Positionspapier der DSK zu empfohlenen technischen und organisatorischen Maßnahmen bei der Entwicklung und dem Betrieb von KI-Systemen

- Fassung vom 6. November 2019 -

1 Allgemeine technische und organisatorische Anforderungen an KI

1.1 Einordnung

In diesem Dokument wird aus technischer Sicht unter dem Begriff KI die Anwendung von Verfahren des maschinellen Lernens und der Einsatz von KI-Komponenten verstanden, mit denen diese Verfahren umgesetzt werden. Es existiert eine Vielzahl unterschiedlicher Verfahren des maschinellen Lernens mit unterschiedlichen Eigenschaften und Spezifika. Für jedes dieser Verfahren ergibt sich jeweils eine mehr oder minder große Menge an möglichen Einsatzgebieten[1].

Für den konkreten Einsatz von KI-Systemen ergeben sich unterschiedliche Möglichkeiten zur Anwendung von KI-Komponenten. Diese erstrecken sich von der Wahl des konkreten Verfahrens des maschinellen Lernens über die konkrete Ausgestaltung der Nutzung des Verfahrens, die entsprechende technische Umsetzung sowie die Auswahl von Trainingsdaten und deren Nutzung bis hin zum tatsächlichen Einsatz der trainierten KI-Komponente. Hinzu kommen Aspekte der Überprüfung und etwaiger Anpassung von KI-Komponenten sowie ggf. weiterer Trainingsphasen.

Die folgende Lebenszyklusbetrachtung von KI-Systemen wird in diesem Positionspapier zugrunde gelegt, die in Abbildung 1 illustriert werden:

1. Design des KI-Systems und deren KI-Komponenten
2. Veredelung der Rohdaten zu Trainingsdaten
3. Training der KI-Komponenten
4. Validierung der Daten und KI-Komponenten sowie angemessene Prüfungsmethoden
5. Einsatz und Nutzung des KI-Systems
6. Rückkopplung von Ergebnissen und Selbstveränderung des Systems

[1] In diesem Dokument wird der Begriff KI-System als übergeordneter Begriff verwendet. Ein KI-System beinhaltet in der Regel mindestens eine KI-Komponente. Unter einer KI-Komponente wird eine Abbildungsvorschrift verstanden, die durch Verfahren maschinellen Lernens (siehe Abbildung 1) gebildet wurde.

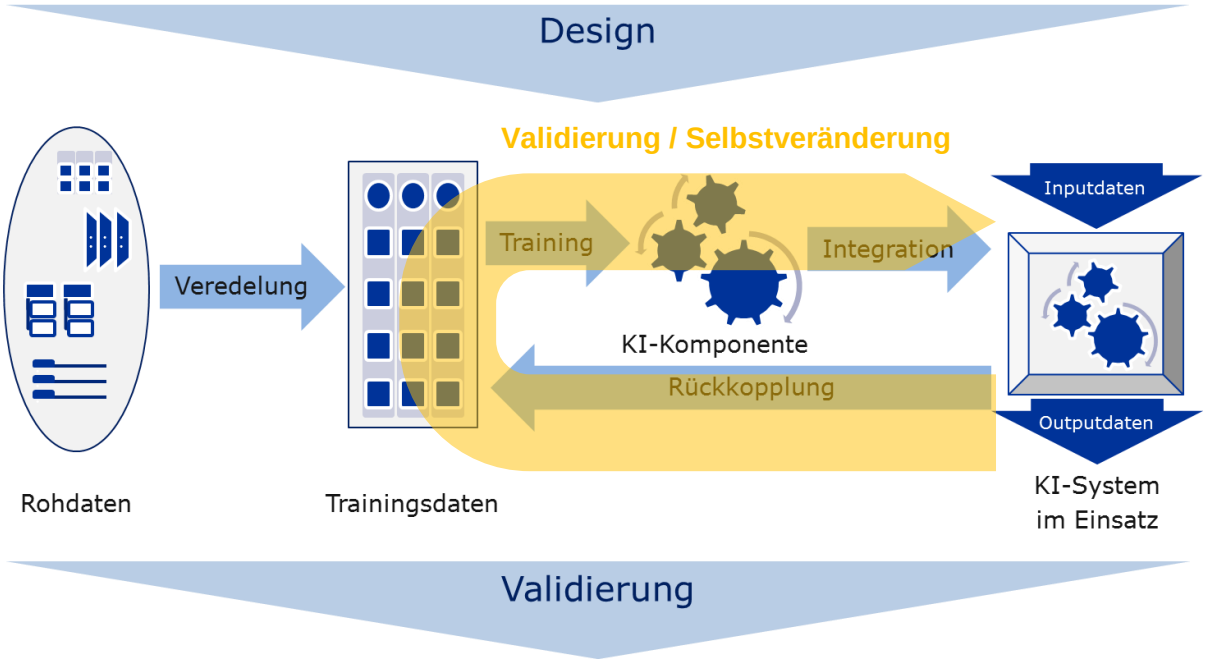

Abbildung 1 - Allgemeiner Lebenszyklus von KI-Systemen

Das Einsatzgebiet und die konkrete Anwendung von KI-Systemen sowie die Ausgestaltung der o. g. Aspekte haben Auswirkungen auf die Verarbeitung personenbezogener Daten. Dies gilt insbesondere auch für Art und Umfang der verarbeiteten personenbezogenen Daten, deren Verarbeitung in KI-Komponenten und -Systemen sowie resultierender Risiken für Rechte und Freiheiten betroffener Personen.

Damit die Risiken für die Rechte und Freiheiten der von KI-Verfahren betroffenen natürlichen Personen identifiziert werden können, ist es erforderlich zu unterscheiden, welche Art von Unterstützung die KI-Systeme liefern sollen. Dies erfordert eine spezielle Art der Risikodefinitionen von KI-Systemen. So sind die Risiken von KI-Systemen, die bspw. lediglich Kaufempfehlungen zu einzelnen Produkten liefern sollen, anders zu bewerten als die Risiken von KI-Systemen, die etwa Entscheidungen im Bereich des autonomen Fahrens treffen. Entsprechend ist vor dem Einsatz einer KI-Komponente zu definieren, ob man sich in Bezug auf relevante Eigenschaften von KI-Systemen auf die Zusicherungen des Lieferanten der Komponente verlassen darf oder ob diese Komponenten selbst geprüft werden müssen.

Um den Rahmen des Dokumentes nicht zu sprengen, kann nicht auf einzelne Verfahren maschinellen Lernens eingegangen werden. Gleiches gilt für konkrete Anwendungsgebiete, Einsatzszenarien sowie Architekturaspekte. Die nachfolgende Tabelle soll lediglich eine Übersicht über die KI-Verfahren bzw. -Methoden geben, die in diesem Dokument datenschutzrechtlich beleuchtet werden. Sie ist der Broschüre „MASCHINELLES LERNEN - EINE ANALYSE ZU KOMPETENZEN, FORSCHUNG UND ANWENDUNG“ der Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V. (München 2018) entnommen. Konkrete KI-Verfahren werden in diesem Dokument nur angesprochen, wenn dies für eine differenzierte datenschutzrechtliche Betrachtung sinnvoll ist.

Lernstil	Lernaufgabe	Lernverfahren	Modell
Überwacht	Regression	Lineare Regression	Regressionsgerade
		Klassifikations- und Regressionsbaum-verfahren (CART)	Regressionsbaum
	Klassifikation	Logistische Regression	Trennlinie
		Iterative Dichotomizer (ID3)	Entscheidungsbaum
		Stützvektormaschine (SVM)	Hyperebene
		Bayessche Inferenz	Bayessche Modelle
Unüberwacht	Clustering	K-Means	Clustermittelpunkte
	Dimensionsreduktion	Kernel Principal Component Analysis (PCA)	Zusammengesetzte Merkmale
Bestärkend	Sequentielles Entscheiden	Q-Lernen	Strategien
Verschiedene	Verschiedene	Rückwärtspropagierung	Künstliche Neuronale Netze

Abbildung 2: Verwendete Methoden der von Kaggle befragten Data Scientists und ML-Fachleute[2]

In der Tabelle in Abbildung 2 werden Verfahren des maschinellen Lernens in Bezug auf den verwendeten Lernstil (überwacht, unüberwacht, bestärkend) unterscheiden.

Ziel des überwachten Lernens ist es, ein KI-System mit Trainingsdaten so lange zu trainieren, bis das erwartete Ergebnis geliefert wird. Die richtige Antwort muss während der Trainingsphase also bereits vorliegen. Beim überwachten Lernen ist eine sogenannte Grundwahrheit (ground truth) bekannt. Ein „Lehrer“ kontrolliert das KI-System, ob es richtig oder falsch gelegen hat. Das KI-System wird so lange trainiert, bis es die gewünschte Leistungsfähigkeit erreicht hat.

