AI-Governance

Der AI-Act in der Praxis

Alexander Thamm GmbH

Lucia Karch, Patrick Zimmermann

März 2024

VORSTELLUNG

MAKE DATA WORK

Patrick Zimmermann

Teamlead & Data/AI Compliance Track Lead bei Alexander Thamm GmbH

Lucia Karch

Director Casebase & Principal Venture Builder

[at] alexanderthamm

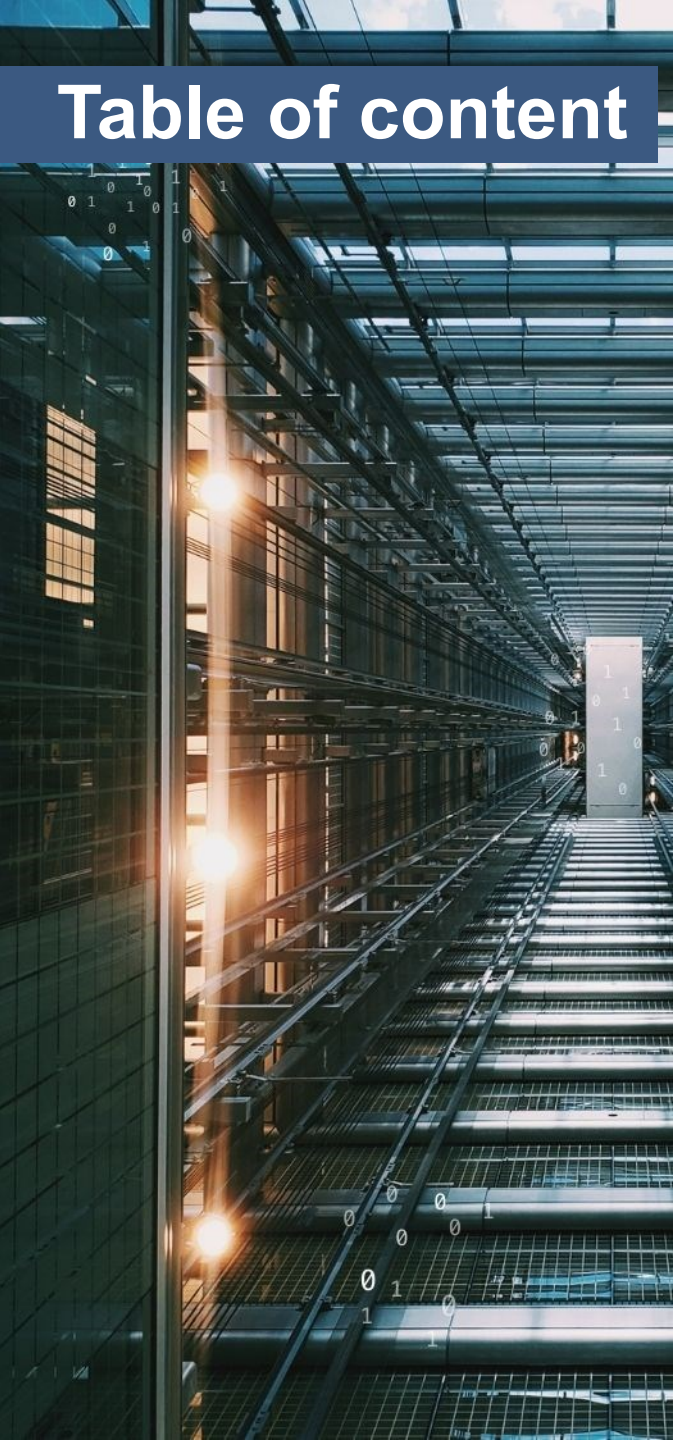


Table of content

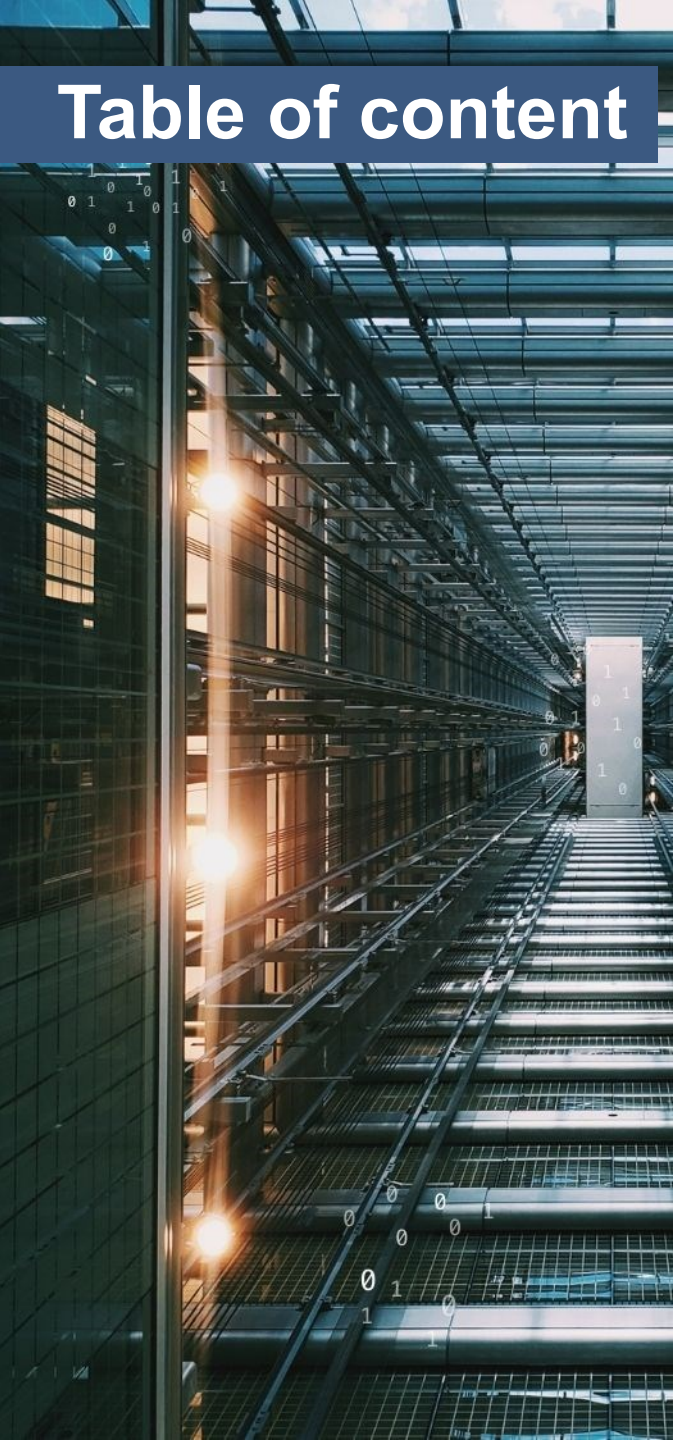
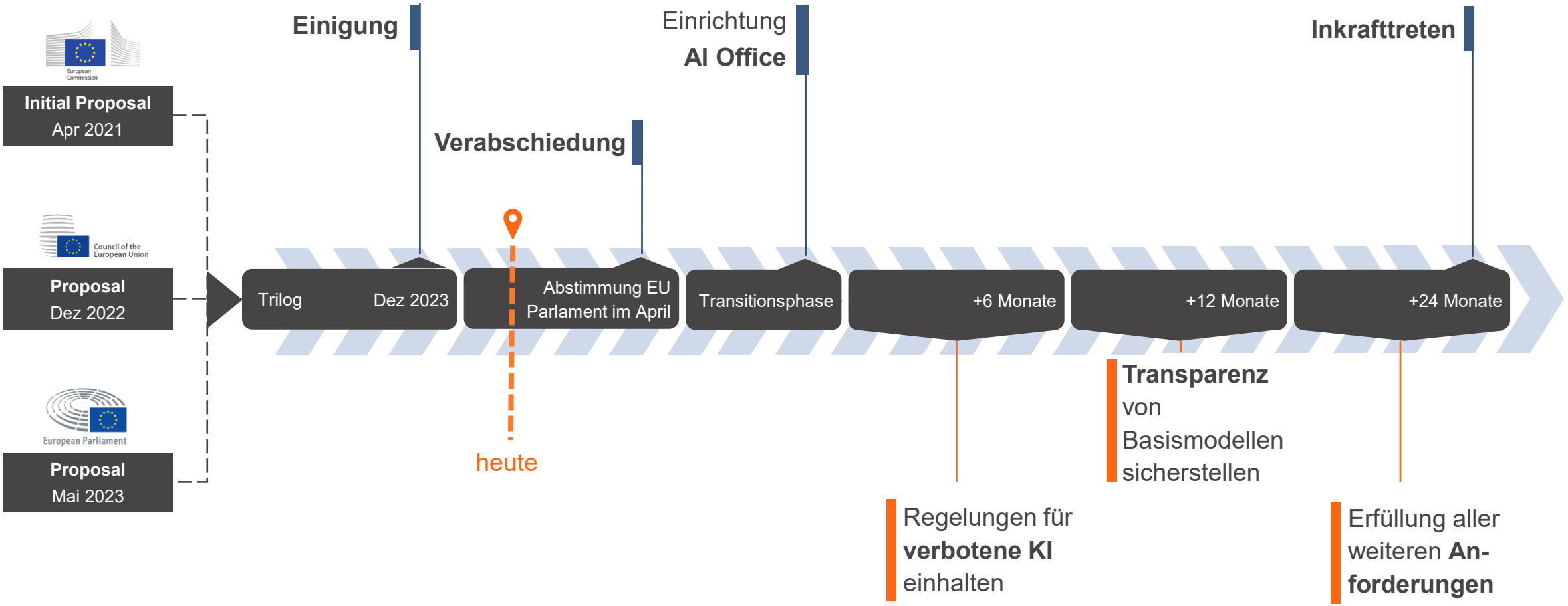


