TÜV AUSTRIA Best Practice Leitfaden:

Technische Dokumentation

Versionsnummer: 1.3

Erscheinungsdatum: 11.03.2025

Diese Dokumentvorlage bietet eine Struktur für die technische Dokumentation des entwickelten KI-Systems in Übereinstimmung mit Artikel 11 des EU AI Acts, der ISO/IEC 42001:2023 und dem TÜV AUSTRIA TRUSTED AI Standard. Es stellt sicher, dass alle relevanten technischen Details erfasst werden, so dass Benutzer und Interessenvertreter das Design und die Funktionalität des KI-Systems sowie die Einhaltung gesetzlicher Standards verstehen können.

Welche Abschnitte der Vorlage anwendbar sind, hängt von den geltenden Vorschriften, der Unternehmensgröße sowie dem Risikoniveau und der Art des entwickelten KI-Systems ab.

RTR	**Hinweis:** Die vorliegende Dokumentvorlage wurde gemeinsam mit der bei der RTR-GmbH eingerichteten KI-Servicestelle abgestimmt und ist nach § 194a Abs 1 Z 2 TKG auch auf der Website der KI-Servicestelle unter https://ki.rtr.at abrufbar. Diese Dokumentvorlage soll als Orientierungshilfe zur Erfüllung der im AI Act normierten Anforderungen an technische Dokumentation von KI-Systemen dienen. Sie ist kein Präjudiz zu erwarteten harmonisierten Normen oder noch erwarteten Leitlinien des AI Office.

Aus Gründen der Verständlichkeit gebraucht der TÜV AUSTRIA Best Practice Leitfaden: Technische Dokumentation *vorrangig den Begriff „KI-System“, bezieht sich jedoch auch an einigen Stellen auf „KI-Modelle“, „GPAI-Modelle“ oder „Modelle“. Die erwähnten Begriffe und deren Derivate nehmen keine rechtliche Qualifikation der jeweiligen technischen Komponente als „KI-System“ im Sinne des Art. 3 Z 1 EU AI Act, „GPAI-Modell“ im Sinne des Art. 3 Z 66 EU AI Act, als „KI-Modell“ (Begriff nicht explizit vom EU AI Act definiert) etc., und damit einhergehende Dokumentationserfordernisse vorweg. Beispielsweise spricht der Leitfaden von „Zuständigkeiten (einschließlich Kontaktangaben) von der Person/en oder Organisation/en, die das KI-System entwickelt haben/hat“. Entsprechende Angaben sind auch bei einer rechtlichen Einstufung als „GPAI-Modell“ anzuführen.* Zur Auslegung des Begriffs „KI-System“ iSd Art. 3 Z 1 EU AI Act können auch die von der Europäischen Kommission veröffentlichten Leitlinien[1] herangezogen werden.

[1] https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-ai-system-definition-facilitate-first-ai-acts-rules-application

1. Wichtigste Fakten

Beschreibung von:

- Zuständigkeiten (einschließlich Kontaktangaben)
    - Person oder Organisation, die das KI-System entwickelt hat, falls zutreffend
    - Person oder Organisation, die für den Betrieb des KI-Systems, und diejenigen, die für die Nutzung des Systems verantwortlich ist, insbesondere für die Behandlung von Systemfehlern oder die Verwaltung von Updates, falls zutreffend
- Veröffentlichungsdatum des KI-Systems
- Versionsnummer, die die Beziehung zu früheren Versionen widerspiegelt
- Modelltyp(en) und Architektur(en) des dem KI-System zugrundeliegenden Modells bzw. der dem KI-System zugrundeliegenden Modelle
- (High-Level-)Informationen über Trainingsalgorithmen, Parameter, Fairness-Beschränkungen oder andere angewandte Ansätze, verwendete Basismodelle und Funktionen
- Wissenschaftliche Publikationen oder andere Ressourcen für weitere Informationen
- Zitiervorschlag für das KI-System
- Lizenzinformationen
- Kontaktdaten für Fragen oder Kommentare zum KI-System