Der Zweck des unüberwachten Lernens besteht in der Regel darin, aus Daten Wissen zu erzeugen. Beim unüberwachten Lernen ist beim Design des KI-Systems zunächst nicht

[2] Kaggle 2017

bekannt, was es erkennen soll. Es erkennt Muster und teilt die Daten in Cluster oder Kategorien auf, jedoch ohne dass vorgegeben wurde, um welche Kategorien es sich handelt bzw. unter welches Label sie fallen. Wenn Daten sich bspw. mit Hilfe einer Faktoren- oder Clusteranalyse zu unterschiedlichen Clustern häufen lassen, stehen diese Cluster für bestimmte Dimensionen bzw. für bestimmte Inhalte, mit denen sich theoretisch gestützt ein Modell für treffsichere Prognosen bilden lässt. In der Praxis ist eine Clusterbildung wichtig, um anhand möglichst weniger aber relevanter Variablen erfolgreiche Prognosen durchzuführen. Dabei ist mit einem definierten Restfehler aufgrund einer bestimmten Annahme bzgl. einer Fehlerverteilung zu rechnen. Anhand solch theoretisch gestützter Einsichten können dann Datensätze bereinigt, ergänzt und vereinheitlicht werden, mit denen die Trainings von KI-Systemen auf den gewünschten Zweck hin optimiert werden.

Eine besondere Form des überwachten Lernens ist das bestärkende Lernen. Es wird genutzt, wenn ein Endergebnis noch nicht bestimmbar ist, jedoch der Trend hin zum Erfolg oder Misserfolg erkennbar wird. In der Trainingsphase werden beim bestärkenden Lernen die korrekten Ergebnisse also nicht zur Verfügung gestellt, jedoch wird jedes Ergebnis bewertet, ob dieses in die richtige oder falsche Richtung geht.

Ein KI-System durchläuft in der Regel mehrere Stadien:

- von der untrainierten, zufällig initialisierten Verarbeitung
- zu einer trainierten Verarbeitung, welche noch verifiziert werden muss,
- über eine verifizierte Verarbeitung im Echtbetrieb
- bis zu einem iterativ verbesserten Verfahren.

Die ersten beiden Stadien werden die Lern-Phase genannt, die letzten beiden Stadien werden Kann-Phase genannt. Im Folgenden soll für alle Methoden des maschinellen Lernens dargestellt werden, welche technischen und organisatorischen Maßnahmen erforderlich sind, um den mit Hilfe der Gewährleistungsziele systematisierten Anforderungen des Datenschutzrechts gerecht zu werden. Dieses Dokument verfolgt das Ziel, allgemeine und grundsätzliche datenschutzrechtliche Anforderungen aus technischer Sicht für den Einsatz von KI-Systemen aufzustellen. Es fokussiert auf spezifische zusätzliche Anforderungen an KI-Systeme und die in diesen genutzten KI-Komponenten. Grundsätzlich gilt, dass datenschutzrechtliche Anforderungen an IT-Systeme ohne KI-Komponenten auch für KI-Systeme gelten. Diese Anforderungen werden in Bezug auf KI-Systeme um die in diesem Dokument enthaltenen Anforderungen ergänzt.

1.2 Rechtliche Bezüge aus technischer Sicht

Für Verarbeitungstätigkeiten mit Personenbezug (Definition: vgl. Art. 4 Nr. 2 DS-GVO), bei denen Komponenten der Künstliche-Intelligenz (KI-Komponenten) zum Einsatz kommen, gelten die in der DS-GVO formulierten Grundsätze (vgl. Art. 5 Abs. 1 DS-GVO).

KI-Systeme können in vielfältiger und teils nur schwer erkennbarer, vorhersehbarer oder beweisbarer Art und Weise Risiken für die Freiheiten und Rechte natürlicher Personen darstellen. Diese Risiken hängen maßgeblich vom Einsatz-Szenario sowie von den verwendeten KI-Komponenten ab. Zur Verringerung dieser Risiken muss der Verantwortliche KI-spezifische Maßnahmen definieren, implementieren und betreiben. Der Zweck einer Verarbeitungstätigkeit mit einer KI-Komponente muss berechtigt sein, die Verarbeitung muss eine Rechtsgrundlage aufweisen und die bei der Verarbeitung entstehenden Risiken müssen minimiert werden. Die Verwendung von KI-Systemen erzeugt in der Regel hohe Risiken für die Rechte und Freiheiten der Betroffenen. Deshalb bestehen hohe Anforderungen an die

technischen und organisatorischen Maßnahmen und dabei insbesondere an die Transparenz der Datenverarbeitung.

Unter Sicherung der Transparenz versteht man im Datenschutz insbesondere die Herstellung der Prüfbarkeit einer Verarbeitungstätigkeit. So muss der Verantwortliche den Nachweis darüber erbringen, dass die von ihm verantwortete Verarbeitungstätigkeit die Anforderungen der DS-GVO wirksam umsetzt. Gegenstand der Transparenz sind die funktionalen Eigenschaften sowie die technischen und organisatorischen Maßnahmen einer Verarbeitungstätigkeit. Zur Durchsetzung der Transparenz gehören u. a. die Erstellung von Datenschutzerklärungen, Einwilligungstexten und Verträge sowie Spezifikationen, Dokumentationen, Logs und Protokolldaten sowie insbesondere aktive Tests der KI-Systeme oder -Komponenten. In KI-Systemen werden in der Regel automatisierte Entscheidungen getroffen oder vorbereitet. Eine sorgfältige Risikobetrachtung ist daher immer erforderlich. Vor allem wenn die Auswirkungen automatisierter Entscheidungen oder Entscheidungsvorbereitungen voraussichtlich hohe Risiken für die Rechte und Freiheiten natürlicher Personen zur Folge haben, muss eine Datenschutz-Folgenabschätzung gem. Art. 35 DS-GVO durchgeführt werden.[3] Im Falle einer systematischen und umfassenden Bewertung persönlicher Aspekte natürlicher Personen wird oftmals auch das Regelbeispiel aus Art. 35 Abs. 3 DS-GVO vorliegen. Grundsätzlich muss nachvollziehbar sein, welche Daten verarbeitet wurden, welche Programme und Systeme zum Einsatz kommen und wie diese organisatorisch in die Verarbeitungstätigkeit eingebunden sind. Die speziellen Funktionsweisen der unterschiedlichen KI-Komponenten müssen erklärt werden können. Hier sind die Hersteller von KI-Systemen besonders gefordert.

Besonders hohe datenschutzrechtliche Anforderungen resultieren aus der Nutzung von Verfahren des maschinellen Lernens, in denen sehr viele Daten einer Wissensdomäne in einem subsymbolischen Aktivitätsmuster abgebildet werden: Die Daten und die erlernten Zusammenhänge sind dort in vielen Zahlen versteckt, die keinen Einblick in die erlernten Lösungswege erlauben. Das betrifft insbesondere Neuronale Netze. Bei derartigen Verfahren kann nicht ausgeschlossen werden, dass Menschen mit bestimmten Eigenschaften benachteiligt (ggf. auch ungerechtfertigt bevorteilt) werden, ohne dass diese unerwünschte Diskriminierungseigenschaft hieraus ersichtlich ist. In entsprechenden KI-Komponenten können negative Diskriminierungen von Personengruppen versteckt sein.

Erstmals zum Zeitpunkt der Festlegung der Zwecke und Mittel der Verarbeitung, also während der Planung und Spezifikation der KI-Systeme, ist zu prüfen, welche Ergebnisse eines KI-Systems als angemessen und korrekt gelten sollen. Es ist für die Spezifikation eines KI-Systems unerlässlich, im Vorhinein die Erwartungen der verschiedenen Beteiligten bei der Nutzung eines KI-Systems eindeutig zu beschreiben. Der Zweck eines KI-Systems muss so eng wie möglich beschrieben sein, vorzugsweise in einer maschinell zugänglichen Policy. Sofern es entsprechend standardisierte Definitionen für derartige Policies gibt, müssten diese dann konkrete funktionale Anweisungen bzgl. Funktionen und technische und organisatorische Maßnahmen ergeben, mit denen sich Regelverstöße, Zweckdehnungen und Zweckverletzungen im Betrieb feststellen und dokumentieren lassen.

Zuvor muss inhaltlich festgelegt sein, welche Diskriminierungen im Sinne negativer Unterscheidungen als rechtlich nicht erlaubt gelten und die sich folglich auch nicht auf Entscheidungen und Prognosen des KI-Systems auswirken dürfen.

[3] Siehe Kurzpapier Nr. 5 (DSFA) der Datenschutzkonferenz

KI-Systeme lassen sich in deterministische (d. h. vollständig vorhersehbare) und nicht-deterministische Systeme unterteilen. Nicht-deterministische Systeme, deren Entscheidungsweise nicht vollständig vorhergesagt werden kann, bergen ein Risiko für unvorhergesehene Fehlentscheidungen. Diese unvollständige Vorhersagbarkeit gilt insbesondere für im laufenden Betrieb weiterlernende Systeme. Aufgrund dieses Risikos ist bei vergleichbarer Effizienz und Effektivität eines deterministischen und eines nicht-deterministischen Systems das deterministische zu verwenden. Dabei kann ein deterministisches System auch mithilfe eines nicht-deterministischen entwickelt oder von diesem abgeleitet werden.