Table of content

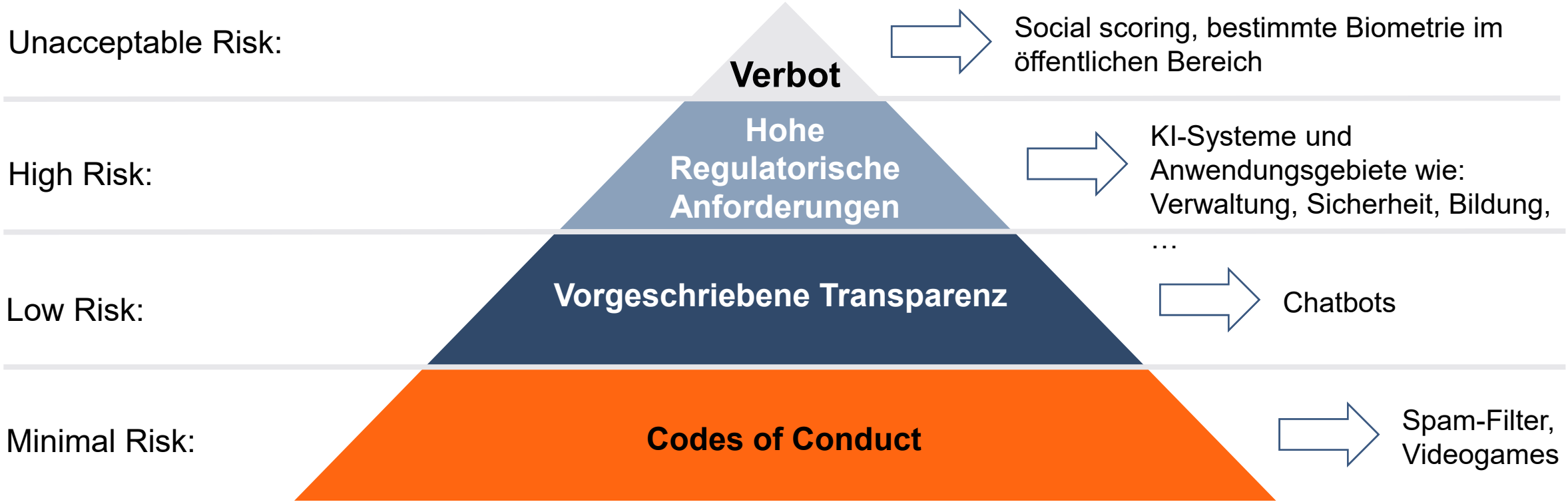
European AI-Act – Setup und Historie

Wir befinden uns bereits in der heißen Phase. Es ist unabdingbar, sich bereits heute mit den Inhalten des AI-Acts zu befassen.



European AI-Act: Risikoklassifizierung

Technologieagnostische Herangehensweise: Use Cases werden in Risikoklassen eingeordnet.



General Purpose AI (GPAI) und Foundation Modelle

Generative Basismodelle werden gesondert betrachtet.

GPAI

Basismodelle (Foundation models)

Anforderungen an **Transparenz**

- Technische Dokumentation
- Einhaltung des EU-Urheberrechts
- Detaillierte Zusammenfassungen über Trainingsdaten

Anbieter von GPAI müssen Informationen an nachgelagerte Betreiber/Einrichter weitergeben

High-Impact GPAI

Zusätzliche Verpflichtungen durch systemisches Risiko

- Modellevaluierungen
- Bewertung und Risikomitigation von systemischen Risiken
- Bericht an die Kommission über schwerwiegende Vorfälle
- Kontradiktorische Tests (Adversarial tests)
- Cybersecurity

Ausnahme: Open-Source-Modelle ohne Bezug zu hochriskanten/verbotenen Anwendungsfällen
→ Es muss der **Zugang**, die **Nutzung**, die **Änderung** und die **Verbreitung** des Modells und der Parameter möglich sein.

Welche Unternehmen betrifft der AI Act?

Anwendungsfälle für Hochrisiko KI.

