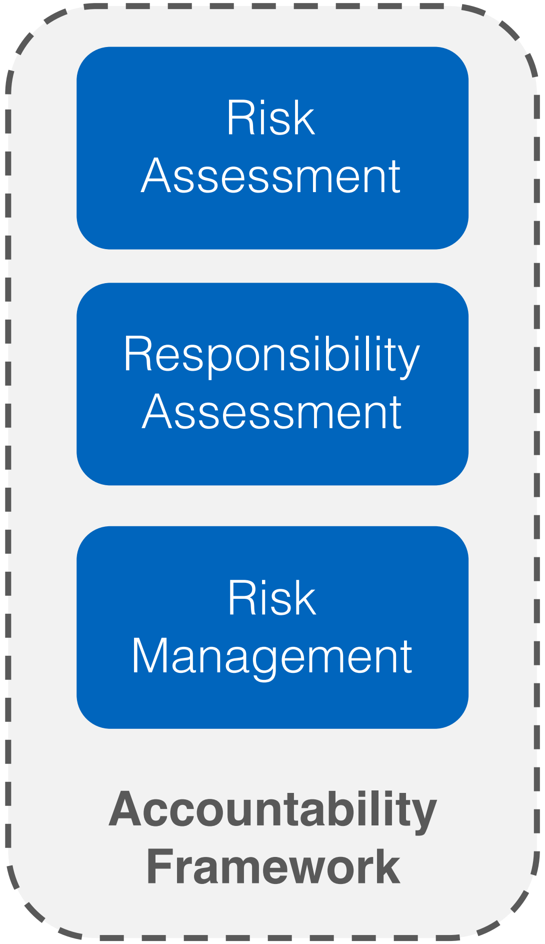
System of AI Accountability in Financial Services

Quantifying AI Ethics Principles

May 2023

Our project

In the joint project between Fujitsu and TUM, we aim at developing a user-centric and practical organizational framework for AI accountability.

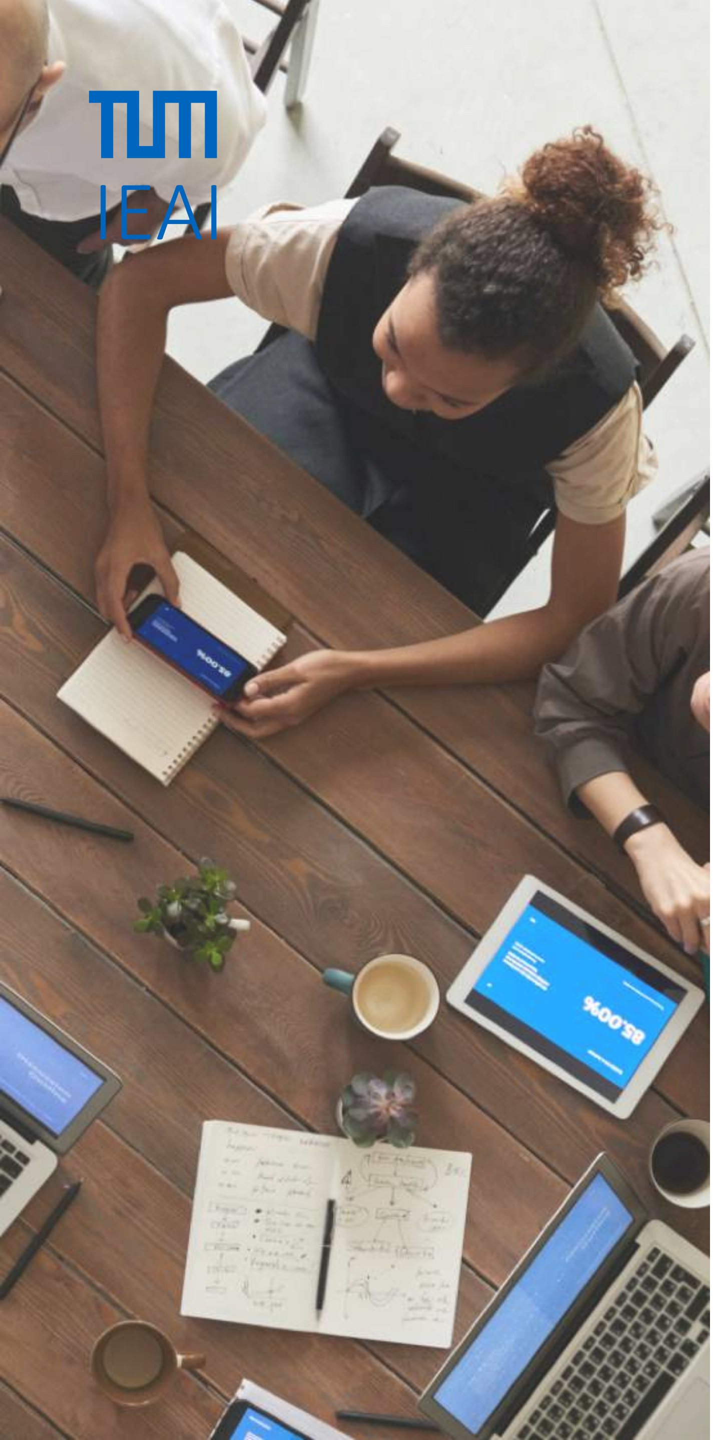


For what is someone accountable and **towards whom**?

Who is accountable?

How can the responsible entity **ensure compliance** with the identified duties?

How can **satisfactory explanation** be given for the measures taken?



Workshop intentions

In our workshop, we want to quantify the degree of adherence to ethicality of AI applications by determining measurable characteristics.

Motivation

In order to effectively manage the ethicality of AI, it's important to establish a **foundation** that allows us to measure and evaluate more **objectively**.

Problem

The concept of AI ethicality as a whole is **not measurable**, but we can decompose it to its **components** and measure those.

Goal

We want to quantify the **degree of adherence to ethicality** of AI applications by …

Objective

… determining **characteristics and criteria** of AI applications to measure the implementation of AI ethics principles.

Preliminaries and Background

IEAI

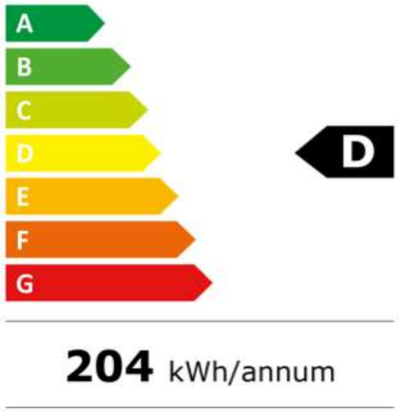
Quantification of abstract concepts

Quantification can give a basis for discussions on how to evaluate elements of abstract concepts and what is detrimental or beneficial.

Motivation behind quantification

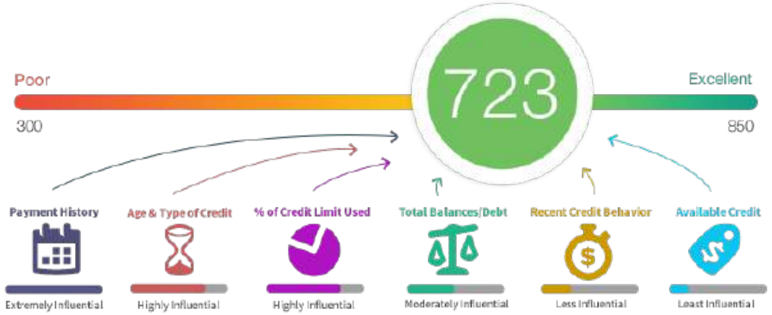
- A number is **more objective** than a perception and can be a basis for discussion
- It can facilitate the **automation** of an evaluation
- It enables **integration** into BI Tools & Dashboards for measurement of corporate strategies

Some examples…



Energy labels can be seen as a numeric criteria for a product's environmental compatibility.

Credit Scores reflect a person's financial trustworthiness.



The fundament of AI ethics

Applying the concept of quantification to ethicality of AI applications, one approach is to measure their adherence to agreed AI ethics principles.

1 **Human agency & oversight**

Ensure fundamental rights, human agency and oversight

2 **Technical robustness & safety**

Ensure resilience to attack & security, fallback-plans & general safety, accuracy and reliability & reproducibility

3 **Privacy & data governance**

Ensure privacy & data protection, quality & integrity of data and access to data

4 **Transparency**

Ensure traceability, explainability and communication

5 **Diversity, non-discrimination & fairness**

Avoid unfair bias, ensure accessibility & universal design and stakeholder participation

6 **Societal and environmental well-being**

Ensure sustainable & environmentally friendly AI, reduce impact on society and democracy

7 **Accountability**

Ensure auditability, minimization and reporting of negative effects, trade-offs and redress

Source: High-Level Expert Group on Artificial Intelligence (2019)

AI ethics principles in finance

Financial, reputational and safety risks make financial companies that use AI even more vulnerable compared to other industries.