Allerdings sind KI-Systeme gerade deshalb so erfolgreich, weil deren Effizienz und/oder Effektivität sowohl klassische Algorithmen als auch menschliche Fähigkeiten übertreffen kann. Liegt die Leistungsfähigkeit des nicht-deterministischen Systems höher, so kann dieses nach sorgfältiger Abwägung der Risiken zum Einsatz kommen. Innerhalb der Begründung sind von dem jeweils Verantwortlichen die Leistungsüberlegenheit quantitativ darzulegen sowie eventuelle Risiken mit deren Eintrittswahrscheinlichkeit.

Bei einer Auswahl zwischen ansonsten gleichwertigen Alternativen sind solche KI-Komponenten zu bevorzugen, die einem menschlichen Verständnis (z.B. in Form von theoretisch gestützten oder kausalen Zusammenhängen) zugänglich sind und somit besser nachvollziehbar sind. Verantwortlichen ist eine Begründung abzuverlangen, wenn beispielsweise anstelle von einfacher nachvollziehbaren Modellen wie etwa Regressionsmodellen oder Entscheidungsbäumen komplexe Modelle wie neuronale Netze als KI-Komponente verwendet werden sollen. Denn deren Funktionieren ist ungleich undurchsichtiger und deren Risiken sind ungleich schwieriger abzuschätzen, als es grundsätzlich bei Regressionsmodellen und Entscheidungsbäumen der Fall ist.

Die Unvorhersehbarkeit von KI-Systemen kann im Zweifel soweit gehen, dass ihr Einsatz auch andere als den vorgesehenen Zweck verfolgt. Damit ein KI-System, das personenbezogene Daten verarbeitet, zuverlässig funktionieren kann, muss es so entwickelt sein und mit Daten trainiert werden, dass es dem ausgewiesenen Zweck folgt. Das schließt nicht aus, dass das KI-System anderen Zwecken dient, wenn die Voraussetzungen der Zweckkompatibilität vorliegen (siehe dazu auch Art. 6 Abs. 4 und Art. 5 Abs. 1 lit. b DS-GVO). Insbesondere bei im laufenden Betrieb noch lernenden Systemen sind schädliche Einflüsse nicht mehr nur durch klassische Maßnahmen der IT-Sicherheit auszuschließen. Auch deshalb ist sicherzustellen, dass es nur durch Befugte konzipiert, programmiert, trainiert, genutzt und überwacht wird. Weiterhin muss es Möglichkeiten zum Eingreifen in die Verarbeitung geben. Eine gravierende Möglichkeit des Eingriffs ist das Stoppen der Verarbeitung, das allerdings nur durch Befugte erfolgen darf. Schließlich muss sichergestellt sein, dass jederzeit Befugten, zu denen gem. Art. 15 Abs. 1 lit. h DS-GVO grundsätzlich auch die von der Entscheidung Betroffenen zählen, Auskunft darüber erteilt werden kann, wie Entscheidungen und Prognosen durch ein KI-System zustande gekommen sind.

2 Anforderungen an KI-Systeme bezogen auf Verarbeitungsschritte

Orientiert an den unter 1.1 eingeführten Lebenszyklen von KI-Systemen werden diese im Folgenden anhand der Gewährleistungsziele analysiert und Anforderungen definiert. Dabei ist zu beachten, dass angesichts der breiten Varianz an KI-Systemen im Einzelfall auch einzelne Aspekte einer anderen Phase zugeordnet werden können.

2.1 Design und Veredelung

In den ersten Phasen des Lebenszyklus eines KI-Systems ist zu klären, mit welchen Eingangs-Daten das KI-System arbeiten soll. Die Aufbereitung von Rohdaten zu Trainingsdaten für die repräsentative Modellierung eines zweckbestimmten, domänenspezifischen Wissens spielt dabei eine wesentliche Rolle.

Bezogen auf das Design eines KI-Systems müssen – analog zu Art. 25 DS-GVO – bereits alle datenschutzrechtlichen Anforderungen bei der Systementwicklung mitgedacht werden. Besonders gilt das für die Transparenzanforderungen.

Im Detail stellen sich aus Datenschutzsicht beim Einsatz von KI-Systemen in personenbezogenen Verarbeitungstätigkeiten neben den konventionellen datenschutzrechtlichen Fragen bzgl. der Verantwortung und der Rechtsgrundlage für eine Verarbeitungstätigkeit zumindest die folgenden Fragen:

- Zu welchen Zwecken werden ein KI-System und dessen KI-Komponenten eingesetzt?
- Welche KI-Modelle werden für die verwendeten KI-Komponenten genutzt?
- Sind an der Entscheidungsfindung bzw. Prognose durch eine KI-Komponente Menschen beteiligt und wenn ja in welcher Form?
- Welche Institutionen haben die Kontrolle über die Auswahl der KI-Modelle, der Implementation und der Trainingsmethoden?
- Welche Zielgrößen sind für eine KI-Komponente festgelegt?
- Wie wurde die KI-Komponente getestet, ob sie die zweckgemäßen Eigenschaften aufweist und wie wird der laufende Betrieb dieser Komponente im Hinblick auf die Einhaltung des Zwecks überwacht?
- Wird die verwendete KI-Komponente mit getestet-gesicherten Eigenschaften „eingefroren“ genutzt oder wird die Komponente im laufenden Betrieb mit fortlaufend eintreffenden Nutzungs-Daten trainiert? Gibt es Teilbereiche, die „eingefroren“ genutzt werden?
- Kann die KI-Komponente rein lokal durch den Anwender genutzt werden, oder sind Onlineverbindungen bspw. zum Verantwortlichen oder zum Hersteller der KI-Komponenten oder zu Dritten, die bspw. als Trainings-Provider agieren, notwendig?

Besteht eine (Online-)Verbindung des KI-Systems oder einzelner Komponenten bspw. zu einem Profiling-Unternehmen, zu Herstellern oder zu Sicherheitsbehörden?

2.1.1 Gewährleistungsziel Transparenz

Für ein KI-System ist zu fordern, dass die Herkunft der Rohdaten ausgewiesen wird: Welche Institutionen haben Rohdaten generiert? Welche Institutionen haben diese Rohdaten zu Trainingsdaten verarbeitet (veredelt, „kuratiert“)? In diesem Kontext ist zudem die Sicherheit der IT-Umgebung zur Produktion der Daten zu spezifizieren bzw. zu dokumentieren, um sicherzustellen, dass die Roh- und Trainingsdaten nicht unbefugt verändert wurden oder unbefugt abgeflossen sind. Ebenso ist der Zweck der Verarbeitungstätigkeit sowie die Rechtsgrundlage zu dokumentieren, mit denen diese Rohdaten erzeugt und gespeichert wurden und an welche Organisationen diese Daten übermittelt wurden bzw. welche Organisationen diese Daten abgerufen haben.

In Bezug auf den ausgewiesen angestrebten Zweck des Einsatzes einer KI-Komponente bzw. eines KI-Systems müssen die zugehörige Wissensdomäne sowie die statistischen

Methoden spezifiziert und dokumentiert werden, auf die hin die Trainingsdaten bearbeitet werden.

Zu spezifizieren und zu dokumentieren sind mindestens die nachfolgend genannten Maßnahmen, die insbesondere der Sicherung der Integrität der Roh- bzw. Trainingsdaten dienen, aber auch der konventionellen IT-Sicherheit in der gesamten Verarbeitungs- und Übermittlungskette. So muss der jeweils Verantwortliche bei der Planung spezifizieren und für den Betrieb zumindest dokumentieren, auf welcher theoretischen Grundlage und mit welcher Methode

- Rohdaten normalisiert und standardisiert werden,
- synthetische Daten generiert werden,
- ein Datenbestand komplettiert oder fehlerbereinigt wird,
- personenbezogene Daten pseudonymisiert und/oder anonymisiert werden,
- die Gesamtmenge an Roh- bzw. Trainingsdaten für das Grundtraining einer KI-Komponente hergestellt wurde und wie dabei sichergestellt wird, dass in der Kern-Entwicklungsphase keine Testdaten herangezogen werden,
- geregelt wird, bis zu welchem Entwicklungsstand die letztlich gültigen Testdaten herangezogen werden dürfen, bevor der Übergang vom Teststatus zum Produktivitätsstatus erfolgt,
- gesetzlich verbotene, negative Diskriminierungen unterbunden werden (zu dokumentieren ist, wenn die zur Diskriminierung führenden Daten gelöscht werden; zu dokumentieren sind auch Regelungen, mit denen die Verwendung hoch-korrelierender „Ersatz“-Daten für die zur Diskriminierung führenden Daten unterbunden wird),
- die Relevanz bzw. Repräsentativität der Trainingsdaten für die Wissensdomäne bestimmt wird, unter Ausweis des Fehlermodells und der daran gemessenen verbliebenen Fehlerrate.

