellung, dass Datenschutzmaßnahmen bereits in der Entwicklung und Implementierung der Systeme integriert sind ("data-protection-by-design")– Implementierung von Verschlüsselungstechniken und Sicherheitsprotokollen, um die Vertraulichkeit und Integrität der Firmendaten zu schützen, insbesondere während der Übertragung und Speicherung.– Einführung von Mechanismen zur Sicherstellung der Integrität und Authentizität von Firmendaten, um Manipulationen zu verhindern und deren Echtheit zu gewährleisten.– Einführung regelmäßiger Datensicherungen und Notfallpläne zur Wiederherstellung von Daten bei Verlust oder Beschädigung.– Sicherstellung der Interoperabilität von Datenformaten und die Möglichkeit, Firmendaten einfach zwischen Systemen oder Anbietern zu übertragen <p>Tech. Maßnahmen – Betrieb</p> <ul style="list-style-type: none">– Regelmäßige Prüfungen und Audits der Datenverarbeitungsprozesse und -praktiken, um sicherzustellen, dass die Datenschutzeranforderungen entsprechend dem Verwendungszweck eingehalten werden.	<p>Orga. Maßnahmen – Governance / Systemnahe Prozesse</p> <ul style="list-style-type: none">– Einführung von Mechanismen zum Speicherfristen-Management, um Daten nach Ablauf der festgelegten Fristen sicher zu löschen.– Klare vertragliche Vereinbarungen mit externen Dienstleistern über den Zugriff, die Nutzung und den Schutz proprietärer Daten– Gewährleistung, dass Unternehmen jederzeit die Kontrolle über ihre Daten behalten, insbesondere bei der Nutzung externer Cloud- oder Speicherdienste.– Implementierung eines Notfallplans für die Erkennung, Meldung und Behebung von Datenschutzverletzungen <p>Tech. Maßnahmen – Daten</p> <ul style="list-style-type: none">– Einführung strenger Zugriffskontrollen und von Berechtigungsmanagement, um sicherzustellen, dass nur autorisierte Personen Zugang zu proprietären Daten haben.– Implementierung von Verschlüsselungstechniken und Sicherheitsprotokollen, um die Vertraulichkeit und Integrität der Firmendaten zu schützen, insbesondere während der Übertragung und Speicherung.– Einführung von Mechanismen zur Sicherstellung der Integrität und Authentizität von Firmendaten, um Manipulationen zu verhindern und deren Echtheit zu gewährleisten.– Einführung regelmäßiger Datensicherungen und Notfallpläne zur Wiederherstellung von Daten bei Verlust oder Beschädigung. <p>Tech. Maßnahmen – Betrieb</p> <ul style="list-style-type: none">– Regelmäßige Prüfungen und Audits der Datenverarbeitungsprozesse und -praktiken, um sicherzustellen, dass die Datenschutzeranforderungen entsprechend dem Verwendungszweck eingehalten werden.	<p>Orga. Maßnahmen – Governance / Systemnahe Prozesse</p> <ul style="list-style-type: none">– Einführung von Mechanismen zum Speicherfristen-Management, um Daten nach Ablauf der festgelegten Fristen sicher zu löschen.– Klare vertragliche Vereinbarungen mit externen Dienstleistern über den Zugriff, die Nutzung und den Schutz proprietärer Daten– Gewährleistung, dass Unternehmen jederzeit die Kontrolle über ihre Daten behalten, insbesondere bei der Nutzung externer Cloud- oder Speicherdienste. <p>Tech. Maßnahmen – Daten</p> <ul style="list-style-type: none">– Einführung von Zugriffskontrollen und von Berechtigungsmanagement, um sicherzustellen, dass nur autorisierte Personen Zugang zu proprietären Daten haben.– Implementierung von Verschlüsselungstechniken, um die Vertraulichkeit und Integrität der Firmendaten zu schützen, insbesondere während der Übertragung und Speicherung.– Einführung regelmäßiger Datensicherungen und Notfallpläne zur Wiederherstellung von Daten bei Verlust oder Beschädigung.	Es wurden keine nicht-KI-spezifischen Maßnahmen zum Schutz proprietärer Daten ergriffen.			System/Komponente	Maßnahme	Maximalwert	DA1.2, DA1.3, MA1, CY1.5	
DA3.3	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung		
Es müssen Maßnahmen und Sicherheitskontrollen vorhanden sein, um KI-spezifische Angriffe die während der Vorbereitungs- und Modelltrainingsphase der KI-Modelle stattfinden zu mitigieren.		Siehe: CY2.3			Aus Plausibilitätsgründen sind nicht zwangsläufig alle Maßnahmen in CY2.3 relevant zum Schutz proprietärer Daten, z.B., wenn es um das Schutzziel Availability geht.	System/Komponente	Maßnahme	Maximalwert			

Datenqualität, -schutz und -Governance									
DA3.4	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Es müssen Maßnahmen und Sicherheitskontrollen vorhanden sein, um KI-spezifische Angriffe die während des Betriebs auftreten können zu mitigieren.	Siehe: CY2.2				Aus Plausibilitätsgründen sind nicht zwangsläufig alle Maßnahmen in CY2.2 relevant zum Schutz proprietärer Daten, z.B., wenn es um das Schutzziel Availability geht.	System/Komponente	Maßnahme	Maximalwert	
DA3.5	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Es müssen Maßnahmen und Sicherheitskontrollen vorhanden sein, um fehlerhafte Nutzung und beabsichtigten Missbrauch des KI-Systems zu verhindern.	Siehe: CY1.4				Dies sollte vor allem die fehlerhafte Nutzung oder den Missbrauch des Systems abdecken, welche proprietäre Daten fälschlich verwendet oder offenlegt.	System/Komponente	Maßnahme	Maximalwert	
DA3.6	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Es muss bewertet werden, ob die ergriffenen Maßnahmen die festgestellten Risiken auf ein annehmbares Maß vermindert haben und die Qualität des KI-Systems den gesetzten Zielvorgaben entspricht.	Bewertung – Zusammenfassende Betrachtung der Effekte der Durchführung der technischen und organisatorischen Maßnahmen – Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen – Ermittlung und Beschreibung des Restrisikos nach Durchführung der technischen und organisatorischen Maßnahmen – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Bewertung – Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Bewertung – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Eine Bewertung wurde nicht durchgeführt.		System	Bewertung	Maximalwert	

Nicht-Diskriminierung										
Vermeidung von ungerechtfertigter Diskriminierung und Verzerrung										
ND1										
ND1.1	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung	
Das Risiko von Diskriminierung im Zusammenhang mit dem beabsichtigten Zweck des KI-Systems muss analysiert werden.	Analyse-Risiko Detaillierte Risikoanalyse einschließlich Quantifizierung von Eintrittswahrscheinlichkeiten und Risiken: <ul style="list-style-type: none">– Identifikation der möglichen Risiken und ihrer Ursachen– Zuweisung der Risikoverantwortung– Schätzung der Eintrittswahrscheinlichkeit– Schätzung der Aufdeckungswahrscheinlichkeit– Schätzung der Auswirkung– Strukturierte Abstufung und Priorisierung der Risiken Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen: <ul style="list-style-type: none">– Historische Verzerrungen– Stichprobenverzerrung– Modellverzerrungen– Attributionsfehler– Impliziter Bias– Bestätigungsfehler– Voreingenommenheit von Entwicklern– Verzerrung durch Nutzerinteraktionen– Missinterpretation von Resultaten– Mangelnde Berücksichtigung der möglichen (Nutzer-)Gruppen– Benachteiligung/Diskriminierung von Gruppen mit geschützten Eigenschaften	Analyse-Risiko Limitierte Risikoanalyse mit Fokus auf den Schutzbedarf ohne Quantifizierung von Wahrscheinlichkeiten: <ul style="list-style-type: none">– Identifikation der möglichen Risiken– Zuweisung der Risikoverantwortung– Schätzung der Auswirkung– Strukturierte Abstufung und Priorisierung der Risiken Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen: <ul style="list-style-type: none">– Historische Verzerrungen– Stichprobenverzerrung– Modellverzerrungen– Attributionsfehler– Impliziter Bias– Bestätigungsfehler– Voreingenommenheit von Entwicklern– Verzerrung durch Nutzerinteraktionen– Missinterpretation von Resultaten– Mangelnde Berücksichtigung der möglichen (Nutzer-)Gruppen– Benachteiligung/Diskriminierung von Gruppen mit geschützten Eigenschaften	Analyse-Risiko Hauptsächlich qualitative Abschätzung ohne Wahrscheinlichkeiten: <ul style="list-style-type: none">– Identifikation der möglichen Gefährdungen– Zuweisung der Risikoverantwortung– Qualitative Schätzung der Auswirkung– Qualitative Abstufung und Priorisierung der Risiken Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen: <ul style="list-style-type: none">– Historische Verzerrungen– Stichprobenverzerrung– Modellverzerrungen– Attributionsfehler– Impliziter Bias– Bestätigungsfehler– Voreingenommenheit von Entwicklern– Verzerrung durch Nutzerinteraktionen– Missinterpretation von Resultaten– Mangelnde Berücksichtigung der möglichen (Nutzer-)Gruppen– Benachteiligung/Diskriminierung von Gruppen mit geschützten Eigenschaften	Es wurde keine Risikoanalyse durchgeführt.		System	Analyse	Maximalwert	TR1.1	
ND1.2	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung	
Das notwendige Maß an Vermeidung von ungerechtfertigter Verzerrung bzw. Diskriminierung muss im Kontext des beabsichtigten Zwecks definiert werden.	Analyse – Definition / Metriken & Schwellenwerte Basierend auf dem festgelegten Zweck des KI-Systems (TR1.1) muss definiert werden welches Maß an Vermeidung von Verzerrung bzw. Schutz vor Nicht-Diskriminierung notwendig ist. Hierzu gehört: <ul style="list-style-type: none">– Identifikation sensibler bzw. schützenswerter Merkmale in den Daten mit Begründung– Identifikation von Gruppen mit geschützten Eigenschaften basierend auf den möglichen Nutzer*innen und betroffenen Personen– Identifikation zusätzlicher Gruppen, die unvorhergesehen betroffen werden können, z.B. da deren Merkmale nicht als Features sondern nur implizit in den Daten enthalten sind)– Definition der beabsichtigten Fairness– Kollaboration mit Repräsentanten von identifizierten Gruppen mit geschützten Eigenschaften– Festlegung von Test, Metriken und Schwellenwerten zur Erfassung der Zielsetzung im Rahmen der Fairness-Definition Die ausgewählten Metriken und Tests sollten mindestens die folgenden Merkmale aufweisen: <ul style="list-style-type: none">– Methoden sollten für jedes KI-Modell im und für das KI-System als Ganzes angewendet werden, falls dieses einen gesammelten Output erzeugt– Wenn durch die Auswahl begründet, sollten die Metriken und Tests sowohl wichtige Basismethoden als auch fortgeschrittene Methoden (hinsichtlich Komplexität, Informationsgehalts, Implementierungsaufwands, z.B. umfangreiches Prüfwerkzeug) umfassen– Die ausgewählten Metriken und Tests sollten wenn möglich durch einen hohen Grad der Automatisierung eine kontinuierliche Evaluation der KI-Modelle erlauben	Analyse – Definition / Metriken & Schwellenwerte Basierend auf dem festgelegten Zweck des KI-Systems (TR1.1) muss definiert werden welches Maß an Vermeidung von Verzerrung bzw. Schutz vor Nicht-Diskriminierung notwendig ist. Hierzu gehört: <ul style="list-style-type: none">– Identifikation sensibler bzw. schützenswerter Merkmale in den Daten mit Begründung– Identifikation von Gruppen mit geschützten Eigenschaften basierend auf den möglichen Nutzer*innen und betroffenen Personen– Definition der beabsichtigten Fairness– Festlegung von Test, Metriken und Schwellenwerten zur Erfassung der Zielsetzung im Rahmen der Fairness-Definition Die ausgewählten Metriken und Tests sollten mindestens die folgenden Merkmale aufweisen: <ul style="list-style-type: none">– Methoden sollten für jedes KI-Modell im und für das KI-System als Ganzes angewendet werden, falls dieses einen gesammelten Output erzeugt– Wenn durch die Auswahl begründet, sollten die Metriken und Tests sowohl wichtige Basismethoden als auch fortgeschrittene Methoden (hinsichtlich Komplexität, Informationsgehalts, Implementierungsaufwands, z.B. umfangreiches Prüfwerkzeug) umfassen	Analyse – Definition / Metriken & Schwellenwerte Basierend auf dem festgelegten Zweck des KI-Systems (TR1.1) muss definiert werden welches Maß an Vermeidung von Verzerrung bzw. Schutz vor Nicht-Diskriminierung notwendig ist. Hierzu gehört: <ul style="list-style-type: none">– Identifikation sensibler bzw. schützenswerter Merkmale in den Daten– Identifikation von Gruppen mit geschützten Eigenschaften basierend auf den möglichen Nutzer*innen und betroffenen Personen– Definition der beabsichtigten Fairness– Festlegung von Test, Metriken und Schwellenwerten zur Erfassung der Zielsetzung im Rahmen der Fairness-Definition Die ausgewählten Metriken und Tests sollten mindestens die folgenden Merkmale aufweisen: <ul style="list-style-type: none">– Methoden sollten für jedes KI-Modell im und für das KI-System als Ganzes angewendet werden, falls dieses einen gesammelten Output erzeugt– Wenn durch die Auswahl begründet, sollten die Metriken und Tests sowohl wichtige Basismethoden als auch fortgeschrittene Methoden (hinsichtlich Komplexität, Informationsgehalts, Implementierungsaufwands, z.B. umfangreiches Prüfwerkzeug) umfassen	Eine Definition und Zielsetzung zur Vermeidung ungerechtfertigter Verzerrung und Diskriminierung ist nicht erfolgt.		System/Komponente	Analyse	Maximalwert	TR1.1	
ND1.3	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung	
Eine Analyse des KI-Systems hinsichtlich bestehender Verzerrungen und möglicher Diskriminierung muss durchgeführt werden.	Tech. Maßnahme – Modell, Daten & Tests Das KI-System und die darin enthaltenen Modelle und Daten wurden in Bezug auf die in ND1.2 festgelegte Zielsetzung hin untersucht. Das beinhaltet mindestens: <ul style="list-style-type: none">– Begründung der Modellauswahl (einschließlich der Aufgabenstellung und Optimierungsstrategie, sowie ggf. Vor- oder Nachverarbeitungsmaßnahmen im KI-System zur Abschwächung von Verzerrungen)– Beschreibung des Umfangs und Durchführung von Test (siehe ND1.2)– Abgleich mit alternativen Tests/Metriken– Bewertung der Einhaltung vorgegebener Schwellen- bzw. Zielwerte– Dokumentation der Testergebnisse und daraus abgeleiteter Schlussfolgerungen– Beschreibung der Grenzen der Aussagekraft der durchgeführten Tests und ihrer Ergebnisse– Kollaboration mit relevanten Gruppen mit geschützten Eigenschaften	Tech. Maßnahme – Modell, Daten & Tests Das KI-System und die darin enthaltenen Modelle und Daten wurden in Bezug auf die in ND1.2 festgelegte Zielsetzung hin untersucht. Das beinhaltet mindestens: <ul style="list-style-type: none">– Beschreibung des Umfangs und Durchführung von Test (siehe ND1.2)– Abgleich mit alternativen Tests/Metriken– Bewertung der Einhaltung vorgegebener Schwellen- bzw. Zielwerte– Dokumentation der Testergebnisse und daraus abgeleiteter Schlussfolgerungen	Tech. Maßnahme – Modell, Daten & Tests Das KI-System und die darin enthaltenen Modelle und Daten wurden in Bezug auf die in ND1.2 festgelegte Zielsetzung hin untersucht. Das beinhaltet mindestens: <ul style="list-style-type: none">– Beschreibung des Umfangs und Durchführung von Test (siehe ND1.2)– Bewertung der Einhaltung vorgegebener Schwellen- bzw. Zielwerte– Dokumentation der Testergebnisse und daraus abgeleiteter Schlussfolgerungen	Es wurde keine Analyse hinsichtlich bestehender Verzerrungen und möglicher Diskriminierung durchgeführt.		System/Komponente	Analyse	Maximalwert		
ND1.4	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung	
Bei der Entwicklung eines KI-Systems müssen die Grundsätze des universellen Designs angewandt werden, um die Zugänglichkeit für Menschen mit Einschränkungen zu gewährleisten.	Tech. Maßnahme – Betrieb Dort wo das KI-System mit natürlichen Personen in Kontakt kommt, wird die Zugänglichkeit für mögliche Personen mit Einschränkungen untersucht und durch Maßnahmen unterstützt. Hierzu gehört die Berücksichtigung der folgenden Maßnahmen: <ul style="list-style-type: none">– Gestaltung der Benutzeroberfläche des KI-Systems berücksichtigt Personen mit Einschränkungen (z.B. Rot-Grün-Schwäche oder motorische Einschränkungen)– Informationen über Ergebnisse und Mechanismen, auch bspw. über eine Benutzeroberfläche, sind adressatengerecht aufbereitet und berücksichtigen mögliche Einschränkungen der Empfänger– Die Gestaltung der Benutzeroberfläche bzw. bereitgestellter Informationen basiert auf Kollaboration mit Nutzergruppen, die Personen mit Einschränkungen umfassen	Tech. Maßnahme – Betrieb Dort wo das KI-System mit natürlichen Personen in Kontakt kommt, wird die Zugänglichkeit für mögliche Personen mit Einschränkungen untersucht und durch Maßnahmen unterstützt. Hierzu gehört die Berücksichtigung der folgenden Maßnahmen: <ul style="list-style-type: none">– Gestaltung der Benutzeroberfläche des KI-Systems berücksichtigt Personen mit Einschränkungen (z.B. Rot-Grün-Schwäche oder motorische Einschränkungen)– Informationen über Ergebnisse und Mechanismen, auch bspw. über eine Benutzeroberfläche, sind adressatengerecht aufbereitet und berücksichtigen mögliche Einschränkungen der Empfänger	Tech. Maßnahme – Betrieb Dort wo das KI-System mit natürlichen Personen in Kontakt kommt, wird die Zugänglichkeit für mögliche Personen mit Einschränkungen untersucht und durch Maßnahmen unterstützt. Hierzu gehört die Berücksichtigung der folgenden Maßnahmen: <ul style="list-style-type: none">– Informationen über Ergebnisse und Mechanismen, auch bspw. über eine Benutzeroberfläche, sind adressatengerecht aufbereitet und berücksichtigen mögliche Einschränkungen der Empfänger	Grundsätze des universellen Designs wurden in der Entwicklung nicht betrachtet.		System	Maßnahme	Normal		

Nicht-Diskriminierung										
ND1.5	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung	
Das Personal, das an der Entwicklung des KI-Systems und an der Umsetzung der Maßnahmen zur Verringerung des Risikos von ungerechtfertigter Verzerrung und Diskriminierung beteiligt ist, muss informiert und geschult werden.	Orga. Maßnahmen – Schulung Ein zielgruppenorientiertes Schulungsprogramm wird für alle mit der Entwicklung des KI-Systems betrauten internen und externen Mitarbeiter betrieben. Dieses Programm umfasst mindestens die folgenden Aspekte: – Vermittlung der Zweckbestimmung des KI-Systems – Sensibilisierung zu unterschiedlichen möglichen Verzerrungen und dem Erkennen von potenziellen Diskriminierungen im Kontext von KI-Systemen – Umgang mit relevanten Datentypen (z.B. Trainings- und Validierungsdaten, Betriebsdaten, Kundendaten) gemäß den geltenden Richtlinien sowie gesetzlichen und regulatorischen Anforderungen – Anforderungen an die Durchführung von Trainings, Validierungen und Tests der KI-Modells im Kontext von Verzerrungen, Fairness und Diskriminierung Das Programm wird regelmäßig basierend auf den gewonnenen Erkenntnissen, Änderungen der Richtlinien aktualisiert. Die verwendeten Inhalte und die Teilnahme am Programm werden dokumentiert.	Orga. Maßnahmen – Schulung Ein Schulungsprogramm wird für alle mit der Entwicklung des KI-Systems betrauten internen und externen Mitarbeiter betrieben. Dieses Programm umfasst mindestens die folgenden Aspekte: – Sensibilisierung zu unterschiedlichen möglichen Verzerrungen und dem Erkennen von potenziellen Diskriminierungen im Kontext von KI-Systemen – Umgang mit relevanten Datentypen (z.B. Trainings- und Validierungsdaten, Betriebsdaten, Kundendaten) gemäß den geltenden Richtlinien sowie gesetzlichen und regulatorischen Anforderungen – Anforderungen an die Durchführung von Trainings, Validierungen und Tests der KI-Modells im Kontext von Verzerrungen, Fairness und Diskriminierung Die verwendeten Inhalte und die Teilnahme am Programm werden dokumentiert.	Orga. Maßnahmen – Schulung Ein Schulungsprogramm wird für alle mit der Entwicklung des KI-Systems betrauten internen und externen Mitarbeiter betrieben. Dieses Programm umfasst mindestens die folgenden Aspekte: – Sensibilisierung zu unterschiedlichen möglichen Verzerrungen und dem Erkennen von potenziellen Diskriminierungen im Kontext von KI-Systemen – Anforderungen an die Durchführung von Trainings, Validierungen und Tests der KI-Modells im Kontext von Verzerrungen, Fairness und Diskriminierung Die verwendeten Inhalte und die Teilnahme am Programm werden dokumentiert.	Es ist kein Anleiten des mit der KI-Entwicklung betrauten Personals im Kontext von Verzerrungen und Diskriminierung erfolgt.		System	Maßnahme	Normal		
ND1.6	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung	
Technische Maßnahmen zur Vermeidung von ungerechtfertigten Verzerrungen und Diskriminierungen müssen während der Entwicklung des KI-Systems ergriffen werden.	Tech. Maßnahmen – Daten Die zum Training, zur Validierung und zum Testen verwendeten Daten spiegeln die Fairness-Zielsetzung wider und die Testdaten sind geeignet, um mögliche Verzerrungen im KI-System aufzudecken. Hierzu zählt mindestens: – Dokumentation der Daten, Auswahl- oder Erhebungsverfahren und Aufbereitungsschritte (siehe DA 1.1) – Durchführung von Aufbereitungsschritten zur Verminderung von ungerechtfertigten Verzerrungen in den Daten – Nutzen- und Angemessenheitsargumentation der Maßnahmen in den Trainings- und Testdaten in Bezug auf das Risiko von ungerechtfertigten Verzerrungen und Diskriminierungen)	Tech. Maßnahmen – Daten Die zum Training, zur Validierung und zum Testen verwendeten Daten spiegeln die Fairness-Zielsetzung wider und die Testdaten sind geeignet, um mögliche Verzerrungen im KI-System aufzudecken. Hierzu zählt mindestens: – Dokumentation der Daten, Auswahl- oder Erhebungsverfahren und Aufbereitungsschritte (siehe DA 1.1) – Durchführung von Aufbereitungsschritten zur Verminderung von ungerechtfertigten Verzerrungen in den Daten	Tech. Maßnahmen – Daten Die zum Training, zur Validierung und zum Testen verwendeten Daten spiegeln die Fairness-Zielsetzung wider und die Testdaten sind geeignet, um mögliche Verzerrungen im KI-System aufzudecken. Hierzu zählt mindestens: – Dokumentation der Daten, Auswahl- oder Erhebungsverfahren und Aufbereitungsschritte (siehe DA 1.1) – Durchführung von Aufbereitungsschritten zur Verminderung von ungerechtfertigten Verzerrungen in den Daten	Es wurden keine Maßnahmen zur Vermeidung von ungerechtfertigten Verzerrungen in der Entwicklung ergriffen.		System/Komponente	Maßnahme	Normal	DA1.1	
ND1.7	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung	
Ein Mechanismus für Nutzerfeedback muss verfügbar sein, um Probleme im Zusammenhang mit möglicher Diskriminierung zu melden	Siehe: MA1.4						System	Maßnahme	Normal	
ND1.8	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung	
Das KI-System muss in Entwicklung und Betrieb zum Zweck der Vermeidung unerwünschter Verzerrung und Diskriminierung überwacht werden (können)	Siehe: MA2.3						System	Maßnahme	Normal	
ND1.9	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung	
Es muss bewertet werden, ob die ergriffenen Maßnahmen die festgestellten Risiken auf ein annehmbares Maß vermindert haben und die Qualität des KI-Systems den gesetzten Zielvorgaben entspricht	Bewertung – Zusammenfassende Betrachtung der Effekte der Durchführung der technischen und organisatorischen Maßnahmen – Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen – Ermittlung und Beschreibung des Restrisikos nach Durchführung der technischen und organisatorischen Maßnahmen – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Bewertung – Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Bewertung – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Eine Bewertung wurde nicht durchgeführt		System	Bewertung	Maximalwert		

Transparenz										
TR1	Rückverfolgbarkeit & Dokumentation									
TR1.1	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Verwendungszweck und Anwendungsbereich des KI-Systems sind klar definiert und beschrieben.	Analyse – Definition Der Verwendungszweck ist definiert im Sinne von: i) Verbesserungen und Vorteile, die mit dem System erreicht werden können und ii) Funktionalitäten des KI-Systems, mithilfe derer diese Verbesserungen und Vorteile realisierbar sein sollen iii) angestrebte/mögliche/erlaubte Nutzergruppen und Betroffene Personen ("Data Subjects") Der Anwendungsbereich ist mit Bezug auf den Verwendungszweck definiert und umfasst: – die zu erwarteten Eingabedaten und angestrebte Systemausgaben (inklusive Format und Inhalt) – eine umfassende Beschreibung von Einsatzkontexten/ Umgebungen, für die das KI-System geeignet ist und auch solche, in denen das KI-System nicht eingesetzt werden darf – Beschreibung von vorhersehbarer missbräuchlicher oder fehlgeleiteter Anwendung des KI-Systems – falls anwendbar: Definition einer ODD Die Einsehbarkeit der Definitionen ist für relevante Interessensgruppen möglich.	Analyse – Definition Der Verwendungszweck ist definiert im Sinne von Funktionalitäten des KI-Systems sowie angestrebte/mögliche/erlaubte Nutzergruppen und Betroffene Personen ("Data Subjects"). Der Anwendungsbereich ist mit Bezug auf den Verwendungszweck definiert und umfasst: – die zu erwarteten Eingabedaten und angestrebte Systemausgaben (inklusive Format und Inhalt) – Grobe Beschreibung von Einsatzkontexten/ Umgebungen, für die das KI-System geeignet ist und auch solche, in denen das KI-System nicht eingesetzt werden darf	Analyse – Definition Der Verwendungszweck ist definiert im Sinne von Funktionalitäten des KI-Systems. Der Anwendungsbereich ist mit Bezug auf den Verwendungszweck definiert und umfasst: – die zu erwarteten Eingabedaten und angestrebte Systemausgaben (inklusive Format und Inhalt) – Grobe Beschreibung von Einsatzkontexten/ Umgebungen, für die das KI-System geeignet ist	Der Verwendungszweck und Anwendungsbereich sind nicht klar definiert.	Siehe auch Glossar	System	Analyse	Maximalwert	global relevant; insbesondere TR1.2, TR2.1, VE1.1, VE2.1, CY1.1, DA2.1, DA3.1, ND1.1, MA1.2, MA2.2	
TR1.2	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Die Risiken im Kontext einer mangelnden Rückverfolgbarkeit und Dokumentation des KI-System müssen analysiert werden.	Analyse-Risiko Detaillierte Risikoanalyse einschließlich Quantifizierung von Eintrittswahrscheinlichkeiten und Risiken: – Identifikation der möglichen Risiken und ihrer Ursachen – Zuweisung der Risikoverantwortung – Schätzung der Eintrittswahrscheinlichkeit – Schätzung der Aufdeckungswahrscheinlichkeit – Schätzung der Auswirkung – Strukturierte Abstufung und Priorisierung der Risiken Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen: – Mangelndes Vertrauen von Nutzer – Unklare Rechenschaft – Verschleiern von Verzerrungen oder Sicherheitsschwachstellen – Nicht-Einhaltung regulatorischer Vorgaben – Probleme mit der Interoperabilität	Analyse-Risiko Limitierte Risikoanalyse mit Fokus auf den Schutzbedarf ohne Quantifizierung von Wahrscheinlichkeiten: – Identifikation der möglichen Risiken – Zuweisung der Risikoverantwortung – Schätzung der Auswirkung – Strukturierte Abstufung und Priorisierung der Risiken Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen: – Mangelndes Vertrauen von Nutzer*innen – Unklare Rechenschaft – Verschleiern von Verzerrungen oder Sicherheitsschwachstellen – Nicht-Einhaltung regulatorischer Vorgaben – Probleme mit der Interoperabilität	Analyse-Risiko Hauptsächlich qualitative Abschätzung ohne Wahrscheinlichkeiten: – Identifikation der möglichen Gefährdungen – Zuweisung der Risikoverantwortung – Qualitative Schätzung der Auswirkung – Qualitative Abstufung und Priorisierung der Risiken Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen: – Mangelndes Vertrauen von Nutzer*innen – Unklare Rechenschaft – Verschleiern von Verzerrungen oder Sicherheitsschwachstellen – Nicht-Einhaltung regulatorischer Vorgaben – Probleme mit der Interoperabilität	Es wurde keine Risikoanalyse durchgeführt.		System	Analyse	Maximalwert	TR1.1	
TR1.3		B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung	
Die Architektur des KI-Systems muss dokumentiert werden.	Tech. Maßnahmen – Sonstige Die Systemarchitektur ist dokumentiert, einschließlich: – KI-Komponenten (siehe TR1.4) – Hard- und Software-Einbettung und Anforderungen an diese – Schnittstellen z.B. für Nutzer*innen oder zu anderen Systemen – Informationsfluss zwischen einzelnen Bestandteilen des Systems – Begründung der Wahl der Architektur mit einer Beschreibung der Rolle der einzelnen Bestandteile im Kontext des Zwecks des KI-Systems – Ggf. Lizenz unter welcher das System verwendet werden darf – Vorgesehene Modalitäten des KI-Systems zur "eigenständigen" Anpassung/Weiterentwicklung während des Betriebs inklusive Online-Learning	Tech. Maßnahmen – Sonstige Die Systemarchitektur ist dokumentiert, einschließlich: – KI-Komponenten (siehe TR1.4) – Hard- und Software-Einbettung und Anforderungen an diese – Schnittstellen z.B. für Nutzer*innen oder zu anderen Systemen – Informationsfluss zwischen einzelnen Bestandteilen des Systems – Ggf. Lizenz unter welcher das System verwendet werden darf	Tech. Maßnahmen – Sonstige Die Systemarchitektur ist dokumentiert, einschließlich: – KI-Komponenten (siehe TR1.4) – Hard- und Software-Einbettung – Schnittstellen z.B. für Nutzer*innen oder zu anderen Systemen – Informationsfluss zwischen einzelnen Bestandteilen des Systems – Ggf. Lizenz unter welcher das System verwendet werden darf	Die Architektur des Systems ist nicht dokumentiert.		System	Analyse	Maximalwert	VE1.1, VE2.1, VE1.2, VE1.3, CY1.5, CY1.6	
TR1.4	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung	
Die Merkmale des/der KI-Modells/Modelle müssen dokumentiert werden.	Tech. Maßnahmen – Modell Eigenschaften werden dokumentiert, einschließlich: – Modellbezeichnung – Modellversion(-shistorie) inklusive Datum – Architekturbeschreibung und -diagramm des KI-Modells – Erwartete Eingabedaten – Erwartete Ausgabedaten – Verwendete Trainings- & Testdaten – Erwartete Leistungsfähigkeit – Durchgeführte Tests, ermittelte Testergebnisse und abgeleitete Schlussfolgerungen Die Wahl der Architektur und des Designs des KI-Systems und/oder -Modells muss begründet werden inklusive der Vorteile des Modells und Abwägungen in Bezug auf mögliche Zielkonflikte sind beschrieben.	Tech. Maßnahmen – Modell Eigenschaften werden dokumentiert, einschließlich: – Modellbezeichnung – Modellversion(-shistorie) inklusive Datum – Architekturbeschreibung und -diagramm des KI-Modells – Erwartete Eingabedaten – Erwartete Ausgabedaten – Verwendete Trainings- & Testdaten – Erwartete Leistungsfähigkeit	Tech. Maßnahmen – Modell Eigenschaften werden dokumentiert, einschließlich: – Modellbezeichnung – Aktuellste Modellversion mit Datum – Grobe Architekturbeschreibung des KI-Modells – Erwartete Eingabedaten – Erwartete Ausgabedaten – Verwendete Trainings- & Testdaten – Erwartete Leistungsfähigkeit	Es ist keine Dokumentation des/der KI-Modells/Modelle vorhanden.	Die Architekturbeschreibung des KI-Modells kann z.B. Typ des Modells, Art und Anzahl der Ebenen/ Schichten bei neuronalen Netzen sowie auch Funktionen zur Aktivierung oder Belohnung umfassen.	Komponente	Analyse	Maximalwert	VE1.2, VE1.3, CY1.5, CY1.6	
TR1.5	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung	
Die Merkmale der verwendeten Datensätze müssen dokumentiert werden.		Siehe: DA1.1					Komponente	Analyse	Maximalwert	
TR1.6	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung	
Der Entwurfs- und Entwicklungsprozess des KI-Systems muss beschrieben werden.	Orga. Maßnahme – Systemnahe Prozesse Die Dokumentation zum Entwurfs- und Entwicklungsprozess umfasst: – Übersicht der verantwortlichen Person(en) und deren Aufgaben – Rückverfolgbarkeit und systematische Vorgehensweise einzelner Arbeitselemente (z.B. durch Projektmanagement-Tool) – Ggfs. Beschreibung und Begründung der Anpassungen der Architektur des KI-Systems im Laufe der Entwicklung Tech. Maßnahmen – Daten Die Dokumentation zum Entwurfs- und Entwicklungsprozess umfasst: – Verwendete Daten und ihre Herkunft – Relevante Datenvorbereitungsprozesse (z.B. Bereinigung, Annotation, Kennzeichnung, Anreicherung, Aggregation, Feature Engineering) – Rückverfolgung der Data-Lineage und zur möglichen Wiederherstellung der Daten Tech. Maßnahmen – Modell Die Dokumentation zum Entwurfs- und Entwicklungsprozess umfasst: – Ein System zur Versionsverwaltung und Nachverfolgung von Änderungen an den KI-Modellen und Aufzeichnung der Trainings und der jeweils eingesetzten Daten – Beschreibung der Änderungen an KI-Modellen im Laufe des Entwicklungsprozesses	Orga. Maßnahme – Systemnahe Prozesse Die Dokumentation zum Entwurfs- und Entwicklungsprozess umfasst: – Übersicht der verantwortlichen Person(en) und deren Aufgaben – Rückverfolgbarkeit und systematische Vorgehensweise einzelner Arbeitselemente (z.B. durch Projektmanagement-Tool) Tech. Maßnahmen – Daten Die Dokumentation zum Entwurfs- und Entwicklungsprozess umfasst: – Verwendete Daten und ihre Herkunft – Relevante Datenvorbereitungsprozesse (z.B. Bereinigung, Annotation, Kennzeichnung, Anreicherung, Aggregation, Feature Engineering) – Rückverfolgung der Data-Lineage Tech. Maßnahmen – Modell Die Dokumentation zum Entwurfs- und Entwicklungsprozess umfasst: – Ein System zur Versionsverwaltung und Nachverfolgung von Änderungen an den KI-Modellen und Aufzeichnung der Trainings und der jeweils eingesetzten Daten	Orga. Maßnahme – Systemnahe Prozesse Die Dokumentation zum Entwurfs- und Entwicklungsprozess umfasst: – Übersicht der verantwortlichen Person(en) und deren Aufgaben – Rückverfolgbarkeit und systematische Vorgehensweise einzelner Arbeitselemente (z.B. durch Projektmanagement-Tool) Tech. Maßnahmen – Daten Die Dokumentation zum Entwurfs- und Entwicklungsprozess umfasst: – Verwendete Daten und ihre Herkunft – Rückverfolgung der Data-Lineage Tech. Maßnahmen – Modell Die Dokumentation zum Entwurfs- und Entwicklungsprozess umfasst: – Ein System zur Versionsverwaltung und Nachverfolgung von Änderungen an den KI-Modellen und Aufzeichnung der Trainings und der jeweils eingesetzten Daten	Der Entwurfs- und Entwicklungsprozess wurde nicht dokumentiert.	Für die Dokumentation von Datensätzen und Modell können z.B. geläufige Formate, wie Data Sheets bzw. Model Cards verwendet werden.	System/Komponente	Analyse	Normal	VE1.4, VE1.5	