1 **Transparency and explainability**

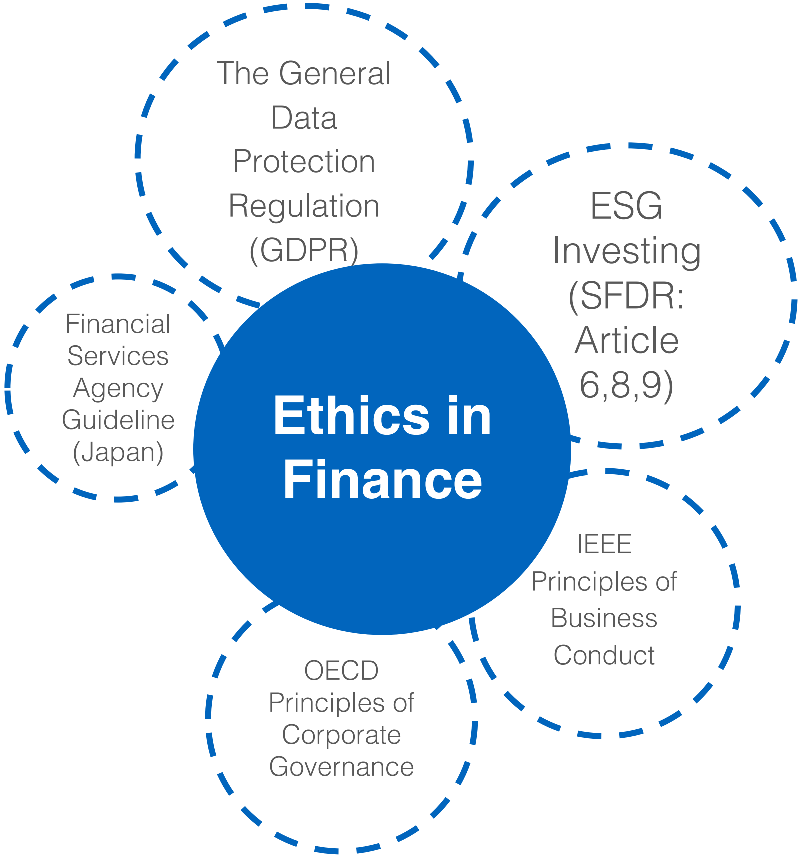
Ensure fundamental rights, human agency and oversight

2 **Fairness & non-discrimination**

Ensure resilience to attack & security, fallback-plans & general safety, accuracy and reliability & reproducibility

3 **Safety & security**

Ensure privacy & data protection, quality & integrity of data and access to data

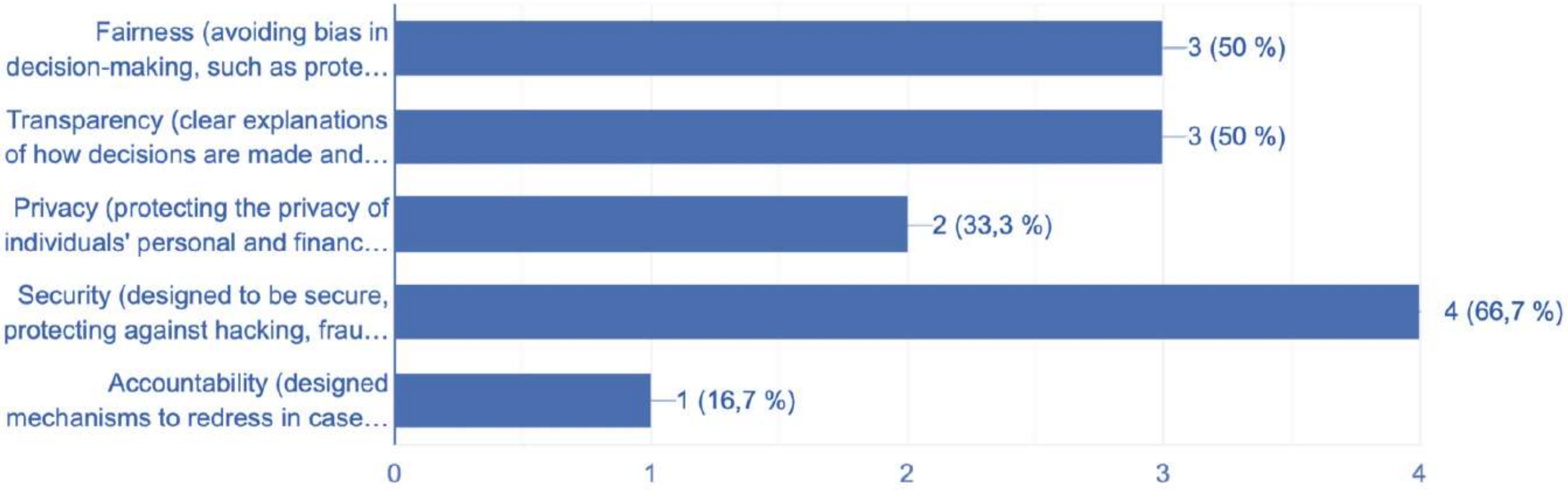

The General Data Protection Regulation (GDPR)
ESG Investing (SFDR: Article 6,8,9)
Financial Services Agency Guideline (Japan)
Ethics in Finance
IEEE Principles of Business Conduct
OECD Principles of Corporate Governance

Quantifying AI ethics

The same three principles have been identified as critical for AI-based finance applications in a short survey with the participants prior to the workshop.

Here are the results from our short survey…

Which two characteristics are most important in AI systems when being applied in the finance industry?

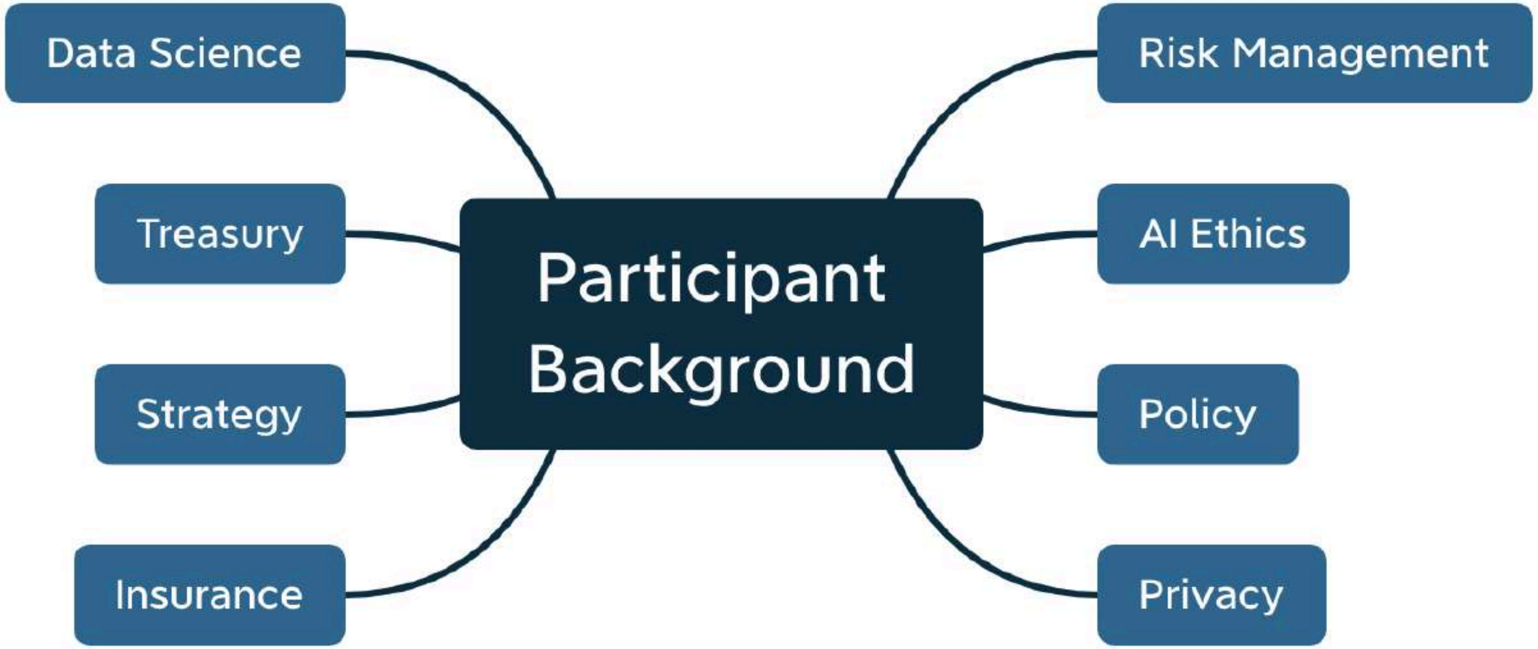


IEAI

Methodology

Participant Background

In total, 13 participants, experts from the finance industry brought different knowledge and perspective to the exercises and discussions.



Participant Background

Data Science

Treasury

Strategy

Insurance

Risk Management

AI Ethics

Policy

Privacy

Workshop methodology

An Expert Workshop methodology was used that requires to simplify the complex phenomena by agreeing on contextual, narrow definitions that can be tested.

Workshop aim

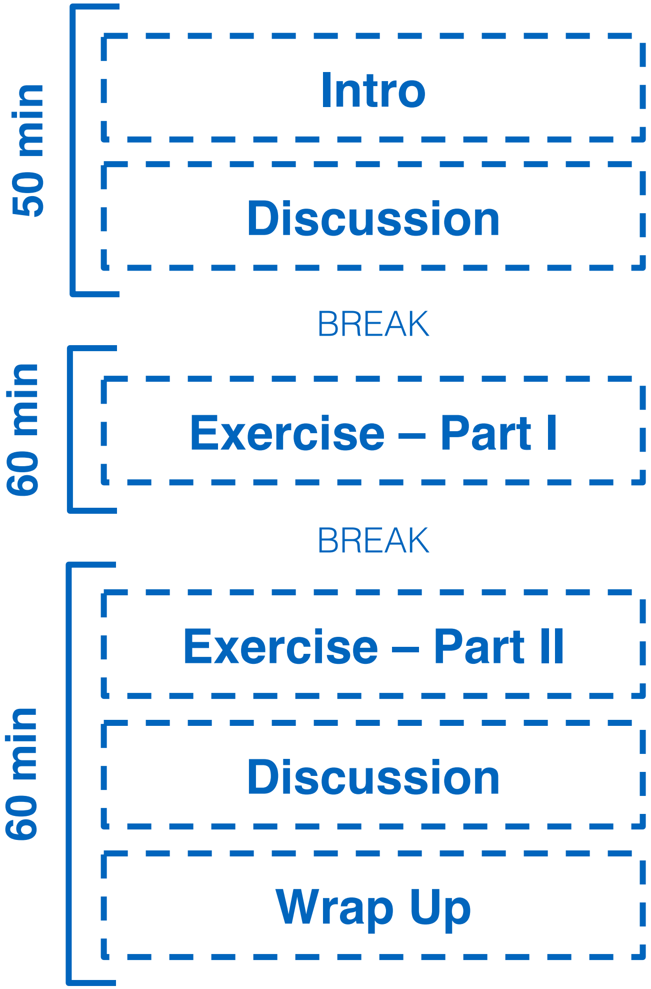
We want to quantify the degree of adherence to ethicality of AI applications by determining criteria of AI applications to measure the implementation of AI ethics principles.

Our Procedure

1. Create contextual definitions about a phenomena in order to **simplify** it
2. Analyze and synthesize the elements of the phenomena, here called **characteristics**, using the opinion of experts in individual and team work
3. Based on expert opinion, assign **numeric values** to the characteristics using and creating a **scale** for each of the found elements
4. **Quantifying generalized perceptive assessment** of the degree of adherence to ethical principles of an AI application

Workshop agenda

The goal of the workshop was to quantify the degree of adherence to ethicality of AI applications.