2.1.2 Gewährleistungsziel Datenminimierung

Soweit möglich sollten Daten ohne Personenbezug (synthetische bzw. hinreichend anonymisierte Daten) verwendet werden.

Bereits im Laufe der Sichtung von Rohdaten sollte auf wenige, inhaltlich gut verstandene Dimensionen reduziert werden. Dabei ist möglichst sicherzustellen, dass nur sehr hoch korrelierende Daten, d. h. solche Daten, die in einer engen Wechselbeziehung stehen, verwendet werden und das Korrelationsmaß nachgewiesen wird. Dabei ist jedoch zu berücksichtigen, dass eine zu einseitig verstandene, nicht ausgewogene Minimierung von Daten die Integrität der Modellierung von KI-Systemen gefährden kann. Die Datenmengen, die für ein Training mit einem akzeptablen Fehler erforderlich sind, um die ausgewiesenen Zielgrößen des Systemverhaltens zu erreichen, sollten bei der Spezifikation eines KI-Systems theoriegestützt abgeschätzt werden. Zwar kann man mit „beliebigen Daten“ in „beliebig großen Mengen“ versuchen, die wesentlichen Merkmale einer nur schlecht verstandenen Wissensdomäne zu identifizieren, jedoch vergrößert dies die Risiken für die Rechte und Freiheiten von betroffenen Personen: Wenn KI-Komponenten mit Kategorien von Daten trainiert werden, deren Relevanz für die Wissensdomäne nicht geklärt ist, können in der Folge bei Ihrem Einsatz in KI-Systemen Risiken entstehen. Diese Risiken können beispielsweise darin bestehen, dass auf Basis der Datenkategorien, wie dem Geschlecht, das KI-System diskriminierende oder fehlerhafte Ergebnisse liefert. Daher sollte der jeweilige Verantwortliche vorab eine Hypothese zum erwarteten Einfluss der verwendeten

Datenkategorien auf das Verhalten der KI-Komponente formulieren und diese im Laufe des Trainings und während des Einsatzes evaluieren.

2.1.3 Gewährleistungsziel Nichtverkettung

Die Verarbeitung von personenbezogenen Daten für das Trainieren von KI-Komponenten stellt einen eigenen Verarbeitungszweck dar. Es dürfen nur solche Daten verwendet werden, die dem unmittelbar ausgewiesenen Zweck dienen. Für einen anderen Zweck dürfen die Daten verwendet werden, wenn die Voraussetzungen einer Zweckänderung vorliegen oder eine ausdrückliche Rechtsgrundlage besteht.Das setzt voraus, dass man die Daten der wesentlichen Dimensionen der zu modellierenden Wissensdomäne statistisch auf einem hinreichend akzeptablen Fehlerniveau erfassen kann bzw. erfasst hat.

Wenn gesetzlich geregelte Diskriminierungsverbote die Verwendung bestimmter Daten nicht zulassen und stattdessen hochkorrelierende Ersatzvariablen verwendet werden, kann negative Diskriminierung dennoch nicht ausgeschlossen werden. Derartige Diskriminierungen sind denkbar, wenn bspw. anstelle des Geschlechts stark korrelierende Merkmale (bspw. Vorname, Gewicht, gekaufte Produkte oder die Kombinationen daraus) herangezogen werden. Um ein derartiges Diskriminierungspotenzial zu erkennen und ggf. vermeiden zu können, müssen KI-Systeme und ihre einzelnen Komponenten frühzeitig und dauerhaft auf ihre Diskriminierungseigenschaften hin geprüft werden.

Zu prüfen ist zudem, mit welcher Arbeitsteilung unterschiedliche KI-Komponenten genutzt werden können.

2.1.4 Gewährleistungsziel Intervenierbarkeit

Es muss gewährleistet werden, dass Rohdaten, die falsch sind oder deren Nutzung erkennbar negativ-diskriminierend wirkt, gar nicht erst zum Generieren von Trainingsdaten herangezogen sondern gelöscht oder ggf. korrigiert werden.

2.1.5 Gewährleistungsziel Verfügbarkeit

Durch konventionelle Maßnahmen der IT-Sicherheit ist sicherzustellen, dass bei der Produktion, dem Speichern oder der Übermittlung die Verfügbarkeit der Roh-, Eingabe- und Trainingsdaten gewährleistet ist.

2.1.6 Gewährleistungsziel Integrität

Die Wissensdomäne, für die ein KI-System eingesetzt werden soll, muss entsprechend definiert und von anderen Wissensdomänen abgegrenzt werden. Eine vollständige Repräsentation durch eine Modellierung ist zwar anzustreben, aber voraussichtlich in den meisten Fällen nicht realistisch. Ebenso wenig wie Komplexität vollständig modelliert werden kann, kann aus Daten der Vergangenheit die Zukunft verlässlich prognostiziert werden. Deshalb muss abgeschätzt werden, mit welchem Fehlerniveau die vorhandenen Rohdaten eine Wissensdomäne repräsentieren. Schon zum Zeitpunkt der Datenveredelung muss beurteilt werden, zu welchen Risiken für die Rechte und Freiheiten natürlicher Personen der Einsatz eines KI-Systems, das mit diesen Daten trainiert wurde, führen wird.

Damit Rohdaten bei der Veredelung zu Trainingsdaten nicht verfälscht werden, sind folgende Aktivitäten erforderlich:

- Normalisierung und Standardisierung der Rohdaten,
- Komplettierung und Fehlerbereinigung des Rohdatenbestands,

- Identifikation oder Herstellung von „störsignalbehafteten Daten“, mit denen KI-Systeme getestet werden müssen, um deren Robustheit („Resilienz“) nachzuweisen,
- Aktivitäten, mit denen auf Änderungen in der Wissensdomäne (Änderungen von Kontexten, rechtlichen Regelungen, technischen Änderungen, Zunahme des Wissens) reagiert werden kann,
- Festlegung von Verfahren, mit denen Änderungen der Wissensdomäne identifiziert werden
- Einteilung der Gesamtmenge an Roh- bzw. Trainingsdaten in
  - Primär-Trainingsdaten (ausschließlich zum Training des KI-Systems),
  - Verifikationsdaten,
  - Sekundär-Trainingsdaten (im Einsatz) und
  - Testdaten,

  mit jeweils theoretisch begründet angemessenen Anteilen von der Gesamtdatenmenge,
- Untersuchung, ob synthetische Daten, sofern sie zum Training verwendet werden sollen, als Ersatz für echte Daten geeignet sind (siehe oben Datenminimierung).

Mit den Primär-Trainingsdaten wird eine Repräsentation des Modells ausgebildet, mit den Verifikationsdaten werden Fehlerraten ermittelt und Korrekturstrategien entwickelt. Testdaten werden verwendet, um im Übergang von der Trainings- zur Produktionsphase eines KI-Systems eine Fehlerquote auszuweisen. Dazu muss eine hinreichend große Menge an Testdaten zur Verfügung stehen. Durch technische und organisatorische Maßnahmen ist sicherzustellen, dass diese weder zum eigentlichen Training noch zum Verifizieren verwendet werden. Im laufenden Einsatz eines KI-Systems werden ggf. die Eingabedaten auch für ein fortlaufendes Training des KI-Systems genutzt (Sekundär-Trainingsdaten), diese unterliegen prinzipiell den gleichen Anforderungen wie Trainingsdaten. Es wäre erstrebenswert, wenn wissenschaftlich abgesichert Standard-Testdatensätze für verschiedene Wissensdomänen entwickelt würden, mit denen unabhängige Institutionen Tests durchführen und Fehlerraten ermitteln.

Beim Design eines KI-Systems ist – bezogen auf die erwarteten Trainings- und Einsatzdaten – eine Robustheit des Systems einzuplanen gegenüber falschen Daten oder solchen, mit denen die Manipulation bzw. Änderung des Systemverhaltens beabsichtigt ist.

Wenn bei einer KI-Komponente keine Fehlerrate ausgewiesen werden kann, sollten nur sehr gezielt relevante Trainingsdaten verwendet werden, um die bis dahin bereits erreichten, stabilen Systemeigenschaften vorausliegender Trainingseinheiten nicht zu gefährden.

Mit einer zu geringen Menge an Roh- bzw. Trainingsdaten kann kein hinreichend integres Systemverhalten erreicht werden, wenn zudem nicht abgeschätzt werden kann, wie vollständig die Grundgesamtheit ist. Die Menge der Trainingsdaten muss hinreichend repräsentativ sein. Repräsentativität zu erreichen ist statistisch möglich, wenn man die Grundgesamtheit und die Fehlerverteilung kennt und dann zufällig eine bestimmte Anzahl an Daten erhebt.

