Eine Prüfplattform für vertrauenswürdige KI - Anforderungen und Umsetzung

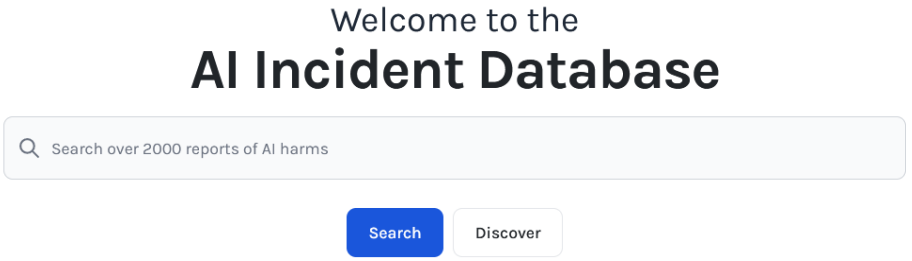
Maximilian Pintz | KI-Absicherung & Zertifizierung | 29.11.23

Sind KI-Anwendungen vertrauenswürdig?

KI-Anwendungen & ihre Vertrauenswürdigkeit

- KI-Software findet immer breiteren Einsatz, auch in **sicherheitskritischen Systemen**
- Fehler in KI-Anwendungen haben bereits Konsequenzen (siehe zB. „AI Incident Database“)
- Beispiel: Kindergeld-Affäre in den Niederlanden
  - KI-Anwendung zum Erkennen von betrügerischen Kindergeld-Anträgen
  - Durch Fehlerkennungen wurde Familien zu Unrecht das Geld verweigert & zur Rückzahlung aufgefordert

➢ **Es bestehen gesellschaftliche & unternehmerische Interessen KI-Anwe ıdungen vertrauenswürdig zu gestalten**



incidentdatabase.ai

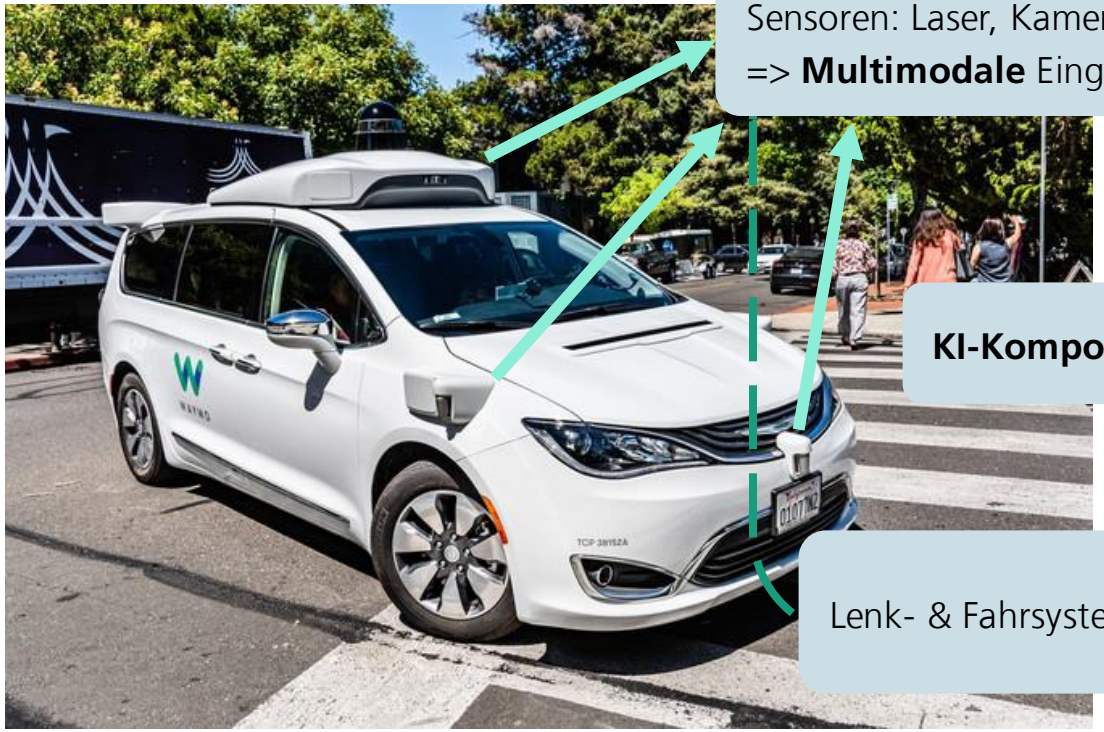

The Dutch Tax Authority Was Felled by AI—What Comes Next? ›European regulation hopes to rein in ill-behaving algorithms