50 min

Intro

Introduction from TUM & **presentation** of project, workshop goals and deliberating on the assumptions

Discussion

Small **exercise** and **survey** on the degree of ethicality in System of AI Accountability in finance today

BREAK

60 min

Exercise – Part I

Brainstorm on the (quantifiable) characteristics of a financial application in **smaller groups**

BREAK

60 min

Exercise – Part II

Assigning numeric values to the determined characteristics in **plenum**

Discussion

Discussion of obstacles for ethicality measurement and recommendations to overcome them in **plenum**

Wrap Up

Wrap-up of workshop and transition to **networking and lunch** until ~14^{00}

Use case

In the workshop, a predefined use case from the finance industry was used as an example for the determination of characteristics and assignment of optimal values.

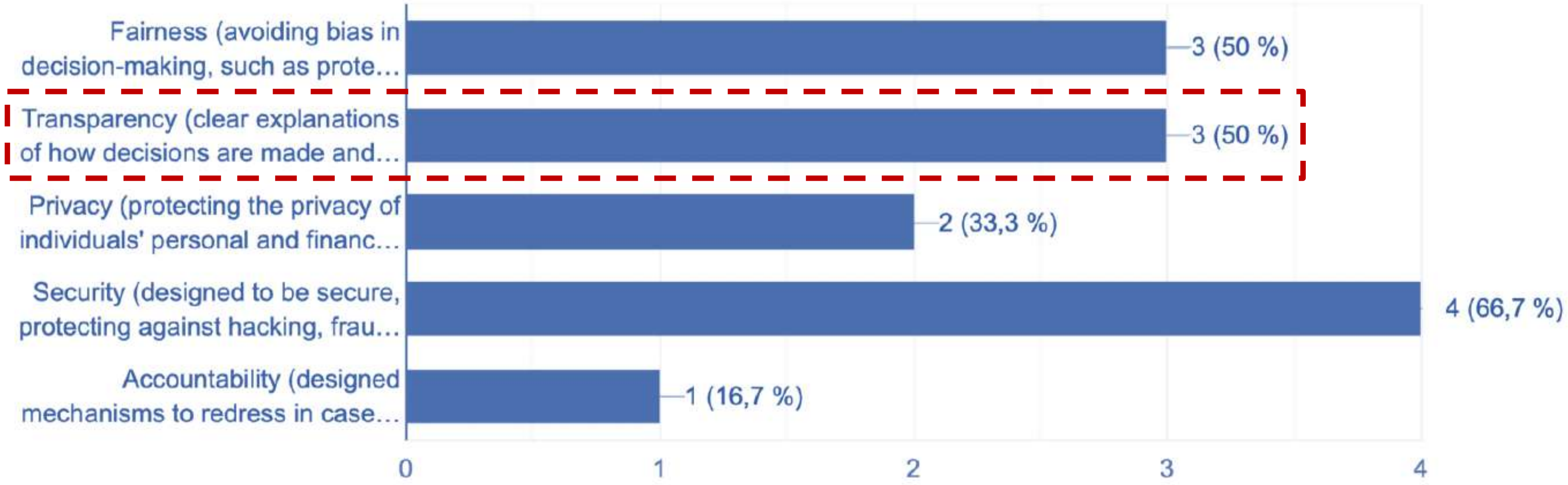
An **AI-based Credit Scoring (CS) tool** developed and tested in the **EU** and based on **Neural Networks** models (making it quite obscure) is put on the market. The company proposing it claims that their tool expands access to capital and financial services for **marginalized communities** and uses both **financial and non-specified alternative data** for decision-making when the client gives a **consent** to disclose its data, as required to comply with GDPR.

Quantifying AI ethics

The proposed concept of AI ethics quantification was tested within the workshop on one AI ethics principle as an example.

Here are the results from our short survey…

Which two characteristics are most important in AI systems when being applied in the finance industry?



Step 1 – Contextual Definitions

TUM IEAI

Assumptions and Definitions

To simplify the complex phenomena of quantifying the abstract concept of ethicality of an AI application, contextual definitions and assumptions need to be agreed.

Assumed conditions

1. An AI application is perceived ethical if it adheres to defined ethical principles of AI.
2. We can measure AI ethics by quantifying characteristics defining the ethicality of an AI application.
3. We can quantify the ethicality of an AI application by measuring its adherence to the given characteristics of the AI ethics principles.

Assumptions and Definitions

To simplify the complex phenomena of quantifying the abstract concept of ethicality of an AI application, contextual definitions and assumptions need to be agreed.

Ethicality of an AI application = the AI product's degree of adherence to the 7 key requirements for trustworthy AI as defined by the AI-HLEG.

Transparency = users of an AI system should understand how they are assessed and how the AI algorithm arrived at a prediction.

Explainability = the AI algorithm should be able to provide a clear explanation of how it arrived at a prediction, including the factors that were considered and how they were weighted.

By implementing these concepts into their policies companies can build customer trust and confidence in their AI-based processes.

Hypothesis

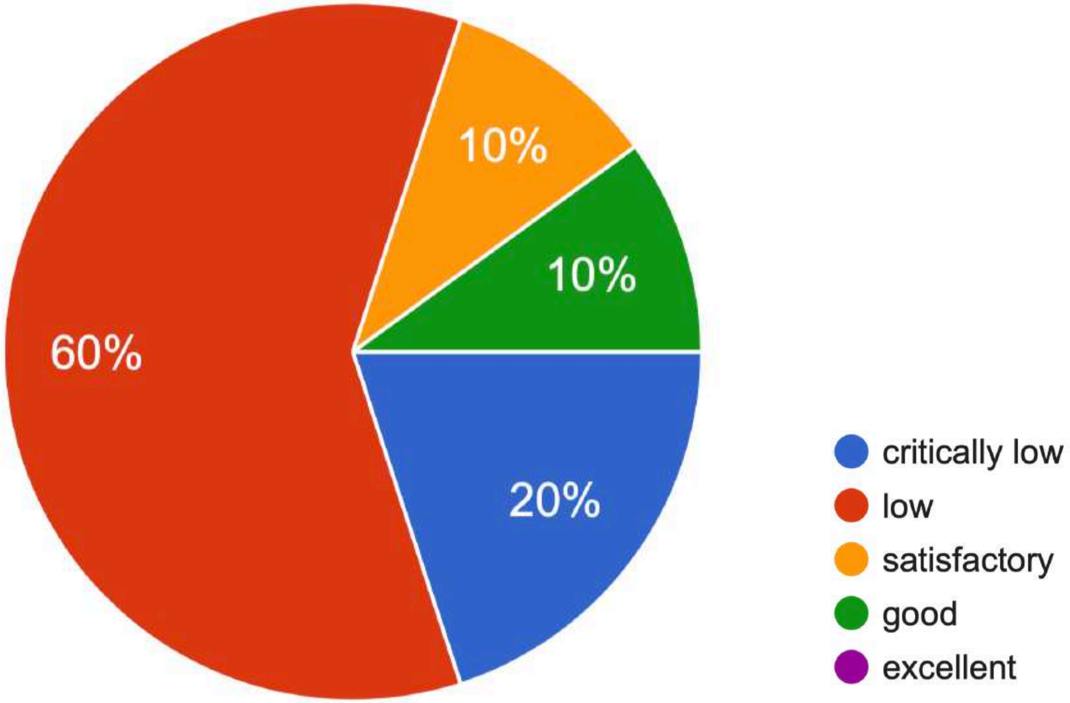
Based on the agreed terminology, the hypothesis on how to quantify the ethicality of an AI application was formulated as follows.

We assume that it is possible to measure the ethicality of an AI application by measuring a set of representative characteristics of the AI application. The characteristics should reflect the degree of adherence to the ethical principle of transparency.

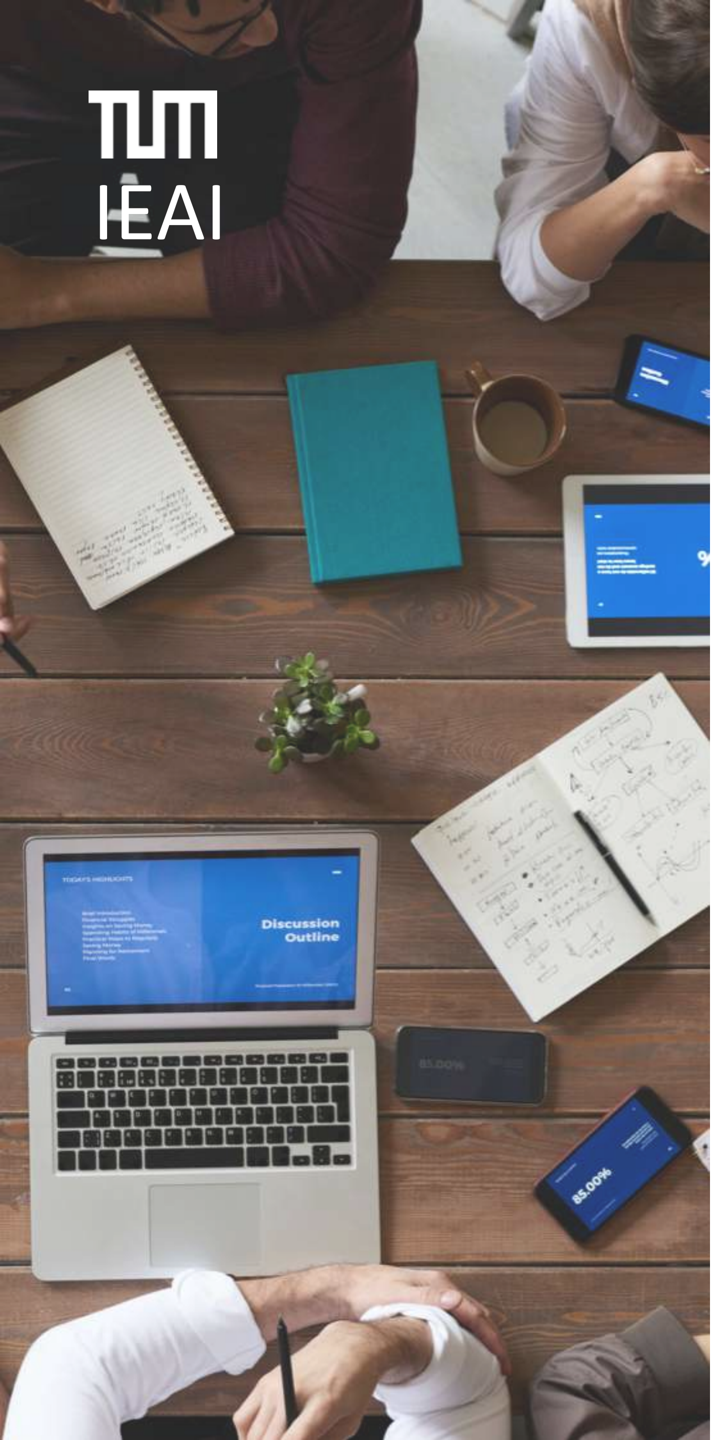
Intuitive assessment

To compare the participants’ intuitive perception to their quantified assessment, their opinion on the ethicality of common credit scoring applications was assessed.

