



KI

ABSICHERUNG

Safe AI for Automated Driving

23.06.2021, Zertifizierte KI

Workshop zur Prüfung von KI-Systemen

Ansätze aus KI-Absicherung Sicherheitsargumentationen für KI-Systeme

Dr. Maram Akila, Dr. Tim Wirtz, Fraunhofer IAIS





KI

ABSICHERUNG

Safe AI for Automated Driving

***KI Absicherung macht die Sicherheit KI-basierter
Funktionsmodule für das hochautomatisierte Fahren
nachweisbar.***



KI WISSEN Entwicklung von Methoden für die Einbindung von Wissen in maschinelles Lernen

KI DELTA LEARNING
Methoden und Werkzeuge zur Erweiterung und Transformation vorhandener KI-Module autonomer Fahrzeuge auf neue Domänen und komplexe Szenarien



KI ABSICHERUNG Methoden und Maßnahmen zur Absicherung von KI-basierten Wahrnehmungsfunktionen für das automatisierte Fahren

KI DATA TOOLING Methoden und Werkzeuge für das Generieren und Veredeln von Trainings-, Validierungs- und Absicherungsdaten für KI-Funktionen autonomer Fahrzeuge

KI Absicherung - Safe AI for Automated Driving



Konsortialleitung: **Volkswagen AG**

Stellv. Konsortialleitung:
Wiss. Koordination: **Fraunhofer IAIS**

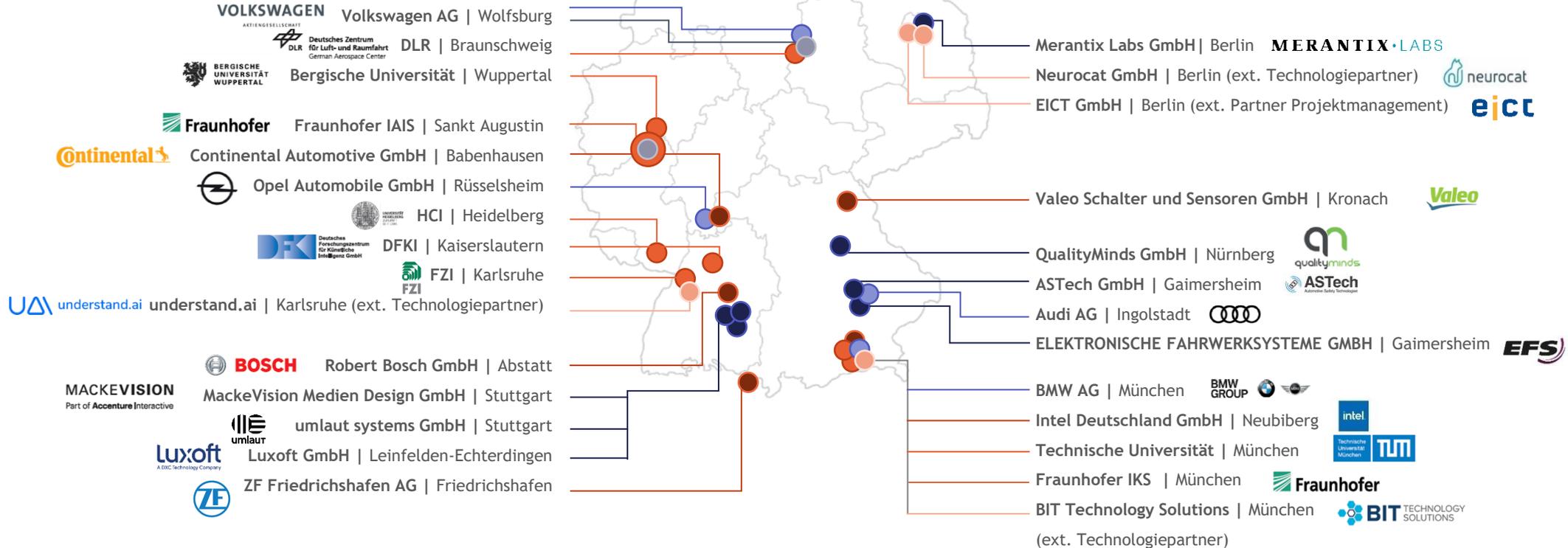
Budget: **41 Mio. €**

Förderung: **19,2 Mio. €**

Laufzeit: **36 Monate**

24 Partner

01.07.2019 - 20.06.2022



● Konsortialleitung ● OEMs ● Zulieferer ● Technologieprovider ● Forschung ● Externe Partner

Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages

Zentrale Ziele in KI Absicherung



1. Trainings- und Testmethoden für KI-basierte Funktionen

KI Absicherung entwickelt und untersucht Methoden und Maßnahmen für die Absicherung KI-basierter Funktionen für das hochautomatisierte Fahren.

2. Absicherungsargumentation

Am Use Case Fußgängererkennung erarbeitet das Projekt eine beispielgebende Argumentations- und Prozesskette zur Absicherung einer komplexen KI-Funktion.

3. Kommunikation mit Standardisierungsgremien zur KI-Zertifizierung

Für die Entwicklung eines Industriekonsenses zur Absicherung von KI-Funktionsmodulen werden die Projektergebnisse in den Dialog mit Standardisierungsgremien eingebracht.



Grundlegende Gedanken zur Motivation

Annahmen zur Fußgängerdetektion:

- KI-System mit Genauigkeit („Accuracy“) von 99,9% → 1 von 1000 Vorhersagen falsch
- Werte jedes 1000. Versagen als „kritisch“ → 1 von 1M Vorhersagen kritisch falsch
- „Kontinuierlicher“ Einsatz, Werte Eingaben alle 20 Sekunden als „neu“
- Verwendung der KI-Anwendung in 100.000 Fahrzeugen à 10 Stunden pro Tag

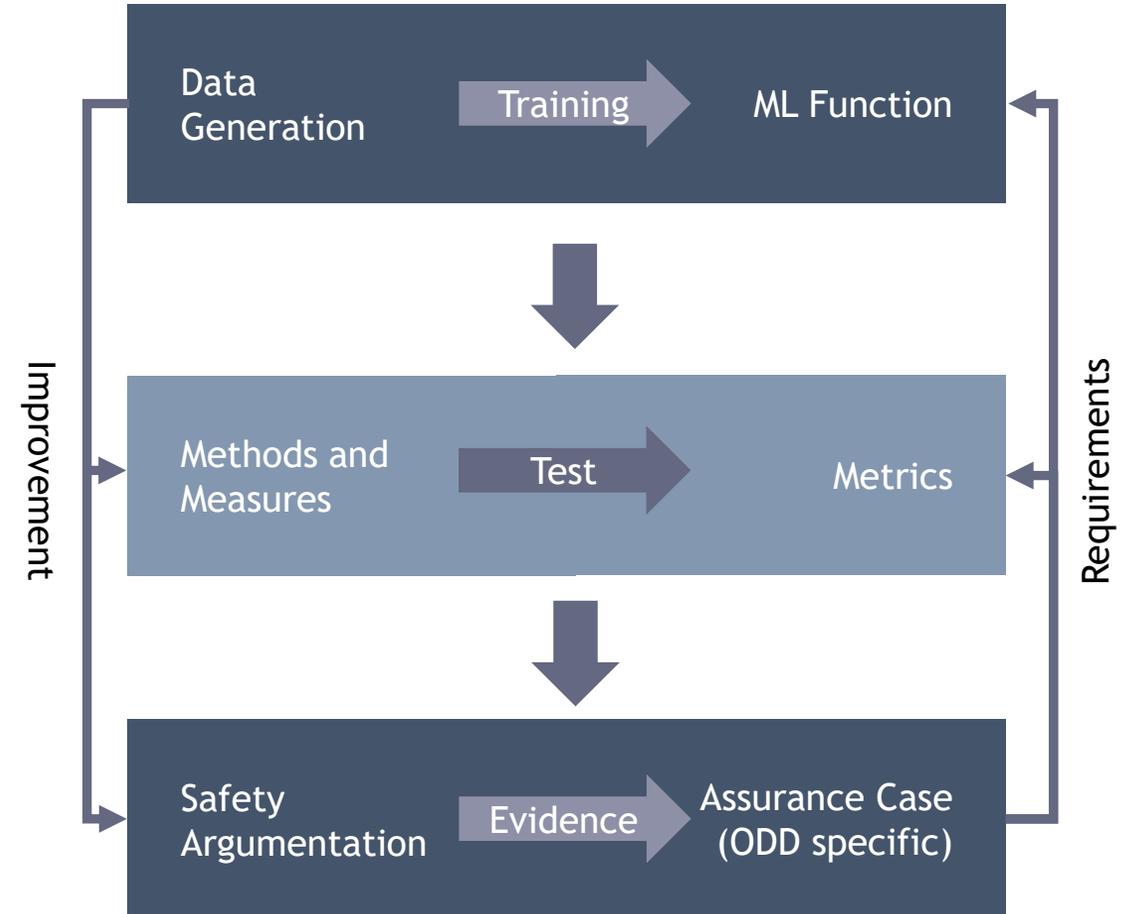
➔ 180 Kritische Vorfälle pro Tag
(Vgl. Verkehrstote: 8-9 pro Tag in DE)