Transparenz										
TR1.7	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung	
Das KI-System muss Funktionen zur Überwachung, Erfassung und Aufzeichnung seines Verhaltens enthalten.	<p>Orga. Maßnahmen – Systemnahe Prozesse</p> <p>Es muss ein zweckorientiertes Konzept zur Protokollierung ("Logging") des KI-Systems sowie der darin verwendeten Daten während der Entwicklung und des Betriebs erstellt werden. Das Konzept umfasst mindestens die folgenden Informationen:</p> <ul style="list-style-type: none">– Definition der zu überwachenden Daten und Metriken– Definition einer dem Zweck entsprechenden Aufbewahrungsfrist der Protokolle– Definition einer dem Zweck entsprechenden Aufzeichnungshäufigkeit– Definition der Aufzeichnungsstruktur– Beschreibung zur Schnittstelle, über die die Protokollierung angesprochen werden kann <p>Tech. Maßnahmen – Betrieb</p> <p>Es ist eine Schnittstelle vorhanden, die es im Betrieb erlaubt mindestens die folgenden Informationen zu erfassen und zu überwachen:</p> <ul style="list-style-type: none">– Falls technisch möglich und sinnvoll im Anwendungskontext, Nutzerinteraktionen und Modellanfragen einschließlich der für die Anfrage verwendeten Modellversion, Eingaben und Ausgaben über einen definierten Zeitraum– Fehlfunktionen (z.B. bei der Verarbeitung von automatischen oder manuellen Aktionen)– Zugriffe auf Daten, Dienste oder Funktionen durch die Nutzer– Änderungen an sicherheitsrelevanten Konfigurationsparametern (z.B. Fehlerbehandlung und Protokollierungsmechanismen, Benutzer-authentifizierung, Aktionsautorisierung, Kryptographie und Kommunikationssicherheit)– Verletzung der Funktionalitäts-, Schutz- oder Qualitätsziele des KI-Systems <p>Sind bestimmte Informationen technisch nicht sinnvoll zu protokollieren, ist dies zu begründen.</p> <p>Die zu jeweils zu erfassenden Metadaten erfassen im Rahmen der technischen Möglichkeiten mindestens Informationen über:</p> <ul style="list-style-type: none">– Art, Zeitpunkt, Dauer, Speicherort und Akteur/System von aufgezeichneten Ereignissen oder Aktionen	<p>Orga. Maßnahmen – Systemnahe Prozesse</p> <p>Es muss ein zweckorientiertes Konzept zur Protokollierung ("Logging") des KI-Systems sowie der darin verwendeten Daten während der Entwicklung und des Betriebs erstellt werden. Das Konzept umfasst mindestens die folgenden Informationen:</p> <ul style="list-style-type: none">– Definition der zu überwachenden Daten und Metriken– Definition einer dem Zweck entsprechenden Aufbewahrungsfrist der Protokolle– Definition einer dem Zweck entsprechenden Aufzeichnungshäufigkeit– Definition der Aufzeichnungsstruktur– Beschreibung zur Schnittstelle, über die die Protokollierung angesprochen werden kann <p>Tech. Maßnahmen – Betrieb</p> <p>Es ist eine Schnittstelle vorhanden, die es im Betrieb erlaubt mindestens die folgenden Informationen zu erfassen und zu überwachen:</p> <ul style="list-style-type: none">– Falls technisch möglich und sinnvoll im Anwendungskontext, Nutzerinteraktionen und Modellanfragen einschließlich der für die Anfrage verwendeten, Modellversion, Eingaben und Ausgaben– Fehlfunktionen (z.B. bei der Verarbeitung von automatischen oder manuellen Aktionen) <p>Sind bestimmte Informationen technisch nicht sinnvoll zu protokollieren, ist dies zu begründen.</p> <p>Die zu jeweils zu erfassenden Metadaten erfassen im Rahmen der technischen Möglichkeiten mindestens Informationen über:</p> <ul style="list-style-type: none">– Art, Zeitpunkt, Dauer, Speicherort und Akteur/System von aufgezeichneten Ereignissen oder Aktionen	<p>Orga. Maßnahmen – Systemnahe Prozesse</p> <p>Es muss ein zweckorientiertes Konzept zur Protokollierung ("Logging") des KI-Systems sowie der darin verwendeten Daten während der Entwicklung und des Betriebs erstellt werden. Das Konzept umfasst mindestens die folgenden Informationen:</p> <ul style="list-style-type: none">– Definition der zu überwachenden Daten und Metriken– Falls technisch möglich und sinnvoll im Anwendungskontext, Nutzerinteraktionen und Modellanfragen einschließlich der für die Anfrage verwendeten Modellversion, Eingaben und Ausgaben– Fehlfunktionen (z.B. bei der Verarbeitung von automatischen oder manuellen Aktionen) <p>Sind bestimmte Informationen technisch nicht sinnvoll zu protokollieren, ist dies zu begründen.</p> <p>Die zu jeweils zu erfassenden Metadaten erfassen im Rahmen der technischen Möglichkeiten mindestens Informationen über:</p> <ul style="list-style-type: none">– Art, Zeitpunkt, Dauer, Speicherort und Akteur/System von aufgezeichneten Ereignissen oder Aktionen	Eine Aufzeichnung wichtiger Daten zum System ist nicht vorbereitet.		System/Komponente	Maßnahme	Normal	VE1.6, DA2.6	
TR1.8	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung	
Es muss bewertet werden, ob die ergriffenen Maßnahmen die festgestellten Risiken auf ein annehmbares Maß vermindert haben und die Qualität des KI-Systems den gesetzten Zielvorgaben entspricht	<p>Bewertung</p> <ul style="list-style-type: none">– Zusammenfassende Betrachtung der Effekte der Durchführung der technischen und organisatorischen Maßnahmen– Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen– Ermittlung und Beschreibung des Restrisikos nach Durchführung der technischen und organisatorischen Maßnahmen– Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist– Begründung der Tolerierbarkeit des Restrisikos	<p>Bewertung</p> <ul style="list-style-type: none">– Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen– Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist– Begründung der Tolerierbarkeit des Restrisikos	<p>Bewertung</p> <ul style="list-style-type: none">– Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist– Begründung der Tolerierbarkeit des Restrisikos	Eine Bewertung wurde nicht durchgeführt.		System	Bewertung	Maximalwert		
TR2	Erklärbarkeit & Interpretierbarkeit									
TR2.1	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Die Risiken im Kontext einer mangelnden Interpretierbarkeit oder Erklärbarkeit des KI-Systems müssen analysiert werden.	<p>Analyse-Risiko</p> <p>Detaillierte Risikoanalyse zur Erklärbarkeit einschließlich Quantifizierung von Eintrittswahrscheinlichkeiten und Risiken:</p> <ul style="list-style-type: none">– Identifikation der möglichen Risiken und ihrer Ursachen– Zuweisung der Risikoverantwortung– Schätzung der Eintrittswahrscheinlichkeit– Schätzung der Aufdeckungswahrscheinlichkeit– Schätzung der Auswirkung– Strukturierte Abstufung und Priorisierung der Risiken <p>Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen:</p> <ul style="list-style-type: none">– Mangelndes Vertrauen von Nutzer*innen– Mangelndes Verständnis der Ausgaben– Fehlerhafte Entscheidungen aufgrund eines unzureichenden Verständnisses der Ausgaben	<p>Analyse-Risiko</p> <p>Limitierte Risikoanalyse zur Erklärbarkeit mit Fokus auf den Schutzbedarf ohne Quantifizierung von Wahrscheinlichkeiten:</p> <ul style="list-style-type: none">– Identifikation der möglichen Risiken– Zuweisung der Risikoverantwortung– Schätzung der Auswirkung– Strukturierte Abstufung und Priorisierung der Risiken <p>Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen:</p> <ul style="list-style-type: none">– Mangelndes Vertrauen von Nutzer*innen– Mangelndes Verständnis der Ausgaben– Fehlerhafte Entscheidungen aufgrund eines unzureichenden Verständnisses der Ausgaben	<p>Analyse-Risiko</p> <p>Hauptsächlich qualitative Abschätzung ohne Wahrscheinlichkeiten zur Erklärbarkeit:</p> <ul style="list-style-type: none">– Identifikation der möglichen Gefährdungen– Zuweisung der Risikoverantwortung– Qualitative Schätzung der Auswirkung– Qualitative Abstufung und Priorisierung der Risiken <p>Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen:</p> <ul style="list-style-type: none">– Mangelndes Vertrauen von Nutzer*innen– Mangelndes Verständnis der Ausgaben– Fehlerhafte Entscheidungen aufgrund eines unzureichenden Verständnisses der Ausgaben	Es wurde keine Risikoanalyse durchgeführt.		System	Analyse	Maximalwert	TR1.1	
TR2.2	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung	
Die Interpretierbarkeit und Erklärbarkeit des KI-Systems muss in Hinblick auf die Nutzergruppen, den Zweck und die Risiken analysiert werden.	<p>Tech. Maßnahme – Tests</p> <p>Der Grad, zu dem das KI-System interpretierbar oder erklärbar für die erlaubten Nutzergruppen ist, muss unter Beachtung der Risikoanalyse (TR2.1) und Zweckbestimmung (TR1.1) analysiert werden.</p> <p>Hierzu muss der Zusammenhang zwischen Eingaben und Ausgaben der KI-Komponenten des Systems anhand der inhärenten Merkmale der Komponente oder geeigneter technischer Tests untersucht werden.</p> <p>Die ausgewählten Tests sollten mindestens die folgenden Merkmale aufweisen:</p> <ul style="list-style-type: none">– Wenn durch die Auswahl begründet, sollten die Metriken und Tests sowohl wichtige Basismethoden als auch fortgeschrittene Methoden (hinsichtlich Komplexität, Informationsgehalt, Implementierungsaufwand, z.B. umfangreiches Prüfwerkzeug) umfassen– ggf. Festlegung von Schwellenwerten (als Mindestanforderungen an die Erklärbarkeit) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks– falls anwendbar: Abstufung der Metriken und Schwellenwerte nach unterschiedlichen Einsatzszenarien mit Bezug zur Definition des Anwendungsbereichs.	<p>Tech. Maßnahme – Tests</p> <p>Der Grad, zu dem das KI-System interpretierbar oder erklärbar für die erlaubten Nutzergruppen ist, muss unter Beachtung der Risikoanalyse (TR2.1) und Zweckbestimmung (TR1.1) analysiert werden.</p> <p>Hierzu muss der Zusammenhang zwischen Eingaben und Ausgaben der KI-Komponenten des Systems anhand der inhärenten Merkmale der Komponente oder geeigneter technischer Tests untersucht werden.</p> <p>Die ausgewählten Tests sollten mindestens die folgenden Merkmale aufweisen:</p> <ul style="list-style-type: none">– Wenn durch die Auswahl begründet, sollten die Metriken und Tests sowohl wichtige Basismethoden als auch fortgeschrittene Methoden (hinsichtlich Komplexität, Informationsgehalt, Implementierungsaufwand, z.B. umfangreiches Prüfwerkzeug) umfassen	<p>Tech. Maßnahme – Tests</p> <p>Der Grad, zu dem das KI-System interpretierbar oder erklärbar für die erlaubten Nutzergruppen ist, muss unter Beachtung der Risikoanalyse (TR2.1) und Zweckbestimmung (TR1.1) analysiert werden.</p> <p>Hierzu muss der Zusammenhang zwischen Eingaben und Ausgaben der KI-Komponenten des Systems anhand der inhärenten Merkmale der Komponente oder geeigneter technischer Tests untersucht werden.</p> <ul style="list-style-type: none">– Es genügen Basismethoden (hinsichtlich Komplexität, Informationsgehalt, Implementierungsaufwand, z.B. eine simple Metrik)	Die Interpretierbarkeit und Erklärbarkeit des KI-Systems wurden nicht analysiert.		System/Komponente	Analyse	Normal		

Transparenz									
TR2.3	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung
Es müssen Maßnahmen ergriffen werden, um das KI-System im Hinblick auf seinen Verwendungszweck interpretierbar oder angemessen erklärbar zu machen.	<p>Tech. Maßnahmen – Modell Das KI-System muss im Kontext seines Verwendungszweckes und mit dem Ziel der Risikominimierung angemessen erklärbar gemacht werden. Das kann beinhalten:</p> <p>i) die Wahl von inhärent interpretierbaren Algorithmen oder</p> <p>ii) die Wahl von Algorithmen zu denen entsprechende Ansätze zur Herstellung der (lokalen) Erklärbarkeit einzelner Ausgaben existieren und deren anschließende Implementierung</p> <p>Tech. Maßnahmen – Benutzerinstruktionen Das KI-System muss für den nichtfachkundigen Nutzer die notwendigen Informationen zur Nachvollziehung von Modellausgaben liefern. Hierzu ist eine Benutzeroberfläche verfügbar, die folgende Aspekte enthält:</p> <p>– Beschreibung der verwendeten Methoden um Systemausgaben verständlich zu machen</p> <p>– Erläuterung der erzeugten Erklärung</p> <p>– Anforderungen an die Nutzer (z.B. informatische Kenntnisse), um diese prinzipiell verstehen zu können.</p> <p>Orga. Maßnahmen – Schulungen Zur Qualifikation und Schulung des Personals, welches mit dem Betrieb und der Aufsicht des KI-Systems betraut ist bzw. menschlicher Endnutzer, sind zielgruppenorientierte Schulungen vorbereitet worden. Diese Schulungen legen einen Fokus auf die Interpretation von Ergebnissen des KI-Systems und möglicher Funktionen, die die Erklärbarkeit der Ergebnisse unterstützen. Dabei sind mindestens folgende Aspekte abgedeckt:</p> <p>– Relevanz und Nutzen der Erklärbarkeit des KI-Systems im Einsatzkontext</p> <p>– Eingesetzte Methoden und Werkzeuge zur Erklärbarkeit</p> <p>– Übung der Anwendung der Methoden und Werkzeuge zur Erklärung</p>	<p>Tech. Maßnahmen – Modell Das KI-System muss im Kontext seines Verwendungszweckes und mit dem Ziel der Risikominimierung angemessen erklärbar gemacht werden. Das kann beinhalten:</p> <p>i) die Wahl von inhärent interpretierbaren Algorithmen oder</p> <p>ii) die Wahl von Algorithmen zu denen entsprechende Ansätze zur Herstellung der (lokalen) Erklärbarkeit einzelner Ausgaben existieren und deren anschließende Implementierung</p> <p>Orga. Maßnahmen – Benutzerinstruktionen Das KI-System muss für den nichtfachkundigen Nutzer die notwendigen Informationen zur Nachvollziehung von Modellausgaben liefern. Hierzu sind folgende Informationen zu liefern:</p> <p>– Beschreibung der verwendeten Methoden um Systemausgaben verständlich zu machen</p> <p>– Erläuterung der erzeugten Erklärung</p> <p>– Anforderungen an die Nutzer (z.B. informatische Kenntnisse), um diese prinzipiell verstehen zu können.</p> <p>Benutzerinstruktionen können inhaltlich auch gänzlich über Schulungen abgedeckt werden.</p>	<p>Tech. Maßnahmen – Modell Das KI-System muss im Kontext seines Verwendungszweckes und mit dem Ziel der Risikominimierung angemessen erklärbar gemacht werden. Das kann beinhalten:</p> <p>i) die Wahl von inhärent interpretierbaren Algorithmen oder</p> <p>ii) die Wahl von Algorithmen zu denen entsprechende Ansätze zur Herstellung der (lokalen) Erklärbarkeit einzelner Ausgaben existieren und deren anschließende Implementierung</p> <p>Orga. Maßnahmen – Benutzerinstruktionen Das KI-System muss für den nichtfachkundigen Nutzer die notwendigen Informationen zur Nachvollziehung von Modellausgaben liefern. Hierzu sind folgende Informationen zu liefern:</p> <p>– Beschreibung der verwendeten Methoden um Systemausgaben verständlich zu machen</p> <p>– Anforderungen an die Nutzer (z.B. informatische Kenntnisse), um diese prinzipiell verstehen zu können.</p> <p>Benutzerinstruktionen können inhaltlich auch gänzlich über Schulungen abgedeckt werden.</p>	Es wurden keine angemessenen Maßnahmen zum Herstellen von Interpretierbarkeit oder Erklärbarkeit vorgenommen.		System/Komponente	Maßnahme	Normal	
TR2.4	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung
Es muss bewertet werden, ob die ergriffenen Maßnahmen die festgestellten Risiken auf ein annehmbares Maß vermindert haben und die Qualität des KI-Systems den gesetzten Zielvorgaben entspricht.	<p>Bewertung</p> <p>– Zusammenfassende Betrachtung der Effekte der Durchführung der technischen und organisatorischen Maßnahmen</p> <p>– Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen</p> <p>– Ermittlung und Beschreibung des Restrisikos nach Durchführung der technischen und organisatorischen Maßnahmen</p> <p>– Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist</p> <p>– Begründung der Tolerierbarkeit des Restrisikos</p>	<p>Bewertung</p> <p>– Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen</p> <p>– Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist</p> <p>– Begründung der Tolerierbarkeit des Restrisikos</p>	<p>Bewertung</p> <p>– Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist</p> <p>– Begründung der Tolerierbarkeit des Restrisikos</p>	Eine Bewertung wurde nicht durchgeführt.		System	Bewertung	Maximalwert	

Menschliche Aufsicht und Kontrolle										
MA1	Menschliche Handlungsfähigkeit									
MA1.1	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Der Grad der Autonomie des KI-Systems muss analysiert und dokumentiert werden.	Analyse – Definition <ul style="list-style-type: none"> – Feststellung der Autonomiestufe bzw. des Grads an menschlicher Kontrolle und Einbindung in Entscheidungsprozesse – Begründete Abgrenzung von anderen Autonomiestufen – Ableitung von spezifischen gesetzlichen Anforderungen in Bezug auf menschliche Aufsicht und Kontrolle 	Analyse – Definition <ul style="list-style-type: none"> – Feststellung der Autonomiestufe bzw. des Grads an menschlicher Kontrolle und Einbindung in Entscheidungsprozesse – Ableitung von spezifischen gesetzlichen Anforderungen in Bezug auf menschliche Aufsicht und Kontrolle 	Analyse – Definition <ul style="list-style-type: none"> – Feststellung der Autonomiestufe bzw. des Grads an menschlicher Kontrolle und Einbindung in Entscheidungsprozesse – Ableitung von spezifischen gesetzlichen Anforderungen in Bezug auf menschliche Aufsicht und Kontrolle 	Es wurde keine Einordnung der Autonomiestufe vorgenommen.	Autonomiestufen werden in der ISO/IEC 22989 vorgestellt	System	Analyse	Maximalwert	MA2.1	
MA1.2	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Die Risiken im Kontext der Einschränkung der menschlichen Handlungsfähigkeit durch das KI-System müssen analysiert werden.	Analyse-Risiko <p>Detaillierte Risikoanalyse einschließlich Quantifizierung von Eintrittswahrscheinlichkeiten und Risiken:</p> <ul style="list-style-type: none"> – Identifikation der möglichen Risiken und ihrer Ursachen – Zuweisung der Risikoverantwortung – Schätzung der Eintrittswahrscheinlichkeit – Schätzung der Aufdeckungswahrscheinlichkeit – Strukturierte Abstufung und Priorisierung der Risiken <p>Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen:</p> <ul style="list-style-type: none"> – Ethische Dilemmata / Entscheidungsfindung – Unklare Rechenschaft – Automation Bias / übermäßige Abhängigkeit – Vortäuschung menschlicher Züge oder unklare Urheberschaft – Mangelnde Nachvollziehbarkeit von Entscheidungen 	Analyse-Risiko <p>Limitierte Risikoanalyse mit Fokus auf den Schutzbedarf ohne Quantifizierung von Wahrscheinlichkeiten:</p> <ul style="list-style-type: none"> – Identifikation der möglichen Risiken – Zuweisung der Risikoverantwortung – Schätzung der Auswirkung – Strukturierte Abstufung und Priorisierung der Risiken <p>Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen:</p> <ul style="list-style-type: none"> – Ethische Dilemmata / Entscheidungsfindung – Unklare Rechenschaft – Automation Bias – Vortäuschung menschlicher Züge oder unklare Urheberschaft – Mangelnde Nachvollziehbarkeit von Entscheidungen 	Analyse-Risiko <p>Hauptsächlich qualitative Abschätzung ohne Wahrscheinlichkeiten:</p> <ul style="list-style-type: none"> – Identifikation der möglichen Gefährdungen – Zuweisung der Risikoverantwortung – Qualitative Schätzung der Auswirkung – Qualitative Abstufung und Priorisierung der Risiken <p>Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen:</p> <ul style="list-style-type: none"> – Ethische Dilemmata / Entscheidungsfindung – Unklare Rechenschaft – Automation Bias – Vortäuschung menschlicher Züge oder unklare Urheberschaft – Mangelnde Nachvollziehbarkeit von Entscheidungen 	Es wurde keine Risikoanalyse durchgeführt.		System	Analyse	Maximalwert	TR1.1	
MA1.3	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Es muss für betroffene Personen und Nutzer möglich sein, ihre mit personenbezogenen Daten einhergehenden Rechte zu Datenverwaltung, -löschung und -benutzung sowie Informationspflichten auch im Betrieb des KI-Systems wahrzunehmen.		DA2.6				System	Maßnahme	Normal	TR1.7, MA1.4, MA1.6, MA2.3	
MA1.4	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Es muss möglich sein, Feedback zum KI-System zu geben, Rückfragen zu stellen und Probleme zu melden.	Tech. Maßnahmen – Betrieb <ul style="list-style-type: none"> – Es gibt mindestens einen Kanal durch den mit dem Anbieter zum Zweck des Feedbacks in Kontakt getreten werden kann – Kanäle sind durch jeden mit begründetem Interesse klar im Kontext des KI-Systems identifizierbar – Es ist erkennbar, dass die möglichen Kanäle zur Abgabe von Feedback vorgesehen sind Orga. Maßnahmen – Systemnahe Prozesse <ul style="list-style-type: none"> – Sichtung und Überprüfung des Feedbacks und die Vergabe der Verantwortung hierfür erfolgt – Garantierte und individualisierte Beantwortung von Anfragen – Prozess zur Weiterleitung und Einarbeitung des Feedbacks in die Weiterentwicklung des KI-Systems 	Tech. Maßnahmen – Betrieb <ul style="list-style-type: none"> – Es gibt mindestens einen Kanal durch den mit dem Anbieter zum Zweck des Feedbacks in Kontakt getreten werden kann – Kanäle sind durch jeden mit begründetem Interesse klar im Kontext des KI-Systems identifizierbar Orga. Maßnahmen – Systemnahe Prozesse <ul style="list-style-type: none"> – Sichtung und Überprüfung des Feedbacks und die Vergabe der Verantwortung hierfür erfolgt 	Tech. Maßnahmen – Betrieb <ul style="list-style-type: none"> – Die Kontaktinformationen sind dem KI-System beigelegt und können zum Zweck des Feedbacks genutzt werden 	Es sind keine Kanäle vorhanden, um Feedback zum KI-System zu geben.	betroffene Personen oder Betreiber sein.	System	Maßnahme	Normal	DA2.6	
MA1.5		B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Betroffene Personen und Nutzer müssen die Möglichkeit haben, sich über die Anwendung des KI-Systems, die Art und Weise, wie das KI-System eine Entscheidung unterstützt, und die Interpretation seiner Ausgaben zu informieren.	Orga. Maßnahmen – Benutzerinstruktion <p>Bereitstellung von adressatengerecht aufbereiteten Instruktionen und Informationen mit mindestens den folgenden Inhalten:</p> <ul style="list-style-type: none"> – Kontaktinformationen – Zweck des KI-Systems – Vorgesehener Anwendungsbereich des KI-Systems – Erwartbare Leistung und Funktionalität des KI-Systems – Bekannte Risiken und Implikationen für die Nutzung des KI-Systems (Risiken durch Nutzung und durch Ergebnisse) – Zusammenfassung der erwarteten Eingabedaten – Beschreibungen zur Interpretation der Ausgaben des KI-Systems – Beschreibung des Grads der Autonomie und der Entscheidungsprozesse des KI-Systems – Notwendige Maßnahmen zur menschlichen Aufsicht über das System – Anforderungen an die Nutzer (z.B. Schulung) für die Verwendung des KI-Systems – Notwendige Instandhaltungsmaßnahmen (inklusive Softwareupdates) Tech. Maßnahmen – Betrieb <ul style="list-style-type: none"> – Integration von Benachrichtigungen an natürliche Personen, die mit dem KI-System interagieren darüber, dass die Ergebnisse und potenzielle Entscheidungsfindung auf dem KI-System basieren – Bereitstellung von Informationen zu Ausgaben oder getroffenen Entscheidungen des KI-Systems, die eine korrekte Interpretation ermöglichen – Warnung vor sogenanntem "Automation Bias" und möglichen Folgen 	Orga. Maßnahmen – Benutzerinstruktion <p>Bereitstellung von adressatengerecht aufbereiteten Instruktionen und Informationen mit mindestens den folgenden Inhalten:</p> <ul style="list-style-type: none"> – Kontaktinformationen – Zweck des KI-Systems – Vorgesehener Anwendungsbereich des KI-Systems – Erwartbare Leistung und Funktionalität des KI-Systems – Beschreibung über die Interpretation der Ausgaben des KI-Systems – Notwendige Maßnahmen zur menschlichen Aufsicht über das System – Notwendige Instandhaltungsmaßnahmen (inklusive Softwareupdates) Tech. Maßnahmen – Betrieb <ul style="list-style-type: none"> – Integration von Benachrichtigungen an natürliche Personen, die mit dem KI-System interagiert darüber, dass die Ergebnisse und potenzielle Entscheidungsfindung auf dem KI-System basieren – Warnung vor sogenanntem "Automation Bias" und möglichen Folgen 	Orga. Maßnahmen – Benutzerinstruktion <p>Bereitstellung von Instruktionen und Informationen mit mindestens den folgenden Inhalten:</p> <ul style="list-style-type: none"> – Kontaktinformationen – Zweck des KI-Systems – Vorgesehener Anwendungsbereich des KI-Systems – Erwartbare Leistung und Funktionalität des KI-Systems – Beschreibung über die Interpretation der Ausgaben des KI-Systems – Notwendige Maßnahmen zur menschlichen Aufsicht über das System Tech. Maßnahmen – Betrieb <ul style="list-style-type: none"> – Integration von Benachrichtigungen an natürliche Personen, die mit dem KI-System interagiert darüber, dass die Ergebnisse und potenzielle Entscheidungsfindung auf dem KI-System basieren – Warnung vor sogenanntem "Automation Bias" und möglichen Folgen 	Es werden keine Hinweise oder Benachrichtigungen an Nutzer oder betroffene Personen ausgespielt	Die Bereitstellung der Benachrichtigungen und Instruktionen kann über eine grafische Schnittstelle erfolgen.	System	Maßnahme	Normal	TR2, MA1.6	
MA1.6	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Es müssen Mechanismen vorhanden sein, die es Nutzern oder betroffenen Personen ermöglichen, die Entscheidungen und Ausgaben des KI-Systems zu verwerfen, anzufechten, zu korrigieren oder zu unterbrechen	Tech. Maßnahmen – Betrieb <p>Die entsprechenden Mechanismen zum Anfechten, Korrigieren oder Unterbrechen des KI-Systems müssen auf die identifizierten Risiken, den Zweck des KI-Systems und die Adressaten der Maßnahme abgestimmt werden.</p> <p>Maßnahmen, die in Betracht gezogen werden, umfassen zumindest:</p> <ul style="list-style-type: none"> – Widerspruchsförmular über das Anfechtungen von Entscheidungen eingereicht werden können – Weiterleitung zu einer menschlichen Überprüfung, wo durch eine qualifizierte und autorisierte natürliche Person eine Entscheidung bestätigt oder eine Ausgabe verarbeitet werden muss – Not-Aus-Funktion welche das System durch eine autorisierte natürliche Person in einen sicheren Zustand versetzt – Interaktive grafische Nutzeroberfläche zur Abbildung der Aktionen 	Tech. Maßnahmen – Betrieb <p>Die entsprechenden Mechanismen zum Anfechten, Korrigieren oder Unterbrechen des KI-Systems müssen auf die identifizierten Risiken, den Zweck des KI-Systems und die Adressaten der Maßnahme abgestimmt werden.</p> <p>Maßnahmen, die in Betracht gezogen werden, umfassen zumindest:</p> <ul style="list-style-type: none"> – Widerspruchsförmular über das Anfechtungen von Entscheidungen eingereicht werden können – Weiterleitung zu einer menschlichen Überprüfung, wo durch eine qualifizierte und autorisierte natürliche Person eine Entscheidung bestätigt oder eine Ausgabe verarbeitet werden muss – Not-Aus-Funktion welche das System durch eine autorisierte natürliche Person in einen sicheren Zustand versetzt – Interaktive grafische Nutzeroberfläche zur Abbildung der Aktionen 	Tech. Maßnahmen – Betrieb <p>Maßnahmen, die in Betracht gezogen werden, umfassen zumindest:</p> <ul style="list-style-type: none"> – Widerspruchsförmular über das Anfechtungen von Entscheidungen eingereicht werden können – Weiterleitung zu einer menschlichen Überprüfung, wo durch eine qualifizierte und autorisierte Person eine Entscheidung bestätigt oder eine Ausgabe verarbeitet werden muss – Not-Aus-Funktion welche das System durch eine natürliche Person in einen sicheren Zustand versetzt – Interaktive grafische Nutzeroberfläche zur Abbildung der Aktionen 	Es sind keine Mechanismen zum Anfechten, Korrigieren oder Unterbrechen des KI-Systems vorhanden.	Sowohl human-in-the-loop (HITL) als auch human-on-the-loop (HOTL) Konstruktionen sind möglich	System	Maßnahme	Maximalwert	MA1.4, VE2.2	