What do you think is the current state of AI Credit Scoring adherence to the ethical principle of transparency?



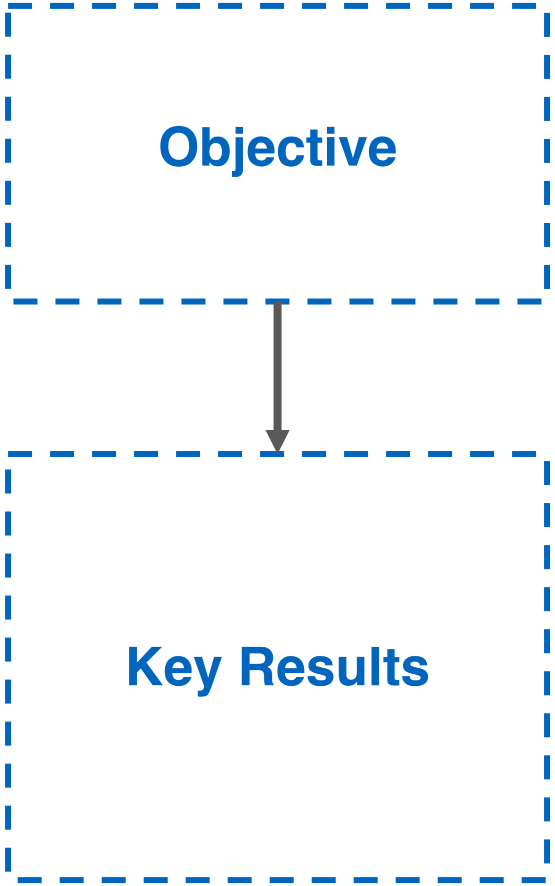
Step 2 – Quantifiable Characteristics



TUM IEAI

Quantification procedure

The procedure to find quantifiable characteristics was done analogous to a common management practice, called OKRs.



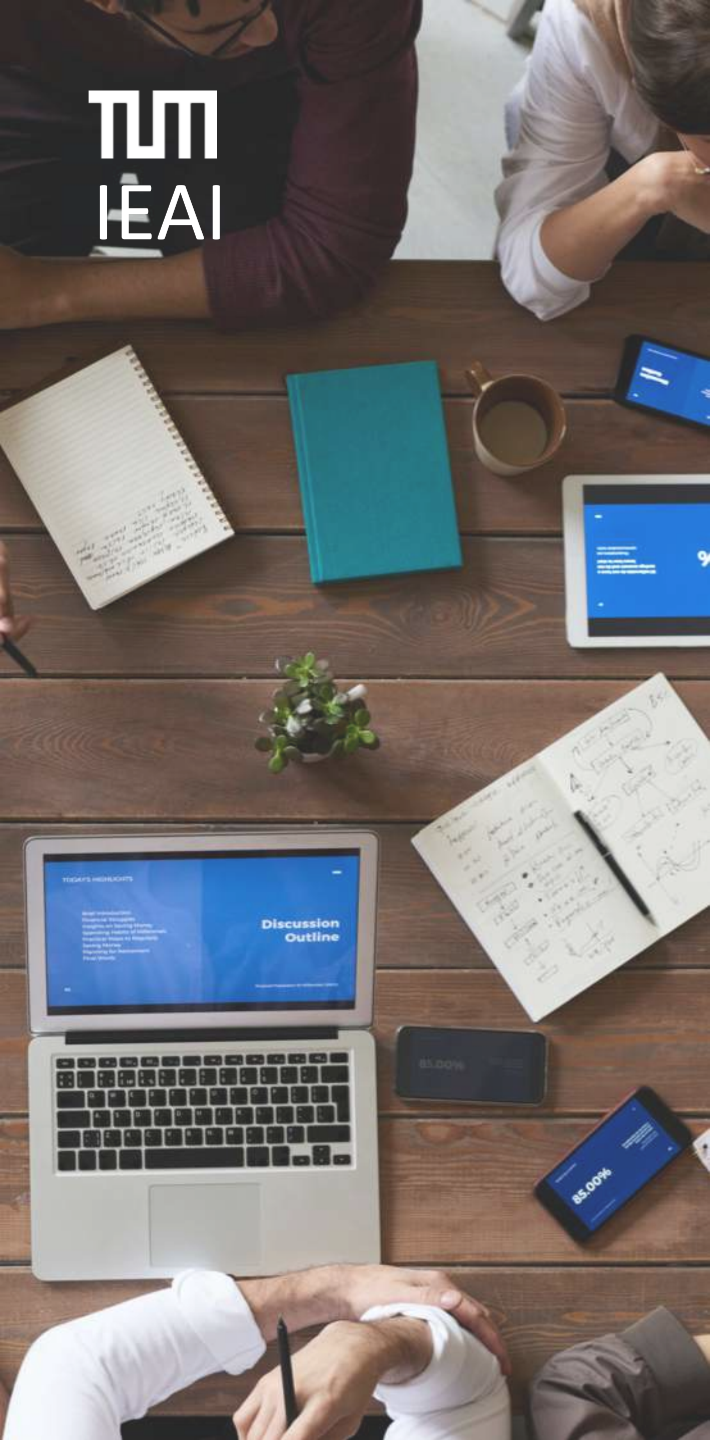
The „What“

An Objective is what you want to do. It describes your **mission-supporting goal** and sets a deadline for achieving it.

The „How“

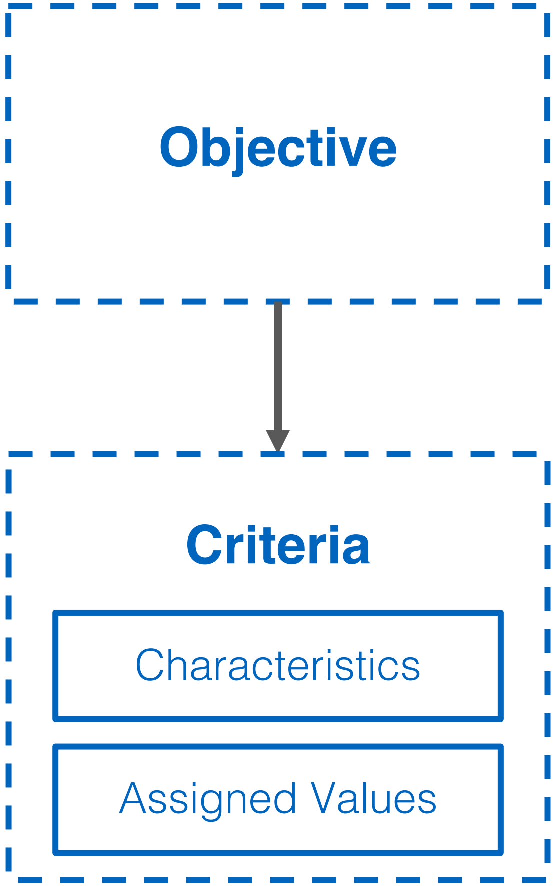
Objectives must be paired with a roadmap that will help you know whether or not you’re on the path to meeting your goals. Key Results are the **benchmarks you can measure that track your progress** toward the Objective.

Source: whatmatters.com



Quantification procedure

While OKRs focus on short-term goal-reaching, the proposed procedure determines characteristics and values to enable continuous measurement of the resulting criteria.



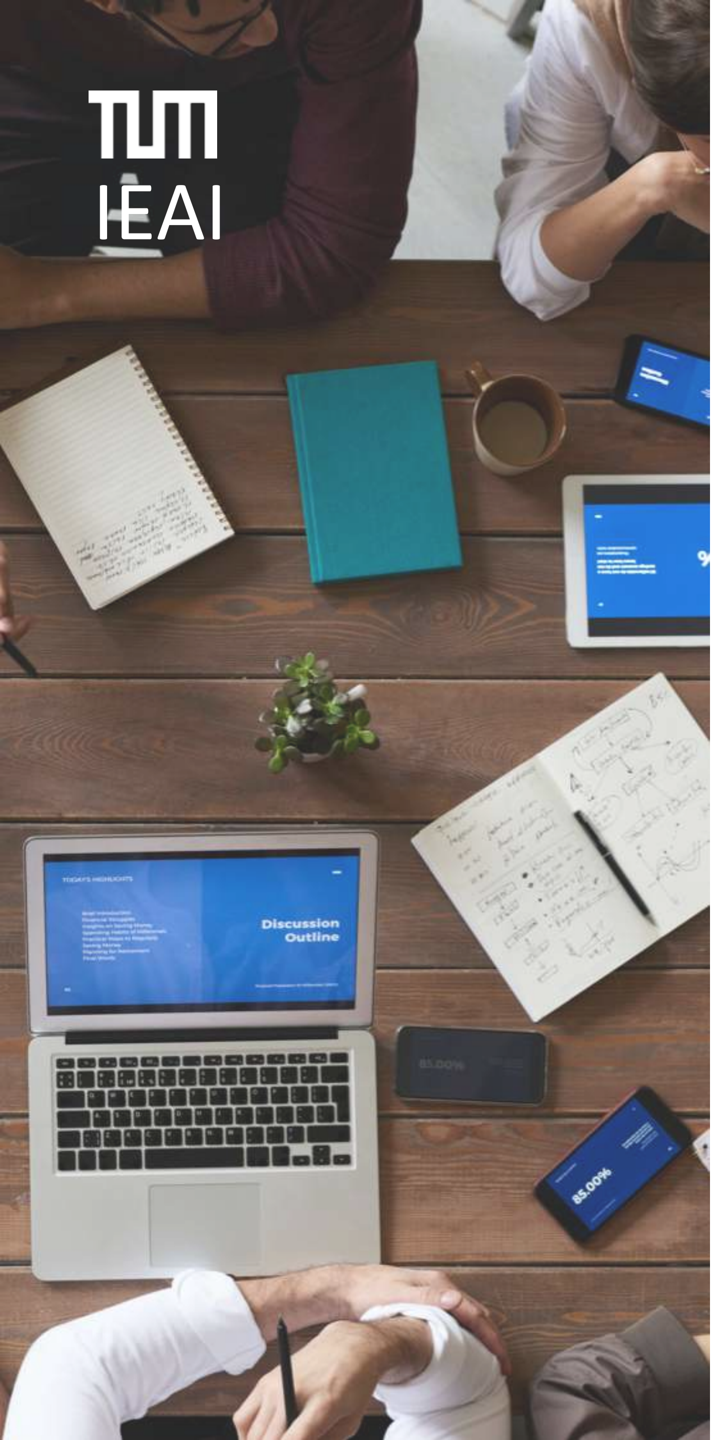
The „What“

To measure the ethicality of an AI application, we set **adherence to defined AI ethics principles** as the underlying objective.