BY RAHUL RAO | 09 MAY 2022 | 4 MIN READ

Aus: IEEE Spectrum Magazine, spectrum-ieee-org.cdn.ampproject.org

Prüfung von KI-Anwendungen

Prüfungen von KI-Anwendungen sind komplex



Sensoren: Laser, Kameras, Radar
=> **Multimodale** Eingaben

KI-Komponenten

Lenk- & Fahrsystem

Daten-Vorverarbeitungsschritte

Nachverarbeitung

Safety- & Kontrollmechanismen

Tests

Prüfwerkzeuge

Zertifizierung

Dokumentation

Wie können diese Komplexitäten beherrschbar gemacht werden?

Software Qualitäts- und Risikomanagement

Software Qualitäts- & Risikomanagement …

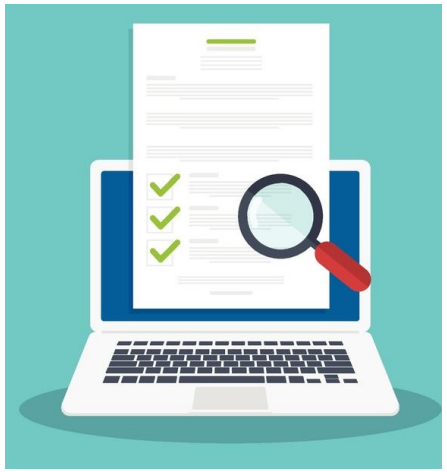
- … wirkt den Gefahren von KI-Software entgegen
- … wird für KI-Software bereits in **Gesetzesentwürfen** gefordert: EU AI Act
- … ist für klassische Software bereits etabliert und **standardisiert**

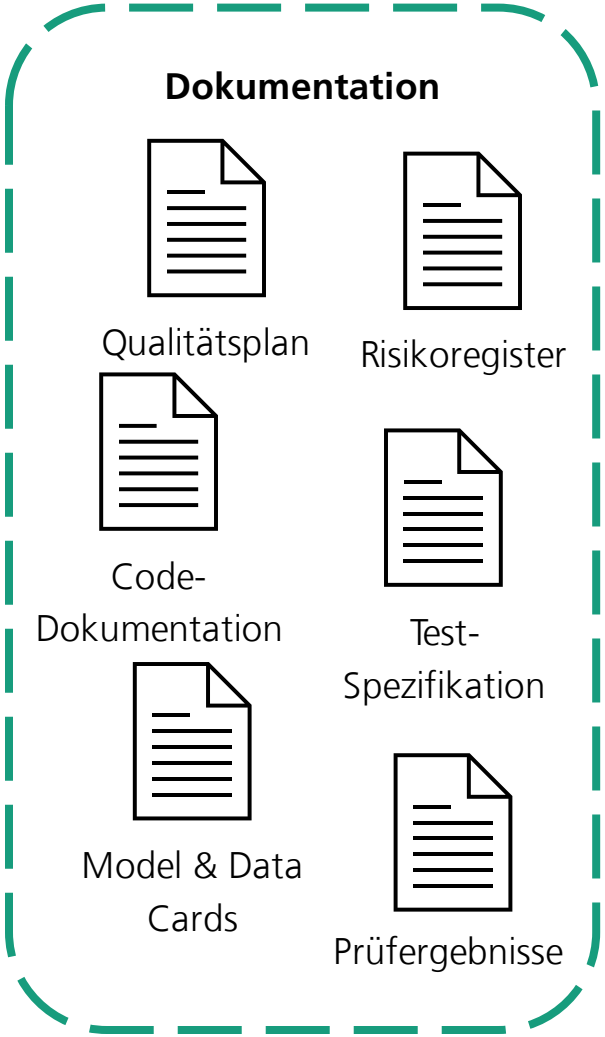
Ziel:

Bestehende Verfahren auf KI-Anwendungen übertragen und neue Entwickeln

Typische Prozesse:

- Analyse von Risiken und Qualitätskriterien, Treffen von Maßnahmen
- Entwicklungsprozesse (z.B. Agile/Testgetriebene Entwicklung)
- Etablieren einer **einheitlichen Dokumentation** der Qualitätsprozesse
- **Software Testing** und Review (insb. Continuous Testing)
- …

