Konferenz der unabhängigen
Datenschutzaufsichtsbehörden des Bundes und der Länder

Orientierungshilfe zu empfohlenen technischen und organisatorischen Maßnahmen bei der Entwicklung und beim Betrieb von KI-Systemen

Version 1.0

Stand:
Juni 2025

Inhalt

1 Datenschutzrechtliche Anforderungen an KI-Systeme

Im Bereich der Künstlichen Intelligenz (KI) werden häufig große Datenmengen, auch personenbezogener Daten, verarbeitet. Wegen des Umfangs der Verarbeitung personenbezogener Daten, aber auch des potenziell damit einhergehenden hohen Risikos für Betroffene, hat der Datenschutz für KI-Systeme eine besondere Relevanz. Daher sollte der Datenschutz nach dem Prinzip „data protection by design“ von Anfang an bei der Entwicklung und Anwendung von KI-Systemen[1] mitgedacht werden.

Dieses Papier richtet sich vorrangig an Hersteller:innen und Entwickler:innen von KI-Systemen und soll diesen als Hilfestellung bei der datenschutzkonformen Entwicklung von datenschutzkonform einsetzbaren KI-Systemen dienen. Verantwortliche, die KI-Systeme einsetzen möchten, können die von der DSK veröffentlichte Orientierungshilfe „Künstliche Intelligenz und Datenschutz“[2] konsultieren und das hier vorliegende Positionspapier zurate ziehen um die technischen Entwicklungsmöglichkeiten im Beschaffungsprozess zu berücksichtigen.

1.1 Einordnung

In diesem Dokument wird unter dem Begriff KI-System ein maschinengestütztes System verstanden, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist, 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 (siehe Art. 3 Nr. 1 der KI-Verordnung). Ein KI-System beruht auf einem oder mehreren KI-Modellen.

Bei der Betrachtung von KI-Systemen in diesem Papier wird die gesamte Bandbreite von solchen mit einem klar definierten Verwendungszweck (z. B. zur Mustererkennung in medizinischen Anwendungen) bis hin zu KI-Systemen mit einem allgemeinen Verwendungszweck abgedeckt. Nicht betrachtet werden solche KI-Systeme, die sicher keinen Personenbezug haben – d. h. weder in den Trainingsdaten noch in der Anwendung (z. B. Vorhersage-Systeme für Naturereignisse).

Aus datenschutzrechtlicher Perspektive liegt der Fokus darauf, wie die Rechte und Freiheiten von natürlichen Personen geschützt werden können. Ein wichtiger von klassischen IT-Systemen abweichender Aspekt bei KI-Systemen ist das Training der verwendeten KI-Modelle und die Nutzung von Trainingsdaten. Für viele KI-Modelle werden sehr große Datenmengen zum Training benötigt. Die datenschutzrechtlichen Fragen, die sich aus der Sammlung dieser Datensätze (z. B. durch Crawling, Scraping) ergeben, sind nicht Gegenstand dieses Papiers.

Die Einbindung von bestehenden KI-Modellen oder -Systemen in neue KI-Systeme – sei es durch die Nutzung und Spezialisierung von KI-Modellen von Dritten oder die Einbindung über

[1] Die in diesem Dokument verwendeten Begriffe orientieren sich an den Definitionen des Art. 3 der KI-Verordnung (KI-VO).

[2] https://www.datenschutzkonferenz-online.de/media/oh/20240506_DSK_Orientierungshilfe_KI_und_Datenschutz.pdf

eine Schnittstelle – wird ebenfalls nicht in diesem Papier erörtert. Viele der in diesem Papier beschriebenen Aspekte sind auch hierfür relevant, allerdings ergeben sich weitere Fragestellungen, wenn das KI-Modell oder -System und die neue Anwendung nicht aus einer Hand stammen.

Für Anwendungen oder Dienste, die ein oder mehrere KI-Systeme oder -Modelle als Komponente eines größeren Systems enthalten, können die in diesem Papier beschriebenen Aspekte ebenfalls hilfreich sein. Für eine Gesamtbetrachtung werden sie aber ggf. nicht ausreichen.

Um den Rahmen des Dokumentes nicht zu sprengen, kann nicht auf einzelne KI-Algorithmen eingegangen werden. Gleiches gilt für konkrete Anwendungsgebiete, Einsatzszenarien sowie Architekturaspekte. Bestimmte Arten von KI-Algorithmen werden in diesem Dokument nur angesprochen, wenn dies für eine differenzierte datenschutzrechtliche Betrachtung sinnvoll ist.

1.2 Abgrenzung der Lebenszyklusphasen

Ein KI-System durchläuft während der Entwicklung und des anschließenden Betriebs verschiedene Phasen, welche dessen Lebenszyklus darstellen können. Für das vorliegende Dokument werden dabei die vier Phasen (1) Design, (2) Entwicklung, (3) Einführung und (4) Betrieb und Monitoring unterschieden, die im Folgenden kurz beschrieben und zueinander abgegrenzt werden. Darüber hinaus wird einführend kurz auf die Bedeutung der Phasen für die Verarbeitung personenbezogener Daten eingegangen. Es ist anzumerken, dass diese Phasen für verschiedene KI-Systeme unterschiedliche Relevanz haben, und auch der Übergang zwischen den Phasen ist teilweise fließend.

1. Design (u. a. Auswahl der Daten, Datensammlung)

Die Designphase umfasst vor allem die Planung und vorbereitende Schritte für die Umsetzung eines neuen KI-Systems. Ziel ist es, alle notwendigen Informationen für die darauffolgende Entwicklung bzw. technische Umsetzung zu sammeln und strukturiert zu dokumentieren. In einem ersten Schritt werden dabei die Anforderungen erhoben, die das System erfüllen soll. Dem Prinzip „data protection by design“ folgend, sollten neben den funktionalen Anforderungen in diesem Schritt auch bereits datenschutzrechtliche Fragestellungen in Erwägung gezogen werden. Dies umfasst auch die frühzeitige Planung von technischen und organisatorischen Maßnahmen, um den Schutz der personenbezogenen Daten in allen Lebenszyklusphasen zu gewährleisten.

Im Vergleich zu klassischen IT-Systemen wird für das in der Entwicklungsphase angesiedelte Training üblicherweise eine große Datenmenge – die sogenannten Trainingsdaten, aber auch Test- und Validierungsdaten – benötigt, deren Beschaffung in Form von Rohdaten ebenfalls Teil der Designphase ist. Die Daten werden entweder in eigener Zuständigkeit erhoben oder aus bereits vorhandenen Datenquellen – bspw. öffentlich zugängliche Datenarchive – bezogen.

2. Entwicklung (Datenaufbereitung, Training und Validierung)

Nachdem die Anforderungen an das KI-System definiert wurden und die erforderlichen Rohdaten für die Trainings-, sowie Test- und Validierungsdaten vorliegen, wird nun mit der konkreten technischen Umsetzung begonnen. Dies umfasst zum einen die Implementierung der verwendeten KI-Algorithmen, als auch das anschließende Training der KI-Modelle mit den Trainingsdaten.

Um die gesammelten Daten für das Training verwenden zu können, müssen diese ggf. noch aufbereitet werden. Dabei handelt es sich um den Prozess, der die Rohdaten zweckbestimmt und domänenspezifisch aufbereitet, sodass diese während des Trainings von den KI-Algorithmen verarbeitet werden können, um die Anforderungen an die KI-Modelle zu erfüllen. Sofern die Rohdaten einen Personenbezug aufweisen, dieser aber für das zu entwickelnde KI-Modell nicht erforderlich ist, sollte dieser spätestens zum jetzigen Zeitpunkt durch die Anwendung geeigneter Maßnahmen entfernt werden.

Die durch das Training erzeugten KI-Modelle müssen in dieser Phase ebenfalls validiert werden, was bedeutet, dass die Qualität der Ausgaben dahingehend geprüft werden muss, dass diese den erwarteten Qualitätsmaßen entsprechen. Qualitätsmaße können z. B. Fehlertoleranzen bei Klassifikationsaufgaben sein oder die Unterdrückung ungewünschter Ausgaben in generativen KI-Modellen. Zu diesem Zweck sollten ergänzend zu der Prüfung der Güte des KI-Modells anhand der Validierungsdaten Testverfahren entwickelt und angewendet werden, die auch die vorgesehene Integration des KI-Modells berücksichtigen.

3. Einführung (Softwareverteilung inkl. Updates)

Wenn die Optimierung des KI-Systems abgeschlossen ist, kann dieses in der Produktivumgebung installiert und für die Anwender:innen zugänglich gemacht werden.

Spezifisch für den Anwendungsfall müssen an dieser Stelle die notwendigen Konfigurationen vorgenommen werden, welche datenschutzfreundliche Voreinstellungen berücksichtigen sollten („data protection by default“).

4. Betrieb und Monitoring

Nachdem ggf. Pilotprojekte erfolgreich durchgeführt und die notwendigen Anpassungen für einen Wirkbetrieb abgeschlossen wurden, kann das KI-System für die Nutzung freigegeben werden und in den Produktivbetrieb übergehen. Anwender:innen können nun auf das System zugreifen und damit Daten für die definierten Zwecke verarbeiten.

Obwohl bereits ausführliche Tests durchgeführt wurden, sollte in regelmäßigen Intervallen eine erneute Prüfung durchgeführt werden, um die Qualität der Ausgaben zu evaluieren.

Ein KI-System kann auch so ausgestaltet sein, dass es selbstständig im produktiven Betrieb lernt und sein Verhalten entsprechend anpasst. Zu diesem Zweck werden die Ein- und Ausgaben verarbeitet, sowie auch mögliche Rückmeldungen zur Qualität der Ausgaben durch die Anwender:innen berücksichtigt. Diese Daten können dabei entweder in eine automatisierte Aktualisierung der KI-Modelle einfließen oder für ein zeitverzögertes, manuell angestoßenes

Neutraining verwendet werden. Dies kann dazu führen, dass die Lebenszyklusphasen erneut durchgeführt werden müssen.

