Trustworthy AI – Obligation or Entrepreneurial Opportunity?

Dr. Maximilian Poretschkin

Fraunhofer Institute for Intelligent Analysis and Information Systems

EAA e-Conference on
Data Science & Data Ethics

29 June 2021

AI IS UBIQUITOUS IN INSURANCE & FINANCIAL INDUSTRY…

Claims processing

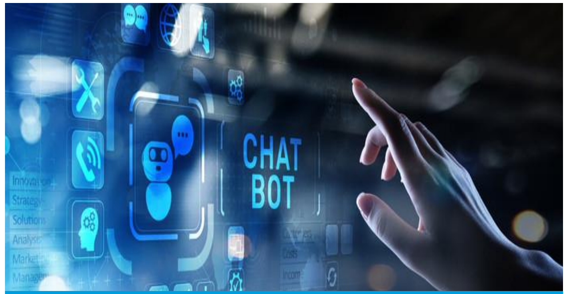
Automated estimation of damage amount

Personalized insurance pricing

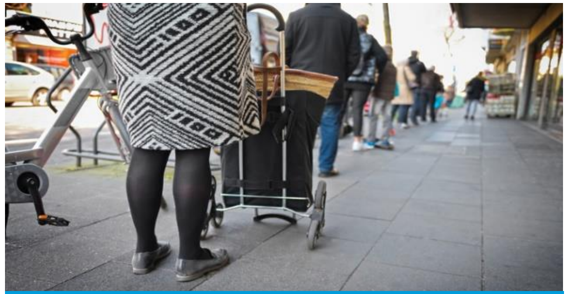
Assessment of the probability of default

Detection of fraudulent claims

Document creation



Responding to customer queries



Churn prediction

... AND CHANGES THE INDUSTRY DISRUPTIVELY

! New providers with expertise in data-driven business models can **occupy the customer interface**

! Insurtechs **unbundle the insurance value chain** by using data and AI use cases

! AI-based **loss prevention**, **differentiated pricing** and **new insurance models** (insurance of AI)

EXAMPLES OF NON-INTENDED OR INCORRECT BEHAVIOR OF AI-APPLICATIONS

„3 crashes, 3 deaths raise questions about Tesla‘s autopilot“
AP News, January 3, 2020

„Apple Card Investigated After Gender Discrimination Complaints“
New York Times, November 10, 2019

„Amazon scraps secret AI recruiting tool that showed bias against women“
Reuters, October 11, 2018

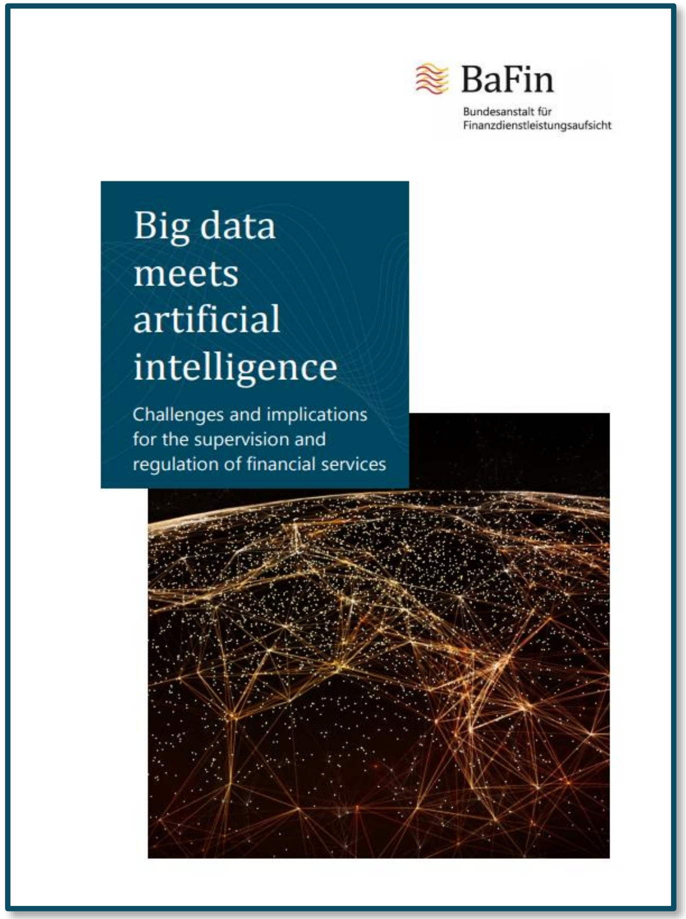
„Alipay responds to risks of facial recognition payment“
China Daily, September 9, 2019

Image source: ©Zerbor - stock.adobe.com

GUARDRAILS FOR ARTIFICIAL INTELLIGENCE

01

INVESTIGATION OF IMPLICATIONS FOR SUPERVISION HAS ALREADY STARTED IN 2018



- Embedding BDAI within **a proper business organization**
- **No black box excuses –** explainability/traceability of models is necessary
- Continuing to develop existing **governance concepts**
- Defining **supervisory requirements for the explainability** and effectiveness of compliance processes
- Defining **prerequisites for BDAI use in models** requiring **supervisory approval**
- Addressing increased **information security risks** and using BDAI to combat them

Source: Big Data meets Artificial Intelligence, Challenges and implications for the supervision and regulation of financial services, BaFin, 2018

european actuarial academy

REQUIREMENTS FOR THE USE OF ARTIFICIAL INTELLIGENCE BY THE EU HLEG AI

Human agency and oversight
- Fundamental rights, human agency and human oversight

Technical robustness and safety
- Resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility

Privacy and data governance
- Respect for privacy, quality and integrity of data, and access to data

Transparency
- Traceability, explainability and communication

Diversity, non-discrimination and fairness
- Avoidance of unfair bias, accessibility and universal design, and stakeholder participation

Societal and environmental wellbeing
- Sustainability and environmental friendliness, social impact, society and democracy

Accountability
- Auditability, minimisation and reporting of negative impact, trade-offs and redress.

"The list of requirements is non-exhaustive."