2. Verwendungszweck

Beschreibung von:

- Primären Verwendungszwecken und der (vereinfachten) Definition des Anwendungsbereichs des KI-Systems
- Vorgesehene Hauptnutzer des KI-Systems
- Vorhersehbare Fehlanwendungen
- Anwendungsfällen, die nicht in den Anwendungsbereich fallen
- Risiken bei der Anwendung des KI-Systems und Maßnahmen zur Risikominderung, einschließlich der Quellen der Risiken für
    - Gesundheit und Sicherheit
    - Grundrechte und
    - Diskriminierung

    falls zutreffend,

und

- Eine Betriebsanleitung für den Betreiber und ggf. eine grundlegende Beschreibung der für den Betreiber zur Verfügung gestellten Benutzerschnittstelle.

3. Ergebnisse der Risikoanalyse

Beschreibung von

- den bekannten und vorhersehbaren Risiken, die das Hochrisiko-KI-System für die Gesundheit, die Sicherheit oder die Grundrechte darstellen kann, wenn das Hochrisiko-KI-System entsprechend seiner Zweckbestimmung verwendet wird.
- der Abschätzung und Bewertung der Risiken, die entstehen können, wenn das Hochrisiko-KI-System entsprechend seiner Zweckbestimmung oder im Rahmen einer vernünftigerweise vorhersehbaren Fehlanwendung eingesetzt wird
- der Bewertung anderer möglicherweise auftretender Risiken, basierend auf der Analyse von Daten, die aus dem in Art. 72 des EU AI Act genannten Post-Market-Monitoring-System erhoben wurden
- der Ergreifung geeigneter und gezielter Risikomanagementmaßnahmen zur Bewältigung der ermittelten Risiken
- Risikomanagementmaßnahmen, die Folgendes sicher stellen:
    - Beseitigung oder Reduzierung der oben genannte identifizierten und bewerteten Risiken, soweit technisch machbar durch geeignete Gestaltung und Entwicklung des Hochrisiko-KI-Systems
    - gegebenenfalls die Umsetzung angemessener Maßnahmen zur Abschwächung und Kontrolle von nicht auszuschließenden Risiken
    - Bereitstellung der gemäß Art. 13 des EU AI Act erforderlichen Informationen und ggf. Schulung der Betreiber

4. Grundrechte-Folgenabschätzung

Beschreibung der möglichen Folgen für Einzelpersonen oder Gruppen von Einzelpersonen oder beides und für die Gesellschaft, die sich aus dem KI-System während seines gesamten Lebenszyklus ergeben können, z. B. durch eine Folgenabschätzung gemäß ISO/IEC 42001:2023 Anhang B.5.

Bestimmte Betreiber müssen nach Art. 27 des EU AI Act eine Abschätzung durchführen und eine Beschreibung errichten, die folgendes umfasst:

- Verfahren des Betreibers, bei denen das Hochrisiko-KI-Systems im Einklang mit seiner Zweckbestimmung verwendet wird
- Zeitraum und Häufigkeit der Verwendung des Hochrisiko-KI-Systems
- Kategorien der natürlichen Personen und Personengruppen, die von der Verwendung des Hochrisiko-KI-Systems betroffen sein könnten
- Spezifische Schadensrisiken, die sich auf die im vorherigen Punkt genannten Personen oder Personengruppen auswirken könnten
- Umsetzung der Maßnahmen der menschlichen Aufsicht entsprechend der Betriebsanleitung
- zu ergreifende Maßnahmen im Falle des Eintretens der oben genannten Risiken (einschließlich Regelungen für die interne Unternehmensführung und Beschwerdemechanismen)