Menschliche Aufsicht und Kontrolle									
MA1.7	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Es muss bewertet werden, ob die ergriffenen Maßnahmen die festgestellten Risiken auf ein annehmbares Maß vermindert haben und die Qualität des KI-Systems den gesetzten Zielvorgaben entspricht.	Bewertung – Zusammenfassende Betrachtung der Effekte der Durchführung der technischen und organisatorischen Maßnahmen – Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen – Ermittlung und Beschreibung des Restrisikos nach Durchführung der technischen und organisatorischen Maßnahmen – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Bewertung – Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Bewertung – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Eine Bewertung wurde nicht durchgeführt.		System	Bewertung	Maximalwert	

Menschliche Aufsicht und Kontrolle										
MA2	Menschliche Aufsicht									
MA2.1	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Der Grad der Autonomie des KI-Systems muss analysiert und dokumentiert werden.	Siehe: MA1.1					System	Analyse	Maximalwert	MA1.1	
MA2.2	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Die Risiken, dass das KI-System ohne menschliche Aufsicht (und Kontrolle) zu Schäden führt, müssen analysiert werden.	<p>Analyse-Risiko</p> <p>Detaillierte Risikoanalyse einschließlich Quantifizierung von Eintrittswahrscheinlichkeiten und Risiken:</p> <ul style="list-style-type: none">– Identifikation der möglichen Risiken und ihrer Ursachen– Zuweisung der Risikoverantwortung– Schätzung der Eintrittswahrscheinlichkeit– Schätzung der Aufdeckungswahrscheinlichkeit– Schätzung der Auswirkung– Strukturierte Abstufung und Priorisierung der Risiken <p>Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen:</p> <ul style="list-style-type: none">– Implikationen (Teil-)autonomer Entscheidungen des Systems ohne Kontrolle, Eingriff oder Korrektur durch natürliche Personen– Ethische Dilemmata / Entscheidungsfindung– Mangelnde Nachvollziehbarkeit von Entscheidungen– Haftung für Entscheidungen des KI-Systems– Urheberschaft für Ausgaben des KI-Systems– Fehleranfälligkeit des KI-Systems ohne das dies durch Menschen erkannt wird– Fehlerhafte/unsachgemäße Nutzung (z.B. durch mangelnde Qualifikation)	<p>Analyse-Risiko</p> <p>Limitierte Risikoanalyse mit Fokus auf den Schutzbedarf ohne Quantifizierung von Wahrscheinlichkeiten:</p> <ul style="list-style-type: none">– Identifikation der möglichen Risiken– Zuweisung der Risikoverantwortung– Schätzung der Auswirkung– Strukturierte Abstufung und Priorisierung der Risiken <p>Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen:</p> <ul style="list-style-type: none">– Implikationen (Teil-)autonomer Entscheidungen des Systems ohne Kontrolle, Eingriff oder Korrektur durch natürliche Personen– Ethische Dilemmata / Entscheidungsfindung– Mangelnde Nachvollziehbarkeit von Entscheidungen– Haftung für Entscheidungen des KI-Systems– Urheberschaft für Ausgaben des KI-Systems– Fehleranfälligkeit des KI-Systems ohne das dies durch Menschen erkannt wird– Fehlerhafte/unsachgemäße Nutzung (z.B. durch mangelnde Qualifikation)	<p>Analyse-Risiko</p> <p>Hauptsächlich qualitative Abschätzung ohne Wahrscheinlichkeiten:</p> <ul style="list-style-type: none">– Identifikation der möglichen Gefährdungen– Zuweisung der Risikoverantwortung– Qualitative Schätzung der Auswirkung– Qualitative Abstufung und Priorisierung der Risiken <p>Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen:</p> <ul style="list-style-type: none">– Implikationen (Teil-)autonomer Entscheidungen des Systems ohne Kontrolle, Eingriff oder Korrektur durch natürliche Personen– Ethische Dilemmata / Entscheidungsfindung– Mangelnde Nachvollziehbarkeit von Entscheidungen– Haftung für Entscheidungen des KI-Systems– Urheberschaft für Ausgaben des KI-Systems– Fehleranfälligkeit des KI-Systems ohne das dies durch Menschen erkannt wird– Fehlerhafte/unsachgemäße Nutzung (z.B. durch mangelnde Qualifikation)	Es wurde keine Risikoanalyse durchgeführt.		System	Analyse	Maximalwert	TR1.1	
MA2.3	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Das KI-System muss in Entwicklung und Betrieb überwacht werden (können).	<p>Tech. Maßnahme – Betrieb</p> <p>Eine kontinuierliche Überwachung muss für die Implementierung des KI-Systems vorgesehen sein und auf Basis entsprechender Protokollierung (siehe TR1.7) umgesetzt werden können. Dies beinhaltet die Erprobung folgender Möglichkeiten:</p> <ul style="list-style-type: none">– Monitoring der Leistung inklusive einer Überwachung der Modelle und Daten (d.h. einkommende Produktionsdaten und ggf. sich erweiternde Trainingsdaten-basis)– Monitoring von Verzerrungen inklusive einer Überwachung der Daten (d.h. einkommende Produktionsdaten und ggf. sich erweiternde Trainingsdatenbasis) im Kontext der Vermeidung von ungerechtfertigter Diskriminierung und Verzerrungen– Durchführung von Tests (z.B. Sanity Checks), die im Rahmen des Monitorings eingesetzt werden, um etwa Model und Concept Drift oder auch schädliche Eingabedaten zu erkennen– Falls anwendbar, Durchführung von Tests zur Erkennung von Missbrauch und schädlichen Eingabedaten– Ggf. Qualitätsüberprüfung der sich erweiternden Trainingsdatenbasis <p>Organisatorische Maßnahme – Systemnahe Prozesse</p> <p>Zusätzlich zur technischen Ermöglichung des Monitorings sollten die folgenden Aspekte vorbereitet werden:</p> <ul style="list-style-type: none">– Konzept zur (automatischen) Überwachung und -prüfung größerer Veränderungen am KI-System, inklusive bei Soft- und Hardware-Komponenten, aber insbesondere im Fall von Online Learning– Empfohlene Tests müssen als Teil eines kontinuierlichen Testplans dokumentiert sein, insbesondere im Fall von Online Learning– Falls möglich, Mechanismen in Form sinnvoller Definition von Schwellwerten bzw. Szenarien, bei denen (menschliche) Überprüfung und Mitigationsmaßnahmen eintreten sollten– Mechanismen zum Teilen von neuen Informationen über mögliche sicherheitsrelevante Vorfälle und ihrer Vermeidung	<p>Tech. Maßnahme – Betrieb</p> <p>Eine kontinuierliche Überwachung muss für die Implementierung des KI-Systems vorgesehen sein und auf Basis entsprechender Protokollierung (siehe TR1.7) umgesetzt werden können. Dies beinhaltet die folgender Möglichkeiten:</p> <ul style="list-style-type: none">– Monitoring der Leistung inklusive einer Überwachung der Modelle und Daten (d.h. einkommende Produktionsdaten und ggf. sich erweiternde Trainingsdaten-basis)– Monitoring von Verzerrungen inklusive einer Überwachung der Daten (d.h. einkommende Produktionsdaten und ggf. sich erweiternde Trainingsdatenbasis) im Kontext der Vermeidung von ungerechtfertigter Diskriminierung und Verzerrungen– Durchführung von Tests (z.B. Sanity Checks), die im Rahmen des Monitorings eingesetzt werden, um etwa Model und Concept Drift, oder auch schädliche Eingabedaten zu erkennen <p>Organisatorische Maßnahme – Systemnahe Prozesse</p> <p>Zusätzlich zur technischen Ermöglichung des Monitorings sollten die folgenden Aspekte vorbereitet werden:</p> <ul style="list-style-type: none">– Konzept zur (automatischen) Überwachung und -prüfung größerer Veränderungen am KI-System, inklusive bei Soft- und Hardware-Komponenten, aber insbesondere im Fall von Online Learning– Empfohlene Tests müssen als Teil eines kontinuierlichen Testplans dokumentiert sein, insbesondere im Fall von Online Learning	<p>Tech. Maßnahme – Betrieb</p> <p>Eine kontinuierliche Überwachung muss für die Implementierung des KI-Systems vorgesehen sein und auf Basis entsprechender Protokollierung (siehe TR1.7) umgesetzt werden können. Dies beinhaltet die folgender Möglichkeiten:</p> <ul style="list-style-type: none">– Monitoring der Leistung inklusive einer Überwachung der Modelle und Daten (d.h. einkommende Produktionsdaten und ggf. sich erweiternde Trainingsdaten-basis)– Monitoring von Verzerrungen inklusive einer Überwachung der Daten (d.h. einkommende Produktionsdaten und ggf. sich erweiternde Trainingsdatenbasis) im Kontext der Vermeidung von ungerechtfertigter Diskriminierung und Verzerrungen– Durchführung von Tests (z.B. Sanity Checks), die im Rahmen des Monitorings eingesetzt werden, um etwa Model und Concept Drift, oder auch schädliche Eingabedaten zu erkennen	Das KI-System kann im Betrieb nicht überwacht werden.		System	Maßnahme	Maximalwert	VE1.7, CY1.5 (Angriffe während der Betriebsphase), DA1.4	
MA2.4	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Die mit dem Betrieb und der Aufsicht betrauten Personen müssen Zugang zu einer verständlichen Beschreibung des KI-Systems haben und angemessen auf die Ausübung ihrer Aufgaben vorbereitet werden.	<p>Orga. Maßnahmen – Benutzerinstruktion</p> <p>Bereitstellung von Adressatengerecht aufbereiteten Instruktionen und Informationen mit mindestens den folgenden Inhalten:</p> <ul style="list-style-type: none">– Kontaktinformationen– Zweck des KI-Systems– Vorgesehener Anwendungsbereich des KI-Systems– Erwartbare Leistung und Funktionalität des KI-Systems– Bekannte Risiken und Implikationen für die Nutzung (der Ergebnisse) des KI-Systems– Zusammenfassung der erwarteten Eingabedaten– Beschreibung über die Interpretation der Ausgaben des KI-Systems– Beschreibung des Grads der Autonomie und der Entscheidungsprozesse des KI-Systems– Notwendige/mögliche Maßnahmen zur menschlichen Aufsicht und Kontrolle über das System, wann diese angewendet werden sollen/können, und wie erkennbar wird, dass ein menschlicher Eingriff gefordert ist– Anforderungen an die Nutzer (z.B. Schulung) für die Verwendung des KI-Systems– Notwendige Instandhaltungsmaßnahmen (inklusive Softwareupdates) <p>Orga. Maßnahmen – Schulungen</p> <p>Zur Qualifikation und Schulung des Personals, welches mit dem Betrieb und der Aufsicht des KI-Systems betraut ist, sind zielgruppenorientierte Schulungsmaterialien erstellt worden, die ein Fokus auf den Sachgemäßen Einsatz und ein erhöhtes Sicherheitsbewusstsein legen. Das Material umfasst mindestens folgende Aspekte:</p> <ul style="list-style-type: none">– Ordnungsgemäßer Umgang mit Systemkomponenten in der Produktionsumgebung– Ordnungsgemäßer Umgang mit relevanten Datentypen (z.B. Trainings- und Validierungsdaten, Betriebsdaten, Kundendaten) auch gemäß den geltenden Richtlinien sowie gesetzlichen und regulatorischen Anforderungen– Angemessene Durchführung von Training, Validierung und Testen im Betrieb des KI-Systems und seiner Komponenten– Überwachung der Leistung des KI-Systems und seiner Komponenten– Informationen über potenzielle Bedrohungsszenarien und dazugehöriger Mitigationsmaßnahmen (z.B. Interventionen)– Verhalten im Falle von Sicherheitsvorfällen	<p>Orga. Maßnahmen – Benutzerinstruktion</p> <p>Bereitstellung von Adressatengerecht aufbereiteten Instruktionen und Informationen mit mindestens den folgenden Inhalten:</p> <ul style="list-style-type: none">– Kontaktinformationen– Zweck des KI-Systems– Vorgesehener Anwendungsbereich des KI-Systems– Beschreibung über die Interpretation der Ausgaben des KI-Systems– Notwendige(mögliche Maßnahmen zur menschlichen Aufsicht und Kontrolle über das System, wann diese angewendet werden sollen/können, und wie erkennbar wird, dass ein menschlicher Eingriff gefordert ist– Notwendige Instandhaltungsmaßnahmen (inklusive Softwareupdates) <p>Orga. Maßnahmen – Schulungen</p> <p>Zur Qualifikation und Schulung des Personals, welches mit dem Betrieb und der Aufsicht des KI-Systems betraut ist, sind zielgruppenorientierte Schulungsmaterialien erstellt worden, die ein Fokus auf den Sachgemäßen Einsatz und ein erhöhtes Sicherheitsbewusstsein legen. Das Material umfasst mindestens folgende Aspekte:</p> <ul style="list-style-type: none">– Ordnungsgemäßer Umgang mit Systemkomponenten in der Produktionsumgebung– Überwachung der Leistung des KI-Systems und seiner Komponenten– Informationen über potenzielle Bedrohungsszenarien und dazugehöriger Mitigationsmaßnahmen (z.B. Interventionen)– Verhalten im Falle von Sicherheitsvorfällen	<p>Orga. Maßnahmen – Benutzerinstruktion</p> <p>Bereitstellung von Instruktionen und Informationen mit mindestens den folgenden Inhalten:</p> <ul style="list-style-type: none">– Kontaktinformationen– Zweck des KI-Systems– Vorgesehener Anwendungsbereich des KI-Systems– Beschreibung über die Interpretation der Ausgaben des KI-Systems– Notwendige(mögliche Maßnahmen zur menschlichen Aufsicht und Kontrolle über das System, wann diese angewendet werden sollen/können, und wie erkennbar wird, dass ein menschlicher Eingriff gefordert ist– Notwendige Instandhaltungsmaßnahmen (inklusive Softwareupdates)	Es sind keine Materialien vorhanden, die die mit dem Betrieb und der Aufsicht betrauten Personen auf einen Umgang mit dem KI-System vorbereiten.		System	Maßnahme	Normal	MA1.5	

Menschliche Aufsicht und Kontrolle									
MA2.5	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Maximalwert	Verknüpfung
Es muss bewertet werden, ob die ergriffenen Maßnahmen die festgestellten Risiken auf ein annehmbares Maß vermindert haben und die Qualität des KI-Systems den gesetzten Zielvorgaben entspricht.	Bewertung – Zusammenfassende Betrachtung der Effekte der Durchführung der technischen und organisatorischen Maßnahmen – Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen – Ermittlung und Beschreibung des Restrisikos nach Durchführung der technischen und organisatorischen Maßnahmen – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Bewertung – Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Bewertung – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Eine Bewertung wurde nicht durchgeführt.		System	Bewertung		

KI-spezifische Cybersicherheit									
CY1	Allgemeine KI-spezifische Cybersicherheit								
CY1.1	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Risiken für die Cybersicherheit des KI-Systems müssen unter Beachtung des Verwendungszwecks analysiert werden.	<p>Analyse – Risiko</p> <p>Detaillierte Risikoanalyse zur Cybersicherheit einschließlich Quantifizierung von Eintrittswahrscheinlichkeiten und Risiken:</p> <ul style="list-style-type: none">– Identifikation der möglichen Risiken und ihrer Ursachen– Zuweisung der Risikoverantwortung– Schätzung der Eintrittswahrscheinlichkeit– Schätzung der Aufdeckungswahrscheinlichkeit– Schätzung der Auswirkung– Strukturierte Abstufung und Priorisierung der Risiken <p>Die Risikobewertung sollte unter Beachtung des Verwendungszwecks und Anwendungsbereichs (siehe TR1.1) mindestens berücksichtigen:</p> <ul style="list-style-type: none">– Alle klassischen Schutzziele der Cybersicherheit, also Vertraulichkeit (z.B. Schutz von Trainingsdaten vor böswilliger Offenlegung oder Missbrauch), Integrität (z.B. Schutz des KI-Systems vor einer böswilligen Einflussnahme auf den Output zum Vorteil des Angreifers) und Verfügbarkeit (z.B. Schutz des KI-Systems vor einer böswillig ausgelösten Überlastung)– Bedrohungsbewertung (Threat Modelling) zur Bestimmung von Eintrittswahrscheinlichkeiten– Schwachstellenanalyse zur Bestimmung von Eintrittswahrscheinlichkeiten– Analyse der zu schützenden System-Assets zur Bestimmung der Auswirkungen <p>Das Risikobewertungsframework sollte möglichst in ein Cybersicherheitsmanagementsystem eingebettet werden können und dabei idealerweise bekannten Standards folgen.</p>	<p>Analyse – Risiko</p> <p>Limitierte Risikoanalyse mit Fokus auf den Schutzbedarf ohne Quantifizierung von Wahrscheinlichkeiten</p> <ul style="list-style-type: none">– Identifikation der möglichen Risiken– Zuweisung der Risikoverantwortung– Schätzung der Auswirkung– Strukturierte Abstufung und Priorisierung der Risiken <p>Die Risikobewertung sollte unter Beachtung des Verwendungszwecks und Anwendungsbereichs (siehe TR1.1) mindestens berücksichtigen:</p> <ul style="list-style-type: none">– Alle klassischen Schutzziele der Cybersicherheit, also Vertraulichkeit (z.B. Schutz personenbezogener Trainingsdaten vor böswilliger Offenlegung oder Missbrauch), Integrität (z.B. Schutz des KI-Systems vor einer böswilligen Einflussnahme auf den Output zum Vorteil des Angreifers) und Verfügbarkeit (z.B. Schutz des KI-Systems vor einer böswillig ausgelösten Überlastung)– Bedrohungsbewertung (Threat Modelling) zur Bestimmung von Eintrittswahrscheinlichkeiten– Schwachstellenanalyse zur Bestimmung von Eintrittswahrscheinlichkeiten– Analyse der zu schützenden System-Assets zur Bestimmung der Auswirkungen <p>Das Risikobewertungsframework sollte möglichst in ein Cybersicherheitsmanagementsystem eingebettet werden können und dabei idealerweise bekannten Standards folgen.</p>	<p>Analyse – Risiko</p> <p>Hauptsächlich qualitative Abschätzung ohne Wahrscheinlichkeiten</p> <ul style="list-style-type: none">– Identifikation der möglichen Risiken– Zuweisung der Risikoverantwortung– Qualitative Schätzung der Auswirkung– Qualitative Abstufung und Priorisierung der Risiken <p>Die Risikobewertung sollte unter Beachtung des Verwendungszwecks und Anwendungsbereichs (siehe TR1.1) mindestens berücksichtigen:</p> <ul style="list-style-type: none">– Alle klassischen Schutzziele der Cybersicherheit, also Vertraulichkeit (z.B. Schutz personenbezogener Trainingsdaten vor böswilliger Offenlegung oder Missbrauch), Integrität (z.B. Schutz des KI-Systems vor einer böswilligen Einflussnahme auf den Output zum Vorteil des Angreifers) und Verfügbarkeit (z.B. Schutz des KI-Systems vor einer böswillig ausgelösten Überlastung)– Bedrohungsbewertung (Threat Modelling) zur Bestimmung von Eintrittswahrscheinlichkeiten– Schwachstellenanalyse zur Bestimmung von Eintrittswahrscheinlichkeiten– Analyse der zu schützenden System-Assets zur Bestimmung der Auswirkungen <p>Das Risikobewertungsframework sollte möglichst in ein Cybersicherheitsmanagementsystem eingebettet werden können und dabei idealerweise bekannten Standards folgen.</p>	Es wurde keine Risikoanalyse durchgeführt.	Wichtige Standards sind zum Beispiel BSI Grundschutz, ISO 27k series, IEC 62443, ISO/SAE 21434, ETSI EN 303 645, NIST AI RMF, ...	System	Analyse	Maximalwert	TR1.1 (Verwendungszweck)
CY1.2	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Für das KI-System, einschließlich aller KI-Komponenten und des eingebetteten IKT-Systems, muss ein Cybersicherheitsmanagementsystem (CSMS) vorgesehen sein.	<p>Orga. Maßnahmen – Governance & Tech. Maßnahmen – Betrieb (Anbindung Cybersicherheitsmanagementsystem)</p> <p>Es muss vorgesehen sein, dass das KI-System in einer Cybersicherheitsmanagementsystem (CSMS) eingebettet werden kann, insbesondere:</p> <ul style="list-style-type: none">– Monitoring und Logging müssen implementiert oder vorgesehen sein und Schnittstellen für ein CSMS bereit stellen– die Dokumentation von Schwachstellen, Assets, Risikoanalyse und den Ergebnissen aus allen Mitigationsmaßnahmen sollte vollständig bereit gehalten sein– eine Strategie für Software -und Sicherheitsupdates sollte vorhanden sein <p>Das KI-System sollte Schnittstellen bereithalten um mindestens die folgenden Aspekte eines CSM abdecken zu können:</p> <ul style="list-style-type: none">– Ein Systemlebenszyklus basiertes System zur regelmäßigen Cybersicherheitsqualitätskontrolle einschließlich vorgesehener Reviews der Sicherheitsmaßnahmen und Protokolle.– Grundlegendes Cybersecurity Asset Management, einschließlich für Sicherheitskontrollen und Überwachungssysteme selber– Management von Schwachstellen (Identifizierung während des gesamten Lebenszyklus, insbesondere Training/Evaluation und Deployment-Phasen; wo nötig, sollten Penetrationstests vorgeschrieben sein)– Ein Cybersicherheits-Monitoringsystem– Zugangsmanagement zu den verschiedenen Assets und Komponenten des Systems– Aspekte des Personalmanagement, wie Schulung, Ausbildung und Rollenzuweisung in der Entwicklung und Begleitung des KI-Produkts– eine Analyse des erwarteten und tatsächlichen Zeitrahmens innerhalb dessen Sicherheitsaktualisierungen für das KI-System zur Verfügung gestellt werden– Strategien für regelmäßigen Review von Sicherheitsmaßnahmen und -protokollen– Dokumentation des CSMS. Die Informationen müssen direkt mit der KI-Anwendung einsehbar sein.– Das CSMS kann durch Compliance mit entsprechenden Cybersicherheitsstandards nachgewiesen sein, ein entsprechender Nachweis für das entwickelte KI-System sollte automatisch als A gewertet sein (z.B., BSI Grundschutz, ISO 27k series, IEC 62443, ISO/SAE 21434, ETSI EN 303 645, NIST AI RMF, ...)	<p>Orga. Maßnahmen – Governance & Tech. Maßnahmen – Betrieb (Anbindung Cybersicherheitsmanagementsystem)</p> <p>Es sollte so viele Information, wie möglich bereit gestellt werden, so dass das KI-System in einer Cybersicherheitsmanagementsystem (CSMS) eingebettet werden kann, insbesondere:</p> <ul style="list-style-type: none">– Monitoring und Logging müssen mindestens vorgesehen sein und Schnittstellen für ein CSMS bereit stellen– die Dokumentation von Schwachstellen, Assets, Risikoanalyse und den Ergebnissen aus allen Mitigationsmaßnahmen sollte vollständig bereit gehalten sein <p>Das KI-System sollte Schnittstellen bereithalten um mindestens die folgenden Aspekte eines CSMS abdecken zu können:</p> <ul style="list-style-type: none">– Ein Systemlebenszyklus basiertes System zur regelmäßigen Cybersicherheitsqualitätskontrolle einschließlich vorgesehener Reviews der Sicherheitsmaßnahmen und Protokolle.– Grundlegendes Cybersecurity Asset Management, einschließlich für Sicherheitskontrollen und Überwachungssysteme selber– Management von Schwachstellen (Identifizierung während des gesamten Lebenszyklus, insbesondere Training/Evaluation und Deployment-Phasen; wo nötig, sollten Penetrationstests vorgeschrieben sein)– Zugangsmanagement zu den verschiedenen Assets und Komponenten des Systems– eine Analyse des erwarteten und tatsächlichen Zeitrahmens innerhalb dessen Sicherheitsaktualisierungen für das KI System/Anwendung zur Verfügung gestellt werden– Strategien für regelmäßigen Review von Sicherheitsmaßnahmen und -protokollen– Dokumentation des CSMS. Die Informationen müssen direkt mit der KI-Anwendung einsehbar sein.– Das CSMS kann durch Compliance mit entsprechenden Cybersicherheitsstandards nachgewiesen sein, ein entsprechender Nachweis für das entwickelte KI-System sollte automatisch als A gewertet sein (z.B., BSI Grundschutz, ISO 27k series, IEC 62443, ISO/SAE 21434, ETSI EN 303 645, NIST AI RMF, ...)	<p>Orga. Maßnahmen – Governance & Tech. Maßnahmen – Betrieb (Anbindung Cybersicherheitsmanagementsystem)</p> <p>– Es sollte so viele Information zur KI-spezifischen Cybersicherheit, wie möglich bereit gestellt werden um das KI-System in größere Systemkontexte einbetten zu können insbesondere die Dokumentation von Schwachstellen, Assets, Risikoanalyse und den Ergebnissen aus allen Mitigationsmaßnahmen</p>	Die Einbettung in ein CSMS wurde in keiner Weise vorgesehen oder erleichtert.	Wichtige Standards sind zum Beispiel BSI Grundschutz, ISO 27k series, IEC 62443, ISO/SAE 21434, ETSI EN 303 645, NIST AI RMF, ...	System/Komponente	Maßnahme	Normal	CY1.5 (Datenzugang), CY2.3, CY2.2 (KI-spezifische Mitigationsmaßnahmen)
CY1.3	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Es müssen Maßnahmen zur Kontrolle der Cybersicherheitsaspekte des KI-Lieferkettenmanagements vorhanden sein, insbesondere im Hinblick auf Schwachstellen oder schädliche Praktiken bei der Verwendung von vortrainierten Modellen, Open-Source-Bibliotheken für maschinelles Lernen und Trainingsdaten von Dritten.	<p>Orga. Maßnahmen – Systemnahe Prozesse</p> <ul style="list-style-type: none">– Es sollte ein Prozess vorhanden sein zur Analyse der Cybersicherheit der Software- und Hardwarelieferketten hinsichtlich der Entwicklungs- und Produktumgebung des KI-Systems– Die Analyse der Lieferkette umfasst vortrainierte Modelle, Datensätze, und genutzte Softwarebibliotheken und die genutzte Hardware für Training- und ggf.. Produktionsphase des KI-Systems.– Es sollte ein Prozess vorgesehen sein zur stetigen weiteren Überwachung der Lieferkette einschließlich eines Plans für Updates und Review nach Auftauchen bisher unentdeckter Schwachstellen in der Lieferkette. <p>Tech. Maßnahmen – Modelle, Daten</p> <ul style="list-style-type: none">– Es muss eine Sicherheitsüberprüfung aller identifizierten genutzten vortrainierten Modelle, Datensätze und Softwarebibliotheken aus der Lieferkette nachgewiesen sein <p>Tech. Maßnahmen – Betrieb</p> <ul style="list-style-type: none">– Es sollten Schnittstellen vorhanden sein, welche die technische Überwachung der benutzten Soft- und Hardware des KI-Systems auch bzgl. Schwachstellen durch ein Cybersicherheitsmonitoring erlauben	<p>Orga. Maßnahmen – Systemnahe Prozesse</p> <ul style="list-style-type: none">– Es sollte ein Prozess vorhanden sein zur Analyse der Cybersicherheit der Software- und Hardwarelieferketten hinsichtlich der Entwicklungs- und Produktumgebung des KI-Systems– Die Analyse der Lieferkette umfasst vortrainierte Modelle, Datensätze, und genutzte Softwarebibliotheken und die genutzte Hardware für Training- und ggf.. Produktionsphase des KI-Systems.– Es sollte ein Prozess vorgesehen sein zur stetigen weiteren Überwachung der Lieferkette einschließlich eines Plans für Updates und Review nach Auftauchen bisher unentdeckter Schwachstellen in der Lieferkette.	<p>Orga. Maßnahmen – Systemnahe Prozesse</p> <ul style="list-style-type: none">– Es sollte ein Prozess vorhanden sein zur Analyse der Cybersicherheit der Software- und Hardwarelieferketten hinsichtlich der Entwicklungs- und Produktumgebung des KI-Systems– Die Analyse der Lieferkette umfasst vortrainierte Modelle, Datensätze, und genutzte Softwarebibliotheken	Es wurden keine Maßnahmen ergriffen zur Kontrolle der Cybersicherheitsaspekte des KI-Lieferkettenmanagements.		System/Komponente	Maßnahme	Normal	