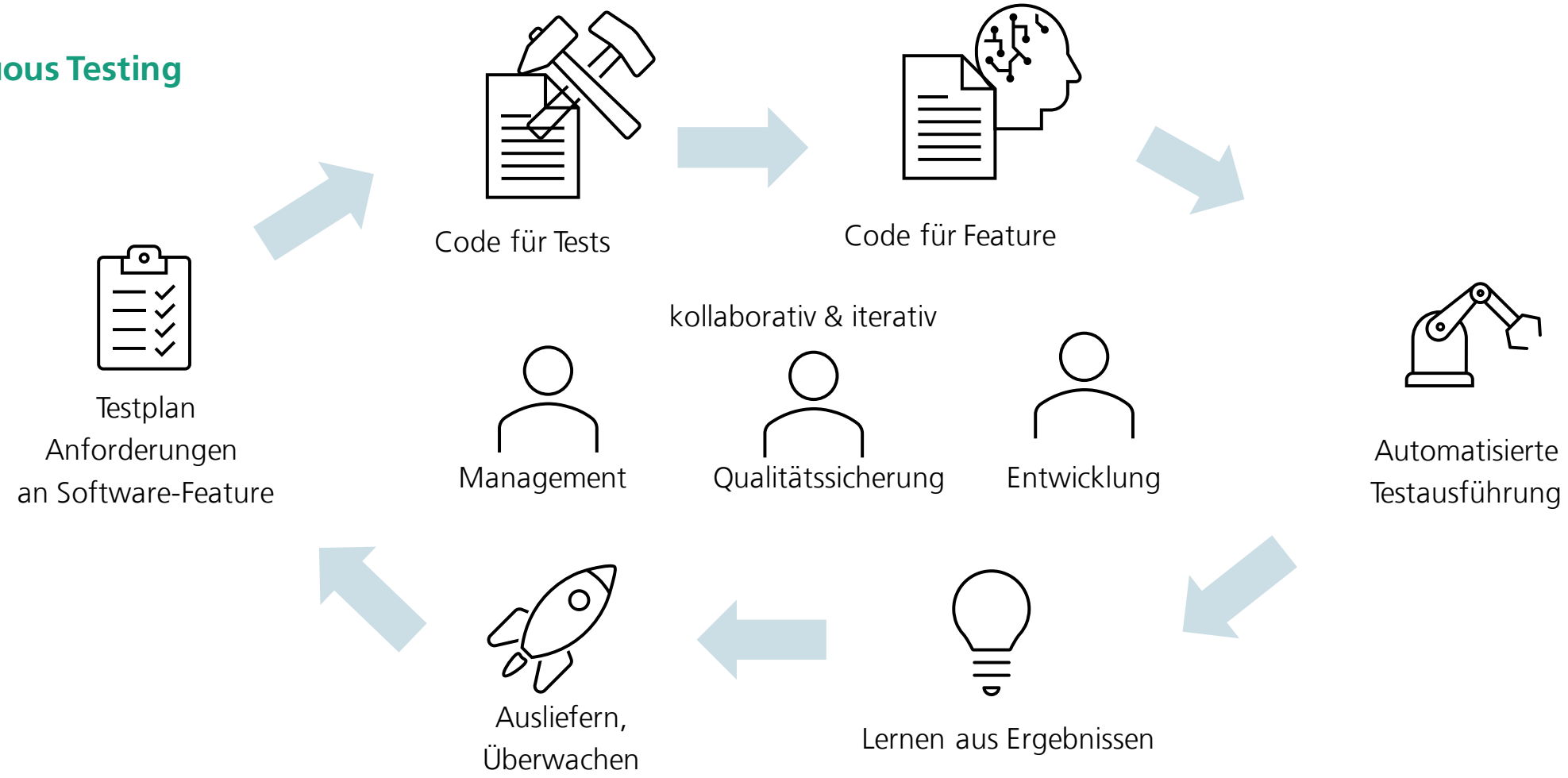




Software Qualitäts- & Risikomanagement

Continuous Testing – ein typischer Prozess des SQRM

Continuous Testing



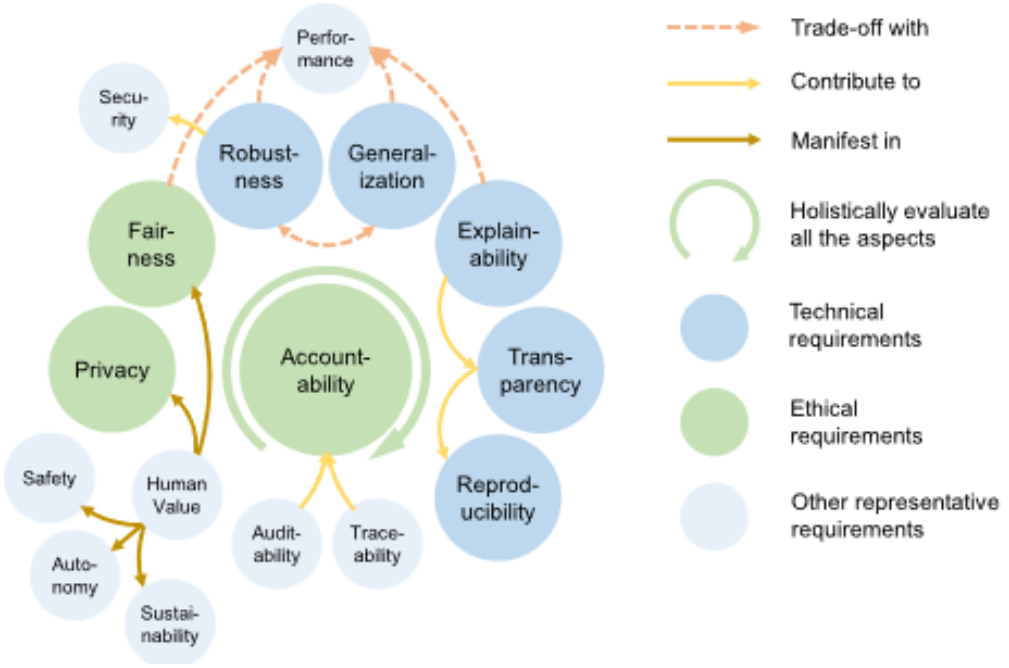
Prüfen und Testen von KI

Qualitätskriterien & Maßnahmen für KI-Anwendungen

Vertrauenswürdige KI

- Entwickelt **Qualitätskritierien & -Maßnahmen** für KI-Anwendung
- Verschiedene **Prüfdimensionen** mit komplexen Abhängigkeiten
- Verschiedene **Prüfverfahren** pro Dimension & Use Case
- Schnelle & stetige Weiterentwicklung und Verbesserung bestehender Verfahren
- Viele Start-ups & Unternehmen bieten eigene **Prüftools** zu Aspekten der vertrauenswürdigen KI an

Leitfaden zur Gestaltung vertrauenswürdiger KI, Fraunhofer IAIS



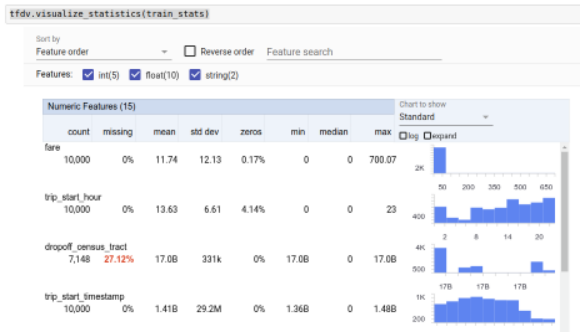
Li et. Al., Trustworthy AI From Principles to Practises, ACM Comput. Surveys 2023