The „How“

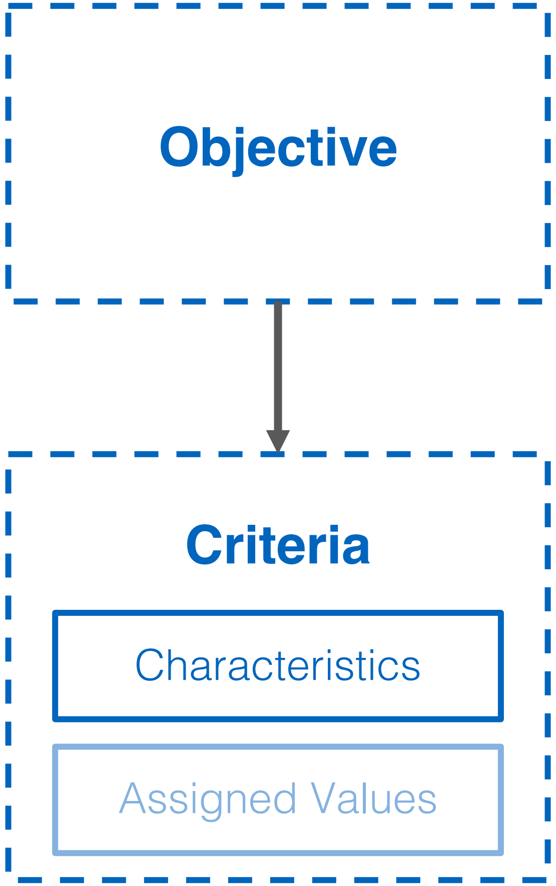
Criteria are used to measure adherence to the defined principles. They consist of **measurable characteristics** of AI applications as well as predefined **values that reflects the optimal state**.

Source: whatmatters.com



Exercise – Part I

In Exercise Part I, we tested for a defined use case and a given ethical principle the determination of characteristics to quantify ethicality.



Objective

Criteria

Characteristics

Assigned Values

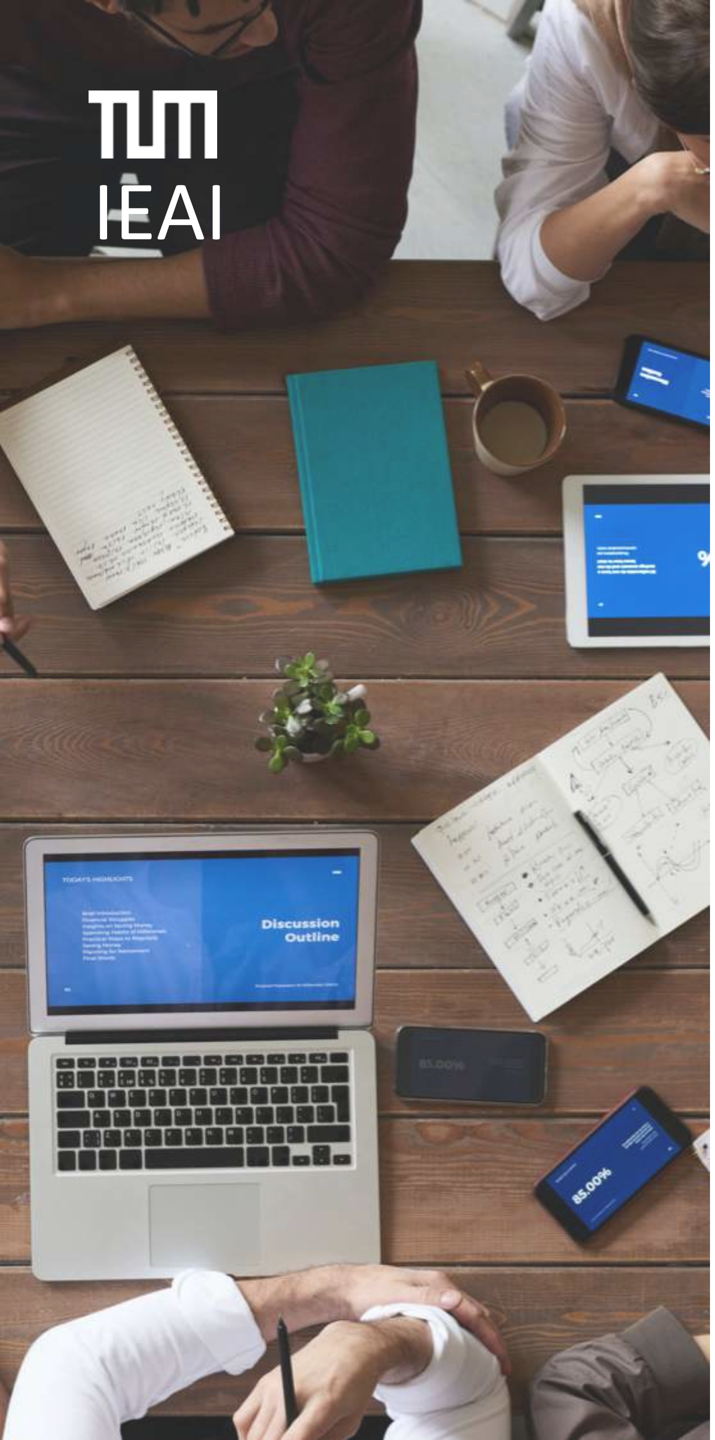
The „What“

Measure an AI applications adherence to **transparency** and **explainability**.

The „How“

Exercise – Part I

Exercise – Part II

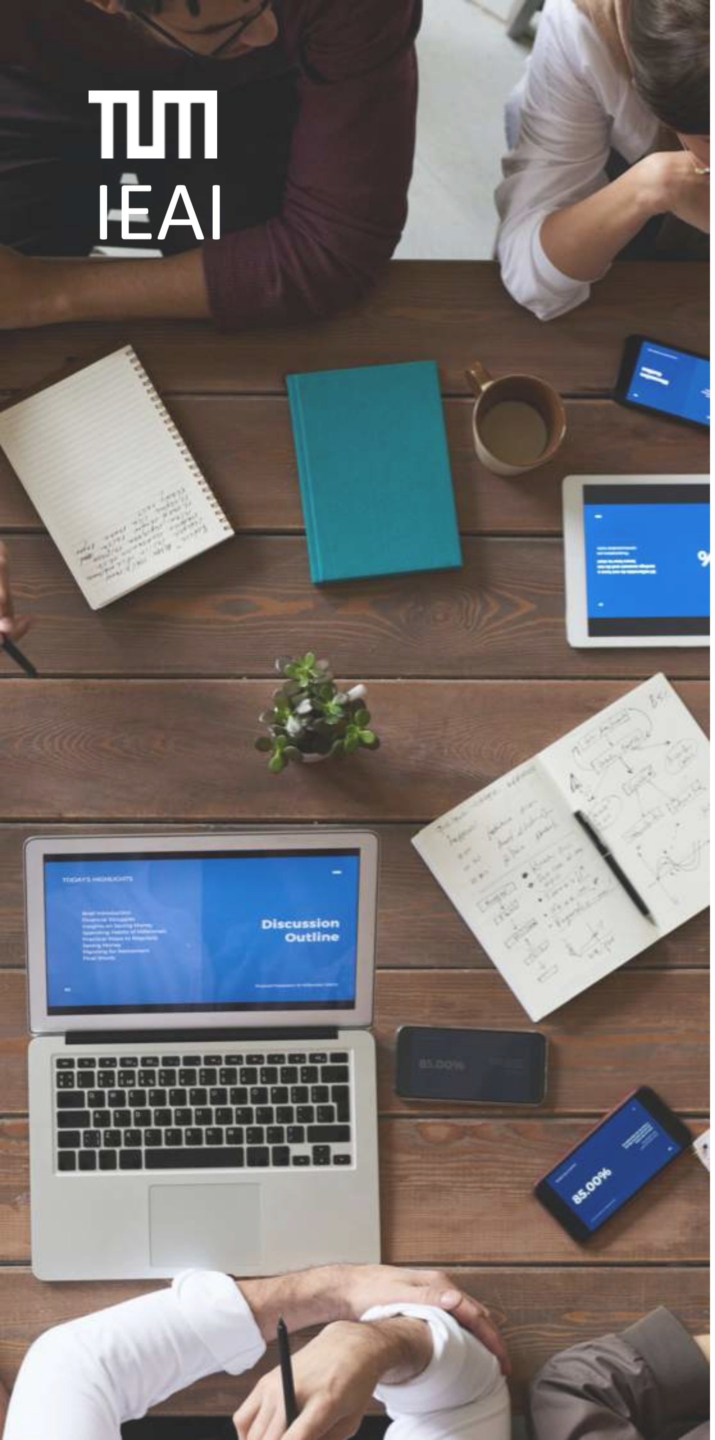


Exercise – Part I

Participants were asked to form 3 groups and brainstorm on quantifiable characteristics that reflect the transparency of the use case application.

Some inputs for characteristics:

- **Direct** - Direct characteristics relate directly to AI Credit Scoring, indirect characteristics- to other objects (indicators), i.e. these are direct signs of other objects.
- **Relative** - Characteristic that can be quantified only as a share of a bigger entity (Share of false decisions out of all decisions, share of mathematical classes out of all classes, share of sugar in a kg of cake etc.)
- **Absolute** - Should be specific, i.e. (time spent per decision; amount of euros per person in a year)



Exercise – Part I: Outcome

Each group presented 5 quantifiable characteristics that reflect the transparency of the use case application.