Rechtspflege und demokratische Prozesse

- z. B. Ermittlung und Auslegung von Sachverhalten und Rechtsvorschriften

Leistungsbewertung in Bildung und Berufsausbildung

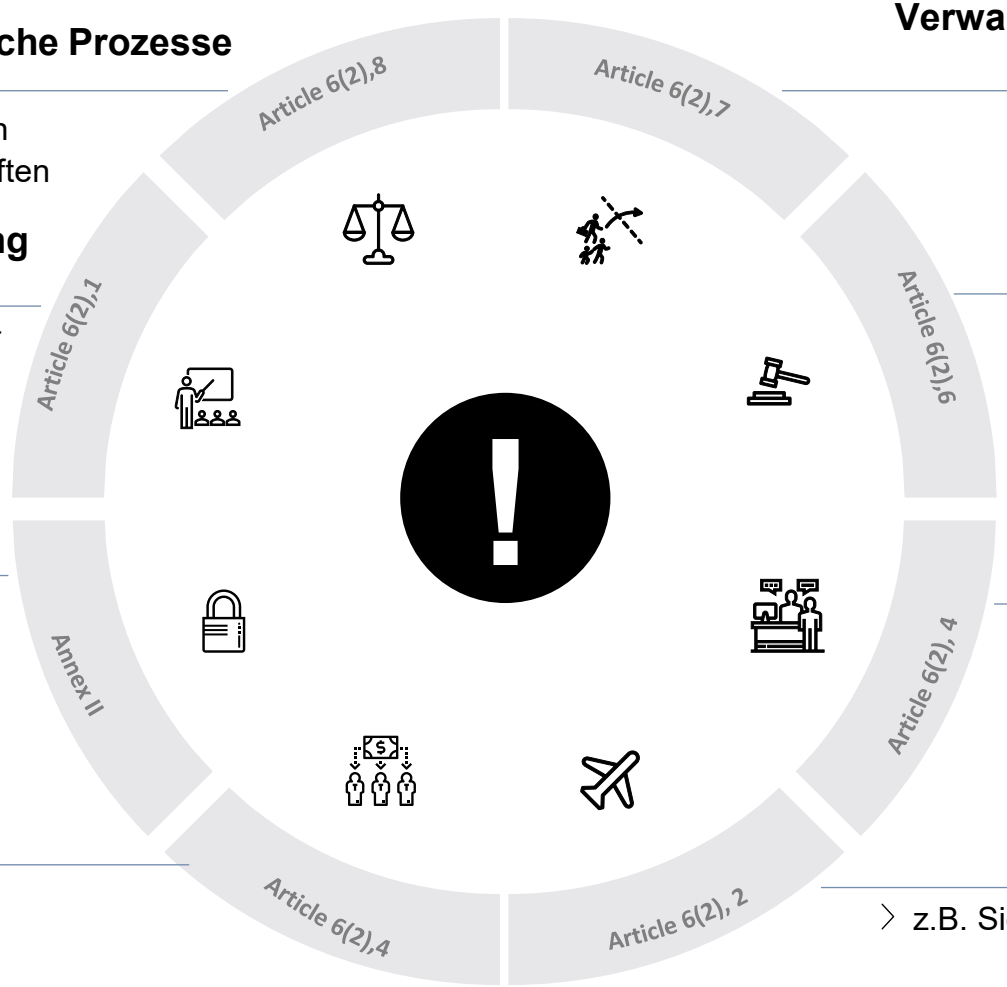
- z. B. Bewertung von Schülern über Kamerasysteme

Produkt-Sicherheits-komponenten

- z. B. Personendetektion durch autonome Logistikroboter

Beschäftigung, Personalmanagement

- Lebenslaufsortiersoftware für Einstellungsverfahren



Verwaltung von Migration, Asyl und Grenzkontrollen

- z. B. Überprüfung der Echtheit von Reisedokumenten

Strafverfolgung

- z. B. Bewertung der Zuverlässigkeit von Beweismitteln

Private und öffentliche Dienste und Leistungen

- Kreditwürdigkeitsprüfung und Kreditpunkte-bewertung natürlicher Personen

Kritische Infrastruktur

- z.B. Sicherheitskomponenten in der Verwaltung der Gasversorgung

[at]

Wen betrifft der AI-Act in einem Unternehmen?

Viele Mitarbeiter sind vom AI-Act betroffen, nicht nur „Techies“.

▪ Positive **wirtschaftliche Auswirkungen** von KI ▪ Compliance, Sicherheit und Vertrauen für **Kunden und Mitarbeiter**	**C-Level** 	**Compliance-Abteilung** 	▪ Sicherstellung der Einhaltung **geltender Gesetze** ▪ Einhaltung von **internen Vorschriften** ▪ **Verständnis** der technischen Besonderheiten von KI
▪ Verantwortlich für die **KI-Bemühungen** in der Organisation ▪ Effiziente KI-Projekte für **hohen ROI** ▪ Verwaltet und beaufsichtigt den **Governance-Prozess** und die **Dokumentation**	**AI-Lead** 	**Dev-Lead** 1010 1010	▪ KI in ein größeres System/Anwendung/Produkt einbetten ▪ **Nutzertransparenz**/Optionen für **Vorfallsberichte** bereitstellen
▪ **Entwickelt** ML-Systeme, hat Zugang zu Informationen, muss **technische Details und Tests** dokumentieren und berichten	**AI-Team** 	**Customer Success** ★★☆	▪ **Kommunikation** mit Kunden ▪ Gewährleistung von **sicherer und zuverlässiger KI** für Kunden ▪ Beantwortung und Weiterleitung von Unfallberichten/Anfragen

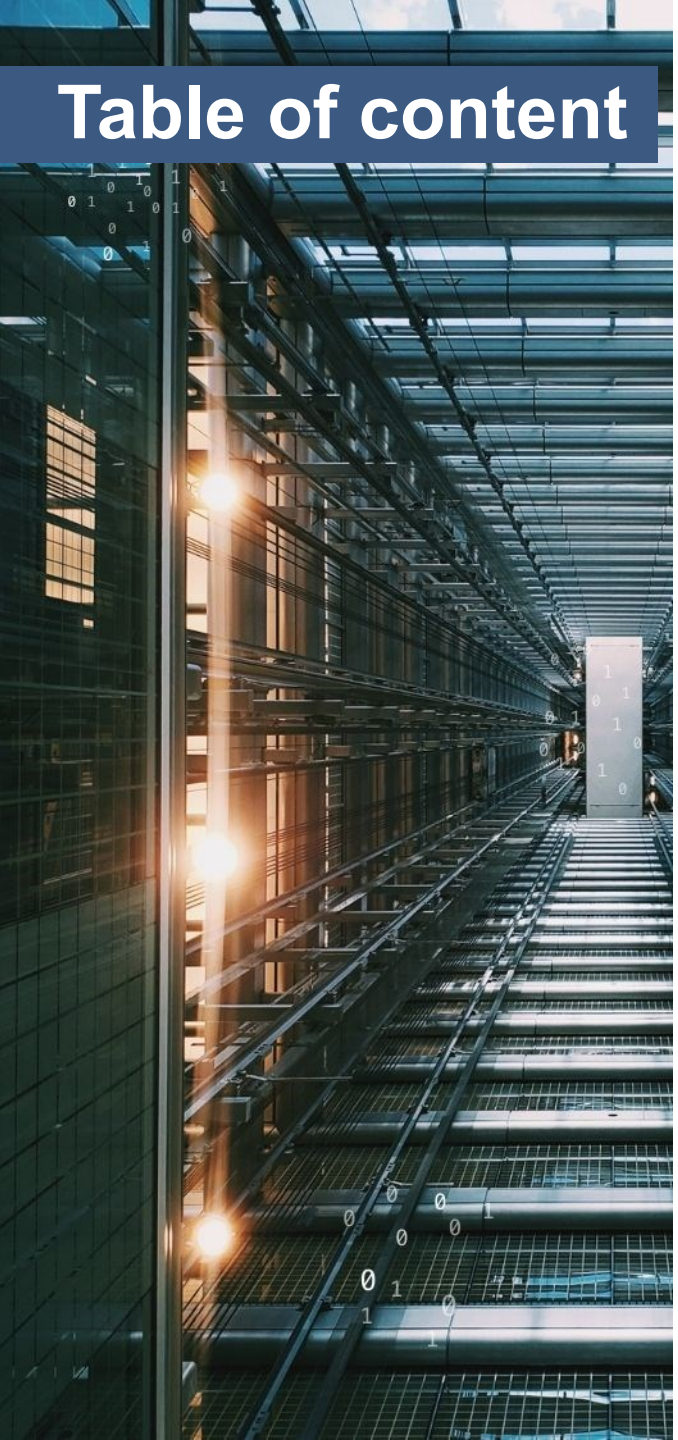


Table of content

Use Case basierter Ansatz als Basis für die Risikobetrachtung

Es kommt auf den Use Case an

Use Case / Anwendungsfall

Formalisierung, dessen, was man erreichen möchte.