1.3 Rechtliche Grundsätze

Für die Verarbeitung personenbezogener Daten (Definition: vgl. Art. 4 Nr. 2 DSGVO) gilt die DSGVO einschließlich der in ihr verankerten Grundsätze (vgl. Art. 5 Abs. 1 DSGVO), u. a. der Rechtmäßigkeit, der Verarbeitung nach Treu und Glauben, der Transparenz, der Zweckbindung, der Datenminimierung und der Richtigkeit. Nach Art. 5 Abs. 2 DSGVO ist der Verantwortliche für die Einhaltung der Grundsätze nach Abs. 1 verantwortlich und muss deren Einhaltung nachweisen können.

Der Zweck einer Verarbeitungstätigkeit muss vorher festgelegt, eindeutig und legitim sein, die Verarbeitung muss eine Rechtsgrundlage aufweisen und die bei der Verarbeitung entstehenden Risiken müssen ausreichend eingedämmt werden. Die Entwicklung und der Betrieb von KI-Systemen können mit hohen Risiken für die Rechte und Freiheiten der Betroffenen einhergehen, z. B. aufgrund der Art, des Umfangs und der Zwecke der Verarbeitung. In solchen Fällen müssen geeignete technische und organisatorische Maßnahmen nach Art. 32 DSGVO getroffen werden, um ein dem Risiko angemessenes Schutzniveau zu gewährleisten.

Hat eine Form der Verarbeitung voraussichtlich ein hohes Risiko für die Rechte und Freiheiten natürlicher Personen zur Folge, muss eine Datenschutz-Folgenabschätzung gem. Art. 35 DSGVO durchgeführt werden.[3]

Darüber hinaus müssen Verantwortliche ihren Informationspflichten nach Art. 13, 14 DSGVO nachkommen und gewährleisten, dass betroffene Personen ihre Rechte ausüben können, insbesondere das Recht auf Auskunft nach Art. 15 DSGVO, das Recht auf Berichtigung nach Art. 16 DSGVO, das Recht auf Löschung nach Art. 17 DSGVO und das Recht nicht einer automatisierten Entscheidung im Einzelfall unterworfen zu werden nach Art. 22 DSGVO.

Hersteller:innen und Entwickler:innen sind in der Regel für die Verarbeitungen personenbezogener Daten in den Phasen Design und Entwicklung Verantwortliche im Sinne des Art. 4 Nr. 7 DSGVO. In den Phasen Einführung und Betrieb bestimmt die Stelle, die das KI-System einsetzt, welche personenbezogenen Daten zu welchen Zwecken mittels eines KI-Systems verarbeitet werden. Die in diesen Phasen Verantwortlichen können nur KI-Systeme bzw. KI-Modelle einsetzen, die von den Hersteller:innen und Entwickler:innen so gestaltet wurden, dass sie datenschutzkonform eingesetzt werden können.[4]

[3] https://www.datenschutzkonferenz-online.de/media/kp/dsk_kpnr_5.pdf

[4] Siehe auch Rz. 130 der Stellungnahme 28/2024 des EDSA vom 17. Dezember 2024: Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_de

1.4 SDM

In der Praxis müssen die rechtlichen Anforderungen aus der DSGVO durch technische und organisatorische Maßnahmen erfüllt werden. Dabei unterstützt das Standard-Datenschutzmodell (SDM)[5], eine Methode zur Datenschutzberatung und -prüfung auf Basis einheitlicher Gewährleistungsziele. Die Systematik und Anwendung des SDM lehnt sich an die Vorgehensweise von Sicherheitsstandards nach IT-Grundschutz des Bundesamtes für Sicherheit in der Informationstechnik an.

Das SDM referenziert auf die DSGVO und bietet geeignete Mechanismen, um rechtliche Anforderungen in technische und organisatorische Maßnahmen zu überführen. Zu diesem Zweck erfasst das SDM die rechtlichen Anforderungen aus den Grundsätzen im Art. 5 sowie Art. 32 DSGVO und ordnet sie anschließend den folgenden Gewährleistungszielen zu:

- **Datenminimierung**: Die verarbeiteten personenbezogenen Daten müssen dem Zweck angemessen und erheblich sowie auf das für die Zwecke der Verarbeitung notwendige Maß beschränkt sein.
- **Verfügbarkeit**: Der Zugriff auf die personenbezogenen Daten muss gesichert sein, und Systeme können dann ihre Verarbeitungen durchführen, wenn diese angefordert werden.
- **Vertraulichkeit**: Unbefugte dürfen nicht auf die personenbezogenen Daten zugreifen können.
- **Integrität**: Die personenbezogenen Daten sind und bleiben unversehrt, vollständig, zurechenbar und aktuell.
- **Intervenierbarkeit**: Verantwortliche müssen jederzeit in der Lage sein, in die Datenverarbeitung vom Erheben bis zum Löschen von Daten einzugreifen, um neben der Umsetzung von Betroffenenrechten, wie die Rechte auf Berichtigung oder Löschung, auch behördliche Anordnungen umsetzen und Datenschutzverletzungen beheben bzw. abmildern zu können.
- **Transparenz**: Einerseits ist Transparenz zur Gewährleistung der Rechenschaftspflicht nach Art. 5 Abs. 2 DSGVO sowie andererseits zur Erfüllung der Informationspflichten nach Art. 12, 13 sowie 14 DSGVO erforderlich.[6]
- **Nichtverkettung**: Die personenbezogenen Daten werden nicht für einen anderen als den festgelegten, eindeutigen und legitimen Zweck erhoben, verarbeitet und genutzt.

Es ist zu beachten, dass die oben genannten Gewährleistungsziele teilweise konträre Anforderungen erfüllen sollen. Daher muss bei einer ganzheitlichen Betrachtung auf ein ausgewogenes Verhältnis der in Abhängigkeit stehenden Gewährleistungsziele geachtet werden.

[5] https://www.datenschutzkonferenz-online.de/media/ah/SDM-Methode-V31.pdf

[6] Die Art der zu übermittelnden Informationen hängt von der Zielgruppe ab (Betroffene, Verantwortliche, Aufsichtsbehörden). Deshalb sollte eine Abstufung in unterschiedliche Level von Transparenz erfolgen.

Die Gewährleistungsziele dienen vor allem der Bündelung und Strukturierung abstrakter rechtlicher Anforderungen, um diese in erforderliche konkrete technische und organisatorische Maßnahmen zur Gewährleistung eines angemessenen Schutzniveaus transferieren zu können. Ein zum SDM gehörender Katalog risikomindernder Referenzmaßnahmen hilft bei diesem Schritt.

Im Folgenden soll dargestellt werden, welche technischen und organisatorischen Maßnahmen dazu beitragen können, um den mit Hilfe der Gewährleistungsziele systematisierten Anforderungen des Datenschutzrechts gerecht zu werden. Viele der bei klassischen IT-Systemen etablierten technischen und organisatorischen Maßnahmen, die in dem Maßnahmenkatalog des SDM[7] und in den Empfehlungen zum IT-Grundschutz des BSI[8] zu finden sind, können und sollten bei der Entwicklung und dem Betrieb von KI-Systemen ebenfalls umgesetzt werden.

Einige dieser Maßnahmen sind bei der Verarbeitung personenbezogener Daten in großem Umfang, wie sie bei der Entwicklung von KI-Systemen häufig vorkommt, von besonderer Relevanz. Dazu zählen etwa Verschlüsselungsmechanismen und klassische Methoden der IT-Sicherheit zum Schutz vor Manipulation, wie z. B. digitale Signaturen, Prüfsummen und Protokollierung. Wichtig sind auch Rollen- und Berechtigungskonzepte, sowie Löschkonzepte mit Aufbewahrungs- und Löschfristen. Regelmäßige Schulungen und Weiterbildung sind bei den schnellen Entwicklungen im Bereich der KI ebenfalls dringend erforderlich.

Alle folgenden Betrachtungen konzentrieren sich auf KI-spezifische Anforderungen und Maßnahmen.

2 Technische und organisatorische Anforderungen an KI-Systeme

Orientiert an den in Abschnitt 1.2 eingeführten Lebenszyklusphasen 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 auch einzelne Aspekte einer anderen Phase zugeordnet werden können.

2.1 Design (u. a. Auswahl der Daten, Datensammlung)

Während der Designphase werden vor allem Entscheidungen bezüglich der Ausgestaltung und Funktionsweise des künftigen KI-Systems getroffen. Neben der Auswahl einer geeigneten Architektur, der technischen Infrastruktur, auf welcher das System betrieben werden soll, sowie der Art des genutzten KI-Modells sollte ein besonderes Augenmerk auf die Erhebung oder Auswahl von Daten gelegt werden, die zum Trainieren des KI-Modells sowie zur Validierung und Testung des Gesamtsystems gebraucht werden. Abhängig vom Zweck und der Art des ausgewählten KI-Systems können hierbei personenbezogene, synthetische oder anonyme Daten genutzt werden.

[7] https://www.datenschutz-mv.de/datenschutz/datenschutzmodell/

[8] https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/IT-Grundschutz/it-grundschutz_node.html

Für die Verarbeitung personenbezogener Daten ist eine Rechtsgrundlage erforderlich. Bei öffentlich zugänglichen Daten muss der Verantwortliche, der den Datensatz zum Trainieren seiner KI-Modelle verwenden möchte, neben einer geeigneten Rechtsgrundlage für das Training selbst, sicherstellen, dass die Erstellung des Datensatzes nicht offensichtlich rechtswidrig war (z. B. durch eine Datenpanne). Dafür ist unter anderem darauf zu achten, dass die Quelle der Daten bei der Beschreibung des Datensatzes angegeben ist, der Datensatz nicht durch eine Straftat erstellt wurde (z. B. Leak aus dem Darknet) und nicht Gegenstand eines öffentlichen Verfahrens war, das die Erstellung des Datensatzes als rechtswidrig einstuft, und kein Zweifel an der Rechtmäßigkeit des Datensatzes besteht (wenn z. B. der Datensatz sensitive Informationen wie präzise Geodaten vom Wohn- und Arbeitsort von Tausenden Menschen beinhaltet, die als Klardaten und ohne Angabe einer Rechtsgrundlage zur Verfügung gestellt werden).[9] Bei der Auswahl und Erhebung personenbezogener Daten müssen außerdem geeignete technische und organisatorische Maßnahmen getroffen werden, die ein angemessenes Schutzniveau gewährleisten. Im Folgenden werden mögliche Maßnahmen bezüglich der einzelnen Gewährleistungsziele erörtert.

