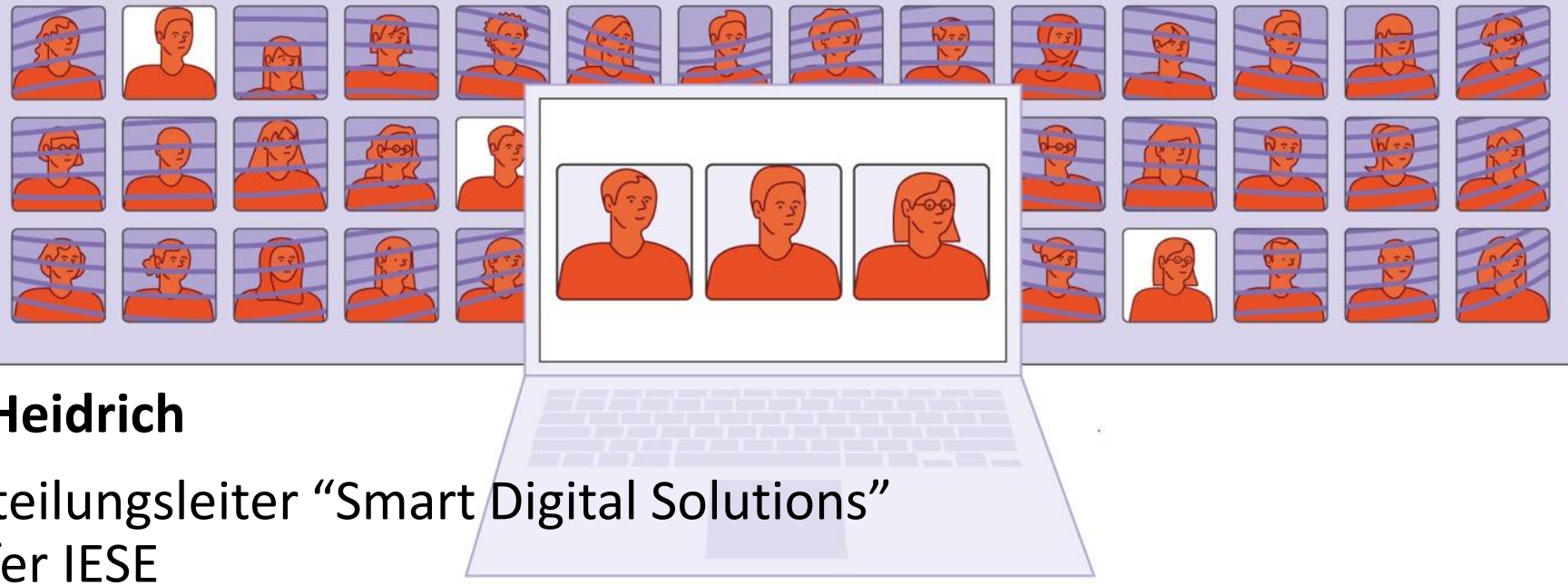
Wie Assurance Cases KI in der Industrie sicherer machen können