Eine Beschreibung des zu lösenden **Problems**, der **Lösung**smöglichkeiten und des resultierenden **Mehrwertes**.

Risikobasiertes Prüfungsschema

Betrachtung des Risikos für den potentiellen Nutzer, das Umfeld oder weitere Betroffene

Lebenszyklus und Reifegrad

Betrachtung des gesamten Lebenszyklus, d.h. unterschiedliche Reifgrade des Use Case von der initialen Idee bis hin zum implementierten Produkt

Wie kann ein Use Case Portfolio Management Tool helfen?

Nachhaltiges Portfolio Management bildet die Grundlage für AI Governance

Zentrale Use Case Library / AI Register

Governance entlang des gesamten Lebenszyklus

Management der Verantwortlichkeiten

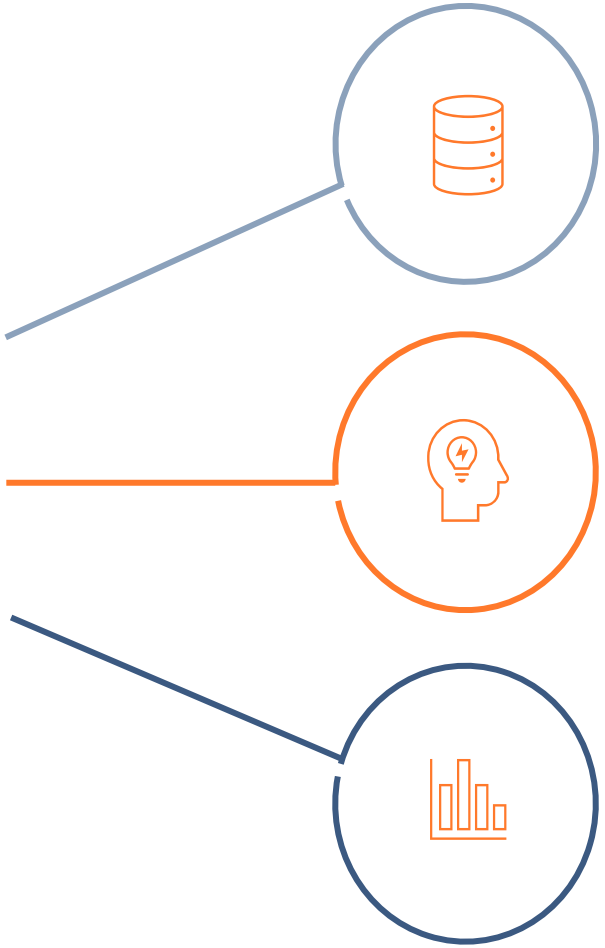
Umfassendes Risikomanagement

Auditfähiges Reporting und Nachvollziehbarkeit

Use Case Beispiel: KI im Recruiting

Effektives & effizientes Recruiting ist ein entscheidender Erfolgsfaktor für Firmen – KI kann hier unterstützen

*HR-Mitarbeiterin Hannah Recrui-Ting ist überfordert mit dem Ansturm an Bewerbungen, die eingehen, daher wünscht sie sich eine **KI**, die ihr hilft eine **Vorauswahl** zu treffen und eine **erste Einschätzung** gibt ob der Bewerber zu der Stelle und zur Firma passt, zudem wäre es perfekt wenn die KI die **erste Kommunikation** übernimmt, ein **Interview** durchführt und **Skills überprüft.***



Problem

- Informationsflut, viele Bewerbungsunterlagen
- selbst nach einer ersten Filterung noch zeitaufwändig eine Vorauswahl zu treffen

Lösungsansatz

- Mit Hilfe eines LLM könnte die Stellenbeschreibung, der Lebenslauf und das ein Bewerberformular analysiert werden, um eine Eignungsbewertung für den Bewerber zu erstellen.
- Wenn die Eignung über einem bestimmten Schwellenwert liegt, könnte dem Bewerber automatisch eine Einladung zu einem ersten Test & Interview geschickt werden.

Mehrwert

- Enormen Zeitersparnis für Hannah und die HR-Abteilung, schnelleres Feedback für Bewerber, besserer Fit, kein menschliches Bias bei der Beurteilung

Potenzielle Chancen & Risiken des Use Cases

KI im Recruiting

Chancen & Mehrwert

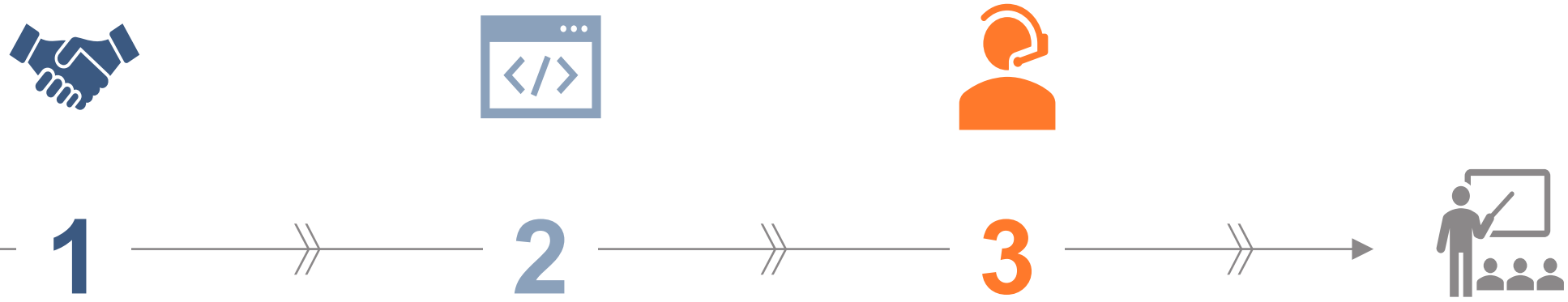
- **Effizienzsteigerung:**
  - Schnellerer Auswahlprozess durch Automatisierung von Prozessschritten
- **Effektivität:**
  - Bessere Fit von Anforderungen und Bewerbern durch matching von „must-have Anforderungen“
- **Fairness:**
  - menschliche Voreingenommenheit/Bias reduzieren
  - Förderung von Gleichberechtigung

Risiken

- **Transparenz und Erklärbarkeit**
  - Entscheidungsfindung der KI lässt sich nicht nachvollziehen (Input / Outputfaktoren unklar)
- **Fairness**
  - Bias / Diskriminierung in der Entscheidungsfindung
- **Funktionale Sicherheit & Data Privacy**
  - Privatsphäre und die Datenrechte sehr sensibel im Bewerbungsprozess
  - Erreichbarkeit und Robustness des zukünftigen Systems ist kritisch für den Erfolg

Was sind Hannahs nächste Schritte?

Aktives Use Case Management und Risikobewertung



1 Use Case Beschreibung

Standardisierte Use Case Definition über ein Use Case Management Tool

2 Identifizierung des Risikos

Klassifizieren Risikogruppe, um Anforderungen & Maßnahmen zu identifizieren

3 Umsetzung der Konformitäts-anforderungen

Schrittweise erfüllen aller relevanten Anforderungen und managen über die Use Case Management Plattform.