Die Verwendung von Trainingsdaten kann zu dem sog. „katastrophischen Vergessen“ führen, d.h. dass ein bislang stabiles Systemverhalten einer KI-Komponente kippt bis dahin stabile Eigenschaften der KI-Komponente, z.B. der Erkennung von Mustern, plötzlich nicht mehr vorhanden sind.

Durch konventionelle Maßnahmen der IT-Sicherheit ist sicherzustellen, dass Unbefugte bei der Produktion, dem Speichern oder der Übermittlung der Rohdaten diese nicht verändern können. Dazu gehört auch, dass nur Befugte Rohdaten zu integren Trainingsdaten veredeln können. Sie müssen in der Lage sein, den Veredelungsprozess zu beurteilen, sowohl wenn er durch Andere durchgeführt wurde als auch wenn sie ihn selber durchgeführt haben. Dies soll sicherstellen, dass die Datenaufbereitung zweckbestimmt, mit relevanten Daten und in einer korrekten Form geschieht. In jedem Fall muss sichergestellt werden, dass ein KI-System nicht mit unbefugt manipulierten Daten trainiert wird.

2.1.7 Gewährleistungsziel Vertraulichkeit

Durch konventionelle Maßnahmen der IT-Sicherheit ist sicherzustellen, dass Unbefugte bei der Produktion, dem Speichern oder der Übermittlung der Rohdaten keinen Zugriff auf diese Daten nehmen und diese zu anderen Zwecken nutzen können.

2.2 Training und Validierung der KI-Komponente

Die nächsten Phasen des Lebenszyklus einesKI-Systems nach der Definition und Vorverarbeitung der Inputs ist die eigentliche Verarbeitung durch das Verfahren des maschinellen Lernens, das sogenannten Training und der damit verbundenen Validierung der KI-Komponente. Eine Übersicht über KI-Verfahren und –Methoden ist der Tabelle in Abbildung 1 im Abschnitt 1 zu entnehmen.

2.2.1 Gewährleistungsziel Transparenz

Der Grad der herzustellenden Transparenz über die Methoden der verwendeten KI-Systeme hängt von der Zielgruppe ab (Betroffene, Verantwortliche, Aufsichtsbehörden). Deshalb sollte eine Abstufung in unterschiedliche Level von Transparenz erfolgen. Während Betroffene meist weniger Detailtiefe benötigen und vor allem die Information benötigen, dass ein KI-System eingesetzt wurde und wo ggf. weitere Details in Erfahrung gebracht werden können, benötigen Verantwortliche nachvollziehbare Erläuterungen für alle Schritte des KI-Systems. Prüfende Institutionen wie Aufsichtsbehörden benötigen hingegen weitere Informationen, etwa um abschätzen zu können, wie sicher das KI-System zu Ergebnissen kommt. Generell wird es daher als wenig zielführend angesehen, ausschließlich über die gewählte Methode des maschinellen Lernens zu informieren. Diese ist für Verfahrensfremde nicht ausreichend. Häufig sind die Systeme nicht ausreichend nachvollziehbar. So verarbeiten z. B. neuronale Netze Informationen verteilt und einzelne Subnetze müssen damit keine erklärbaren Zwischenergebnisse erzeugen. Vielmehr ist es sinnvoll, Transparenz im Sinne der Güte und Erklärbarkeit des KI-Systems herzustellen. Dies charakterisiert das genutzte KI-System unabhängig von den eigentlich eingesetzten Lernverfahren.

Die Güte des KI-Systems kann im Fall des überwachten Lernens über Fehlerraten angegeben werden. Da hier die Soll-Ergebnisse in jedem Fall bekannt sind, kann auch der Fehler des KI-Systems gut bestimmt werden. So muss einerseits angegeben werden, welcher Datensatz zur Ermittlung des Fehlers genutzt worden ist (also ob es einen speziellen Verifikationsdatensatz für diesen Zweck gibt, wie groß dieser ist und evtl. wie sich dieser von den Trainingsdaten unterscheidet) und andererseits eine Beschreibung der erzielten Güte.

Im Fall der Nutzung von unüberwachten Lernverfahren ist eine Fehlerangabe oft schwieriger. Hier wird im Entwurfsprozess meist nur die Ausgabedimension festgelegt, nicht jedoch die Ergebnisklassen oder eine andere Ausgabestruktur. Die Ergebnisse müssen daher durch menschliche Aktion noch nachträglich interpretiert oder geprüft werden. Hier muss im Einzelfall ein Verifikationsmodell erstellt werden, wie die Güte des Systems beurteilt werden kann.

Dem Betroffenen sollte sowohl die Beschreibung des Verfahrens der Fehlerbestimmung als auch das entstandene Fehlermaß transparent gemacht werden.

Wie oben bereits erwähnt, ist die Erklärbarkeit eines Lernverfahrens nicht in jedem Fall gegeben. So ist für Lernsysteme häufig nur die Konvergenz der mathematischen Abbildung bewiesen. Allerdings ist dies nicht mit einer semantischen Erklärbarkeit gleichzusetzen, in welcher klar erkennbar ist, welche Teile einer Methode welche Zusammenhänge abbilden. Ansätze, die dies ermöglichen könnten, sind noch Forschungsgegenstand. Es sollte jedoch geprüft werden, ob eine Möglichkeit besteht, die Erklärbarkeit durch die Annäherung eines komplexen Systems an ein einfacheres System herzustellen. In einfachen Fällen kann so eine Erklärbarkeit der Abbildungsfunktion des Lernsystems herbeigeführt werden. Sollte dies nicht möglich sein, sollte zumindest die Stabilität des KI-Systems nachgewiesen sein sowie erklärt werden können, dass es keine semantisch sinnvoll interpretierbaren Teilergebnisse gibt und welche Analysemethode zu diesem Ergebnis geführt hat.

Wenn nicht erklärt werden kann, welche Teile einer KI-Komponente zur Entscheidung bzw. Entscheidungsvorbereitung beitragen – bspw. aufgrund der Struktur oder Komplexität der KI-Komponente, sollten die Eigenschaften der KI-Komponente mit sogenannten Black-Box-Tests untersucht werden. Beim Black-Box-Test einer KI-Komponente werden synthetische Testdaten erzeugt, mit denen geprüft wird, welchen Einfluss die Eingabeparameter auf die Ausgabe der KI-Komponente haben. Auf diese Weise können statistische Aussagen zum Verhalten einer KI-Komponente getroffen werden. Mit diesen statistischen Aussagen können auch die Hypothese zum Zusammenhang zwischen Eingabe- und Ausgabeparametern, das von den Nutzern erwartete Verhalten des KI-Systems und diskriminierende oder andere unerwünschte Eigenschaften des KI-Systems beurteilt werden. Der Vorteil von Black-Box-Tests ist, dass diese unabhängig vom eingesetzten KI-Verfahren sind.

In der Praxis werden oft Kombinationen aus verschiedenen Verfahren maschinellen Lernens und ggf. konventioneller Programmierung angewendet. In diesen Fällen kann das Zustandekommen eines Ergebnisses u. U. zumindest teilweise erklärt werden oder es könnten Zwischenergebnisse der einzelnen Verfahren der betroffenen Person offenbart werden - z. B. wie eine mögliche Mehrheitsentscheidung verschiedener Verfahren ausgefallen ist oder ob Schwellwerte nur knapp (nicht) erreicht wurden.

Zusammenfassend ist - abhängig von der Eingriffsintensität der Entscheidungen eines KI-Systems - zu fordern, dass in angemessenem Maße Maßnahmen ergriffen und ggf. auch Forschungsaktivitäten ausgeweitet werden, um das Zustandekommen von Entscheidungen erklären, nachvollziehen und prüfen zu können.

Der Verantwortliche muss spezifizieren und dokumentieren, auf welcher theoretischen Grundlage und mit welcher Methode

- die relevanten Soll-Werte bestimmt und die aktuell gemessenen Ist-Werte ermittelt werden, inkl. der Protokollierung der Abweichungen,
- „störsignalbehaftete Daten“ identifiziert bzw. generiert werden und diese in Tests verwendet werden,
- auf Änderungen in der Wissensdomäne (Ausweitung des Zwecks, Änderung rechtlicher Regelungen, Änderungen von Kontexten oder technischen Details) in Form von Daten reagiert wird, wie also eine KI-Komponente aktuell und korrekt gehalten wird,

2.2.2 Gewährleistungsziel Datenminimierung

Bereits bei der Auswahl der Input-Daten sollte die Vollständigkeit des Trainingsdatensatzes in der Wissensdomäne geprüft werden. Während der Verarbeitung können zur Verbesserung der Trainingsergebnisse allerdings noch weitere Trainingsdaten gewonnen werden (z. B. durch Ermittlung von Beispielen, welche zu Falsch-Klassifikationen führen).