Source: ETHICS GUIDELINES FOR TRUSTWORTHY AI, High-Level Expert Group on Artificial Intelligence, 2019

european actuarial academy

GERMAN STANDARDIZATION ROADMAP HAS BEEN PUBLISHED IN DECEMBER 2020

WORKING GROUP QUALITY, CONFORMITY ASSESSMENT AND CERTIFICATION

Gefördert durch:
Bundesministerium für Wirtschaft und Energie
aufgrund eines Beschlusses des Deutschen Bundestages

Two deliverables for establishing a testing procedure:

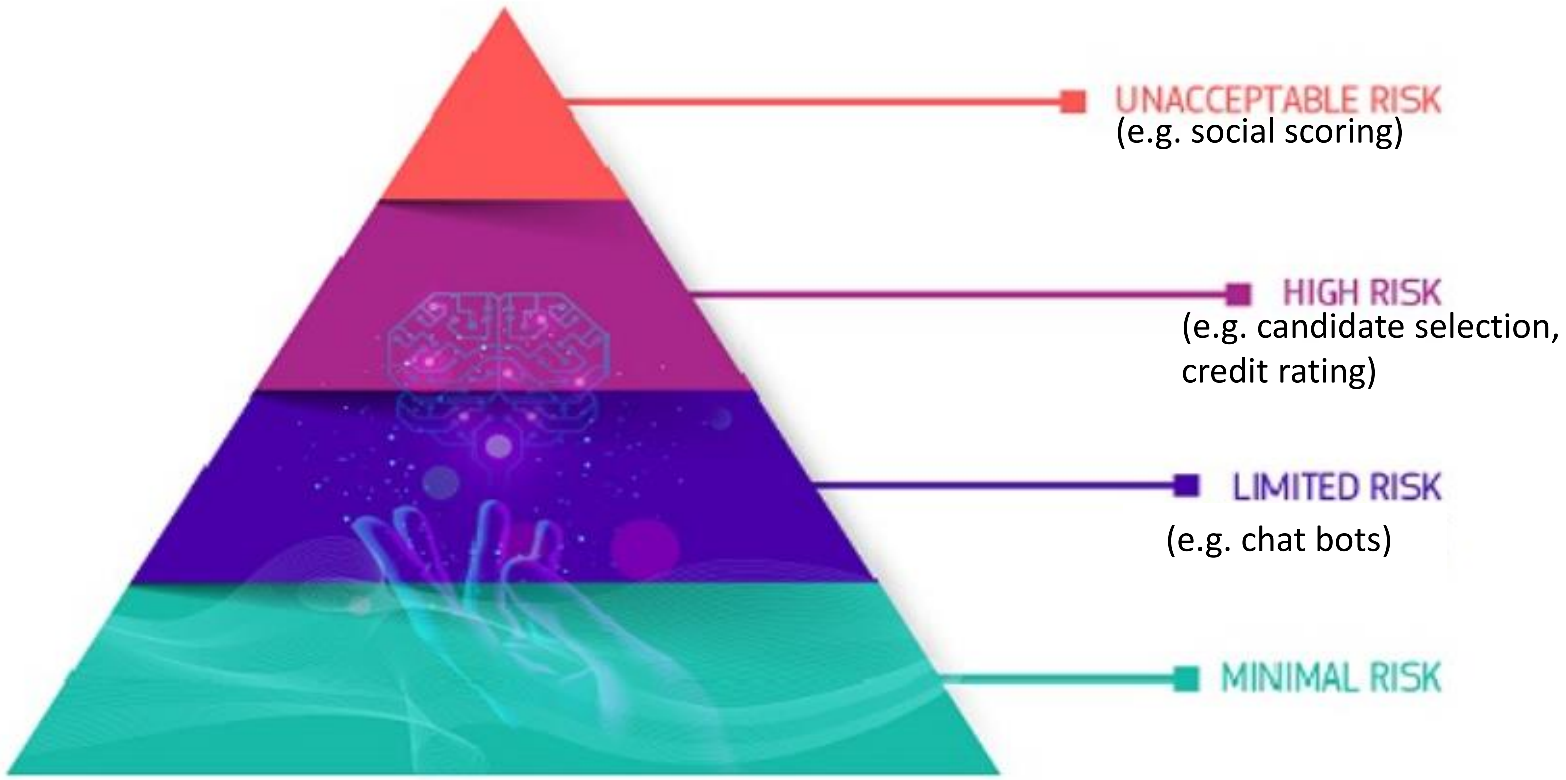
1) Testing framework that guarantees comparability of tests (and is compatible with existing IT testing procedures!).
   - Process testing (standards for the development and operation of AI systems)
   - Product testing (verification of assured properties)
   - Differentiated assurance levels / testing depths

2) Criteria frameworks that operationalize trustworthiness requirements and map AI-specific challenges.
   - Use case dependency in formulation is challenge (metrics, thresholds)
   - Completely new testing tools and methods

Source: GERMAN STANDARDIZATION ROADMAP ON ARTIFICIAL INTELLIGENCE, DIN & DKE, 2019

EU HAS RECENTLY PUBLISHED A PROPOSAL TO REGULATE AI-SYSTEMS

MANY INSURANCE USE CASES ARE AFFECTED



Procedure for high risk AI-systems

1) Development of a high-risk system
2) Conformity assessment and potentially notification of a notified body
3) Registration within an EU data base
4) Declaration of conformity & placing on the market

Procedure to systematically evaluate risks of AI-Systems is crucial!

