HK2
Rechtsanwälte

Datenschutzrechtliche und sicherheitstechnische Vorgaben für KI-Systeme

Bernhard Kloos, 8. März 2023

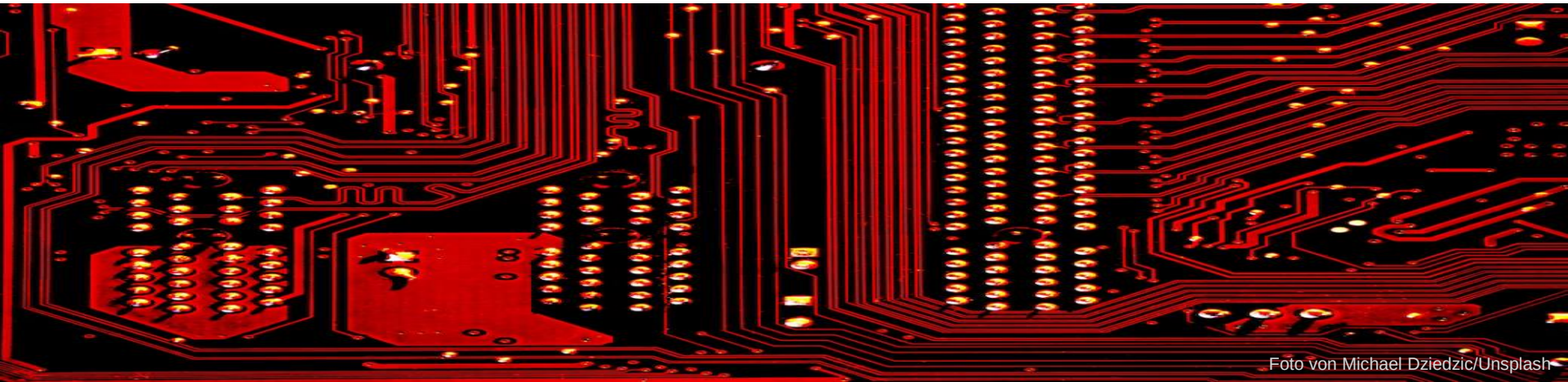


Foto von Michael Dziedzic/Unsplash

Bernhard Kloos

- Partner/ Rechtsanwalt, HK2 Rechtsanwälte
- Geschäftsführer HK2 Comtection GmbH
- Autor zu Datenschutz in KI & Recht kompakt
- IT-Recht
- Datenschutz
- Werberecht

Was machen wir?

- Rechtsrahmen für KI
- Datenschutzrechtliche Anforderungen
- Maßnahmen
- Risiken

Vorschlag einer EU-Verordnung zur Regulierung Künstlicher Intelligenz vom 21.04.2021 (KI-Verordnung)

HK2 Rechtsanwälte

- Verordnungsentwurf zur Förderung sicherer und rechtskonformer KI im gesamten Binnenmarkt
- Harmonisierte und unmittelbar wirkende Vorschriften für die Entwicklung, das Inverkehrbringen und die Verwendung von KI-Systemen in der Europäischen Union
- Status: Finale Abstimmung im Europäischen Parlament läuft, anschließend Trilog, Abschluss 2023/2024 erwartet
- Risikobasierter Ansatz:
  - unannehmbares Risiko / Verbot (z.B. Beeinflussung mit möglichen Gesundheitsschäden, biometrische Echtzeit-Fernidentifikation im öffentlichen Raum zur Strafverfolgung (mit Ausnahmen))
  - Hochrisikosysteme (z.B. Recruiting, Personalmanagement, KRITIS, aber auch GPAIS / LGAIM bei entsprechender Einsatzmöglichkeit)
  - geringes oder minimales Risiko (z.B. Chatbots, Spamfilter)