AI-Act konform agieren

Einhaltung der Vorschriften und überwachen der Produkte nach der Markteinführung.

Vorgehen zur Risikoklassifizierung

Das Assessment deckt insgesamt 4 Fragekategorien ab

Definition von KI	♦ Ist das System ein KI-System?
Markteilnahme	♦ Beeinflusst das KI-System den europäischen Markt? ♦ Was ist die Marktteilnehmerrolle in Bezug auf das KI-System?
Scope	♦ Fällt das KI-System unter den AI Act?
Klassifikation in Risikoklassen	♦ Welche Risikoklasse besteht?

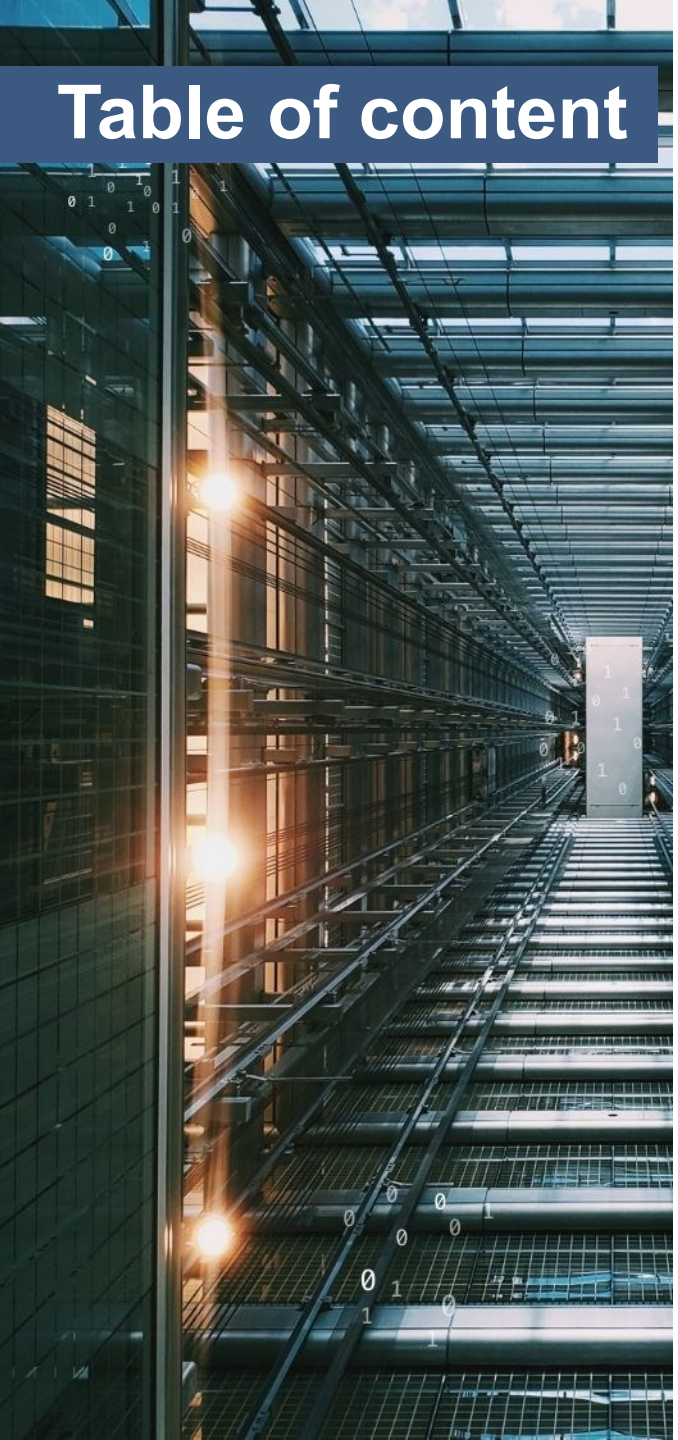


Table of content

Demo in Casebase

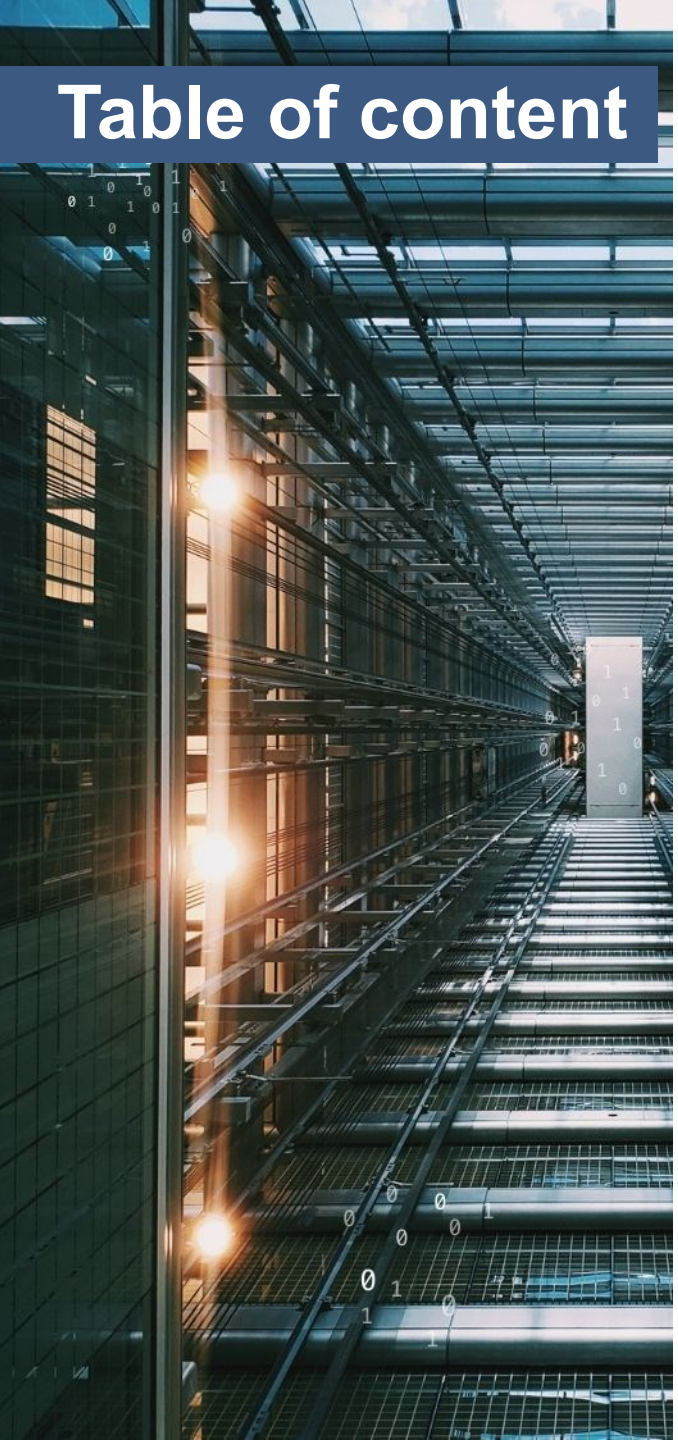


Table of content

Big Picture zur AI-Act Readiness

Die Vorbereitung für den AI-Act erfordert einen individuellen, abgestimmten Maßnahmenkatalog.

Gängige AI Governance Rahmenwerke

Es gibt unterschiedliche Rahmenwerke zum Planen und Gestalten der AI Governance.