Für diese zusätzlichen Daten muss dokumentiert werden, aus welchen Gründen diese Daten in den Trainingsdatensatz mit aufgenommen werden, welche zusätzlichen (personenbezogenen) Eigenschaften in diesen Daten enthalten sind und ob sich hier evtl. auch neue Erkenntnisse zur Vollständigkeitsbetrachtung der Wissensdomäne ergeben. Je nach Anwendungszweck muss sichergestellt werden, dass das erlernte Modell nur die minimal notwendigen personenbezogenen Daten zum Training enthält bzw. reproduzieren kann, um ein festgelegtes Qualitätsmaß zu erreichen. So darf beispielsweise ein KI-System, dessen Zweck die Unterscheidung von Menschen zu anderen Objekten ist, nicht zusätzlich Personen aus dem Trainingsdatensatz identifizieren können.

2.2.3 Gewährleistungsziel Nichtverkettung

KI-Systeme sind für Verkettung sehr gut geeignet. So ist es möglich, aus dem gewählten Input weitere Abbildungen zu lernen. Leicht nachvollziehbar sind solche Möglichkeiten bei einem System zum Vorschlagen von Musikstücken basierend auf den bisher gewählten Musikstücken. Es ist denkbar, auf Basis dieser Vorschläge ein System zu trainieren, welches auch Aussagen zur politischen Orientierung liefert. Diese Verkettung kann man technisch nicht immer verhindern. Dennoch müssen Verantwortliche Maßnahmen ergreifen, dass zumindest die im operativen Einsatz befindlichen Systeme keine derart problematischen, vom Ursprungszweck nicht gedeckten, Ergebnisse liefern. Ist dies nicht möglich muss ausgeschlossen werden, dass diese Ergebnisse explizit weiter verarbeitet werden. Hier ist die Zweckbindung entscheidend, für welche Zwecke Systeme trainiert werden dürfen. Jedes Lernsystem muss dem definierten Zweck folgen. Dies muss schriftlich dokumentiert sein.

2.2.4 Gewährleistungsziel Intervenierbarkeit

Basierend auf der Risikoanalyse der Outputdaten, muss der Verantwortliche entscheiden, inwieweit Ergebnisse und Ausgaben des Lernsystems durch den Nutzer in Frage gestellt werden können. Es wird empfohlen, einen solchen Mechanismus zu implementieren, da darüber recht komfortabel Falsch-Positiv-Klassifikationen bzw. Falsch-Negativ-Klassifikationen erfasst werden können, welche dann wieder in den Trainingsdatensatz fließen bzw. zur Verifikation genutzt werden können. Werden unter Einsatz eines KI-Systems Entscheidungen getroffen, welche die Rechte und Freiheiten betroffener Personen in besonderem Maße einschränken, so ist zwingend ein einfach zu nutzender Interventionsmechanismus zu implementieren, welcher einen manuellen Eingriff durch einen menschlichen Kontrolleur ermöglicht. Auf diesen Mechanismus sind Betroffene hinzuweisen, insbesondere um ihre Rechte gem. Art. 22 Abs. 1 DS-GVO durchsetzen zu können.

2.2.5 Gewährleistungsziel Verfügbarkeit

In klassischen IT-Verfahren wird die Verfügbarkeit durch Maßnahmen unterstützt, welche eine maximale Antwortzeit des Systems auch in Teilausfallsituationen bzw. durch Backupmechanismen eine maximale Wiederherstellungszeit garantieren. Ebenso sollte der Schutz vor Angriffen in dieser Kategorie abgesichert werden. All diese Mechanismen aus der Betrachtung klassischer IT-Systeme gelten auch für KI-Systeme.

2.2.6 Gewährleistungsziel Integrität

Ein KI-System kann dann als integer betrachtet werden, wenn in der Verarbeitungsstufe die erlernte Parametrisierung des Systems vor ungewollter Manipulation geschützt wird (betrifft die Kann-Phase) und wenn die schädliche Manipulation des Systems durch Trainingsdaten wirksam verhindert wird. Die Parametrisierung beschreibt im Kern die mathematische Abbildung, welche das Lernsystem erlernt hat. Hierfür können klassische Methoden der IT-Sicherheit zum Schutz vor Manipulation genutzt werden, wie z. B. die digitale Signatur von Trainingsdaten oder von Systemparametern. Ebenso sollte über ein Rechte- und Rollenkonzept festgelegt werden, welcher Systembetreuer welche Rechte am System hat. Systemveränderungen (d. h. Initialtraining, Nachtraining, Schlüsseltausch etc.) müssen protokolliert werden.

Zusätzlich kommt bei KI-Systemen das Kriterium der inhaltlichen Stabilität hinzu. Da das System die mathematische Abbildung erst lernen muss und diese mit menschlichen Begriffen oder Analogien häufig nicht erklärbar ist, kann ein KI-System unvorhergesehene Ergebnisse liefern. Dazu gehören unter anderem Ergebnisse, die eine diskriminierende Wirkung haben. Wie in 2.1.2 beschrieben muss der den erwarteten Einfluss der verwendeten Datenkategorien auf das Verhalten des KI-Systems abschätzen und im Laufe des Trainings evaluieren.

Problematisch ist weiterhin, dass der Eingaberaum eines Lernverfahrens so groß sein kann, dass es nicht möglich ist, alle möglichen Eingaben auf ihre Ausgabe zu testen. Daher sollten bezüglich der Stabilitätsbetrachtung der erlernten Abbildung mindestens beschrieben werden, inwieweit Untersuchungen zu Inputs durchgeführt wurden, welche sich vom Trainingsdatensatz stark unterscheiden, inwieweit die Reaktion des Systems auf Störungen (z.B. „katastrophisches Vergessen“, Manipulation) untersucht wurde und welche Methodik (inkl. der erzielten Ergebnisse) zur Untersuchung der Systemstabilität genutzt wurde.

2.2.7 Gewährleistungsziel Vertraulichkeit

Wenn ein KI-System mit Interaktions- und nutzerspezifischen Rohdaten fortlaufend trainiert wird, die im laufenden Betrieb anfallen, ist sicherzustellen, dass diese Daten möglichst nur auf einer vom Nutzer verwendeten lokalen KI-Komponente bzw. dessen Client verarbeitet und nach dem Training gelöscht werden. Werden die Daten jedoch an einen KI-Server übermittelt, muss die Vertraulichkeit durch eine Ende-zu-Ende-Verschlüsselung gewährleistet werden. Auch serverseitig dürfen die Daten nur durch Befugte weiter verarbeitet werden.

In KI-Systemen können Zwischenergebnisse entstehen, die personenbezogen sind, entweder weil diese eine Identifizierung von Personen ermöglichen oder eine semantische Bedeutung haben können, die u. U. ungewollte, sensible Rückschlüsse auf die Person ermöglichen, deren Daten verarbeitet werden (siehe auch die Ausführungen zur Erklärbarkeit im Abschnitt 2.2.1 - Transparenz). Es ist technisch sicherzustellen, dass diese Zwischenergebnisse nicht langfristig gespeichert werden und nur ein fest definierter Personenkreis zu vorher festgelegten Zwecken Zugriff auf diese Zwischenergebnisse hat. Die Zugriffe müssen protokolliert werden.

Da das Training der Lernalgorithmen sehr aufwendig sein kann, werden häufig schon Cloud-Services für die Zwecke des Trainings, aber auch der Kann-Phase, genutzt. Hierbei muss der Verantwortliche klären, welche Zugriffsmöglichkeiten der Cloudbetreiber auf die Trainingsdaten und auf die Outputs und Zwischenergebnisse hat und wie diese

organisatorisch geregelt sind. Ist die Möglichkeit der Kenntnisnahme der personenbezogenen Daten durch den Cloud-Betreiber zu risikobelastet für die Rechte und Freiheiten der betroffenen Person, so könnte eine Risikobetrachtung (bzw. die Datenschutzfolgenabschätzung) zum Ergebnis haben, dass das Training und auch die Kann-Phase auf Geräten des Verantwortlichen durchgeführt werden müssen.

Wird das Training in einem Cloud-Bereich durchgeführt, müssen die Trainingsdaten, Testdaten und Verifikationsdaten auf verschlüsseltem Weg in diesen Bereich transportiert werden.

Mit fortschreitender Technik ist zu erwarten, dass insbesondere die Kann-Phase vermehrt auf lokalen oder gar mobilen Geräten direkt ausgeführt wird, statt die Verarbeitung in einer Cloud durchzuführen. Wann immer diese Möglichkeit besteht, sollte hiervon Gebrauch gemacht werden, um die Entstehung eingriffsintensiver Datensammlungen z. B. erlernter Gewohnheiten oder Vorlieben einer großen Anzahl von Personen zu vermeiden.