KI-Verordnung – Anforderungen an Hochrisiko-KI

- Risikomanagementsystem (Analyse, Bewertung, Maßnahmen) während des Lebenszyklus der KI -> Risiken müssen vertretbar sein
- Obligatorisches Testen zu jedem geeigneten Zeitpunkt
- Data Governance: Verwendung qualitätsgesicherter Trainings-, Validierungs- und Testdatensätze (einschließlich auf Relevanz, Verfügbarkeit, Bias und Lücken oder Mängel)
- Dokumentation und Aufzeichnung / Protokollierung
- Transparente Information der Nutzer
- Menschliche Aufsicht
- Genauigkeit, Robustheit (gegen Fehler, Störungen und Unstimmigkeiten) und Cybersicherheit
- Qualitätsmanagementsystem (einschließlich Beobachtung, Meldung von Vorfällen und Rechenschaft / persönliche Verantwortung)
- Konformitätsbewertung

Datenschutzrechtliche Eckpunkte
(gemäß DSK)

HK2 Rechtsanwälte

- Für KI-Systeme gelten die Grundsätze für die Verarbeitung personenbezogener Daten, etwa Transparenz und Zweckbindung (Art. 5 DSGVO)
- Grundsätze müssen durch frühzeitig geplante technische und organisatorische Maßnahmen (TOM) von Verantwortlichen umgesetzt werden (Art. 25 DSGVO)
- Die Sicherheit der Verarbeitung muss gewährleisten sein unter Berücksichtigung des Stands der Technik, der Implementierungskosten und der Art, des Umfangs, der Umstände und der Zwecke der Verarbeitung sowie der unterschiedlichen Eintrittswahrscheinlichkeit und Schwere des Risikos (z.B., durch Pseudonymisierung, Verschlüsselung die Fähigkeit, die Vertraulichkeit, Integrität, Verfügbarkeit und Belastbarkeit der Systeme und Dienste im Zusammenhang mit der Verarbeitung auf Dauer sicherzustellen, Art. 32 DSGVO)
- Gewährleistungsziele sind Transparenz und Erklärbarkeit, Datenminimierung, Nichtverkettung, Intervenierbarkeit, Verfügbarkeit, Integrität und Vertraulichkeit

Rechtsgrundlagen

- Einwilligung, d.h.
  - freiwillig ohne Druck oder Einfluss
  - spezifisch
  - informiert (Zweck, Funktion, Dauer, Dritte)
  - aktiv und unzweideutig
  - jederzeit widerruflich
- Vertragsdurchführung
- Berechtigtes Interesse
- Rechtliche Pflicht

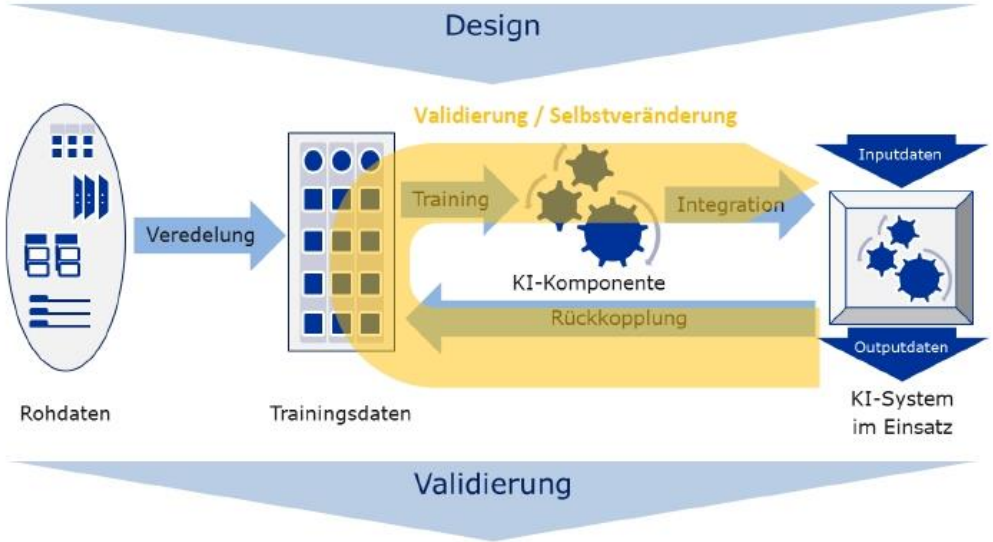
Technische und organisatorische Maßnahmen (TOM) bei Entwicklung und Betrieb von KI-Systemen