Image source: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence

AN UNFAIR RACE? WE HAVE TO FIND A MIDDLE GROUND

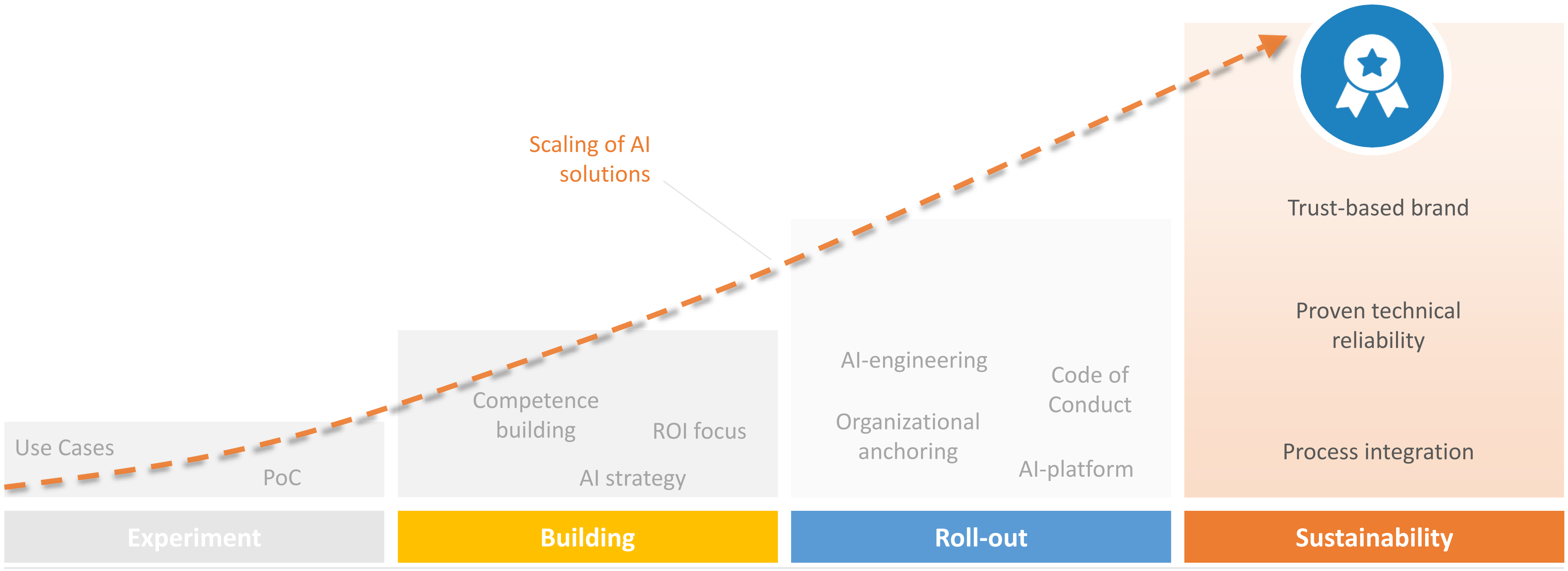
Image source: Adopted from Prof. Dr. Dr. Wolfgang Wahlster, Plattform Lernende Systeme