Group 1

1 Share of relevant data points that were used in a decision-making of AI CS that was disclosed and explained to the customer

2 Share of AI CS decisions that was reviewed by a credit analysis' domain expert

3 Share of reviewed decisions by a AI CS, explanations on which were found satisfactory by a domain expert

4 Share of predictions correctly explained by a local interpretation method

5 Share of complaints/incidents asked on a AI CS decision after a customer asked for clarification on his/her decision

Group 2

1 Weight of data source and type

2 Share of cases where human intervention was needed

3 Share of (sensitive) features used

4 Model metrics (accuracy, confidence level, fairness metrics)

5 Number of different data sources / share of trustworthy data sources

Group 3

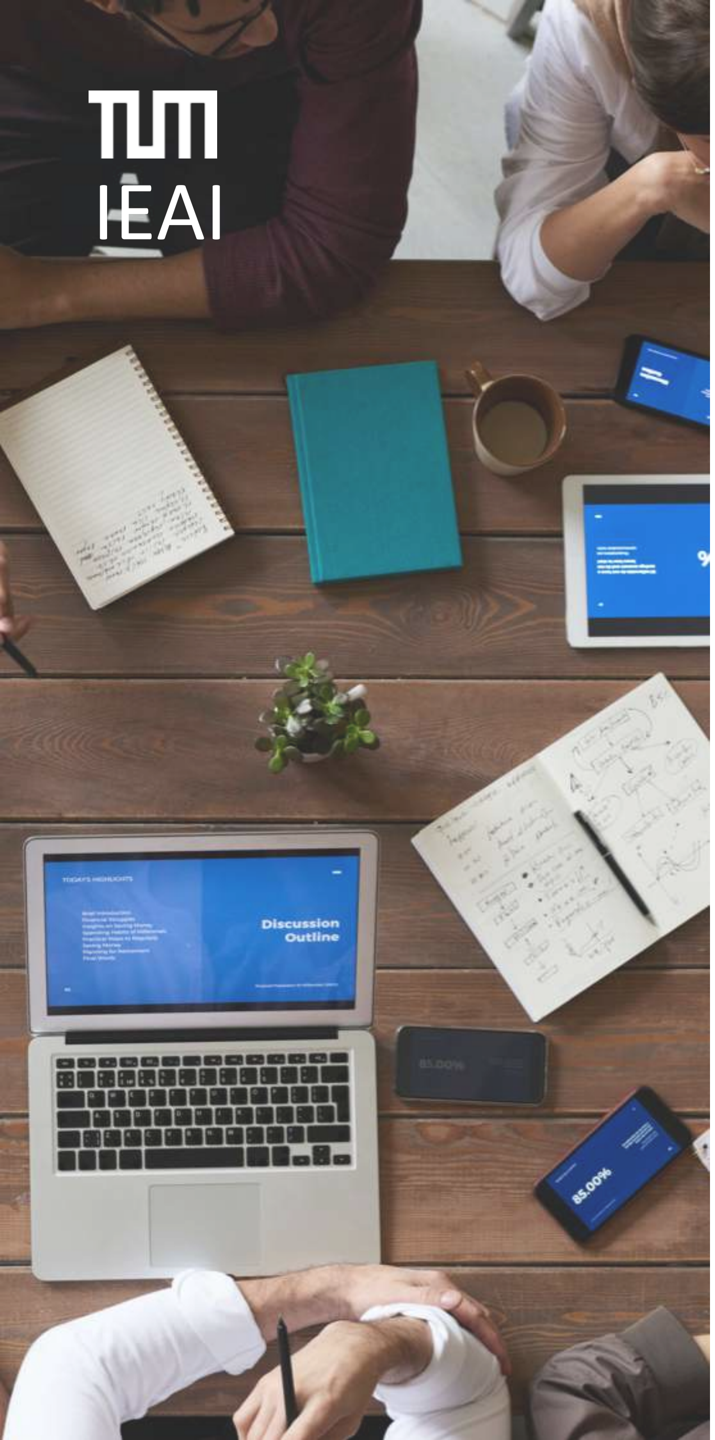
1 Share of documentation of relevant steps in the AI tool lifecycle (defined by standards and including post-hoc adjustments)

2 Share of cases for which output is reproducible within acceptable standards (defined by standards)

3 Share of group of users (reporting) understanding of the tool (UX research)

4 Share of known potential limitations presented to the public

5 Share of information about the system that is publically available (based on internal documentation)



Exercise – Part I: Outcome

Of the resulting 15 characteristics, 5 were chosen in a discussion as the most representative ones.

5 characteristics

1 Share of relevant features that are involved in the AI CS decision that were disclosed and explained to the customers

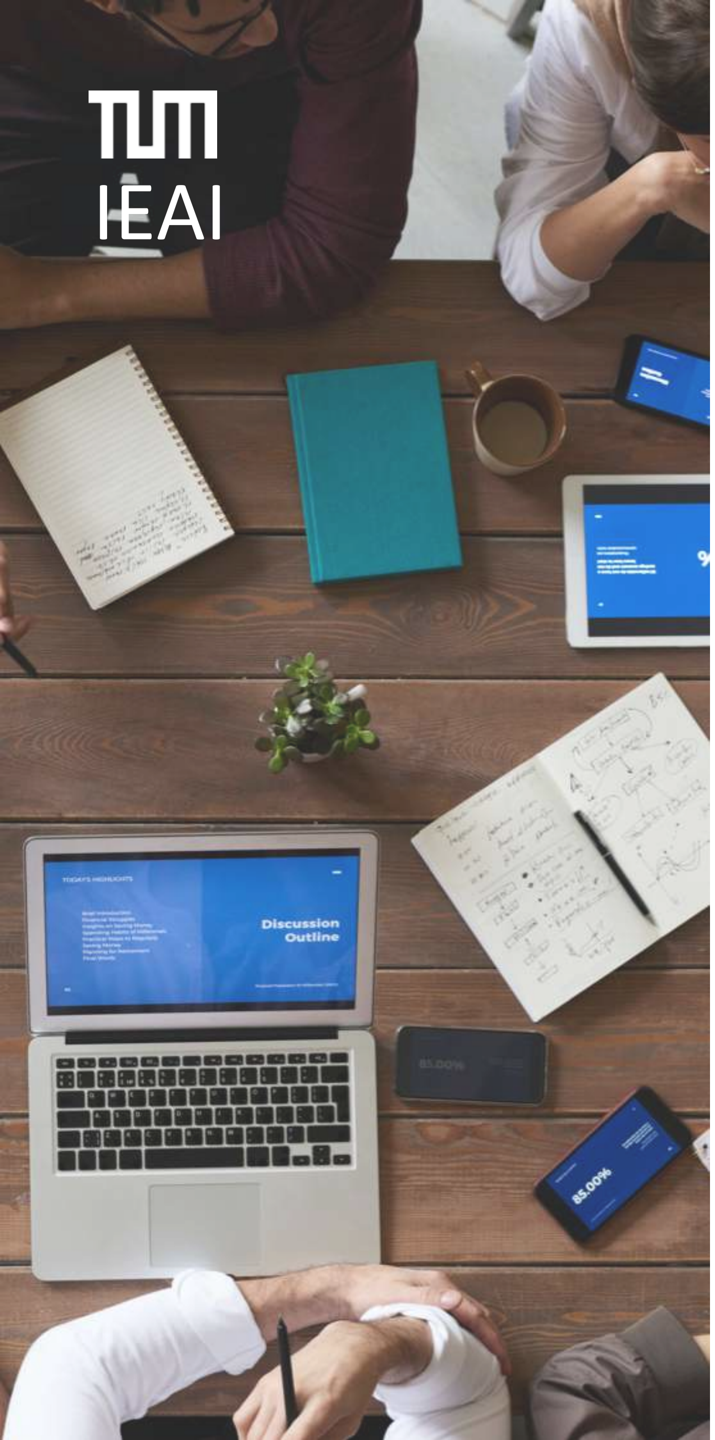
2 Share of relevant data that comes from trustworthy data sources

3 Share of prediction performance metrics and limitations correctly explained to the target group

4 Ratio of inquiries on AI CS relating to understandability

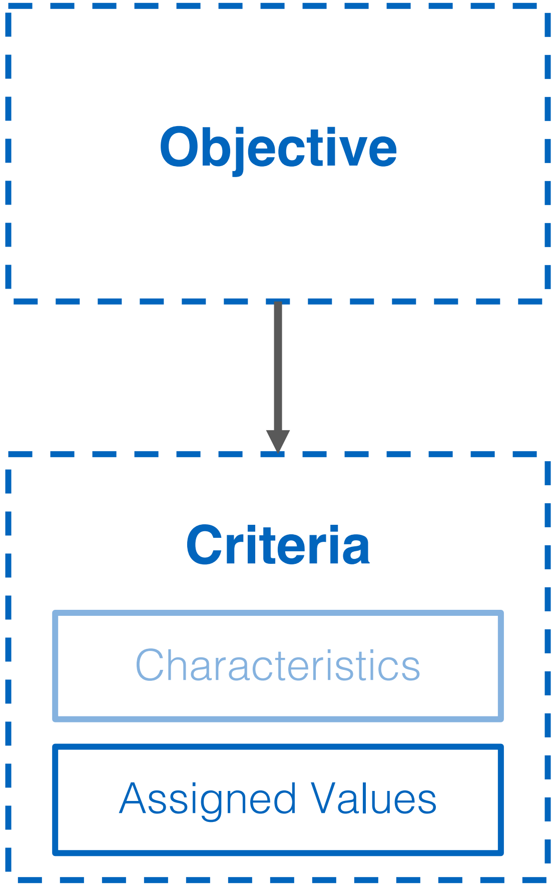
5 Share of AI CS decisions that were reviewed by a domain expert (credit analyst)

Step 3 – Numeric Values



Exercise – Part II

In Exercise Part II, optimal values and ranges for the resulting 5 characteristics were defined by the participants.