Prüfen und Testen von KI

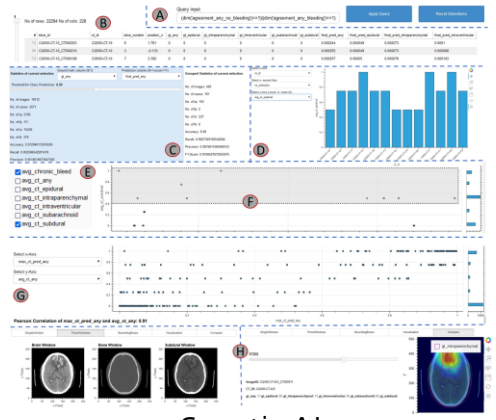
Prüftools in der vertrauenswürdigen KI

Prüftools

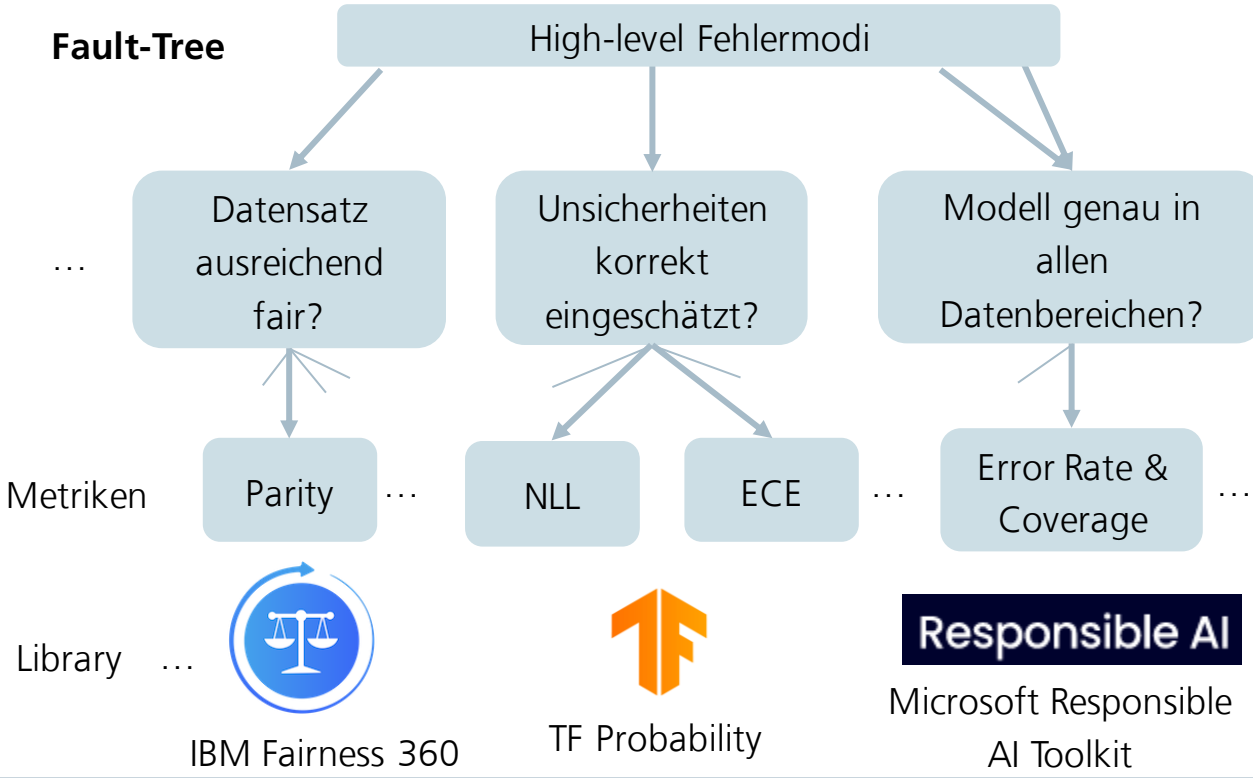
- Üblicherweise **Software-Libraries**, die eine Menge von Funktionen zu **spezifischen Teilaspekten** einer KI-Prüfung bereitstellen
  - Programmierschnittstellen für den Entwickler
  - Experten-Visualisierungen
- **Zusammenschalten** solcher Werkzeuge notwendig, aber oft **schwierig**
  - Tool-spezifische Schnittstellen, **keine Einheitlichkeit** zwischen verschiedenen Werkzeugen
  - Einbinden von Eingabedaten & KI-Modellen auf verschiedene Arten (unterschiedliche Formate, ML-Frameworks, Cloud-spezifisch …)



Tensorflow Data Validation



ScrutinAI



IBM Fairness 360

TF Probability

Microsoft Responsible AI Toolkit

Prüfen und Testen von KI

Herausforderungen in der vertrauenswürdigen KI

Herausforderungen in der vertrauenswürdigen KI

- Je nach Use Case sind die Konzepte, Prüfverfahren & Prüftools noch **lückenhaft oder unausgereift**
- Zusammenführung von verschiedenartigen Datensätzen, Modellen & Tools ist **technisch komplex**
- Prüftools allein helfen typischerweise wenig bezüglich Reproduzierbarkeit, Nachvollziehbarkeit, Dokumentation …

Wie kann man diese Herausforderungen adressieren?