5. Einhaltung der Vorschriften

Beschreibung

- Eines Risikomanagementsystems gemäß Art. 9 des EU AI Act
- der ganz oder teilweise angewandten harmonisierten Normen, die im Amtsblatt der Europäischen Union veröffentlicht worden sind. Wurden keine solchen harmonisierten Normen angewandt, so ist eine ausführliche Beschreibung der Lösungen vorzulegen, mit denen die Anforderungen gemäß Kapitel III Abschnitt 2 des EU AI Act erfüllt werden, einschließlich eines Verzeichnisses der sonstigen angewandten einschlägigen Normen und technischen Spezifikationen
    - Einhaltung der Anforderungen (EU AI Act, Art. 8)
    - Risikomanagementsystem (EU AI Act, Art. 9)
    - Daten und Daten-Governance (EU AI Act, Art. 10)
    - Technische Dokumentation (EU AI Act, Art. 11)
    - Aufzeichnungspflichten (EU AI Act, Art. 12)
    - Transparenz und Bereitstellung von Informationen für die Betreiber (EU AI Act, Art. 13)
    - Menschliche Aufsicht (EU AI Act, Art. 14)
    - Genauigkeit, Robustheit und Cybersicherheit (EU AI Act, Art. 15)
- ein Verzeichnis sonstiger angewandter einschlägiger Normen und technischer Spezifikationen

und

- eine Kopie der EU-Konformitätserklärung gemäß Art. 47 des EU AI Acts

6. Entwicklungsprozess

Beschreibung von:

- der Verwendung von vortrainierten Systemen oder Tools, die für die Entwicklung des KI-Systems verwendet wurden, und Details wie diese integriert oder verändert wurden
- der allgemeinen Logik des KI-Systems
    - einschließlich der wichtigsten Entwurfsentscheidungen mit den Gründen und getroffenen Annahmen, der Frage, wofür das System optimiert werden soll, und der Relevanz verschiedener Parameter für das Erreichen dieser Ziele
- den verwendeten Trainingsdatensätzen, insbesondere
    - eine allgemeine Beschreibung des Datensatzes (Datenquellen und Daten-erhebungsverfahren, Zusammensetzung des Datensatzes)
    - die Herkunft der Daten und Informationen über das Auswahl- und Kennzeichnungsverfahren
    - die Datenbereinigungsmethoden
- Schritte der Datenvorverarbeitung nach Art. 10 des EU AI Act
- die Trainingsmethoden und -techniken, einschließlich
    - Informationen über die Feinabstimmung von Hyperparametern und der zugehörigen Kostenfunktion, falls zutreffend

- Entscheidungen über mögliche Kompromisse in Bezug auf die technischen Lösungen, die zum Ausgleich verschiedener Leistungskennzahlen gewählt wurden (bspw. in Bezug auf Fairness der Anwendung)

- die für das Training, die Validierung und den Test des KI-Systems erforderlichen Rechenressourcen