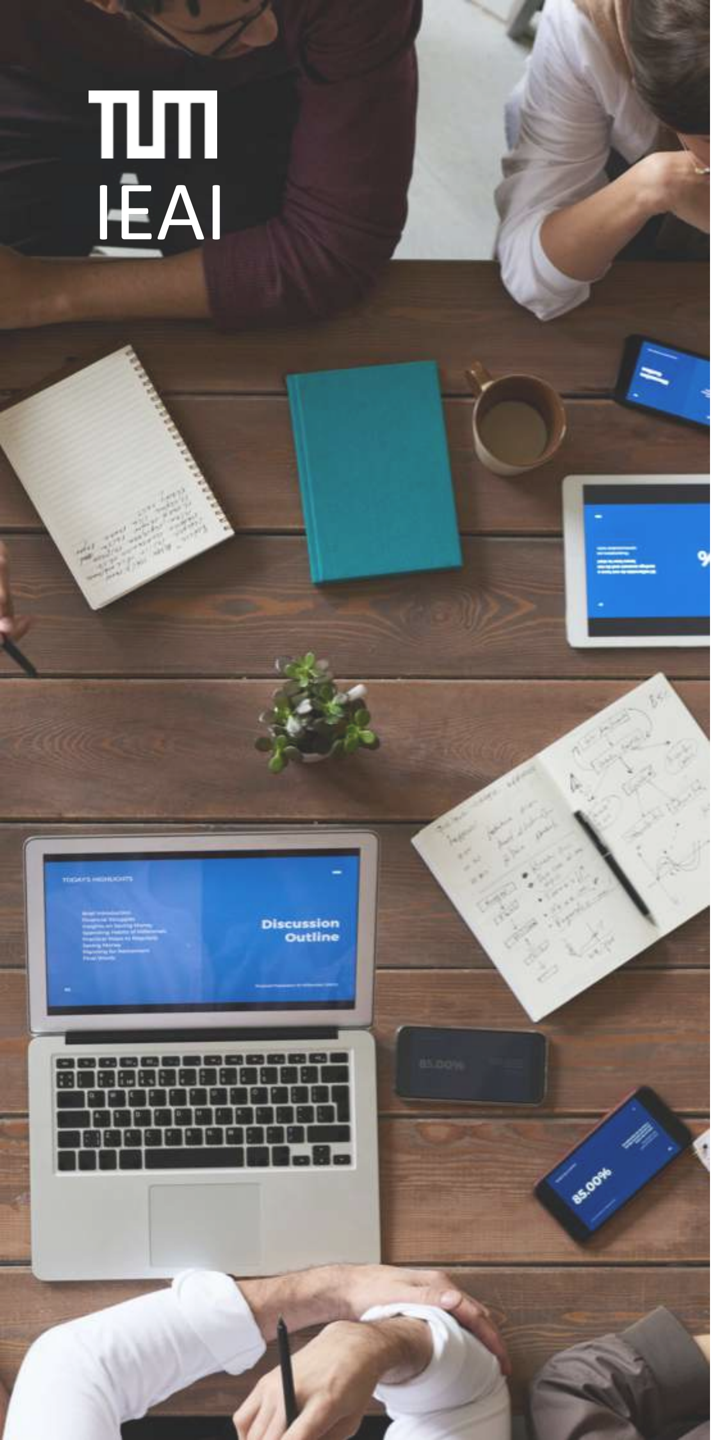
The „What“

Measure an AI applications adherence to **transparency** and **explainability**.

The „How“

Exercise – Part I

Exercise – Part II



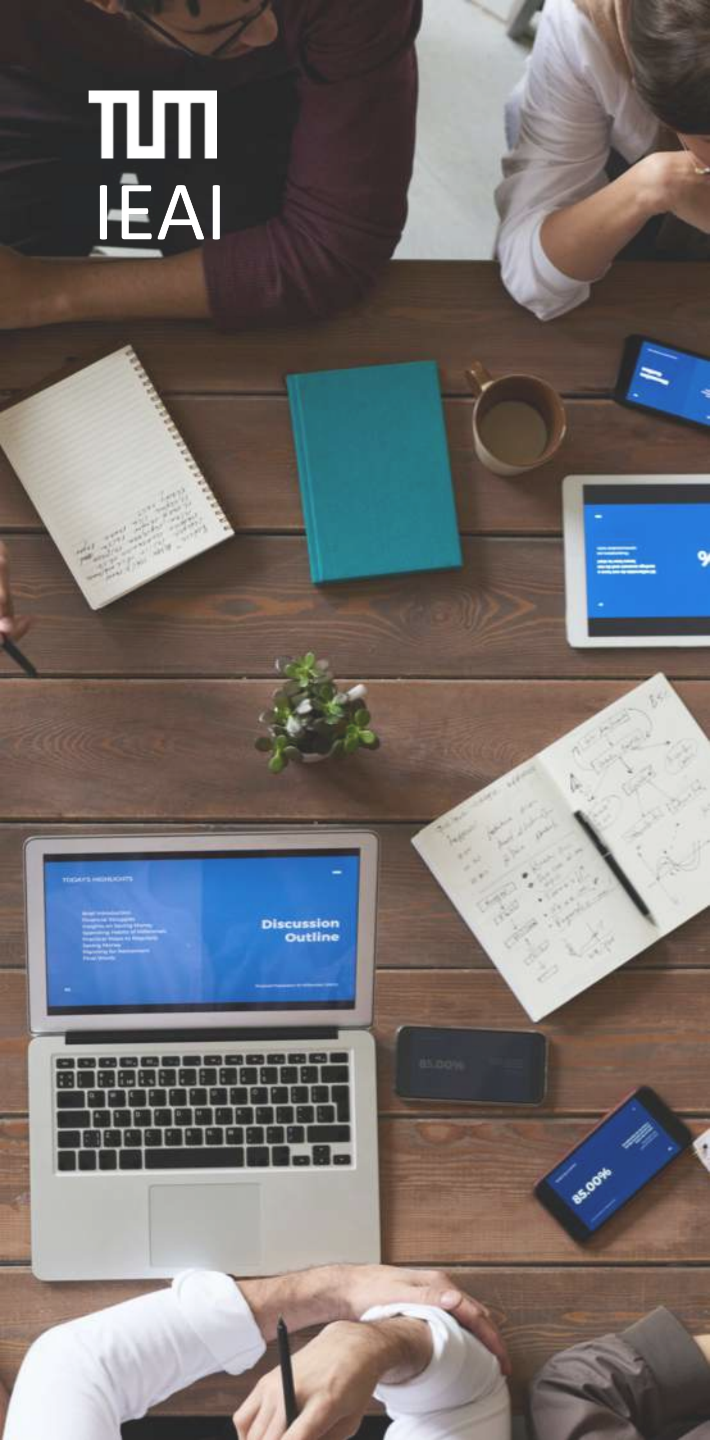
Quantification Matrix

Participants were asked to fill in values in the quantification matrix to assess the scale and current state of the determined characteristics.

Ratio of Importance – relative contribution of a characteristic to the overall score	Characteristics	scale				
		critically low	low	satisfactory	good	excellent
	(1) Share of relevant features that are involved in the AI CS decision that were disclosed and explained to the customers (2) Share of relevant data that comes from trustworthy data sources (3) Share of prediction performance metrics and limitations correctly explained to the target group (4) Ratio of inquires on AI CS relating to understandability (5) Share of AI CS decisions that were reviewed by a domain expert (credit analyst)					

TUM IEAI

Step 4 – Quantified Assessment



Quantification Matrix

The participants' answers were evaluated to create a generalized scale to rate the adherence of credit scoring tools to the principles of transparency & explainability.

Ratio of importance	Assessment of the state – Characteristics	critically low	low	satisfac-tory	good	excellent
0,27	(1) Share of relevant features that are involved in the AI CS decision that were disclosed and explained to the customers	0,2	0,4	0,5	0,6	0,8
0,25	(2) Share of relevant data that comes from trustworthy data sources	0,3	0,5	0,6	0,7	0,9
0,18	(3) Share of prediction performance metrics and limitations correctly explained to the target group	0,34	0,43	0,52	0,62	0,77
0,13	(4) Ratio of inquiries on AI CS relating to understandability	0,4	0,5	0,7	0,8	0,9
0,17	(5) Share of AI CS decisions that were reviewed by a domain expert (credit analyst)	0,0	0,1	0,1	0,2	0,3
	Generalized Scale	**0,26**	**0,36**	**0,48**	**0,61**	**0,75**

Conclusion

Conclusion

The analysis has revealed insights regarding the quantification of the degree of adherence to ethicality of AI applications.

Outcomes

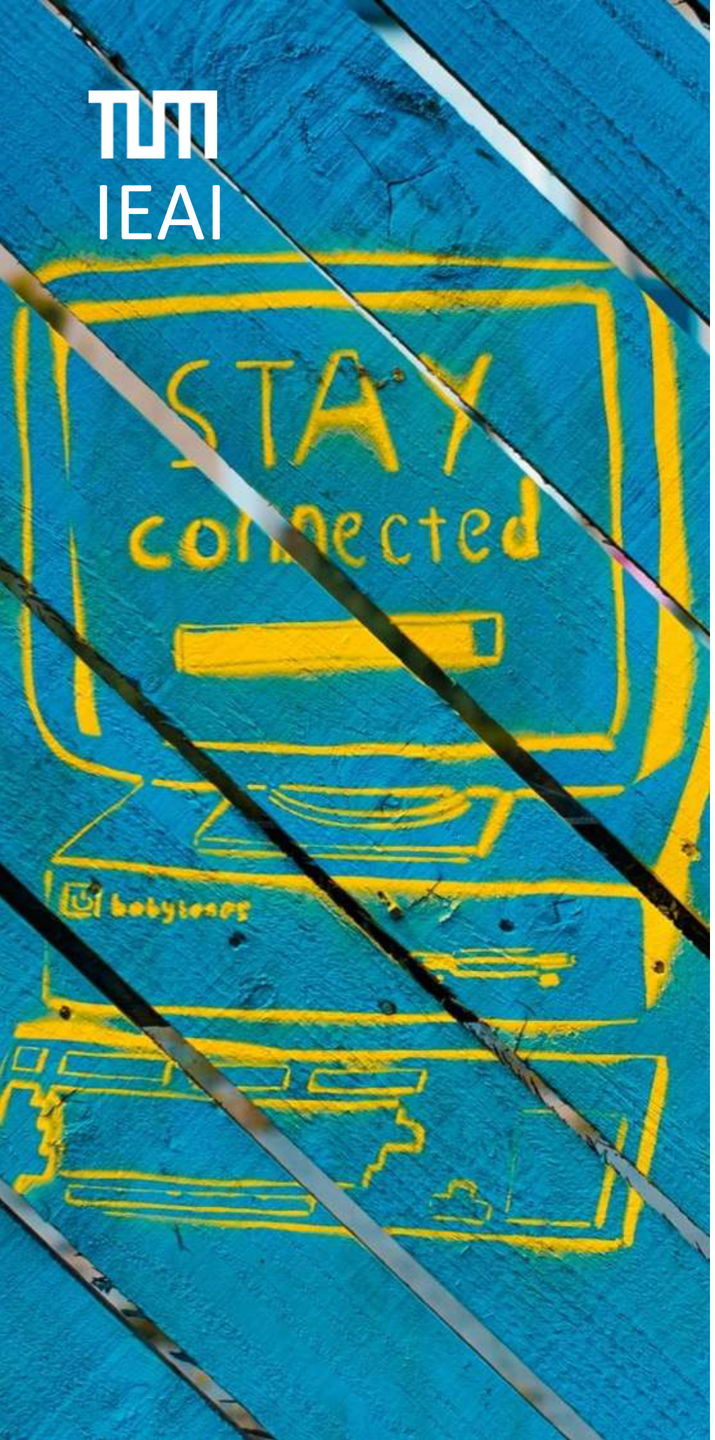
- Five distinct **characteristics** that exemplify compliance with the principle of transparency &explainability
- Quantifiable scale to assess the extent of implementation of each of these characteristics