KI-spezifische Cybersicherheit									
CY1.4	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Es müssen Maßnahmen und Sicherheitskontrollen vorhanden sein, um fehlerhafte Nutzung und Missbrauch des KI-Systems zu verhindern.	<p>Org. Maßnahmen – Systemnahe Prozesse</p> <p>– Prozesse müssen eingeführt sein um den möglichen beabsichtigten Missbrauchs des KI-Systems im Betrieb adressieren zu können, einschließlich der Einbindung möglicher Abschaltungsmechanismen im Risikomanagement.</p> <p>– Prozesse sollten eingeführt sein, um eine durch Monitoringsysteme und Logging festgestellte fehlerhafte Nutzung im Betrieb adressieren zu können</p> <p>– Prozesse sollten insbesondere darauf achten, dass neben der Funktionalität, Integrität und Verfügbarkeit des KI-Systems auch die unter den unterschiedlichen Qualitätsdimensionen definierten Ziele geschützt sind bei fehlerhafter Nutzung oder Missbrauch, also Wahrung des Schutzes personenbezogener oder proprietärer Daten, des Schutzes vor Nicht-Diskriminierung, des Autonomiegrades und der dazugehörigen menschlichen Kontrollmechanismen</p>	<p>Org. Maßnahmen – Systemnahe Prozesse</p> <p>– Prozesse müssen eingeführt sein um den möglichen beabsichtigten Missbrauchs des KI-Systems im Betrieb adressieren zu können, einschließlich der Einbindung möglicher Abschaltungsmechanismen</p> <p>– Prozesse sollten eingeführt sein, um eine durch Monitoringsysteme und Logging festgestellte fehlerhafte Nutzung im Betrieb adressieren zu können</p> <p>– Prozesse sollten insbesondere darauf achten, dass neben der Funktionalität, Integrität und Verfügbarkeit des KI-Systems auch die unter den unterschiedlichen Qualitätsdimensionen definierten Ziele geschützt sind bei fehlerhafter Nutzung oder Missbrauch, also Wahrung des Schutzes personenbezogener oder proprietärer Daten, des Schutzes vor Nicht-Diskriminierung, des Autonomiegrades und der dazugehörigen Menschlichen Kontrollmechanismen</p>	<p>Org. Maßnahmen – Systemnahe Prozesse</p> <p>– Prozesse müssen eingeführt sein um den möglichen beabsichtigten Missbrauchs des KI-Systems im Betrieb adressieren zu können, einschließlich der Einbindung möglicher Abschaltungsmechanismen</p> <p>– Prozesse sollten eingeführt sein, um eine durch Monitoringsysteme und Logging festgestellte fehlerhafte Nutzung im Betrieb adressieren zu können</p> <p>– Prozesse sollten insbesondere darauf achten, dass neben der Funktionalität, Integrität und Verfügbarkeit des KI-Systems auch die unter den unterschiedlichen Qualitätsdimensionen definierten Ziele geschützt sind bei fehlerhafter Nutzung oder Missbrauch, also Wahrung des Schutzes personenbezogener oder proprietärer Daten, des Schutzes vor Nicht-Diskriminierung, des Autonomiegrades und der dazu-gehörigen Menschlichen Kontrollmechanismen</p>	Es sind keine Maßnahmen vorhanden um fehlerhafte Nutzung und Missbrauch des KI-Systems zu verhindern.		System	Maßnahme	Normal	VE2.5 (derselbe Indikator)
CY1.5	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Es müssen Maßnahmen und Sicherheitskontrollen vorhanden sein, die den Zugang zu Trainings-, Test-, Validierungs-, und allen anderen KI-spezifischen Daten regeln.	<p>Org. Maßnahme – Governance</p> <p>– Die individuellen Zugriffsrechte auf die Daten müssen festgelegt und dokumentiert werden und mit dem Zweck des KI-Systems übereinstimmen (siehe auch CY1.2)</p> <p>– ein entsprechendes Zugriffsmanagement sollte mindestens beinhalten:</p> <p>a) Gewährung und Änderung (Provisionierung) von Zugriffsberechtigungen basierend auf dem Prinzip der minimalen Rechtevergabe und dem Need-to-know-Prinzip;</p> <p>b) Trennung von Aufgaben;</p> <p>c) Zugang zu Daten für unbefugte Subjekte wird verweigert;</p> <p>d) Regelmäßige Überprüfung der gewährten Berechtigungen;</p> <p>e) Entzug von Berechtigungen bei Änderungen im Beschäftigungsverhältnis oder der Rolle des Mitarbeiters</p> <p>f) Im Falle von personenbezogenen Daten, die Möglichkeit der betroffenen Personen Zugriff auf ihre Daten zu erhalten, siehe auch MA1.3</p> <p>g) Die Dokumentation der Daten sollte Details zum Zugangsmanagement beinhalten, siehe DA1.1</p>	<p>Org. Maßnahme – Governance</p> <p>– Die individuellen Zugriffsrechte auf die Daten müssen festgelegt und dokumentiert werden und mit dem Zweck des KI-Systems übereinstimmen (siehe auch CY1.2)</p> <p>– ein entsprechendes Zugriffsmanagement sollte mindestens beinhalten:</p> <p>a) Gewährung und Änderung (Provisionierung) von Zugriffsberechtigungen basierend auf dem Prinzip der minimalen Rechtevergabe und dem Need-to-know-Prinzip;</p> <p>b) Trennung von Aufgaben;</p> <p>c) Zugang zu Daten für unbefugte Subjekte wird verweigert;</p> <p>d) Regelmäßige Überprüfung der gewährten Berechtigungen;</p> <p>e) Entzug von Berechtigungen bei Änderungen im Beschäftigungsverhältnis oder der Rolle des Mitarbeiters</p> <p>f) Im Falle von personenbezogenen Daten, die Möglichkeit der betroffenen Personen Zugriff auf ihre Daten zu erhalten, siehe auch MA1.3</p> <p>g) Die Dokumentation der Daten sollte Details zum Zugangsmanagement beinhalten, siehe DA1.1</p>	<p>Org. Maßnahme – Governance</p> <p>– Die individuellen Zugriffsrechte auf die Daten müssen festgelegt und dokumentiert werden und mit dem Zweck des KI-Systems übereinstimmen (siehe auch CY1.2)</p>	Es wurden keine Maßnahmen ergriffen oder Sicherheitskontrollen eingeführt, um den Zugang zu KI-spezifischen Daten zu regeln.		Komponente	Maßnahme	Maximalwert	CY1.2 (CSMS), CY2.3 (Angriffe auf die Trainingsphase) DA2.1, DA2.2, DA3.2, MA1.3
	<p>Tech. Maßnahme – Daten</p> <p>– technische Maßnahmen wie ein Zugriffsmanagementsystem müssen definiert und vorhanden sein, um den Zugriff auf personenbezogene und proprietäre Trainings-, Test-, Validierungs-, und anderen KI-spezifischen Daten durch nicht autorisierte Personen zu verhindern</p> <p>– Die Übermittlung von personenbezogenen oder proprietären Daten zwischen Parteien oder in eine Cloud muss gesichert werden, mindestens durch</p> <p>a) Nutzung und Dokumentierung von Verschlüsselungsverfahren für die Übertragung (Daten in Bewegung)</p> <p>b) Implementierung technischer Schutzmaßnahmen für die Kommunikationssicherheit</p>	<p>Tech. Maßnahme – Daten</p> <p>– technische Maßnahmen wie ein Zugriffsmanagementsystem müssen definiert und vorhanden sein, um den Zugriff auf personenbezogene und proprietäre Trainings-, Test-, Validierungs-, und anderen KI-spezifischen Daten durch nicht autorisierte Personen zu verhindern</p> <p>– Die Übermittlung von personenbezogenen oder proprietären Daten zwischen Parteien oder in eine Cloud muss gesichert werden, mindestens durch</p> <p>a) Nutzung und Dokumentierung von Verschlüsselungsverfahren für die Übertragung (Daten in Bewegung)</p> <p>b) Implementierung technischer Schutzmaßnahmen für die Kommunikations-sicherheit</p>	<p>Tech. Maßnahme – Daten</p> <p>– technische Maßnahmen wie ein Zugriffsmanagementsystem müssen definiert und vorhanden sein, um den Zugriff auf personenbezogene und proprietäre Trainings-, Test-, Validierungs-, und anderen KI-spezifischen Daten durch nicht autorisierte Personen zu verhindern</p> <p>– Die Übermittlung von personenbezogenen oder proprietären Daten zwischen Parteien oder in eine Cloud muss gesichert werden, mindestens durch</p> <p>a) Nutzung und Dokumentierung von Verschlüsselungsverfahren für die Übertragung (Daten in Bewegung)</p> <p>b) Implementierung technischer Schutzmaßnahmen für die Kommunikations-sicherheit</p>						
CY1.6	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung
Es muss bewertet werden, ob die ergriffenen Maßnahmen die festgestellten Risiken auf ein annehmbares Maß vermindert haben und die Qualität des KI-Systems den gesetzten Zielvorgaben entspricht.	<p>Bewertung</p> <p>– Zusammenfassende Betrachtung der Effekte der Durchführung der technischen und organisatorischen Maßnahmen</p> <p>– Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen</p> <p>– Ermittlung und Beschreibung des Restrisikos nach Durchführung der technischen und organisatorischen Maßnahmen</p> <p>– Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist</p> <p>– Begründung der Tolerierbarkeit des Restrisikos</p>	<p>Bewertung</p> <p>– Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen</p> <p>– Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist</p> <p>– Begründung der Tolerierbarkeit des Restrisikos</p>	<p>Bewertung</p> <p>– Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist</p> <p>– Begründung der Tolerierbarkeit des Restrisikos</p>	Eine Bewertung wurde nicht durchgeführt.		System	Bewertung	Maximalwert	
CY2									
Widerstandsfähigkeit gegen KI-spezifische Angriffe									
CY2.1	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Risiken für die Cybersicherheit des KI-Systems durch KI-spezifische Angriffe müssen unter Beachtung des Verwendungszwecks analysiert werden.	<p>Analyse – Risiko</p> <p>Detaillierte Risikoanalyse zur Cybersicherheit einschließlich Quantifizierung von Eintrittswahrscheinlichkeiten und Risiken:</p> <p>– Identifikation der möglichen Risiken und ihrer Ursachen</p> <p>– Zuweisung der Risikoverantwortung</p> <p>– Schätzung der Eintrittswahrscheinlichkeit</p> <p>– Schätzung der Aufdeckungswahrscheinlichkeit</p> <p>– Schätzung der Auswirkung</p> <p>– Strukturierte Abstufung und Priorisierung der Risiken</p>	<p>Analyse – Risiko</p> <p>Limitierte Risikoanalyse mit Fokus auf den Schutzbedarf ohne Quantifizierung von Wahrscheinlichkeiten</p> <p>– Identifikation der möglichen Risiken</p> <p>– Zuweisung der Risikoverantwortung</p> <p>– Schätzung der Auswirkung</p> <p>– Strukturierte Abstufung und Priorisierung der Risiken</p>	<p>Analyse – Risiko</p> <p>Hauptsächlich qualitative Abschätzung ohne Wahrscheinlichkeiten</p> <p>– Identifikation der möglichen Risiken</p> <p>– Zuweisung der Risikoverantwortung</p> <p>– Qualitative Schätzung der Auswirkung</p> <p>– Qualitative Abstufung und Priorisierung der Risiken</p>	Es wurde keine Risikoanalyse durchgeführt.	Wichtige Standards sind zum Beispiel BSI Grundschutz, ISO 27k series, IEC 62443, ISO/SAE 21434, ETSI EN 303 645, NIST AI RMF, ...	System	Analyse	Maximalwert	TR1.1 (Verwendungszweck)
	<p>Die Risikobewertung sollte unter Beachtung des Verwendungszwecks und Anwendungsbereichs (siehe TR1.1) mindestens berücksichtigen:</p> <p>– Analyse aller potenziellen Angriffsszenarien und Angriffsvektoren auf die KI-Elemente des Systems, einschließlich KI-spezifische Angriffe (z.B. Evasion Attacks oder Data Poisoning) oder klassische Angriffe (z.B. in KI-Softwarebibliotheken eingebettete Malware)</p>	<p>Die Risikobewertung sollte unter Beachtung des Verwendungszwecks und Anwendungsbereichs (siehe TR1.1) mindestens berücksichtigen:</p> <p>– Analyse aller potenziellen Angriffsszenarien und Angriffsvektoren auf die KI-Elemente des Systems, einschließlich KI-spezifische Angriffe (z.B. Evasion Attacks oder Data Poisoning) oder klassische Angriffe (z.B. in KI-Softwarebibliotheken eingebettete Malware) .</p>	<p>Die Risikobewertung sollte unter Beachtung des Verwendungszwecks und Anwendungsbereichs (siehe TR1.1) mindestens berücksichtigen:</p> <p>– Analyse aller potenziellen Angriffsszenarien und Angriffsvektoren auf die KI-Elemente des Systems, einschließlich KI-spezifische (z.B. Evasion Attacks oder Data Poisoning) oder klassische Angriffe (z.B. in KI-Softwarebibliotheken eingebettete Malware)</p>	Das Risikobewertungsframework sollte möglichst in ein Cybersicherheits-managementsystem eingebettet werden können und dabei idealerweise bekannten Standards folgen.					

KI-spezifische Cybersicherheit									
CY2.2	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Es müssen Maßnahmen und Sicherheitskontrollen vorhanden sein, um KI-spezifische Angriffe die während des Betriebs auftreten können zu mitigieren.	<p>Tech. Maßnahmen – Test (Metriken KI-Modelle Angriffe in Inferenz- und Deploymentphase):</p> <p>Prüfung von durch den Anwendungsbereich begründeter Auswahl und Anzahl an Metriken und Tests der adversariellen Robustheit in Bezug auf Angriffsszenarien in der Inferenz- und Deploymentphase, die gegen Vertraulichkeit, Integrität oder Verfügbarkeit des KI-Systems gerichtet sein können.</p> <p>– Dies schließt eine Prüfung mit Basis- und fortgeschrittenen Methoden mindestens der folgende Angriffsarten ein:</p> <p>a) evasion attacks, b) latency attacks c) model extraction attacks d) data reconstruction oder property inference attacks e) prompt injection attacks oder jailbreaks</p> <p>– Penetrationstests von Systemaspekten mit Bezug auf KI-spezifische Angriffe während des Betriebs sollten durchgeführt werden.</p> <p>Die ausgewählten Metriken und Tests sollten mindestens die folgenden Merkmale aufweisen:</p> <p>– Methoden sollten für jedes KI-Modell im und für das KI-System als Ganzes angewendet werden, falls dieses einen gesammelten Output erzeugt</p> <p>– wenn durch die Auswahl begründet, sollten die Metriken und Tests sowohl wichtige Basis- methoden als auch fortgeschrittene Methoden (hinsichtlich Komplexität, Informationsgehalt, Implementierungsaufwand, z.B. umfangreiches Prüfwerkzeug) umfassen</p> <p>– die ausgewählten Metriken und Tests sollten wenn möglich einen hohen Grad der Automatisierung erlauben</p> <p>– ggf. Festlegung von Schwellenwerten (als Mindestanforderungen an die Funktionalität/Leistung) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks</p> <p>– falls anwendbar: Abstufung der Metriken und Schwellenwerte nach unterschiedlichen Einsatzszenarien mit Bezug zur Definition des Anwendungsbereichs.</p> <p>Tech. Maßnahmen – Modell</p> <p>– angemessene KI-spezifische Sicherheitskontrollen und technische Maßnahmen zur Mitigation des identifizierten Risikos von Angriffen in der Inferenzphase und während des Deployment des KI-Systems müssen auf Modellebene eingeführt werden (z.B. adversariales Training, Methoden zur Stärkung der Modellrobustheit oder Unsicherheitsmethoden)</p> <p>– es sollten Robustheitsnachweise für jedes KI-Modell geführt werden zu verschiedenen relevanten Angriffsarten</p> <p>Tech. Maßnahmen – System</p> <p>– angemessene KI-spezifische oder nicht-KI-spezifische Sicherheitskontrollen zur Mitigation des identifizierten Risikos von Angriffen in der Inferenzphase und während des Deployment des KI-Systems müssen auf Systemebene eingeführt werden (z.B. Monitoring von Inputs, siehe MA2.3)</p> <p>– klassisches Sicherheitskontrollen des gesamten Systems im Betrieb sollten vorhergesehen und eingeplant sein einschließlich einer Einbindung in ein existierendes CSMS.</p> <p>– Red Teaming des gesamten Systems mit Bezug auf KI-spezifische Angriffe während der Betriebsphase sollten durchgeführt werden</p>	<p>Tech. Maßnahmen – Test (Metriken KI-Modelle Angriffe in Inferenz- und Deploymentphase):</p> <p>Prüfung von durch den Anwendungsbereich begründeter Auswahl und Anzahl an Metriken und Tests der adversariellen Robustheit in Bezug auf Angriffsszenarien in der Inferenz- und Deploymentphase, die gegen Vertraulichkeit, Integrität oder Verfügbarkeit des KI-Systems gerichtet sein können.</p> <p>– Dies schließt eine Prüfung mit Basis-Methoden mindestens der folgende Angriffsarten ein:</p> <p>a) evasion attacks, b) latency attacks c) model extraction attacks d) data reconstruction oder property inference attacks e) prompt injection attacks oder jailbreaks</p> <p>– Penetrationstests des gesamten Systems mit Bezug auf KI-spezifische Angriffe während des Betriebs sollten durchgeführt werden.</p> <p>Die ausgewählten Metriken und Tests sollten mindestens die folgenden Merkmale aufweisen:</p> <p>– Methoden sollten für jedes KI-Modell im und für das KI-System als Ganzes angewendet werden, falls dieses einen gesammelten Output erzeugt</p> <p>– wenn durch die Auswahl begründet, sollten die Metriken und Tests sowohl wichtige Basis- methoden als auch fortgeschrittene Methoden (hinsichtlich Komplexität, Informationsgehalt, Implementierungsaufwand, z.B. umfangreiches Prüfwerkzeug) umfassen</p> <p>– die ausgewählten Metriken und Tests sollten wenn möglich einen hohen Grad der Automatisierung erlauben</p> <p>– ggf. Festlegung von Schwellenwerten (als Mindestanforderungen an die Funktionalität/Leistung) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks</p> <p>– falls anwendbar: Abstufung der Metriken und Schwellenwerte nach unterschiedlichen Einsatzszenarien mit Bezug zur Definition des Anwendungsbereichs.</p> <p>Tech. Maßnahmen – Modell</p> <p>– angemessene KI-spezifische Sicherheitskontrollen zur Mitigation des identifizierten Risikos von Angriffen in der Inferenzphase und während des Deployment des KI-Systems müssen auf Modellebene eingeführt werden (z.B. adversariales Training, Methoden zur Stärkung der Modellrobustheit oder Unsicherheitsmethoden)</p> <p>– es sollten möglichst Robustheitsnachweise für jedes KI-Modell geführt werden zu verschiedenen relevanten Angriffsarten</p> <p>Tech. Maßnahmen – System</p> <p>– angemessene KI-spezifische oder nicht-KI-spezifische Sicherheitskontrollen zur Mitigation des identifizierten Risikos von Angriffen in der Inferenzphase und während des Deployment des KI-Systems müssen auf Systemebene eingeführt werden (z.B. Monitoring von Inputs, siehe MA2.3)</p> <p>– klassisches Sicherheitskontrollen des gesamten Systems im Betrieb sollten vorhergesehen und eingeplant sein einschließlich einer Einbindung in ein existierendes CSMS.</p>	<p>Tech. Maßnahmen – Test (Metriken KI-Modelle Angriffe in Inferenz- und Deploymentphase):</p> <p>Prüfung von durch den Anwendungsbereich begründeter Auswahl und Anzahl an Metriken und Tests der adversariellen Robustheit in Bezug auf Angriffsszenarien in der Inferenz- und Deploymentphase, die gegen Vertraulichkeit, Integrität oder Verfügbarkeit des KI-Systems gerichtet sein können.</p> <p>– Dies schließt eine Prüfung mit Basis-Methoden möglichst der folgende Angriffsarten ein:</p> <p>a) evasion attacks, b) latency attacks c) model extraction attacks d) data reconstruction oder property inference attacks e) prompt injection attacks oder jailbreaks</p> <p>Die ausgewählten Metriken und Tests sollten möglichst die folgenden Merkmale aufweisen:</p> <p>– Methoden können entweder für jedes KI-Modell oder für das KI-System als Ganzes angewendet werden, je nachdem, was sinnvoller und machbar ist</p> <p>– es genügen Basismethoden (hinsichtlich Komplexität, Informationsgehalt, Implementierungs-aufwand, z.B. eine simple Metrik)</p> <p>– ggf. Festlegung von Schwellenwerten (als Mindestanforderungen an die Funktionalität/Leistung) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks</p> <p>Tech. Maßnahmen – Modell</p> <p>– angemessene KI-spezifische Sicherheitskontrollen zur Mitigation des identifizierten Risikos von Angriffen in der Inferenzphase und während des Deployment des KI-Systems müssen auf Modellebene eingeführt werden (z.B. adversariales Training, Methoden zur Stärkung der Modellrobustheit oder Unsicherheitsmethoden)</p> <p>– es sollten Robustheitsnachweise für jedes KI-Modell geführt werden zu verschiedenen relevanten Angriffsarten</p> <p>Tech. Maßnahmen – System</p> <p>– angemessene KI-spezifische oder nicht-KI-spezifische Sicherheitskontrollen zur Mitigation des identifizierten Risikos von Angriffen in der Inferenzphase und während des Deployment des KI-Systems müssen auf Systemebene eingeführt werden (z.B. Monitoring von Inputs, siehe MA2.3)</p> <p>– klassisches Sicherheitskontrollen des gesamten Systems im Betrieb sollten vorhergesehen und eingeplant sein einschließlich einer Einbindung in ein existierendes CSMS.</p>	Es wurden keine Maßnahmen ergriffen, um KI-spezifische Angriffe während des Betriebs zu mitigieren.	<p>Sicherheitskontrollen und Maßnahmen auf Modellebene können im Detail zum Beispiel sein:</p> <p>Maßnahmen in der Modelarchitektur wie Model hardening durch Distillation, Maßnahmen im Training, wie Adversariales Training, oder Unsicherheitsmethoden zur Erkennung adversarialer Inputs.</p> <p>Sicherheitskontrollen auf Systemeben können zum Beispiel sein: Zugangs-Management, Benutzungsrestriktionen für Anzahl der Aufrufe des Systems, Out-of-distribution Detection, Prompt Filtering</p> <p>Für Details zur Taxonomy verschiedener Angriffsarten und Mitigationen siehe zum Beispiel MITRE ATLAS oder NIST AI 100-2 E2023 zu "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations"</p>	System/Komponente	Maßnahme	Maximalwert	CY1.1 (Risiko), CY1.2 (CSMS), MA2.3 (Monitoring), VE1.3 (Tests und Maßnahmen zur allgemeinen Robustheit), VE2.2 (Maßnahmen zur Mitigation von Systemfehler- und Ausfall), TR1.4, TR1.3
CY2.3	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Es müssen Maßnahmen und Sicherheitskontrollen vorhanden sein, um KI-spezifische Angriffe, die während der Vorbereitungs- und Modelltrainingsphase der KI-Modelle stattfinden zu mitigieren.	<p>Tech. Maßnahmen – Test (Metriken KI-Modelle Angriffsszenarien Vorbereitungs- und Modelltrainingsphase):</p> <p>Prüfung von durch den Anwendungsbereich begründeter Auswahl und Anzahl an Metriken und Tests der adversariellen Robustheit in Bezug auf Angriffsszenarien der Vorbereitungs- und Modelltrainingsphase, die gegen Vertraulichkeit, Integrität oder Verfügbarkeit des KI-Systems gerichtet sein können.</p> <p>– Dies schließt eine Prüfung mit Basis- und fortgeschrittenen Methoden mindestens der folgende Angriffsarten ein:</p> <p>a) data poisoning b) label poisoning c) KI-spezifische backdoor attacks</p> <p>– Penetrationstests von Systemaspekten mit Bezug auf KI-spezifische Angriffe während der Vorbereitungs- und Modelltrainingsphase sollten durchgeführt werden.</p> <p>Die ausgewählten Metriken und Tests sollten mindestens die folgenden Merkmale aufweisen:</p> <p>– Methoden sollten für jedes KI-Modell im und für das KI-System als Ganzes angewendet werden, falls dieses einen gesammelten Output erzeugt</p> <p>– wenn durch die Auswahl begründet, sollten die Metriken und Tests sowohl wichtige Basismethoden als auch fortgeschrittene Methoden (hinsichtlich Komplexität, Informationsgehalt, Implementierungsaufwand, z.B. umfangreiches Prüfwerkzeug) umfassen</p> <p>– die ausgewählten Metriken und Tests sollten wenn möglich einen hohen Grad der Automatisierung erlauben</p> <p>– ggf. Festlegung von Schwellenwerten (als Mindestanforderungen an die Funktionalität/Leistung) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks</p> <p>– falls anwendbar: Abstufung der Metriken und Schwellenwerte nach unterschiedlichen Einsatzszenarien mit Bezug zur Definition des Anwendungsbereichs.</p> <p>Tech. Maßnahmen – Modell</p> <p>– angemessene KI-spezifische Sicherheitskontrollen zur Mitigation des identifizierten Risikos von Angriffen in der Vorbereitungs- und Modelltrainingsphase des KI-Systems müssen auf Modellebene eingeführt werden (z. B Datenfiltermethoden oder Suchsysteme für Backdoors, Unsicherheitsmethoden)</p> <p>– es sollten Robustheitsnachweise für jedes KI-Modell geführt werden zu verschiedenen relevanten Angriffsarten</p> <p>Tech. Maßnahmen – System</p> <p>– angemessene KI-spezifische oder nicht-KI-spezifische Sicherheitskontrollen zur Mitigation des identifizierten Risikos von Angriffen von Angriffen in der Vorbereitungs- und Modelltrainingsphase des KI-Systems müssen auf Systemebene eingeführt werden.</p> <p>– klassisches Sicherheitskontrollen des gesamten Systems im Betrieb sollten vorhergesehen und eingeplant sein einschließlich einer Einbindung in ein existierendes CSMS.</p> <p>– Red Teaming des gesamten Systems mit Bezug auf KI-spezifische Angriffe während der Vorbereitungs- und Modelltrainingsphase sollten durchgeführt werden.</p>	<p>Tech. Maßnahmen – Test (Metriken KI-Modelle Angriffsszenarien Vorbereitungs- und Modelltrainingsphase):</p> <p>Prüfung von durch den Anwendungsbereich begründeter Auswahl und Anzahl an Metriken und Tests der adversariellen Robustheit in Bezug auf Angriffsszenarien der Vorbereitungs- und Modelltrainingsphase, die gegen Vertraulichkeit, Integrität oder Verfügbarkeit des KI-Systems gerichtet sein können.</p> <p>– Dies schließt eine Prüfung mit Basis-Methoden mindestens der folgende Angriffsarten ein:</p> <p>a) data poisoning b) label poisoning c) KI-spezifische backdoor attacks</p> <p>Die ausgewählten Metriken und Tests sollten mindestens die folgenden Merkmale aufweisen:</p> <p>– Methoden sollten für jedes KI-Modell im und für das KI-System als Ganzes angewendet werden, falls dieses einen gesammelten Output erzeugt</p> <p>– wenn durch die Auswahl begründet, sollten die Metriken und Tests sowohl wichtige Basismethoden als auch fortgeschrittene Methoden (hinsichtlich Komplexität, Informationsgehalt, Implementierungsaufwand, z.B. umfangreiches Prüfwerkzeug) umfassen</p> <p>– ggf. Festlegung von Schwellenwerten (als Mindestanforderungen an die Funktionalität/Leistung) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks</p> <p>Tech. Maßnahmen – Modell</p> <p>– angemessene KI-spezifische Sicherheitskontrollen zur Mitigation des identifizierten Risikos von Angriffen in der Vorbereitungs- und Modelltrainingsphase des KI-Systems müssen auf Modellebene eingeführt werden (z. B Datenfiltermethoden oder Suchsysteme für Backdoors, Unsicherheitsmethoden)</p> <p>– es sollten möglichst Robustheitsnachweise für jedes KI-Modell geführt werden zu verschiedenen relevanten Angriffsarten</p> <p>Tech. Maßnahmen – System</p> <p>– angemessene KI-spezifische oder nicht-KI-spezifische Sicherheitskontrollen zur Mitigation des identifizierten Risikos von Angriffen von Angriffen in der Vorbereitungs- und Modell-trainingsphase des KI-Systems müssen auf Systemebene eingeführt werden.</p>	<p>Tech. Maßnahmen – Test (Metriken KI-Modelle Angriffsszenarien Vorbereitungs- und Modelltrainingsphase):</p> <p>Prüfung von durch den Anwendungsbereich begründeter Auswahl und Anzahl an Metriken und Tests der adversariellen Robustheit in Bezug auf Angriffsszenarien der Vorbereitungs- und Modelltrainingsphase, die gegen Vertraulichkeit, Integrität oder Verfügbarkeit des KI-Systems gerichtet sein können.</p> <p>– Dies schließt eine Prüfung mit Basis-Methoden möglichst der folgende Angriffsarten ein:</p> <p>a) data poisoning b) label poisoning c) KI-spezifische backdoor attacks</p> <p>Die ausgewählten Metriken und Tests sollten möglichst die folgenden Merkmale aufweisen:</p> <p>– Methoden können entweder für jedes KI-Modell oder für das KI-System als Ganzes angewendet werden, je nachdem, was sinnvoller und machbar ist</p> <p>– es genügen Basismethoden (hinsichtlich Komplexität, Informationsgehalt, Implementierungs-aufwand, z.B. eine simple Metrik)</p> <p>– ggf. Festlegung von Schwellenwerten (als Mindestanforderungen an die Funktionalität/Leistung) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks</p> <p>Tech. Maßnahmen – Modell</p> <p>– angemessene KI-spezifische Sicherheitskontrollen zur Mitigation des identifizierten Risikos von Angriffen in der Vorbereitungs- und Modelltrainingsphase des KI-Systems müssen auf Modellebene eingeführt werden (z. B Datenfiltermethoden oder Suchsysteme für Backdoors, Unsicherheitsmethoden)</p> <p>– es sollten möglichst Robustheitsnachweise für jedes KI-Modell geführt werden zu verschiedenen relevanten Angriffsarten</p> <p>Tech. Maßnahmen – System</p> <p>– angemessene KI-spezifische oder nicht-KI-spezifische Sicherheitskontrollen zur Mitigation des identifizierten Risikos von Angriffen von Angriffen in der Vorbereitungs- und Modell-trainingsphase des KI-Systems müssen auf Systemebene eingeführt werden.</p>	Es wurden keine Maßnahmen ergriffen um KI-spezifische Angriffe während der Vorbereitungs- und Modelltrainingsphase zu mitigieren.	<p>Sicherheitskontrollen auf Modell-ebene/Dateneben können zum Beispiel sein:</p> <p>Mitigationen, wie zum Beispiel das Filtern auf Data Poisioning oder Suchsysteme für Modell-Backdoors</p> <p>Sicherheitskontrollen auf Systemeben können zum Beispiel sein: Access Management, Datensicherheitsmanagement, Out-of-distribution Detection,</p> <p>Für Details zur Taxonomy verschiedener Angriffsarten und Mitigationen siehe zum Beispiel MITRE ATLAS oder NIST AI 100-2 E2023 zu "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations"</p>	System/Komponente	Maßnahme	Maximalwert	CY1.1 (Risiko), DA1.1 (Risiko), CY1.2 (CSMS), CY1.8 (Datenzugang), VE1.3 (Tests und Maßnahmen zur allgemeinen Robustheit), VE2.2 (Maßnahmen zur Mitigation von Systemfehler- und Ausfall), TR1.4, TR1.3
CY2.4	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung
Es muss bewertet werden, ob die ergriffenen Maßnahmen die festgestellten Risiken auf ein annehmbares Maß vermindert haben und die Qualität des KI-Systems den gesetzten Zielvorgaben entspricht.	<p>Bewertung</p> <p>– Zusammenfassende Betrachtung der Effekte der Durchführung der technischen und organisatorischen Maßnahmen</p> <p>– Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen</p> <p>– Ermittlung und Beschreibung des Restrisikos nach Durchführung der technischen und organisatorischen Maßnahmen</p> <p>– Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist</p> <p>– Begründung der Tolerierbarkeit des Restrisikos</p>	<p>Bewertung</p> <p>– Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen</p> <p>– Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist</p> <p>– Begründung der Tolerierbarkeit des Restrisikos</p>	<p>Bewertung</p> <p>– Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist</p> <p>– Begründung der Tolerierbarkeit des Restrisikos</p>	Eine Bewertung wurde nicht durchgeführt.		System	Bewertung	Maximalwert	