Verlässliche KI



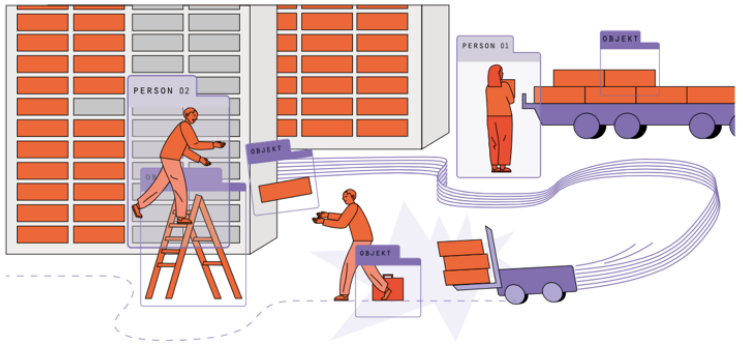
Dr. Jens Heidrich

Hauptabteilungsleiter “Smart Digital Solutions”
Fraunhofer IESE

Jens.Heidrich@iese.fraunhofer.de

ExamAI

Anwendungen und Gefahren von KI in der Produktionsautomatisierung



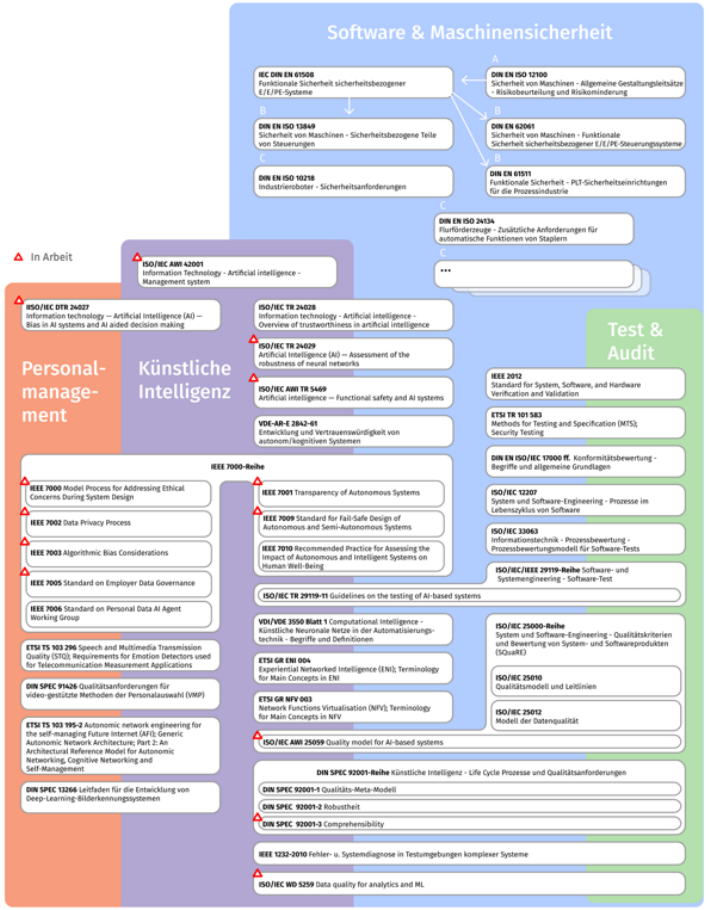
- Nutzung von KI in der Produktionsautomatisierung:
  - Steigerung der Produktivität
  - Ressourcenschonung
  - Qualitätsverbesserungen
- Fünf Anwendungsbereiche:
  1. Prozessplanung und -automatisierung
  2. Wartungs- und Designvorschläge
  3. Wissensbereitstellung
  4. Produkt-Qualitätssicherung
  5. **Maschinelle Unterstützung**

- Arten von Schädigungen bei Fehlverhalten der KI
  - **Personenschäden**
  - Finanzielle Schäden
  - Diskriminierung
- Fokus auf **funktionale Sicherheit = Safety**
- Beispiel-Anwendungsszenarien
  - **Cobots:** Kollaborierende Roboter
  - **Fahrerlose Transportfahrzeuge** (FTFs)

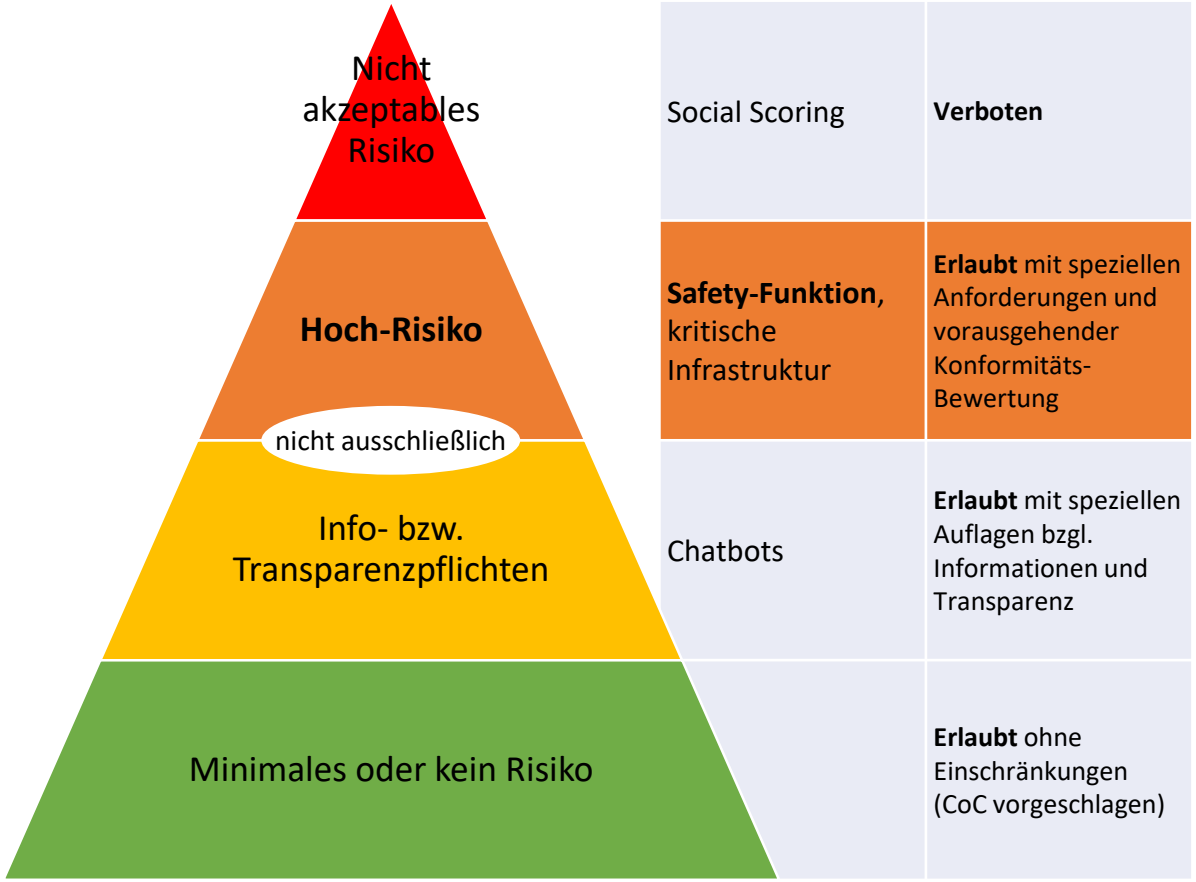
Gefördert durch:
Im Rahmen des:

Verordnungen und Normen

- Mit Maschinenrichtlinie harmonisierte Normen zur **Maschinensicherheit** (wie ISO 12100) und **Safety** (wie ISO 13849) adressieren KI-Systeme nur unzureichend
- KI-Komponenten (auf Basis von z.B. ML) funktionieren fundamental anders als klassische Software und erfordern angepasste Verfahren und Normen
- Zurzeit werden an **EU AI Act** und **Maschinenverordnung** (Nachfolge der Maschinenrichtlinie) gearbeitet
- Der direkte Nachweis der **Compliance** zu den Verordnungen ist extrem zeit- und kostenaufwendig für Unternehmen
- Ziel von Standardisierungsorganisationen (wie DIN, DKE, CEN, CENELEC, ETSI) ist es, **harmonisierte Normen** abzuleiten, um den **Nachweis der Compliance** zu vereinfachen



Einordnung von KI-Systemen und Anforderungen im EU AI Act



Beispiele	Verordnung
Social Scoring	**Verboten**
Safety-Funktion, kritische Infrastruktur	**Erlaubt** mit speziellen Anforderungen und vorausgehender Konformitäts-Bewertung
Chatbots	**Erlaubt** mit speziellen Auflagen bzgl. Informationen und Transparenz
	Erlaubt ohne Einschränkungen (CoC vorgeschlagen)