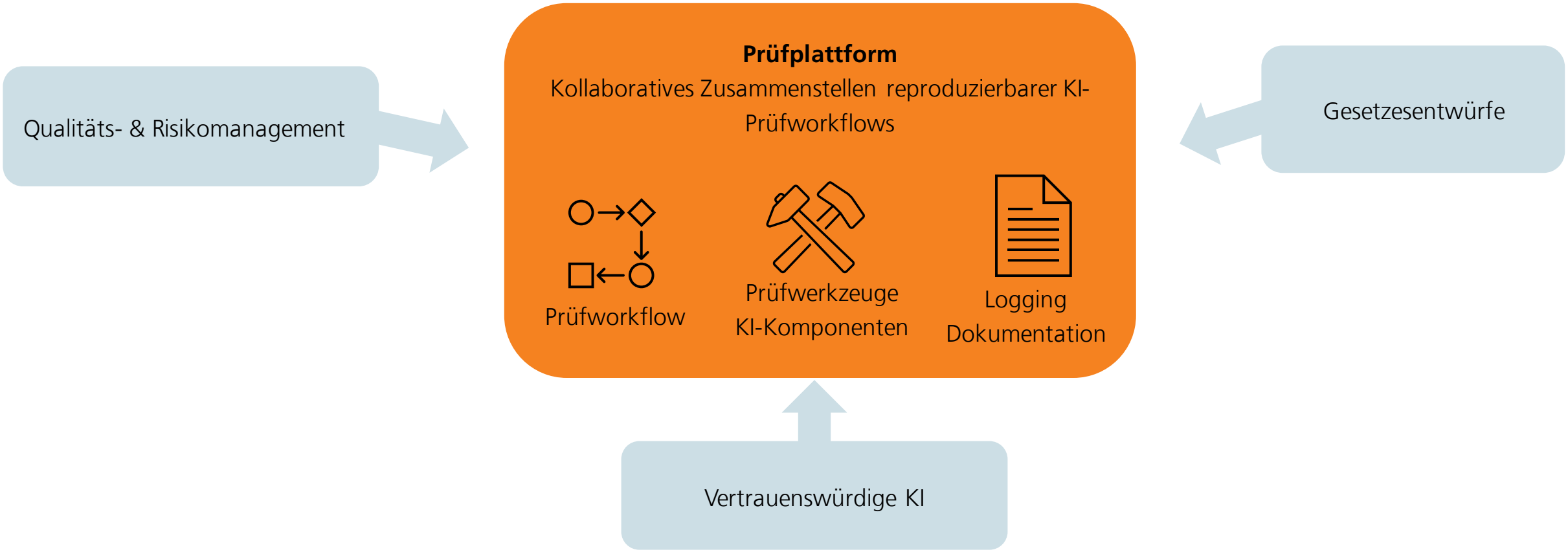
- Einerseits: Forschung & Weiterentwicklung von Konzepten, Prüfverfahren & Tools
- Andererseits: (Weiter-)entwicklung von **Prüfplattformen** für vertrauenswürdige KI
  - **Technischer Rahmen** für die **Integration von Prüftools** zur Sicherstellung von Robustness, Fairness, Verlässlichkeit, …
  - Tragen aktiv zu den Dimensionen **Reproduzierbarkeit & Accountability** bei

Prüfplattformen sind essentiell für die Entwicklung & Prüfung vertrauenswürdiger KI-Anwendungen