Hinweis:
Lediglich Modellbeispiel
(keine echten Zahlen)

Herangehensweise im Projekt („Projektstruktur“)



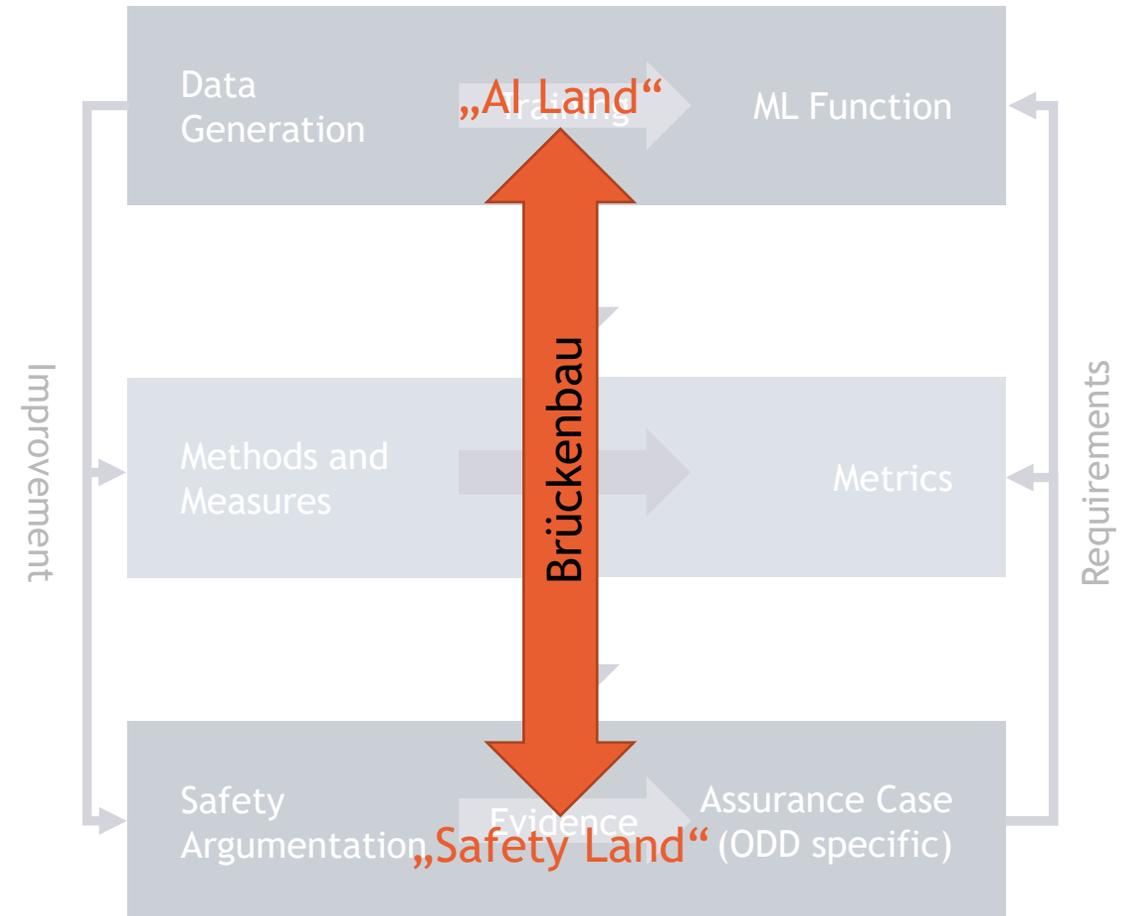
- Prozessbegleitende Generierung von **synthetischen** Lern-, Test- und Absicherungsdaten.
- Entwicklung von Methoden und Maßnahmen, die die KI-Funktion bzgl. eines breiten Spektrums von Metriken verbessern.
- Entwicklung und Validierung von Testmethoden für diese Metriken.
- Stringente Argumentationskette für die KI-Funktion und ihre Operational Design Domain (ODD).