2.1.1 Gewährleistungsziel Transparenz

Um Prüfbarkeit herzustellen, sollte folgendes durch den Verantwortlichen dokumentiert werden:

1. Der Zweck der Verarbeitung und die Rechtsgrundlage für die Erhebung oder die Weiternutzung der Rohdaten ist anzugeben.
2. Ist für das Training die Verarbeitung personenbezogener Daten erforderlich? Lässt sich das Ziel und der Zweck des KI-Systems evtl. auch mit synthetischen oder anonymisierten Daten erreichen?
3. Die Methodik, welche in „Datasheets for datasets“[10] beschrieben wird, ist eine standardisierte Möglichkeit Datensätze zu dokumentieren. Sofern die Datensätze von externen Stellen bezogen werden, kann diese Dokumentation bereits durch diese Institutionen erstellt werden. Beim Bezug von Trainingsdaten kann dann bereits auf die vorliegenden Informationen zurückgegriffen werden. Es ist zu beachten, dass die Dokumentation nach Änderung des Datensatzes aktualisiert werden muss, auch während der darauffolgenden Phasen. Die Dokumentation sollte vor allem die folgenden Aspekte berücksichtigen:
    a. Woraus besteht der Datensatz? (Datentyp (Bilder, Videos, tabellarische Daten usw.), Datenkategorien (Attribute, Klassen usw.), Metadaten usw.)

[9] Siehe hierzu Abschnitt 3.4 der Stellungnahme 28/2024 des EDSA vom 17. Dezember 2024: Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_de

[10] Gebru, Timnit, et al.: „Datasheets for datasets“, Communications of the ACM 64.12 (2021), 86-92, https://arxiv.org/abs/1803.09010

b. Herkunft/Quellen des Datensatzes (bei bereits verfügbaren Datensätzen: Version und Datum des Datensatzes und Link zum Datensatz angeben; bei selbst erhobenen Datensätzen: Methodik der Erhebung detailliert beschreiben, z. B. Skript zur Datenerhebung aufbewahren und auf Anfrage zur Verfügung stellen)
c. Kontext der Erhebung (für jede Quelle ist die Methode der Erhebung, die Betroffenenkreise (z. B. Kundendaten von Firma xy) und der Zeitraum der Datensammlung anzugeben)
d. (Ursprünglicher) Zweck der Datenerhebung
e. Version der Dokumentation und letzte Aktualisierung

4. Zum geplanten KI-System:
   a. Zielsetzung: Ziel und Funktion des KI-Systems
   b. System-Architektur (Woraus soll das KI-System bestehen: vortrainiertes KI-Modell? Federated Learning Architektur? RAG? API?)
   c. Auswahl der KI-Algorithmen, die in Frage kommen und die bei der Entwicklungsphase trainiert werden sollen. Vor der Entwicklungsphase ist häufig nicht klar, welches KI-Modell später zum Einsatz kommen wird, da dies von der Performance Evaluation abhängt, deren Ergebnisse erst nach der Entwicklungsphase feststehen. Es ist trotzdem eine gute Praxis, zumindest die Gruppe von KI-Algorithmen zu definieren, die dem Zweck und dem Ziel der Verarbeitung entsprechen.
   d. Ausgeführte Experimente, die die Designauswahl begründen.
5. Maßnahmen zur Erreichung der anderen Gewährleistungsziele, um Prüfbarkeit in Bezug auf ebendiese herzustellen (z. B. Maßnahmen zur Datenminimierung oder Integrität).

Außer den oben genannten Dokumentationen, ist die Frage nach der Transparenz eines KI-Systems bzw. KI-Modells zu klären. KI-Systeme, die auf Open-Source-Ansätzen basieren, können mehr Transparenz gewährleisten. Erklärbare KI-Methoden können nachvollziehbarer machen, wie KI-Systeme zu Ergebnissen kommen und können daher sowohl für die Betroffenen als auch für die Hersteller:innen und Entwickler:innen hilfreich sein. Des Weiteren können Visualisierungen in der Dokumentation helfen, die Funktionsweise eines KI-Modells verständlich zu machen.

2.1.2 Gewährleistungsziel Datenminimierung

Der Verantwortliche muss bereits vor der Datenerhebung bzw. der Weiternutzung bereits erhobener Daten festlegen, was der Zweck des KI-Systems ist und welche Daten dafür benötigt werden könnten. Auf Grundlage dieser Überlegungen sollten genaue Kriterien für die Datenerhebung definiert werden. Dabei ist darauf zu achten, dass nicht notwendige Daten gar nicht erst erhoben werden. Um die notwendigen personenbezogenen Daten zu identifizieren, sollten die folgenden Aspekte berücksichtigt werden:

System-Design: Nachdem man die Zielsetzung des KI-Systems definiert hat, kommen oft verschiedene KI-Algorithmen, welche unterschiedliche Mengen an Trainingsdaten benötigen, in

Frage. Wenn ein KI-System dieselbe Funktion erfüllt und eine ähnliche Performance mit weniger personenbezogenen Daten erreichen kann, sollte es grundsätzlich bevorzugt werden. Auch wenn der Verantwortliche einige Fragen bzgl. der Performance des zu entwickelnden KI-Systems erst in der Entwicklungsphase genau beantworten kann, nachdem er genügend Experimente durchgeführt hat, ist es trotzdem eine gute Praxis sich im Vorfeld zu überlegen, welche KI-Algorithmen in Frage kommen. Hier können Untersuchungen der aktuellen Erkenntnisse aus der wissenschaftlichen Literatur, der Praktiken der Open-Source Community und auszuführende Pilotstudien helfen, eine reflektierte Wahl für das KI-System zu treffen. Ein Beispiel für eine solche technische Maßnahme kann Federated Learning sein, welches Training eines globalen KI-Modells auf mehreren Datenquellen ermöglicht, wobei die lokalen Daten nicht zusammengeführt oder zwischen den beteiligten Institutionen ausgetauscht werden.

Volumen: Die Anzahl der Datenpunkte muss in Bezug auf die Zielsetzung des KI-Systems und die Wahl des verwendeten KI-Algorithmus gerechtfertigt werden. Dabei spielen insbesondere die Qualität der Daten und die Repräsentativität des Datensatzes eine Rolle. Darüber hinaus kann eine unausgewogene Minimierung von Daten die Integrität der Modellierung von KI-Modellen gefährden, d.h. zu Bias führen. Historische Tiefe und die Vertretung verschiedener Personengruppen in dem Datensatz sind Faktoren, die für Richtigkeit und als Maßnahme gegen Diskriminierung zu berücksichtigen sind. Die Durchführung von Pilotstudien mit kleineren Datensätzen, die schrittweise Vergrößerung des Datensatzes – falls erforderlich – und das Löschen nicht mehr benötigter Daten ist eine gute Praxis.

Kategorien: Auf welche Kategorien von Daten soll sich die Entscheidung des KI-Modells stützen und wovon soll die Entscheidung potentiell abhängig gemacht werden (z. B. Alter, Geschlecht, Gesichtsbilder, Aktivität in einem sozialen Netzwerk)? Die Nutzung von besonderen Kategorien personenbezogener Daten (sogenannte „Art. 9 Daten“) muss geprüft und begründet werden. Attribute mit generalisiertem Charakter sollten bevorzugt werden (z. B. nur das Jahr statt das Bündel Tag-Monat-Jahr für das Geburtsdatum). Verschiedene Feature Selection- und Dimensionsreduzierungs-Techniken können dabei helfen, die Dimension der Attribute zu minimieren.[11] Außerdem sollten Attribute entfernt werden, die zu Bias und Diskriminierung führen können, wenn sie für die Verarbeitung nicht zwingend erforderlich sind. Hier sollte der Verantwortliche darauf achten, dass auch sogenannte Proxy Features indirekt auf sensible Attribute verweisen können (z. B. Postleitzahl als Proxy für die Herkunft). 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.[12]

[11] Mittels Feature Selection Techniken werden stark korrelierende Attribute identifiziert und nur die übrig gebliebene Teilmenge der Features verwendet. Dimensionsreduzierungs-Techniken wie PCA projizieren jeden Datenpunkt auf einen Punkt mit weniger Merkmalen, während wichtige Informationen erhalten bleiben.

[12] 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-Modelle mit Kategorien von Daten trainiert werden,

Typologie: Sind für das Training personenbezogene Daten notwendig, oder können synthetische oder anonymisierte Daten verwendet werden? Hier muss auf die Re-Identifizierbarkeit der Daten geachtet werden. Die Nutzung augmentierter Daten kann in Frage kommen, um Diversität in einem Datensatz zu erlangen, ohne dass neue Datenpunkte erforderlich werden.

Quelle: Können die Daten aus bereits vorhandenen Quellen bezogen werden oder müssen diese für den geplanten Zweck neu erhoben werden?

2.1.3 Gewährleistungsziel Nichtverkettung

Wenn gesetzliche Regelungen die Verwendung bestimmter Daten nicht zulassen (z. B. bei besonders sensiblen Daten nach Art. 9 DSGVO) und stattdessen hochkorrelierende Ersatzvariablen verwendet werden sollen, stellt dies eine Verkettung dar. Es ist zu beachten, dass bei einem Verarbeitungsverbot des direkten Datums auch eine Herleitung des Datums aus scheinbar unkritischen personenbezogenen Daten unzulässig ist – selbst, wenn dies nur als Zwischenergebnis für den eigentlichen Verarbeitungszweck genutzt werden soll.