(gemäß DSK
https://www.datenschutzkonferenz-online.de/media/en/20191106_positionspapier_kuenstliche_intelligenz.pdf, Grafik S. 2)

HK2 Rechtsanwälte

Grundlage ist der allgemeine Lebenszyklus von KI-Systemen:

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. **Nutzung** des KI-Systems
6. **Rückkopplung** von Ergebnissen und Selbstveränderung des Systems



TOM Design

Festzulegen und zu dokumentieren sind:

- **Zweck** des KI-Systems und der KI-Komponenten
- angestrebte **Güte** des KI-Systems
- **Eingabe- und Ausgabepara**meter
- **Hypothese des gewünschten Zusammenhangs** zwischen Eingabe- und Ausgabeparametern
- **Festlegung von unerwünschtem Verhalten**
- Festlegung eines geeigneten KI-Verfahrens und Dokumentation dessen Auswahl
- Beteiligte Personen und Rechte

TOM Veredelung

In dieser Phase sind zu dokumentieren:

- **Herkunft der Rohdaten** klären, normalisieren, standardisieren, komplettieren, fehlerbereinigen und möglichst reduzieren und / oder pseudonymisieren
- Festlegung des **Verfahrens zur Datenveredelung**
- **Ausschluss** diskriminierender oder ungewünschter Einflussfaktoren (**Bias**)
- **Relevanz und Repräsentativität der Trainingsdaten**
- Möglichkeiten zur Anonymisierung oder Synthetisierung
- Abschätzung der Datenmengen, um angestrebte Güte zu erreichen
- **Erkennung und Korrektur** von falschen oder diskriminierenden Roh- und Trainingsdaten, Testdaten und Testverfahren
- **Manipulationsfreiheit, Vertraulichkeit und Verfügbarkeit von Roh- und Trainingsdaten**

TOM Training

HK2 Rechtsanwälte

- Prüfung der **Kompatibilität der Zwecke** der KI-Komponente mit den Verwendungszwecken der Trainingsdaten
- **Dokumentation der Trainingsdaten** und Klärung ihrer Herkunft
- Festlegung von Primär-Trainings-, Verifikations- und Testdatensätzen einschließlich des Verfahrens zur Einteilung und den Regeln zur Verwendung
- **Verfahren zur Ermittlung der Güte** einschließlich ggf. Prozess zur Erweiterung der Trainingsdaten, um die Güte zu erreichen
- **Verhinderung von unbefugten Manipulationen** an KI-Komponenten

TOM Validierung

HK2 Rechtsanwälte

- **Test mit gewünschten und mit störsignalbehafteten (synthetischen) Testdaten** auf Basis des erwarteten und unerwünschten Verhaltens
- **Prüfung der Hypothese und der Erwartungen**
- **Dokumentation der Güte** der KI-Komponente inklusive der ermittelten Fehlerrate und Systemstabilität,
- Untersuchung der KI-Komponente auf **Erklärbarkeit und Nachvollziehbarkeit**
- **Evaluation** des ausgewählten KI-Verfahrens bezüglich alternativer, erklärbarerer KI-Verfahren

TOM Einsatz I

- **Überwachung des Verhaltens** der KI-Komponente und des -Systems
- Möglichkeit des **Eingreifens einer Person** in den Entscheidungsprozess
- Entscheidungen mit hohen Risiken dürfen von KI-Systemen nur vorbereitet werden, Möglichkeit des Stoppens oder des Ersetzens einer KI-Komponente durch Fall-Back-Lösung
- **Protokollierung von finalen Entscheidungen**, deren Freigabe/Bestätigung/Ablehnung, Zeitpunkt und ggf. entscheidende Person
- **Prüfung und Bewertung** der Hypothese und der Erwartungen auf Basis des Verhaltens im Betrieb sowie der Güte des KI-Systems und seiner KI-Komponenten