ENTRE-
PRENEURIAL
OPPORTUNITY

02

AI ENTERS THE SCALING PHASE - SUSTAINABILITY REQUIRES TRUST

FROM AI EXPERIMENT TO SUSTAINABLY SCALABLE AI SOLUTION



GOVERNANCE, MANAGEMENT & TECHNICAL-ORGANIZATIONAL MEASURES

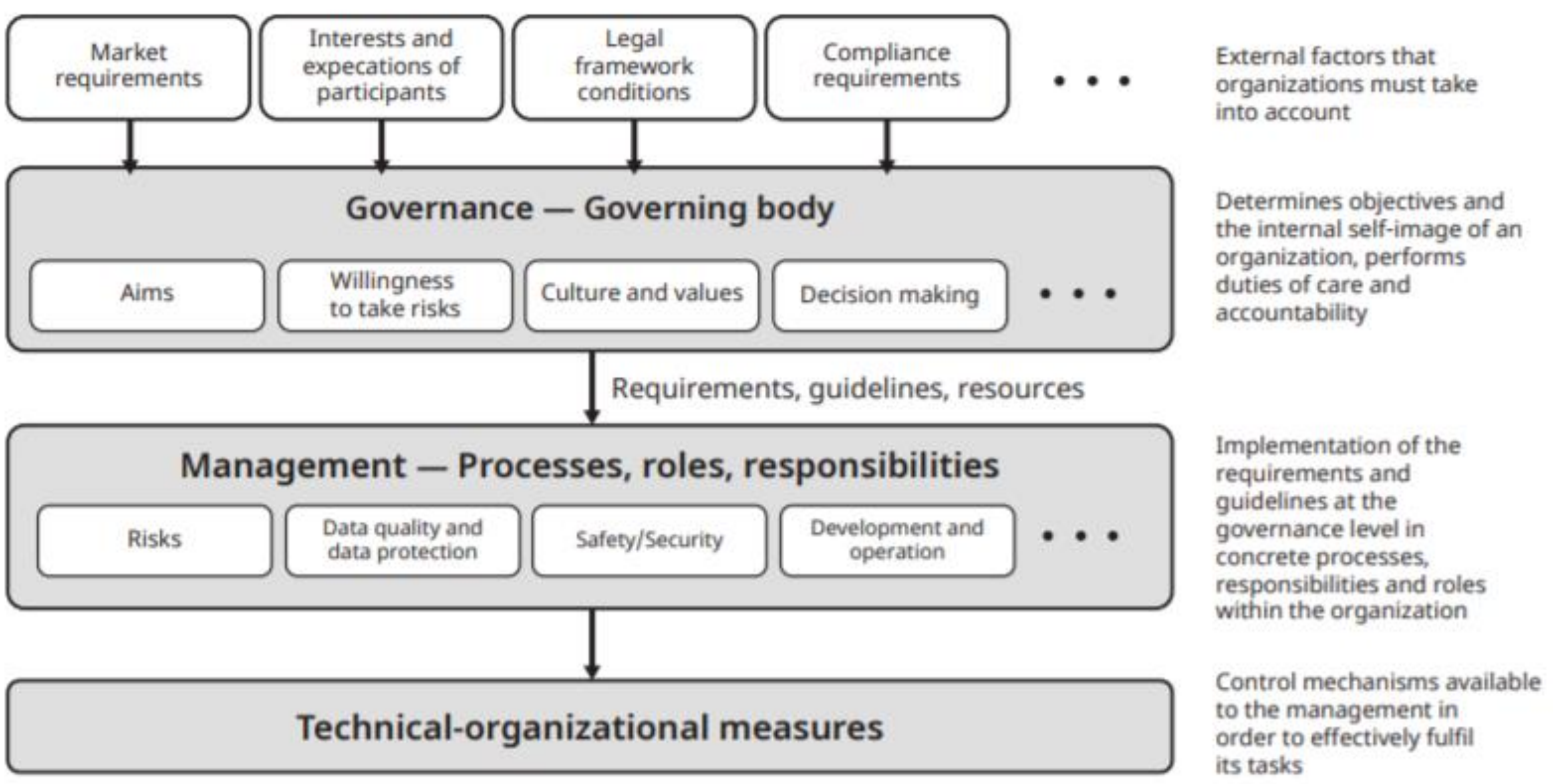


Image source: GERMAN STANDARDIZATION ROADMAP ON ARTIFICIAL INTELLIGENCE, DIN & DKE, 2019

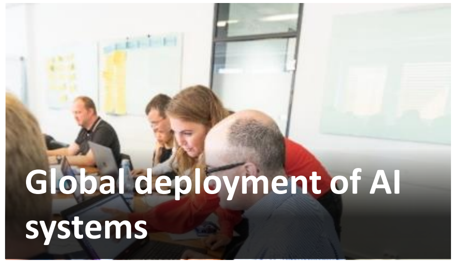
european actuarial academy

THE MAIN USE CASES CAN BE DIVIDED INTO 4 CATEGORIES

Building internal trust

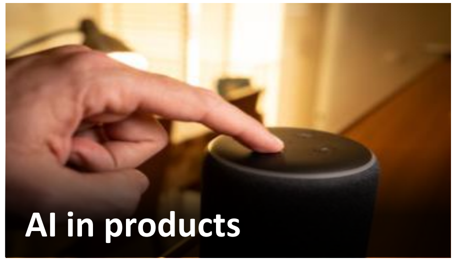
Are the AI-based recommendations comprehensible and trustworthy?

Can malfunctions cause significant (personal and/or financial) damage?



Is the AI system reliable enough to be rolled out globally?

Building external trust



Can a competitive advantage be generated through proven technical reliability?

How can a trusted brand be maintained for products with AI components?

Understanding risks

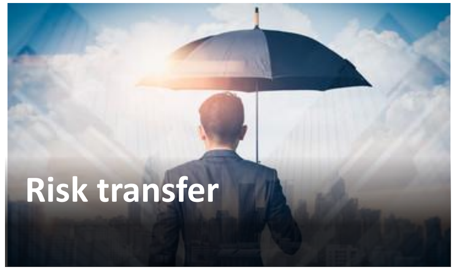
Does the purchased AI solution meet the required characteristics?

Does a company takeover entail technical risks? Does an acquired AI solution meet the expected requirements?

Are AI risks recorded and assessed transparently? Are internal AI guidelines implemented?

Risk transfer

Can proof of technical reliability reduce the insurance premium?



Can the residual risk be covered by AI insurance?

Image sources: fizkes - stock.adobe.com, Nataliya Hora - stock.adobe.com, Fraunhofer IAIS, tanaonte - stock.adobe.com, Jacob Lund - stock.adobe.com, Looker_Studio - stock.adobe.com, amnaj - stock.adobe.com, Song_about_summer - stock.adobe.com, Zerbor - stock.adobe.com, TimeStopper - stock.adobe.com

CLASSICAL APPROACHES ARE NOT TRANSFERABLE

- Data from past damage caused by AI applications is scarce
- Risks are AI application-specific and can only be generalized to a limited extent
- AI risks and damage scenarios are continuously changing. Approaches and competencies must be continuously updated
- "Silent AI risks" in existing policies must be made transparent

Icon source: flaticon

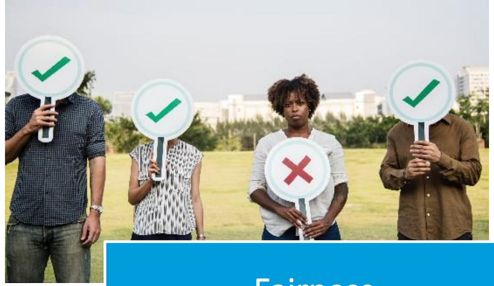
SYSTEMATIC EVALUATION OF AI-RISKS

03

TYPICAL RISK FIELDS OF AI-SYSTEMS

Ethics & Law

Key questions concerning ethical issues

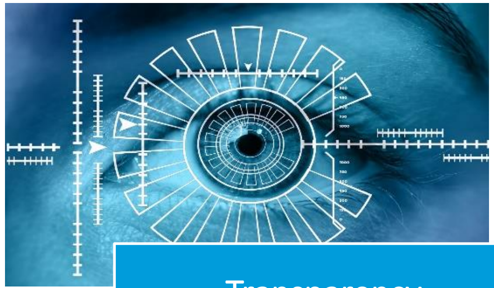


Fairness

Historically unbalanced data

Autonomy & Control

Appropriate degree of autonomy



Transparency

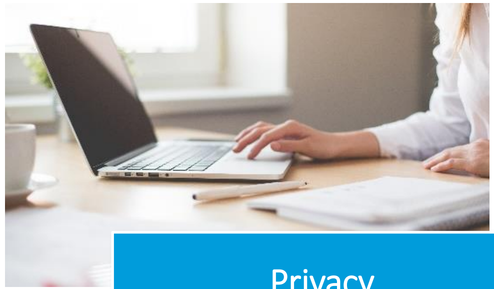
Incomprehensibility of results from neural networks

Reliability

Robustness of results processed by AI-systems

Safety & Security

Safety risks due to probabilistic output from AI component



Privacy

New types of personal data through AI

Image sources: https://www.pexels.com/de-de/foto/afroamerikaner-betrubt-draussen-farbige-frau-1656594; https://www.pexels.com/photo/person-using-white-tablet-computer-displaying-location-text-1305305; https://pixabay.com/de/illustrations/auge-iris-biometrie-iriserkennung-2771174; https://pixabay.com/de/illustrations/sicherheit-schloss-sicher-internet-1202344; https://www.pexels.com/photo/ballpen-blur-close-up-computer-461077

ENTIRE LIFECYCLE OF AI-SYSTEM NEEDS TO BE TAKEN INTO ACCOUNT

Design		▪ The conception and architecture of the AI-system which ensures that certain characteristics are fulfilled „by design“, like Privacy-by-Design, Safety-by-Design and Verifiability-by-Design.
Develop-ment	**Data**	▪ The selection, augmentation, pre-processing of the training-, test- and input data of the AI–system as a key pre-requisite for a high quality of the AI-system.
	AI-Com-ponent	▪ The selection of a method/algorithm, the training and test/validation of the model, aspects of transparency and explainability. The implementation into (standard) software.
	Embed-ding	▪ The embedding of the AI-component into the AI-system with a focus on those aspects of the AI-system whose behaviour is based on the AI component.
Operation		▪ Application-related testing and assurance of the model’s quality during operations. Verifiability and logging of the behaviour which is based on the AI-component.