Verlässlichkeit										
VE1										
Leistungsfähigkeit und Robustheit										
VE1.1	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Risiken, die zu unzureichender Leistungsfähigkeit und Robustheit von KI-Komponenten des KI-Systems führen können, müssen unter Beachtung des Verwendungszwecks analysiert werden.	<p>Analyse – Risiko</p> <p>Detaillierte Risikoanalyse einschließlich Quantifizierung von Eintrittswahrscheinlichkeiten und Risiken:</p> <ul style="list-style-type: none">– Identifikation der möglichen Risiken und ihrer Ursachen– Zuweisung der Risikoverantwortung– Schätzung der Eintrittswahrscheinlichkeit– Schätzung der Aufdeckungswahrscheinlichkeit– Schätzung der Auswirkung– Strukturierte Abstufung und Priorisierung der Risiken <p>Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen:</p> <ul style="list-style-type: none">– Fehlerhaftes Modelltraining, die falsche Wahl des Modells oder ungeeignete Optimierungs- und Validierungsstrategien– schlecht gewählte Metriken und Tests– Limits im Modelltraining, z.B. durch zu schlechte oder zu wenige Trainings- oder Validierungs- oder Testdaten, Overfitting– Limitierungen durch Hardware bei Training und Inferenz– Notwendigkeit einer gut kalibrierten Unsicherheitsabschätzung– Auswirkung einer Änderung der Zusammensetzung des KI-Systems (Soft- und Hardwareebene), insbesondere Risiken durch Neutraining oder Online-Learning	<p>Analyse – Risiko</p> <p>Limitierte Risikoanalyse mit Fokus auf den Schutzbedarf ohne Quantifizierung von Wahrscheinlichkeiten</p> <ul style="list-style-type: none">– Identifikation der möglichen Risiken– Zuweisung der Risikoverantwortung– Schätzung der Auswirkung– Strukturierte Abstufung und Priorisierung der Risiken <p>Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen:</p> <ul style="list-style-type: none">– Fehlerhaftes Modelltraining, die falsche Wahl des Modells oder ungeeignete Optimierungs- und Validierungsstrategien– schlecht gewählte Metriken und Tests– Limits im Modelltraining, z.B. durch zu schlechte oder zu wenige Trainings- oder Validierungs- oder Testdaten, Overfitting– Limitierungen durch Hardware bei Training und Inferenz– Notwendigkeit einer gut kalibrierten Unsicherheitsabschätzung– Auswirkung einer Änderung der Zusammensetzung des KI-Systems (Soft- und Hardwareebene), insbesondere Risiken durch Neutraining oder Online-Learning	<p>Analyse – Risiko</p> <p>Hauptsächlich qualitative Abschätzung der Gefährdungen ohne Wahrscheinlichkeiten</p> <ul style="list-style-type: none">– Identifikation der möglichen Gefährdungen:– Zuweisung der Risikoverantwortung– Qualitative Schätzung der Auswirkung– Qualitative Abstufung und Priorisierung der Gefährdungen <p>Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen:</p> <ul style="list-style-type: none">– Fehlerhaftes Modelltraining, die falsche Wahl des Modells oder ungeeignete Optimierungs- und Validierungsstrategien– schlecht gewählte Metriken und Tests– Limits im Modelltraining, z.B. durch zu schlechte oder zu wenige Trainings- oder Validierungs- oder Testdaten, Overfitting– Limitierungen durch Hardware bei Training und Inferenz– Notwendigkeit einer gut kalibrierten Unsicherheitsabschätzung– Auswirkung einer Änderung der Zusammensetzung des KI-Systems (Soft - und Hardwareebene), insbesondere Risiken durch Neutraining oder Online-Learning	Es wurde keine Risikoanalyse oder Gefährdungsabschätzung durchgeführt.	Bei Verwendung von Modellen externer Anbieter Dokumentationen und klare Begründungen zur Validierung bereitstellen.	Komponente	Analyse	Maximalwert	TR1.1, TR1.3	
					Im Gegensatz zu VE2.1 geht es hier nicht um die Risiken die aus einem KI-System mit unzureichender Leistungsfähigkeit entstehen, sondern um die Risiken die zu unzureichender Leistungsfähigkeit selber führen und damit den Verwendungszweck gefährden können.					
					Risiko = Gefährdung x Wahrscheinlichkeit des Eintritts.					
VE1.2	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Es müssen geeignete Metriken und Tests definiert werden, um zu bewerten, ob die Leistung des KI-Systems die beabsichtigte Funktionsweise realisiert.	<p>Analyse – Metriken / Schwellenwerte [Leistungsfähigkeit]</p> <p>Durch den Anwendungsbereich begründete Auswahl und Anzahl an Metriken und Tests, je nach Nutzen möglichst unterschiedlicher Kategorien. Dies schließt ein:</p> <ul style="list-style-type: none">– statistische Auswertung des KI-Modells oder empirische Auswertung der Systemfunktionalität auf das gesamte KI-System z.B. im Sinne von Nutzerexperimenten oder Befragungen– abhängig vom KI-Modell gewählte passende Methoden zur Unsicherheitsbestimmung oder die Nutzung probabilistischer KI-Modellarchitekturen; Metriken für Kalibrierung von Unsicherheitsbestimmung (z.B. über Konfidenzwerte) <p>Die ausgewählten Metriken und Tests sollten möglichst die folgenden Merkmale und Rahmenbedingungen aufweisen:</p> <ul style="list-style-type: none">– Methoden sollten für jedes KI-Modell im und für das KI-System als Ganzes angewendet werden, falls dieses einen gesammelten Output erzeugt– wenn durch die Auswahl begründet, sollten die Metriken und Tests sowohl Basismethoden (simple Metrik) und einige fortgeschrittene Methoden (hinsichtlich Informationsgehalt, Aussagekräftigkeit, Implementierungsaufwand, z.B. umfangreiches Prüfwerkzeug – siehe zusätzliche Information) umfassen– Festlegung von Schwellenwerten (als Mindestanforderungen an die Funktionalität/Leistung) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks (falls anwendbar: Abstufung der Metriken und Schwellenwerte nach unterschiedlichen Einsatzszenarien mit Bezug zur Definition des Anwendungsbereichs) <p>Die Begründung der Metriken, Tests und Methoden geht auf folgende Aspekte ein:</p> <ul style="list-style-type: none">– Anwendungsbereich, Zweck des KI-Systems,– Modelltyp,– Zusammensetzung des KI-Systems (d.h., Zusammenspiel der Komponenten bzw. Zusammenhang von ML-Modell und Gesamtsystem), Aufgabenbereich der KI-Systems (z.B. Klassifikation vs. Regression)	<p>Analyse – Metriken / Schwellenwerte [Leistungsfähigkeit]</p> <p>Durch den Anwendungsbereich begründete Auswahl und Anzahl an Metriken und Tests, je nach Nutzen möglichst unterschiedlicher Kategorien. Dies schließt ein:</p> <ul style="list-style-type: none">– statistische Auswertung des KI-Modells oder empirische Auswertung der Systemfunktionalität auf das gesamte KI-System z.B. im Sinne von Nutzerexperimenten oder Befragungen– abhängig vom KI-Modell gewählte passende Methoden zur Unsicherheitsbestimmung oder die Nutzung probabilistischer KI-Modellarchitekturen; Metriken für Kalibrierung von Unsicherheitsbestimmung (z.B. über Konfidenzwerte) <p>Die ausgewählten Metriken und Tests sollten möglichst die folgenden Merkmale und Rahmenbedingungen aufweisen:</p> <ul style="list-style-type: none">– Methoden sollten für jedes KI-Modell im und für das KI-System als Ganzes angewendet werden, falls dieses einen gesammelten Output erzeugt– wenn durch die Auswahl begründet, sollten die Metriken und Tests sowohl Basismethoden (simple Metrik) und einige fortgeschrittene Methoden (hinsichtlich Informationsgehalt, Aussagekräftigkeit, Implementierungsaufwand, z.B. umfangreiches Prüfwerkzeug – siehe zusätzliche Information) umfassen– Festlegung von Schwellenwerten (als Mindestanforderungen an die Funktionalität/Leistung) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks <p>Die Begründung der Metriken, Tests und Methoden geht auf folgende Aspekte ein:</p> <ul style="list-style-type: none">– Anwendungsbereich, Zweck des KI-Systems,– Modelltyp,– Zusammensetzung des KI-Systems (d.h., Zusammenspiel der Komponenten bzw. Zusammenhang von ML-Modell und Gesamtsystem), Aufgabenbereich der KI-Systems (z.B. Klassifikation vs. Regression)	<p>Analyse – Metriken / Schwellenwerte [Leistungsfähigkeit]</p> <p>Durch den Anwendungsbereich begründete Auswahl und Anzahl an Metriken und Tests. Dies schließt je nach KI-Modell mindestens eine Methode ein aus:</p> <ul style="list-style-type: none">– statistische Auswertung des KI-Modells oder empirische Auswertung der Systemfunktionalität auf das gesamte KI-System z.B. im Sinne von Nutzerexperimenten oder Befragungen <p>Die ausgewählten Metriken und Tests sollten möglichst die folgenden Merkmale und Rahmenbedingungen aufweisen:</p> <ul style="list-style-type: none">– Methoden können entweder für jedes KI-Modell oder für das KI-System als Ganzes angewendet werden, je nachdem, was sinnvoller und machbar ist– Es genügen Basismethoden (hinsichtlich Informationsgehalt, Aussagekräftigkeit, Implementierungsaufwand, z.B. umfangreiches Prüfwerkzeug – siehe zusätzliche Information)– Festlegung von Schwellenwerten (als Mindestanforderungen an die Funktionalität/Leistung) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks <p>Die Begründung der Metriken, Tests und Methoden geht auf mindestens einen der folgenden Aspekte ein:</p> <ul style="list-style-type: none">– Anwendungsbereich, Zweck des KI-Systems,– Modelltyp,– Zusammensetzung des KI-Systems (d.h., Zusammenspiel der Komponenten bzw. Zusammenhang von ML-Modell und Gesamtsystem), Aufgabenbereich der KI-Systems (z.B. Klassifikation vs. Regression)	Es wurden systematisch keine Metriken oder Tests festgelegt, die zur Untersuchung der Leistungsfähigkeit des Systems dienen.	Bei Verwendung von Modellen externer Anbieter Dokumentationen und klare Begründungen zur Validierung bereitstellen.	Komponente	Analyse	Maximalwert	TR1.3, TR1.4	
					Die beabsichtigte Funktionsweise sollte sich aus dem Verwendungszweck und Anwendungsbereich ergeben.					
					Unterscheidung zwischen Basismethoden und fortgeschrittene Methoden:					
					- In der "Technische Prüfmethodensammlung.xlsx" sind einige gängige Methoden gelistet und in Basis- und fortgeschrittene Methoden kategorisiert					
					- Selbst entwickelte Testmethoden werden als fortgeschrittene Methoden anerkannt					
					Die ausgewählten Metriken und Tests sollten wenn möglich einen hohen Grad der Automatisierung erlauben.					
VE1.3	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Es müssen geeignete Metriken und Tests definiert werden, um zu bewerten, ob die Robustheit des KI-Systems die beabsichtigte Funktionsweise realisiert.	<p>Analyse – Metriken / Schwellenwerte [Robustheit]</p> <p>Durch den Anwendungsbereich begründete Auswahl und Anzahl an Metriken, Tests und Methoden zur Analyse der Robustheit des Systems, je nach Nutzen möglichst unterschiedlicher Kategorien. Dies schließt ein:</p> <ul style="list-style-type: none">– statistische Auswertung der Robustheit des KI-Modells, z.B. zur Bestimmung von Edge-Cases oder empirische Auswertung der Systemfunktionalität– abhängig vom KI-Modell gewählte passende Methoden zur Unsicherheitsbestimmung oder die Nutzung probabilistischer KI-Modellarchitekturen; Metriken für Kalibrierung von Unsicherheitsbestimmung– z.B. zur Bewertung von Outliereffekten– Detektion von fehlerhaften Eingaben oder Fehlfunktionen auf Modellebene <p>Die ausgewählten Metriken und Tests sollten mindestens die folgenden Merkmale aufweisen:</p> <ul style="list-style-type: none">– Methoden sollten für jedes KI-Modell im und für das KI-System als Ganzes angewendet werden, falls dieses einen gesammelten Output erzeugt– wenn durch die Auswahl begründet, sollten die Metriken und Tests sowohl Basismethoden (simple Metrik) und einige fortgeschrittene Methoden (hinsichtlich Informationsgehalt, Aussagekräftigkeit, Implementierungsaufwand, z.B. umfangreiches Prüfwerkzeug – siehe zusätzliche Information) umfassen– Festlegung von Schwellenwerten (als Mindestanforderungen an die Robustheit) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks (falls anwendbar: Abstufung der Metriken und Schwellenwerte nach unterschiedlichen Einsatzszenarien mit Bezug zur Definition des Anwendungsbereichs) <p>Die Begründung der Metriken, Tests und Methoden geht auf folgende Aspekte ein:</p> <ul style="list-style-type: none">– Anwendungsbereich, Zweck des KI-Systems,– Modelltyp,– Zusammensetzung des KI-Systems (d.h., Zusammenspiel der Komponenten bzw. Zusammenhang von ML-Modell und Gesamtsystem), Aufgabenbereich der KI-Systems (z.B. Klassifikation vs. Regression)	<p>Analyse – Metriken / Schwellenwerte [Robustheit]</p> <p>Durch den Anwendungsbereich begründete Auswahl und Anzahl an Metriken und Tests, je nach Nutzen möglichst unterschiedlicher Kategorien. Dies schließt ein:</p> <ul style="list-style-type: none">– statistische Auswertung des KI-Modells oder empirische Auswertung der Systemfunktionalität auf das gesamte KI-System z.B. im Sinne von Nutzerexperimenten oder Befragungen– abhängig vom KI-Modell gewählte passende Methoden zur Unsicherheitsbestimmung oder die Nutzung probabilistischer KI-Modellarchitekturen; Metriken für Kalibrierung von Unsicherheitsbestimmung z.B. zur Bewertung von Outliereffekten– Detektion von fehlerhaften Eingaben oder Fehlfunktionen auf Modellebene <p>Die ausgewählten Metriken und Tests sollten möglichst die folgenden Merkmale aufweisen:</p> <ul style="list-style-type: none">– Methoden sollten für jedes KI-Modell im und für das KI-System als Ganzes angewendet werden, falls dieses einen gesammelten Output erzeugt– wenn durch die Auswahl begründet, sollten die Metriken und Tests sowohl Basismethoden (simple Metrik) und einige fortgeschrittene Methoden (hinsichtlich Informationsgehalt, Aussagekräftigkeit, Implementierungsaufwand, z.B. umfangreiches Prüfwerkzeug – siehe zusätzliche Information) umfassen– Festlegung von Schwellenwerten (als Mindestanforderungen an die Funktionalität/Leistung) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks <p>Die Begründung der Metriken, Tests und Methoden geht auf folgende Aspekte ein:</p> <ul style="list-style-type: none">– Anwendungsbereich, Zweck des KI-Systems,– Modelltyp,– Zusammensetzung des KI-Systems (d.h., Zusammenspiel der Komponenten bzw. Zusammenhang von ML-Modell und Gesamtsystem), Aufgabenbereich der KI-Systems (z.B. Klassifikation vs. Regression)	<p>Analyse – Metriken / Schwellenwerte [Leistungsfähigkeit]</p> <p>Durch den Anwendungsbereich begründete Auswahl und Anzahl an Metriken und Tests. Dies schließt je nach KI-Modell mindestens eine Methode ein aus:</p> <ul style="list-style-type: none">– statistische Auswertung des KI-Modells oder empirische Auswertung der Systemfunktionalität auf das gesamte KI-System z.B. im Sinne von Nutzerexperimenten oder Befragungen <p>– Detektion von fehlerhaften Eingaben oder Fehlfunktionen auf Modellebene</p> <p>Die ausgewählten Metriken und Tests sollten möglichst die folgenden Merkmale aufweisen:</p> <ul style="list-style-type: none">– Methoden können entweder für jedes KI-Modell oder für das KI-System als Ganzes angewendet werden, je nachdem, was sinnvoller und machbar ist– Es genügen Basismethoden (hinsichtlich Informationsgehalt, Aussagekräftigkeit, Implementierungsaufwand, z.B. umfangreiches Prüfwerkzeug – siehe zusätzliche Information)– Festlegung von Schwellenwerten (als Mindestanforderungen an die Funktionalität/Leistung) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks <p>Die Begründung der Metriken, Tests und Methoden geht auf folgende Aspekte ein:</p> <ul style="list-style-type: none">– Anwendungsbereich, Zweck des KI-Systems,– Modelltyp,– Zusammensetzung des KI-Systems (d.h., Zusammenspiel der Komponenten bzw. Zusammenhang von ML-Modell und Gesamtsystem), Aufgabenbereich der KI-Systems (z.B. Klassifikation vs. Regression)	Es wurden systematisch keine Metriken oder Tests festgelegt, die zur Untersuchung der Robustheit des Systems dienen.	Die beabsichtigte Funktionsweise sollte sich aus dem Verwendungszweck und Anwendungsbereich ergeben.	Komponente	Analyse	Normal	TR1.3, TR1.4, CY2.2, CY2.3 (Tests und Mitigationen zur Robustheit gegen Angriffe)	
					Die Begründung der Metriken geht auf zum Beispiel auf folgende Aspekte ein:					
					– Anwendungsbereich, Zweck des KI-Systems,					
					– Modelltyp,					
					– Zusammensetzung des KI-Systems (d.h., Zusammenspiel der Komponenten bzw. Zusammenhang von ML-Modell und Gesamtsystem), Aufgabenbereich der KI-Systems (z.B. Klassifikation vs. Regression)					
					(Classification/Regression/Generation/ Unsupervised (z.B. Clustering, Anomaly Detection...)/Reinforcement Learning etc.)					
					Für die statistische Auswertung: falls vorhanden auf etablierten Benchmark-Datensätzen, z.B. mit passenden Augmentierungen zur Abdeckung von Robustheitstests					
					Unterscheidung zwischen Basis-methoden und fortgeschrittene Methoden:					
					– In der "Technische Prüfmethodensammlung.xlsx" sind einige gängige Methoden gelistet und in Basis- und fortgeschrittene Methoden kategorisiert					
					– Selbst entwickelte Testmethoden werden als fortgeschrittene Methoden anerkannt					
					Die ausgewählten Metriken und Tests sollten wenn möglich einen hohen Grad der Automatisierung erlauben.					

Verlässlichkeit										
VE1.4	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Es muss ein Testplan entwickelt und implementiert sein, der das Prüfen aller vorgesehenen Metriken und Tests umfasst, einschließlich einer Prüfung des KI-Systems unter repräsentativen Bedingungen des Anwendungsbereiches.	Orga. Maßnahmen – Systemnahe Prozesse Dokumentierter Testplan für Leistungsfähigkeit und Robustheit, der mindestens die folgenden Punkte definiert (z.B. tabellarisch): – Testobjekte müssen definiert werden (d.h., zu testendes Modell (mit Versionsnummer) oder andere zu testende Komponente/Algorithmus wie etwa eine Unsicherheitsschätzung zur Bewertung von Outliereffekten) – vorgesehene Testmethoden müssen mindestens die in VE1.2-VE1.3 zuvor definierten Metriken und Tests beinhalten und müssen, soweit möglich Leistungsfähigkeit, Robustheit und Unsicherheitsabschätzung abdecken. Je nach zuvor festgestellter Analyse müssen möglicherweise auch fortgeschrittene Methoden im Testplan vorgesehen werden (zum Beispiel hinsichtlich Implementierungsaufwand und Hardwareanforderungen) und die Eigenschaften der Testmethoden klar und im Detail festgehalten sein hinsichtlich Komplexität und Informationsgehalt – ggf. müssen zusätzliche Vorgaben zu den Testparametern festgelegt werden – Verwendeten Testdaten müssen beschrieben werden – die Testumgebung muss eine Prüfung des KI-Systems unter repräsentativen Bedingungen des Anwendungsbereichs erlauben, mit einer hohen Nähe zur späteren Produktivumgebung – Zeitpunkt bzw. Regelmäßigkeit der Tests, einschließlich Vorbereitung von Testplänen für eine zukünftige Betriebsphase des KI-Systems, um die Leistungsfähigkeit und Robustheit des KI-Systems fortwährend testen zu können, insbesondere hinsichtlich auf eine Änderung der Systemzusammensetzung oder sich verändernder Trainingsdaten, siehe VE1.6 – Beachtung der Auswirkung möglicher signifikanter Änderungen des KI-System in Robustheitstests – die für die Durchführung benötigten Testressourcen (Software- und Hardwareanforderungen) müssen bestimmt und festgehalten sein – verantwortliche Person zur Durchführung und Dokumentation der Tests muss festgelegt sein – Begründung des Testplans mit Bezug auf den Anwendungskontext (einschließlich ggf., einer ODD) und auf VE1.2-VE1.3 mit einer Argumentation, dass der Testplan alle wichtigen Aspekte der Leistungsfähigkeit und Robustheit (bzgl. aller notwendigen, repräsentativen Szenarien) abdeckt.	Orga. Maßnahmen – Systemnahe Prozesse Dokumentierter Testplan für Leistungsfähigkeit und Robustheit, der mindestens die folgenden Punkte definiert (z.B. tabellarisch): – Testobjekte müssen definiert werden (d.h., zu testendes Modell (mit Versionsnummer) oder andere zu testende Komponente/Algorithmus wie etwa eine Unsicherheitsschätzung zur Bewertung von Outliereffekten) – Vorgesehene Testmethoden müssen mindestens die in VE1.2-VE1.3 zuvor definierten Metriken und Tests beinhalten und müssen, soweit möglich Leistungsfähigkeit, Robustheit und Unsicherheitsabschätzung abdecken. Je nach zuvor festgestellter Analyse müssen möglicherweise auch fortgeschrittene Methoden im Testplan vorgesehen werden (zum Beispiel hinsichtlich Implementierungsaufwand und Hardwareanforderungen) und die Eigenschaften der Testmethoden klar und im Detail festgehalten sein hinsichtlich Komplexität und Informationsgehalt – ggf. müssen zusätzliche Vorgaben zu den Testparametern festgelegt werden – Verwendeten Testdaten müssen beschrieben werden – Die Testumgebung muss eine Prüfung des KI-Systems unter möglichst repräsentativen Bedingungen des Anwendungsbereichs erlauben, aber nicht unbedingt die endgültige Produktivumgebung genau widerspiegeln – Zeitpunkt bzw. Regelmäßigkeit der Tests, einschließlich Vorbereitung von Testplänen für eine zukünftige Betriebsphase des KI-Systems, um die Leistungsfähigkeit und Robustheit des KI-Systems fortwährend testen zu können, insbesondere hinsichtlich auf eine Änderung der Systemzusammensetzung oder sich verändernder Trainingsdaten, siehe VE1.6 – Die für die Durchführung benötigten Testressourcen (Software- und Hardwareanforderungen) müssen bestimmt und festgehalten sein – Verantwortliche Person zur Durchführung und Dokumentation der Tests muss festgelegt sein – Begründung des Testplans mit Bezug auf den Anwendungskontext (einschließlich ggf., einer ODD) und auf VE1.2-VE1.3	Orga. Maßnahmen – Systemnahe Prozesse Dokumentierter Testplan für Leistungsfähigkeit und Robustheit, der mindestens die folgenden Punkte definiert (z.B. tabellarisch): – Testobjekte müssen definiert werden (d.h., zu testendes Modell (mit Versionsnummer) oder andere zu testende Komponente/Algorithmus wie etwa eine Unsicherheitsschätzung zur Bewertung von Outliereffekten) – Vorgesehene Testmethoden müssen mindestens die in VE1.2-VE1.3 zuvor definierten Metriken und Tests beinhalten und müssen, soweit möglich Leistungsfähigkeit, Robustheit und Unsicherheitsabschätzung abdecken. Die Eigenschaften der Testmethoden müssen klar und im Detail festgehalten sein hinsichtlich Komplexität und Informationsgehalt – ggf. müssen zusätzliche Vorgaben zu den Testparametern festgelegt werden – Verwendeten Testdaten müssen beschrieben werden – Zeitpunkt bzw. Regelmäßigkeit der Tests, einschließlich Vorbereitung von Testplänen für eine zukünftige Betriebsphase des KI-Systems, um die Leistungsfähigkeit und Robustheit des KI-Systems fortwährend testen zu können, insbesondere hinsichtlich auf eine Änderung der Systemzusammensetzung oder sich verändernder Trainingsdaten, siehe VE1.6 – Die für die Durchführung benötigten Testressourcen (Software- und Hardwareanforderungen) müssen bestimmt und festgehalten sein – Verantwortliche Person zur Durchführung und Dokumentation der Tests muss festgelegt sein	Es wurden keine systematischen Testpläne definiert.	Die Einteilung der Metriken und Methoden in einfach bis fortgeschritten hängt von vielen Details ab, aber orientiert sich grob and Komplexität und erwarteten Informationsgehalt (z.B. einfache Metrik, Benchmark, bis hin zu Expertengetriebene Validierungsansätze wie etwa systematische Schwachstellensuche, visuelle Exploration, Anwendung von XAI-Methoden, etc.) Der Grad der Automatisierung der Tests, wenn möglich sollten automatisierbare Methoden bevorzugt eingesetzt werden.	System/Komponente	Maßnahme	Normal	TR1.6 (Entwicklungsprozessdokumentation)	
VE1.5	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Das KI-System muss gemäß dem Testplan mit unterschiedlichen Eingaben, Bedingungen und Umgebungen getestet werden, um seine Leistungsfähigkeit und Robustheit sicherzustellen.	Tech. Maßnahme – Tests – Formelle Beschreibung des Anwendungsbereichs als Eingabe für Testmethoden (Eingaberaum, Anwendungsbereich mit Verteilung, Anwendungsgrenze, ggfs. Beschreibung einer ODD) – Entlang des Testplans (siehe VE1.4), Analyse der Abdeckung des Anwendungsbereichs (oder ggfls. einer ODD) durch die vorhandenen Testdaten – Beschreibung des gegebenen Formats d.h. die Eingaben, Bedingungen und Umgebungen, auf denen auf Leistungsfähigkeit und Robustheit direkt oder indirekt getestet wird – Durchführung des Tests entlang des Testplans und Dokumentation aller Testergebnisse – Dokumentation von Schwachstellen, dabei mindestens a) Jede aufgrund des Systemdesigns und in Bezug auf den Verwendungszweck unerwartete Minderung in der tatsächlichen Leistungsfähigkeit des KI-Systems b) Grenzen des Eingabebereichs (In welchen Situationen/ bei welchen Eingaben funktioniert das KI-System nur eingeschränkt oder gar nicht mehr? c) Shortcuts in Modellen (falls keine identifiziert wurden ist dies so zu dokumentieren)	Tech. Maßnahme – Tests – Formelle Beschreibung des Anwendungsbereichs als Eingabe für Testmethoden (Eingaberaum, Anwendungsbereich mit Verteilung, Anwendungsgrenze, ggfls. Beschreibung einer ODD) – Durchführung des Tests entlang des Testplans und Dokumentation aller Testergebnisse – Dokumentation von Schwachstellen, dabei mindestens a) Jede aufgrund des Systemdesigns und in Bezug auf den Verwendungszweck unerwartete Minderung in der tatsächlichen Leistungsfähigkeit des KI-Systems b) Grenzen des Eingabebereichs (In welchen Situationen/ bei welchen Eingaben funktioniert das KI-System nur eingeschränkt oder gar nicht mehr?	Tech. Maßnahme – Tests – Formelle Beschreibung des Anwendungsbereichs als Eingabe für Testmethoden (Eingaberaum, Anwendungsbereich mit Verteilung, Anwendungsgrenze, ggfls. Beschreibung einer ODD) – Durchführung des Tests entlang des Testplans und Dokumentation aller Testergebnisse	Es wurden keine dokumentierten Tests durchgeführt.	Für die technische Details zur Beschreibung von Anwendungs-bereichen siehe z.B. der Fraunhofer KI-Prüfkatalog z.B. [VE-R-RE-RI-01] [VE-R-RE-KR-02], [VE-R-RO-RI-01], [VE-R-RO- KR-01], [VE-R-RO-KR-03] und generell die Testmaßnahmen im Kapitel Verlässlichkeit.	System/Komponente	Maßnahme	Maximalwert	TR1.6 (Entwicklungsprozessdokumentation)	
VE1.6	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung	
Ein Monitoring der Leistungsfähigkeit und Robustheit des KI-Systems muss möglich sein.		MA2.3				System/Komponente	Maßnahme	Normal	VE1.6	
VE1.7	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung	
Es muss bewertet werden, ob die ergriffenen Maßnahmen die festgestellten Risiken auf ein annehmbares Maß vermindert haben und die Qualität des KI-Systems den gesetzten Zielvorgaben entspricht.	Bewertung – Zusammenfassende Betrachtung der Effekte der Durchführung der technischen und organisatorischen Maßnahmen – Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen – Ermittlung und Beschreibung des Restrisikos nach Durchführung der technischen und organisatorischen Maßnahmen – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Bewertung – Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Bewertung – Abwägung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Eine Bewertung wurde nicht durchgeführt.		System	Bewertung	Maximalwert		