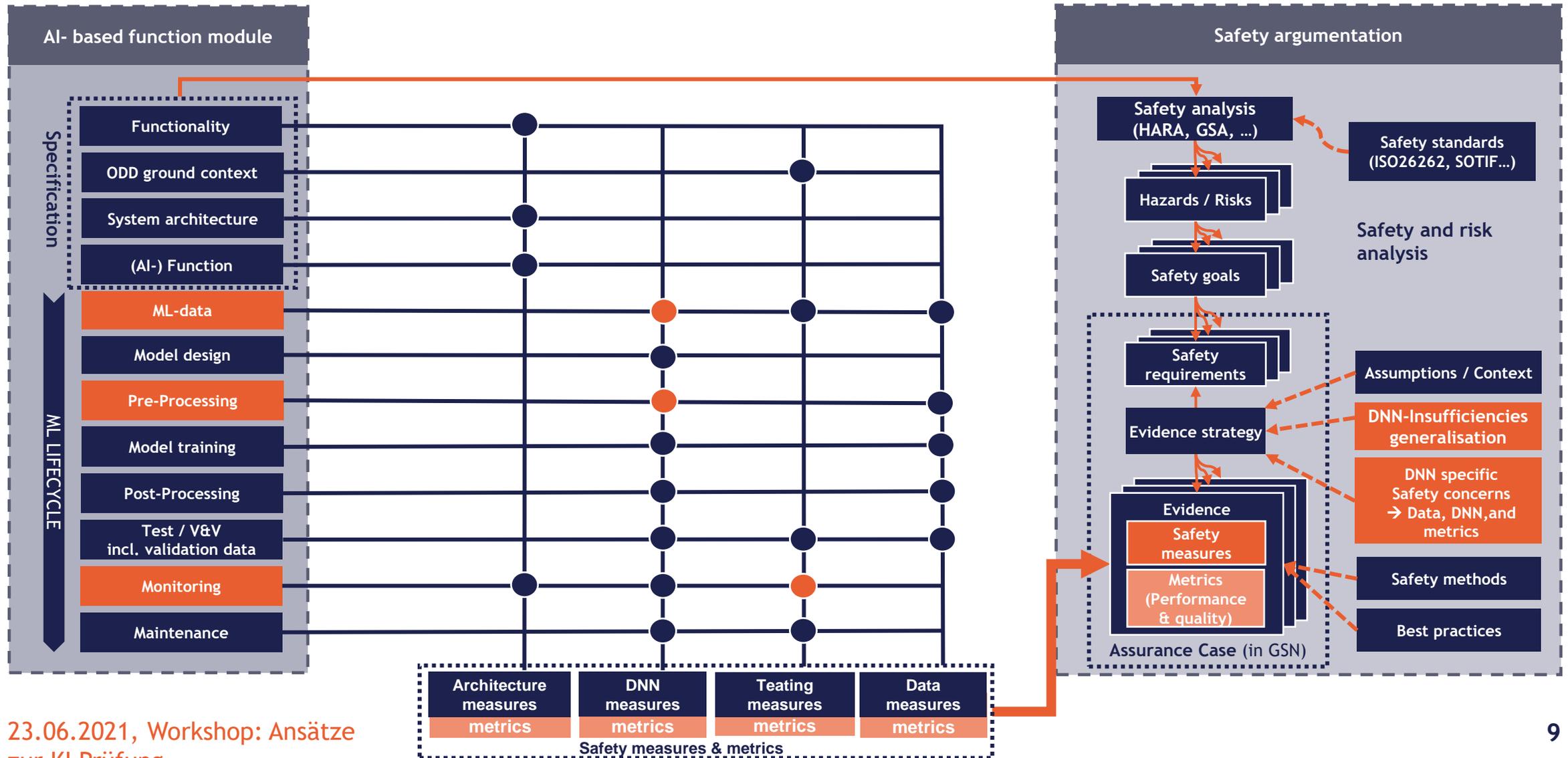
Herangehensweise im Projekt („Projektstruktur“)



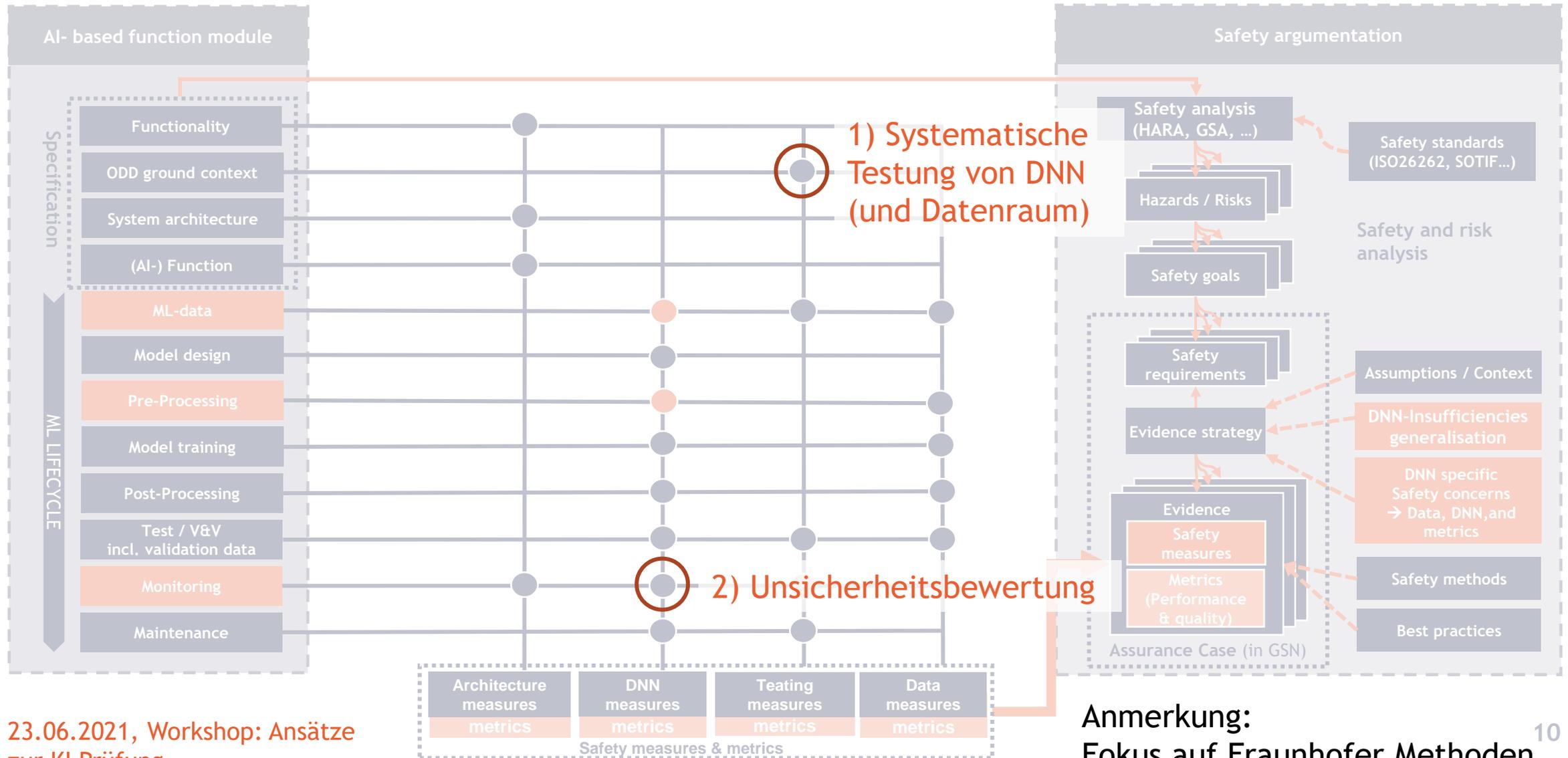
- Prozessbegleitende Generierung von **synthetischen** Lern-, Test- und Absicherungsdaten.
- Entwicklung von Methoden und Maßnahmen, die die KI-Funktion bzgl. eines breiten Spektrums von Metriken verbessern.
- Entwicklung und Validierung von Testmethoden für diese Metriken.
- Stringente Argumentationskette für die KI-Funktion und ihre Operational Design Domain (ODD).



KI spezifische Sicherheitsargumentation



Beispielmethoden im Fokus des Vortrags



Neue Methoden und Werkzeuge für die Entwicklung und Prüfung von KI-Systemen benötigt



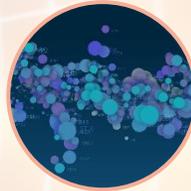
Realistische Unsicherheitsbewertung

Anwendung von Methoden zur Unsicherheitsbewertung bei Neuronalen Netzen



Transparenz-Framework

Nachvollziehbare Entscheidungsmodelle für Blackbox-Modelle erstellen



Explorative Fehleranalyse mit ScrutinAlze

Visuelle Analyse semantischer Hypothesen von Modellen und deren Performance



Regelbasierte Fehlererklärung bei Neuronalen Netzen

Auffinden von Regeln, unter denen Neuronale Netze mit hoher Wahrscheinlichkeit Fehler machen



Automatische Testgenerierung mit Carla

Testen von Modellen an menschlich verständlichen semantischen Konzepten, um Schwachstellen der Modelle aufzudecken

11

Notwendigkeit für Qualitätssicherung auch für andere Use Cases

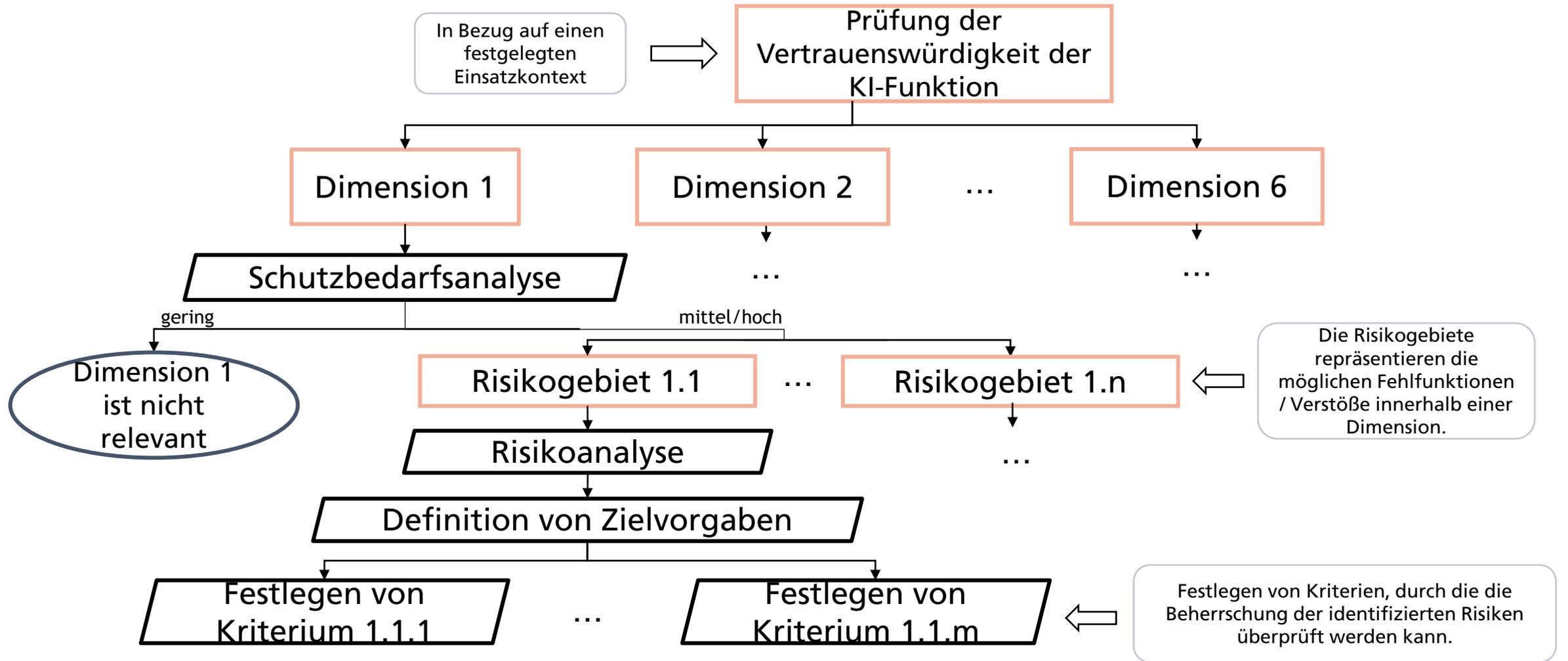
Prüfkatalog bietet Anleitung für systematische Risikobewertung und Mitigation