Icon source: https://iconmonstr.com/construction-35-png; https://iconmonstr.com/folder-20-png/

WHAT IS THE STARTING POINT OF DISCUSSING FAIRNESS?

- Case 1: There is a **commonly preferred label**
  - We don`t want to be refused that label due to
    - gender, ethnicity, …
  - Such attributes should not play a role in the decision process


DHH
@dhh

The @AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

9:34 nachm. · 7. Nov. 2019 · Twitter for iPhone

12.826 Retweets 28.631 „Gefällt mir"-Angaben

- Case 2: There is **no commonly preferred label**
  - But we care whether we are **assigned the correct label**
  - We don`t want to be treated with less care / worse service quality due to
    - gender, ethnicity, …
  - Such attributes should not influence the model performance

Gender was misidentified in **up to 12 percent of darker-skinned males** in a set of 318 photos.

Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

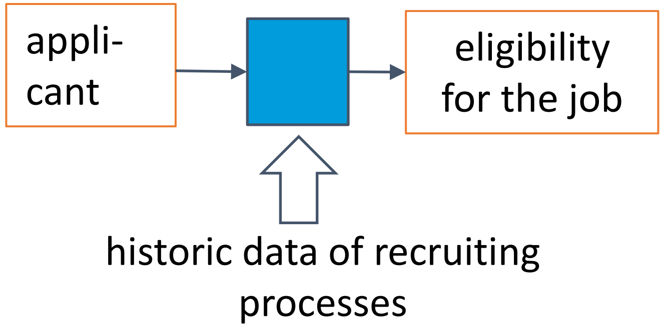
Icon sources: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; https://www.pexels.com/de-de/foto/afroamerikaner-betrubt-draussen-farbige-frau-1656594; Joy Buolamwini, M.I.T. Media Lab

european actuarial academy

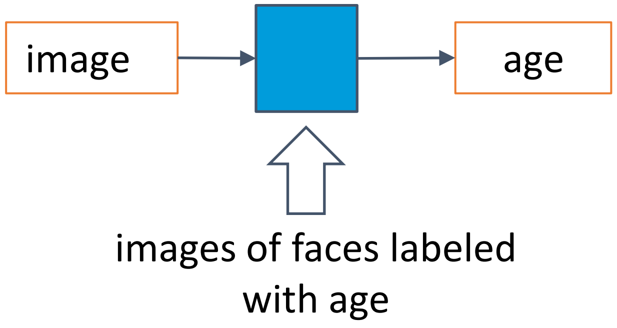
OUTPUT OF AI-APPLICATION REFLECTS BIAS IN DATA

DATA IS A MAJOR CAUSE OF DISCRIMINATION BY AI-APPLICATIONS

Example: Recruiting Tool
Historic bias: Men were systematically preferred → ML-model could **learn negative correlation** of *woman* and *eligibility*



Example: Age Predictor
Certain group **underrepresented** in training data (or only **incomplete/ inaccurate data** available)
→ **higher error rate** wrt this group



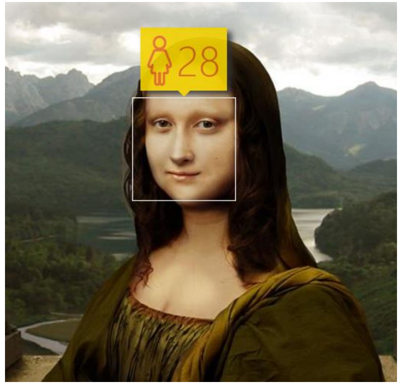


Image sources: https://www.pexels.com/de-de/foto/afroamerikaner-betrubt-draussen-farbige-frau-1656594; www.how-old.net

FAIRNESS NEEDS A QUANTITATIVE DEFINITION

CHOICE OF FAIRNESS METRIC IS CRUCIAL FOR SAFEGUARDING

Risk analysis

- What are the **risks** regarding fairness?
- Which **concept of fairness** is appropriate in the given context?
- To what extent is there a **trade-off** between fairness and utility of the application?

→ Choice of fairness metric **determines the procedure** of safeguarding and **enables objective evaluation** of measures

Icon sources: Flaticon.com icons by srip: Scale blue, risk blue, collaboration blue

SELECTION OF MOST COMMON FAIRNESS METRICS AT A GLANCE

WIDE RANGE OF METRICS / FAIRNESS CONCEPTS

Concepts of group fairness		Other concepts
Statistical/Demographic Parity	Overall Accuracy Equality	Individual Fairness
Predictive Rate Parity	Treatment Equality	Causal Discrimination
Equalized Odds	Well-Calibration	Counterfactual Fairness
Equal Opportunity	Test-fairness	…

SELECTION OF METHODS TO MITIGATE UNFAIRNESS AT A GLANCE

MITIGATE UNFAIRNESS BY MODIFYING DATASETS FOR TRAINING

Pre-Processing		In-Processing	Post-Processing
Massaging	Unawareness	Adversarial Debiasing	Equalized Odds Threshold Predictor
Uniform Sampling	Disparate Impact Remover	Classifier without Disparate Mistreatment	Reject Option Classification
Preferential Sampling	Optimized Preprocessing	Prejudice Remover Regularizer	Calibrated Equalized Odds Postprocessing
Reweighing	Learning Fair Representations		