2.3 Einsatz, Rückkopplung und Selbstveränderung des KI-Systems

Die letzten Phasen des Lebenszyklus eines KI-Systems nach dem Training seiner KI-Komponenten umfassen den Einsatz und die Nutzung des KI-Systems, ggf. die Rückkopplung von erzeugten Outputdaten zum weiteren Training der KI-Komponenten und die fortwährende Validierung des KI-Systems. Um die Anforderungen an KI-Systeme für diese Phasen zu systematisieren, wurden folgende Kriterien herangezogen:

- Verhinderung vollständiger automatisierter Entscheidungen,
- Kontrollierbarkeit / Intervenierbarkeit,
- Zweckbindung,
- Transparenz / Nachvollziehbarkeit / Erklärbarkeit,
- Diskriminierungsverbot / Risikoüberwachung /Risikobewertung,
- Datenminimierung / Pseudonymisierung / Anonymisierung,
- Beständigkeit,
- Safety und
- Rückkopplung von Outputdaten.

Die nachfolgenden stichwortartigen Empfehlungen gehen von einem modellhaften KI-System ohne konkretes Anwendungsszenario aus und besitzen demzufolge zum Teil noch einen gewissen Abstraktionsgrad. Ihre Ausprägung in konkreten KI-Systemen ist daher ggf. weiter zu spezifizieren.

2.3.1 Gewährleistungsziel Transparenz („Erklärbarkeit“)

Die Entscheidungssituationen des KI-Systems sind konzeptionell zu kategorisieren nach Vor- oder Teilentscheidungen und finaler/finalen Entscheidung(en).

Eine relevante finale Entscheidung, deren Freigabe/Bestätigung/Ablehnung (siehe unten), Zeitpunkt und die entscheidende Person sowie ggf. die Gründe, sind revisionssicher (automatisiert) zu dokumentieren.

Für das KI-System müssen Testfälle erstellt werden, bei denen ein definierter Input zu einem plausiblen bzw. innerhalb gewisser definierter Grenzen möglichen Output führt, um die grundsätzliche ordnungsgemäße Arbeitsweise des Systems abschätzen zu können. Diese Testszenaren müssen in regelmäßigen Abständen wiederholt und die Ergebnisse sowie etwaige daraus resultierende Anpassungen dokumentiert werden.

Die maßgeblichen Parameter (z. B. Entscheidungsbäume) und Verarbeitungsschritte für das zustande kommen des Outputs sind revisionssicher zu dokumentieren.

Für ein KI-System sind geeignete Parameter zu definieren, die einen Rückschluss auf die Qualität (Güte, Fehlerrate) der Verarbeitung und damit den nachfolgenden Output zulassen. Die Einhaltung der jeweiligen Parameter bzw. Sollvorgaben ist regelmäßig zu evaluieren.

Dokumentiert und laufend untersucht werden muss, wie gesetzlich verbotene, negative Diskriminierungen unterbunden werden.

2.3.2 Gewährleistungsziel Datenminimierung

Wenn im Laufe des Einsatzes eines KI-Systems für den Output erkennbar irrelevante Daten verarbeitet werden oder Daten die zur Erfüllung des festgelegten Zwecks nicht (mehr) erforderlich sind, sollte das KI-System mit entsprechend reduzierten Trainingsdaten erneut trainiert werden.

Wenn der Output eines KI-Systems mehr Daten umfasst als für den vorgegebenen Zweck erforderlich sind, sind letztere für die weitere Verarbeitung zu verwerfen.

Darüber hinaus ist zu prüfen, ob eine Anpassung des Systems dahingehend erfolgen muss, dass es künftig nicht mehr zu diesem Output kommt.

Wenn der Personenbezug eines Outputs für den definierten Zweck nicht erforderlich ist, so ist der Output in geeigneter Weise zu anonymisieren, z. B. durch Datenreduktion/-aggregation.

Wenn die Rückkopplung des Outputs für eine qualitative Verbesserung des KI-Systems im Rahmen seiner Lernfähigkeit verwendet wird, ist ein etwaiger Personenbezug zu entfernen (Anonymisierung/Pseudonymisierung).

2.3.3 Gewährleistungsziel Nichtverkettung

Durch konventionelle Maßnahmen der IT-Sicherheit ist sicherzustellen, dass beim Umgang mit den Verarbeitungsergebnissen (Output) des KI-Systems die Zweckbindung gewährleistet wird und die widerrechtliche Zusammenführung von Output-Daten ausgeschlossen wird.

Outputdaten unterliegen der Zweckbindung und folgen den Regeln für eine rechtmäßige Weiterverarbeitung.

2.3.4 Gewährleistungsziel Intervenierbarkeit

Soweit an eine finale Entscheidung Folgemaßnahmen geknüpft werden, die ein hohes Risiko für Betroffene bergen, bedarf es eines Freigabe-/Bestätigungs- bzw. Ablehnungsmechanismus durch menschliches Zutun (Art. 22 Abs. 1 DS-GVO). Das KI-System muss dazu solange in einem Warte-Status verbleiben („pending“), bis der weitere Fortgang (manuell) angestoßen wird.

Die für die Wahrung der Betroffenenrechte (Auskunft, Berichtigung, Löschung, Einschränkung) nach Art. 15, 16, 17 und 18 DS-GVO erforderlichen Funktionen müssen

vorhanden sein.

Es muss erkennbar sein, dass der Output eines KI-Systems in die unter Art. 9 Abs. 1 DS-GVO genannten Kategorien fällt, insbesondere wenn dies erstmals der Fall ist. Es muss in regelmäßigen Abständen überprüft werden, ob der Output eines KI-Systems - unabhängig von den zugrundeliegenden Daten - zu einer Diskriminierung nach den in Art. 9 Abs. 1 genannten Kategorien führt. In diesem Fall müssen geeignete Gegenmaßnahmen eingeleitet werden.

Bei der Anwendung einer KI-Komponente kann es leicht zu einer indirekt nutzbaren negativ diskriminierenden Auswertung kommen, indem „unverdächtig“ erscheinende Variablen genutzt werden, von denen jedoch bekannt ist, dass sie mit einer diskriminierenden Variablen korrelieren können (z.B. „Geschlecht“/“Kaufverhalten“). Die vom KI-System genutzten Variablen müssen daher auf entsprechende Korrelationen und eine etwaige Diskriminierungswirkung hin untersucht werden.

Wenn ein KI-System z. B. durch selbstlernende Mechanismen oder Veränderungen beim Input zu einem neuen oder deutlich veränderten Output kommt, ist regelmäßig zu überprüfen, ob sich dadurch neue oder geänderte Risiken für den Betroffenen ergeben.

Soweit der Output eines KI-Systems das Risiko birgt, dass die körperliche Unversehrtheit beeinträchtigt sein kann, bedarf es vor einer entsprechenden Maßnahme einer menschlichen Einwirkung (siehe 2.3.1). In diesen Fällen sollten etablierte Verfahren aus dem Safetybereich wie ein zweiter Durchlauf mit einem Alternativsystem erfolgen, um zu prüfen, ob mit gleichen Eingangsparametern der gleiche Output erzeugt wird.

2.3.4 Gewährleistungsziel Verfügbarkeit

Der Zugriff auf Inputdaten, Verarbeitungsvariablen oder Steuerungs-/Entscheidungsparameter muss an ein Rollen- und Berechtigungskonzept gebunden sein.

Es muss die Möglichkeit bestehen, einen Output im Fall des Verlusts wiederherzustellen.

Es muss die Möglichkeit bestehen, einen Ausfall des KI-Systems bei Bedarf zeitnah zu kompensieren.

2.3.5 Gewährleistungsziel Integrität

Die Integrität des Outputs sollte durch geeignete Maßnahmen (Signatur, Prüfsumme) sichergestellt sein.

Der Zugriff auf Inputdaten, Verarbeitungsvariablen oder Steuerungs-/Entscheidungsparameter muss an ein Rollen- und Berechtigungskonzept gebunden sein.

Änderungen in der Wissensdomäne (z.B. Änderungen von Kontexten, rechtlichen Regelungen, technischen Änderungen, Zunahme des Wissens) müssen durch die unter 2.1.6 beschriebenen Verfahren identifiziert werden. Auf die Änderungen muss mit geeigneten Gegenmaßnahmen reagiert werden.

2.1.7 Gewährleistungsziel Vertraulichkeit

Soweit den Outputdaten Vertraulichkeit zukommt, muss der Zugriff darauf an ein Rollen- und Berechtigungskonzept gebunden sein.

Soweit bei den vorstehenden Dokumentations- oder Protokollierungsanforderungen ein Personenbezug besteht, muss der Zugriff auf die entsprechenden Daten an ein Rollen- und Berechtigungskonzept gebunden sein.