NIST Artificial Intelligence Risk Management Framework

ISO/IEC 42001:2023

The **IEEE** Global Initiative on Ethics of Autonomous and Intelligent Systems

The **European Union's** Ethics Guidelines for Trustworthy AI

Anforderungen an Hochrisiko-Systeme

Acht Bereiche müssen bespielt werden, um High-Risk Use-Cases AI-Act-konform zu betreiben.

Risikomanagement-System
(Titel III, Kap. 2, Art. 9)

Data & Data Governance
(Titel III, Kap. 2, Art. 10)

Technische Dokumentation
(Titel III, Kap. 2, Art. 11)

Aufzeichnungspflichten
(Titel III, Kap. 2, Art. 12)

Transparenz und Bereitstellung von Informationen für die Nutzer
(Titel III, Kap. 2, Art. 13)

Menschliche Aufsicht
(Titel III, Kap. 2, Art. 14)

Genauigkeit, Robustheit und Cybersicherheit
(Titel III, Kap. 2, Art. 15)

Grundrechte Assessment
(Titel III, Kap 3, Art.29a)

Zusätzlich im neuen Gesetzestext

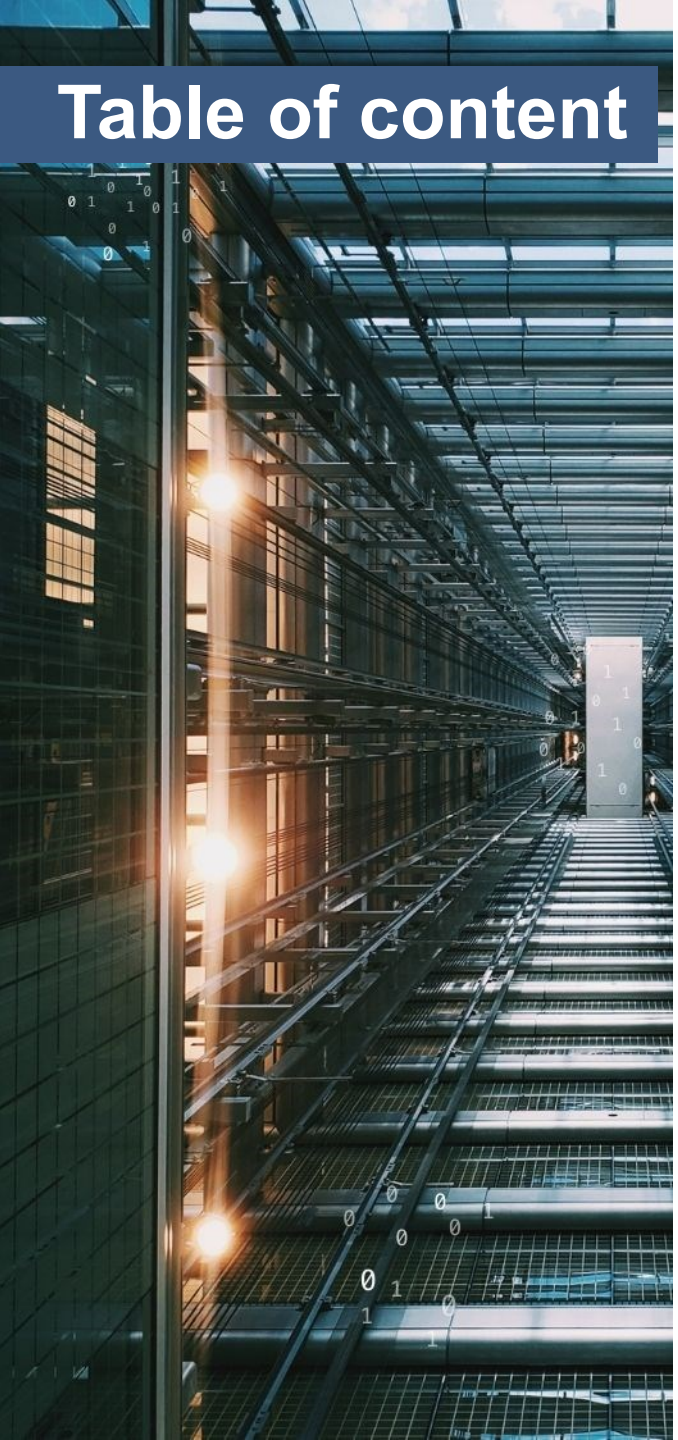


Table of content

Vorgehen zur AI-Act-Konformität

Klar definierte Schritte führen zum rechtskonformen Umgang mit AI.

1. AI-Act-Reifegrad und Verantwortlichkeiten

Schaffung bzw. Assessment der **AI Compliance-Organisation** mit **Verantwortlichkeiten**

2. Identifizierung, Dokumentation und Klassifizierung

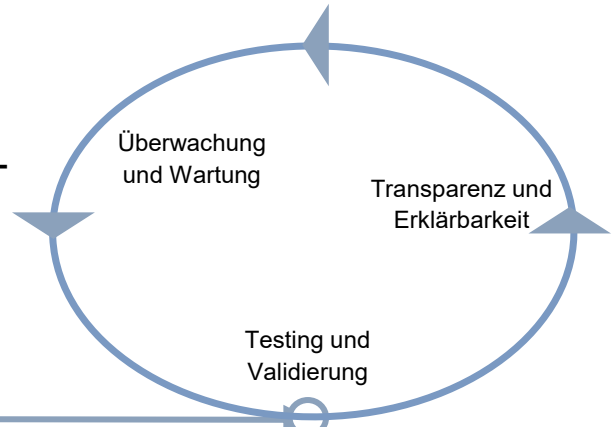
Einführung eines **AI-Use-Case-Registers/Management-Tools** und Durchführung der **Risikoeinstufung** nach den AI-Act-Kriterien

3. AI Governance und AI-Act konforme Systeme

Implementierung und **kontinuierliche Anpassung** der notwendigen **Prozesse und Systeme**