Verlässlichkeit									
VE2									
VE2	Rückfallpläne und funktionale Sicherheit								
VE2.1	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Risiken für mögliche Folgen einer Fehlfunktion oder eines Ausfalls und für die funktionale Sicherheit des KI-System müssen unter Beachtung des Verwendungszwecks analysiert werden.	Analyse – Risiko Die Risikoanalyse zu möglichen Folgen eine (Teil-) Ausfalls des KI-Systems sowie für die funktionale Sicherheit soll sich mit durch das KI-System erzeugten Gefährdungen und schädlichen Auswirkungen für die Außenwelt durch Fehlfunktion oder Ausfall aufgrund unzureichender Leistungsfähigkeit oder Robustheit beschäftigen. Die Risikoanalyse sollte beinhalten: Detaillierte Risikoanalyse einschließlich Quantifizierung von Eintrittswahrscheinlichkeiten und Risiken: – Identifikation der möglichen Risiken und ihrer Ursachen – Zuweisung der Risikoverantwortung – Schätzung der Eintrittswahrscheinlichkeit – Schätzung der Aufdeckungswahrscheinlichkeit – Schätzung der Auswirkung – Strukturierte Abstufung und Priorisierung der Risiken Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen: – für den Verwendungszweck unzureichende Leistungsfähigkeit und Robustheit der KI-Komponenten des KI-Systems (siehe VE1) – Kritikalität des Verwendungszwecks und Anwendungsbereichs in Bezug auf mögliche Schäden und Folgen durch unzureichende Leistungsfähigkeit und Robustheit – mögliche funktionale Gründe für den Ausfall oder Teilausfall des KI-Systems mit deren Eintrittswahrscheinlichkeiten – fehlerhafter (systematisch und zufällig) oder missbräuchlicher Einsatz des KI-Systems außerhalb des Verwendungszwecks – Systemkomponenten mit unmittelbaren Schnittstellen mit dem KI-System (z.B. die Inputdaten liefern oder Outputdaten des KI-Systems verwenden) Mögliche Folgen und Schäden die aus Unfallrisiken herrühren können, sollten mindestens berücksichtigt werden für: – Leib & Leben und die körperliche Gesundheit von Menschen – Grundrechte – Eigentum und Sachen	Analyse – Risiko Die Risikoanalyse zu möglichen Folgen eine (Teil-) Ausfalls des KI-Systems sowie für die funktionale Sicherheit soll sich mit durch das KI-System erzeugten Gefährdungen und schädlichen Auswirkungen für die Außenwelt durch Fehlfunktion oder Ausfall aufgrund unzureichender Leistungsfähigkeit oder Robustheit beschäftigen. Die Risikoanalyse sollte beinhalten: Limitierte Risikoanalyse mit Fokus auf den Schutzbedarf ohne Quantifizierung von Wahrscheinlichkeiten – Identifikation der möglichen Risiken – Zuweisung der Risikoverantwortung – Schätzung der Auswirkung – Strukturierte Abstufung und Priorisierung der Risiken Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen: – für den Verwendungszweck unzureichende Leistungsfähigkeit und Robustheit der KI-Komponenten des KI-Systems (siehe VE1) – Kritikalität des Verwendungszwecks und Anwendungsbereichs in Bezug auf mögliche Schäden und Folgen durch unzureichende Leistungsfähigkeit und Robustheit – mögliche funktionale Gründe für den Ausfall oder Teilausfall des KI-Systems mit deren Eintrittswahrscheinlichkeiten – fehlerhafter (systematisch und zufällig) oder missbräuchlicher Einsatz des KI-Systems außerhalb des Verwendungszwecks – Systemkomponenten mit unmittelbaren Schnittstellen mit dem KI-System (z.B. die Inputdaten liefern oder Outputdaten des KI-Systems verwenden) Mögliche Folgen und Schäden die aus Unfallrisiken herrühren können, sollten mindestens berücksichtigt werden für: – Leib & Leben und die körperliche Gesundheit von Menschen – Grundrechte – Eigentum und Sachen	Analyse – Risiko Die Risikoanalyse zu möglichen Folgen eine (Teil-) Ausfalls des KI-Systems sowie für die funktionale Sicherheit soll sich mit durch das KI-System erzeugten Gefährdungen und schädlichen Auswirkungen für die Außenwelt durch Fehlfunktion oder Ausfall aufgrund unzureichender Leistungsfähigkeit oder Robustheit beschäftigen. Die Risikoanalyse sollte beinhalten: Hauptsächlich qualitative Abschätzung ohne Wahrscheinlichkeiten – Identifikation der möglichen Gefährdungen – Zuweisung der Risikoverantwortung – Qualitative Schätzung der Auswirkung – Qualitative Abstufung und Priorisierung der Gefährdungen Mindestens die folgenden Risikoquellen sind unter Beachtung des Verwendungszwecks (siehe TR1.1) zu berücksichtigen: – für den Verwendungszweck unzureichende Leistungsfähigkeit und Robustheit der KI-Komponenten des KI-Systems (siehe VE1) – Kritikalität des Verwendungszwecks und Anwendungsbereichs in Bezug auf mögliche Schäden und Folgen durch unzureichende Leistungsfähigkeit und Robustheit – mögliche funktionale Gründe für den Ausfall oder Teilausfall des KI-Systems mit deren Eintrittswahrscheinlichkeiten – fehlerhafter (systematisch und zufällig) oder missbräuchlicher Einsatz des KI-Systems außerhalb des Verwendungszwecks – Systemkomponenten mit unmittelbaren Schnittstellen mit dem KI-System (z.B. die Inputdaten liefern oder Outputdaten des KI-Systems verwenden) Mögliche Folgen und Schäden die aus Unfallrisiken herrühren können, sollten mindestens berücksichtigt werden für: – Leib & Leben und die körperliche Gesundheit von Menschen – Grundrechte – Eigentum und Sachen	Es wurde keine Risikoanalyse durchgeführt.	Hier kann zusätzlich auf die Schutzbedarfsanalyse hingewiesen werden. Es sollte aber beachtet werden, dass hier nicht eine Risikoanalyse vollständig vom Prüfenden durchgeführt werden soll sondern die Existenz einer solchen und deren Umfang bewertet werden soll. Hier wäre es möglich auf gängige Standards, z.B. zu funktionalen Sicherheit und Safety hinzuweisen so wie ISO/IEC Guide 51 und das Durchführen einer FMEA (die falls vorhanden hier eigentlich angerechnet werden sollte) Mehr Details können im KI-Prüfkatalog des Fraunhofer IAIS Kapitel 8 gefunden werden. Klassische Referenz zu funktionaler KI-Sicherheit Amodel 2006.	System	Analyse	Maximalwert	
VE2.2	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Es müssen geeignete technische Maßnahmen im Betrieb des KI-Systems zur Mitigation des Risikos einer Fehlfunktion oder eines Ausfalls des KI-System definiert werden.	Techn. Maßnahme – Betrieb – systemische Redundanz, und Rückfallmechanismen sollten im KI-System für den Betrieb vorgesehen sein (z.B. Umschaltung auf einen sicheren Modus, Not-Aus-Schalter), – Mechanismen für den sicheren Ausfall sollten geplant und wenn möglich schon implementiert sein (z.B. Manipulationsschutz, sicherer Modus), – Schnittstellen des in VE1.6 vorgesehenen Monitoringsystems mit den Rückfall- und Ausfallsystemen sollten sichergestellt sein um das Monitoring von KI-Komponenten mit denen eines größeren IT-Systems verbinden zu können – Schnittstellen für ein Alarmsystem sollten geplant sein (Endbenutzer, Anbieter, zuständige Behörde), – Ausfallsicheres Logging eines unter Betrieb befindlichen Produkts muss unterstützbar sein durch das KI-System (z.B. Blackbox), siehe auch nach TR1.7 – Interventionsmaßnahmen ("incident response") zur Fehlerbehebung und eine Systemwiederherstellung aus dem Betrieb sollte möglich sein	Techn. Maßnahme – Betrieb Durch den Anwendungsbereich begründete Auswahl und Anzahl an Metriken und Tests, je nach Nutzen möglichst unterschiedlicher Kategorien, Dies schließt ein: – Mechanismen für den sicheren Ausfall sollten geplant und wenn möglich schon implementiert sein (z.B. Manipulationsschutz, sicherer Modus), – Schnittstellen des in VE1.6 vorgesehenen Monitoringsystems mit den Rückfall- und Ausfallsystemen sollten sichergestellt sein um das Monitoring von KI-Komponenten mit denen eines größeren IT-Systems verbinden zu können, siehe auch nach TR1.7	Techn. Maßnahme – Betrieb – Mechanismen für den sicheren Ausfall sollten geplant und wenn möglich schon implementiert sein (z.B. Manipulationsschutz, sicherer Modus),	Es wurden keine Maßnahmen ergriffen, es zu ermöglichen im Betrieb das Risiko von Fehlfunktionen und Ausfall zu mindern.		System	Analyse	Normal	MA2.3 (Monitoring), TR1.7(Logging), CY2.2, CY2.3 (Maßnahmen zur Mitigation von Angriffen auf das System)
VE2.3	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Es müssen geeignete technische Maßnahmen in Form von Metriken und Tests zur Mitigation des Risikos einer Fehlfunktion oder eines Ausfalls des KI-System definiert werden.	Techn. Maßnahmen – Metriken & Schwellenwerte Durch den Anwendungsbereich begründete Auswahl und Anzahl an Metriken und Tests, je nach Nutzen möglichst unterschiedlicher Kategorien. Dies schließt ein: – Methoden, die die Verlässlichkeit der KI-Komponenten fördern/sicherstellen, z.B. gut gewählte Optimierungsmethode, Unsicherheitsbestimmungen (z.B. Konfidenzwerte) – Detektionsmethoden und Abfangen von Fehlern in Checks auf Daten, Modellebene oder auf den Ausgaben einschließlich sinnvoller Schwellenwerte. – Detektionsmethoden und Abfangen von Fehlern in Checks auf Systemebene einschließlich nicht KI-spezifischer Maßnahmen und sinnvoller Schwellenwerte Die ausgewählten Metriken und Tests sollten mindestens die folgenden Merkmale aufweisen: - wenn durch die Auswahl begründet, sollten die Metriken und Tests sowohl Basismethoden (simple Metrik) und einige fortgeschrittene Methoden (hinsichtlich Informationsgehalt, Aussagekräftigkeit, Implementierungsaufwand, z.B. umfangreiches Prüfwerkzeug - siehe zusätzliche Information) umfassen – ggf. Festlegung von Schwellenwerten (als Mindestanforderungen an die Funktionalität/Leistung des gesamten Systems) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks	Techn. Maßnahmen – Metriken & Schwellenwerte Durch den Anwendungsbereich begründete Auswahl und Anzahl an Metriken und Tests, je nach Nutzen möglichst unterschiedlicher Kategorien. Dies schließt ein: – Methoden, die die Verlässlichkeit der KI-Komponenten fördern/sicherstellen, z.B. gut gewählte Optimierungsmethode, Unsicherheitsbestimmung (z.B. Konfidenzwerte) – Detektionsmethoden und Abfangen von Fehlern in Checks auf Daten, Modellebene oder auf den Ausgaben einschließlich sinnvoller Schwellenwerte. Die ausgewählten Metriken und Tests sollten möglichst die folgenden Merkmale aufweisen: – wenn durch die Auswahl begründet, sollten die Metriken und Tests sowohl Basismethoden (simple Metrik) und einige fortgeschrittene Methoden (hinsichtlich Informationsgehalt, Aussagekräftigkeit, Implementierungsaufwand, z.B. umfangreiches Prüfwerkzeug - siehe zusätzliche Information) umfassen – ggf. Festlegung von Schwellenwerten (als Mindestanforderungen an die Funktionalität/Leistung des gesamten Systems) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks	Techn. Maßnahmen – Metriken & Schwellenwerte Durch den Anwendungsbereich begründete Auswahl und Anzahl an Metriken und Tests. Dies schließt mindestens eine Methode ein: – Methoden, die die Verlässlichkeit der KI-Komponenten fördern/sicherstellen, z.B. gut gewählte Optimierungsmethode, Unsicherheitsbestimmung (z.B. Konfidenzwerte) – Detektionsmethoden und Abfangen von Fehlern in Checks auf Daten, Modellebene oder auf den Ausgaben einschließlich sinnvoller Schwellenwerte. Die ausgewählten Metriken und Tests sollten möglichst die folgenden Merkmale aufweisen: – es genügen Basismethoden (hinsichtlich Informationsgehalt, Aussagekräftigkeit, Implementierungsaufwand, z.B. eine simple Metrik - siehe zusätzliche Information) – ggf. Festlegung von Schwellenwerten (als Mindestanforderungen an die Funktionalität/Leistung des gesamten Systems) und Begründung der Schwellenwerte unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks	Es wurden keine technischen Methoden oder Metriken implementiert um das Risiko von Fehlfunktionen und Ausfall zu mindern.	Unterscheidung zwischen Basis-methoden und fortgeschrittene Methoden: – In der Prüfmethodensammlung sind einige gängige Methoden gelistet und in Basis- und fortgeschrittene Methoden kategorisiert – Selbst entwickelte Testmethoden werden als fortgeschrittene Methoden anerkannt Die ausgewählten Metriken und Tests sollten wenn möglich einen hohen Grad der Automatisierung erlauben.	System	Analyse	Normal	
VE2.4	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Es müssen geeignete organisatorische Maßnahmen zur Mitigation des Risikos einer Fehlfunktion oder eines Ausfalls des KI-System definiert werden.	Orga. Maßnahmen – Governance – Planung von, oder, wo möglich, Einbindung von organisatorischen oder technischen Maßnahmen zur Mitigation von Fehlern in ein Management System, insbesondere hinsichtlich menschlich bedingter systematischer Fehler und fehlerhafter Handlungen während des Lebenszyklus – Planung von Maßnahmen zur Mitigation der Auswirkung zweiter Ordnung auf Stakeholder, einschließlich Kommunikation – Einrichtung eines Incident Response Kanals für zukünftige Kunden, über den sofort (ohne schuldhaftes Verzögerung) weitere Mitigationsmaßnahmen ergriffen werden können. – die Mitigationsmaßnahmen sollten Teil einer definierten Strategie sein, die in ein fortlaufendes Governancesystem zum Management von Sicherheitsrisiken eingebunden werden kann	Orga. Maßnahmen – Governance – Planung von, oder, wo möglich, Einbindung von organisatorischen oder technischen Maßnahmen zur Mitigation von Fehlern in ein Management System, insbesondere hinsichtlich menschlich bedingter systematischer Fehler fehlerhafter Handlungen während des Lebenszyklus – die Mitigationsmaßnahmen sollten Teil einer definierten Strategie sein, die in ein fortlaufendes Governancesystem zum Management von Sicherheitsrisiken eingebunden werden kann	Orga. Maßnahmen – Governance – Planung von organisatorischen oder technischen Maßnahmen zur Mitigation zufälliger Fehler in ein Management System, insbesondere hinsichtlich menschlich bedingter systematischer Fehler und fehlerhafter Handlungen während des Lebenszyklus	Es wurden keine organisatorischen Methoden vorgesehen, um das Risiko von Fehlfunktionen und Ausfall zu mindern.		System	Analyse	Normal	

Verlässlichkeit									
VE2.5	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Im Rahmen der beabsichtigen Funktionsweise und unter Berücksichtigung des Anwendungs-bereichs müssen die definierten Maßnahmen vor dem Betrieb getestet werden.	Tech. Maßnahme – Test Die in VE2.2-VE2.4 definierten Mitigationsmaßnahmen und -strategien müssen soweit möglich unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks getestet und bewertet werden. Dies beinhaltet: – Testen und Bewerten von Monitoring und Logging (siehe auch MA2.3 und TR1.7) hinsichtlich der in VE2.1 identifizierten Risiken, insbesondere in Bezug auf fehlerhafter Eingaben, falscher Anwendungsumgebungen, externer Änderungen und Extremfälle – Testen von Rückfallmechanismen, Fail-safe Modi und Alarmsystemen definiert in VE2.2, mindestens aber muss die Möglichkeit getestet werden solche Systeme an das KI-System anbinden zu können – wenn anwendbar, Prüfen von Metriken und Tests zur Mitigation des Risikos einer Fehlfunktion oder eines Ausfalls des KI-System definiert werden, definiert in VE2.3 – Dokumentation aller Testergebnisse	Tech. Maßnahme – Test Die in VE2.2-VE2.4 definierten Mitigationsmaßnahmen- und Strategien müssen soweit möglich unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks getestet und bewertet werden. Dies beinhaltet: – Testen und Bewerten von Monitoring und Logging (siehe auch MA2.3 und TR1.7) hinsichtlich der in VE2.1 identifizierten Risiken, insbesondere in Bezug auf fehlerhafter Eingaben, falscher Anwendungsumgebungen, externen Änderungen und Extremfälle – wenn anwendbar, Prüfen von Metriken und Tests zur Mitigation des Risikos einer Fehlfunktion oder eines Ausfalls des KI-System definiert werden, definiert in VE2.3 – Dokumentation aller Testergebnisse	Tech. Maßnahme – Test Die in VE2.2-VE2.4 definierten Mitigationsmaßnahmen- und strategien müssen soweit möglich unter Berücksichtigung des Anwendungsbereichs und Verwendungszwecks getestet und bewertet werden. Dies beinhaltet: – Bewerten von Monitoring und Logging (siehe auch MA2.3 und TR1.7) und der Mitigationsmaßnahmen in VE2.2-VE2.4 hinsichtlich der in VE2.1 identifizierten Risiken, insbesondere in Bezug auf fehlerhafter Eingaben, falscher Anwendungsumgebungen, externen Änderungen und Extremfälle – Dokumentation aller Testergebnisse	Es wurden keine Test der technischen Maßnahmen durchgeführt.		System/Komponente	Maßnahme	Maximalwert	
VE2.6	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verknüpfung
Es müssen Maßnahmen und Sicherheitskontrollen vorhanden sein, um fehlerhafte Nutzung oder Missbrauch des KI-Systems zu verhindern.	Siehe: CY1.4					System/Komponente	Maßnahme	Normal	
VE2.7	A	B	C	D	Zusätzliche Informationen	Bezugsebene	Typ	Gewichtung	Verlinkung
Es muss bewertet werden, ob die ergriffenen Maßnahmen die festgestellten Risiken auf ein annehmbares Maß vermindert haben und die Qualität des KI-Systems den gesetzten Zielvorgaben entspricht.	Bewertung – Zusammenfassende Betrachtung der Effekte der Durchführung der technischen und organisatorischen Maßnahmen – Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen – Ermittlung und Beschreibung des Restrisikos nach Durchführung der technischen und organisatorischen Maßnahmen – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Bewertung – Abwägung von Wechselwirkungen der ergriffenen Maßnahmen, und Abwägung der Mitigation der Risiken im Zusammenspiel mit den Risiken anderer Qualitätsdimensionen – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Bewertung – Einschätzung durch eine qualifizierte und autorisierte Person, ob das Restrisiko tolerierbar ist – Begründung der Tolerierbarkeit des Restrisikos	Eine Bewertung wurde nicht durchgeführt.		System	Bewertung	Maximalwert	

4.5 Prozessablauf der Gesamtbewertung

In dieser Anleitung wird der Prozessablauf der Gesamtbewertung im Detail schematisch in fünf Schritten beschrieben (siehe Abbildung 6).



Abbildung 6: Prozessablauf der Gesamtbewertung

Um zu einer Einstufung auf Ebene der Kriterien zu gelangen, werden die individuellen Einstufungen der jeweils zugehörigen Indikatoren über eine Aggregationsfunktion („Rating“) zusammengefasst. Die Aggregation erfolgt auf Basis eines Durchschnitts der gleichgewichteten Indikatoreneinstufungen unter Berücksichtigung einzelner Maximalwertindikatoren. Deren Einstufung stellt aufgrund ihrer Kritikalität für den ermittelten Schutzbedarf für den Durchschnitt einen bindenden Maximalwert dar. Im Ergebnis entsteht für jedes Kriterium eine Einstufung, die durch ein Scoring von A (am besten) bis C (am niedrigsten) bzw. D (nicht vorhanden) den Umfang der ergriffenen Qualitätsmaßnahmen in dem KI-System aufzeigt. Eine detaillierte Beschreibung der Aggregationsfunktion, der Auswahl und Gewichtung durch die Maximalwertindikatoren findet sich in den folgenden Abschnitten.

Schritt 1: Bestimmung der Indikatoreneinstufung durch Observablen

Zuerst werden die Einstufungen der Indikatoren über die Observablen bestimmt und in die vier Einstufungen von D bis A eingeteilt. Dies ist inhaltlich ein großer Prozessschritt und wird in Abbildung 7 konzeptionell beschrieben.

Aggregation der Einstufung von Indikatoren

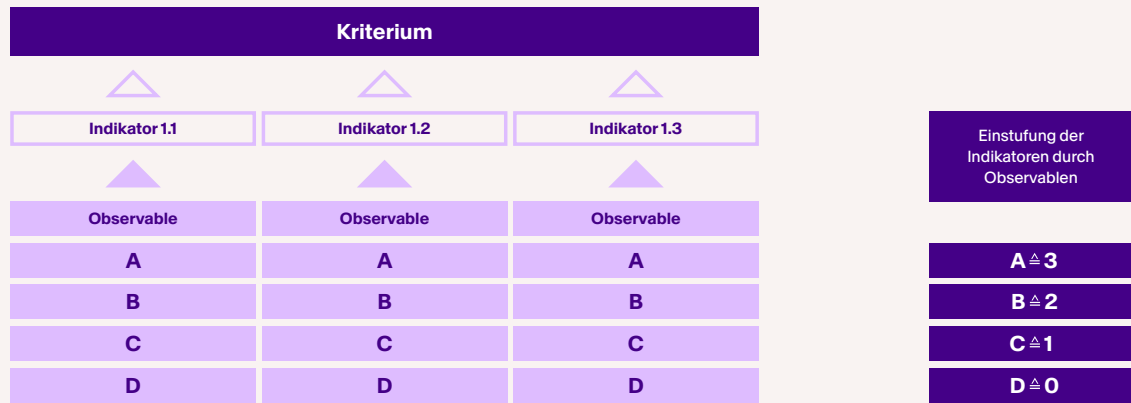


Abbildung 7: Darstellung des grundlegenden Ansatzes zur Aggregation der Einstufung von Indikatoren durch Observablen auf die Kriterienebene.

Schritt 2: Bestimmung des Mittelwerts der Indikatoreneinstufungen

Im nächsten Schritt wird das arithmetische Mittel der Indikatoreneinstufungen bestimmt. Die genaue Formel für die Ermittlung findet sich in Abbildung 9. Die Abbildung 8 zeigt einen beispielhaften Fall mit drei Indikatoren.

Durchschnittsbildung

Kriterium	
Indikator 1.1	B \triangleq 2
Indikator 1.2	C \triangleq 1
Indikator 1.3	A \triangleq 3
\emptyset (Kriterium) = $6/3 = 2$	

Abbildung 8: Beispielhafte Darstellung der Durchschnittsbildung auf Basis der Indikatoreneinstufung.

Schritt 3: Überprüfung der Maximalwertindikatoren

Nachfolgend sind die Maximalwertindikatoren zu überprüfen und gegebenenfalls nach den Beschreibungen in Abbildung 10 anzuwenden, um das Ergebnis der Einstufung zu ändern. Wie Abbildung 8 zeigt, gibt es konzeptionell drei unterschiedliche Möglichkeiten:

1. Der Mittelwert ist größer als der kleinste Maximalwertindikator. Das bedeutet, dass die Kriterieneinstufung gleich dem kleinsten Maximalwertindikator wird.
2. Der Mittelwert ist kleiner oder gleich als der kleinste Maximalwertindikator. In diesem Fall bestimmt der Mittelwert die Kriterieneinstufungen.
3. Es gibt keinen Maximalwertindikator. Auch hier bestimmt der Mittelwert die Kriterieneinstufung.

Detaillierte Beschreibung der Bewertungsaggregation

Formale Berechnung

Im Detail wird die Aggregation von Indikatoren- auf Kriterienebene durchgeführt, indem jeder möglichen Einstufung absteigend eine ganze Zahl zwischen 3 und 0 zugeordnet wird (also A=3, ..., D=0). Dies ermöglicht die Berechnung einer aggregierten Einstufung für jedes Kriterium L_{Krit} durch ein einfaches arithmetisches Mittel über alle zugehörigen Indikatoreinstufungen $L_{\text{Ind}, i}$:

$$L_{\text{Krit}} = \frac{\sum_{i=0}^n L_{\text{Ind}, i}}{n}$$

Hierbei wird normal gerundet, also z. B. führen ein „A“ und ein „B“ auf Indikatorebene zu einer Gesamteinstufung von „A“ ($A=3, B=2 \rightarrow L_{\text{Krit}} = 5/2 = 2,5 \rightarrow 3$). Um zu ermöglichen, dass einzelne Indikatoren besonders kritisch für den Schutzbedarf eines Kriteriums sein können, werden solche Indikatoren für die Aggregation im VCIO als „Extremalwertindikatoren“ kategorisiert. Dies hat zur Folge, dass deren individuelle Einstufung jeweils einen maximalen Wert für die Kriterieneinstufung L_{Krit} vorgibt, d. h. für die Einstufung bei Maximalwertindikator $L_{\text{Ind}, j}^{\text{max}}$ die Berechnung gilt:

$$L_{\text{Krit}} = \min \left(L_{\text{Ind}, j}^{\text{max}}, \frac{\sum_{i=0}^n L_{\text{Ind}, i}}{n} \right)$$

Die Maximalwertindikatoren wurden im Detail nach zwei Prinzipien identifiziert, sie sind in den Prüfanforderungen als solche gekennzeichnet:

- 1 Indikatoren, die als besonders kritisch für die zugeordneten Schutzkategorien der Schutzbedarfsanalyse angesehen werden
- 2 Indikatoren, die sich inhaltlich nicht gut durch die Maßnahmen in anderen Indikatoren ausgleichen lassen, beispielsweise nach einem Prinzip der Komplementarität.

Abbildung 9: Detaillierte Beschreibung der Bewertungsaggregation von Indikatoren- auf Kriterienebene.

Schritt 4: Bestimmung der Kriterieneinstufung

Formal wird im letzten Schritt nach den Ergebnissen aus Schritt 3 eine Einstufung für das ganze Kriterium festgelegt. Siehe dazu Abbildung 10. Die Schritte 1–4 werden für alle Kriterien wiederholt.

Einstufung der Kriterien

Kriterieneinstufung anhand des gerundeten numerischen Wertes	
Gerundete numerische Werte	3 \triangleq A
	2 \triangleq B
	1 \triangleq C
	0 \triangleq D

Abbildung 10: Einstufung der Kriterien anhand des gerundeten numerischen Wertes.

Schritt 5: Gegenüberstellung zur Schutzbedarfsanalyse

Wurde die Einstufung aller Kriterien bestimmt, kann wie oben beschrieben die Gegenüberstellung zu den Schutzeinstufungen der Schutzbedarfsanalyse vorgenommen werden (siehe Abbildung 11).

Das Ergebnis ist je nach Vergleich pro Kriterium entweder das Bestehen oder das Nichtbestehen der Anforderungen des Mindeststandards.

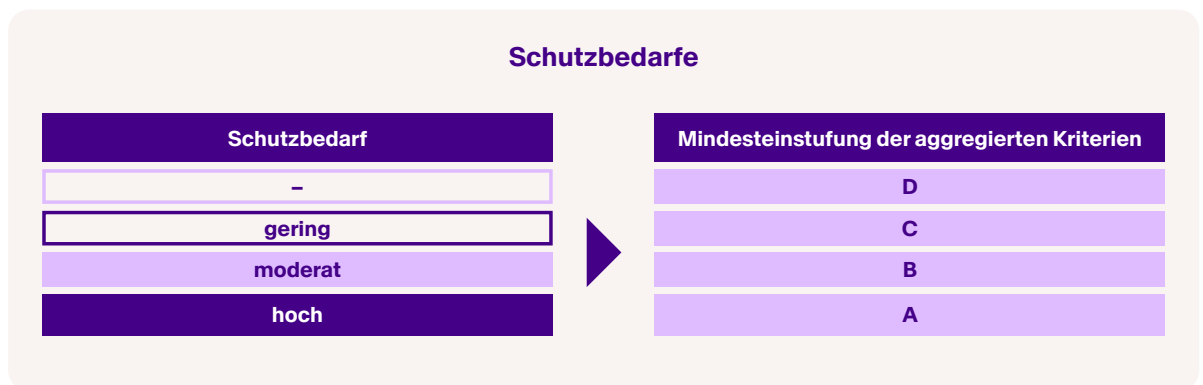


Abbildung 11: Schutzbedarfe als implizite Mindeststufen in der Gesamtbewertung, die die Prüftiefe des Mindeststandards für ein KI-System in mit konkreten Anwendungsfall festlegen.