2.1.4 Gewährleistungsziel Intervenierbarkeit

Schon in der Designphase müssen Maßnahmen zur Umsetzung der Betroffenenrechte und behördlicher Anordnungen getroffen werden, sowohl in Bezug auf die Trainingsdaten als auch ggf. bezüglich der KI-Modelle. Dafür muss ein interner Prozess definiert werden.

Intervenierbarkeit an den Rohdaten: Wenn die betroffenen Personen über die Erhebung der Rohdaten nach Art. 13 oder 14 DSGVO informiert werden müssen, muss ein angemessener Zeitraum zwischen dem Zeitpunkt der Erhebung der Rohdaten, der Information der betroffenen Personen und dem Zeitpunkt des Trainings des KI-Modells definiert werden, um den betroffenen Personen die Möglichkeit zu geben, ihre Rechte im Vorfeld des Trainings auszuüben. Es dürfen keine Daten von Webseiten gesammelt werden, die sich eindeutig gegen die Wiederverwendung ihrer Inhalte für den Zweck zum Aufbau eines Trainingsdatensatzes aussprechen (z. B. durch robots.txt oder ai.txt).

Die Gewährleistung der Ausübung von Betroffenenrechten entbindet den Verantwortlichen nicht davon, Methoden mit möglichst geringer Eingriffstiefe für die Verarbeitung zu wählen.

Maßnahmen zur Auffindbarkeit von personenbezogenen Daten im Rohdatensatz sollten durch Datenmanagement-Systeme und Suchfunktionen, z. B. durch Big Data Datenbanken, implementiert werden.

Intervenierbarkeit an dem KI-Modell: KI-Modelle, die den Betroffenen besseren Zugang zur Ausübung ihrer Rechte ermöglichen, sind zu bevorzugen. Z. B. ist ein KI-Modell, das im Falle einer Löschanfrage personenbezogener Daten ohne die zu löschenden Daten schneller neu trainiert werden kann, zu bevorzugen, falls es eine ausreichende Performance erreicht. In einigen Fällen kann es sogar einfach möglich sein, personenbezogene Daten in dem KI-Modell

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 beispielsweise dem Geschlecht, das KI-System diskriminierende oder fehlerhafte Ergebnisse liefert.

direkt zu identifizieren und diese nach Anfrage zu entfernen (z. B. bei Support-Vector-Maschinen (SVMs)). Techniken wie Machine Unlearning oder Ansätze wie das Fine-Tuning von KI-Modellen ohne die zu löschenden Daten sollten daraufhin geprüft werden, ob sie im jeweiligen Fall effektiv sind.

2.1.5 Gewährleistungsziel Verfügbarkeit

Um den Zugriff auf personenbezogene Daten und ihre Verarbeitung unverzüglich möglich zu machen, integriert der Verantwortliche Datenmanagement-Systeme, z. B. Big Data Datenbanken.

2.1.6 Gewährleistungsziel Integrität

Der Verantwortliche prüft die Richtigkeit der Rohdaten, indem er insbesondere auf die Qualität der Daten und ihrer Annotationen, Vertrauenswürdigkeit der Quelle der Daten und vorhandenen Bias in dem Datensatz achtet. Um eine mögliche Verfälschung und Veränderung der Rohdaten zu erkennen, können Hash-Werte über den gesamten Rohdatensatz oder einzelne Datenpunkte generiert und gespeichert werden.

Es ist eine gute Praxis, die Rohdaten statistisch zu untersuchen, sowie die einzelnen Datenpunkte individuell zu betrachten. Die sich daraus ergebende Erkenntnisse sind vor allem für die Implementierung der Aufbereitung in der Entwicklungsphase wichtig. Hier sind Data Validation-Methoden von großer Bedeutung, was auch – mit Untersuchung, ob die Quelle der Daten vertrauenswürdig ist – gegen Data Poisoning[13] helfen kann.

Genaue Untersuchungen der Verteilung der Daten und Pilotstudien, wo z. B. die Performance eines KI-Modells über verschiedene Personengruppen hinweg untersucht werden, ist ein erster Schritt, um Diskriminierungen durch KI-Systeme vorzubeugen. Maßnahmen für eine repräsentative Verteilung der Daten und weitere Präventionsmaßnahmen und Abmilderungstechniken sollten umgesetzt werden.

Falls vortrainierte KI-Modelle benutzt werden sollten, ist auf Backdoor Poisoning[14] zu achten, was zu ungewollten Ausgaben des KI-Modells führen kann. Hierzu prüft der Verantwortliche die Vertrauenswürdigkeit der vortrainierten KI-Modelle, bevor diese weitertrainiert und eingesetzt werden.

[13] Goldblum, Micah, et al.: „Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses“, IEEE Transactions on Pattern Analysis and Machine Intelligence 45.2 (2022), 1563-1580

[14] Backdoor Poisoning Attacken können zum Beispiel durch Manipulierung der Modellgewichte in dem vortrainierten KI-Modell unter Umständen dazu führen, dass es zu ungewollten Voraussagen des KI-Modells kommt, siehe dazu https://arxiv.org/abs/2004.06660.

2.1.7 Gewährleistungsziel Vertraulichkeit

KI-Modelle und -Systeme können personenbezogene Daten preisgeben, die zum Training benutzt worden sind.[15] So kann es einem Angreifer gelingen, personenbezogene Daten aus einem KI-Modell oder -System zu extrahieren. Ein KI-System kann aber auch unbeabsichtigt personenbezogene Daten an Anwender:innen ausgeben. Einige der Gründe dafür sind strukturell und haben mit der Art und Weise zu tun, wie KI-Modelle konstruiert sind, während andere auf Faktoren wie schlechte Generalisierung oder Memorisierung des KI-Modells zurückzuführen sind.[16] Einige generelle Gegenmaßnahmen sind Privacy-Preserving-Techniken wie Differential Privacy, sowie Regularisationstechniken, um die Generalisierung des KI-Modells zu verbessern.

2.2 Entwicklung (Datenaufbereitung, Training und Validierung)

In der Entwicklungsphase erfolgt in der Regel die Aufbereitung von Rohdaten zu Trainingsdaten für die repräsentative Modellierung eines zweckbestimmten, domänenspezifischen Wissens. Häufig erfolgt dabei eine Datenfilterung, eine Datentransformation, das Labeln von Daten (d.h. Anreichern der Daten mit Zusatzinformationen) und eine Datennormalisierung. Anschließend erfolgen das Training sowie eine damit verbundene Validierung des KI-Modells. Hierzu werden drei Arten von Daten genutzt: Trainings-, Validierungs- und Testdaten.

Aufbereitung, Training, Validierung und Testung stellen in dieser Phase die relevanten Verarbeitungstätigkeiten dar. Aus Datenschutzsicht sollten insbesondere die folgenden Punkte Beachtung finden:

- Auswahl und Dokumentation verwendeter KI-Algorithmen für die entsprechenden KI-Modelle
- Implementierung der Umsetzung der Interventionsmöglichkeiten aus Abschnitt 2.1.4
- Festlegung von Zielgrößen für KI-Modelle zu Validierungszwecken
- Entwicklung von Testverfahren, um sicherzustellen, dass die KI-Modelle ihre Zweckbestimmung erfüllen
- Zur Unterstützung der Transparenz- und Rechenschaftsplichten nach Art. 5 Abs. 2 DSGVO sowie Art. 13 und 14 DSGVO, sollten bewusste Entscheidungen des Verantwortlichen über Möglichkeiten zur lokalen Nutzung des KI-Systems oder zur Notwendigkeit von Onlineverbindungen beispielsweise zu Hersteller:innen und Entwickler:innen getroffen werden

[15] Rigaki, Maria, and Sebastian Garcia: „A survey of privacy attacks in machine learning“, ACM Computing Surveys 56.4 (2023), 1-34; Carlini, Nicholas, et al.: „The secret sharer: Evaluating and testing unintended memorization in neural networks“, 28th USENIX security symposium (USENIX security 19), 2019; Haim, Niv, et al.: „Reconstructing training data from trained neural networks“, Advances in Neural Information Processing Systems 35 (2022), 22911-22924; Neel, Seth, and Peter Chang: „Privacy issues in large language models: A survey“, arXiv preprint arXiv:2312.06717 (2023)

[16] Rigaki, Maria, and Sebastian Garcia: „A survey of privacy attacks in machine learning“, ACM Computing Surveys 56.4 (2023), 1-34

- In Zusammenhang mit dem vorherigen Punkt: Dokumentation des Einsatzes von externen Diensten, welche nicht unter direkter Kontrolle der Hersteller:innen und Entwickler:innen stehen, sowie deren potenzielle Verbindungen z. B. zu Dienstanbindung über API-Anknüpfungen

2.2.1 Gewährleistungsziel Transparenz

Der für das Training Verantwortliche hat zu dokumentieren, welche Institutionen Rohdaten zu Trainingsdaten verarbeitet haben. Weiterhin müssen in Bezug auf den ausgewiesenen angestrebten Zweck eines KI-Modells bzw. eines KI-Systems die statistischen Methoden spezifiziert und dokumentiert werden, auf deren Basis die Trainingsdaten bearbeitet werden. Art. 5 Abs. 2 DSGVO fordert Rechenschaft u. a. über die Richtigkeit, Speicherbegrenzung und Datenminimierung der Verarbeitung. Die Informationspflicht über Verarbeitungszwecke, Datenempfänger sowie Speicherdauer wird zudem in Art. 13 DSGVO festgelegt. Es gibt also Transparenzpflichten, welche sich aus der DSGVO ergeben. Diese müssen auch beim Training beachtet werden. Zur Herstellung der notwendigen Transparenz wird als wenig zielführend angesehen, ausschließlich die gewählte Methode des maschinellen Lernens zu dokumentieren. Vielmehr ist es sinnvoll, Transparenz im Sinne der Güte und Erklärbarkeit des KI-Systems herzustellen. Die gewählten Validierungsmethoden schaffen hier das nötige Niveau an Transparenz.[17]

Darüber hinaus sind in Bezug auf Transparenz folgende Fragen zu beantworten und zu dokumentieren: Wo werden Daten vorverarbeitet? Wo werden Daten gespeichert? Wie wird die Integrität der Trainingsdaten und Trainingsergebnisse (d.h. des KI-Modells) sichergestellt? Wo findet das Training statt und ist es evtl. verteilt organisiert? Auf welchen Systemen werden die KI-Modelle validiert und getestet? Wer hat Zugriff auf Trainings-, Validierungs- und Testdaten? Wer hat Zugriff auf die KI-Modelle und die zur Validierung nötigen Ausgaben? Auch die aus diesen Überlegungen resultierenden technischen und organisatorischen Maßnahmen sind zu dokumentieren, werden aber in den nachfolgenden Gewährleistungszielen wieder aufgegriffen.