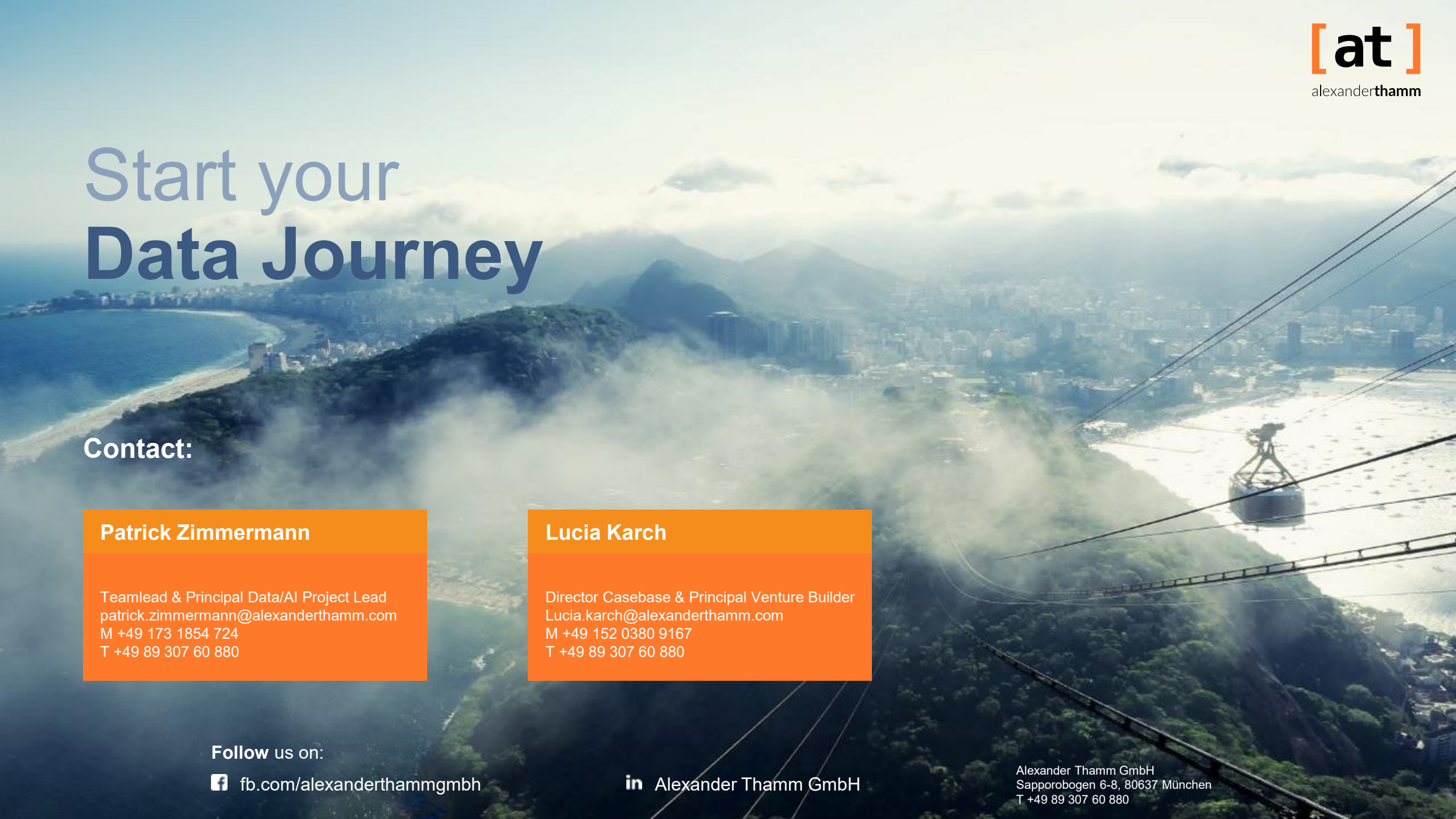
4. Konformitätserklärung und Zertifizierungen

Konformitätsbewertung für die AI-Systeme (**CE-Kennzeichnung)** und **Zusammenarbeit** mit den **Aufsichtsbehörden**



5. Monitoring des Betriebes

Automatisch erzeugte Protokolle über den Betrieb (**Bestandteile sind aktuell zu halten)**



[at]
alexanderthamm

Start your
Data Journey

Contact:

Patrick Zimmermann

Teamlead & Principal Data/AI Project Lead
patrick.zimmermann@alexanderthamm.com
M +49 173 1854 724
T +49 89 307 60 880

Lucia Karch

Director Casebase & Principal Venture Builder
Lucia.karch@alexanderthamm.com
M +49 152 0380 9167
T +49 89 307 60 880

Follow us on:

fb.com/alexanderthammgmbh

Alexander Thamm GmbH

Alexander Thamm GmbH
Sapporobogen 6-8, 80637 München
T +49 89 307 60 880

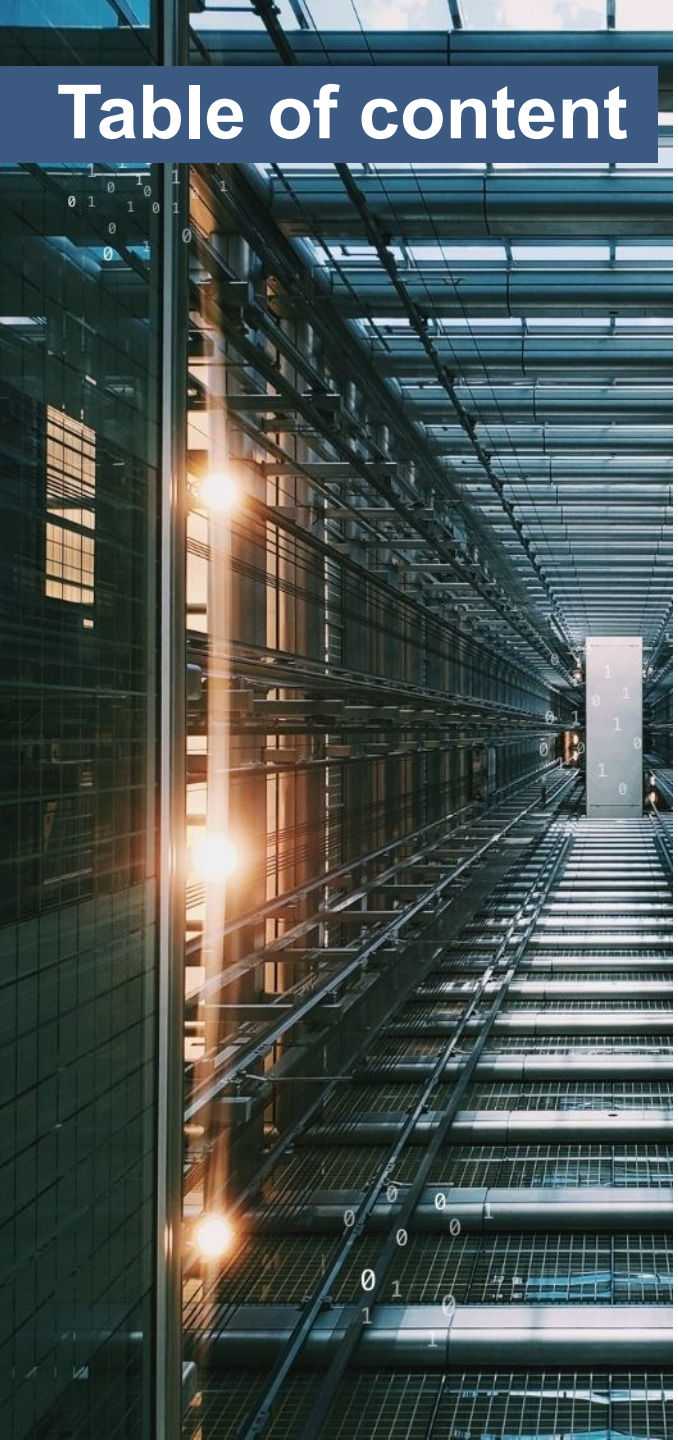


Table of content

Risikomanagement-System

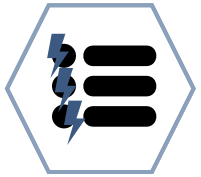
Ein Risikomanagementsystem muss eingerichtet, umgesetzt, dokumentiert und gepflegt werden.

Identifikation möglicher Risiken

Etablierung eines Identifikations-prozesses, durch den sichergestellt wird, dass die Risiken gemanaged werden

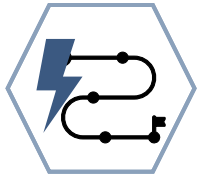
Evaluierung und Bewertung der Risiken

Bewertung der Risiken nach einheitlichen Standards uns Schaffung von Vorlagen



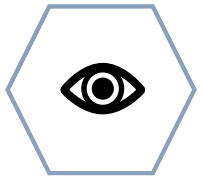
Kategorisierung / Priorisierung der Risken

Einheitliche Priorisierung nach Auswirkung und Eintritts-wahrscheinlichkeit



Maßnahmen zum Umgang mit den Risiken

Definition von konkreten Maßnahmen zur Risikomitigierung mit Zuordnung einer Verantwortlichkeit und einem Abschluss-Zeitpunkt



Monitoring

Nachhalten der Maßnahmen sowie kontinuierliche Verbesserung des Risikomanagement-prozesses

Data & Data Governance

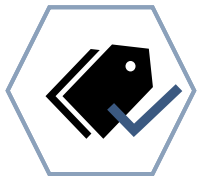
Trainings-, Validierungs- und Testdaten müssen definierten Qualitätskriterien entsprechen.