STRUCTURED RISK ASSESSMENT NECESSARY FOR ALL AI RISK DIMENSIONS

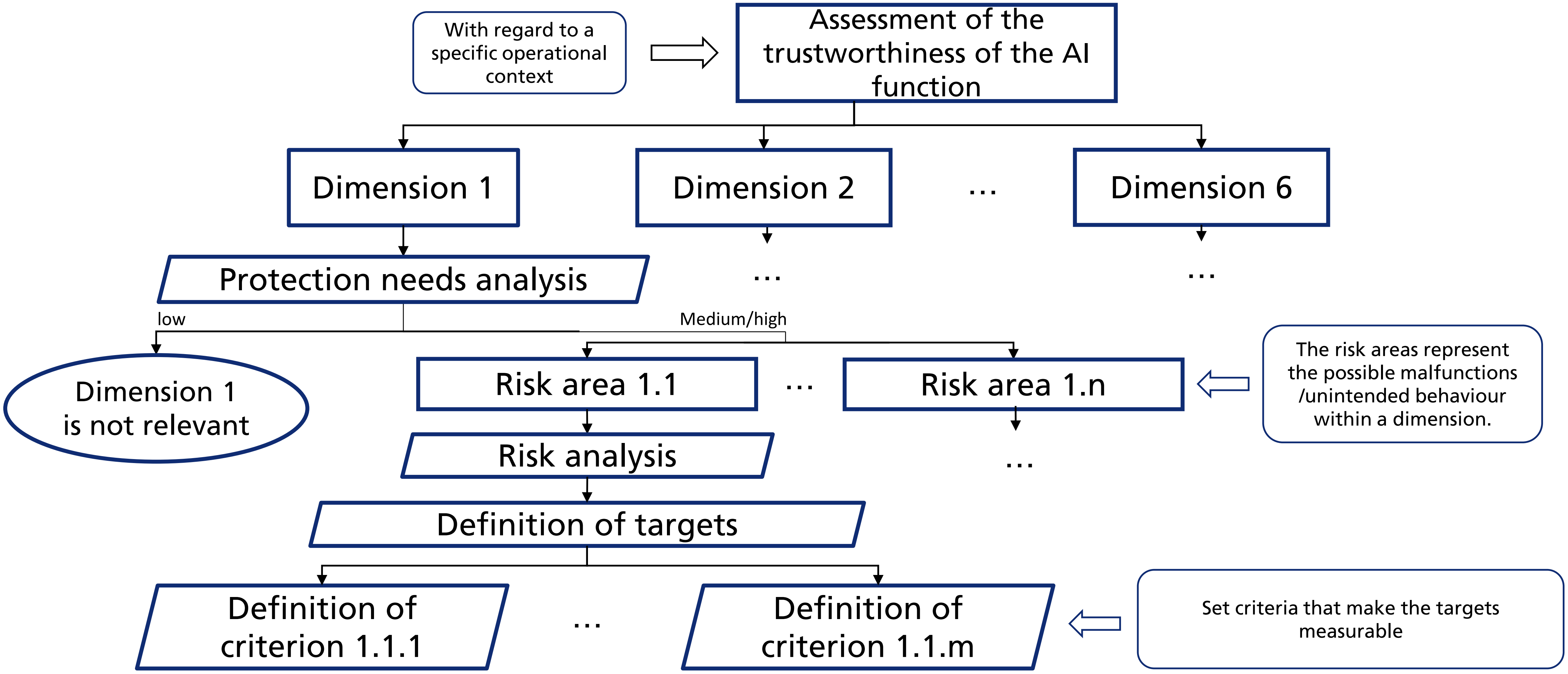
ASSESSMENT CATALOGUE PROVIDES GUIDANCE

Dimension	Risk area
Fairness	Fairness
	Control of dynamics
Autonomy and Control	Distribution of tasks between human and AI-system
	Information and empowerment of users and stakeholders
	Control of dynamics
Transparency	Explainability to users
	Interpretability for experts
	Auditability
	Control of dynamics

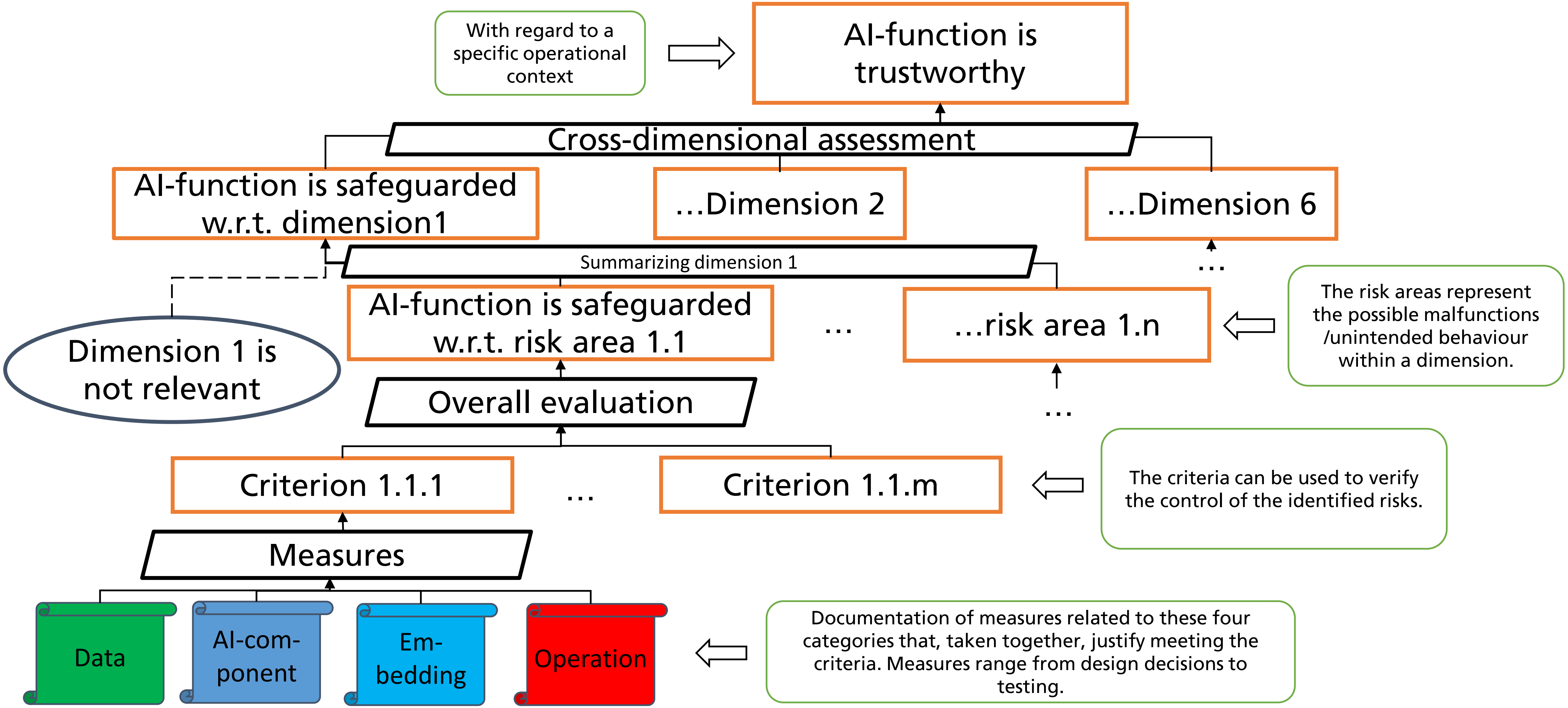
Dimension	Risk area
Privacy	Protection of personal data
	Protection of business-relevant information
	Control of dynamics
Reliability	Reliability during regular operation
	Robustness
	Evasion strategies
	Estimation of uncertainty
	Control of dynamics
Safety and Security	Functional safety
	Integrity and confidentiality
	Availability
	Control of dynamics