7. Technische Details

Beschreibung von

- dem Zusammenspiel des Systems mit Hardware und Software:
  - Informationen darüber, wie das KI-System mit anderer Hardware/Software interagiert, einschließlich anderer KI-Systeme, falls zutreffend
  - der Interaktion zwischen den datengetrieben programmierten Teilen und den nicht datengetrieben programmierten Teilen,
- der Systemarchitektur, einschließlich:
  - High-Level-Architektur: Ein (z.B. UML) Diagramm und eine Beschreibung der gesamten Systemarchitektur
  - System-Integration: Wie das KI-System mit anderen Systemen oder Komponenten integriert wird (REST, usw.)
    - Wesentliche Designentscheidungen und -konzepte einschließlich einer Erklärung für die betroffenen Systembenutzer
- den minimalen und empfohlenen Hardware-Spezifikationen, einschließlich:
  - CPU: Die minimalen und empfohlenen Prozessorspezifikationen, die für den Betrieb des KI-Systems erforderlich sind. Dazu können die Anzahl der Kerne, die Taktfrequenz und gegebenenfalls bestimmte Prozessormodelle gehören,
  - GPU: Für Systeme, die GPUs zur Beschleunigung nutzen, die minimalen und empfohlenen GPU-Spezifikationen an, z. B. den Typ (z. B. NVIDIA), das Modell, die Größe des VRAM (Video RAM) und die CUDA-Kerne.
  - RAM: Die minimale und empfohlene Menge an Systemspeicher, die für einen effizienten Betrieb des KI-Systems erforderlich ist
  - Speicherplatz: Minimale und empfohlene Speicheranforderungen, einschließlich der Art des Speichers (HDD, SSD) und des erforderlichen Platzes für die Datenspeicherung, Modelldateien und Protokolle
- den Software-Anforderungen
  - Betriebssystem: Unterstützte Betriebssysteme (z. B. Windows, Linux, macOS) und spezifische Versionen.
  - Abhängigkeiten: Liste aller Software-Abhängigkeiten , die für den Betrieb des KI-Systems erforderlich sind, einschließlich:
    - Programmiersprachen: Geben Sie die verwendeten Programmiersprachen an
    - Bibliotheken und Frameworks: Detaillierte Liste der Bibliotheken und Frameworks, einschließlich Versionsnummern
    - Datenbanken: Einzelheiten zu den benötigten Datenbanken (Engines, Datenformate usw.)
    - Andere Tools: Zusätzliche Tools oder Software, die für einen Betrieb erforderlich sind

- Installationsanweisungen: Schritt-für-Schritt-Anweisungen für die Installation der einzelnen Abhängigkeiten, einschließlich der erforderlichen Befehle und Konfigurationen, notwendigen Zugriffsberechtigungen und Integrationstests

- Input/Output, einschließlich, aber nicht beschränkt auf:
  - Datenquellen
    - Informationen über sensorische oder andere Zugangspunkte, über die das System auf Eingaben zugreift
  - Anforderungen an die Eingabedaten (Eingabeformat und mögliche Einschränkungen für die Eingabedaten)
  - Datenspeicherung (wie und wo die Daten gespeichert werden, einschließlich Datenbankschemata oder Dateistrukturen)
  - Ausgabeformat des Systems (Datentypen, etc.)
  - Interpretation der Ergebnisse (Wahrscheinlichkeiten, Logits usw.)
- den Software/Firmware-Versionen:
  - Informationen über die relevanten Software-/Firmwareversionen und Aktualisierungsanforderungen
- der Verfügbarkeit:
  - Beschreibung aller Varianten, in denen das KI-System verfügbar ist (z. B. Pakete, Downloads, APIs).
- den Details der Produktkomponenten:
  - Falls es sich bei dem KI-System um eine Produktkomponente handelt, sind Fotos oder Abbildungen vorzulegen, die die äußeren Merkmale, die Kennzeichnungen und den inneren Aufbau darstellen.
- den Maßnahmen zur menschlichen Überwachung:
  - Eine Beschreibung der Werkzeuge für die Mensch-Maschine-Schnittstelle, mit denen sichergestellt wird, dass das KI-System wirksam überwacht werden kann
  - Eine Beschreibung, wie Betreiber in der Lage sind
    - die einschlägigen Fähigkeiten und Grenzen des Hochrisiko-KI-Systems angemessen zu verstehen und seinen Betrieb ordnungsgemäß zu überwachen
    - die Ergebnisse des KI-Systems zu interpretieren (erkennen von Anomalien, Fehlfunktionen und unerwarteter Leistung)
    - sogenannten “Automatisierungsbias” (übermäßiges Vertrauen in die vom KI-System hervorgebrachte Ausgabe) zu erkennen und sich dessen bewusst sein
    - in bestimmten Situationen die Ausgaben eines Hochrisiko-KI-Systems außer Acht zu lassen bzw. diese außer Kraft zu setzen oder rückgängig zu machen, oder
    - bei Bedarf einzugreifen (zB Verfahren zu unterbrechen).