Veröffentlichung
am 8. Juli 2021

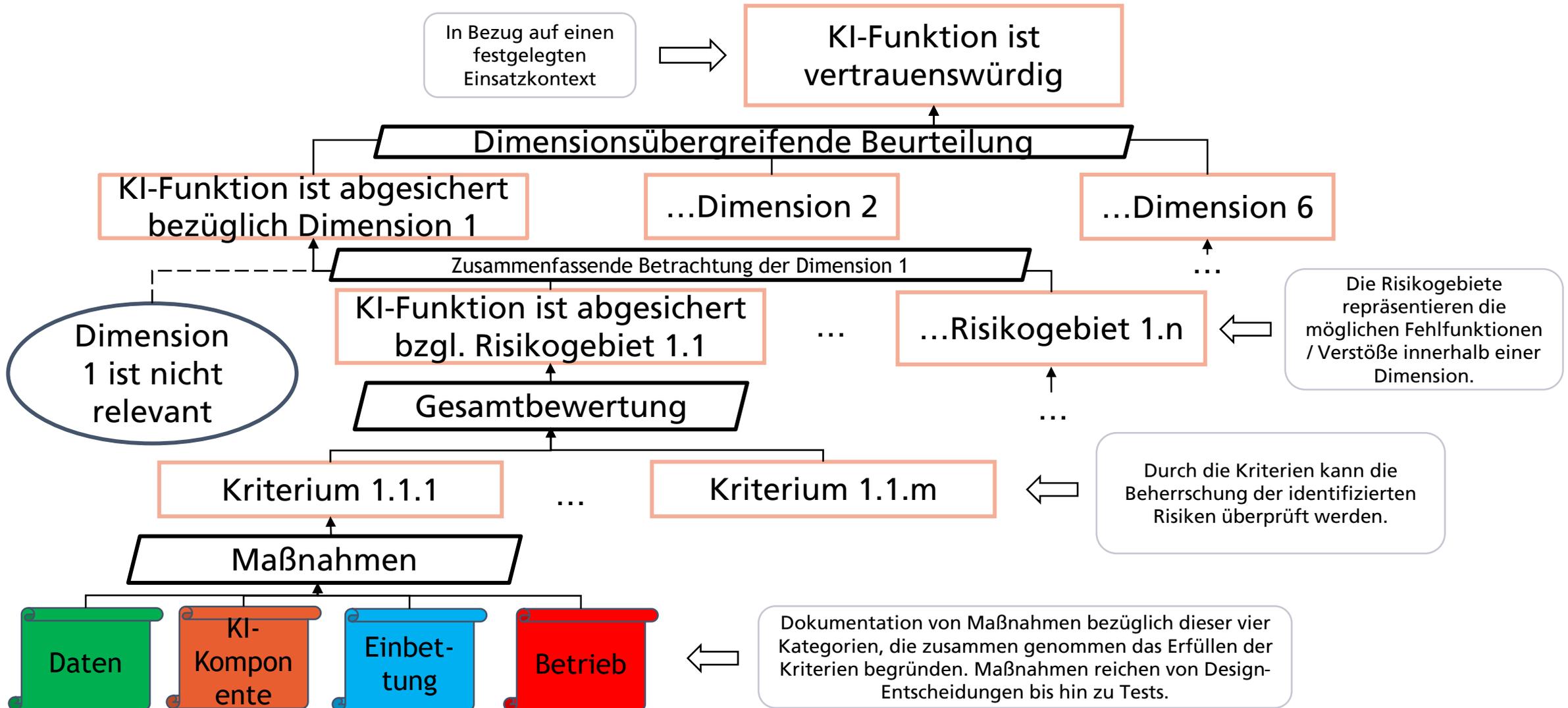
Handlungsfeld	Prüfziel
Fairness	Fairness
	Beherrschung der Dynamik
Autonomie und Kontrolle	Aufgabenverteilung zwischen Mensch und KI-Anwendung
	Sicherstellung der Information und Befähigung von Nutzern und Betroffenen
	Beherrschung der Dynamik
Transparenz	Erklärbarkeit gegenüber Nutzern
	Interpretierbarkeit für Experten
	Auditfähigkeit
	Beherrschung der Dynamik

Handlungsfeld	Prüfziel
Datenschutz	Schutz personenbezogener Daten
	Schutz geschäftsrelevanter Information
	Beherrschung der Dynamik
Verlässlichkeit	Verlässlichkeit im Regelfall
	Robustheit des Modells
	Ausweichstrategien
	Einschätzung v. Unsicherheit
Sicherheit	Beherrschung der Dynamik
	Funktionale Sicherheit
	Integrität und Vertraulichkeit
	Verfügbarkeit
	Beherrschung der Dynamik

Top-down-approach mit Risikoanalyse für spezifischen Use Case



Bottom-up-approach zur Erstellung einer Absicherungsargumentation





KI

ABSICHERUNG

Safe AI for Automated Driving

Dr. Maram Akila | Fraunhofer IAIS | maram.akila@iais.fraunhofer.de

Dr. Tim Wirtz | Fraunhofer IAIS | tim.wirtz@iais.fraunhofer.de

KI Absicherung ist ein Projekt der KI Familie und wurde aus der VDA Leitinitiative autonomes und vernetztes Fahren heraus entwickelt.



KI FAMILIE

www.ki-absicherung.vdali.de  @KI_Familie  KI Familie

Gefördert durch:



aufgrund eines Beschlusses
des Deutschen Bundestages