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

Im Juni 2025 hat die DSK eine Orientierungshilfe zu empfohlenen technischen und organisatorischen Maßnahmen bei der Entwicklung und beim Betrieb von KI-Systemen im Sinne des Art 3 Nr. 1 KI-Verordnung mit Personenbezug veröffentlicht.

Das Papier stellt dabei datenschutzrechtliche Anforderungen/Empfehlungen, mit Hauptaugenmerk auf die datenschutzrechtlichen Grundsätze Art. 5, 32 DSGVO, hinsichtlich der unterschiedlichen Schritte im Lebenszyklus einer KI auf und unterscheidet zwischen folgenden vier Phasen:

1. **Designphase**
   Umfasst Planung vorbereitender Schritte, Aufstellung Anforderungskatalog, Auswahl der Daten, Datensammlung (Trainingsdaten)

2. **Entwicklungsphase**
   Umfasst konkrete Umsetzung, insb. Aufbereitung und Verarbeitung erhobene Daten/ Trainingsdaten, Training und Validierung

3. **Einführung**
   Softwareverteilung inkl. Updates

4. **Betrieb und Monitoring**
   Freigabe der Nutzung, Übergang in Produktivbetrieb

Die Handlungsempfehlungen sind auf die jeweiligen Phasen bezogen und richten sich nach dem Transparenzgebot, Grundsatz der Datenminimierung, der Intervenierbarkeit, Nichtverkettung, Verfügbarkeit, Vertraulichkeit und Integrität (Gewährleistungsziele der SDM):

- **Transparenzgebot**
  In allen Phasen soll ausführliche und nachvollziehbare Dokumentation datenschutzerheblicher Umstände stattfinden.
  Das betrifft insb. Ergebung, Auswahl, Nutzung, Verarbeitung sowie Erforderlichkeit von Daten. Für Nutzer soll nachvollziehbar sein, welche grundlegenden Entscheidungen zur KI-Nutzung geführt haben und auf welcher Basis die KI ihre Entscheidungen trifft.

- **Datenminimierung**
  Während aller Phasen des Lebenszyklus sollen nur die Daten verwendet werden, die für die KI/Zweckerreichung erforderlich sind.
  Dies kann durch die zielgerichtete Festsetzung von Verarbeitungswecken, anhand derer die Auswahl von Algorithmen und qualitativ hochwertiger Daten erfolgen soll, umgesetzt werden. Der Verzicht auf Daten soll nur in solch einem Umfang stattfinden, dass es nicht zu einer Gefährdung der Zwecke sowie der ordnungsgemäßen Funktion der KI kommt. Zudem hat der Verantwortliche zu

allen Zeitpunkten zu prüfen, ob die Verwendung von bestimmtem Daten zur Zweckerreichung (noch) nötig ist, und ggf. die KI anzupassen.

- **Nichtverkettung**
  Eine Verwendung hochkorrelierender Ersatzvariablen stellt eine Verkettung dar, mithin eine unerlaubte Verwendung von Daten. Soweit direktes Verarbeitungsverbot nach Art. 9 DSGVO besteht, gilt dies auch für die Herleitung des Datums aus scheinbar unkritischen personenbezogenen Daten. Verwender muss Sorge tragen, dass Training nur hinsichtlich der festgelegten Zwecke erfolgt.

- **Intervenierbarkeit**
  Maßnahmen bzw. Implementierung von Möglichkeiten zur Wahrnehmung von Betroffenenrechte sowie behördlicher Anordnungen müssen zu jeder Zeit getroffen werden und die Geltendmachung – auch hinsichtlich Rohdaten – möglich sein. Es wird Empfehlung zur Verwendung von Technologien wie „Machine Unlearning“ gegeben, wo dies möglich ist. Die Intensität der Umsetzung ist von der Gefahr für die Betroffenen abhängig.

- **Verfügbarkeit**
  In Designphase sollen Konzepte zur störungsarmen Entwicklung sowie Betrieb erstellt werden, welches über die jeweiligen Phasen über umgesetzt sowie verbessert werden soll.

- **Integrität**
  Verantwortlicher muss schon bei der Auswahl der Rohdaten auf die Qualität dieser ein besonderes Augenmerk legen. (Quelle, vorhandener Bias, etc.). Empfehlung zur statistischen Untersuchung und individuellen Betrachtung der Daten sowie genaue Untersuchung der Verteilung der Daten und Pilotstudien. Bei vortrainierten Modellen Rücksicht auf etwaiges „Backdoor Poisoning“. Während des Trainings müssen Maßnahmen zum Schutz der erlernten KI-Modellparameter sowie der Trainings-, Validierungs- und Testparameter getroffen werden. Es muss Sorge für eine ausreichende Grundlage an korrekten Daten getragen werden.
  In der Betriebsphase soll die KI auf Wissensänderungen, Verhaltensänderungen o.Ä. kontrolliert und entsprechend reagiert werden.

- **Vertraulichkeit**
  Es sollen ab Designphase Maßnahmen zur Verhinderung von unerlaubtem Zugriff auf Daten sowie ungewollten Abfluss von Daten ergriffen werden. Hinweis auf generelle Gegenmaßnahmen wie Privacy-Preserving-techniken sowie Regularisationstechniken zur Generalisierung des KI-Modells. Insb. bei generativer KI muss verhindert werden, dass Testdaten in der Anwendung wiedergegeben werden. Es muss sichergestellt werden, dass Zwischenergebnisse aus Testphasen, die kritische Daten enthalten können, nicht langfristig gespeichert werden und der Zugriff eingeschränkt ist. Schwere der Maßnahmen sowie Abwägungen von Gefahr für pers. Daten abhängig.
  In der Anwendung muss verhindert werden, dass es zu Zugriff/Extraktion der Trainingsdaten kommt.