2.2.2 Gewährleistungsziel Datenminimierung

Bezüglich Datenminimierung im Kontext von Aufbereitung, Training und Validierung sind zwei Aspekte besonders zu beachten. Zum einen dürfen insgesamt nur diejenigen personenbezogenen Daten verarbeitet werden, die zur Erreichung des Verarbeitungszwecks (die definierte Leistung des KI-Modells) notwendig sind. Zum anderen dürfen die beteiligten KI-Systeme lediglich die Daten zur Verarbeitung erhalten, die diese für ihren konkreten Verarbeitungsschritt benötigen. Gerade bei „Compound AI Systems“ (d.h. KI-Systemen, die nicht als Monolith entworfen sind, sondern aus abgegrenzten Subsystemen bestehen), sind die Komponenten modular und wiederverwendbar gestaltet. Damit sind diese oftmals nicht nur für einen bestimmten Zweck einsetzbar, sondern für diverse Zwecke. Dies wiederum bewirkt, dass auch Daten

[17] Dies charakterisiert das genutzte KI-System unabhängig von den für das Training eingesetzten KI-Algorithmen. Artikel 11, 18, Anhang VI, Anhang VII KI-VO spezifizieren hier näheres (meist nur für Hochrisiko-KI-Systeme) im Rahmen eines Qualitätsmanagementprozesses, daran kann sich orientiert werden.

verarbeitet werden, welche nicht für den angedachten Zweck notwendig sind, aber für den technischen Betrieb der Komponente. Daher ist die Notwendigkeit zur Nutzung von Daten sehr unterschiedlich ausgeprägt. Dies beeinflusst die Wahl der Komponenten eines modularen KI-Systems.

Der Verantwortliche muss sicherstellen, dass das KI-Modell nur personenbezogene Daten enthält bzw. reproduzieren kann, wenn dies zur Zweckerreichung erforderlich ist. Auch sollte der jeweilige Verantwortliche vorab eine Hypothese zum erwarteten Einfluss verwendeter Datenkategorien auf das Verhalten des KI-Modells formulieren und diese im Laufe des Trainings (und während des Einsatzes) evaluieren.

2.2.3 Gewährleistungsziel Nichtverkettung

KI-Systeme sind für Verkettung sehr gut geeignet. So ist es möglich, aus den gewählten Trainingsdaten weitere funktionale Zusammenhänge zu lernen. So stecken in Bildern, in Sprache oder auch in Text häufig viel mehr Informationen, als für den angedachten Zweck notwendig sind. Aus Gesichts-Bildern kann mit dem gleichen Trainingsmaterial z. B. das Geschlecht der Person klassifiziert werden, aber auch die ethnische Herkunft, die Stimmung oder evtl. Erkrankungen (z. B. Magersucht/Fettleibigkeit). Durch Kombination verschiedener Datensätze können während des Trainings Zusammenhänge hergestellt und damit Informationen gelernt werden, welche in den ursprünglichen Trainingsdatensätzen nicht unmittelbar ersichtlich waren. Daher muss der Verantwortliche darauf achten, dass KI-Systeme nur zu den in der Designphase festgelegten Zwecken trainiert werden. Es ist insbesondere zu prüfen, ob ein KI-System ungewollte Zwischenergebnisse produziert bzw. ungewollte Antworten zu Anfragen außerhalb des festgelegten Zwecks liefert.

2.2.4 Gewährleistungsziel Intervenierbarkeit

Wenn ein KI-System selbst oder wahrscheinliche Szenarien, in denen das KI-System von Verantwortlichen eingesetzt werden kann, in den Anwendungsbereich von Art. 22 DSGVO fallen, müssen Hersteller:innen und Entwickler:innen Interventionsmöglichkeiten vorsehen. Hierzu gehört es, dass Ausgaben des KI-Systems durch Anwender:innen oder menschliche Kontrolleur:innen in Frage gestellt und von diesen – soweit möglich – nachvollzogen werden können. Hierzu sind entsprechende Funktionen und Ausgabemöglichkeiten vorzusehen.

2.2.5 Gewährleistungsziel Verfügbarkeit

Die Verfügbarkeit von KI-Systemen während der Entwicklungsphase ist eher untypisch definiert. Es ist üblich, dass Trainingsdurchläufe Monate andauern und große Mengen an Rechenressourcen und Energie benötigen. Während dieser Trainingsphase kann ein Verantwortlicher auf das KI-Modell an sich nicht zugreifen, da es sich laufend verändert. Die Verfügbarkeit ist damit nicht durch die schnelle Verfügbarkeit eines Systems an sich charakterisiert, sondern eher durch den zuverlässigen Dauerbetrieb und die Resilienz des Trainings gegen Störungen wie Defekte oder Stromausfälle.

Der Verantwortliche sollte daher planen, wie die Systeme für Training, Validierung und Testung zu gestalten sind, um eine störungsarme Entwicklung zu gewährleisten. Hierbei sind z. B.

Überlegungen zur redundanten Auslegung von Komponenten, zu Backup-Strategien bzgl. der für die Prozesse notwendigen Daten oder zum möglichst reibungsfreien Wechsel auf ein Ersatzsystem bzw. -komponente im Falle einer Störung anzustellen. Im Vorfeld muss dazu durch die Festlegung der Architektur, der Methoden der Trainings-, Validierungs- und Testprozesse sowie der Menge an Daten dafür Sorge getragen werden, dass Informationen zum Gesamtrechen- und -speicher-aufwand vorliegen. Anschließend sind Maßnahmen zu treffen, die gewährleisten, dass für das Training genügend Speichermöglichkeiten zur Verfügung stehen, um auch Störungen während der Trainings- und Validierungsphase tolerieren zu können.

2.2.6 Gewährleistungsziel Integrität

Die Eigenschaft der Integrität für die umfassten Verarbeitungsprozesse Training, Validierung und Testung muss bezüglich zweier Aspekte sichergestellt werden. Zum einen muss die Integrität der Trainings-, Validierungs-, und Testdaten gewährleistet sein, d.h. diese dürfen nicht verfälscht werden oder zur Unterrepräsentation des Eingaberaumes führen. Zum anderen muss auch die Integrität des trainierten KI-Modells betrachtet werden. Dieses muss zum Zeitpunkt des Abschlusses von Training, Validierung und Testung den gewünschten Ausgaberaum ausreichend abbilden und auch die antrainierte Abbildung von Eingaberaum zu Ausgaberaum mit den geplanten Qualitätsmaßen erbringen. Insbesondere die Robustheit gegenüber fehlerhaften oder seltenen Eingaben trägt zur Integrität des KI-Modells im Sinne der Zuverlässigkeit bzw. Richtigkeit des KI-Modells bei. Daher müssen durch den Verantwortlichen Maßnahmen getroffen werden, die den Schutz sowohl der gelernten KI-Modellparameter als auch der Trainings-, Validierungs- und Testdaten vor Manipulation umfassen.

Maßnahmen zur Sicherstellung der **Integrität des trainierten KI-Modells** können unter anderem die Normalisierung und Standardisierung von Rohdaten sowie deren Komplettierung und Fehlerbereinigung umfassen. Auch die Identifikation oder Herstellung von „störsignalbehafteten Daten“, mit denen die Robustheit („Resilienz“) von KI-Systemen getestet und nachgewiesen werden kann, stellen Maßnahmen zur Sicherstellung der Integrität eines KI-Modells dar. Weiterhin ist sicherzustellen, dass Daten zu Trainingszwecken innerhalb eines Trainingsprozesses nicht gleichzeitig zu Validierungs- oder Testzwecken genutzt werden.

Die **Integrität der Trainingsdaten** ist aus zwei Perspektiven zu betrachten. Erstens kann eine nicht ausbalancierte Verteilung der Trainingsdaten die Integrität des KI-Systems gefährden und zweitens können manipulierte Trainingsdaten das Trainingsergebnis verfälschen. Beides führt zu einem KI-System, welches den festgelegten Zweck nicht erreichen kann, da fehlerhafte Ausgaben entstehen. Daher hat der Verantwortliche dafür zu sorgen, dass durch eine adäquate Menge an Trainingsdaten ein hinreichend integres Systemverhalten erreicht wird, welches die Grundgesamtheit aller möglichen Trainingsdaten für den angestrebten Verarbeitungszweck ausreichend repräsentieren kann.

Der Verantwortliche muss darüber hinaus sicherstellen, dass die Datenaufbereitung zweckbestimmt, mit relevanten Daten und in einer korrekten Form geschieht. In jedem Fall muss ausgeschlossen werden, dass ein KI-Modell mit unbefugt manipulierten Daten trainiert wird. Gerade die Technik des Data Poisoning muss wirksam verhindert werden.

2.2.7 Gewährleistungsziel Vertraulichkeit

Insbesondere bei generativen KI-Modellen ist bekannt, dass diese Trainingsdaten unter Umständen unverändert wieder ausgeben können. Daher ist durch die Hersteller:innen und Entwickler:innen zu prüfen, ob das entwickelte KI-Modell für solche Ausgaben durch zufälliges Anwender:innenverhalten bzw. entsprechende Angriffe anfällig ist und welche Gegenmaßnahmen getroffen werden können.[18]

Im Verlauf des Trainings können Zwischenversionen eines KI-Modells erzeugt werden, welche noch anfällig für solche Angriffe sind. Abhängig von der Bewertung des Risikos kann eine Verarbeitung nur auf Infrastrukturkomponenten des Verantwortlichen auszuführen oder eine Verteilung der Verarbeitung unter Einbeziehung von Auftragsverarbeitern möglich sein. Für ein eventuell geplantes Nachtraining bzw. Fine-Tuning des KI-Modells ist die lokale Verarbeitung auf dem genutzten Endgerät zu bevorzugen.