Datenqualitäts-sicherung

Implementierung von Prozessen zur Überwachung und Verbesserung der Datenqualität während des gesamten Lebenszyklus

Datenzugriffs-kontrolle

Etablieren von Zugriffskontrollen und Berechtigungen, um sicherzustellen, dass nur autorisierte Benutzer auf sensible Daten zugreifen können

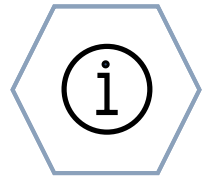


Konsistente Metadaten-verwaltung

Strukturierte Verwaltung von Metadaten, um Datenherkunft, -qualität und -verwendung zu dokumentieren

Datenvalidierung und Säuberung

Implementierung von Validierungsprozessen und Säuberungs-mechanismen für die Datenvorverarbeitung



Data Privacy und Compliance

Einhaltung aller Datenschutz-bestimmungen und Compliance-Anforderungen während der gesamten Datenverarbeitung

Technische Dokumentation

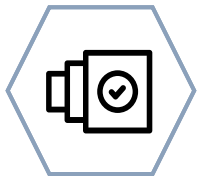
Eine technische Dokumentation ist zu erstellen bevor das System in Betrieb genommen wird und ist aktuell zu halten.

Umfassende Modell-Dokumentation

Erstellung detaillierter Dokumentationen für AI-Modelle, einschließlich Architektur, Trainingsdaten, Parameter und Evaluationsmetriken.

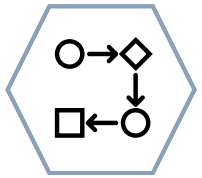
Transparente Code-Dokumentation

Dokumentation des Codes mit klaren Erklärungen, Kommentaren und Erklärungen zur Interpretation.



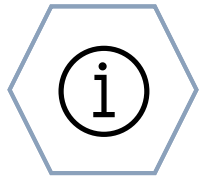
Versionierung und Tracking

Etablieren von Versionskontroll-mechanismen und Verfolgung von Änderungen im Code und Modell.



Dokumentation von Abhängigkeiten

Aufzeichnung und Dokumentation aller externen Abhängigkeiten, Bibliotheken und Infrastruktur.



Erklärbarkeit und Interpretierbarkeit

Integration von Techniken zur Erklärbarkeit und Interpretierbarkeit von Modellentscheidungen.

Aufzeichnungspflichten / Record Keeping

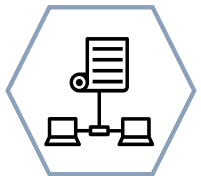
Automatische Aufzeichnung von Ereignissen ("Logs") während des Betriebs nach anerkannten Normen und Standards.

Logging und Audit-Trail

Implementierung von Logging-Funktionen und Audit-Trails für Interaktionen und Entscheidungen des Modells

Langzeit-speicherung von Ergebnissen

Speicherung von Vorhersagen und Ergebnissen über einen definierten Zeitraum für Rückverfolgbarkeit und Analyse

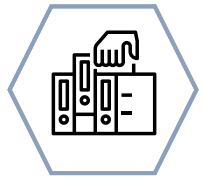


Verlauf und Änderungs-protokolle

Aufzeichnung und Speicherung von Modellverläufen und Änderungen für Rückverfolgbarkeit und Nachvollziehbarkeit

Compliance-orientierte Datenspeicherung

Einhaltung gesetzlicher Anforderungen für die Speicherung und Verarbeitung von Daten



Berichterstattung und Dokumentation

Erstellung und Archivierung von Berichten und Dokumentationen über den Betrieb und die Ergebnisse der AI-Systeme.

Transparenz und Informationsbereitstellung für die Nutzer

Nutzer müssen die Ergebnisse des Systems interpretieren und angemessen nutzen können.

Erklärung der Verarbeitung

Bereitstellung klarer Erklärungen über die Datenverarbeitung an Endnutzer und Art/Typ des AI-Modells

Benutzerzugang zu Informationen

Bereitstellung von Informationen über ihre eigenen Daten und Modellinteraktionen.

Transparenz über Vorhersagen

Bereitstellung von Informationen über Vorhersagen und deren Genauigkeit für Endbenutzer.

Verständliche Anwenderdokumentation

Erstellung von benutzerfreundlichen Dokumentationen und Anleitungen zur Nutzung der AI-Systeme

Feedback-Mechanismen

Implementierung von Systemen zur Erfassung und Berücksichtigung von Benutzerfeedback zur Verbesserung der Systemtransparenz.

Menschliche Aufsicht

Während des Zeitraums der Nutzung muss das System wirksam von natürlichen Personen kontrolliert werden können.

Monitoring und Management-schnittstellen

Bereitstellung von Benutzerschnittstellen für die Überwachung und Steuerung von AI-Systemen durch menschliche Aufsicht. Regelmäßige Evaluierung:

Regelmäßige Evaluierung

Etablieren von regelmäßigen Überprüfungen und Evaluierungen der AI-Systeme durch menschliche Aufsicht

Notfallpläne und Intervention

Entwicklung von Notfallplänen und Möglichkeiten für menschliche Intervention bei unvorhergesehenen Ereignissen

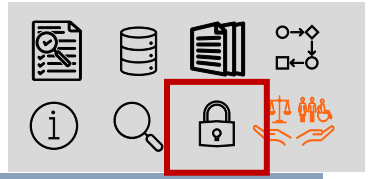
Fachwissen und Schulungen

Bereitstellung von Schulungen und Fachwissen für Techniker und Domänenexperten, um eine angemessene Beurteilung und Reaktion sicherzustellen.

Klare Verantwortlich-keiten

Festlegung klarer Verantwortlichkeiten für menschliche Aufsichtspersonen im Betrieb und der Überwachung der AI-Systeme.

Genauigkeit, Robustheit und Cybersicherheit



Ein angemessenes Maß an Genauigkeit, Robustheit und Cybersicherheit während des gesamten Lebenszyklus ist notwendig.

Robustheits-prüfungen

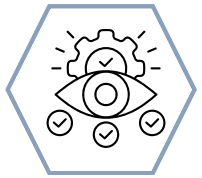
Durchführung umfassender Robustheitsprüfungen, um die Leistung des Modells unter verschiedenen Bedingungen zu testen.

Cybersecurity-Maßnahmen

Implementierung von Sicherheitsmaßnahmen, um die AI-Systeme vor externen Angriffen und Datenverletzungen zu schützen

Verifikation und Validierung

Durchführung von regelmäßigen Verifizierungs- und Validierungsprozessen, um die Genauigkeit und Zuverlässigkeit der Modelle zu überprüfen.



Kontinuierliche Modellüberwachung

Implementierung von Überwachungsmechanismen, um die Modellleistung kontinuierlich zu überwachen und Anomalien zu erkennen.

Sicherstellung von Datenintegrität

Maßnahmen zur Sicherstellung der Integrität von Daten und Verarbeitungsprozessen, um die Genauigkeit der Ergebnisse zu gewährleisten

Umsetzung in der Praxis

AI-Act konformer Betrieb dank der Umsetzung entlang des gesamten ML Lifecycles