Outlook

- Continued **evaluation and refinement** of the defined characteristics is needed to develop a **comprehensive framework** for assessing the ethicality of AI applications in chosen sectors and use cases

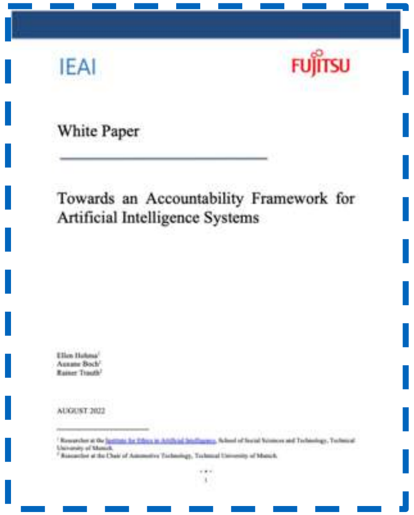
Findings

- Intuitive assessment from expert participants revealed a **strong lack of transparency and explainability** of current AI Credit Scoring tools
- The need to **develop clear scalable characteristics** to evaluate at which level of ethicality in a given context a tool is has been confirmed
- Our methodology can propose a first step towards a solution in **systematically evaluating the ethicality of AI technologies** by developing clear scalable and context-dependent characteristics



Stay connected!

If you are interested in the topic, you can find additional information, material and future updates on our project webpage, or reach out to us.



IEAI FUJITSU

White Paper

Towards an Accountability Framework for Artificial Intelligence Systems

Ellen Hohma[1]
Auxane Boch[1]
Rainer Trauth[2]

AUGUST 2022

[1] Researcher at the Institute for Ethics in Artificial Intelligence, School of Social Sciences and Technology, Technical University of Munich.
[2] Researcher at the Chair of Automotive Technology, Technical University of Munich.

1

frontiers

Investigating Accountability for Artificial Intelligence through Risk Governance: A Workshop-based Exploratory Study

Ellen Hohma[1*], Auxane Boch[1], Rainer Trauth[2], Christoph Lütge[1]

[1]Institute for Ethics in AI, School of Social Sciences and Technology, Technical University of Munich, Munich, Germany

[2]Chair of Automotive Technology, School of Engineering and Design, Technical University of Munich, Munich, Germany

*** Correspondence:**
Ellen Hohma
ellen.hohma@tum.de

Keywords: Accountability₁, artificial intelligence₂, organizational framework₃, risk governance₄, workshop₅

Number of words: 9.926

Number of figures and tables: 13

Abstract

With the growing prevalance of AI-based systems and the development of specific regulations in response, accountability for consequences resulting from the development or use of these technologies becomes increasingly important. However, concrete strategies and approaches of solving related challenges seem to not have been suitably developed for or communicated with AI practitioners. Studying how risk governance methods can be used to administer AI accountability, we aim at contributing to closing this gap. We chose an exploratory workshop-based methodology to investigate current challenges in accountability and risk management approaches raised by AI practitioners from academia and industry. Our interactive study design revealed various insights on which aspects do or do not work for handling risks of AI in practice. From the gathered perspectives, we derived required characteristics for AI risk management methodologies and determined demands for clarification and action in the strive to move AI accountability from a conceptual stage to industry practice.

1 Introduction

The influence of Artificial Intelligence (AI) in our society is growing. Companies are increasingly considering the use of AI models, as they offer completely new application possibilities. Therefore, simple approaches are being replaced by complex systems that can complement, or even surpass, human capabilities in some areas. This leads to a shift of company processes through AI systems that act independently and autonomously. However, these increases in both capability and complexity

FUJITSU TUM

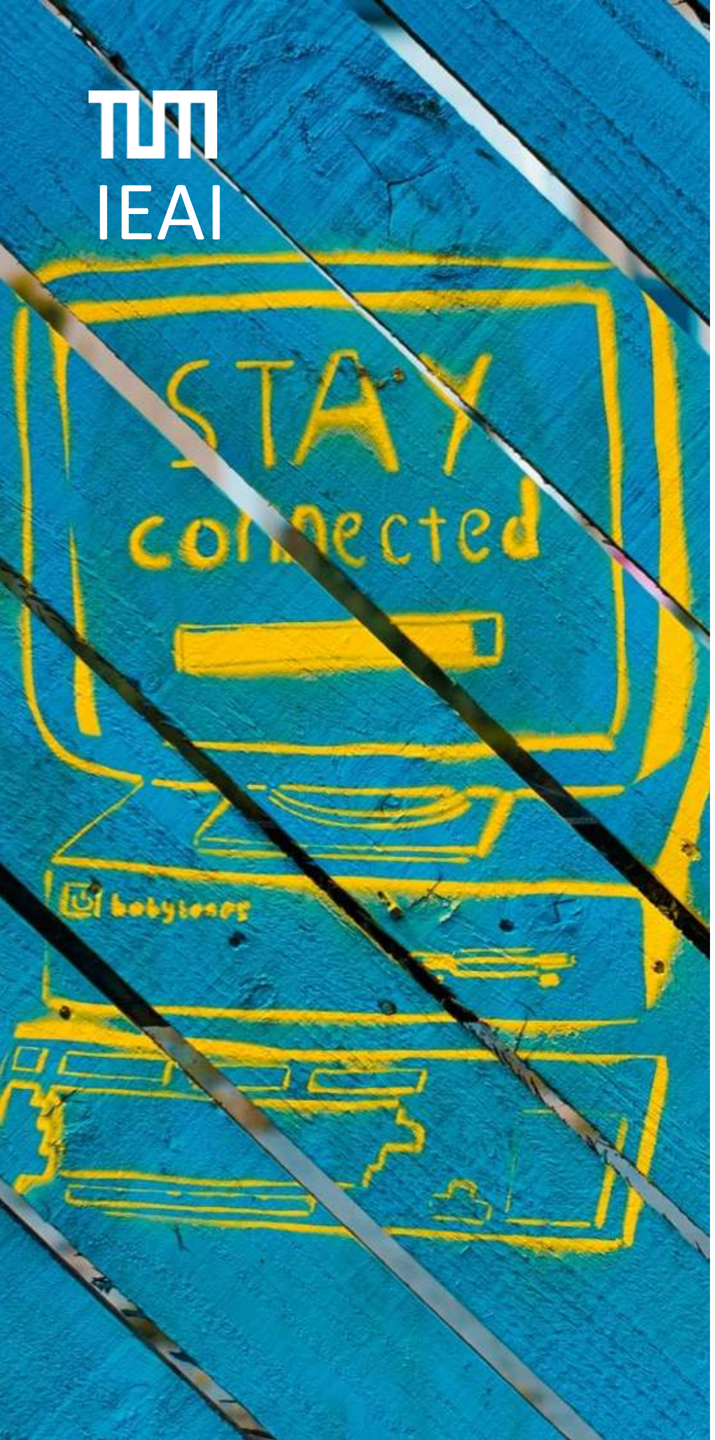
Workshop – Summary & Outcomes

Risk Management and Responsibility Assessment for AI Systems

June 2022

For more material:

ellen.hohma@tum.de | auxane.boch@tum.de | maria.pokholkova@tum.de



Stay connected!

We are happy to see you again.

Stay connected through our website ieai.sot.tum.de, subscribe to our newsletter or follow us on twitter, LinkedIn and YouTube.

References

High-Level Expert Group on Artificial Intelligence (AI HLEG). (2019). Ethics Guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

The General Data Protection Regulation (GDPR) (2016). Regulation (EU) 2016/679.

Financial Services Agency Guidelines of Japan: Legislation and Guidelines.(2003). Guidelines for Personal Information Protection in the Financial Field. https://www.fsa.go.jp/frtc/kenkyu/event/20070424_02.pdf

Sustainable Finance Disclosure Regulation (2019). Articles 6, 8 and 9: Disclosure requirements https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019R2088

Institute of Electrical and Electronics Engineers: the Principles of Business Conduct (2022). https://www.ieee.org/content/dam/ieee-org/ieee/web/org/audit/ieee-principles-of-business-conduct.pdf

OECD,G20/OECD Principles of Corporate Governance, OECD Publishing, Paris (2015). https://www.oecd-ilibrary.org/docserver/9789264236882-en.pdf?expires=1685443351&id=id&accname=guest&checksum=2C025CDB0730F1ACBCC96856BDFAFD36

Hallensleben et al. (2020). From Principles to Practice:An interdisciplinary framework to operationalise AI ethics. https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf

What is the Credit Score? (2022) Retrieved from: https://www.myfico.com/credit-education/credit-scores

Tolkacheva, Rozhkova, Devyashina (2016.) Expert assessment of mathematics teaching abstraction level. http://sefibenvwh.cluster023.hosting.ovh.net/wp-content/uploads/2017/09/rozhkova-expert-assessment-of-mathematics-teaching-abstraction-level-150_a.pdf

What Matters. OKR Approach. https://www.whatmatters.com