Während des Trainings 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. So kann z. B. ein KI-System zur Blickrichtungsschätzung als Zwischenergebnis die genauen Koordinaten der Augen, Nase und Mundpartie erzeugen (also biometrische Daten), welche durch einen weiteren Verarbeitungsschritt in eine Blickrichtung weiterverarbeitet werden. Es ist daher 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 sollten protokolliert werden.

Ebenfalls muss das Need-to-Know-Prinzip durch den Verantwortlichen sichergestellt werden. Findet ein verteiltes Training statt, benötigen die beteiligten Infrastrukturkomponenten nur die für den jeweiligen Trainingslauf notwendigen Daten. Gleiches gilt für die Validierung und Testung.

2.3 Einführung (Softwareverteilung inkl. Updates)

Im Folgenden wird der Schwerpunkt auf die Endanwendung von Verantwortlichen gelegt, bei denen, wie in Abschnitt 1.1 bereits erwähnt, Entwicklung und Anwendung in einer Hand liegen. Komponenten/Funktionen für ein KI-System können auch im Rahmen von AI-as-a-Service zur Verfügung gestellt werden. Entsprechende Auswirkungen sind beim Gewährleistungsziel Vertraulichkeit (Abschnitt 2.3.7) ausgeführt.

Die Einführung im Sinne der Softwareverteilung eines KI-Systems erfolgt gewöhnlich ohne Einbindung der Anwender:innen. Sie ist nur datenschutzrelevant, wenn in dieser Phase personenbezogene Daten verarbeitet werden. In diesem Fall bedarf es einer rechtlichen Grundlage

[18] Siehe Bundesamtes für Sicherheit in der Informationstechnik (BSI): AI security concerns in a nutshell - Practical AI-Security guide, Kap. 4, https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Practical_AI-Security_Guide_2023.html

und einer Zweckbindung. Die im Folgenden genannten Maßnahmen sind aber auch sinnvoll, wenn unklar oder nicht sicher ist, ob personenbezogene Daten verarbeitet werden (sollen).

2.3.1 Gewährleistungsziel Transparenz

Transparenz trägt neben der Erfüllung gesetzlicher Pflichten auch erheblich zur Vertrauenswürdigkeit bei. Um dieses Gewährleistungsziel zu erreichen, ist es wichtig im Rahmen der Einführung die grundlegenden Entscheidungen, welche dazu geführt haben, ein KI-System zu nutzen, nachvollziehbar zu dokumentieren. Solche Dokumente sollten explizit für die Kommunikation mit den Betroffenen freigegeben sein. Um keine weiteren Gewährleistungsziele zu gefährden, sollten informationssicherheitsrelevante Informationen nicht veröffentlicht werden.

Die Konfiguration und die Bereitstellung von Informationen über

- die Funktionsweise, insbesondere bei automatisierter Entscheidungsfindung (Art. 15 Abs. 1 lit. h DSGVO),
- die menschlichen Eingriffsmöglichkeiten, vergleiche z. B. Art. 22 DSGVO, und
- die Betroffenenrechte

sind für potentiell Betroffene wichtig und sollten daher dokumentiert werden.

Es ist außerdem zu dokumentieren, welche Elemente, die das KI-System für die Entscheidungsfindung verwendet, im Rahmen der Einführung ausgeliefert werden. Das können zum Beispiel Parameter für Neuronale Netze, für Rückschlüsse (Inference) verwendete Daten, Vertrauensindikatoren (AI-Alignment) oder Systemversionen sein. Es ist auch wichtig zu dokumentieren, welche Konfigurationsoptionen („data protection by default“) übergeordnet festgelegt wurden und welche durch Anwender:innen selbständig ausgewählt werden können.

Die Dokumentation sollte in einer Art und Weise erfolgen, dass sie möglichst auch für Nicht-Entwickler:innen verständlich ist. Auch aus Gründen der Risikoeinschätzung empfiehlt es sich für den Verantwortlichen, Dokumentationen übersichtlich und leicht nachvollziehbar zu gestalten.

2.3.2 Gewährleistungsziel Datenminimierung

Um Widersprüche zu anderen Gewährleistungszielen zu vermeiden, sollte auf eine datenminimierende Konfiguration geachtet werden, soweit diese z. B. nicht der Rechenschafts- und Nachweispflicht entgegensteht.

Im Rahmen der Einführung sollten nur Daten bereitgestellt werden, die für den jeweiligen Einsatzzweck/Anwendungsfall benötigt werden. Für die Umsetzung dieser Anforderung sollten Informationen aus geeigneten Einführungstests gewonnen und dokumentiert werden.

Dabei muss an die personenbezogenen Daten gedacht werden, die sich im KI-System befinden und mit ihm zusammen verteilt werden. Hierbei ist ggf. zwischen verschiedenen im KI-System verwendeten KI-Algorithmen (z. B. parametrische und nicht-parametrische KI-Algorithmen) zu unterscheiden. Es muss bedacht werden, dass diese unterschiedlichen Arten von KI-Systemen personenbezogene Daten in unterschiedlicher Art und Weise enthalten können.

Wenn personenbezogene Daten in den Trainingsdaten enthalten waren, können damit entwickelte parametrische KI-Modelle (z. B. Neuronale Netze) selbst ebenfalls einen Personenbezug haben. Diese KI-Modelle benötigen für den Betrieb keine Trainingsdaten, weshalb diese nicht mit ausgebracht werden sollten. Im Gegensatz dazu müssen z. B. bei einem nicht-parametrischen KI-Modell (z. B. K-Nächster-Nachbar) die Trainingsdaten zwingend im Rahmen der Einführung zusammen mit dem KI-Modell verteilt werden. Ohne die Trainingsdaten ist ein solches KI-Modell nicht funktionsfähig.

2.3.3.-2.3.6. Gewährleistungsziele Nichtverkettung, Intervenierbarkeit, Verfügbarkeit und Integrität

Für die Einführung von KI-Systemen sind keine speziellen Anforderungen für diese Gewährleistungsziele ersichtlich, die nicht auch für die Einführung anderer IT-Systeme gelten würden (siehe für diese Abschnitt 1.4).

2.3.7 Gewährleistungsziel Vertraulichkeit

Wie bereits unter Abschnitt 2.3.2 erläutert, kann es notwendig sein, auch Trainingsdaten bereitzustellen. Dabei handelt es sich auch um eine Frage der Vertraulichkeit. In solchen Fällen ist zu berücksichtigen, wie das KI-System den Anwender:innen zur Verfügung gestellt wird. Wird das KI-System den Anwender:innen zentral, z. B. in Form einer Web-Anwendung, zur Verfügung gestellt, so dürfen die benötigten personenbezogenen Daten bei der Bereitstellung nur an die Web-Anwendung bzw. jeweilige Web-Session übertragen werden. Wird das KI-System ggf. offline auf dem Endgerät der Anwender:innen betrieben, müssen diese personenbezogenen Daten ggf. an alle Endgeräte verteilt werden. Dies führt zu einer anderen Bewertung der Risiken für die Betroffenen bezüglich der Vertraulichkeit und der Datenminimierung.

Die Auswirkungen der oben beschriebenen Effekte können sich auch auf personenbezogene Daten erstrecken, die im Rahmen des Trainings im KI-Modell verarbeitet wurden und ggf. im KI-Modell verbleiben (Trainingsdaten, KI-Modellparameter etc.). Bei der Bewertung der Risiken für die Betroffenen ist daher auch an personenbezogene Daten zu denken, wie sie je nach Einzelfall z. B. bei Großen Sprachmodellen (LLMs) zur Tokenisierung im Rahmen der Einbettung geschriebener natürlicher Sprache verwendet werden.[19] Um Datenschutzverstößen bei der Verteilung von KI-Modellen oder -Systemen mit Personenbezug vorzubeugen, sollte bei der Verteilung z. B. an den Einsatz kryptographischer Verfahren gedacht werden.

Es muss bedacht werden, dass ein KI-System mehr als nur ein KI-Modell enthalten kann. Abhängig vom Anwendungsbereich beispielsweise eines LLM kommen gegebenenfalls unterschiedliche KI-Algorithmen zum Einsatz, so dass die Verarbeitung personenbezogener Daten in den einzelnen KI-Modellen nicht pauschal ausgeschlossen werden kann (z. B. Text-to-Image-System).

[19] Der Begriff „Tokenisierung“ stammt aus der Computerlinguistik und bezeichnet die Aufteilung eines Textes in einzelne Wörter oder Buchstabengruppen. Die Tokenisierung stellt eine grundlegende Voraussetzung für die weitere Verarbeitung des Textes z. B. bei der Einbettung dar. Der Begriff der Einbettung bezeichnet in der Informatik eine Einbindung oder Integration in einen bestimmten Kontext.

2.4 Betrieb und Monitoring

Die letzte Phase des Lebenszyklus eines KI-Systems umfasst die produktive Nutzung des KI-Systems, ggf. die Rückkopplung von Ein- und Ausgaben, sowie ggf. Rückmeldungen der Anwender:innen zu der Qualität der Ausgaben, zum weiteren Training der KI-Modelle und die fortwährende Validierung des KI-Systems. Die folgenden Anforderungen und Maßnahmen richten sich primär an Hersteller:innen und Entwickler:innen von KI-Systemen, die diese auch als Verantwortliche betreiben und weiterentwickeln. Weiterhin sollten Hersteller:innen und Entwickler:innen die Anforderungen und Maßnahmen umsetzen, um einen datenschutzkonformen Einsatz ihrer KI-Systeme zu ermöglichen.