Anlage: Übersicht über technische und organisatorische Maßnahmen für KI-Komponenten und KI-Systeme

Maßnahme	**Zeitpunkt im Lebenszyklus**	**Gewährleistungsziel**
Festlegung der Zwecke des KI-Systems und dessen KI-Komponenten	Design	Nichtverkettung
Dokumentation der Zwecke des KI-Systems und dessen KI-Komponenten	Design	Transparenz
Festlegung der Eingabe- und Ausgabeparameter der KI-Komponente	Design	Integrität
Festlegung einer Hypothese zum Zusammenhang zwischen Eingabe- und Ausgabeparametern	Design	Integrität
Festlegung von unerwünschtem Verhalten der KI-Komponente	Design	Integrität
Festlegung der Erwartungen der verschiedenen Beteiligten an das KI-System	Design	Integrität
Beschreibung des KI-Systems in einer (maschinell lesbare) Policy	Design	Transparenz
Festlegung eines geeigneten KI-Verfahrens die KI-Komponente	Design	Integrität
Dokumentation der Auswahl des KI-Verfahrens (Abwägung zwischen Nachvollziehbarkeit und benötigter Mächtigkeit)	Design	Transparenz
Dokumentation der angestrebten Güte des KI-Systems	Design	Transparenz
Festlegung welche Personen an der Entwicklung beteiligt sind und welche Rechte diese haben.	Design	Transparenz
Dokumentation der Nutzung von KI-Komponenten in Entscheidungsprozessen zur Entscheidungsfindung und Entscheidungsvorbereitung	Design	Transparenz
Herkunft der Rohdaten klären	Veredelung	Transparenz
Festlegung des Verfahrens zur Datenveredelung (normalisieren, standardisieren, komplettieren, fehlerbereinigen)	Veredelung	Integrität
Reduktion der Rohdaten möglichst auf wenige, inhaltlich gut verstandene Dimensionen	Veredelung	Datenminimierung

Dokumentation der statistischen Methoden und relevanten Aspekte der Wissensdomäne	Veredelung	Transparenz
Dokumentation wie diskriminierende oder andere ungewünschte Einflussfaktoren aus den Rohdaten entfernt wurden	Veredelung	Transparenz
Dokumentation der Relevanz bzw. der Repräsentativität der Trainingsdaten inklusive des Fehlermodells	Veredelung	Transparenz
Risikobeurteilung des Einsatzes einer mit den veredelten Daten trainierten KI-Komponente	Veredelung	Transparenz
Anonymisierung der Roh- bzw. Trainingsdaten, außer Personenbezug ist erforderlich	Veredelung	Datenminimierung
Theoriegestützte Abschätzung der benötigten Datenmengen, um angestrebte Güte des KI-Systems zu erreichen	Veredelung	Datenminimierung
Dokumentation für welche Verwendungszwecke die Trainingsdaten hergestellt werden	Veredelung	Nichtverkettung
Erkennung und Korrektur von falschen oder negativ diskriminierenden Roh- bzw. Trainingsdaten	Veredelung	Integrität
Herstellung von (synthetischen) Testdaten auf Basis der Hypothese und Erwartungen	Veredelung	Integrität
Herstellung von störsignalbehafteten (synthetischen) Testdaten auf Basis des unerwünschten Verhaltens	Veredelung	Integrität
Verhinderung von unbefugten Manipulationen an Roh- und Trainingsdaten	Veredelung	Integrität
Wahrung der Vertraulichkeit von Roh- und Trainingsdaten	Veredelung	Vertraulichkeit
Wahrung der Verfügbarkeit von Roh- und Trainingsdaten	Veredelung	Verfügbarkeit
Prüfung der Kompatibilität der Zwecke der KI-Komponente mit den Verwendungszwecken der Trainingsdaten	Training	Nichtverkettung
Dokumentation der zum Training verwendeten Roh- bzw. Trainingsdaten	Training	Transparenz
Herkunft der Trainingsdaten klären	Training	Transparenz
Festlegung von Primär-Trainings-, Verifikations- und Testdatensätzen	Training	Integrität
Dokumentation des Verfahrens zur Einteilung der Primär-Trainings-, Verifikations-, Sekundär-Trainings- und Testdatensätze	Training	Transparenz
Regelung zur Verwendung von Trainings-, Verifikations- und Testdatensätzen	Training	Integrität
Festlegung des Verfahrens zur Ermittlung der Güte der KI-Komponente	Training	Integrität
Dokumentation des Verfahrens zur Ermittlung der Güte der KI-Komponente	Training	Transparenz
Gegebenenfalls Prozess zur Erweiterung der Trainingsdaten, um gesteckte Güte zu erreichen	Training	Datenminimierung
Verhinderung von unbefugten Manipulationen an	Training	Integrität

KI-Komponenten		
Test mit (synthetischen) Testdaten auf Basis der Hypothese und Erwartungen	Validierung	Integrität
Test mit störsignalbehafteten(, synthetischen) Daten auf Basis des unerwünschten Verhaltens	Validierung	Integrität
Prüfung der Hypothese und der Erwartungen auf Basis des Testdatensatzes	Validierung	Integrität
Dokumentation der Güte der KI-Komponente inklusive der ermittelten Fehlerrate und Systemstabilität	Validierung	Transparenz
Untersuchung der KI-Komponente auf Erklärbarkeit und Nachvollziehbarkeit	Validierung	Transparenz
Evaluation des ausgewählten KI-Verfahrens bezüglich alternativer, erklärbarerer KI-Verfahren	Validierung	Transparenz
Prüfung auf Neben- bzw. Zwischenergebnisse und Bewertung dieser	Validierung	Nichtverkettung
Möglichkeit des Eingreifens einer Person in den Entscheidungsprozess	Einsatz	Intervenierbarkeit
Entscheidungen, die hohe Risiken für Betroffene bergen, dürfen von KI-Systemen nur vorbereitet werden.	Einsatz	Intervenierbarkeit
Möglichkeit des Stoppens einer KI-Komponente, von der potentiell Risiken für die Rechte und Freiheiten natürlicher Personen ausgehen, oder des Ersetzens der KI-Komponente durch eine Fall-Back-Lösung	Einsatz	Intervenierbarkeit
Auskunftsmöglichkeit für Betroffene zum Zustandekommen von Entscheidungen und Prognosen	Einsatz	Transparenz
Überwachung des Verhaltens der KI-Komponente	Einsatz	Transparenz
Protokollierung von finalen Entscheidungen, deren Freigabe/Bestätigung/Ablehnung, Zeitpunkt und ggf. entscheidende Person	Einsatz	Transparenz
Prüfung und Bewertung der Hypothese und der Erwartungen auf Basis des Verhaltens im Betrieb	Einsatz	Transparenz
Einhaltung der Policy überwachen und sicherstellen	Einsatz	Integrität
Regelmäßige Prüfung der KI-Komponente auf Diskriminierungen und anderes unerwünschtes Verhalten	Einsatz	Integrität
Regelmäßige Evaluierung der Hypothese, Erwartungen, unerwünschtem Verhalten bezüglich der Wissensdomäne und den sonstigen Rahmenbedingungen	Einsatz	Integrität
Regelmäßige Evaluierung welche Eingabe- und Ausgabeparameter der KI-Komponenten für das gewünschte Verhalten des KI-Systems relevant und erforderlich sind und wenn möglich Anpassung der KI-Komponenten zur Verarbeitung nur relevanter und erforderlicher Daten	Einsatz	Datenminimierung

Regelmäßige Prüfung der Güte des KI-Systems und seiner KI-Komponenten auf Basis der Betriebsdaten	Einsatz	Integrität
Berechnungen der KI-Komponente möglichst auf Endgerät des Nutzer ohne Übermittlung der Daten	Einsatz	Vertraulichkeit
Wenn Berechnungen der KI-Komponente nicht auf Endgerät des Nutzer durchgeführt werden kann, dann möglichst auf Geräten unter der Kontrolle des Verantwortlichen	Einsatz	Vertraulichkeit
Wenn Übermittlung der Daten an KI-Komponente erforderlich, dann Ende-zu-Ende-Verschlüsselung einsetzen	Einsatz	Vertraulichkeit
Betroffene über das Verfahren zur Ermittlung der Güte des KI-Systems und die festgestellte Güte informieren	Einsatz	Transparenz
Betroffene und eingreifende Personen möglichst über Teilergebnisse und, ob Schwellwert nur knapp (nicht) erreicht wurde, informieren.	Einsatz	Transparenz
Ausschluss der Nutzung von Neben- bzw. Zwischenergebnissen durch Unbefugte	Einsatz	Vertraulichkeit
Ausschluss der Nutzung von Neben- bzw. Zwischenergebnissen zu nicht vorgesehenen Zwecken	Einsatz	Nichtverkettung
Mechanismen zum Anfechten und Korrigieren von Entscheidungen einer KI-Komponente vorsehen	Einsatz	Intervenierbarkeit
Verwendung von angefochtenen und korrigierten Entscheidungen zur Weiterentwicklung der KI-Komponente	Einsatz	Integrität
Hinweis auf die Möglichkeit zur Anfechtung und Korrektur von Entscheidungen geben	Einsatz	Transparenz
Verhinderung von unbefugten Manipulationen an KI-Komponenten	Einsatz	Integrität
Bei der weiteren Verarbeitung von Daten (bspw. direkte Rückkopplung in die KI-Komponente oder Speicherung für erneute Trainingsphasen), die beim Betrieb der KI-Komponenten anfallen, sollten möglichst anonymisiert werden.	Rückkopplung	Datenminimierung