Fraunhofer IAIS

Prof. Dr. Stefan Wrobel | December 13, 2022

Assessing AI Trustworthiness - Necessity, Potential, or Illusion

Intelligent Systems that Work!

Fraunhofer IAIS – Artificial Intelligence, Machine Learning and Big Data from Bonn

- Research in the paradigm of »hybrid AI« in partnership with University of Excellence Bonn and HBRS
- National Competence Center Lamarr, PhenoRob Cluster of Excellence
- Comprehensive, immediately deployable, proven high-performance technology and IP portfolio
- Consulting, 24/7 implementation, software, licensing, innovation partnerships, trainings
- Customers and partners from DAX30 to medium-sized businesses
- Network management KI.NRW, Fraunhofer-Alliance Big Data and Artificial Intelligence, AI4Europe
- Particular focus on AI safeguarding and certification

350+
Researchers

180+
Research and industrial projects per year

20+
Years of experience

UNIVERSITÄT BONN

LAMARR

KI NRW

Fraunhofer BIG DATA AI

AI4EU

Our focus on Trustworthy AI

EU

HLEG on AI
Collaboration on the European guidelines for trustworthy AI in the HLEG on AI

EU

EU Commission
Dialogue on AI-certification

BSI

BSI
Cooperation to develop certification standards for Germany

CERT·AI

CertAI
Technical partner of »CertAI« (by MunichRE) for the development of a quality assessment of AI solutions

DIN

Standardization Roadmap AI
Member of the steering group of the roadmap, head of the working group certification

Development of AI Systems

Development of and consulting for reliable AI Systems

Assessment of AI Systems

Independent testing and evaluation of AI Systems

Safeguarding AI Systems

Mitigation of identified AI risks

Design of Standards

Support of standardization initiatives and activities

Fraunhofer IAIS

The importance of trust in AI

Fraunhofer IAIS

Why is trustworthy AI important for your business?

Reliably generate business value

Sources-Icons: https://nucleoapp.com/app/

Fraunhofer IAIS

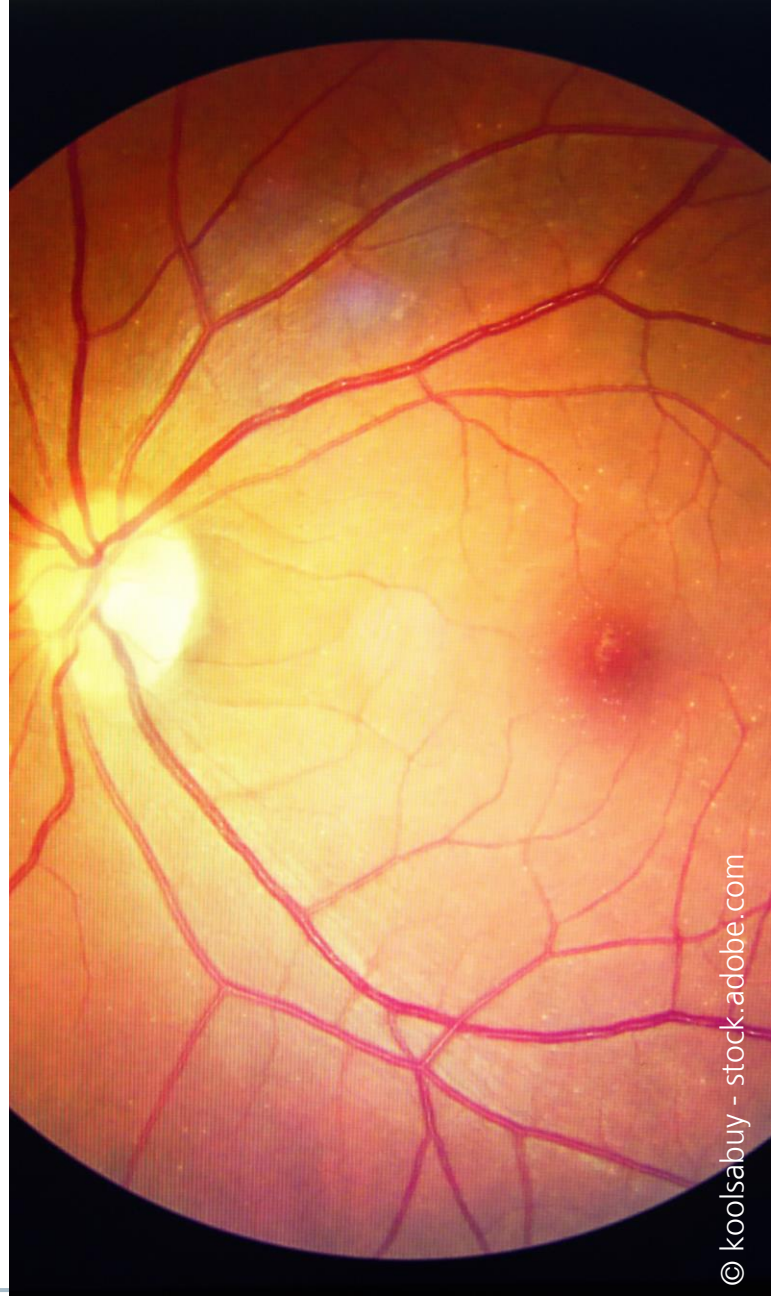
Lessons Learned: Diabetic retinopathy

Lab vs. Deployment

- Diagnostic system developed by Google AI, using image analysis with deep learning
- Accuracy at human expert level (more than 90%)
- System deployed for screening in Thailand in partnership with Ministry of Public Health
- Evaluation in 11 clinics in two provinces over eight months
  - High rejection rate of over 20% due to poor scans
  - Additional personnel resources consumed for retaking images or taking care of patients
  - Significant delays due to cloud-based processing