2.4.1 Gewährleistungsziel Transparenz

Entscheidungen, die Betroffenen gegenüber rechtliche Wirkung entfalten oder sie in ähnlicher Weise erheblich beeinträchtigen, dürfen nicht zufallsbasiert (gleichwohl aber wahrscheinlichkeitsbasiert) zustande kommen. Bei entscheidungsunterstützenden KI-Systemen ist es daher erforderlich, dass die entscheidungsrelevanten Inhalte der Ausgaben deterministisch und reproduzierbar sind.

Die bekannten maßgeblichen KI-Modellparameter (z. B. bei Entscheidungsbäumen) und Verarbeitungsschritte für das Zustandekommen der Ausgaben eines KI-Systems sind revisionssicher zu dokumentieren. Die in Phase 2 definierten Tests zur Überprüfung der ordnungsgemäßen Arbeitsweise des KI-Systems müssen bei Veränderungen und Updates wiederholt und die Ergebnisse sowie etwaige daraus resultierende Anpassungen dokumentiert werden.

Es muss festgehalten werden, welche Daten durch welches KI-System verarbeitet wurden und ob diese Daten für weiteres Training zur Verfügung gestellt werden.

2.4.2 Gewährleistungsziel Datenminimierung

Wenn die Ausgabe eines spezifischen KI-Systems mehr personenbezogene Daten umfasst als für den vorgegebenen Zweck erforderlich sind, ist zu prüfen, ob eine Anpassung des KI-Systems dahingehend erfolgen muss, dass es künftig nicht mehr zu derartigen Ausgaben kommt. Wenn beispielsweise ein Empfehlungssystem ein ideales Baugrundstück für ermittelte Präferenzen ausgibt, ist für einen potentiellen Kauf wichtig, wem das Grundstück gehört bzw. Kontaktdaten des Verkäufers zu erhalten. Irrelevant ist jedoch, wo der Verkäufer wohnt oder wann dieser Geburtstag hat. Diese Daten sollten bei einem solchen Empfehlungssystem nicht ausgegeben werden. Wenn im Laufe des Einsatzes eines KI-Systems erkennbar wird, dass Daten verarbeitet werden, die zur Erfüllung des festgelegten Zwecks nicht (mehr) erforderlich sind, beispielsweise da sich äußere Umstände geändert haben, muss gegebenenfalls das betreffende KI-Modell mit entsprechend reduzierten Trainingsdaten erneut zu trainieren. Dies ist insbesondere auch dann der Fall, wenn eine diskriminierende Wirkung einzelner Attribute erkannt wird.

Das obsolete KI-Modell muss dann im Nachgang gelöscht werden, sofern dem nicht Protokollierungspflichten oder Transparenzanforderungen entgegenstehen.

Die Verarbeitung von personenbezogenen Ein- oder Ausgaben zum weiteren Training eines KI-Modells stellt eine zweckändernde Weiterverarbeitung dar, die ebenfalls den Anforderungen der DSGVO entsprechen muss. Wenn die Rückkopplung der Ein- und Ausgaben für eine qualitative Verbesserung des KI-Systems im Rahmen seiner Lernfähigkeit verwendet wird, ist eine Anonymisierung dieser Daten, z. B. durch Datenreduktion/-aggregation, zu prüfen. Es müssen für diese Verarbeitung vergleichbare Kriterien wie für das initiale Training festgelegt werden. Zu den nötigen Maßnahmen können z. B. Pseudonymisierung oder eine nachträgliche Prüfung des Personenbezugs des entsprechenden KI-Modells gehören. Ggf. kann bei der Rückkopplung verteiltes Training (engl. Federated Learning) angewendet werden, sodass nur einzelne Infrastrukturkomponenten die jeweiligen Daten verarbeiten und nicht die Daten selbst, sondern nur damit erzielte Trainingsergebnisse geteilt werden.

2.4.3 Gewährleistungsziel Nichtverkettung

Für den Betrieb und das Monitoring von KI-Systemen sind keine speziellen Anforderungen für dieses Gewährleistungsziel ersichtlich, die nicht auch für den Betrieb und das Monitoring anderer IT-Systeme gelten würden (siehe für diese Abschnitt 1.4).

2.4.4 Gewährleistungsziel Intervenierbarkeit

Unterstützung bei Entscheidungen: Ist vorgesehen, das KI-System zur Entscheidungsunterstützung einzusetzen, können technische Maßnahmen zu einer fundierten menschlichen Einwirkung beitragen. Beispielsweise kann das KI-System solange in einem Warte-Status verbleiben („pending“), bis der weitere Fortgang nach menschlicher Kontrolle angestoßen wird, es können feste „menschliche“ Bearbeitungszeiten vorgeschrieben werden bevor eine Bestätigung erfolgen kann um die Auseinandersetzung mit der Ausgabe zu unterstützen oder es kann auch eine regelmäßige Freigabe-Erfordernis geben. Aber auch Hinweise in dem KI-System, dass die Ausgaben nicht perfekt sind, und ggf. die zusätzliche Angabe eines „Unsicherheitsfaktors“, der die Verlässlichkeit der jeweiligen Ausgabe abschätzt, können hilfreich sein. Letzteres kann auch direkt in dem KI-System verwendet werden, sodass Ergebnisse ab einem bestimmten Wert gar nicht erst ausgegeben werden.

Betroffenenrechte: Die Betroffenenrechte nach Art. 16, 17 und 18 DSGVO sind auch bei KI-Modellen und -Systemen von besonderer Relevanz, wenn diese nicht anonym sind. Konkret ergeben sich folgende Anforderungen:

- **Berichtigung**: Unrichtige personenbezogene Daten in Trainingsdaten sind auf Verlangen zu berichtigen, das gilt insbesondere für Ein- und Ausgaben, die für das weitere Training genutzt werden. Auch bei unrichtigen personenbezogenen Ausgaben eines KI-Systems besteht ein Berichtigungsanspruch. Die Berichtigung muss vor der Weiterverarbeitung der Ausgabe erfolgen und die unrichtigen Ausgaben müssen durch geeignete technische oder organisatorische Maßnahmen verhindert werden, bspw. indem sie innerhalb des KI-Systems vor einer Ausgabe an die Anwender:innen gefiltert und entsprechend berichtigt werden.

- **Löschung**: Wenn eine Löschung nach Art. 17 Abs. 1 DSGVO erforderlich ist, ist technisch eine vollumfängliche Löschung der betreffenden Daten notwendig. Das umfasst Ein- und Ausgaben, insbesondere, wenn diese als Trainingsdaten weiterverwendet werden, zusätzlich aber auch das KI-Modell, wenn dieses die zu löschende Information „beinhaltet". Davon ist in aller Regel zumindest dann auszugehen, wenn die Information in den verwendeten Trainingsdaten und einer Ausgabe des KI-Modells enthalten war. In diesem Fall muss regelmäßig ein neues KI-Modell trainiert werden, wobei die zu löschenden Daten dann nicht mehr in den Trainingsdaten enthalten sein dürfen, oder das bestehende KI-Modell muss geeignet nachtrainiert werden (engl. Machine Unlearning). Der Erfolg des Machine Unlearning muss nachgewiesen werden, indem die betreffenden Daten nicht mehr basierend auf dem KI-Modell als in den Trainingsdaten enthalten identifiziert werden können. Geeignete technische und organisatorische Maßnahmen, z. B. Ein- und Ausgabefilter, sind als Übergangslösung einzusetzen. Allerdings stellen solche mitigierenden Maßnahmen, insbesondere auch Filter, an sich keine Löschung dar. Es besteht nicht zwingend in jedem Fall ein Anspruch auf die Löschung personenbezogener Daten, sondern in Art. 17 Abs. 2 und 3 DSGVO sind Einschränkungen und Ausnahmen formuliert.[20]
- **Einschränkung**: Dabei handelt es sich laut Art. 4 Nr. 3 DSGVO um die „Markierung gespeicherter personenbezogener Daten mit dem Ziel, ihre zukünftige Verarbeitung einzuschränken". Was die Trainingsdaten und Protokolldaten betrifft, sind diese wie herkömmliche Daten zu behandeln. Was KI-Modelle betrifft, kann sich an den Ausführungen zur Löschung orientiert werden. Geeignete Filter können angewendet werden, um die Verarbeitung der entsprechenden Daten einzuschränken, indem einzuschränkende Ein- und Ausgaben erkannt und blockiert werden.

Bei KI-Systemen, die mit einem hohen Risiko für Betroffene einhergehen, werden höhere Anforderungen an risikoverringernde Maßnahmen gestellt. In einigen Fällen kann es erforderlich sein, den Stand der Wissenschaft (wie bspw. Machine Unlearning) zu implementieren und nicht nur den Stand der Technik, um überhaupt ein vertretbares Risiko zu erreichen. In besonders risikobehafteten Fällen kann auch das nicht ausreichend sein, so dass bspw. eine Interessenabwägung zu Ungunsten des Betriebs eines KI-Systems ausfällt.

2.4.5 Gewährleistungsziel Verfügbarkeit

Für den Betrieb und das Monitoring von KI-Systemen sind keine speziellen Anforderungen für dieses Gewährleistungsziel ersichtlich, die nicht auch für den Betrieb und das Monitoring anderer IT-Systeme gelten würden (siehe für diese Abschnitt 1.4).

[20] Beispielsweise können Protokollierungspflichten und andere gesetzliche Vorgaben Löschersuchen entgegenstehen. Auch öffentliche Interessen im Bereich der öffentlichen Gesundheit gemäß Art. 9 Abs. 2 lit. h und i sowie Art. 9 Abs. 3 DSGVO können einer Löschung entgegenstehen.

2.4.6 Gewährleistungsziel Integrität

Änderungen in der Wissensdomäne (z. B. Änderungen von Kontexten, rechtlichen Regelungen, technischen Änderungen, Zunahme des Wissens) müssen identifiziert werden, da diese dazu führen können, dass ein KI-Modell die veränderte Wissensdomäne nicht mehr adäquat repräsentiert. Das kann die Risiken für die Rechte und Freiheiten Betroffener erhöhen. Auf die Änderungen muss in dem Fall mit geeigneten Maßnahmen reagiert werden, um die Integrität zu gewährleisten. In einigen Fällen kann es erforderlich sein, ein neues KI-Modell mit aktualisierten Trainingsdaten zu entwickeln.