Eine Plattform für Prüfwerkzeuge

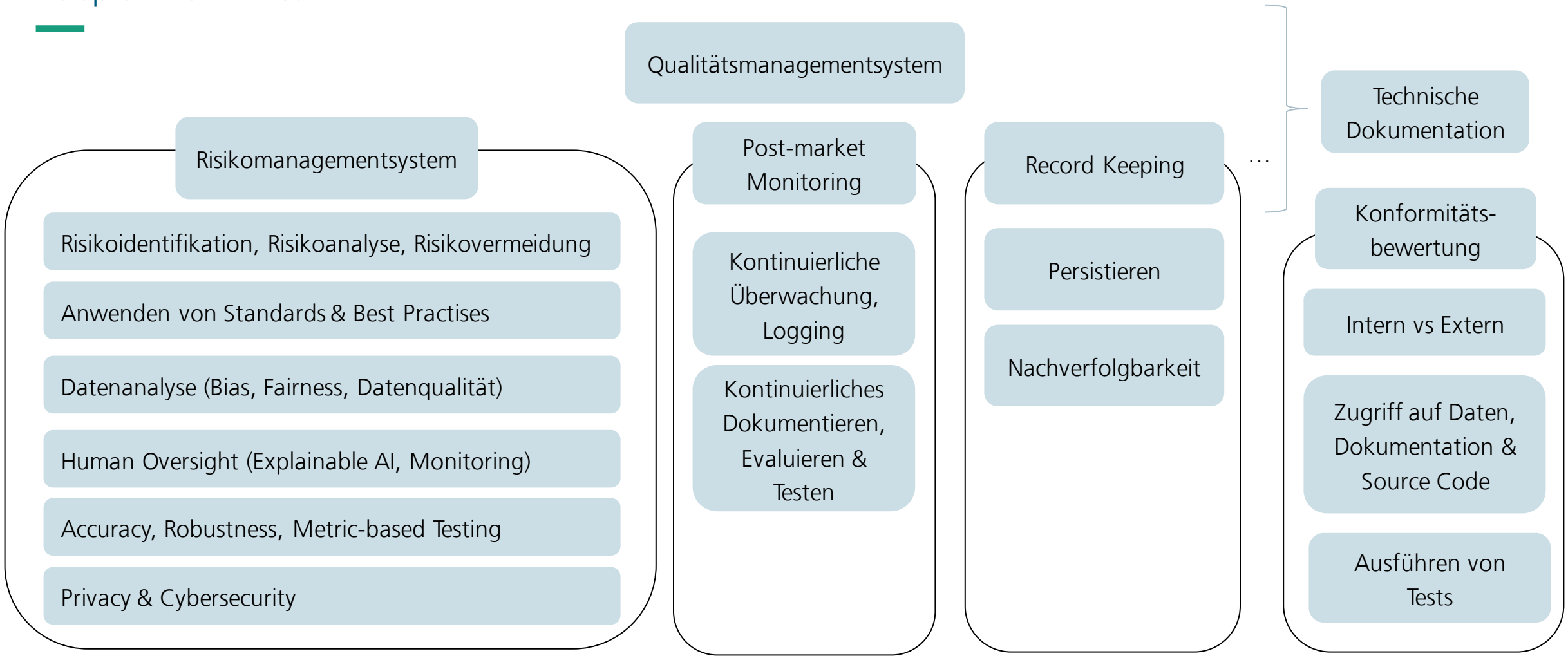
Ziel

Eine Plattform für das Qualitäts- & Risikomanagement von vertrauenswürdiger KI



Gesetzesanforderungen

Beispiel: EU AI Act



Eine Plattform für Prüfwerkzeuge

Eine große Bandbreite von technischen Anforderungen

Technische Anforderungen an eine Prüfplattform

- **Erstellen** von Daten & **Dokumentation** & **Austausch** zwischen verschiedenen Parteien
- **Software-Testing / Continuous Testing**
  - **Automatisierung** von Build-Prozessen, **Reproduzierbarkeit** von Tests & Prüfungen
  - Aufteilung der Software in einzelne **Komponenten** / unterschiedliche Abstraktionsebenen
  - Anwendbarkeit über den **gesamten Lebenszyklus**
- **Prüfungen durch Externe**
  - Abruf von Dokumentation / Ausführung von Prüfworkflows => **Remote-Ausführbarkeit**
  - **Verifizierbarkeit von Ergebnissen**
- Unterstützung & Einbezug von diversen Datensätzen, Modellen & KI-Prüfwerkzeugen => **Integration auf einer Plattform**
- **Logging** => Automatisches Erstellen von Prüf-Reports
- Live-Monitoring
- …

Qualitäts- & Risikomanagement

Gesetzesentwürfe

Vertrauenswürdige KI

Eine Plattform für Prüfwerkzeuge

Bestehende technische Systeme

Datenmanagement

Software-Management

Deployment-Systeme

KI-Spezifische Tracking- & Analysesoftware

...

Bestehende Software-Systeme können allein diese Anforderungen nicht alle erfüllen. Es ist notwendig verschiedene Systeme im Rahmen einer Architektur zusammenzuschalten und zu erweitern.

Eine Plattform für Prüfwerkzeuge

Architektur einer Plattform

Komponenten der Plattform

Artefaktmanagement

Artefakt: Abstraktion für **Daten**, insb. Dokumente, Tabellen, Bilder, Source Code, …

Modulmanagement

Modul: Abstraktion für ausführbare Software-Komponenten, insb. **KI-Modelle & Prüftools**