[Beede et.al, CHI 2020]

https://dl.acm.org/doi/abs/10.1145/3313831.3376718, https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/



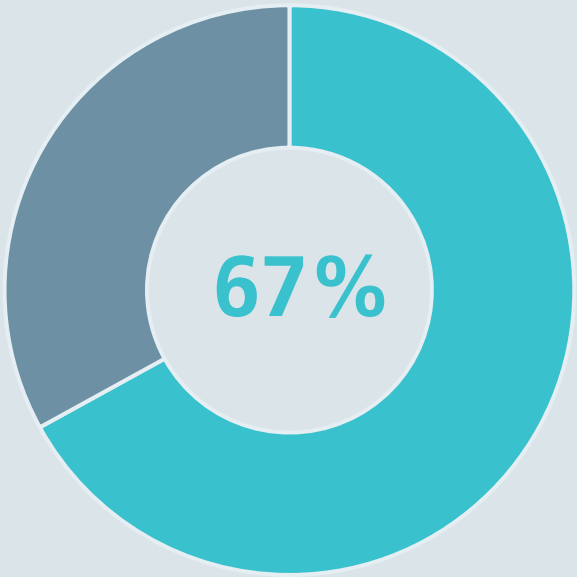
Fraunhofer IAIS

Lack of trust in AI as a top 3 obstacle to its deployment

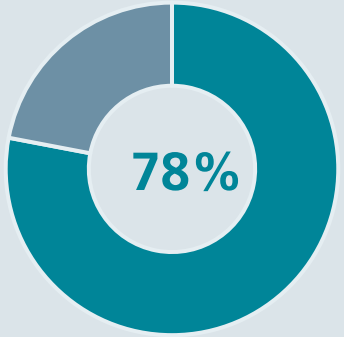
The major obstacles to deployment of AI (answers of respondents, %)

1.005 international managers from public and private sectors respond

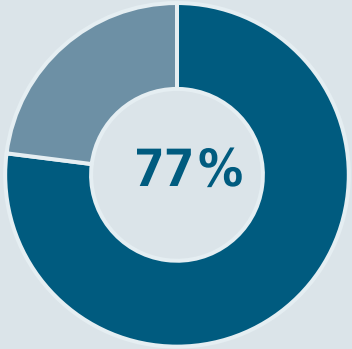
Lack of trust in AI applications and analyses



Insufficient access to AI expertise within or outside of their organisation



Limited availability of AI solutions and products



Source: : *BCG Climate AI survey, May 2022; Remark: Respondents from 14 nations. All respondents have decision-making authority on topics of climate or AI in their organizations. Respondents could give more than one answer.*

Fraunhofer IAIS

Why is trustworthy AI important for your business?

Reliably generate business value

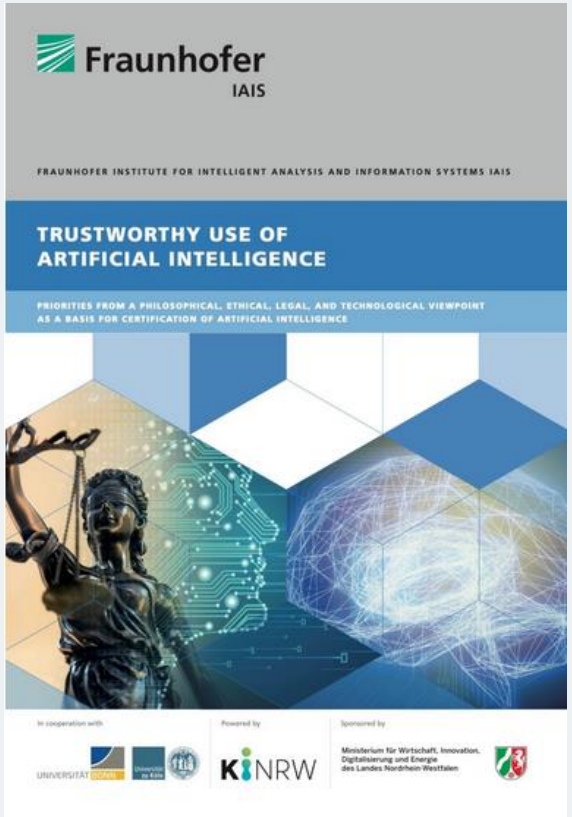
Meet societal expectations, protect brand

Sources-Icons: https://nucleoapp.com/app/

Fraunhofer IAIS

High Expectations

AI use in companies must meet societal expectations



Fairness	Is the AI-system free of discrimination?
Autonomy & Control	Is the degree of autonomy appropriate?
Transparency	Are the functioning and decisions of the AI comprehensible?
Reliability	Is the AI-system reliable?
Safety & Security	Is the AI-system protected against attacks, accidents and errors?
Privacy	Does the AI-system protect sensitive information?

Source: https://www.ki.nrw/wp-content/uploads/2020/03/Whitepaper_Thrustworthy_AI.pdf

Sources-Icons: https://nucleoapp.com/app/

Why is trustworthy AI important for your business?

- Reliably generate business value
- Meet societal expectations, protect brand
- Comply with legal regulations

Sources-Icons: https://nucleoapp.com/app/

Fraunhofer IAIS

Building trust through a legal framework on AI

Upcoming regulation at the European level: AI Act

Proposal for regulation by the EU Commission

Initiation: 2021 proposed by the EU Commission

Goal: Regulate AI harmonized with European values to enable and enforce development of AI systems and ensure trust in AI systems

Current Status: EU Parliament and Council are negotiating the draft within their institutions → next step: trilogue negotiations

Risk based approach:

Level	Description of AI system	Regulation	Example
Unacceptable	Contravening Union values (e.g., fundamental rights)	Prohibited	AI-based social scoring …
High-Risk	High risk to the health and safety or fundamental rights of natural persons	Conformity assessment with specific rules required	Candidate selection…
Limited	Interaction with humans, detection of emotion or association based on biometric data, deep fakes	Transparency obligations	Chat bots …
Minimal	Represent only minmal or no risk for citizens or safety	No regulation	Spam filter …

Sources-Icons: https://nucleoapp.com/app/

High-risk systems

Strong requirements expected

1. ? Conformity assessment
2. 💻 Registration in database
3. ✓ Declaration of conformity

Requirements for High-Risk AI Systems

Article 9: Risk management system

Article 10: Data and data governance

Article 11: Technical documentation

Article 12: Record-keeping obligation

Article 13: Transparency and provision of information

Article 14: Human supervision

Article 15: Accuracy, robustness, cybersecurity

Examples for partly affected areas (according to current status of the proposal):

Biometric identification and categorisation of natural persons

Assess and enjoyment of essential private & public services or benefits

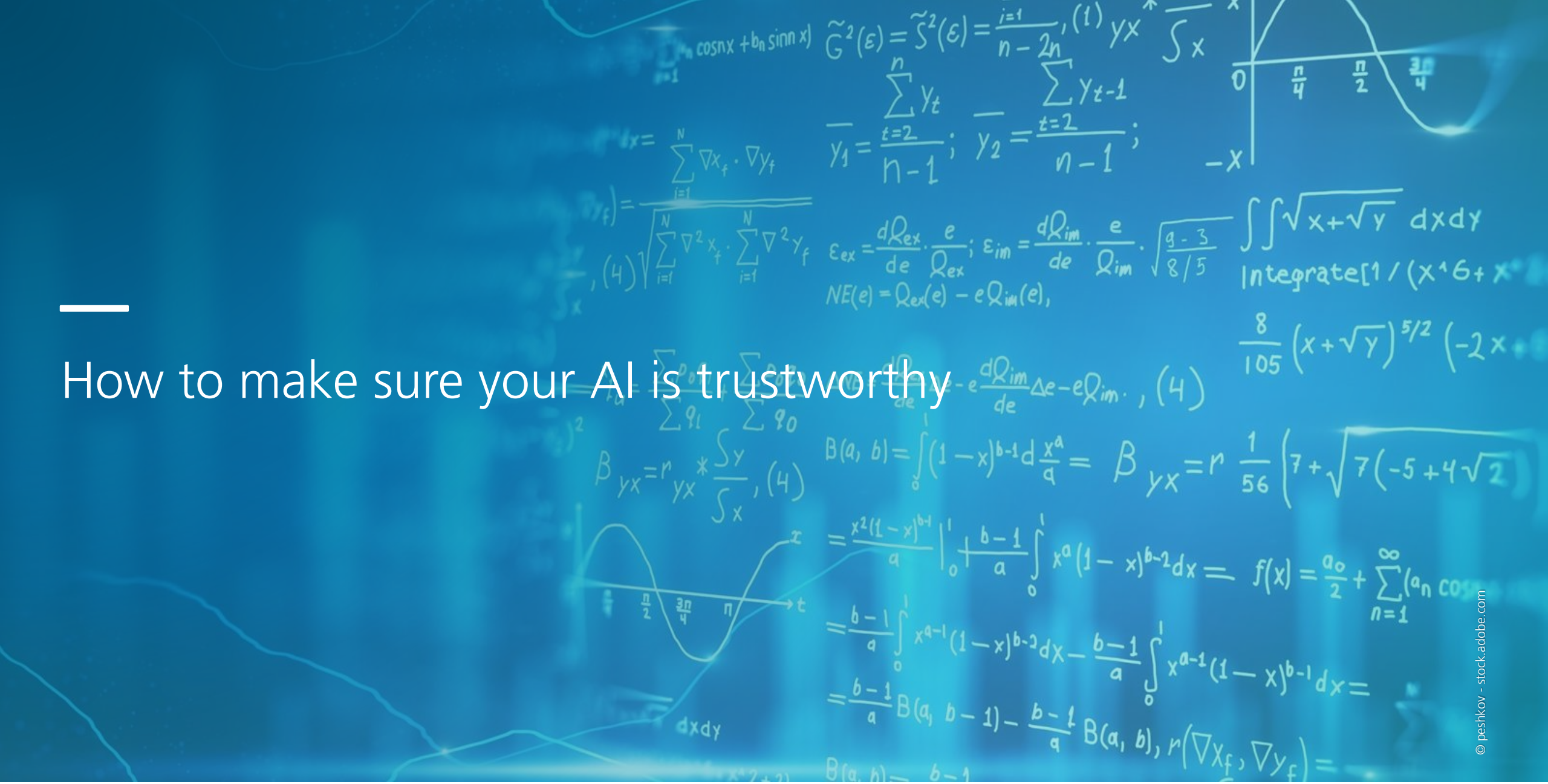
Law enformcement

Administration of justice and democratic processes

....

Fraunhofer IAIS

How to make sure your AI is trustworthy



Fraunhofer IAIS

German standardization roadmap on AI

Two editions as implementation measure of the German government's AI strategy

570 Experts

9 Working groups

18 Coordination members

318 Authors

123 Action recommen-datinos

6 Fields of action

341 Pages

Image Source: DIN

- Development of a framework for action for standardization
- Section on **AI certification**:
  - **Testing framework** that guarantees **comparability** of assessments
  - **Criteria framework** that operationalize trustworthiness requirements and map AI-specific challenges

Source:
GERMAN STANDARDIZATION ROADMAP ON ARTIFICIAL INTELLIGENCE 1, DIN & DKE, 2020
GERMAN STANDARDIZATION ROADMAP ON ARTIFICIAL INTELLIGENCE 2, DIN & DKE, 2022

AI suitability of norms

Review of existing norms and standards with regards to their compatibility concerning the use of AI

> 30.000 norms are being reviewed for their AI readiness

- Identification of content-related connections of relevant norms to AI and need for revision
- Support of technical feasibility of norms
- Contributions of Fraunhofer IAIS:
  - Refining the terms »AI suitability« and »AI relevance«
  - Identification of general categories of norms
  - Development of a prototypical AI tool for machine supported assessment of norms for AI suitability
- Project term 01/2022 – 12/2023

AI suitable set of standards as basis for development and use of high-quality and trustworthy AI

Project partners:

DIN
Beuth Verlag
DIN Software

Fraunhofer IAIS

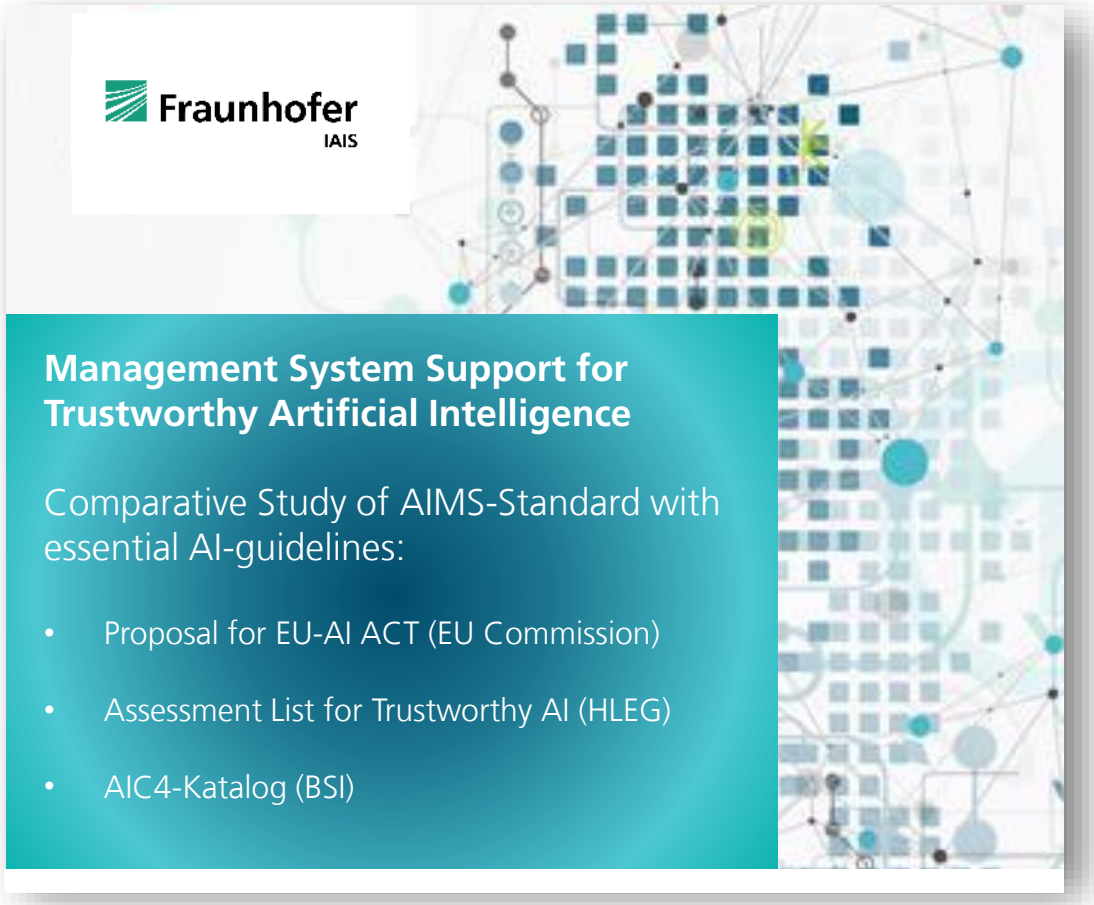
Gefördert durch:

Bundesministerium für Wirtschaft und Klimaschutz

aufgrund eines Beschlusses des Deutschen Bundestages

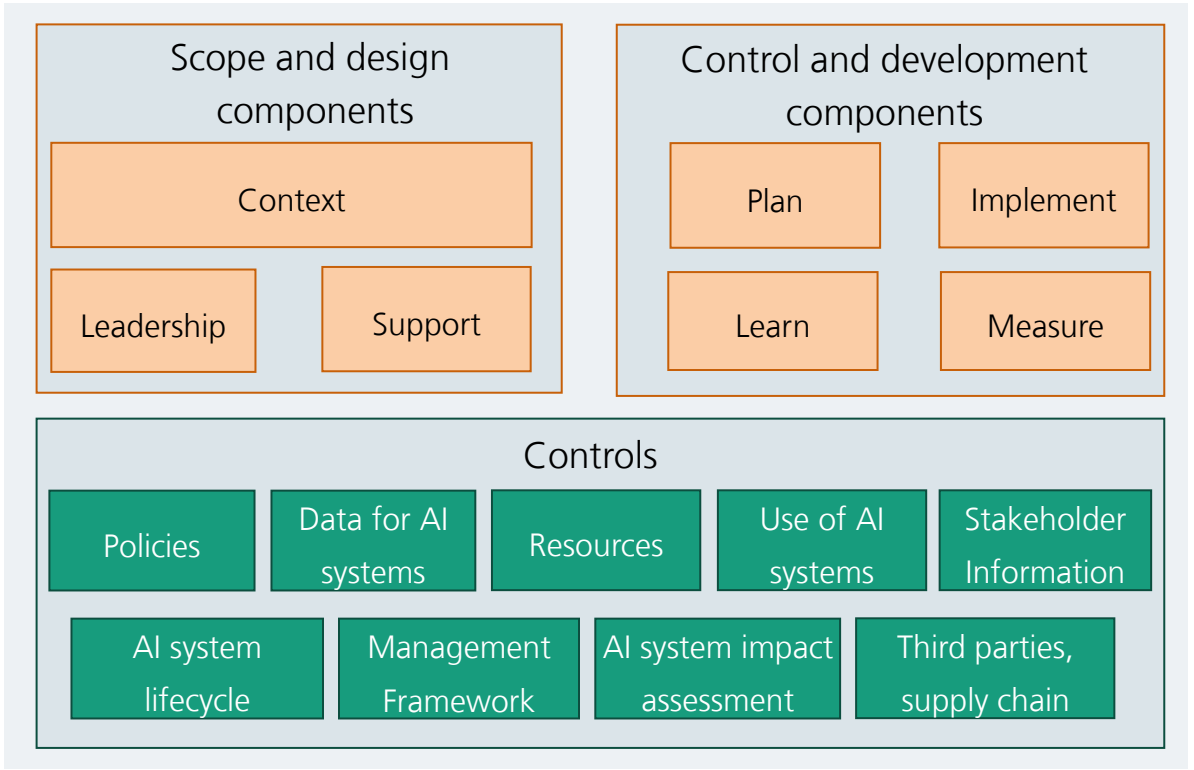
Organizational trust through management systems

Governance, management and technical-organizational measures for AI



Source: Mock, Michael, et al. "Management System Support for Trustworthy Artificial Intelligence." (2021)

Structure of » AIMS«-Standard



Fraunhofer IAIS

Fraunhofer AI Assessment Catalog

Guidelines for a structured evaluation of AI to develop trustworthy AI

Step 1: Risk analysis

- Comprehensive risk analysis along the dimensions of fairness, autonomy and control, transparency, reliability, safety and security and data protection

Step 2: Definition of targets

- Definition of objectives and - preferably measurable - target criteria to mitigate the risks identified in step 1

Step 3: Documentation of measures

- Guidance to systematically list measures along the lifecycle of the AI application to achieve the targets set in step 2

Step 4: Assurance argumentation

- Guidance to develop a stringent argumentation based on the measures of step 3 to demonstrate that the objectives formulated in step 2 have been achieved

Assessment Catalog is freely available at:
https://www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-pruefkatalog.html (English version is in preparation)

Areas of Application

The Assessment Catalog supports

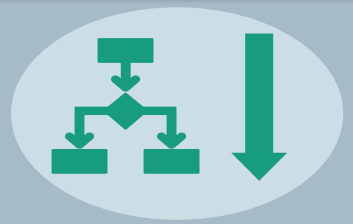
- Developers in the design
  and
- AI assessors in the evaluation and quality assurance

of AI applications.

Logic of the assessment procedure

Fraunhofer AI Assessment Catalog

Step 1: Top-down approach

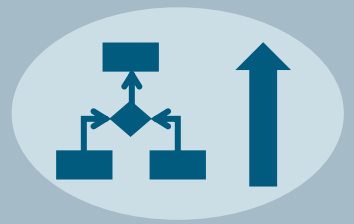


Goal: Operationalize quality criteria

Steps:

- Identify risks
- Analyse risks
- Derive application-specific quality criteria

Step 2: Bottom-up approach



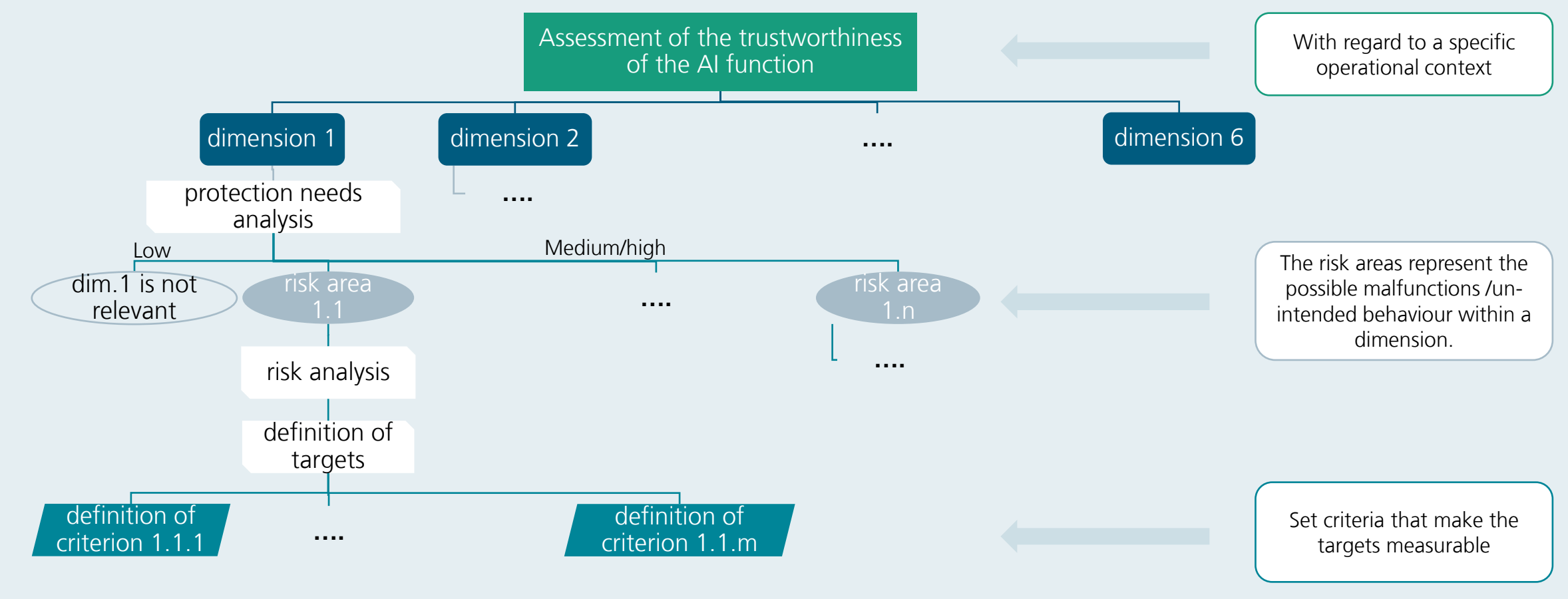
Goal: develop safeguarding argumentation for the trustworthiness of an AI System

Steps:

- Take risk mitigating measures
- Perform tests
- Justify quality criteria

Top-down Approach with Risk Analysis for specific Use Case

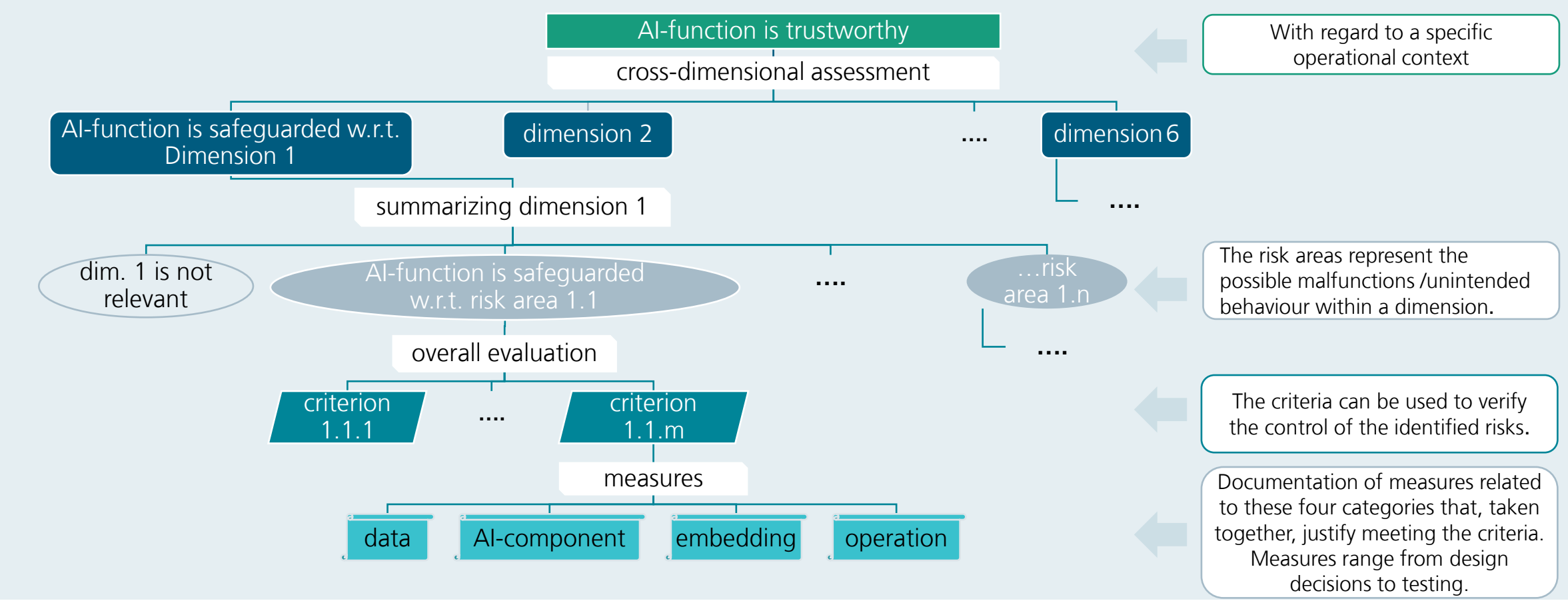
Fraunhofer AI Assessment Catalog



Fraunhofer IAIS

Bottom-up approach for creating a safeguarding argumentation

Fraunhofer AI Assessment Catalog



Fraunhofer IAIS

Sample Project: Synthetic Media Generation

Assessment of a proof-of-concept for AI-based generation of football reports for a media company

- Innovation project of a large media company: AI-based generation of texts for football reports
  - Support for journalists for faster publishing
  - Use of real-time match data for the generation of perspective summaries of matches

- Assessment of trustworthiness of the demonstrator
  - Focus: Reliability, transparency, autonomy & control
  - Identification of risks and weaknesses
  - Requirements and recommendations for assurance of trustworthiness in the use of synthetic media

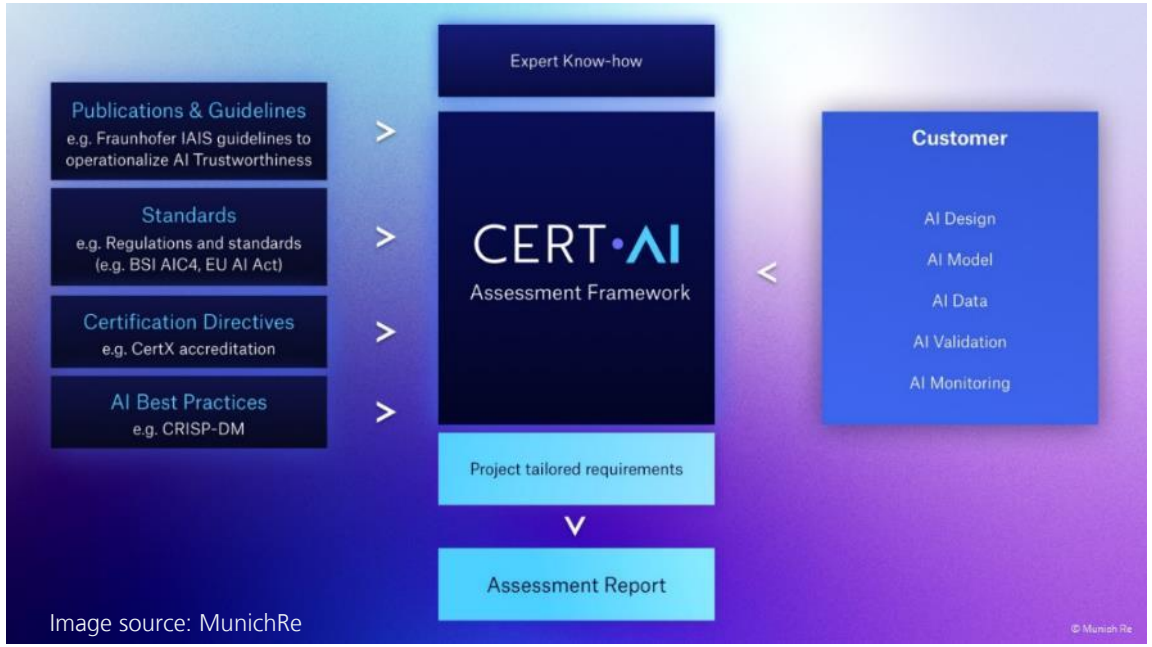
Fraunhofer IAIS

Sample project: AI assessment service for Munich Re

Assessment service CertAI aims to increase acceptance of AI

Munich Re is building a new business area for quality assessment of AI solutions under the brand of »CertAI« – Fraunhofer IAIS is the technology partner

- Subject of the assessment are fully developed or already actively deployed AI systems
- Two assessment dimensions:
  Assessment of the process and assessment of the product. The results are a quality seal and a detailed assessment report
- Assessment service based upon the Fraunhofer IAIS »AI Assessment Catalog«
- Fraunhofer IAIS assists Munich Re with technical product assessments



https://www.certai.com/en.html

Fraunhofer IAIS Assessment Catalog sets standards for AI product assessments on the market

Tool support for assessment

Semantic Analysis of DNN Predictions with Visual Analytics

#	slice_id	ct_id	slice_number	position_z	gt_any	gt_epidural	gt_intraparenchymal	gt_intraventricular	gt_subarachnoid	gt_subdural	final_pred_any	final_pred_epdiural	final_pred_intraparenchymal	f		
0	CQ500-CT-0_CT000028	CQ500-CT-0	0	-11.702	0	0	0	0	0	0	0.000707	0.000107	0.000181			
1	CQ500-CT-0_CT000029	CQ500-CT-0	1	-6.64	0	0	0	0	0	0	0.000635	0.000066	0.00014	0.000115	0.000232	0.000387
2	CQ500-CT-0_CT000026	CQ500-CT-0	2	-1.577	0	0	0	0	0	0	0.000826	0.000064	0.000149	0.000143	0.000237	0.000456
3	CQ500-CT-0_CT000025	CQ500-CT-0	3	3.485	0	0	0	0	0	0	0.001137	0.000067	0.00018	0.000157	0.000278	0.000615
4	CQ500-CT-0_CT000012	CQ500-CT-0	4	8.548	0	0	0	0	0	0	0.00239	0.000092	0.000287	0.000251	0.000416	0.001105