Assessment catalogue is available here: www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-pruefkatalog

TOP-DOWN APPROACH WITH RISK ANALYSIS FOR SPECIFIC USE CASE



BOTTOM-UP APPROACH FOR CREATING A SAFEGUARDING ARGUMENTATION



STRATEGIC PARTNERSHIP BETWEEN FRAUNHOFER AND BSI ON TRUSTWORTHY AI

JOINT PROGRAMME TO DEVELOP TESTING METHODS FOR AI-SYSTEMS

FLAGSHIP-PROJECT ZERTIFIZIERTE KI

ZERTIFIZIERTE KI

Qualität sichern. Fortschritt gestalten.

ZERTIFIZIERTE KI

Testing principles	Requirement Assessment	Use Cases	Testing ecosystem	Societal dialogue
▪ Testing scope ▪ Criteria ▪ Depth of testing ▪ Requirements for testing tools ▪ Concept for infrastructure	▪ Customer analysis ▪ Impact analysis ▪ Development of business models	▪ Image recognition ▪ Natural Language Understanding ▪ Natural language processing ▪ Informed Machine Learning	▪ Platform for testing tools ▪ Testing labs ▪ Development of safeguarding methods	▪ Considerations of ethical, legal and philosophical topics ▪ Public events

Broad-based Participation Process

Key Partners:

- Fraunhofer IAIS
- Bundesamt für Sicherheit in der Informationstechnik
- UNIVERSITÄT BONN
- Universität zu Köln
- RWTH AACHEN UNIVERSITY
- DIN

AUTOMATED AI QUALITY ASSESSMENT

04

ASSESSING THE QUALITY OF AI-APPLICATIONS

Quality estimation despite **restricted test coverage**

Quantitative minimal requirements are highly use case specific

Manual tests are **time consuming** or **uncomplete**

Dynamics of AI-systems
(continuous learning -> continuous assessment)

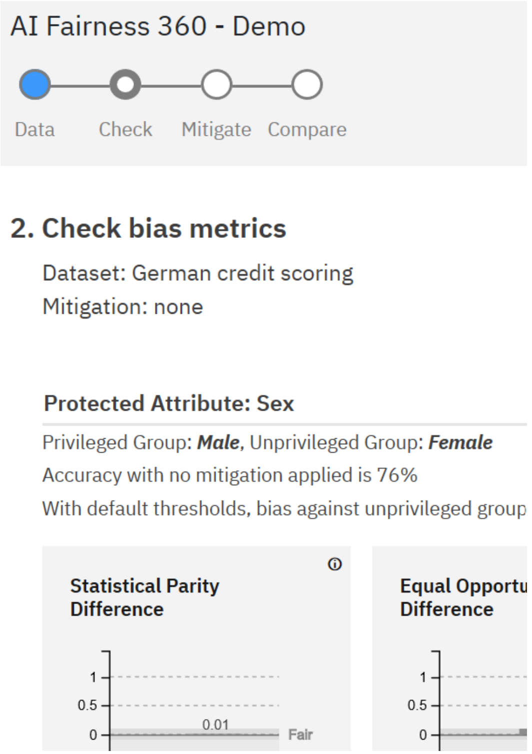
Comprehensibility or explainability for human auditors/ assessors

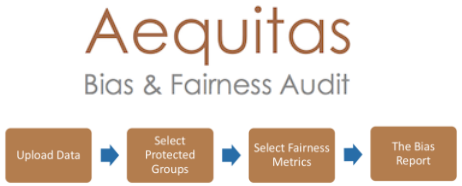
Automation ⟷ Auditability

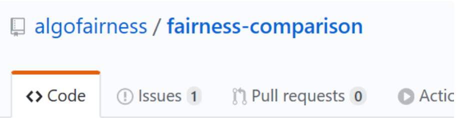
Icon source: https://iconmonstr.com/text-25-png; https://iconmonstr.com/school-7-png; https://iconmonstr.com/networking-7-png; https://iconmonstr.com/gear-11-png; https://iconmonstr.com/search-thin-png