			Siehe Prüf-kriterien-katalog.	Zur Nutzung der Metrik benötigende (Eingabe-) Daten.	Detailgrad der Metrik.	Erforderliches Wissen und Zugriff auf das Modell.		Automatisierte Prüfung der Metrikergebnisse gegen Grenzwerte.	Aussage-kräftigkeit, Umfang und In-formationsgewinn der Methode.	Verfügbare Frameworks, Tools oder Git-Repos.	Beispiele für die beschreibende Webseiten.
Test- / Methodenname	Kurzbeschreibung	Prüfwerkzeuge	Indi-katoren	Daten-anforderungen	Ergebnis-komplexität	Modellzugriff	Aufgaben-anwendbarkeit	Automatisierungs-möglichkeit	Tiefe der Testmethode	Bekannte Implementationen	Referenz
Accuracy	Misst die Genauigkeit eines (binären) Klassifikators als Verhältnis richtiger Vorhersagen zu Gesamt-vorhersagen	Amazon Sagemaker Suite, Azure ML, FairLearn, What If-Tool, Fairness Indicators, Citadel RADAR, ML Privacy Meter, AI Qualify, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, LangChain OpenEvals, Der HELM Benchmark, GAIA-Benchmarks, Robustness Gym, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html	https://en.wikipedia.org/wiki/Accuracy_and_precision
Accuracy Equality	Genauigkeit zwischen Gruppen		ND1.2 ND1.3 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen		agnostisch	ja	Basismethode		
Adversarial accuracy	Metrik für Robustheit von Entscheidungen	Adversarial-Attacks-PyTorch (Foolbox), AutoAttack, Alpha-Beta-CROWN, AI Qualify, AI4CYBER, IBM Adversarial robustness, Robustness Gym	VE1.3 VE1.4 VE1.5 CY2.2 CY2.3 MA2.3	erweiterte oder neue Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	ja	Basismethode	Empirical: https://github.com/fra31/auto-attack https://github.com/Harry24k/adversarial-attacks-pytorch Formal: https://github.com/Verified-Intelligence/alpha-beta-CROWN	
ALE (Accumulated Local Effects)	Misst die Auswirkungen eines Merkmals auf das vorhergesagte Ergebnis, während der durchschnittliche Effekt anderer Merkmale berücksichtigt wird.		TR2.2 TR2.3	vorhandene Testdaten mit GT	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	wahrscheinlich	Fortgeschrittene Methode		
Alpha-Feature Importance	Anzahl der Features, die notwendig sind, um eine Erklärung mit Güte abhängig von Alpha zu erhalten		TR2.2 TR2.3	vorhandene Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	ja	Basismethode		
Amount of leaked information	Metrik für die Anzahl der Informationen, die ein System durch Angriffe freigibt	Preamble	CY2.2 CY2.3 MA2.3	alle vorhandenen gelabelten Daten	einzelne/mehrere reelle Zahlen	modellunabhängig	agnostisch	nein	Basismethode		https://dl.acm.org/doi/10.1145/1242572.1242598
Anonymity Set Size	Größe der Eingabemenge, die anhand der Modell-ausgaben nicht von einer einzelnen Eingabe x unter-schieden werden kann	Coding-Benchmark	DA1.2 DA1.3	alle vorhandenen gelabelten Daten	einzelne/mehrere reelle Zahlen	Inferenzzugriff	agnostisch	ja	Basismethode		https://link.springer.com/article/10.1007/BF00206326
APPS			VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode		
bAbl	Reasoning-Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode	https://github.com/facebookarchive/bAbl-tasks	
BLEU	Metrik zur Ähnlichkeit von Texten, benutzt zur Evaluierung von Freitextantworten in Benchmarks; in Kombina-tion mit Benchmark zu ver-wenden	Azure Machine Learning, Moonshot, RAGAS, LangChain OpenEvals, Robustness Gym,	VE1.2 VE1.4 VE1.5 MA2.3	erweiterte oder neue Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode	https://github.com/bangoc123/BLEU	https://dl.acm.org/doi/10.3115/1073083.1073135
BoolQ	Question-Answering Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode	https://github.com/google-research-datasets/boolean-questions	https://arxiv.org/abs/1905.10044
Brier Score	Misst die mittlere quadratische Differenz zwischen der vorhergesagten Wahr-schein-lichkeit und dem tatsächlichen Ergebnis	IBM UQ360,Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	agnostisch	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.brier_score_loss.html	https://en.wikipedia.org/wiki/Brier_score
Calibration Error	Unterschied zwischen errechneten Wahr-schein-lichkeiten und Akkuranz	Truera, Zeno, Der HELM Benchmark, IBM UQ360, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	agnostisch	ja	Basismethode	https://torchmetrics.readthedocs.io/en/v0.8.0/classification/calibration_error.html	https://towardsdatascience.com/expected-calibration-error-ece-a-step-by-step-visual-explanation-with-python-code-c3e9aa12937d
CIDeR: Consensus-based Image Description Evaluation	Erfasst den Konsens über die Qualität der von genAI generierten Bildbeschrei-bungen		VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	GenAI	ja	Basismethode	https://github.com/vrama91/cider	https://arxiv.org/abs/1411.5726
CLIP Image Quality Assessment	Messung des visuellen Inhalts von Bildern		VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	GenAI	ja	Basismethode	https://lightning.ai/docs/torchmetrics/stable/multimodal/clip_iqua.html	

			Siehe Prüf-kriterien-katalog.	Zur Nutzung der Metrik benötigende (Eingabe-) Daten.	Detailgrad der Metrik.	Erforderliches Wissen und Zugriff auf das Modell.	Prüfbare Tasks mit der Metrik.	Automatisierte Prüfung der Metrikergebnisse gegen Grenzwerte.	Aussage-kraftigkeit, Umfang und In-formationsgewinn der Methode.	Verfügbare Frameworks, Tools oder Git-Repos.	Beispiele für die beschreibende Webseiten.
Test- / Methodenname	Kurzbeschreibung	Prüfwerkzeuge	Indi-katoren	Daten-anforderungen	Ergebnis-komplexität	Modellzugriff	Aufgaben-anwendbarkeit	Automatisierungs-möglichkeit	Tiefe der Testmethode	Bekannte Implementationen	Referenz
Accuracy	Misst die Genauigkeit eines (binären) Klassifikators als Verhältnis richtiger Vorhersagen zu Gesamt-vorhersagen	Amazon Sagemaker Suite, Azure ML, FairLearn, What If-Tool, Fairness Indicators, Citadel RADAR, ML Privacy Meter, AI Qualify, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, LangChain OpenEvals, Der HELM Benchmark, GAIA-Benchmarks, Robustness Gym, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html	https://en.wikipedia.org/wiki/Accuracy_and_precision
CLIP score	Text-Bild-Ähnlichkeitsmetrik, z. B. Qualität von Gen-AI-Bildern.		VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	GenAI	ja	Basismethode	https://lightning.ai/docs/torchmetrics/stable/multimodal/clip_score.html	https://huggingface.co/docs/diffusers/v0.21.0/conceptual/evaluation
CNN/DailyMail	Zusammenfassungs-Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode	https://huggingface.co/datasets/abisee/cnn_dailymail	
Cohen Kappa score	Ein Maß für die Übereinstimmung zwischen den Kommentatoren.	Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html	https://en.wikipedia.org/wiki/Cohen%27s_kappa
Combinatorial Testing	Metrik für die Abdeckung eines Datensatzes zu bewerten, indem überprüft wird, ob alle relevanten Kombinationen von Merkmalswerten ausreichend im Datensatz vertreten sind.		MA2.3 DA1.2 DA1.3	alle vorhandenen gelabelten Daten	komplexe Ergebnisse	modellunabhängig	agnostisch	wahrscheinlich	Fortgeschrittene Methode		https://openaccess.thecvf.com/content_CVPRW_2020/html/w20/Gladisch_Leveraging_Combinatorial_Testing_for_Safety-Critical_Computer_Vision_Datasets_CVPRW_2020_paper.html
Cosine Similarity	Ein Maß für die Ähnlichkeit zwischen zwei von Null verschiedenen Vektoren, die in einem inneren Produktraum definiert sind.	RAGAS, LangChain OpenEvals, Scikit-Learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	agnostisch	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html	https://en.wikipedia.org/wiki/Cosine_similarity
Counterfactual Explanations	Metrik für die Auswirkungen von Änderungen in den Eingangsmerkmalen auf die Vorhersage durch Generierung von alternativen Szenarien	What If-Tool, FairLearn,Microsoft Responsible AI Dashboard	TR2.2 TR2.3	vorhandene Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	wahrscheinlich	Fortgeschrittene Methode		
Coverage Error	Berechnet, wie weit wir durch Ranglistenpunkte gehen müssen, um alle wahren Bezeichnungen zu erfassen	Revise	ND1.2 ND1.3 MA2.3 DA1.2 DA1.3	vorhandene Testdaten mit GT	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	wahrscheinlich	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.coverage_error.html	
Data completeness (Match Rate)	Vergleich der Einträge eines Datensatzes zu einem Referenzdatensatz	Citadel RADAR, Citadel Lens, Google ML Test Score, ScrutinAI	DA1.2 DA1.3 VE2.3 VE2.5	alle vorhandenen gelabelten Daten	einzelne/mehrere reelle Zahlen	modellunabhängig	agnostisch	ja	Basismethode		
Data consistency	Prüfen, ob alle Daten aus verschiedenen Quellen im selben Format vorliegen und ob Widersprüche zwischen Quellen auftreten	System Center Data Protection Manager (DPM), Preamble, AI4CYBER ,Google ML Test Score, ScrutinAI	VE1.6 VE2.3 VE2.5 CY2.3 MA2.3 DA1.2 DA1.3	alle vorhandenen gelabelten Daten	einzelne/mehrere reelle Zahlen	modellunabhängig	agnostisch	ja	Basismethode		
Data Coverage (Match Rate)	Maß für die Abdeckung aller als relevant befundenen Eigenschaften durch die Daten	System Center Data Protection Manager (DPM), AI4CYBER, Google ML Test Score	VE1.4 VE1.5 VE2.3 VE2.5 MA2.3 DA1.2 DA1.3	alle vorhandenen gelabelten Daten	einzelne/mehrere reelle Zahlen	modellunabhängig	agnostisch	ja	Basismethode		
Data imputation	Benchmark für Datenvervollständigung		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	GenAI	ja	Basismethode		
Data timeliness	Check, ob alle Daten aktuell sind	System Center Data Protection Manager (DPM), Preamble, Google ML Test Score	VE1.6 VE2.3 VE2.5 CY2.3 MA2.3 DA1.2 DA1.3	alle vorhandenen gelabelten Daten	einzelne/mehrere reelle Zahlen	modellunabhängig	agnostisch	ja	Basismethode		

Technische Prüfmethodensammlung											
Dieses Dokument ist Teil des MISSION KI Qualitätsstandards. ©acatech – Deutsche Akademie der Technikwissenschaften e.V. Dieses Werk ist lizenziert unter der Creative Commons Lizenz Namensnennung – Keine Bearbeitungen 4.0 International (CC BY-ND 4.0). https://creativecommons.org/licenses/by-nd/4.0/deed.de											
			Siehe Prüf-kriterien-katalog.	Zur Nutzung der Metrik benötigende (Eingabe-) Daten.	Detailgrad der Metrik.	Erforderliches Wissen und Zugriff auf das Modell.	Prüfbare Tasks mit der Metrik.	Automatisierte Prüfung der Metrikergebnisse gegen Grenzwerte.	Aussage-kräftigkeit, Umfang und In-formationsgewinn der Methode.	Verfügbare Frameworks, Tools oder Git-Repos.	Beispiele für die beschreibende Webseiten.
Test- / Methodenname	Kurzbeschreibung	Prüfwerkzeuge	Indi-katoren	Daten-anforderungen	Ergebnis-komplexität	Modellzugriff	Aufgaben-anwendbarkeit	Automatisierungs-möglichkeit	Tiefe der Testmethode	Bekannte Implementationen	Referenz
Accuracy	Misst die Genauigkeit eines (binären) Klassifikators als Verhältnis richtiger Vorhersagen zu Gesamt-vorhersagen	Amazon Sagemaker Suite, Azure ML, FairLearn, What If-Tool, Fairness Indicators, Citadel RADAR, ML Privacy Meter, AI Qualify, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, LangChain OpenEvals, Der HELM Benchmark, GAIA-Benchmarks, Robustness Gym, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html	https://en.wikipedia.org/wiki/Accuracy_and_precision
Data Uniqueness	Metrik für das Auftreten von Duplikaten in einem Datensatz	Citadel RADAR, Citadel Lens, Google ML Test Score	VE2.3 VE2.5 DA1.2 DA1.3	alle vorhandenen gelabelten Daten	einzelne/mehrere reelle Zahlen	modellunabhängig	agnostisch	ja	Basismethode		
Data validity	Check, ob alle Daten im benötigten Format vorliegen	Citadel RADAR, Citadel Lens, System Center Data Protection Manager (DPM), AI4CYBER, Google ML Test Score, ScrutinAI	VE1.6 VE2.3 VE2.5 CY2.3 MA2.3 DA1.2 DA1.3	alle vorhandenen gelabelten Daten	einzelne/mehrere reelle Zahlen	modellunabhängig	agnostisch	ja	Basismethode		
DICE Score	Die Ähnlichkeit zwischen einer vorhergesagten Segmentierungsmaske und der wahren Segmentierungsmaske bewerten	ScrutinAI	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Segmentierung	ja	Basismethode	https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.dice.html	https://en.wikipedia.org/wiki/Dice-S%C3%B8rensen_coefficient
Domain integrity	Check, ob alle Daten in einem sinnvollen vorgegebenen Bereich liegen	Preamble	VE1.6 VE2.3 VE2.5 CY2.3 MA2.3 DA1.2 DA1.3	alle vorhandenen gelabelten Daten	einzelne/mehrere reelle Zahlen	modellunabhängig	agnostisch	ja	Basismethode		
Domino	Suche nach schwachen Abschnitten von unstrukturierten Daten, die im Einbettungsraum näher beieinander liegen und bei denen DNN eine geringe Leistung aufweist		VE1.3 VE1.4 VE1.5	vorhandene Testdaten mit GT	komplexe Ergebnisse	Fester Testsatz (nur Vorhersagen)	agnostisch	wahrscheinlich	Fortgeschrittene Methode		https://arxiv.org/abs/2203.14960
Dyck	Reasoning-Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode		
Entity – and referential Integrity	Check für Eindeutigkeit von Referenzen in relationalen Datenbanken	Preamble	VE1.6 CY2.3 MA2.3 DA1.2 DA1.3	alle vorhandenen gelabelten Daten	einzelne/mehrere reelle Zahlen	modellunabhängig	agnostisch	ja	Basismethode		
Equal Opportunity	Maß der True-positive-Rate zwischen Gruppen	Astraea, Revise, Microsoft Responsible AI Dashboard, AIF360	ND1.2 ND1.3 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	agnostisch	ja	Basismethode		
Equalized Odds	True-Positivrate und False-Positiverate zwischen Gruppen	AIF 360, FairLearn, FairTest, FairDM / FairSight, Fairness Indicators, Robuscope,Microsoft Responsible AI Dashboard, AIF360, CheckList	ND1.2 ND1.3 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	agnostisch	ja	Basismethode	https://aif360.readthedocs.io/en/stable/modules/generated/aif360.sklearn.postprocessing.CalibratedEqualizedOdds.html	https://en.wikipedia.org/wiki/Equalized_odds
Error rate balance	Fehlerrate zwischen Gruppen	NeMo Guardrails, AI Qualify, AIF360, Robustness Gym	ND1.2 ND1.3 MA2.3 DA1.2 DA1.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	agnostisch	ja	Basismethode		
Error ratio	Fehler zwischen Gruppen	AIF360	ND1.2 ND1.3 MA2.3 DA1.2 DA1.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	agnostisch	ja	Basismethode		
Explainability Ease Score	Metrik für die Komplexität der Input-Output-Beziehung	Truera, LIME, SHAP, AI4CYBER, ScrutinAI	TR2.2 TR2.3	vorhandene Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	ja	Fortgeschrittene Methode		

			Siehe Prüf-kriterien-katalog.	Zur Nutzung der Metrik benötigende (Eingabe-) Daten.	Detailgrad der Metrik.	Erforderliches Wissen und Zugriff auf das Modell.		Automatisierte Prüfung der Metrikergebnisse gegen Grenzwerte.	Aussage-kräftigkeit, Umfang und In-formationsgewinn der Methode.	Verfügbare Frameworks, Tools oder Git-Repos.	Beispiele für die beschreibende Webseiten.
Test- / Methodenname	Kurzbeschreibung	Prüfwerkzeuge	Indi-katoren	Daten-anforderungen	Ergebnis-komplexität	Modellzugriff	Aufgaben-anwendbarkeit	Automatisierungs-möglichkeit	Tiefe der Testmethode	Bekannte Implementationen	Referenz
Accuracy	Misst die Genauigkeit eines (binären) Klassifikators als Verhältnis richtiger Vorhersagen zu Gesamt-vorhersagen	Amazon Sagemaker Suite, Azure ML, FairLearn, What If-Tool, Fairness Indicators, Citadel RADAR, ML Privacy Meter, AI Qualify, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, LangChain OpenEvals, Der HELM Benchmark, GAIA-Benchmarks, Robustness Gym, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html	https://en.wikipedia.org/wiki/Accuracy_and_precision
F1-Score	Harmonisches Mittel aus Precision und Recall	Amazon Sagemaker Suite, Azure ML, FairLearn, What If-Tool, Fairness Indicators, Citadel RADAR, Robuscope, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, Der HELM Benchmark, Robustness Gym, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html	https://en.wikipedia.org/wiki/F-score
False Negative Rate		Azure Machine Learning, ML Privacy Meter, TruEra, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html	https://en.wikipedia.org/wiki/False_positives_and_false_negatives#False_positives_and_false_negative_rates
False Omission Rate			VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html	https://en.wikipedia.org/wiki/Positive_and_negative_predictive_values
False Positive Rate		Azure Machine Learning, ML Privacy Meter, TruEra, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html	https://en.wikipedia.org/wiki/False_positive_rate
Feature importance spread	Maß für Streuung der wichtigen Features für einen gegebenen Output	Truera, LIME, SHAP, ELI5,ScrutinAI	TR2.2 TR2.3	vorhandene Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	ja	Basismethode		
Feature importance stability	Varianz der Feature-Importance	LIME, ELI5	TR2.2 TR2.3	vorhandene Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	ja	Basismethode		
Fraction of toxic output	Misst den Anteil toxischen Outputs	Citadel Lens, NeMo Guardrails, Llama Guard 3-8B, Guardrails AI, Der HELM Benchmark	ND1.2 ND1.3 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	GenAI	ja	Fortgeschrittene Methode	https://developers.perspectiveapi.com/s/about-the-api?language=en_US	https://arxiv.org/abs/2106.10328
Fréchet Inception Distance (FID)	Misst die Qualität generativer Bildmodelle.		VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	GenAI	ja	Basismethode	https://pytorch.org/ignite/generated/ignite.metrics.FID.html	https://en.wikipedia.org/wiki/Fr%C3%A9chet_inception_distance
Fuzzy Testing	Fuzzy Testing ist ein Ansatz, der unsichere oder ungenaue Eingabedaten nutzt, um die Robustheit und Leistungsfähigkeit von maschinellen Lernmodellen unter variierenden Bedingungen zu überprüfen.	Astraea	VE1.2 VE1.3 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	vollständiger Modellzugang	agnostisch	ja	Fortgeschrittene Methode		
Gender-based Illicit Proximity Estimate	Embedding-basierte sprachliche Diskriminierungs-detektion		ND1.2 ND1.3 MA2.3	vorhandene Testdaten mit GT	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	wahrscheinlich	Basismethode	https://github.com/vaibkumr/RAN-Debias	https://arxiv.org/abs/2006.01938
Grad-CAM	Erstellt Heatmaps, die die Bereiche eines Bildes hervor-heben, die zu den Vorhersagen eines neuronalen Netzwerks beitragen.	pytorch-grad-cam, Captum, Robuscope	TR2.2 TR2.3	vorhandene Testdaten mit GT und Input	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	wahrscheinlich	Fortgeschrittene Methode		
GSM8K	Mathematik-Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode	https://huggingface.co/datasets/openai/gsm8k	
HellaSwag	Question-Answering Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode	https://rowanzellers.com/hellaswag/	
Homogeneity Score	Homogenitätsmetrik einer Clusterbeschriftung bei gegebener Grundwahrheit.	Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Clustering	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.homogeneity_score.html	
HOTA (Higher Order Tracking Accuracy)	Einheitliche Metrik für den Vergleich von Trackern in Bezug auf die Genauigkeit der Erkennung, Zuordnung und Lokalisierung.		VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	komplexe Ergebnisse	Fester Testsatz (nur Vorhersagen)	Erkennung	ja	Basismethode	https://github.com/JonathonLuiten/TrackEval	https://autonomousvision.github.io/hota-metrics/

Technische Prüfmethodensammlung

Dieses Dokument ist Teil des MISSION KI Qualitätsstandards. ©acatech – Deutsche Akademie der Technikwissenschaften e.V.
Dieses Werk ist lizenziert unter der Creative Commons Lizenz Namensnennung – Keine Bearbeitungen 4.0 International (CC BY-ND 4.0). <https://creativecommons.org/licenses/by-nd/4.0/deed.de>

			Siehe Prüf-kriterien-katalog.	Zur Nutzung der Metrik benötigende (Eingabe-) Daten.	Detailgrad der Metrik.	Erforderliches Wissen und Zugriff auf das Modell.	Prüfbare Tasks mit der Metrik.	Automatisierte Prüfung der Metrikergebnisse gegen Grenzwerte.	Aussage-kräftigkeit, Umfang und In-formationsgewinn der Methode.	Verfügbare Frameworks, Tools oder Git-Repos.	Beispiele für die beschreibende Webseiten.
Test- / Methodenname	Kurzbeschreibung	Prüfwerkzeuge	Indi-katoren	Daten-anforderungen	Ergebnis-komplexität	Modellzugriff	Aufgaben-anwendbarkeit	Automatisierungs-möglichkeit	Tiefe der Testmethode	Bekannte Implementationen	Referenz
Accuracy	Misst die Genauigkeit eines (binären) Klassifikators als Verhältnis richtiger Vorhersagen zu Gesamt-vorhersagen	Amazon Sagemaker Suite, Azure ML, FairLearn, What If-Tool, Fairness Indicators, Citadel RADAR, ML Privacy Meter, AI Qualify, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, LangChain OpenEvals, Der HELM Benchmark, GAIA-Benchmarks, Robustness Gym, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html	https://en.wikipedia.org/wiki/Accuracy_and_precision
HumanEval	Coding-Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode		
IMDB	Sentiment-Analyse-Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode	https://huggingface.co/datasets/stanfordnlp/imdb	
Inception Score	Misst die Realitätsnähe erzeugter Bilder		VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	GenAI	ja	Basismethode	https://lightning.ai/docs/torchmetrics/stable/image/inception_score.html	https://en.wikipedia.org/wiki/Inception_score
Integrated Gradients	Metrik/Visualisierung für Zusammenhang zwischen Inputfeatur und Vorhersage	Captum, Amazon Sagemaker Suite, Azure ML	TR2.2 TR2.3	vorhandene Testdaten mit GT und Input	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	wahrscheinlich	Fortgeschrittene Methode		
IoU	Intersection over Union, Maß für das Zusammenfallen zweier Mengen (bspw. Segmentierungspixel).	ScrutinAI	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Segmentierung	ja	Basismethode	https://lightning.ai/docs/torchmetrics/stable/detection/intersection_over_union.html	
Kolmogorov-Smirnov Test for Drift Distribution k-Projection Coverage	Test, ob zwei Datensätze derselben Verteilung entsprechen Misst, wie gut ein generierter Datensatz die Verteilung eines Referenzdatensatzes in mehreren zufälligen Teil-räumen abdeckt.	Robuscope	VE1.6 MA2.3	alle vorhandenen gelabelten Daten	einzelne/mehrere reelle Zahlen	modellunabhängig	agnostisch	ja	Basismethode		
			MA2.3 DA1.2 DA1.3	alle vorhandenen gelabelten Daten	einzelne/mehrere reelle Zahlen	modellunabhängig	agnostisch	wahrscheinlich	Fortgeschrittene Methode		https://arxiv.org/pdf/1805.04333
LegalSupport	Juristischer Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode		
LIME	Lokales, interpretierbares Modell um eine spezifische Ausgabe	LIME, Captum, Amazon Sagemaker Suite, Azure ML	TR2.2 TR2.3	vorhandene Testdaten mit GT und Input	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	nein	Fortgeschrittene Methode		https://arxiv.org/abs/1602.04938
LORE	Lokale, regelbasierte Erklärungen mit Black-Box-Zugriff		TR2.2 TR2.3	vorhandene Testdaten mit GT und Input	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	nein	Fortgeschrittene Methode		
LRP	"Erklärbare KI"-Methode, die darauf abzielt, den Entscheidungsfindungsprozess von Modellen zu interpretieren, indem sie den Eingabemerkmalen Wichtigkeitspunkte zuweist.		TR2.2 TR2.3	vorhandene Testdaten mit GT und Input	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	wahrscheinlich	Fortgeschrittene Methode		https://www.tensorflow.org/tutorials/interpretability/integrated_gradients
LSAT	Juristischer Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode	https://github.com/zhongwanjun/AR-LSAT	
MAE	Mittlerer absoluter Fehler	Azure Machine Learning, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Regression	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html	https://en.wikipedia.org/wiki/Mean_absolute_error
Mahalanobis Distance	Abstands zwischen einer Probe und einer Verteilung	Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	agnostisch	ja	Basismethode	https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.mahalanobis.html	https://en.wikipedia.org/wiki/Mahalanobis_distance
mAP	Mittlere durchschnittliche Precision	ScrutinAI	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Erkennung	ja	Basismethode	https://lightning.ai/docs/torchmetrics/stable/detection/mean_average_precision.html	
MAPE	Prozentualer absoluter Fehler	DeepChecks, Zeno	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Regression	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_percentage_error.html	https://en.wikipedia.org/wiki/Mean_absolute_percentage_error

			Siehe Prüf-kriterien-katalog.	Zur Nutzung der Metrik benötigende (Eingabe-) Daten.	Detailgrad der Metrik.	Erforderliches Wissen und Zugriff auf das Modell.	Prüfbare Tasks mit der Metrik.	Automatisierte Prüfung der Metrikergebnisse gegen Grenzwerte.	Aussage-kräftigkeit, Umfang und In-formationsgewinn der Methode.	Verfügbare Frameworks, Tools oder Git-Repos.	Beispiele für die beschreibende Webseiten.
Test- / Methodenname	Kurzbeschreibung	Prüfwerkzeuge	Indi-katoren	Daten-anforderungen	Ergebnis-komplexität	Modellzugriff	Aufgaben-anwendbarkeit	Automatisierungs-möglichkeit	Tiefe der Testmethode	Bekannte Implementationen	Referenz
Accuracy	Misst die Genauigkeit eines (binären) Klassifikators als Verhältnis richtiger Vorhersagen zu Gesamt-vorhersagen	Amazon Sagemaker Suite, Azure ML, FairLearn, What If-Tool, Fairness Indicators, Citadel RADAR, ML Privacy Meter, AI Qualify, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, LangChain OpenEvals, Der HELM Benchmark, GAIA-Benchmarks, Robustness Gym, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html	https://en.wikipedia.org/wiki/Accuracy_and_precision
MATH	Mathematik-Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode		
Matthews correlation coefficient		TruEra	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.matthews_corrcoef.html	https://en.wikipedia.org/wiki/Phi_coefficient
Maximum Mean Discrepancy	Metrik für die Übereinstimmung der Verteilung eines Datensatzes mit einer referenzierten Datensatzbeschreibung	Robuscope	VE1.6 MA2.3 DA1.2 DA1.3	alle vorhandenen gelabelten Daten	einzelne/mehrere reelle Zahlen	modellunabhängig	agnostisch	wahrscheinlich	Fortgeschrittene Methode		https://jmlr.org/papers/volume13/gretton12a/gretton12a.pdf
Minimum Distortion	Metrik für adversale Attacken	AI Qualify, IBM Adversarial robustness, Robustness Gym	VE1.3 VE1.4 VE1.5 CY2.2 CY2.3 MA2.3	erweiterte oder neue Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	ja	Basismethode		
mIoU	Über einen Datensatz gemittelte IoU Werte.	ScrutinAI	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Segmentierung	ja	Basismethode		
MMLU	Question-Answering Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode	https://github.com/hendrycks/test	https://en.wikipedia.org/wiki/MMLU
Monte Carlo Dropout	Methode für Unsicherheits-schätzung bei neuronalen Netzen		VE1.2 VE1.3 VE1.4 VE1.5 VE2.5 MA2.3	vorhandene Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	vollständiger Modellzugang	agnostisch	ja	Fortgeschrittene Methode		
Mutual information score	Maß für die gegenseitige Abhängigkeit zwischen den beiden Variablen	Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Clustering	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mutual_info_score.html	https://en.wikipedia.org/wiki/Mutual_information
NarrativeQA	Question-Answering Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode	https://huggingface.co/datasets/deepmind/narrativeqa	
NaturalQuestions	Question-Answering Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode	https://github.com/google-research-datasets/natural-questions	
Neuron Coverage	Metrik zur Bewertung der Testabdeckung von neuronalen Netzwerken, die misst, wie viele Neuronen in einem Netzwerk während der Tests aktiviert wurden, um die Funktionsweise des Modells zu verstehen und zu verbessern	DeepChecks	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	vollständiger Modellzugang	agnostisch	ja	Fortgeschrittene Methode		
OpenBookQA	Question-Answering Benchmark	Citadel Lens	VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode	https://huggingface.co/datasets/allenai/openbookqa	https://arxiv.org/abs/1809.02789
Out-of-distribution (OOD) generalization	Evaluierung der Fähigkeiten des Modells auf ungesehenen Daten durch neue synthetische oder reale Testdaten	Astraea, AI Qualify	VE1.2 VE1.4 VE1.5 MA2.3	erweiterte oder neue Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	wahrscheinlich	Basismethode		[2403.01874] A Survey on Evaluation of Out-of-Distribution Generalization (arxiv.org)

Technische Prüfmethodensammlung											
Dieses Dokument ist Teil des MISSION KI Qualitätsstandards. ©acatech – Deutsche Akademie der Technikwissenschaften e.V. Dieses Werk ist lizenziert unter der Creative Commons Lizenz Namensnennung – Keine Bearbeitungen 4.0 International (CC BY-ND 4.0). https://creativecommons.org/licenses/by-nd/4.0/deed.de											
			Siehe Prüf-kriterien-katalog.	Zur Nutzung der Metrik benötigende (Eingabe-) Daten.	Detailgrad der Metrik.	Erforderliches Wissen und Zugriff auf das Modell.			Aussage-kräftigkeit, Umfang und In-formationsgewinn der Methode.		
Test- / Methodenname	Kurzbeschreibung	Prüfwerkzeuge	Indi-katoren	Daten-anforderungen	Ergebnis-komplexität	Modellzugriff	Aufgaben-anwendbarkeit	Automatisierungs-möglichkeit	Tiefe der Testmethode	Bekannte Implementationen	Beispiele für die beschreibende Webseiten.
Accuracy	Misst die Genauigkeit eines (binären) Klassifikators als Verhältnis richtiger Vorhersagen zu Gesamt-vorhersagen	Amazon Sagemaker Suite, Azure ML, FairLearn, What If-Tool, Fairness Indicators, Citadel RADAR, ML Privacy Meter, AI Quality, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, LangChain OpenEvals, Der HELM Benchmark, GAIA-Benchmarks, Robustness Gym, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html	https://en.wikipedia.org/wiki/Accuracy_and_precision
Page-Hinkley Test	Metrik für Änderungen des Mittelwertes einer Zeitreihe, verwendet für Drift Detection	Robuscope	VE1.6 MA2.3	alle vorhandenen gelabelten Daten	einzelne/mehrere reelle Zahlen	modellunabhängig	agnostisch	ja	Basismethode		
Panoptic Quality	Eine Metrik, die Segmentierungsqualität und Erkennungsqualität kombiniert.		VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Segmentierung	ja	Basismethode	https://lightning.ai/docs/torchmetrics/stable/detection/panoptic_quality.html	
Perceptual evaluation of Speech Quality (PESQ)	Anerkannter Industriestandard für Audioqualität, der verschiedene Merkmale berücksichtigt		VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Audio	ja	Basismethode	https://lightning.ai/docs/torchmetrics/stable/audio/perceptual_evaluation_speech_quality.html	https://en.wikipedia.org/wiki/Perceptual_Evaluation_of_Speech_Quality
Population Stability Index	Vergleich der Distribution in einer Variable über zwei Datensätze, für Drift Detection		VE1.6 MA2.3	alle vorhandenen gelabelten Daten	einzelne/mehrere reelle Zahlen	modellunabhängig	agnostisch	ja	Basismethode		
Precision	Anzahl der korrekt wiedererkannten Positivfälle (gemessen an der Gesamtzahl der positiven Vorhersagen)	Amazon Sagemaker Suite, Azure ML, FairLearn, What If-Tool, Fairness Indicators, Citadel RADAR, Robuscope, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, RAGAS, GAIA-Benchmarks, Robustness Gym, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html	https://en.wikipedia.org/wiki/Precision_and_recall
Predictions Groups Contrast	Differenz in Features bei der Erklärung einer Untergruppe zum Gesamtdurchschnitt		TR2.2 TR2.3	vorhandene Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	ja	Fortgeschrittene Methode		
Predictive Rate Parity	Maß der Genauigkeitsrate positiver Vorhersagen zwischen Gruppen	AIF 360, FairLearn, FairTest, FairDM / FairSight, Fairness Indicators, Microsoft Responsible AI Dashboard, AIF360	ND1.2 ND1.3 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	agnostisch	ja	Basismethode		
QuAC	Question-Answering Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode	https://quac.ai/	
R2	Maß der Varianz in der abhängigen Variable, die durch eine unabhängige Variable erklärt wird	Amazon SageMaker Suite, Azure Machine Learning, Deepchecks, Zeno, Microsoft Responsible AI Dashboard, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Regression	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html	https://en.wikipedia.org/wiki/Coefficient_of_determination
RAFT	Textklassifikation		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode	https://huggingface.co/datasets/ought/raft	
Recall	Anzahl der korrekt wieder-erkannten Positivfälle (gemessen an der Gesamtzahl der tatsächlichen Positivfälle)	Amazon Sagemaker Suite, Azure ML, FairLearn, What If-Tool, Fairness Indicators, Citadel RADAR, Robuscope, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, RAGAS, Robustness Gym, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html	https://en.wikipedia.org/wiki/Precision_and_recall
RMSE	Root mean square error	Amazon SageMaker Suite, Azure Machine Learning, Zeno, Microsoft Responsible AI Dashboard, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Regression	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html	https://en.wikipedia.org/wiki/Root_mean_square_deviation
Robustness radius	Metrik für Robustheit von Entscheidungen	Adversarial-Attacks-PyTorch (Foolbox), Alpha-Beta-CROWN, AI Qualify, AI4CYBER, IBM Adversarial robustness	VE1.3 VE1.4 VE1.5 CY1.5 CY1.6	erweiterte oder neue Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	ja	Fortgeschrittene Methode		
ROC-AUC	Receiver-Operator-Curve	Amazon Sagemaker Suite, Azure ML, FairLearn, What If-Tool, Fairness Indicators, Citadel RADAR, ML Privacy Meter, Robuscope, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	komplexe Ergebnisse	Fester Testsatz (nur Vorhersagen)	Klassifizierung	wahrscheinlich	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html	https://en.wikipedia.org/wiki/Receiver_operating_characteristic