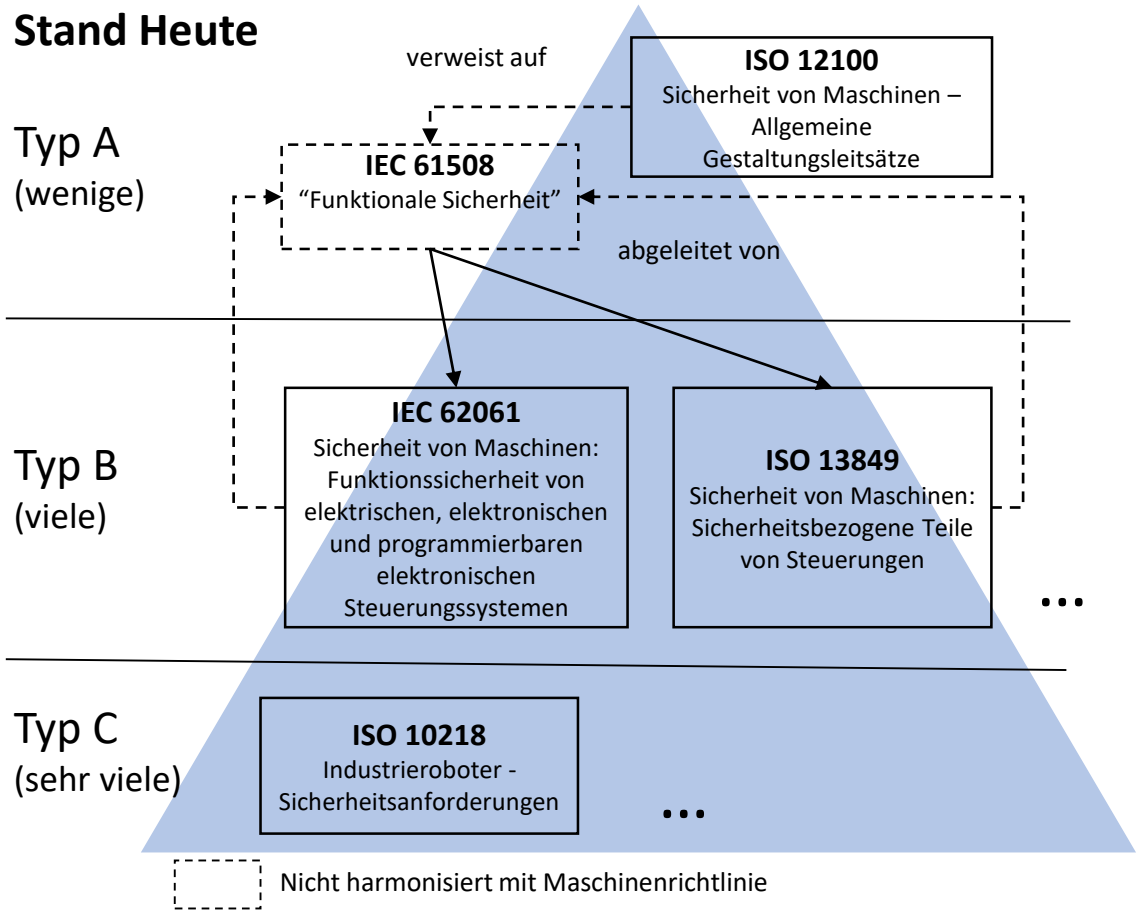
Anforderungen an Hoch-Risiko-KI-Systeme

- Umsetzung eines Risiko-Management-Systems
- Datenqualität (Relevanz, Repräsentativität)
- Technische Dokumentation und Logging
- Transparenz über Fähigkeiten und Einschränkungen
- Menschliche Aufsicht
- Robustheit, Genauigkeit und Cyber-Sicherheit

Gefördert durch:
Im Rahmen des:

Wie bringen wir KI in die Safety-Normen?

- Aktuelle Diskussion, wie **abstrakt bzw. konkret** die Vorgabe in Normen sein soll
- Problem: Aktuell gibt es **keinen Konsens**, wie man mit KI in den Safety-Normen umgehen sollte und welche Methoden anzuwenden wären (wird erforscht)
- Lösungsraum: Erarbeitung konkreter **Checklisten** vs. **zielgerichtete Argumentationsmethode**
- Checklisten erst dann sinnvoll definierbar,, wenn Untersuchungsgegenstand verstanden und Methoden etabliert wurden
- In aktuellen Forschungsprojekten erweisen sich **Assurance Cases** als Mittel der Wahl

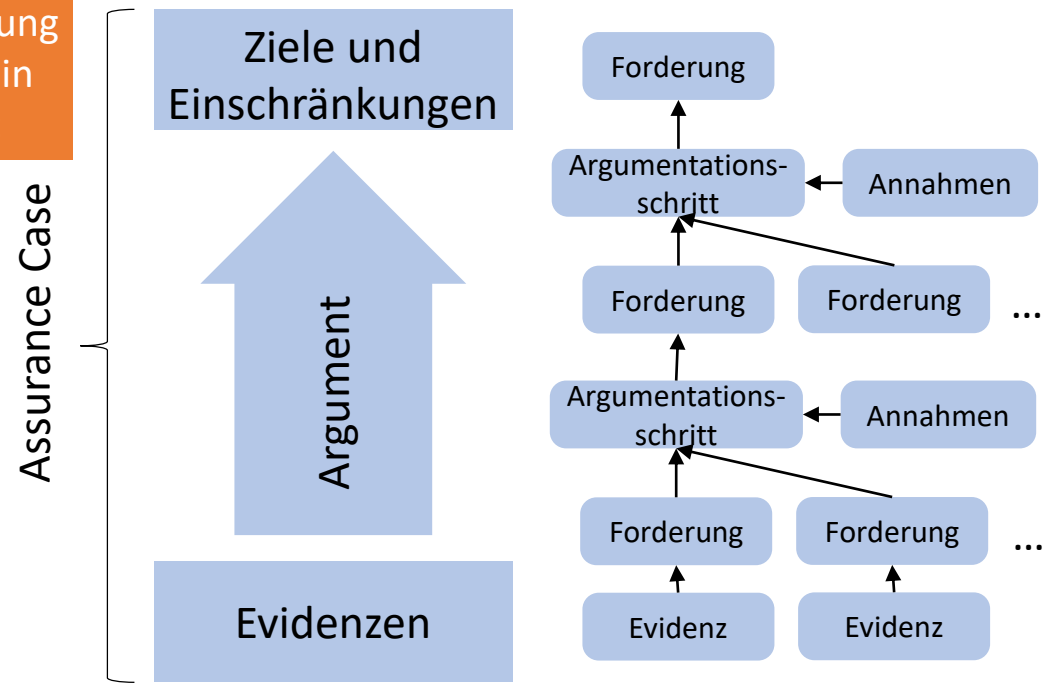


Gefördert durch:
Im Rahmen des:

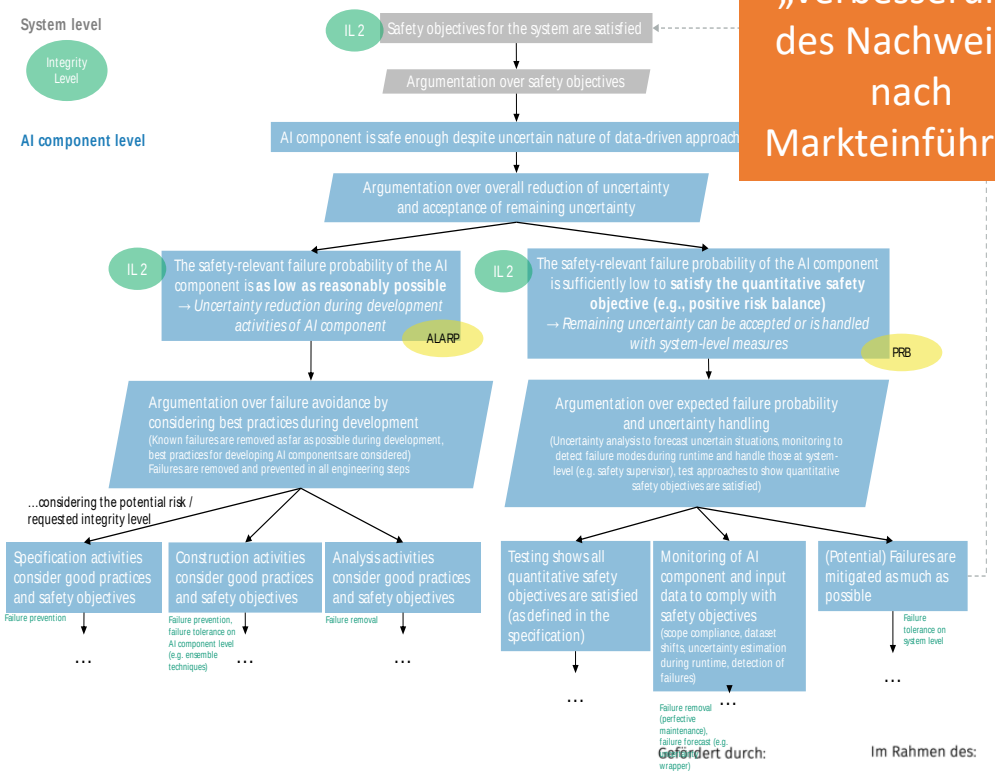
Assurance Cases als Basis des Nachweises

Ein **Assurance Case** ist eine strukturierte Argumentationskette mit zugehörigen Evidenzen, welcher die Annahme erlaubt, dass ein Produkt in einer Nutzungsumgebung die gesetzten Ziele erfüllt (sicher ist).

Erstellung und Prüfung durch ein **Audit**



„Verbesserung“ des Nachweises nach Markteinführung



Gefördert durch:

Im Rahmen des:

Fazit

- **Assurance Cases** eignen sich als zentrales Element für Auditing und Zertifizierung von KI in Bezug auf Safety
  - Geeignet, wenn es kein „Schema F“ gibt, um Safety nachzuweisen
  - Konkrete Checklisten können daraus abgeleitet werden
  - Kernelement neuer Safety-Normen für autonome Systeme mit KI (UL 4600, VDE-AR-E 2842-61)
  - Assurance Cases könnten eine modulare Zertifizierung unterstützen
- Offene Punkte
  - „**Bauanleitung**“ und „**Prüfanleitung**“ für ACs muss in Normen definiert werden
  - ACs in der Normierung aktuell noch **zu generisch**, um einfach umsetzbar zu sein
  - Konkrete **Methoden und Werkzeuge** für Sicherheitsnachweis benötigt
  - **Breitere Evaluierung** des Ansatzes in der Praxis nötig

Gefördert durch:
Im Rahmen des: