Workshop – Risks of AI Systems

Determining Responsibilities within Organizations

January 2023

Workshop Details

TUM
IEAI

Our project

The workshop was part of a joint project between Fujitsu and TUM, where we aim at developing an organizational, risk-based framework for AI accountability.

Risk Assessment

Responsibility Assessment

Risk Management

Accountability Framework

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?

Our workshop

The workshop focused on responsibility assessment and risk management for AI systems within and outside of AI providing organizations.

Risk Assessment

Responsibility Assessment

Risk Management

Accountability Framework

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?

What we expect

Our goal for the workshop was to identify discrepancies between the theoretical requirements for accountability of AI versus their practical applications.

consistency

contribution to the workshop outcome by all stakeholders from a variety of fields and, therefore, stronger reliability of our results

diversification

to ensure the applicability of our Accountability Framework by validating the theory and methods on the example of use cases in different industrial applications

concreteness

to operationalize the principles of AI ethics we make our judgements based on the knowledge of the concrete circumstances (time, trends, interaction between the concepts)

Icons from flaticon.com (Smashicons, Freepik)

What you can expect

The workshop was designed to support participants’ activities through acquiring new insights and exchanging ideas.

hear

what theory says about implications of AI and how to manage them

apply

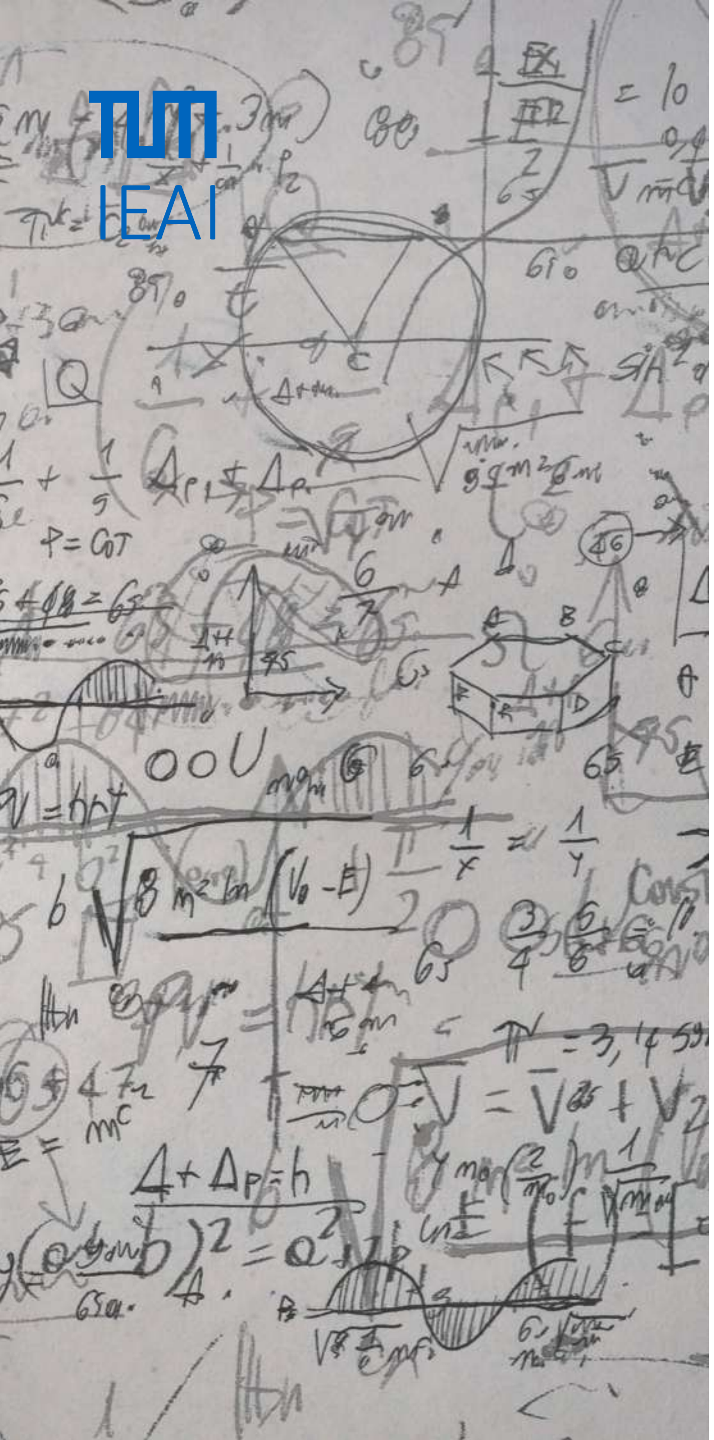
these concepts on practical cases and identify gaps in the knowledge and its transfer to practice

conclude

from the exercises which concrete principles and steps on how to reduce the negative impact of AI technologies are needed

Icons from flaticon.com (Freepik)

Preliminaries and Background

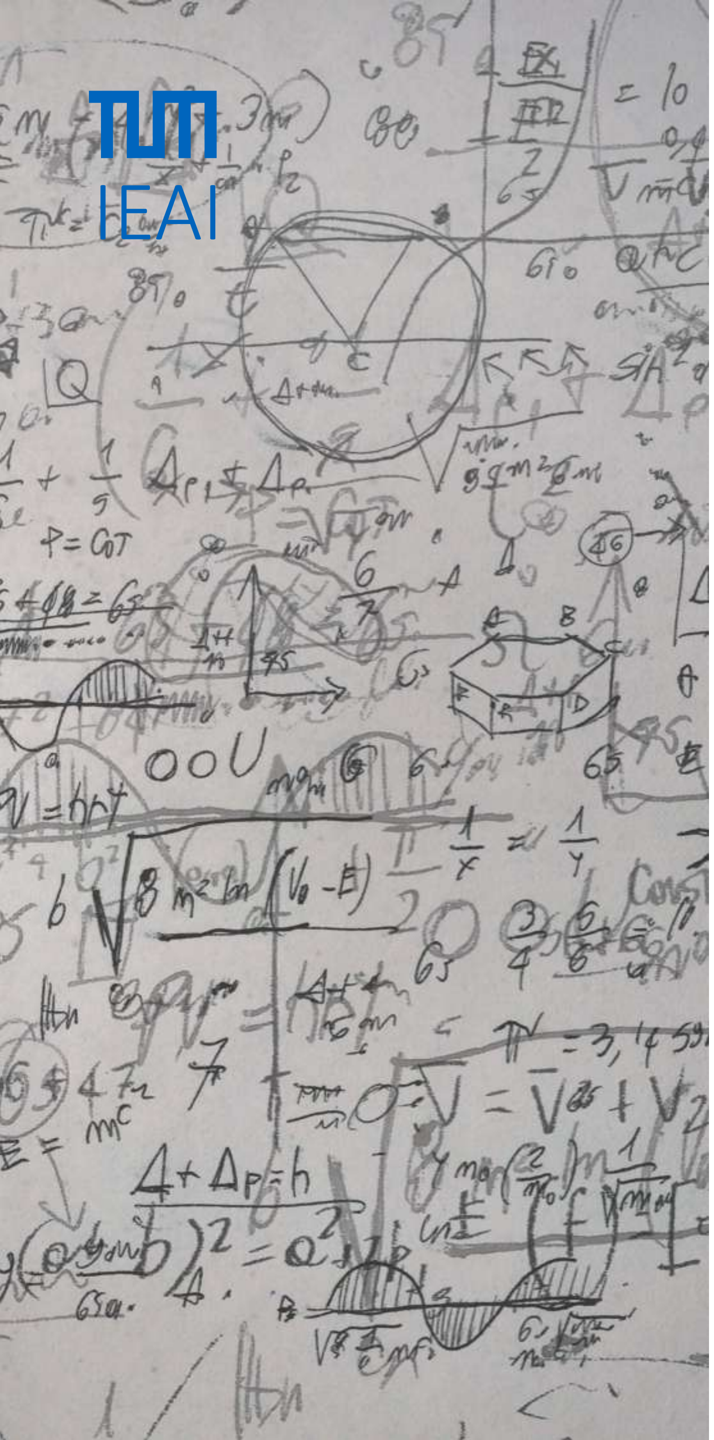


Definition: AI

Some preliminary definitions to kick-off from the same baseline.

Artificial Intelligence (AI) – multidisciplinary field of technologies linked to computer science, data science and statistics, often linked to problem-solving through machine and deep learning

Negative Impacts of AI – a set of risks embodied in the violation of regulations or norms causing harms to individuals or society



Definition: Accountability

Some preliminary definitions to kick-off from the same baseline.

I

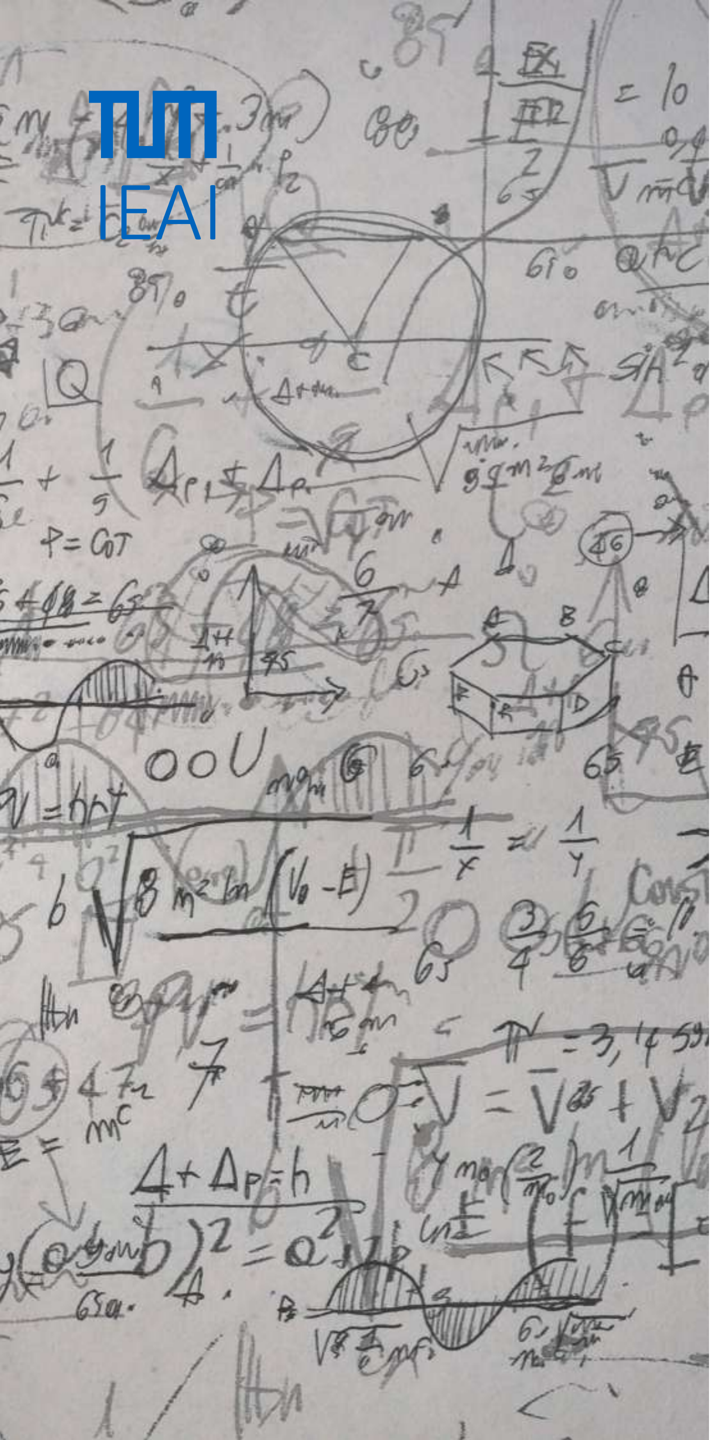
Accountability – “the fact of being responsible for what you do and able to give a satisfactory reason for it”

II

I **Responsibility** = “something that it is your job or duty to deal with”

II **Reasoning** = “the process of thinking about something in order to make a decision“

Source: Cambridge Dictionary (2022)



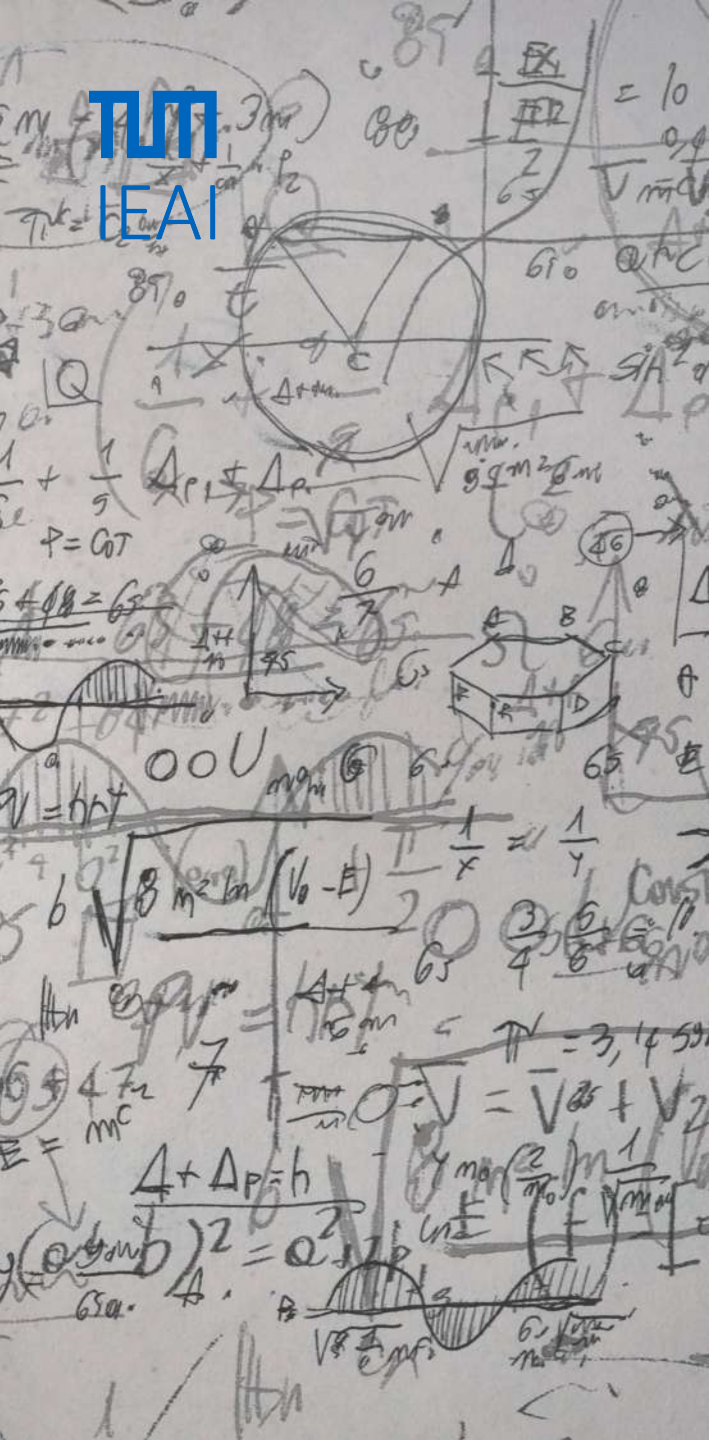
Algorithmic accountability

The definition of accountability can be specified in the context of algorithms, but its two components remain.

Algorithmic accountability comprises 5 key elements:

1. the **actor** – who is responsible
2. the **forum** – to whom is the account directed
3. the **accountability relationship** – the relationship between actor and forum that justifies the account
4. the **account** and its criteria – what does the account entail
5. the **consequences** that are imposed by the forum if the account is not fulfilled

Sources: Wieringa (2020), Bovens (2007)



Algorithmic accountability

The definition of accountability can be detailed regarding the responsible actors involved in AI development projects.

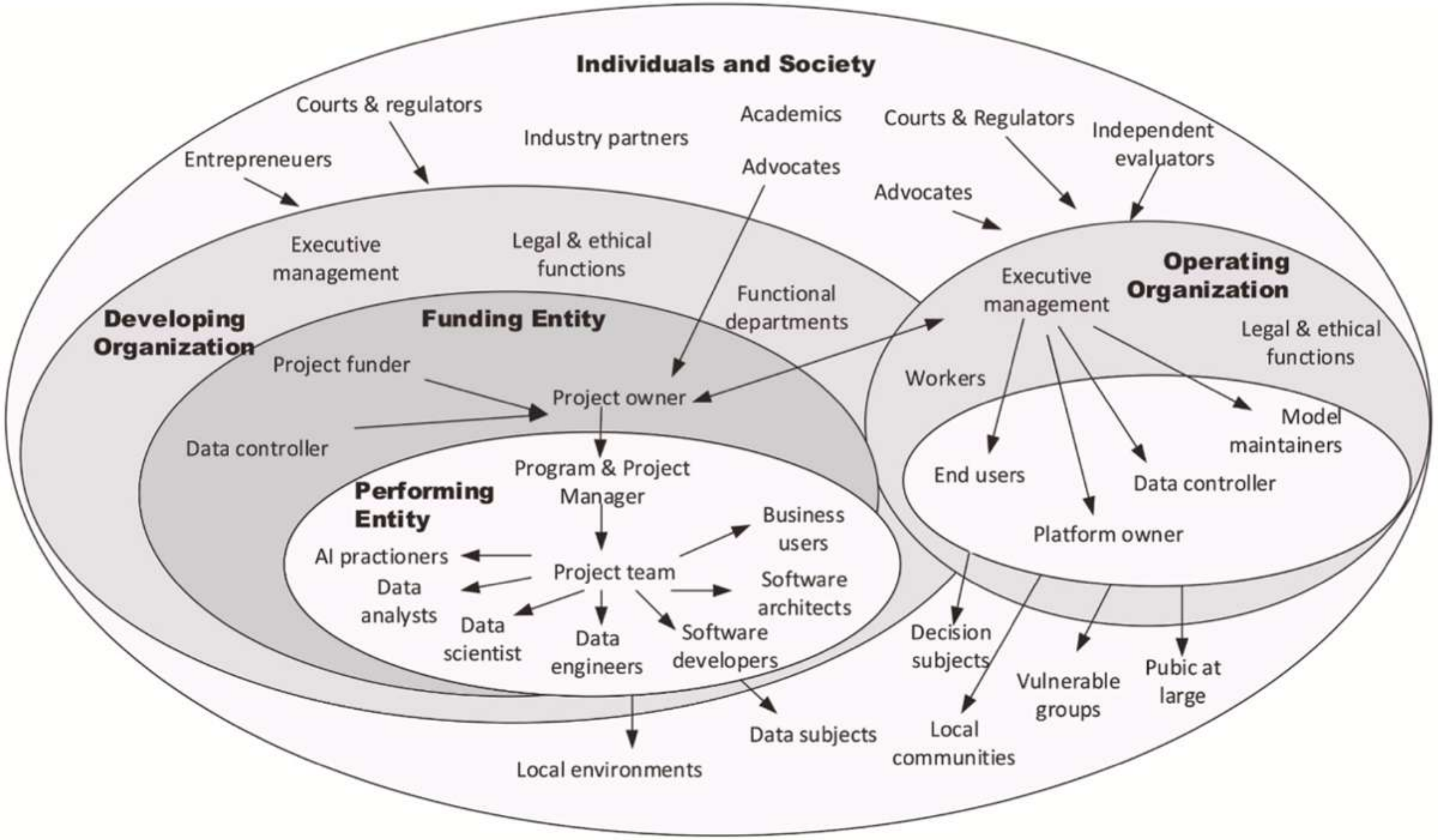
Algorithmic accountability comprises 5 key elements:

1. the **actor** – who is responsible
2. the **forum** – to whom is the account directed
3. the **accountability relationship** – the relationship between actor and forum that justifies the account
4. the **account** and its criteria – what does the account entail
5. the **consequences** that are imposed by the forum if the account is not fulfilled

Sources: Wieringa (2020), Bovens (2007)

AI actor's ecosystem

Recent research has mapped stakeholders of AI development projects into 5 major contributor categories.



Source: Miller (2022)

AI stakeholder roles

Stakeholders may participate in the AI lifecycle at different stages and to varying degrees, resulting in unbalanced levels of power and different roles.

1. **Decision-makers** – decide about the system, its specifications and crucial factors
2. **Designers** – translate requirements and implement the system
3. **Clients** – administer and operate the system
4. **References** – support the project through advice or topic-specific expertise (e.g., law or ethics)
5. **Representatives** – not directly affected but represent passive stakeholders (e.g., policy, civil society communities, media)
6. **Passive** – no option to directly influence projects (e.g., general public)

Sources: Miller (2022), Wieringa (2020)

TUM IEAI

Challenges for AI accountability

Defining why someone should be held accountable for certain actions is challenging in the context of AI.

Responsibility Gap

a manufacturer or operator cannot be held morally responsible if they are not capable of predicting a machine's behavior

Culpability gap

Blameworthiness for wrongdoing based on intention, knowledge or control

Moral accountability gap

Duty of human persons to explain one's reasons and actions to others (under some circumstances)

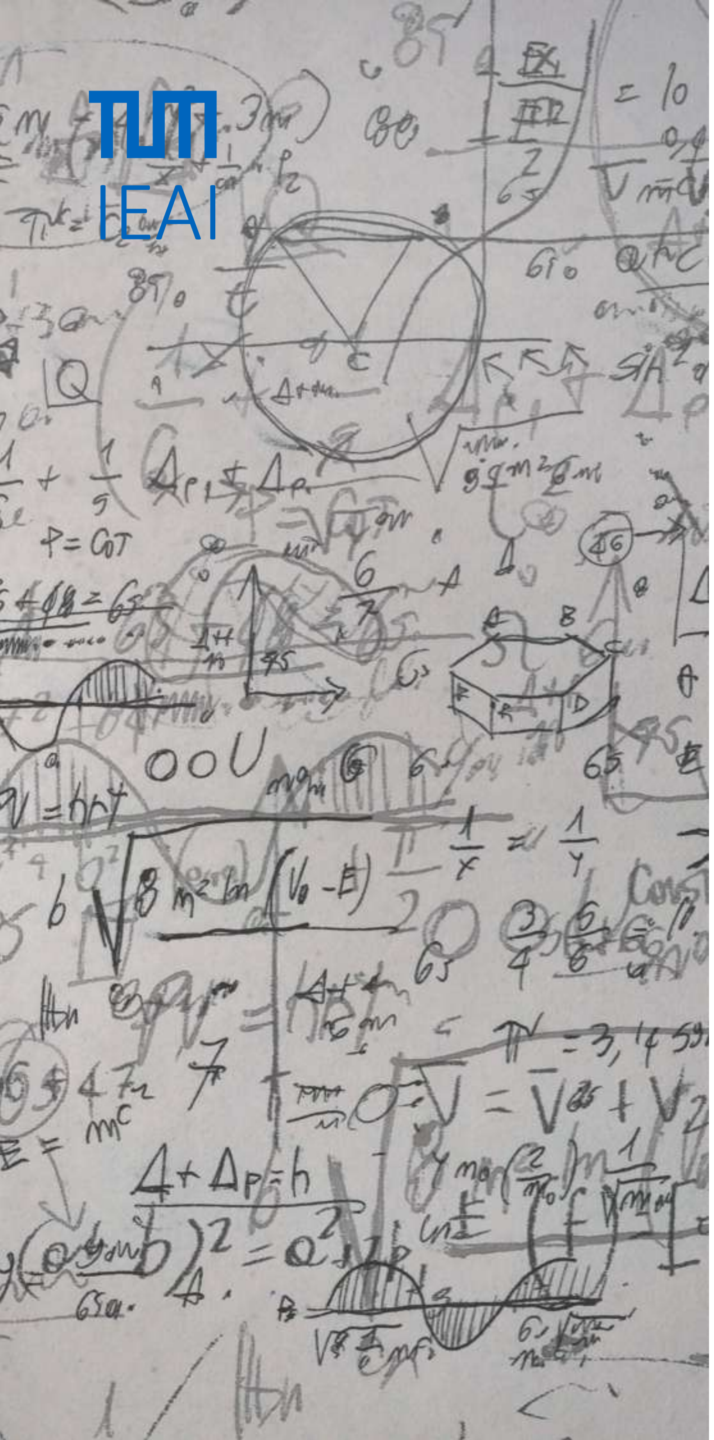
Public accountability gap

Duty of public agents to explain their actions to a public forum

Active responsibility gap

Duty to promote and achieve certain societally shared goals and values

Sources: Matthias (2004), Santoni de Sio & Mecacci (2021)



Algorithmic accountability

The definition of accountability can be detailed regarding what to hold stakeholders responsible for.

Algorithmic accountability comprises 5 key elements:

1. the **actor** – who is responsible
2. the **forum** – to whom is the account directed
3. the **accountability relationship** – the relationship between actor and forum that justifies the account
4. the **account** and its criteria – what does the account entail
5. the **consequences** that are imposed by the forum if the account is not fulfilled

Sources: Wieringa (2020), Bovens (2007)

Accountability clarification efforts

Regulations and policy papers have been published by the EU indicating which objectives and core values justify the need for accountability.

The High-Level Expert Group on Artificial Intelligence has defined **4 ethical principles** for trustworthy AI:

- Respect for human autonomy
- Prevention of harm
- Fairness
- Explicability

The AI Act mentions specific objectives that indicate key risks to be mitigated:

- ensure that AI systems on the Union market are **safe** and respect existing law on **fundamental rights** and **Union values**
- facilitate the development of a single market for **lawful, safe and trustworthy AI** applications

Source: High-Level Expert Group on Artificial Intelligence (2019), AI Act (Regulatipn 2021/0106)

Accountability clarification efforts

These fundamental values are expressed with 3 main pillars for trustworthy AI by the High-Level Expert Group on AI.

Trustworthy AI

Lawful

- EU primary law
- EU secondary law
- UN Human Rights treaties and the Council of Europe conventions
- EU Member State laws

Ethical

- Ethical norms

Robust

- No unintentional harm
- Perform in safe, secure and reliable manner
- Safeguards to prevent unintended adverse impacts
- Robust from technical perspective and societal perspective

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

Ethical & robust AI

Multiple studies and research groups, such as the AI HLEG, have identified key requirements for the ethical or trustworthy design of AI systems.

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)

Principle operationalization

The EU AI Act is the first legal document to specifically address, regulate and govern risks of AI.

PROHIBITED	Unacceptable Risk (Art. 5)	e.g., social scoring, facial recognition, dark-pattern AI, manipulation
CONFORMITY ASSESMENT	High Risk (Art. 6 ss.)	e.g., technology within the field of education, employment, justice, immigration, law, …
TRANSPARENCY	Limited Risk (Art. 52)	e.g., chat bots, deep fakes, emotion recognition systems
CODES OF CONDUCT	Minimal Risk (Art. 69)	e.g., spam filters, video games, …

Source: AI Act (Regulation 2021/0106)

Tools for operationalization

A variety of tools has been proposed by research aiming at managing and mitigating risks by design.

	Use-case development	Design Phase	Training and test data	Building	Testing	Deployment	Monitoring
Beneficence	12	4	2	1	1		1
Non-Maleficence	4	3	7	5	2	4	2
Autonomy	1	4				3	
Justice	4		6	7	9	2	5
Explicability	1	2		4	13	5	4

Number of tools proposed along the AI lifecycle by ethical principle according to research by Morley et al. (2021)

Source: Morley et al. (2021)

Tools for operationalization

However, often these tools have proved impractical in practice and are therefore rarely used.

Weaknesses of practicability have been spotted regarding:

Usability

- Tools do not come as **„off-the-shelf“ methods**
- Additional **effort** required to adapt them to organization's use case, context and needs

Suitability

- Additional effort and consequences of use of tools must be **proportionate** to their benefits
- Currently an **overreliance on explainability** can be observed

Comprehensiveness

- Using technical tools **might not cover all** ethics-related concerns
- Tools' effectiveness **depends on its user's** willingness and correct application

Sources: Vakkuri et al (2020), Morley et al. (2021)

Tools requirements

Our previous workshop results revealed several requirements for risk management methods and tools in order to be regarded useful in practice.

Balanced
Balanced between specialization and generalization, therefore, holistic fundament but adaptable per sector

Extendable
Easily updatable for new regulations & recommendations

Representative
Considering feedbacks from different stakeholders, e.g., field experts or the global population

Transparent
Transparent and understandable by all as well as broadly available and accessible

Long-term oriented
Considering long-term and preventing unexpected or unintended effects

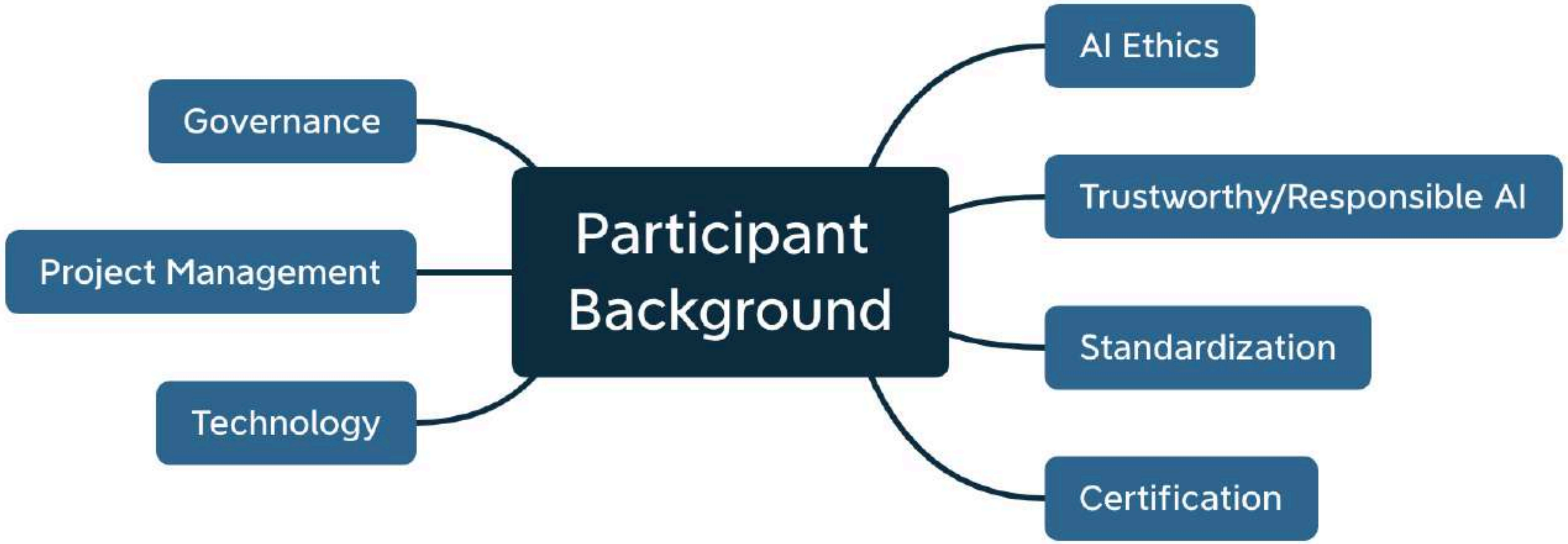
Icons from flaticon.com (Freepik) | Source: Hohma et al. (2023)

Methodology

IEAI

Participant Background

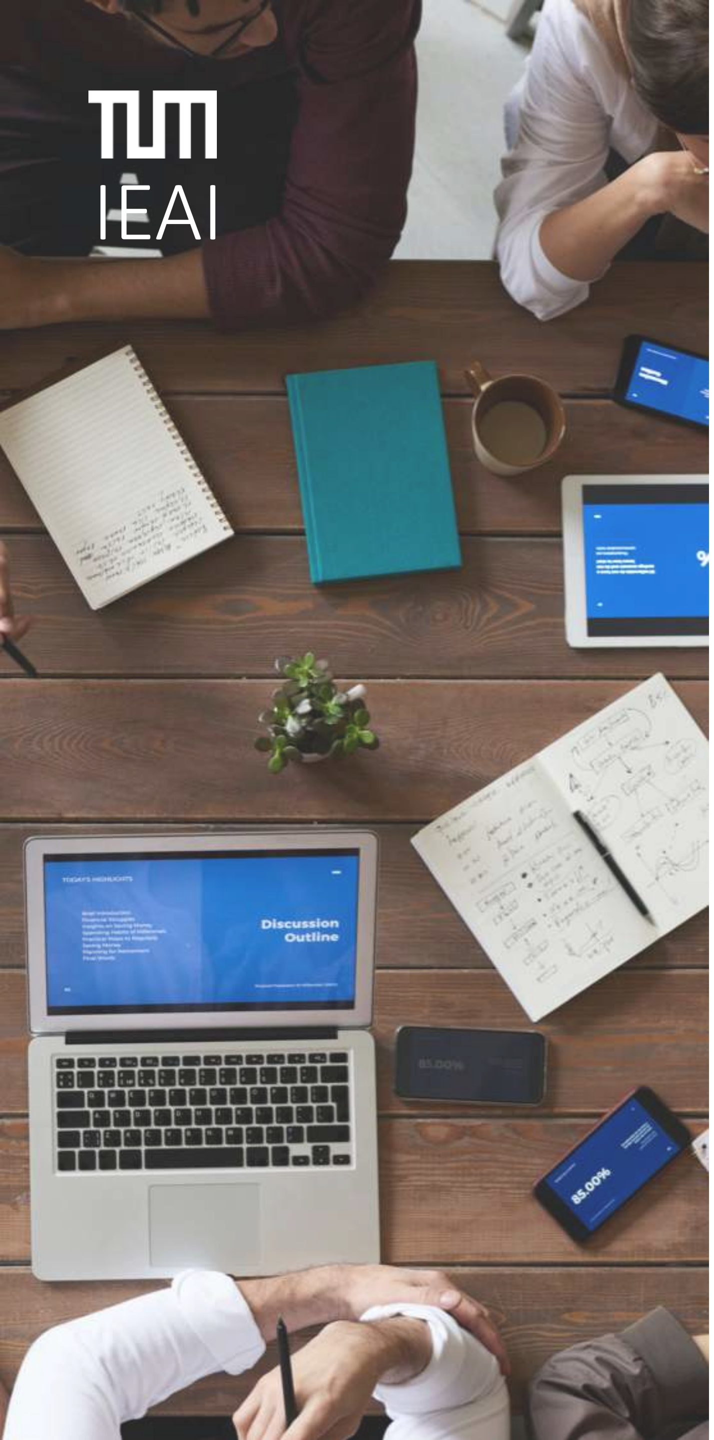
In total, 13 participants brought a great variety and diversity to the discussions, polls and exercises during the workshop.



Workshop agenda

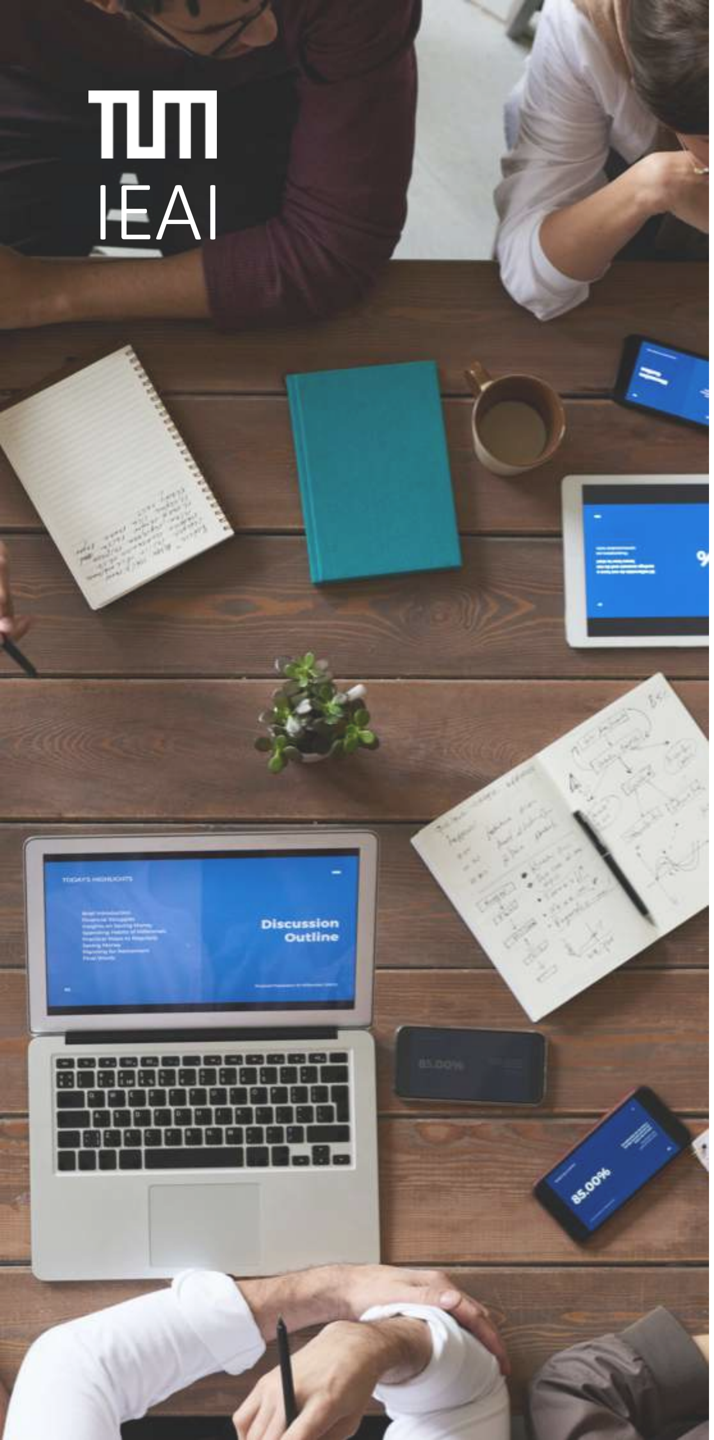
The goal of this workshop was to compare insights on accountability and responsibility for AI providing organizations from theory and practice.

10:00 – 10:45	**Intro**	Introduction from TUM & **presentation** of project, workshop goals and underlying theory
10:45 – 11:00	**Discussion**	Presentation and **discussion** of AI project stakeholders and roles
11:10 – 12:00	**Exercise**	**Mapping** of obligations and measures of responsible AI to AI lifecycle
12:10 – 12:25	**Survey**	**Survey** to assess which measures are currently applied in practice
12:25 – 13:00	**Exercise**	**Discussion** and **rearrangement** of identified tasks according to use cases



Discussion

AI project stakeholders and their roles were discussed on the basis of a presented stakeholder map.



Exercise 1

In two groups, participants were asked to model obligations and measures for reaching trustworthy AI to the AI lifecycle and discuss responsible stakeholders.

Problem Understanding → **Data Handling** → **Model Building** → **Model Monitoring**

Problem Understanding
- Understand the problem
- Identify business metrics

Data Handling
- Data acquisition
- Data labeling
- Data exploration
- Data structuring
- Feature engineering

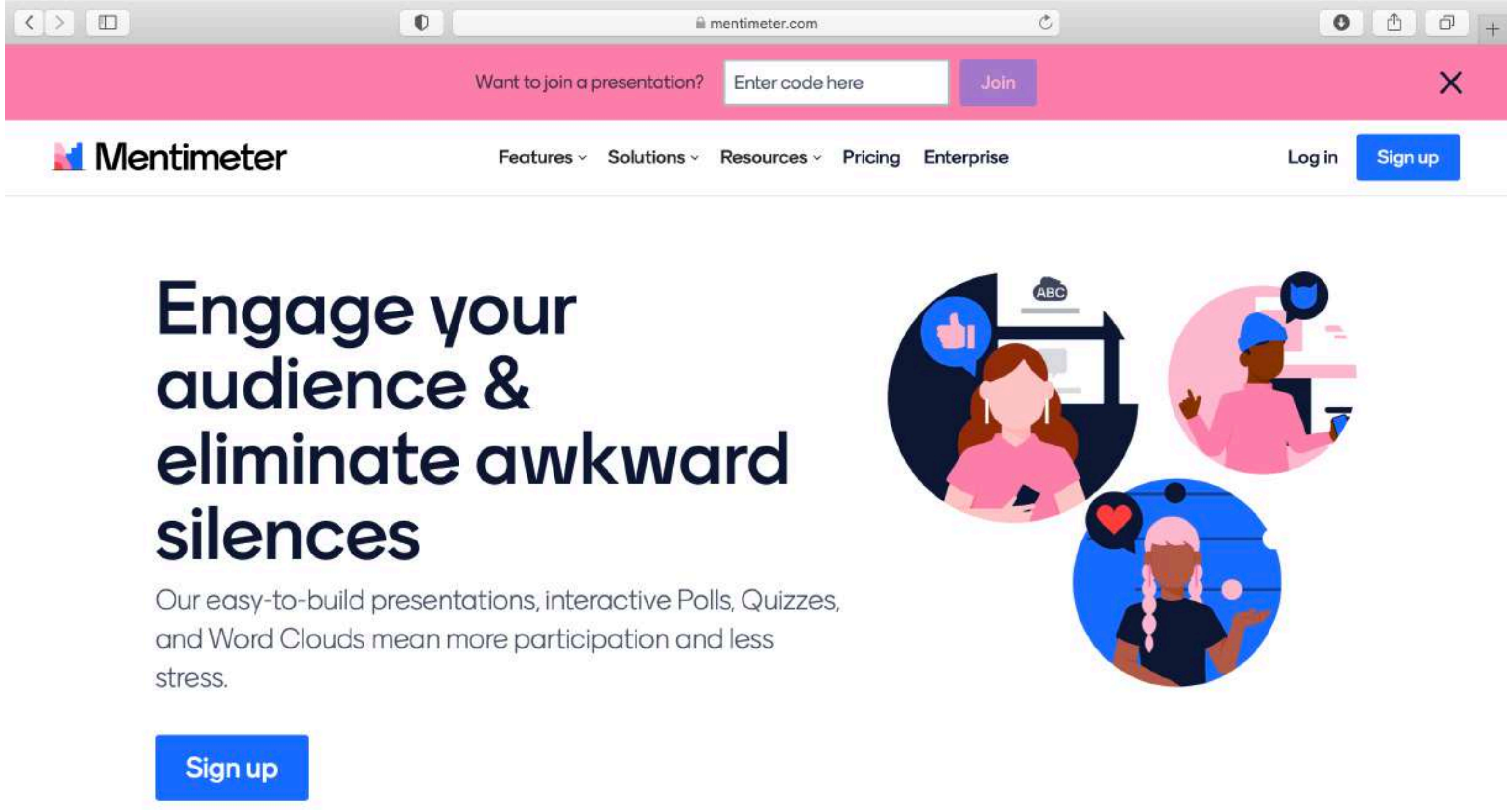
Model Building
- Model training
- Model evaluation
- Model deployment

Model Monitoring
- Evaluate model performance
- Evolve model

Source: de Souza Nascimento et al. (2019)

Survey

Mentimeter, an online tool for interactive polls and word clouds, was used to asses which of the discussed measures participants already implement in practice.



Source screenshot: https://www.mentimeter.com

Exercise 2: healthcare use case

How to implement the determined measures of Exercise 1 in practice was discussed in Exercise 2 on the basis of two real use cases.

Artificial Intelligence can be used in the medical sector to support practitioners in their endeavors. As a specific example, an AI-powered diagnosis assistant building on known medical history of the patient, biological tests, imaging and current course of treatment is used in a general hospital.

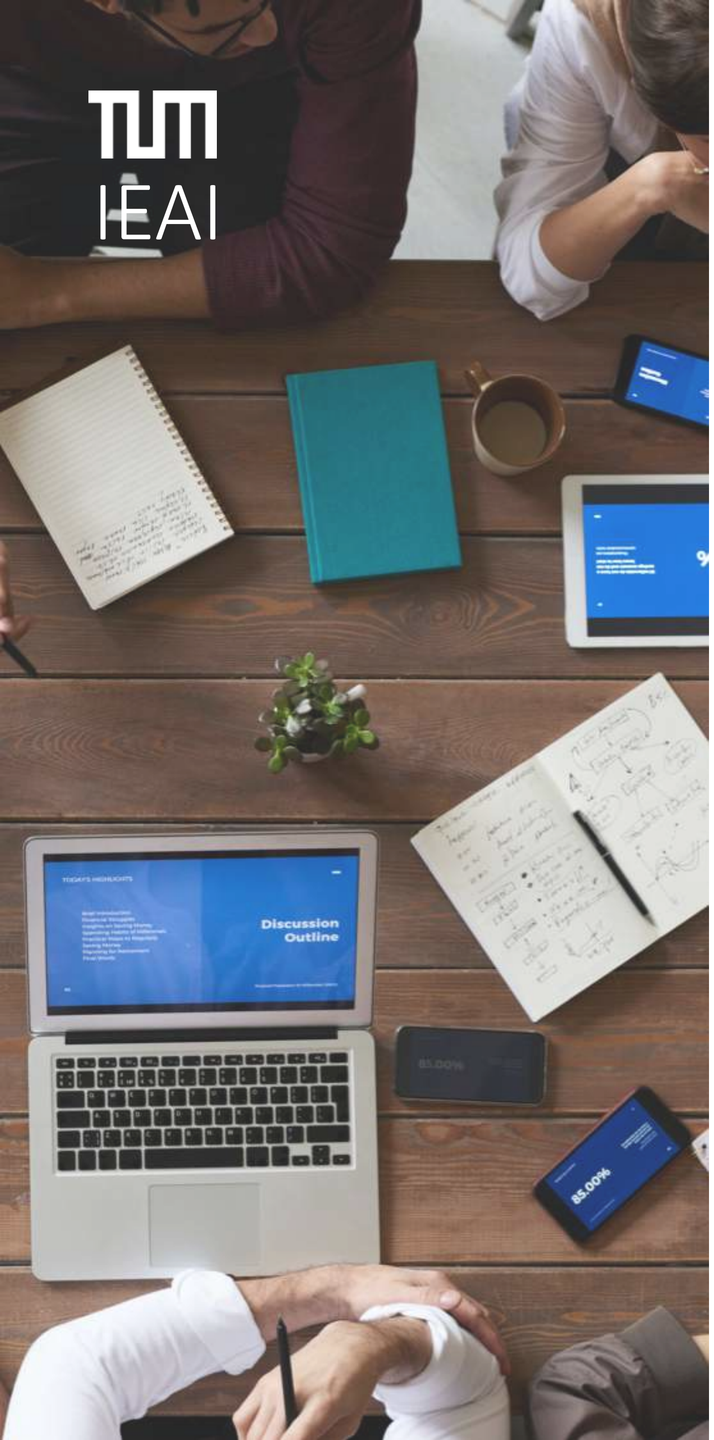
Exercise 2: manufacturing use case

How to implement the determined measures of Exercise 1 in practice was discussed in Exercise 2 on the basis of two real use cases.

Behavioral analysis technologies can detect body actions such as a human’s posture (e.g., sitting or walking), movements (e.g., walking directions) or body part movements (e.g., right arm up).

Using such technologies with real-time video monitoring, for example, in manufacturing facilities can help to verify assembly behaviors, detect workers’ intrusion into dangerous areas or prevent accidents with real-time alerts.

Outcomes



Exercise 1 results (1/2)

In two groups, participants were asked to model obligations and measures for reaching trustworthy AI to the AI lifecycle and discuss responsible stakeholders.

Problem Understanding

- Define **environment & implementation requirements**
  - Legal (e.g., AI act risk classification)
  - Ethical
- Define the **use context** and its **context**, incl.
  - Problem
  - Intended use
  - Scope
  - Stakeholders
- Put up **adequate team**
  - Hire people with right competences and authority
  - Enable internal participation mechanisms
  - Ensure interdisciplinarity

…

Data Handling

- Assess **data quality,** regarding:
  - official requirements (source of the data, legality)
  - ethicality(data bias and fairness, traceability)
  - correctness/adequacy of the data
- Provide **documentation**
- Assess and minimize **risks (incl. unknown)**

Model Building

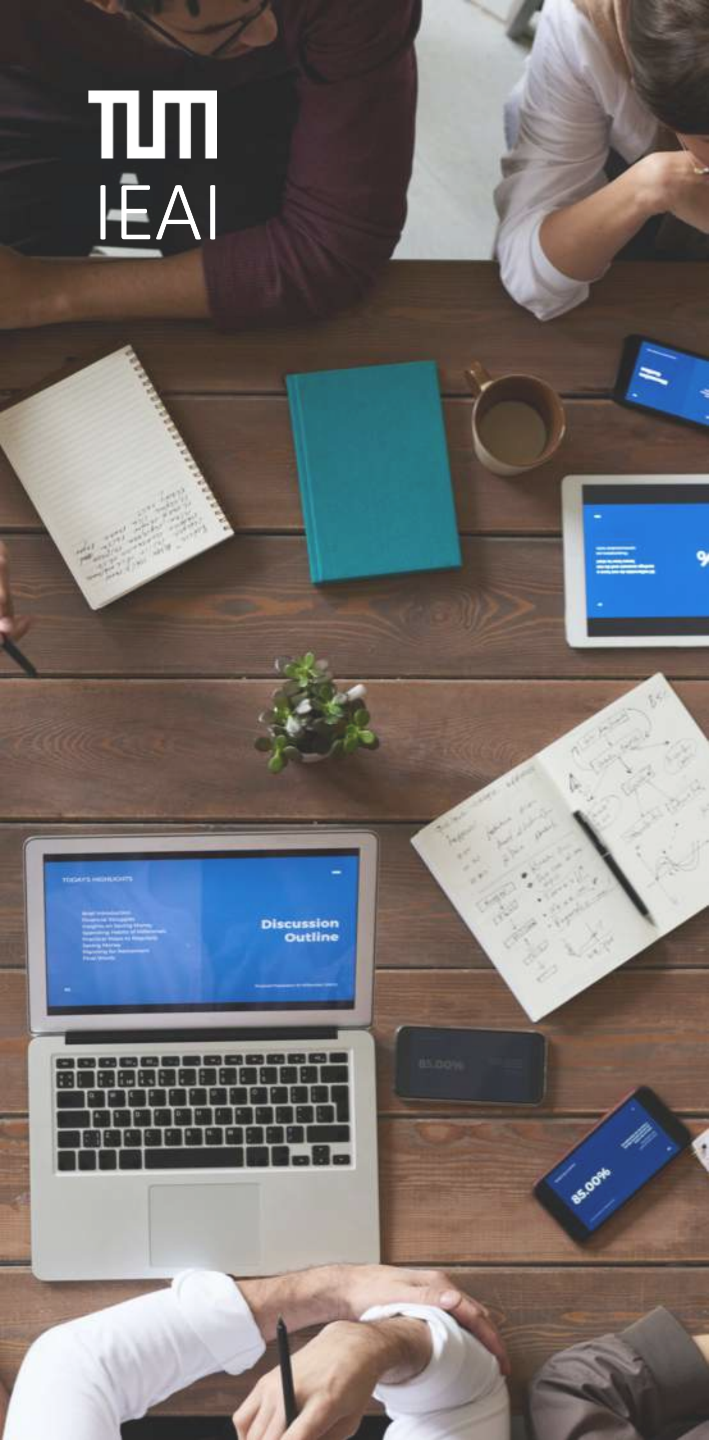
- **Understand**
  - System features / variables
  - Data correlations
  - Third-party components
- Define **target**
  - Define target variables
- **Model evaluation** w.r.t.
  - Costs vs. quality
  - Model vs. performance objectives
  - Model vs. user needs
- **Monitor**
  - risks and risk management
  - KPIs
  - create management dashboard

…

Model Monitoring

- Determine **monitoring criteria**, **KPIs** and **objectives**
- Determine **Resolution Strategies** & **Fallback Plan**
- **Constant monitoring** through
  - Technical tests
  - Compliance tests
- **Re-evaluation** of
  - (Initial) objectives
  - Conflicts of interest
- **Transparently communicate**
  - Information about model and risks
  - Monitoring reports
  - Conflicts of interest

…



Exercise 1 results (2/2)

In two groups, participants were asked to model obligations and measures for reaching trustworthy AI to the AI lifecycle and discuss responsible stakeholders.

Problem Understanding

…

- **Plan processes**, incl.
  - Goals
  - Roles and responsibilities
- Initiate **risk assessment** for
  - Human rights
  - Health and safety
  - Technology standards
  - Ethics technology
- Ensure **outside engagement**
  - Participatory design
  - Understandability
  - Transparent exchange with stakeholders
  - Share knowledge on best ethical practice

Data Handling

Model Building

…

- **Solve identified issues**
  - E.g. with the use of technical tools
- **Document**
  - for marketing claims
  - e.g., using XAI
- **Reevaluate team adequacy**
- Ensure **adaptability** of system to new directives, rules and regulations
- Ensure **participation and collaboration** regarding ethics

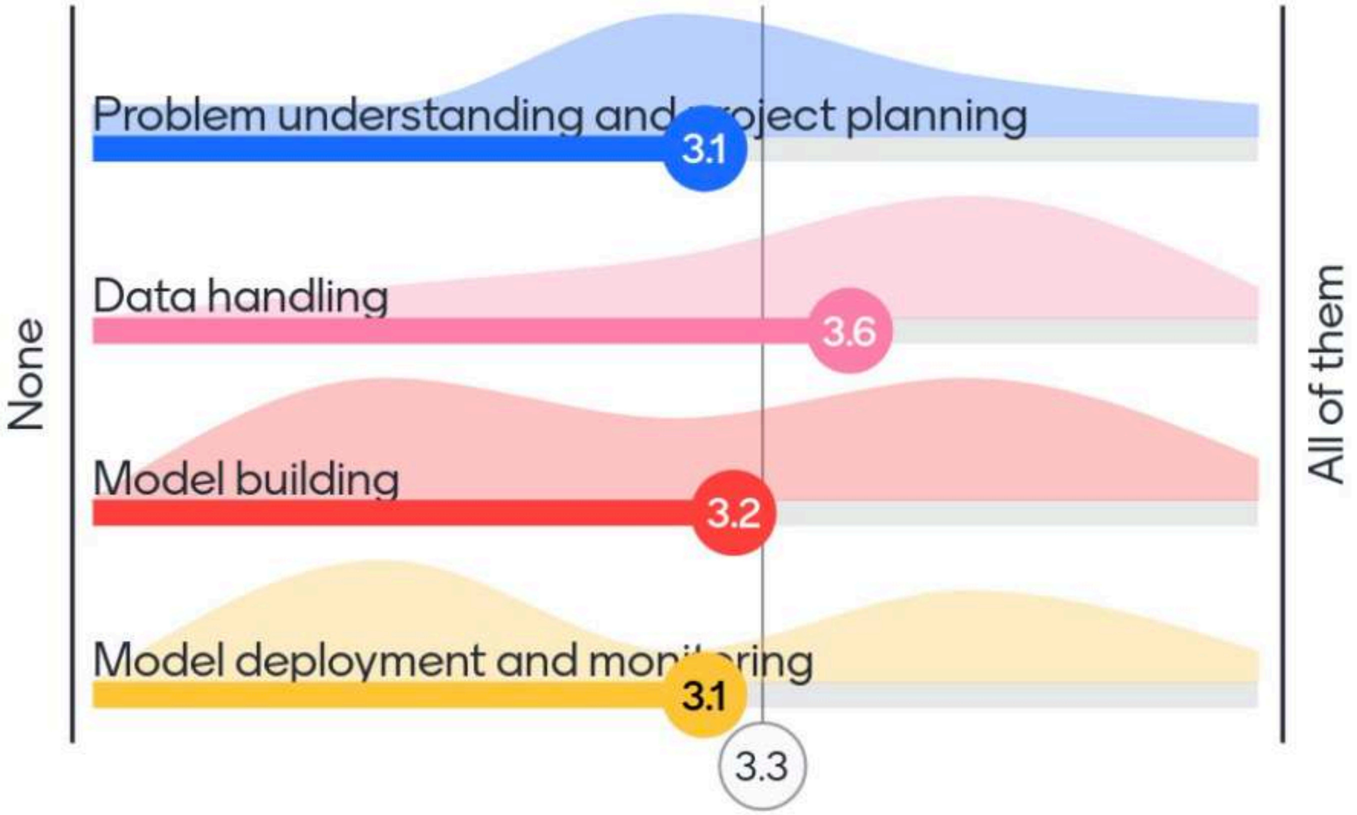
Model Monitoring

…

- **Change mindset** and **guide people**
- **Enable auditing**

Survey results (1/8)

How many of the determined tasks are already executed within your organization (in %)?



Survey results (2/8)

Regarding ‘**Problem understanding and project planning**’, which tasks do you know are planned to be executed in your company within the coming two years?

Survey results (3/8)

Regarding ‘**Data handling**’, which tasks do you know are planned to be executed in your company within the coming two years?

Survey results (4/8)

Regarding ‘**Model Building**’, which tasks do you know are planned to be executed in your company within the coming two years?

Survey results (5/8)

Regarding ‘**Model deployment and monitoring**’, which tasks do you know are planned to be executed in your company within the coming two years?

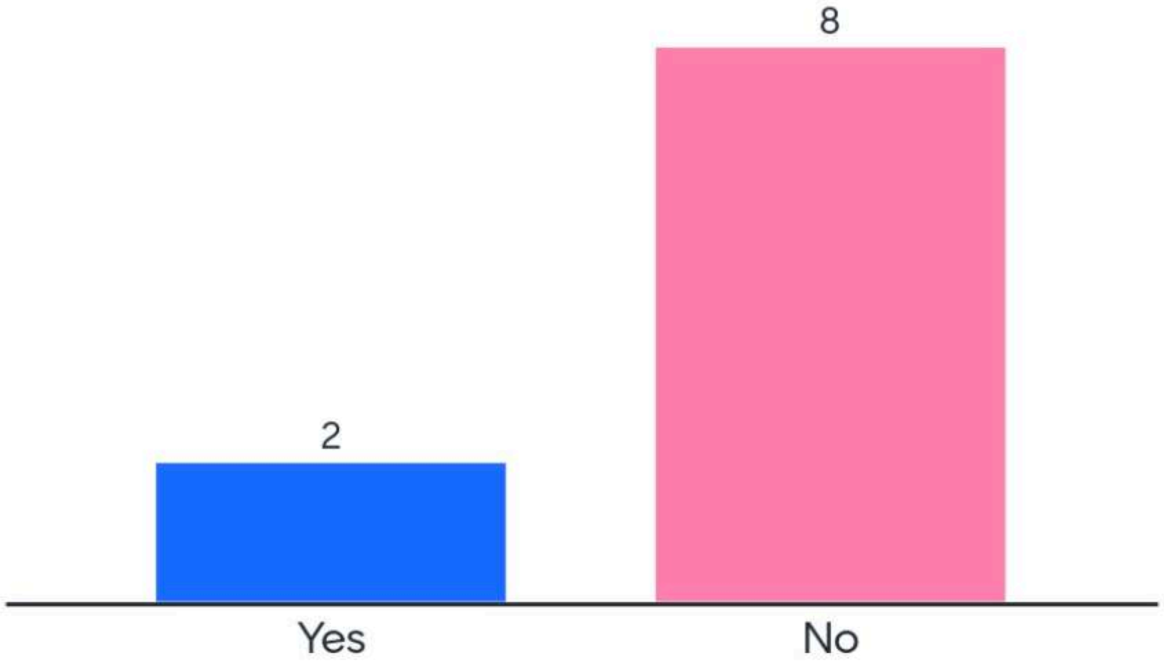
post-market analysis
finding the best tests
live updates
sandbox testing
change control
explainability
incident tracking
traceability
record keeping
working with p sp on uc
problem record keeping

Survey results (6/8)

What do you think are the main challenges impeding the correct execution of those tasks?

Survey results (7/8)

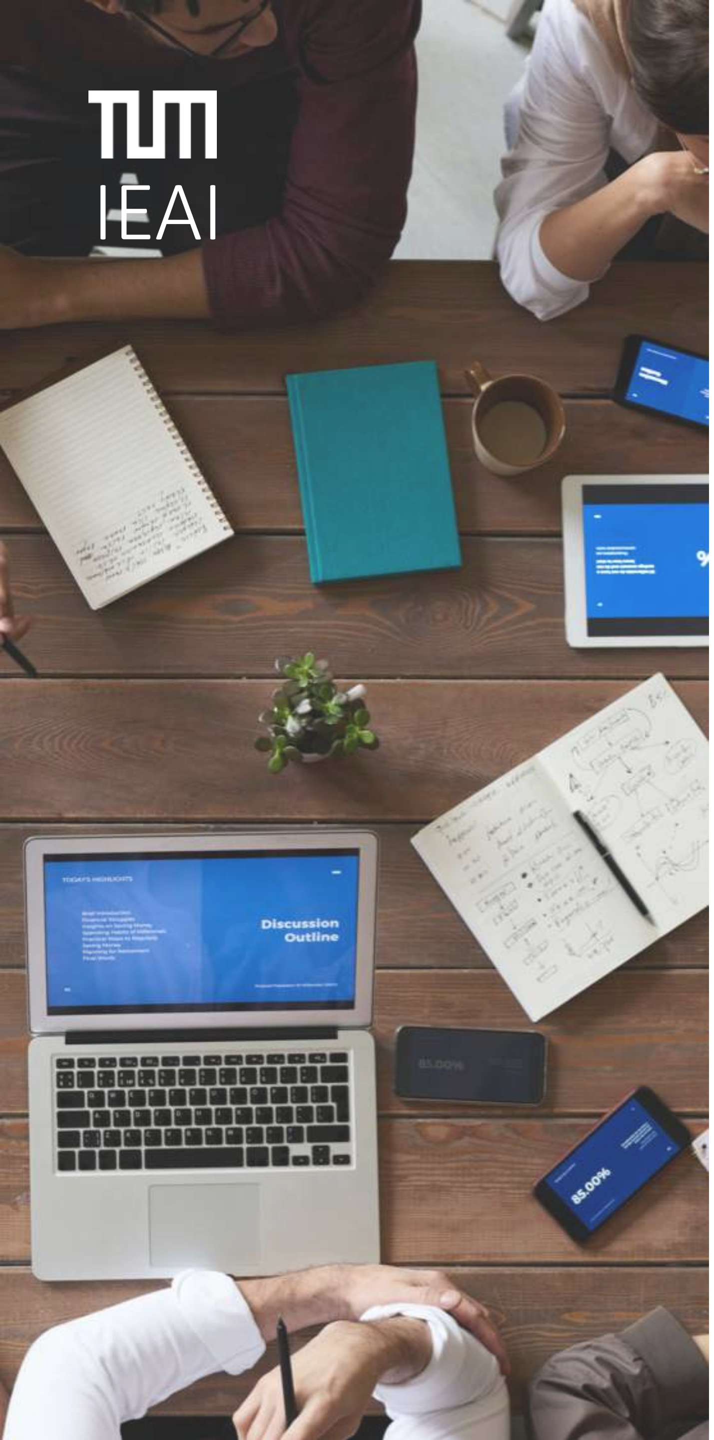
Do you think executing the determined tasks during system development and use (proportionate to respective use case) would result in responsible AI systems?



TUM IEAI

Survey results (8/8)

If you answered no, what would be missing in your opinion?



Exercise 2 results

How to implement the determined measures of Exercise 1 in practice was discussed in Exercise 2 on the basis of two use cases.