Die Einhaltung der in Phase 2 definierten Qualitätsanforderungen ist regelmäßig zu evaluieren. Bei KI-Systemen, die kontinuierlich im Rahmen ihrer Lernfähigkeit weiterentwickelt und angepasst werden, muss die Aufrechterhaltung der Qualität im Laufe der Zeit besonders aufmerksam verfolgt werden. Es müssen plötzliche und kontinuierliche Verhaltensänderungen von KI-Systemen festgestellt und deren Auswirkungen auf die Risiken bewertet werden.

Eingaben außerhalb des zulässigen Bereichs sollten durch Eingabefilter erkannt werden und zu einem deutlich erkennbaren Hinweis führen. Zusätzlich sollte erkannt werden, wenn Eingaben (von einem Angreifer) so verändert wurden, dass sie eine Fehlentscheidung des KI-Systems verursachen (engl. Evasion Attacks).

Es sollten regelmäßige Risikoeinschätzungen zum Beispiel durch Red Teaming vorgenommen werden, insbesondere bei öffentlich verfügbaren KI-Systemen.

Die Integrität der Daten für die Rückkopplung muss sichergestellt sein. Angriffe durch manipulierte Ein- und Ausgaben oder inkorrekte Rückmeldungen zur Qualität der Ausgaben, die für weiteres Training genutzt werden, müssen unterbunden werden. Insbesondere beim verteilten Lernen, wobei die Trainingsdaten nicht geteilt werden, kann es schwierig sein, die Qualität der gesamten Daten zu bestimmen. Hierbei können die in Phase 1 bereits genannten Maßnahmen eingesetzt werden.

2.4.7 Gewährleistungsziel Vertraulichkeit

Abhängig von dem Einsatzzweck eines KI-Systems können sich unterschiedliche Maßnahmen ergeben. Wenn beispielsweise die Hersteller:innen und Entwickler:innen ein KI-System auch Personen zur Verfügung stellen, die keinen Zugriff auf die Trainingsdaten haben, z. B. über eine API, muss die Extraktion der Trainingsdaten verhindert werden. Sofern das KI-Modell selbst personenbezogen ist, müssen Hersteller:innen und Entwickler:innen auch Maßnahmen zur Verhinderung der Modellextraktion ergreifen.

Bei einem Update von zusätzlichen Datenquellen, wie z. B. bei RAG-Systemen, muss jedes Mal betrachtet werden, ob diejenigen, die Zugriff auf das KI-System haben, auch Zugriff auf die neuen Daten haben dürfen.

3 Fazit

Insgesamt sollten Hersteller:innen und Entwickler:innen in jedem Entwicklungsschritt ihres KI-Modells oder -Systems die sieben Gewährleistungsziele Datenminimierung, Verfügbarkeit,

Vertraulichkeit, Integrität, Intervenierbarkeit, Transparenz und Nichtverkettung von Anfang an mitbedenken. Insbesondere in Bezug auf die in der Regel sehr große Mengen an Daten, die für die Entwicklung aber auch das regelmäßige Fine-Tuning oder die Updates von KI-Systemen benötigt werden, muss von Beginn an das Datenschutzrecht eingehalten werden, um die Rechte und Freiheiten natürlicher Personen zu schützen.

4 Glossar

Das folgende Glossar bezieht sich auf die Verwendung der Begriffe in diesem Dokument. Das Ziel ist dem Lesenden das Durchdringen des Textes zu erleichtern. Deshalb bedient sich das Glossar zum Teil vereinfachter Darstellungen, um das Verständnis für ausgewählte Aspekte des Textes zu erleichtern.

AI-as-a-Service	Unter AI-as-a-Service versteht man das Bereitstellen von KI-Modellen und -Systemen, aber auch bspw. KI-Algorithmen für Maschinelles Lernen, als cloudbasierte Dienstleistung.
Anwender:innen	Anwender:innen sind natürliche Personen, die ein KI-System verwenden. Anwender:innen müssen nicht zwangsläufig Betroffene sein.
Design	Das Design umfasst vor allem die Planung und vorbereitende Schritte für die Umsetzung eines neuen KI-Systems und beschreibt als Ergebnis die Anforderungen an die Entwicklung.
Entwicklung	Die Entwicklung umfasst die Implementierung der verwendeten KI-Algorithmen und das anschließende Training der KI-Modelle mit den Trainingsdaten.
Federated Learning Architektur	Eine Federated Learning Architektur ist eine Sammlung an KI-Systemarchitektur-Variationen, bei denen mehrere Geräte unter der Aufsicht eines zentralen Servers gemeinsam ein KI-Modell trainieren, ohne ihre jeweiligen privaten Trainingsdaten zu teilen.
KI-Algorithmus	Ein KI-Algorithmus ist eine geschlossene mathematische Handlungsvorschrift zur Lösung eines Problems aus dem Bereich der KI, die sich in einem Computerprogramm verwenden lässt.
KI-Modell	Ein KI-Modell ist das Ergebnis, das sich aus der Anwendung eines KI-Algorithmus auf Trainingsdaten ergibt. Es dient als Komponente eines KI-Systems, die aus Eingaben Schlussfolgerungen zieht, um Ausgaben zu erzeugen. Ein KI-Modell kann auch für weiteres Training, Fine-Tuning oder die weitere Entwicklung bestimmt sein. – (Nach der Stellungnahme 28/2024 des EDSA und der OECD-Definition)
KI-System	Ein KI-System ist ein maschinengestütztes System, das auf einem oder mehreren KI-Modellen beruht und sowohl für die direkte Verwendung als auch für die Integration in andere KI-Systeme dient. Es ist für einen in unterschiedlichem Grade autonomen Betrieb

	ausgelegt und kann nach seiner Betriebsaufnahme anpassungsfähig sein. Es leitet aus den erhaltenen Eingaben für explizite oder implizite Ziele ab, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können. – (Nach Art. 3 Nr. 1 KI-VO, Art. 3 Nr. 66 KI-VO)
Konfiguration	Eine Konfiguration ist eine Anpassung an ein vorhandenes System mittels konkreter Einstellungen, die gewöhnlich in Dateien oder Datenbanken gespeichert werden.
KI-Modellparameter	KI-Modellparameter sind die lernbaren Parameter innerhalb eines KI-Modells (bspw. Gewichte eines Neuronalen Netzes). Diese werden durch die Anwendung eines KI-Algorithmus auf Trainingsdaten bestimmt.
Performance Evaluation	Messung der Performanz eines KI-Systems anhand vorher festgelegter Parameter. In Bezug auf die Reliabilität eines KI-Systems werden hierfür primär Genauigkeit, Richtigkeit und Präzision aber auch Verteilung der Datendichte und Abweichungen vom richtigen Wert mit unterschiedlichen Modell-abhängigen Metriken gemessen. Abseits der technischen Performanz sind auch Metriken zur Berechnung der Fairness, Interpretierbarkeit, Sicherheit, Robustheit, des Datenschutzes, der Rechtmäßigkeit oder des Schutzes weiterer Ziele relevant.
RAG	Retrieval-augmented generation (RAG) ist ein Verfahren, das über eine Suchfunktion einem LLM zusätzliche, für die jeweilige Eingabe relevante Informationen aus einem Datenpool zur Verfügung stellt.
Red Teaming	Anwendung manueller oder automatisierter Methoden, um ein Sprachmodell auf schädliche Ausgaben zu untersuchen und zu trainieren, um solche Ausgaben zu vermeiden. Anwendungsfälle sind z. B. Bias-Erkennung, Robustheitstests oder Konformitätsprüfungen. Das Konzept wird besonders in selbstlernenden KI-Systemen verwendet.
Rohdaten	Bei Rohdaten kann es sich um in eigener Zuständigkeit erhobene oder aus bereits vorhandenen Datenquellen gewonnene Daten in Ihrer ggf. unveränderten Form handeln. Vor der Verwendung als Trainingsdaten werden diese durch eine Vorverarbeitung aufbereitet.

Rückkopplung	Ein Prozess bei dem die Ein- oder Ausgaben eines KI-Systems und ggf. Rückmeldungen der Anwender:innen zur Qualität der Ausgaben zur Anpassung des KI-Systems oder darin enthaltener KI-Modelle verarbeitet werden.
Synthetische Daten	Synthetische Daten sind maschinell generierte Daten, die reale Daten nachahmen und dieselben relevanten Eigenschaften wie diese aufweisen, ohne dieselben Informationen zu beinhalten. Eine Möglichkeit zur Erzeugung synthetischer Daten sind generative KI-Systeme.
Testdaten	Testdaten werden für eine unabhängige Bewertung (Evaluation) eines KI-Modells oder Systems verwendet, um die erwartete Leistung dieses KI-Modells oder -Systems vor dessen Betriebsaufnahme zu bestätigen. – (Nach Art. 3 Nr. 32 KI-VO)
Trainingsdaten	Trainingsdaten werden aus Rohdaten aufbereitet. Bei Trainingsdaten handelt es sich um Daten, die zum Trainieren eines KI-Modells verwendet werden. Abhängig vom gewählten KI-Modell werden diese Daten dauerhaft dem KI-Modell zur Verfügung gestellt oder durch diese Daten dessen KI-Modellparameter angepasst. – (Nach Art. 3 Nr. 29 KI-VO)
Validierungsdaten	Validierungsdaten werden während der Entwicklung zur Evaluation eines KI-Modells und zur Einstellung seiner nicht erlernbaren Parameter und seines Lernprozesses verwendet, um unter anderem eine Unter- oder Überanpassung zu vermeiden. Häufig werden Validierungsdaten von dem Training den Trainingsdaten entnommen und werden nicht für das Training verwendet. – (Nach Art. 3 Nr. 30, 31 KI-VO)