TOM Einsatz II & Rückkopplung

HK2 Rechtsanwälte

- regelmäßige **Evaluierung welche Eingabe- und Ausgabeparameter** der KI-Komponenten für das gewünschte Verhalten des KI-Systems relevant und erforderlich sind, ggf. Anpassung
- Berechnungen der KI-Komponente möglichst auf Endgerät des Nutzer ohne Übermittlung der Daten (**Federated Learning**), ansonsten auf Geräten des Verantwortlichen
- bei Übermittlung der Daten an KI-Komponente **Ende-zu-Ende-Verschlüsselung** einsetzen
- **Auskunftsmöglichkeit für Betroffene** zum Zustandekommen von Entscheidungen, Prognosen sowie über das Verfahren zur Ermittlung der Güte des KI-Systems und die festgestellte Güte
- **Ausschluss der Nutzung durch Unbefugte** und zu nicht vorgesehenen Zwecken (Berechtigungskonzept)
- **Mechanismen zum Anfechten und Korrigieren** vorsehen und darauf hinweisen
- **Verhinderung von unbefugten Manipulationen**
- **Rückkopplung von erzeugten Outputdaten** zum weiteren Training, möglichst anonymisiert oder pseudonymisiert

TOM Datenschutz

Generell: Die Maßnahmen sind so zu gestalten, dass der Verantwortliche seinen datenschutzrechtlichen Pflichten in allen Verarbeitungsphasen und -schritten nachkommen kann, etwa

- Information (z.B. Hinweisschild auf Sensoren beim autonomen Fahren)
- Auskunft
- Berichtigung
- Löschung
- Einschränkung der Verarbeitung
- Widerspruchsrecht
- Widerruf der Einwilligung
- Verbot automatisierter Entscheidung

Sicherheit von KI-Systemen - Angriffstypen (BSI)

HK2 Rechtsanwälte

Bundesamt für Sicherheit und Informationstechnik (BSI) hat mehrere Studien auf dem Weg zu Prüfkriterien für KI-Systeme in Auftrag gegeben.
Klassische Angriffstypen:

- Backdoor Attacks

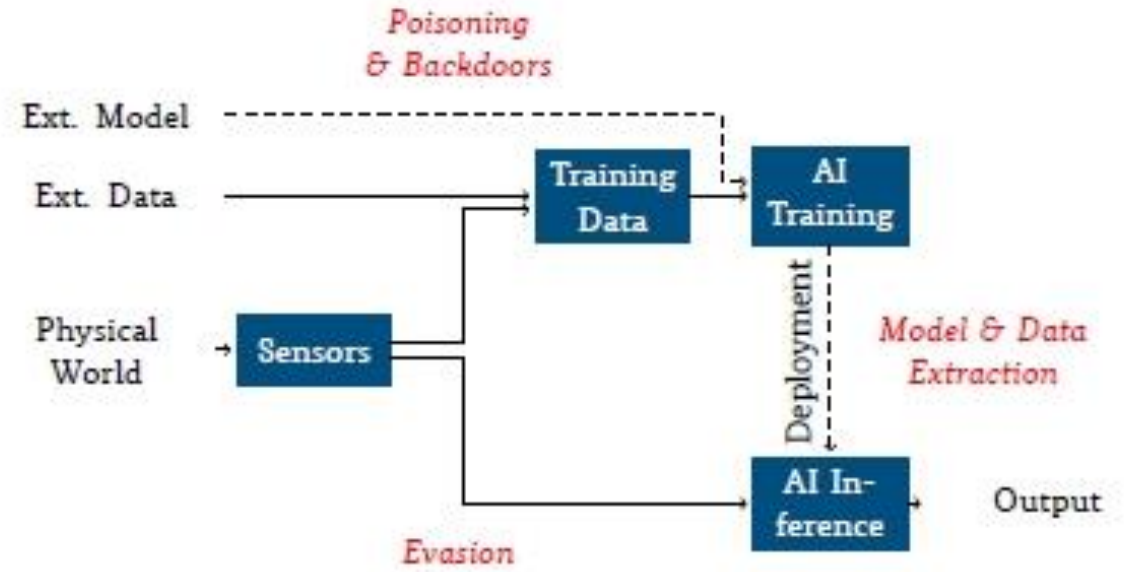
  Während des Trainings werden Schwachstellen in das KI-System eingefügt, die dann vom Angreifer ausgelöst werden können.
- Poisoning Attacks/ Vergiftungsangriffe:

  Die Vergiftungsproben werden während des Trainings der KI injiziert und führen zu einer Beeinträchtigung
- Evasion/Umgehungsangriffe:

  Durch Manipulation von Eingabedaten verleiten Angreifer das KI-Modell im Betrieb zu vom Entwickler nicht vorgesehenen Ausgaben.
- Data & Model Extraction Attacks:

  Angreifer extrahieren Informationen hinsichtlich der Trainingsdaten bis hin kompletten Trainingsmuster aus dem angewandten Modell oder die Funktionalität des Modells.

Sicherheit von KI-Systemen
(BSI)



Quelle: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Security-of-AI-systems_fundamentals.pdf
S. 10

Risikomatrix

Quelle: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Security-of-AI-systems_fundamentals.pdf
S. 4

HK2 Rechtsanwälte

Risk	Low	Medium	High
Deployment Environment	Known lab environment	Confined open-world deployment	Publicly available online service
Training Data Origins	Well established data sets	Self-created data sets	Data sets of unknown origin
Inference Data Origins	Local, e.g., a fixed data set	Air gap, e.g., a camera system	External queries, e.g., the internet
Degree of Autonomy	Recommendation to human expert	Decision observed by human expert	Autonomous decision by the AI
Output Type	System decision	Discretized model output	Full model output
User Access Origins	Local admins	All internal users	Global access
User Access	Single query	Limited number of queries	Unlimited queries
Model Origins	Locally trained model with custom architecture	Locally retrained public architecture	Publicly available model
Learning Strategy	Single training	Continuous and regular updates	Continuous and spontaneous updates

Sicherheit von KI-Systemen
(BSI)

HK2 Rechtsanwälte

Unterschiedliche Kenntnisstände des Angreifers sind zu betrachten:

Beobachtung	Teilkenntnis	Vollständige Kenntnis
KI-Entscheidung, Output Verteilung	Architektur, Trainingsaufbau	Verteidigungen, Architektur, Trainingsaufbau
Black-Box	Gray-Box	White-Box

In a nutshell

HK2 Rechtsanwälte

- Rechts- und Branchenentwicklungen beobachten (KI-Verordnung, BSI Prüfkriterien)
- Wissen was reinkommt (Data Governance)
- Datenschutzrechtliches Konzept
  - Synthetische Daten, Anonymisierung, Pseudonymisierung
  - Rechtsgrundlage passend zum Verwendungszweck
  - Risikoschonende Struktur (föderiert, verschlüsselt oder (erst einmal) mit unsensiblen Daten)
  - Datentöpfe getrennt halten
- Verstehen, was rauskommt: Erklärbarkeit und Nachvollziehbarkeit, Ausschluss von Bias
- Risikomanagement von Beginn und durch alle Phasen implementieren
  - Welche gibt es?
  - Wie gravierend sind sie?
  - Maßnahmen zur Eindämmung
- Testen auf erwartete Ergebnisse, auch mit falschen Testdaten
- Überwachung und menschliche Aufsicht
- Dokumentation (einschließlich QS-Maßnahmen, Zuständigkeiten) und Protokollierung
- Transparente Information

Haben Sie Fragen?

Rechtsanwalt

Bernhard Kloos

HK2 Rechtsanwälte

Hausvogteiplatz 11 A
10117 Berlin

Telefon +49 (0)30 27 89 00 – 0
Telefax +49 (0)30 27 89 00 – 10
E-Mail kloos@hk2.eu

www.hk2.eu