8. Test- und Validierungsdatensatz und Methodik

Beschreibung von:

- der Zusammensetzung des zur Bewertung der Modellleistung verwendeten Testdatensatzes, einschließlich einer Beschreibung der Vorverarbeitung

- den verwendeten Validierungs- und Testverfahren, einschließlich der Merkmale der in diesen Prozessen verwendeten Daten, sowie die zur Messung der Genauigkeit, der Robustheit und der Einhaltung der Anforderungen verwendeten Metriken (sog. Leistungskennzahlen), insbesondere:
    - Validierungs- und Testdaten: Beschreibung der Validierungs- und Testdatensätze, einschließlich ihrer Quellen, Zusammensetzung und Merkmale.
    - Prüfmethodik: Erläuterung der Testverfahren, einschließlich aller Tests unter Realbedingungen und Bewertungen.
    - Metriken und Protokolle: Detaillierte Beschreibung der Leistungskennzahlen (z. B. Genauigkeit, Robustheit, mögliche diskriminierende Auswirkungen) und Aufzeichnung der Testprotokolle und -berichte. Stellen Sie sicher, dass alle Testprotokolle mit Datum versehen und von den verantwortlichen Personen unterzeichnet sind.

9. Leistungskennzahlen des KI-Systems

Beschreibung der Metriken („Leistungskennzahlen“) für das spezifische KI-System:

- Beschreibung der Angemessenheit jeder Leistungskennzahl und des akzeptablen Zielbereichs
- Bewertung geeigneter Leistungskennzahlen, die auf potenzielle Verzerrungen und deren Auswirkungen hinweisen
- Unsicherheit der Leistungsmessungen, d. h. Konfidenzintervalle oder Ergebnisse statistischer Tests, falls zutreffend
- Systemfähigkeiten und -beschränkungen in Bezug auf die Leistungskennzahlen und die Robustheit
- Interpretation der Leistungskennzahlen in Bezug auf die Zuverlässigkeit des Systems
- Public Evaluation Protocols: Falls zutreffend alle verwendeten Protokolle oder Tools (z. B. GLUE, SuperGLUE, HELM) anführen und beschreiben
- Alternative Bewertungsmethoden: Falls zutreffend, andere Methoden, die bei der Bewertung verwendet wurden, einschließlich der Leistung in aufgabenspezifischen und allgemeinen Fällen, erläutern.

10. Überwachung des KI-Systems

Beschreibung von:

- den Plänen für den Fall von Systemausfällen
- den Rollback-Plänen für das KI-System, Verfahren für
    - das Deaktivieren von Funktionen
    - Aktualisierungsprozesse und Updates
    - die Benachrichtigung der Kunden und Nutzer über Änderungen oder Systemausfälle, und
    - Risikominderungsstrategien
- den Verfahren zur Überwachung des Zustands des KI-Systems in der Post-Market Phase. Dies beinhaltet:
    - Sicherstellung des bestimmungsgemäßen Betriebs des KI-Systems innerhalb der normalen Betriebsspannen (Beobachtbarkeit) und Behebung von Ausfällen des KI-Systems

- Überwachung relevanter Ereignisse, Priorisierung und Überprüfung von Ereignisprotokollen, Untersuchung von Fehlern und Maßnahmen zur Fehlervermeidung

- den ergriffenen Maßnahmen zur Cybersicherheit (EU AI Act, Annex IV, 2h) insbesondere technische und organisatorische Maßnahmen zur Absicherung der Entwicklungs- und Betriebsinfrastruktur, sowie Maßnahmen gegen Adversarial Attacks

- den relevanten Änderungen, die der Anbieter im Laufe seines Lebenszyklus am System vorgenommen hat

11. Lebenszyklus des KI-Systems

Beschreibung von:

- den vorher festgelegten Änderungen, einschließlich:
  - einem Überblick über im Voraus festgelegte Änderungen (d. h. Änderungen, Aktualisierungen oder Anpassungen, die geplant und in den Entwurfs- oder Betriebslebenszyklus des KI-Systems integriert sind) und welche Maßnahmen ergriffen werden, um die kontinuierliche Übereinstimmung des Systems mit den einschlägigen technischen und regulatorischen Normen im Hinblick auf alle erwarteten Aktualisierungen oder Änderungen zu gewährleisten.
- den Verfahren, mit denen die Organisation auf betriebliche Veränderungen reagiert, einschließlich der Kommunikation mit den Nutzern und der internen Bewertung.
- Die Dokumentation sollte
  - aktuell
  - genau und
  - von der zuständigen Leitung genehmigt

  sein.