Prüfmanagement

Management von **Prüfworkflows**

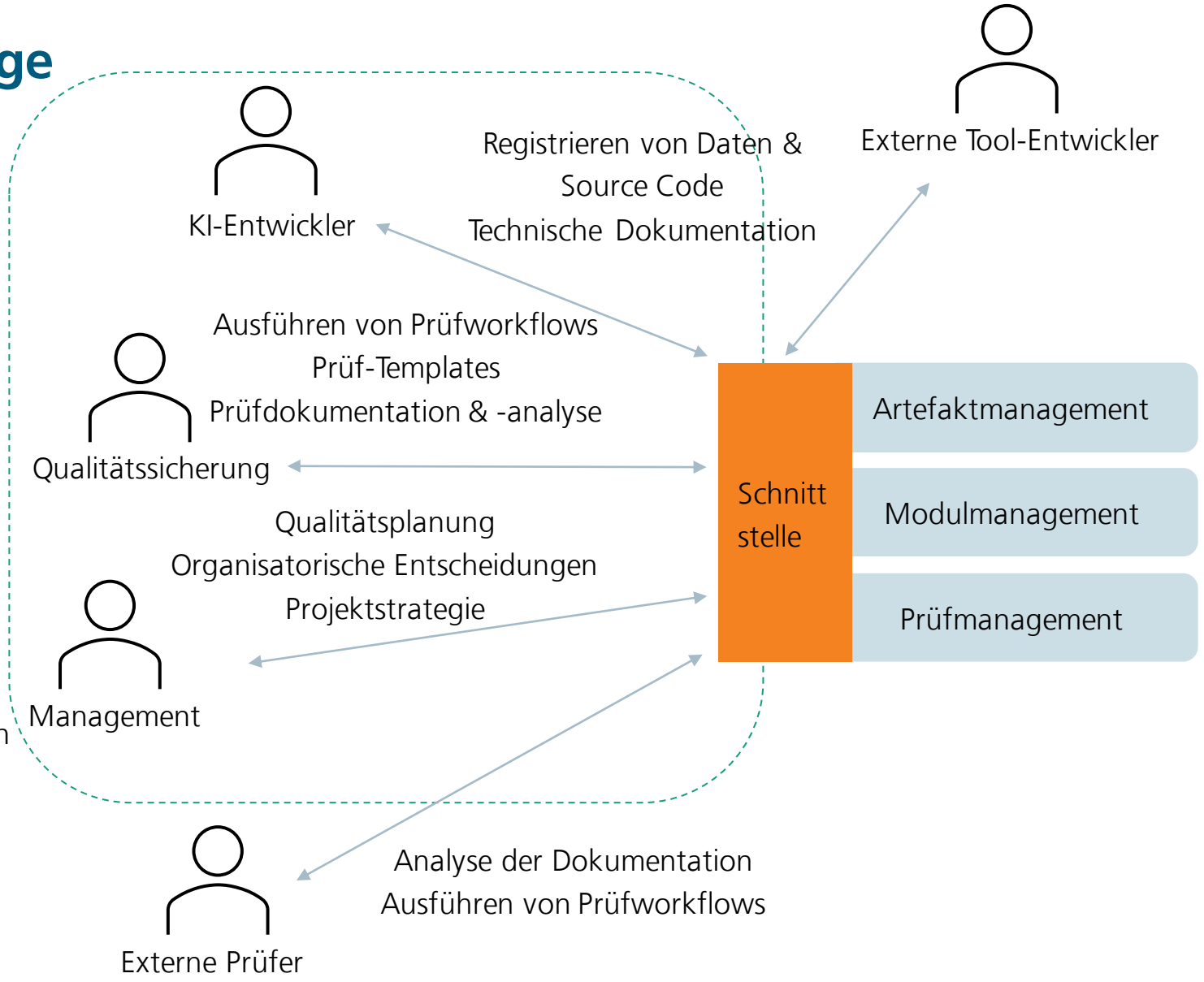
Kollaboratives Qualitäts- & Risikomanagement

Austausch von Dokumentation & Prüftemplates
Analyse von Prüfergebnissen

Eine Plattform für Prüfwerkzeuge

Kollaboratives Qualitäts- & Risikomanagement

- Erzeugen einer einheitlichen & umfassenden **Dokumentation**
  - Dokumentations-Templates
  - Planungsdokumente
  - Dokumentierte Daten & Source Code
  - Prüf-Workflow Spezifikationen
  - Prüf-Reports
- Austausch von **Prüf-Templates** & **Datentypen**
- **Reproduzierbares** Ausführen von **Prüfworkflows** (auch durch **Externe**) mit **verifizierbaren Ergebnissen**
- **Logging** & Live-Monitoring



Zusammenfassung

Anforderungen & Umsetzung einer KI-Prüfplattform

- Qualitäts- & Risikomanagement für KI-Anwendungen ist notwendig, aber komplex
- Eine Prüf-Plattform hilft dabei, diese Komplexitäten beherrschbar zu machen
- Es gibt eine große Bandbreite von Anforderungen an eine solche Plattform seitens SQRM, Vertrauenswürdige KI & Gesetzesentwürfen
- Bestehende Software-Systeme können allein alle diese Anforderungen nicht alle erfüllen. Es ist notwendig verschiedene Systeme im Rahmen einer Architektur zusammenzuschalten (und teilweise zu erweitern)
- Eine Architektur aus Artefaktmanagement, Modulmanagement & Prüf-Management kann diese Anforderungen erfüllen
- Weitere Details in **Whitepaper „KI-Anwendungen systematisch prüfen und absichern“**

Ausblick & Nächste Schritte

Nächste Schritte zur KI-Prüfplattform

- Wir sehen den **Aufbau und das Entwickeln** einer KI-Prüfplattform als **Community-Effort**
- **Zusammenarbeit mit Projektpartnern**
- Wissenschaftliche Publikation zur **Architektur einer Prüfplattform** im Veröffentlichungsprozess
- Demonstrator zur Prüfplattform
- **DIN Spec** zum Thema Prüfwerkzeuge
- **Workshops** & Community-Building