			Siehe Prüf-kriterien-katalog.	Zur Nutzung der Metrik benötigende (Eingabe-) Daten.	Detailgrad der Metrik.	Erforderliches Wissen und Zugriff auf das Modell.		Automatisierte Prüfung der Metrikergebnisse gegen Grenzwerte.	Aussage-kräftigkeit, Umfang und In-formationsgewinn der Methode.	Verfügbare Frameworks, Tools oder Git-Repos.	Beispiele für die beschreibende Webseiten.
Test- / Methodenname	Kurzbeschreibung	Prüfwerkzeuge	Indi-katoren	Daten-anforderungen	Ergebnis-komplexität	Modellzugriff	Aufgaben-anwendbarkeit	Automatisierungs-möglichkeit	Tiefe der Testmethode	Bekannte Implementationen	Referenz
Accuracy	Misst die Genauigkeit eines (binären) Klassifikators als Verhältnis richtiger Vorhersagen zu Gesamt-vorhersagen	Amazon Sagemaker Suite, Azure ML, FairLearn, What If-Tool, Fairness Indicators, Citadel RADAR, ML Privacy Meter, AI Qualify, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, LangChain OpenEvals, Der HELM Benchmark, GAIA-Benchmarks, Robustness Gym, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html	https://en.wikipedia.org/wiki/Accuracy_and_precision
ROUGE	Metrik zur Ähnlichkeit von Texten, benutzt zur Evaluierung von Freitextantworten in Benchmarks, in Kombination mit Benchmark zu verwenden	Azure Machine Learning, Moonshot, RAGAS, LangChain OpenEvals, HELM Benchmark, Robustness Gym	VE1.2 VE1.4 VE1.5 MA2.3	erweiterte oder neue Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI		Basismethode		
SHAP	Bietet konsistente und interpretierbare Merkmale zur Erklärung von Modellvorhersagen basierend auf Shapley-Werten.	SHAP, Captum, Amazon Sagemaker Suite, Azure ML, Robuscope, Microsoft Responsible AI Dashboard	TR2.2 TR2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	wahrscheinlich	Fortgeschrittene Methode		
Short-Time Objective Intelligibility (STOI)	Bewertung von Sprachsignalen anhand eines Verständlichkeitsmaßes, das in hohem Maße mit der Verständlichkeit von verschlechterten Sprachsignalen korreliert		VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Audio	ja	Basismethode	https://lightning.ai/docs/torchmetrics/stable/audio/short_time_objective_intelligibility.html	
Signal-to-Noise Ratio (SNR)	Ein Maß, dass das Level eines gewünschten Signals mit dem Level des Hintergrundrauschens vergleicht.		VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Audio	ja	Basismethode	https://lightning.ai/docs/torchmetrics/stable/audio/signal_noise_ratio.html	https://en.wikipedia.org/wiki/Signal-to-noise_ratio
Silhouette Score	Mittelwert des Maßes, das angibt, wie ähnlich ein Objekt seinem eigenen Cluster (Kohäsion) im Vergleich zu anderen Clustern (Separation) ist.	TrojanZoo,Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	komplexe Ergebnisse	Fester Testsatz (nur Vorhersagen)	Clustering	nein	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html	https://en.wikipedia.org/wiki/Silhouette_(clustering)
Sliceline	Suche nach schwachen Abschnitten von strukturierten Daten, die semantisch kohärent sind und bei denen DNN eine geringe Leistung aufweist		VE1.3 VE1.4 VE1.5	vorhandene Testdaten mit GT	komplexe Ergebnisse	Fester Testsatz (nur Vorhersagen)	agnostisch	wahrscheinlich	Fortgeschrittene Methode		https://mboehm7.github.io/resources/sigmod2021b_sliceline.pdf
Spearman's rank correlation coefficient	Misst, wie gut zwei Variablen bpsw. Input und Output über eine monotone Funktion dargestellt werden können	SciPy	TR2.2 TR2.3	vorhandene Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	ja	Basismethode		
Spotlight	Suche nach schwachen Abschnitten in unstrukturierten Daten unter Verwendung des Einbettungsraums von DNNs		VE1.3 VE1.4 VE1.5	vorhandene Testdaten mit GT	komplexe Ergebnisse	Fester Testsatz (nur Vorhersagen)	agnostisch	wahrscheinlich	Fortgeschrittene Methode		https://dl.acm.org/doi/abs/10.1145/3531146.3533240
Statistical Parity	Misst, ob unterschiedliche Gruppen die gleiche Wahrscheinlichkeit haben, als positiv vorhergesagt zu werden, unabhängig von der tatsächlichen Klasse.	AIF 360, FairLearn, FairTest, FairDM / FairSight, Fairness Indicators, Robuscope,Microsoft Responsible AI Dashboard, AIF360	ND1.2 ND1.3 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	agnostisch	ja	Basismethode	https://aif360.readthedocs.io/en/latest/modules/generated/aif360.sklearn.metrics.statistical_parity_difference.html	https://dataplatfom.cloud.ibm.com/docs/content/wsi/model/wos-stat-parity-diff.html?context=cpdaas
Structural similarity index measure (SSIM)	Methode zur Einschätzung der wahrgenommenen Qualität von Bildern und Videos (Bildähnlichkeit oder -erzeugung)		VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	GenAI	ja	Basismethode	https://pytorch.org/ignite/generated/ignite.metrics.SSIM.html	https://en.wikipedia.org/wiki/Structural_similarity_index_measure
Success Rate of Backdoor Attacks	Anteil erfolgreicher Backdoor-Attacken	TrojanZoo, IBM Adversarial robustness	CY2.3 MA2.3	erweiterte oder neue Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	nein	Basismethode		
Success Rate of Data Poisoning Attacks	Anteil erfolgreicher Data Poisoning Attacken	auto-attack, DeepChecks, IBM Adversarial robustness	CY2.3 MA2.3	erweiterte oder neue Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	nein	Basismethode		
Success Rate of Label Poisoning Attacks	Anteil erfolgreicher LabelPoisoning Attacken	IBM Adversarial robustness	CY2.3 MA2.3	erweiterte oder neue Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	nein	Basismethode		
Success Rate of Membership Inference Attacks	Anteil erfolgreicher Membership Inference Attacken	ml_privacy_meter, DeepChecks, IBM Adversarial robustness	CY2.2 CY2.3 MA2.3	erweiterte oder neue Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	nein	Basismethode	https://github.com/privacytrustlab/ml_privacy_meter	

Technische Prüfmethodensammlung

Dieses Dokument ist Teil des MISSION KI Qualitätsstandards. ©acatech – Deutsche Akademie der Technikwissenschaften e.V.
Dieses Werk ist lizenziert unter der Creative Commons Lizenz Namensnennung – Keine Bearbeitungen 4.0 International (CC BY-ND 4.0). <https://creativecommons.org/licenses/by-nd/4.0/deed.de>

			Siehe Prüf-kriterien-katalog.	Zur Nutzung der Metrik benötigende (Eingabe-) Daten.	Detailgrad der Metrik.	Erforderliches Wissen und Zugriff auf das Modell.	Prüfbare Tasks mit der Metrik.	Automatisierte Prüfung der Metrikergebnisse gegen Grenzwerte.	Aussage-kraftigkeit, Umfang und In-formationsgewinn der Methode.	Verfügbare Frameworks, Tools oder Git-Repos.	Beispiele für die beschreibende Webseiten.
Test- / Methodenname	Kurzbeschreibung	Prüfwerkzeuge	Indi-katoren	Daten-anforderungen	Ergebnis-komplexität	Modellzugriff	Aufgaben-anwendbarkeit	Automatisierungs-möglichkeit	Tiefe der Testmethode	Bekannte Implementationen	Referenz
Accuracy	Misst die Genauigkeit eines (binären) Klassifikators als Verhältnis richtiger Vorhersagen zu Gesamt-vorhersagen	Amazon Sagemaker Suite, Azure ML, FairLearn, What If-Tool, Fairness Indicators, Citadel RADAR, ML Privacy Meter, AI Qualify, DeepChecks, Zeno, Microsoft Responsible AI Dashboard, LangChain OpenEvals, Der HELM Benchmark, GAIA-Benchmarks, Robustness Gym, Scikit-learn SHAP, ELI5	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html	https://en.wikipedia.org/wiki/Accuracy_and_precision
Surrogacy Efficacy Score	Anteil der Ausgaben, die gut erklärbar sind		TR2.2 TR2.3	vorhandene Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	ja	Basismethode		
SVM Failure Directions	Suche nach schwachen Abschnitten in unstrukturierten Daten unter Verwendung des CLIP-Einbettungsraums und eines SVM-Modells		VE1.3 VE1.4 VE1.5	vorhandene Testdaten mit GT	komplexe Ergebnisse	Fester Testsatz (nur Vorhersagen)	agnostisch	wahrscheinlich	Fortgeschrittene Methode		https://arxiv.org/abs/2206.14754
Synthetic reasoning	Reasoning-Benchmark	Citadel Lens	VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode		
Systematic Weakness search	Suche nach schwachen Slices von unstrukturierten Daten durch Umwandlung unstrukturierter Daten in strukturierte Daten, die an ODD ausgerichtet sind, und Anwendung von Sliceline, um Regionen zu finden, in denen DNN eine geringe Leistung aufweist	ELI5	VE1.3 VE1.4 VE1.5	vorhandene Testdaten mit GT	komplexe Ergebnisse	Fester Testsatz (nur Vorhersagen)	agnostisch	wahrscheinlich	Fortgeschrittene Methode		https://openaccess.thecvf.com/content/CVPR2023W/SAIAD/html/Gannamani_Invstigating_CLIP_Performance_for_Meta-Data_Generation_in_AD_Datasets_CVPRW_2023_paper.html
Time until Adversary's Success	Metrik für adversale Attacken	AI Qualify	VE1.3 VE1.4 VE1.5 CY2.2 CY2.3 MA2.3	erweiterte oder neue Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	nein	Basismethode		
True Negative Rate		Azure Machine Learning, ML Privacy Meter, TruEra, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html	https://en.wikipedia.org/wiki/Sensitivity_and_specificity
True Positive Rate		Azure Machine Learning, ML Privacy Meter, TruEra, Scikit-learn	VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	Klassifizierung	ja	Basismethode	https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html	https://en.wikipedia.org/wiki/Sensitivity_and_specificity
TruthfulQA	Question-Answering Benchmark	Citadel Lens	VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode	https://huggingface.co/datasets/truthfulqa/truthful_qa	
T-SNE	Visualisiert hochdimensionale Daten in einer niedrigdimensionalen Darstellung, um Muster oder Cluster zu identifizieren.		TR2.2 TR2.3	vorhandene Testdaten mit GT und Input	komplexe Ergebnisse	modellunabhängig	agnostisch	nein	Fortgeschrittene Methode		
unc	Unsicherheitsschätzung durch Modell-Ensembling		VE2.3 VE2.5	vorhandene Testdaten mit GT und Input	einzelne/mehrere reelle Zahlen	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	agnostisch	ja	Basismethode		
Wasserstein metric	Auch "Earth mover's distance" oder optimale Transportdistanz genannt, ist eine Ähnlichkeitsmetrik zwischen zwei Wahrscheinlichkeitsverteilungen		VE1.2 VE1.4 VE1.5 MA2.3	vorhandene Testdaten mit GT	einzelne/mehrere reelle Zahlen	Fester Testsatz (nur Vorhersagen)	agnostisch	ja	Basismethode	https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wasserstein_distance.html	https://en.wikipedia.org/wiki/Wasserstein_metric
WikiFact	Wissens-Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode	https://github.com/google-research-datasets/wikifact	
XSUM	Zusammenfassungs-Benchmark		VE1.2 VE1.4 VE1.5 MA2.3	Benchmarkdatensatz	komplexe Ergebnisse	Inferenzzugang (vom Prüfer bereitgestellte Eingaben)	GenAI	ja	Basismethode	https://huggingface.co/datasets/EdinburghNLP/xsum	

MISSION KI

Prüfbericht

Ergebnisse der Selbstbewertung von
[Name des Unternehmens] des KI-Systems
„[Name des KI-Systems]“ auf Grundlage
des MISSION KI Qualitätsstandards

Prüfbericht

[Name des Unternehmens]
„[Name des KI-Systems]“ KI-System

Selbstprüfung gemäß MISSION KI Qualitätsstandard

Datum: TT.MM.JJJJ

Prüf-ID: XXXXXX

Typ: Erstprüfung/Folgeprüfung

MISSION KI
c/o acatech
Deutsche Akademie der Technikwissenschaften
Karolinenplatz 4, D-80333 München

mission-ki.de

Gefördert durch:



Bundesministerium
für Digitales und
Staatsmodernisierung

aufgrund eines Beschlusses
des Deutschen Bundestages

Inhalt

1. Allgemeine Informationen zur Prüfung	2
1.1. Ziele, Umfang und Verantwortung	2
1.2. Validierung	2
1.3. Gültigkeit der Prüfung	2
2. Übersicht des KI-Systems und seiner Komponente(n)	3
3. Bewertung der Qualitätsmaßnahmen und Vorkehrungen zur Absicherung des KI-Systems	4
DA: Datenqualität, -schutz und -Governance	4
ND: Nicht-Diskriminierung (inkl. Bias)	6
TR: Transparenz	7
MA: Menschliche Aufsicht und Kontrolle	8
VE: Verlässlichkeit (inkl. Robustheit & Performance)	9
CY: KI-spezifische Cybersicherheit	10
Allgemeine Kommentare zur Prüfung des KI-Systems	11
Abschließendes Fazit	11
4. Bestätigungen	12

1. Allgemeine Informationen zur Prüfung

1.1. Ziele, Umfang und Verantwortung

Dieser Bericht präsentiert die Ergebnisse einer freiwilligen Selbstprüfung, die durchgeführt wurde, um den Umfang und die Effektivität der Maßnahmen zur Qualitätssicherung des (Name des KI-Systems) nach dem **MISSION KI** Qualitätsstandard festzustellen. Die Bewertung des KI-Systems ergibt sich aus der Gegenüberstellung der festgestellten Schutzbedarfe im Kontext der spezifischen Zweckbestimmung und den durch Evidenzen belegten, ergriffenen Maßnahmen. Die zu prüfende Organisation trägt die Verantwortung dafür, dass alle relevanten Informationen vollständig und wahrheitsgemäß bereitgestellt werden sowie entsprechend der Vorgaben validiert wurden.

1.2. Validierung

Die Validierung der Ergebnisse geschieht im Regelfall intern. Nur in Fällen, in denen es explizit genannt ist, wurde eine externe Validierung der Ergebnisse zur Erhöhung der Belastbarkeit durchgeführt. Je nach gewählter Einstufung gelten folgende Anforderungen an die Prüfenden:

Schutzbedarf	Einstufung	Beurteilungsansatz	Qualifikationsanforderung
–	D	Keine spezifische Anforderung.	Keine spezifische Qualifikation erforderlich.
gering	C	Keine Anforderung. Evidenzen und Einstufungen können durch Personen geprüft werden, die an der Entwicklung beteiligt waren.	(Berufs-)Erfahrung in Entwicklung oder Betrieb von KI-Systemen oder vergleichbare Qualifikation.
moderat	B	Die Personen sind nicht an der Entwicklung des KI-Systems beteiligt.	(Berufs-)Erfahrung in Entwicklung oder Betrieb von KI-Systemen oder vergleichbare Qualifikation.
hoch	A	Das 4-Augen-Prinzip wird angewendet, mit organisatorisch unabhängigen Personen.	Mindestens eine Person mit Auditerfahrung. Mindestens eine Person mit (Berufs-) Erfahrung in Entwicklung oder Betrieb von KI-Systemen oder vergleichbare Qualifikation.

1.3. Gültigkeit der Prüfung

Die Prüfaussage ist damit grundsätzlich nur gültig für die eindeutig bestimmte Version des KI-Systems, die zum Zeitpunkt der Prüfung vorlag. Sie verliert ihre Gültigkeit, sobald wesentliche Änderungen an der Zweckbestimmung, an der technischen Umsetzung oder durch äußere Bedingungen auftreten. Solange keine solche Änderungen vorliegen und dies durch regelmäßige Überprüfungen oder Monitoring sichergestellt wird, bleibt die Prüfaussage gültig.

2. Übersicht des KI-Systems und seiner Komponente(n)

Branche:	[X]
Rolle:	[Anbieter/Betreiber]
Betriebsort:	[X]
Kontakt:	[X]
Anwendungsfall:	1.1 [Titel]
Zweck:	1.2 [Wofür wird das System eingesetzt, in welchem Kontext]
Aufgabe:	1.3 [Welche Entscheidung/Vorhersage/Klassifikation]
Eingaben:	1.4 [Datenarten/Situationen]
Grenzen:	1.5 [Nicht-abgedeckte Fälle/Qualitätsgrenzen]
Ausgaben/Nutzung:	1.6 [Output-Form]
Nutzer und betroffene Personen	1.7 [Nutzergruppe/betroffene Personen]
Maß menschlicher Kontrolle	1.8. [Beteiligung an Betrieb/Aufsicht des KI-Systems, z. B. HIC, HITL, HOTL]
Betrieb	1.9 [Cloud/on-prem/hybrid]
Regulatorische Anforderungen	1.10 [Besondere Anforderungen für den Einsatzkontext]
Veränderungen im Betrieb	1.11 [Weiteres Training im Betrieb]

3. Bewertung der Qualitätsmaßnahmen und Vorkehrungen zur Absicherung des KI-Systems

Das Prüfverfahren für KI-Systeme beginnt mit einer detaillierten Beschreibung des Anwendungsfalls, gefolgt von einer Schutzbedarfsanalyse, die ermittelt, welche Qualitätsdimensionen und Kriterien am relevantesten sind und wie kritisch ihr Schutz in Abhängigkeit vom beabsichtigten Zweck und Kontext des Systems ist.

Jedes Kriterium wird zunächst auf seine Anwendbarkeit geprüft – wird ein Kriterium als nicht anwendbar eingestuft, muss im Feld „*Kommentar zum Schutzbedarf*“ eine Begründung angegeben werden, warum es nicht zutrifft; die weitere Bewertung dieses Kriteriums entfällt. Für anwendbare Kriterien wird der Schutzbedarf in drei Stufen eingeteilt: gering, moderat oder hoch, die den Schweregrad potenzieller Schäden widerspiegeln.

Anschließend wird das System eingestuft, wobei jedes anwendbare Kriterium anhand der umgesetzten Maßnahmen und Schutzvorkehrungen beurteilt wird. Diese Bewertung wird durch Evidenzen wie technische Tests, Dokumentationen und Zertifizierungen unterstützt. Der dabei umgesetzte Grad der Validierung der Evidenzen muss aufgeführt werden. Maßnahmen, die bei der Güte des KI-Systems besonders hervorzuheben sind, können in diesem Prüfbericht ebenfalls aufgeführt werden.

Die Ergebnisse der Schutzbedarfsanalyse legen die jeweils pro Kriterium erforderliche Mindeststufe fest und werden systematisch mit der tatsächlichen Einstufung des Systems verglichen. Die Bewertung prüft damit, ob die umgesetzten Maßnahmen die geforderten Schutzbedarfe erfüllen.

DA: Datenqualität, -schutz und -Governance		
Datenqualität		
Schutzbedarf:		<p>Gering</p> <p>Immer anwendbar. Ein geringer Schutzbedarf erfordert mindestens Maßnahmen auf Stufe C.</p>
Erreichte Stufe:		<p>C</p> <p>Evidenzen wurden intern validiert.</p>
Ausgewählte Maßnahmen	Organisatorische Maßnahmen:	z.B. Benutzerinstruktion, Schulung, Fachliche und technische Dokumentation, Berichte
	Technische Maßnahmen:	z.B. Testergebnisse, Logging, Monitoring, Datenverarbeitungsschritte, Not-Aus-Knopf
Kommentar:		
Kriterium erfüllt		

Schutz personenbezogener Daten		
Schutzbedarf:		Choose an item.
		Choose an item.
Erreichte Stufe:		Choose an item.
		Choose an item.
Ausgewählte Maßnahmen	Organisatorische Maßnahmen:	z.B. Benutzerinstruktion, Schulung, Fachliche und technische Dokumentation, Berichte
	Technische Maßnahmen:	z.B. Testergebnisse, Logging, Monitoring, Datenverarbeitungsschritte, Not-Aus-Knopf
Kommentar:		
Kriterium erfüllt		

Schutz proprietärer Daten		
Schutzbedarf:		Choose an item.
		Choose an item.
Erreichte Stufe:		Choose an item.
		Choose an item.
Ausgewählte Maßnahmen	Organisatorische Maßnahmen:	z.B. Benutzerinstruktion, Schulung, Fachliche und technische Dokumentation, Berichte
	Technische Maßnahmen:	z.B. Testergebnisse, Logging, Monitoring, Datenverarbeitungsschritte, Not-Aus-Knopf
Kommentar:		
Kriterium erfüllt		

ND: Nicht-Diskriminierung (inkl. Bias)

Vermeidung von ungerechtfertigter Diskriminierung und Verzerrung

Schutzbedarf:		Choose an item.
		Choose an item.
Erreichte Stufe:		Choose an item.
		Choose an item.
Ausgewählte Maßnahmen	Organisatorische Maßnahmen:	z.B. Benutzerinstruktion, Schulung, Fachliche und technische Dokumentation, Berichte
	Technische Maßnahmen:	z.B. Testergebnisse, Logging, Monitoring, Datenverarbeitungsschritte, Not-Aus-Knopf
Kommentar:		

Kriterium erfüllt

TR: Transparenz

Rückverfolgbarkeit und Dokumentation

Schutzbedarf:	Choose an item.	
	Choose an item.	
Erreichte Stufe:	Choose an item.	
	Choose an item.	
Ausgewählte Maßnahmen	Organisatorische Maßnahmen:	z.B. Benutzerinstruktion, Schulung, Fachliche und technische Dokumentation, Berichte
	Technische Maßnahmen:	z.B. Testergebnisse, Logging, Monitoring, Datenverarbeitungsschritte, Not-Aus-Knopf
Kommentar:		

Kriterium erfüllt

Erklärbarkeit & Interpretierbarkeit

Schutzbedarf:	Choose an item.	
	Choose an item.	
Erreichte Stufe:	Choose an item.	
	Choose an item.	
Ausgewählte Maßnahmen	Organisatorische Maßnahmen:	z.B. Benutzerinstruktion, Schulung, Fachliche und technische Dokumentation, Berichte
	Technische Maßnahmen:	z.B. Testergebnisse, Logging, Monitoring, Datenverarbeitungsschritte, Not-Aus-Knopf
Kommentar:		

Kriterium erfüllt

MA: Menschliche Aufsicht und Kontrolle

Menschliche Handlungsfähigkeit

Schutzbedarf:		Choose an item.
		Choose an item.
Erreichte Stufe:		Choose an item.
		Choose an item.
Ausgewählte Maßnahmen	Organisatorische Maßnahmen:	z.B. Benutzerinstruktion, Schulung, Fachliche und technische Dokumentation, Berichte
	Technische Maßnahmen:	z.B. Testergebnisse, Logging, Monitoring, Datenverarbeitungsschritte, Not-Aus-Knopf
Kommentar:		

Kriterium erfüllt

Menschliche Aufsicht

Schutzbedarf:		Choose an item.
		Choose an item.
Erreichte Stufe:		Choose an item.
		Choose an item.
Ausgewählte Maßnahmen	Organisatorische Maßnahmen:	z.B. Benutzerinstruktion, Schulung, Fachliche und technische Dokumentation, Berichte
	Technische Maßnahmen:	z.B. Testergebnisse, Logging, Monitoring, Datenverarbeitungsschritte, Not-Aus-Knopf
Kommentar:		

Kriterium erfüllt

VE: Verlässlichkeit (inkl. Robustheit & Performance)

Leistungsfähigkeit und Robustheit

Schutzbedarf:		Choose an item.
		Choose an item.
Erreichte Stufe:		Choose an item.
		Choose an item.
Ausgewählte Maßnahmen	Organisatorische Maßnahmen:	z.B. Benutzerinstruktion, Schulung, Fachliche und technische Dokumentation, Berichte
	Technische Maßnahmen:	z.B. Testergebnisse, Logging, Monitoring, Datenverarbeitungsschritte, Not-Aus-Knopf
Kommentar:		

Kriterium erfüllt

Rückfallpläne und funktionale Sicherheit

Schutzbedarf:		Choose an item.
		Choose an item.
Erreichte Stufe:		Choose an item.
		Choose an item.
Ausgewählte Maßnahmen	Organisatorische Maßnahmen:	z.B. Benutzerinstruktion, Schulung, Fachliche und technische Dokumentation, Berichte
	Technische Maßnahmen:	z.B. Testergebnisse, Logging, Monitoring, Datenverarbeitungsschritte, Not-Aus-Knopf
Kommentar:		

Kriterium erfüllt

CY: KI-spezifische Cybersicherheit

Allgemeine KI-Spezifische Cybersicherheit

Schutzbedarf:		Choose an item.
		Choose an item.
Erreichte Stufe:		Choose an item.
		Choose an item.
Ausgewählte Maßnahmen	Organisatorische Maßnahmen:	z.B. Benutzerinstruktion, Schulung, Fachliche und technische Dokumentation, Berichte
	Technische Maßnahmen:	z.B. Testergebnisse, Logging, Monitoring, Datenverarbeitungsschritte, Not-Aus-Knopf
Kommentar:		

Kriterium erfüllt

Widerstandsfähigkeit gegen KI-spezifische Angriffe

Schutzbedarf:		Choose an item.
		Choose an item.
Erreichte Stufe:		Choose an item.
		Choose an item.
Ausgewählte Maßnahmen	Organisatorische Maßnahmen:	z.B. Benutzerinstruktion, Schulung, Fachliche und technische Dokumentation, Berichte
	Technische Maßnahmen:	z.B. Testergebnisse, Logging, Monitoring, Datenverarbeitungsschritte, Not-Aus-Knopf
Kommentar:		

Kriterium erfüllt

Allgemeine Kommentare zur Prüfung des KI-Systems

(Bitte geben Sie in diesem Feld alle allgemeinen Beobachtungen, methodischen Anmerkungen oder andere Besonderheiten ein, die während der Prüfung festgestellt wurden. Dazu können beispielsweise Anmerkungen zu durchgeführten oder nicht durchgeführten Prüfungsverfahren, zur Verfügbarkeit und Qualität von Nachweisen, zu methodischen Einschränkungen oder zu allgemeinen Stärken und Schwächen des KI-Systems gehören. Gegebenenfalls können Sie auch Aspekte wie „Änderungen seit der letzten Bewertung“, „Festgestellte Lücken“ und „Geplante Verbesserungen“ hinzufügen.)

[Vorlage für „Änderungen seit der letzten Bewertung“, „Festgestellte Lücken“ und „Geplante Verbesserungen“:]

Änderungen seit der letzten Bewertung

[Kurz die Bereiche nennen, in denen sich Bewertung oder Evidenzlage verbessert hat.]

Festgestellte Lücken (nur wenn es nicht erfüllte Kriterien gibt)

Einzelne Kriterien, wie [Kriteriennamen], konnten das geforderte Schutzniveau nicht vollständig erreichen. Die Gründe für die Nichterfüllung sind in den jeweiligen Abschnitten bzw. im Feld „Kommentar“ dokumentiert und sollten im Rahmen der Nachbesserungen berücksichtigt werden. Diese Lücken sind vorrangig zu adressieren.

Geplante Verbesserungen

[In diesem Abschnitt könnten neben Maßnahmen zur Schließung identifizierter Lücken auch weitere geplante Verbesserungen dokumentiert werden, die unabhängig von bestehenden Lücken zur Optimierung des KI-Systems oder des Prüfprozesses beitragen sollen.]

- [Sofern Lücken festgestellt wurden:] Zur Schließung der identifizierten Lücken sind gezielte Maßnahmen geplant. Dazu gehören weitere technische Tests, Prozessoptimierungen und eine erweiterte Dokumentation. Die Umsetzung dieser Verbesserungen wird nachverfolgt und in der nächsten Überprüfung verifiziert.
- [Sofern **keine** Lücken festgestellt wurden:] Im Rahmen der Prüfung wurden keine Lücken festgestellt. Entsprechend sind keine weiteren Verbesserungsmaßnahmen zur Schließung von Lücken geplant.

Abschließendes Fazit

Für das KI-System [Name der KI-Anwendung] wurden für [alle/nicht alle] anwendbaren Kriterien Maßnahmen getroffen, die geeignet sind, um den identifizierten Schutzbedarf zu adressieren.

Kriterien (Kriteriennamen) wurden nicht erfüllt. Kriterien [Kriteriennamen] wurden übertroffen.

4. Bestätigungen

Verantwortlich für die Durchführung dieser Bewertung sind die nachfolgend aufgeführten Personen. Sie bestätigen mit ihrer Unterschrift, dass die in diesem Prüfbericht beschriebenen Prüfhandlungen entsprechend ihrer jeweiligen Zuständigkeit (siehe Abschnitt 1.2) ordnungsgemäß durchgeführt wurden und die getätigten Aussagen wahrheitsgetreu sind. Sie bestätigen weiterhin für ihre jeweilige Zuständigkeit, dass relevante und zutreffende Nachweise erbracht wurden aus denen hervorgeht, dass der Umfang und die Effektivität der Maßnahmen ausreichen, um die identifizierten Schutzbedarfe plausibel abzudecken. Sie übernehmen jedoch im Einzelnen keine Verantwortung für Aussagen, die über ihre Zuständigkeit hinausgehenden getätigt wurden.

Mehrere Unterschriften sind erforderlich, wenn unterschiedlich hohe Einstufungen vorgenommen werden. Dies ist beispielsweise der Fall, wenn zur Erfüllung unterschiedlich hoher Schutzbedarfe verschieden hohe Mindeststufen erfüllt werden müssen. In diesem Fall müssen alle relevanten in Abschnitt 1.2 genannten Personen mit Prüfverantwortung unterzeichnen.

Datum:
[Datum]

Datum:
[Datum]

Datum:
[Datum]

Name:
[Name der Prüfenden]

Name:
[Name der Prüfenden]

Name:
[Name der Prüfenden]

Abteilung:
[Name der Abteilung]

Abteilung:
[Name der Abteilung]

Abteilung:
[Name der Abteilung]

An der Entwicklung des
bewerteten KI-Tools beteiligt:
ja/nein

An der Entwicklung des
bewerteten KI-Tools beteiligt:
ja/nein

An der Entwicklung des
bewerteten KI-Tools beteiligt:
ja/nein

Prüfungs-/Auditerfahrungen:
ja/nein

Prüfungs-/Auditerfahrungen:
ja/nein

Prüfungs-/Auditerfahrungen:
ja/nein

Unterschrift

Unterschrift

Unterschrift

Haftungsausschluss

Die angebotene Selbstprüfung dient ausschließlich der freiwilligen internen Bewertung durch die teilnehmende Organisation. Sie stellt keine behördliche Prüfung, Zertifizierung oder rechtsverbindliche Bewertung dar. acatech übernimmt keine Gewähr für Vollständigkeit, Richtigkeit oder rechtliche Wirkung der Ergebnisse dieser Selbstprüfung. Die Nutzung erfolgt auf eigenes Risiko und in eigener Verantwortung des teilnehmenden Unternehmens. acatech haftet nur bei Vorsatz oder grober Fahrlässigkeit sowie in Fällen gesetzlicher Haftung.

Autorenverzeichnis

acatech

Carolin Anderson
Simon Boffen
Dr. Philipp Heß
Adrian Meisner

AI Quality & Testing Hub

Dr. Simone Amoroso
Paul Luca Palupski
Dr. Cord Schlötelburg
Hosei Halim
Paula Hoffmann

Fraunhofer IAIS

Dr. Maram Akila
Dr. Daniel Becker
Rebekka Göрге
Dr. Henrik Junklewitz
Dr. Michael Mock
Dr. Maximilian Poretschkin
Anna Schmitz
Sebastian Schmidt

PricewaterhouseCoopers

Lina Antje Gühne
Alina Kudanova
Nan-Hee Kang
Laszlo Kühl
Jan-Niklas Nieland
Hendrik Reese

TÜV AI.Lab

Dr.-Ing. Marc P. Hauer
Leonie Löbenberg
Matthias König
Dr. Christoph Poetsch
Franziska Weindauer

VDE

Nora Dörr
Andreas Hauschke
Dr. Thorsten Prinz