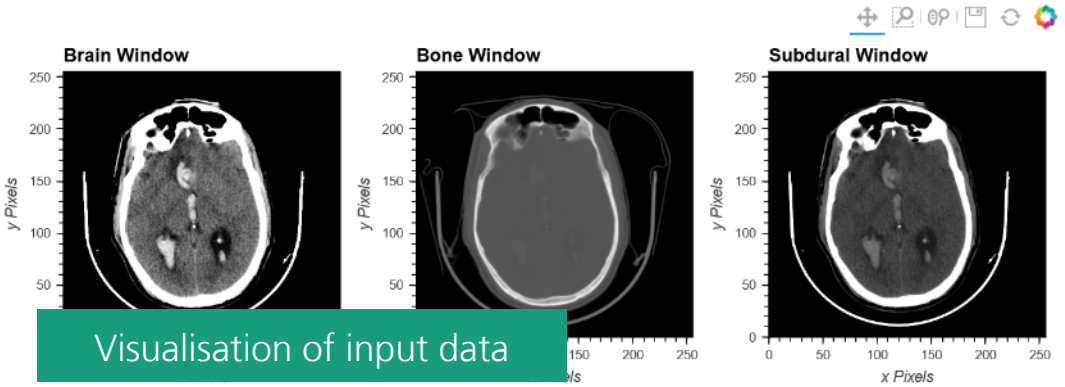
Display metadata

Query input

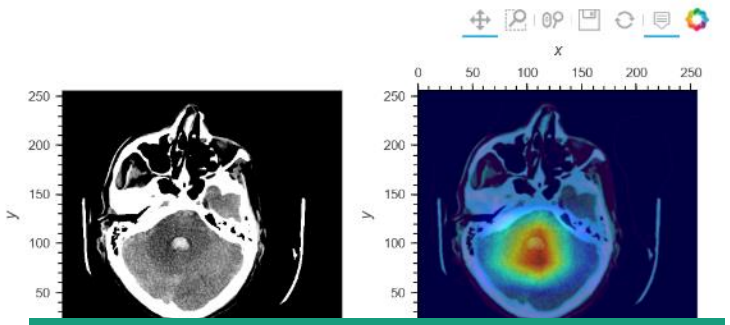
(dim('gt_any')==1)&(dim('final_pred_any')>=0.9)

Apply Query

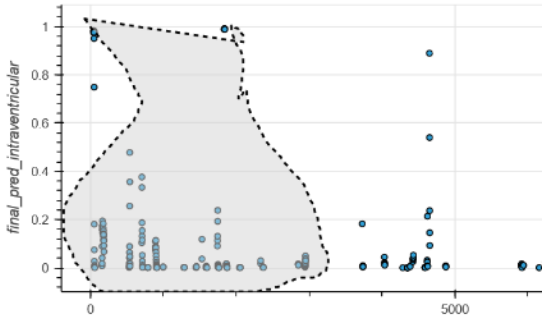
Textual queries



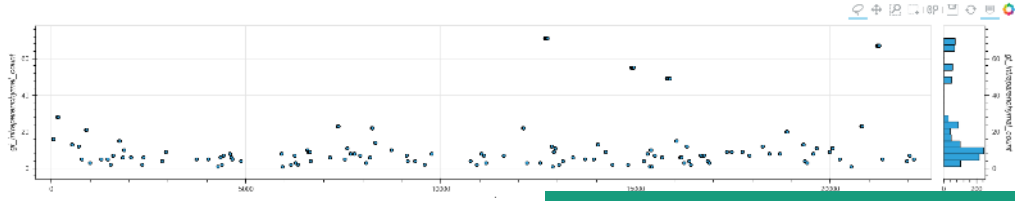
Visualisation of input data



Comparision with local heatmaps



Selection of subgroups



Visual pattern exploration

Sources of Head CT-Scans: http://headctstudy.qure.ai/dataset

Fraunhofer IAIS

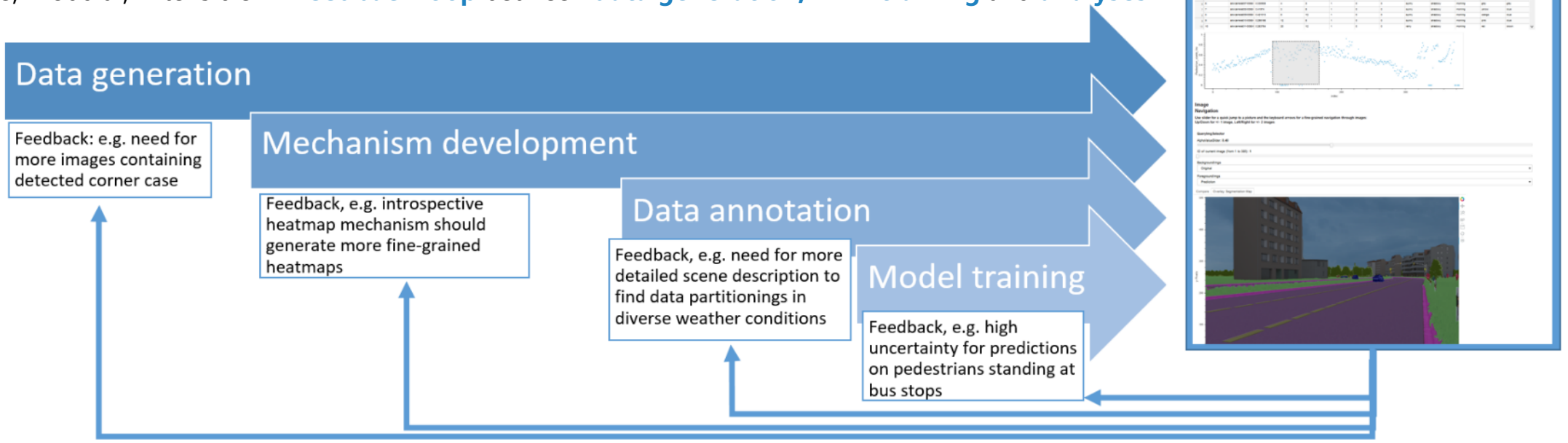
Semantic Analysis of DNN Predictions with Visual Analytics

Establishing a feedback loop

Development of a visual interactive interface

Inspection of **DNN predictions** and **data sets** w.r.t. pre-computed **meta data (semantics)**

Interactive, Modular, Extensible → **Feedback loop** between **data generation, DNN training** and **analyses**



Reference: Haedecke et. al „ScrutinAI: A Visual Analytics Approach for the Semantic Analysis of Deep Neural Network Predictions", EuroVA 2022.

Fraunhofer IAIS

Fraunhofer IAIS

Prof. Dr. Stefan Wrobel | December 13, 2022

Assessing AI Trustworthiness - Necessity, Potential, or Illusion