12. Zusätzliche Anforderungen an KI-Modelle mit allgemeinem Verwendungszweck (*General Purpose KI-Modelle)*

Dieser Abschnitt folgt Art. 53 und Annex XI bis XIII im EU AI Act.

- Allgemeine Beschreibung des KI-Modells mit allgemeinem Verwendungszweck

  - Beabsichtigte Aufgaben und Integration: Ein umfassender Überblick über die Aufgaben, die das Modell erfüllen soll, und die Arten und Merkmale von KI-Systemen, in die es integriert werden kann

  - Anwendbaren Regelungen der akzeptablen Nutzung, Richtlinien oder Beschränkungen, die vom KI-Anbieter festgelegt werden, um zu spezifizieren, wie das KI-Modell genutzt werden sollte und wie nicht

  - Datum der Freigabe und die Vertriebsmethoden

- Detaillierte Beschreibung und genutzte Ressourcen im Entwicklungsprozess

  - Eine ausreichend detaillierte Zusammenfassung der Datenbasis, die für das Trainieren, Testen und Validieren des KI-Modells mit allgemeinem Verwendungszweck eingesetzt wurde (Art und Herkunft der Daten, Aufbereitungsmethoden, Anzahl der verfügbaren Datenpunkte und ihre Hauptmerkmale, Kennzahlen für die Datenqualität). Diese Zusammenfassung sollte öffentlich zugänglich sein.

  - Technische Mittel für die Integration: Detaillierte Informationen über die technischen Mittel, die für die Integration des KI-Modells mit allgemeinem Verwendungszweck in andere KI-Systeme erforderlich sind (z. B. Gebrauchsanweisungen, Infrastruktur, etc.)

- Verwendete Rechenressourcen:
    - Trainingsressourcen: Einzelheiten wie die Anzahl der Gleitkommaoperationen und die Trainingszeit. Nennen Sie andere relevante Informationen
- Energieverbrauch:
    - Bekannter oder geschätzter Energieverbrauch: Der bekannte oder geschätzte Energieverbrauch des Modells (z. B. auf der Grundlage der verwendeten Rechenressourcen)

- Eine Richtlinie zur Einhaltung des Unionsrechts in Bezug auf Urheberrecht und verwandte Schutzrechte gemäß Art. 53 Abs 1 lit c des EU AI Acts.
- Kriterien für die Bewertung systemischer Risiken
    - Anzahl der Parameter: Geben Sie die Anzahl der Parameter für das Modell an und erläutern Sie deren Bedeutung für die Fähigkeiten des Modells.
    - Unterstützte Modalitäten: Listen Sie alle Modalitäten auf, die das Modell unterstützt (z. B. Text-zu-Text, Text-zu-Bild).
    - Schwellenwerte für Fähigkeiten mit hoher Wirkkraft: Beschreiben Sie die dem Stand der Technik entsprechenden Benchmarks und Schwellenwerte für Fähigkeiten mit hoher Wirkkraft in jeder Modalität.
    - Benchmark-Ergebnisse: Geben Sie alle durchgeführten Benchmark-Tests und die Anpassungsfähigkeit des Modells an neue Aufgaben an.
    - Reichweite und Nutzerbasis: Geben Sie den Umfang der Marktreichweite an, einschließlich Statistiken über registrierte geschäftliche Nutzer in der EU (Schwellenwert von 10.000+ als Kriterium).
    - Anzahl der registrierten Endnutzer: Dokumentieren Sie die Anzahl der Endnutzer, die mit dem Modell interagieren oder von ihm profitieren.