EXAMPLE 1: FAIRNESS-TOOLS FOR GROUP FAIRNESS

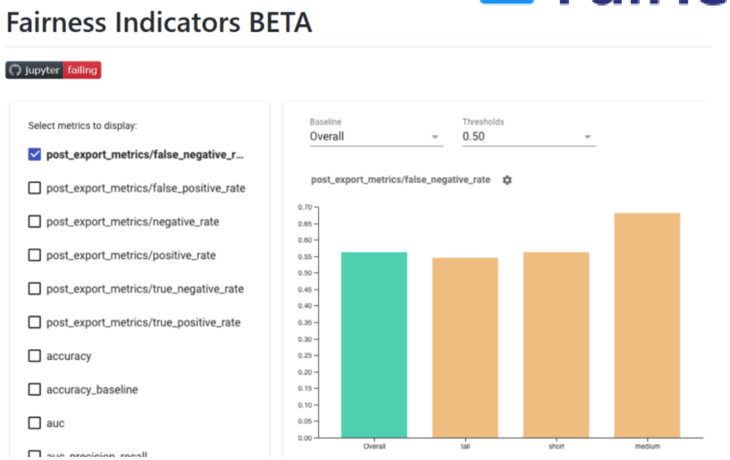
OPEN-SOURCE PACKAGES FOR METRICS, ALGORITHMS AND LONG-TERM SIMULATION





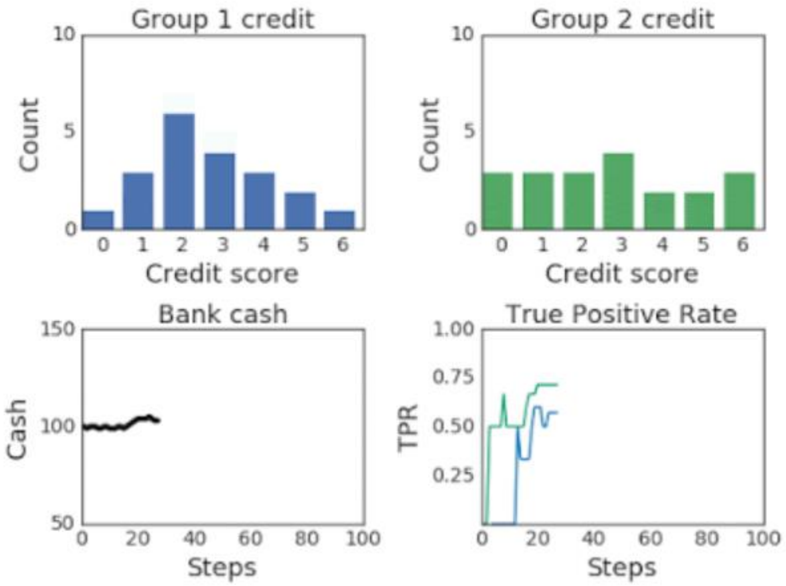


Fairlearn



ML-fairness-gym: A Tool for Exploring Long-Term Impacts of Machine Learning Systems

Wednesday, February 5, 2020

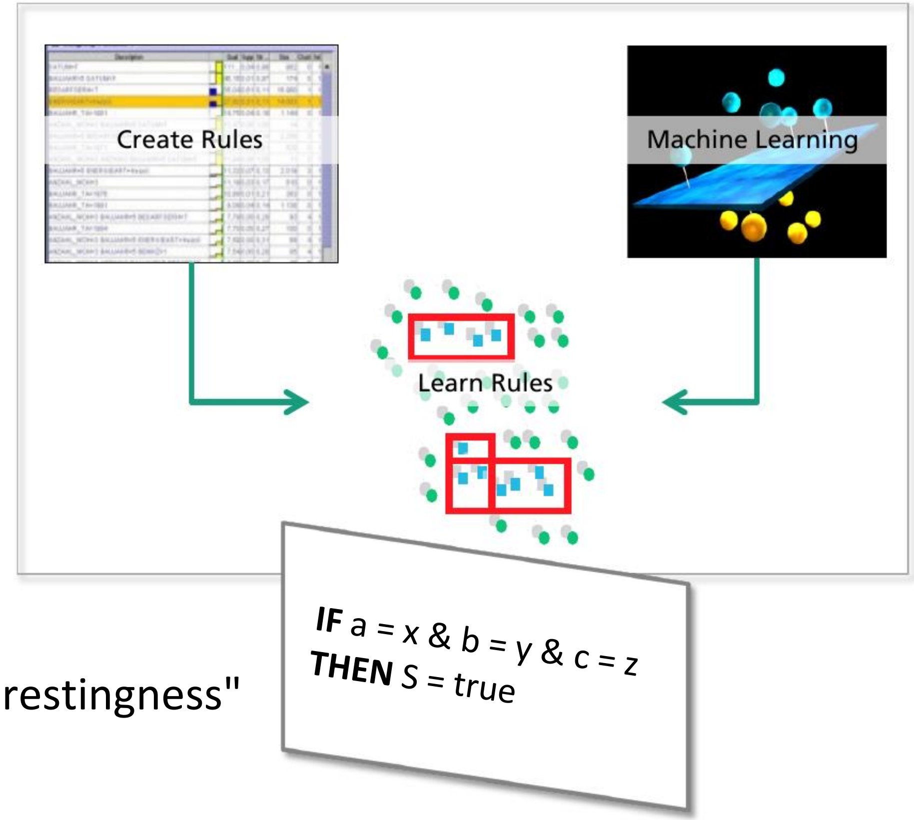


EXAMPLE 2: WEAKNESS ANALYSIS FOR BLACK BOX MODELS (1/2)

AUTOMATED RULE SEARCH FOR ERROR CONDITIONS

Challenges

- Black-box models complex and powerful, but predictions **difficult or impossible to understand**
- **Huge search space** for potential failure modes
- **Reason for failure** in individual cases **difficult to explain** or generalize



Approach

- Subgroup search finds rules for error cases with maximum "interestingness" (size and accuracy) regarding search criterion
- Prerequisite: meaningful **metadata** available or generatable

EXAMPLE 2: WEAKNESS ANALYSIS FOR BLACK BOX MODELS (2/2)

CONTENT ANALYSIS OF CUSTOMER REQUESTS

Use Case

- Model classifies customer requests

Meta data

- Lots of customer data & data concerning claim / request
- Extensible (manually or automatically generated attributes)

Example

- Request not recognized as damage claim, if
  - Subject line is missing
  - Claim request < 10€


```
{
  "Documents": [
    {
      "Title": "De Finibus Bonorum et Malorum",
      "Type": "Article",
      "Author": "Cicero",
      "Date": "45 BC",
      "Keywords": [ "fake", "Latin", "filler", "text" ],
      ...
      "Content": "Lorem ipsum dolor sit amet, consectetur ad
    },
    {
      ...
```

european actuarial academy

KEY TAKE AWAYS

Insurers need to get ready for upcoming regulation and new business opportunities in AI

Therefore a systematic evaluation of AI risks is required

AI risks are use case specific – so are countermeasures

AI quality assessments require new methods and tools

Image sources: Fraunhofer IAIS

EAA e-Conference on
Data Science & Data Ethics

29 June 2021

Contact

Dr. Maximilian Poretschkin

Fraunhofer Institute for Intelligent Analysis and Information Systems

+49 2241 14-1984

maximilian.poretschkin@iais.fraunhofer.de