Falls die GPAI ein systemisches Risiko darstellt:

- Adversarial Testing und Modellanpassungen
    - Beschreibung der Adversarial Testing Verfahrens: Erläutern Sie, ob Adversarial Testing durchgeführt wurden, und beschreiben Sie die Verfahren, z. B. interne Tests, Red-Teaming.
    - Maßnahmen für Red Teaming: Beschreiben Sie den spezifischen Aufbau und die getesteten Szenarien, wie z. B. Edge-Case-Eingaben oder Situationen zur Ermittlung von Schwachstellen.
    - Maßnahmen zur Anpassung und Feinabstimmung: Beschreiben Sie die Techniken, die eingesetzt werden, um die Übereinstimmung des Modells mit ethischen, rechtlichen und Leistungsstandards zu verbessern, z. B. Verstärkungslernen mit menschlichem Feedback (RLHF).
    - Externe Zusammenarbeit beim Testen: Wenn externe Tester oder Feedback aus der Gemeinschaft genutzt wurden, geben Sie Einzelheiten zu den Beiträgen und der Integration an.

Glossar: Begriffe und Definitionen

Konfidenzintervall
Ein Konfidenzintervall ist ein Bereich, der den zu schätzenden Parameter mit einer bestimmten Konfidenz umfasst.

Wird häufig in statistischen Tests verwendet.

Grenzfall (Edge Case)
Ein Beispiel, das nahe an der Entscheidungsgrenze bei Klassifikationsproblemen liegt.

Erklärbarkeit
Eigenschaft eines ML-Systems, die es Menschen ermöglicht, die Faktoren zu verstehen, die das Ergebnis des ML-Systems beeinflussen.

Erklärbarkeit kann durch den Einsatz eines Methodensets wie LIME, SHAP usw. erreicht werden.

Hyperparameter
Eigenschaften eines maschinellen Lernalgorithmus, die den Lernprozess beeinflussen. Sie werden vor dem Training festgelegt.

Beispiele für Hyperparameter sind: Anzahl der Schichten eines Neuronalen Netzes, Breite jeder Schicht, Art der Aktivierungsfunktion, Optimierungsmethode, Lernrate für neuronale Netze; Wahl der Kernelfunktion in einer Support-Vektor-Maschine; Anzahl der Blätter oder die Tiefe eines Entscheidungsbaums; das K für K-Means-Clustering; die maximale Anzahl von Iterationen des Erwartungs-Maximierungs-Algorithmus; die Anzahl der Gauss-Funktionen in einer Gaussian-Mischung.

Robustheit
Die Fähigkeit eines KI-Moduls, mit fehlerhaften, verrauschten, unbekannten oder absichtlich manipulierten Eingabedaten umzugehen.

Die zwei Arten der Robustheit umfassen adversarielle Robustheit (AR) und Korruptionsrobustheit (CR). AR: Qualität eines Modells, mit adversariellen Angriffen oder Störungen (böswilligen Änderungen der Daten durch Dritte) umzugehen. CR: Fähigkeit des Modells, mit unbeabsichtigten Abweichungen umzugehen, z. B. Unterschieden zwischen Trainingsdaten und Einsatzdaten.

Testdatensatz
Datensatz, der verwendet wird, um die Leistung eines finalen ML-Modells zu bewerten.

Training
Prozess zur Bestimmung der Parameter eines ML-Modells, basierend auf einem ML-Algorithmus, der auf den Trainingsdatensatz angewendet wird.

Trainingsdatensatz
Datensatz, der verwendet wird, um das ML-Modell zu trainieren.

Validierungsdatensatz
Datensatz, der vor, während und nach dem Training verwendet wird, um die Modellklasse auszuwählen und Hyperparameter zu optimieren.