Problem Understanding

- Define **environment / context** regarding
  - Legal requirements
  - Ethical requirements
  - Business requirements
- Define **use case**
  - Goals
  - Limitations
  - Market value
  - Stakeholders
- Initiate **risk assessment**

Data Handling

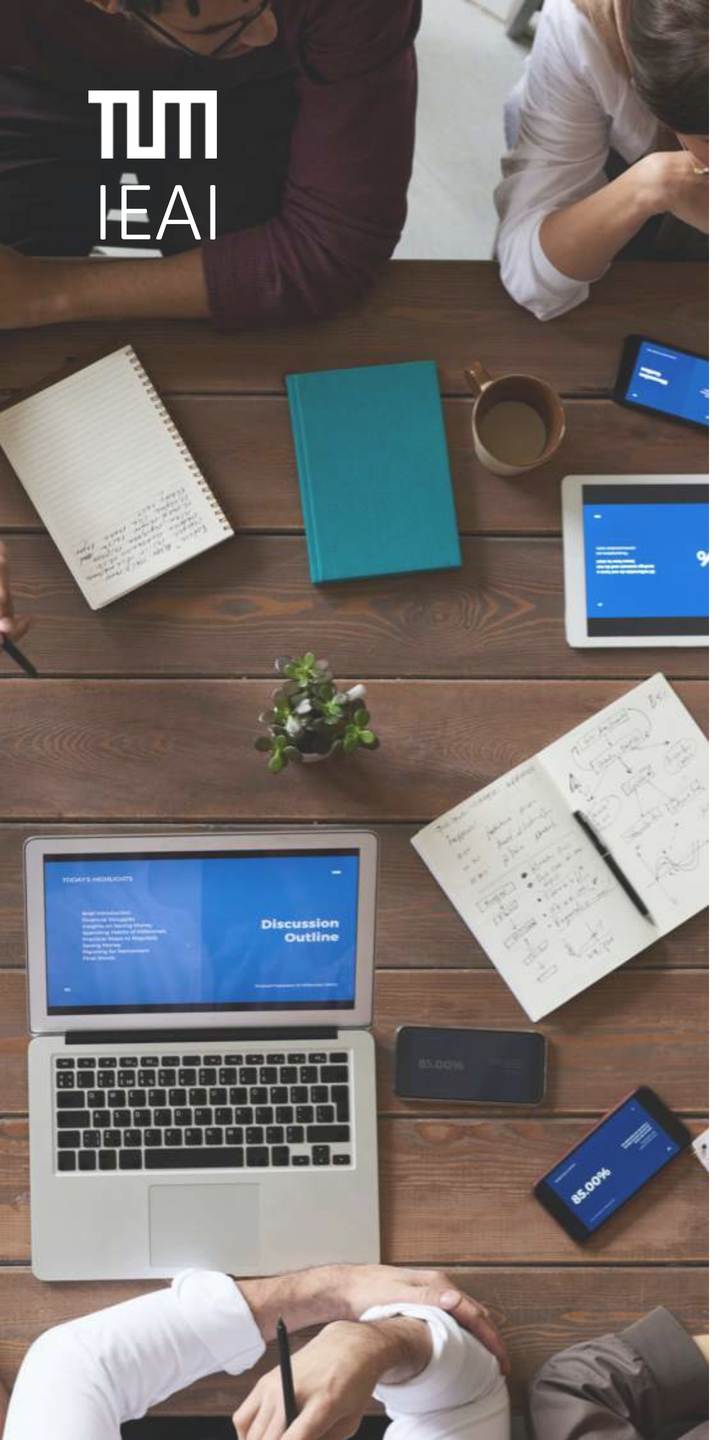
- **Privacy**
  - Remove personal identifier
  - Avoid collecting personal features if possible
- Respect needs of **minorities**
- Ensure **transparency on data collection** processes
- Respect **data security and regulations**
- Definition of **dataset limitations**

Model Building

- Consider **new responsibilities** and **process/lifecycle changes** that arise with self-learning systems
- Build **knowledge** and clarify **technical objectives**
- Understand the **limitations** of the model

Model Monitoring

- Ensure **transparency** of system's purpose and operations
- Avoid **reputational issues** (FDA wall of shame)
- Run **tests** to ensure usefulness, usability and accuracy with participatory design



Conclusion

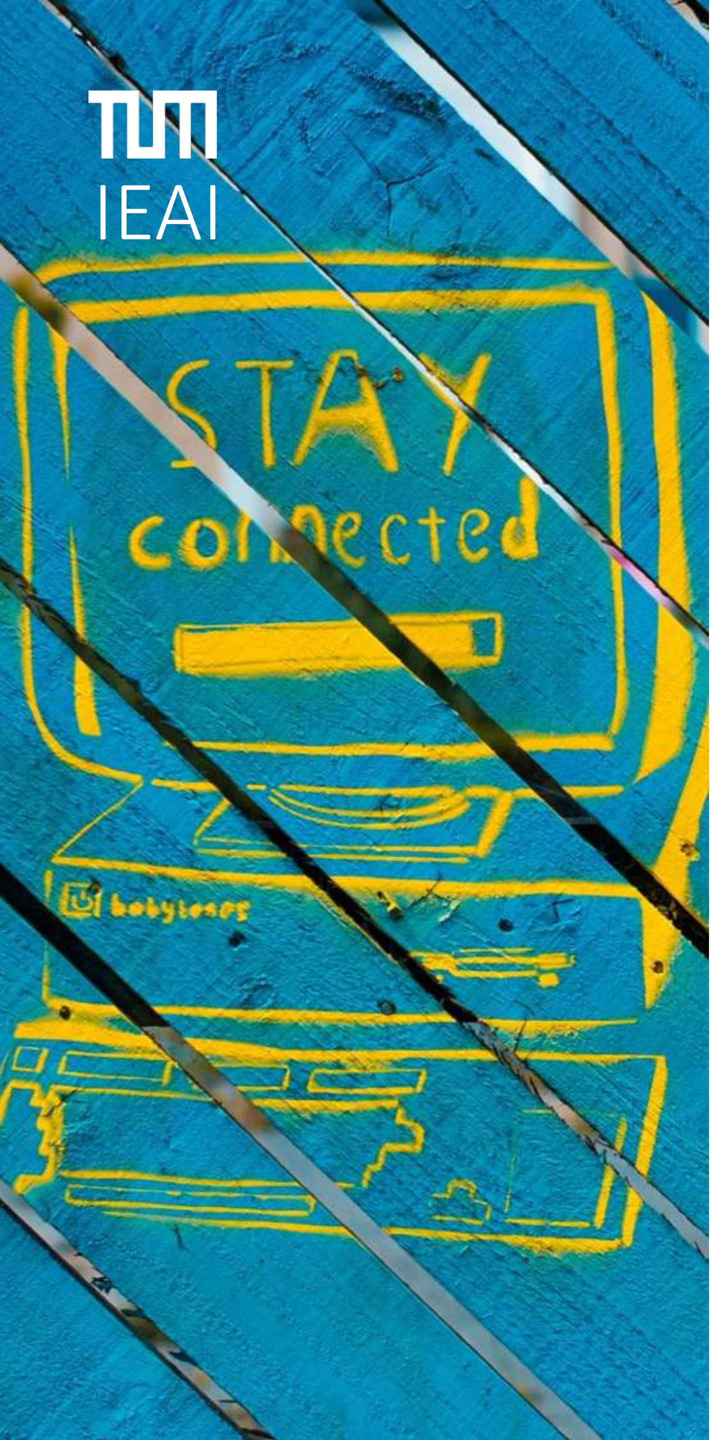
The analysis has revealed interesting commonalities and differences between how measures for AI governance are adopted in theory and practice.

Theory

- Measures can be **determined and distributed** along the AI Lifecycle.
- For more specific use cases, tasks can be concretized but in principle **don’t differ much from the ‘ideal’ considerations**.

Reality

- While the AI lifecycle can give good indications which steps are required, in reality, **stages are more intertwined** and therefore **measures less easily integratable**.
- Companies are moving forward with the implementation of measures, however, **missing standardization and best practices** are still a major point of concern.
- The tasks that are proposed in theory seem to help making AI applications more responsible, but **continuous monitoring and improvement will always be required**.



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

TUM IEAI

Cambridge Dictionary. (2022, May 11). *Accountability*. https://dictionary.cambridge.org/dictionary/english/accountability?q=Accountability

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework 1. *European law journal*, 13(4), 447-468.

de Souza Nascimento, E., Ahmed, I., Oliveira, E., Palheta, M. P., Steinmacher, I., & Conte, T. (2019). Understanding development process of machine learning systems: Challenges and solutions. *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*.

European Commission. (2021). Regulation of the Euopean Parliament and of the Council: Laying down harmonised rules on artificial intelligence (*Artificial Intelligence Act*) and amending certain Union legislative acts.

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

Hohma, E., Boch, A., Trauth, R., & Lütge, C. (2023). Investigating accountability for Artificial Intelligence through risk governance: A workshop-based exploratory study. *Front. Psychol.* 14(1073686), 1-17. https://doi.org/10.3389/fpsyg.2023.1073686

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata*. Ethics and information technology*, 6(3), 175-183.

Miller, G. J. (2022). Stakeholder roles in artificial intelligence projects. *Project Leadership and Society*, 3(100068), 1-15. https://doi.org/10.1016/j.plas.2022.100068

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2021). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Ethics, Governance, and Policies in Artificial Intelligence,* Springer, 153-183.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, 1-28.

Vakkuri, V., Kemell, K.-K., Kultanen, J., & Abrahamsson, P. (2020). The current state of industrial practice in artificial intelligence ethics. *IEEE Software*, 37(4), 50-57.

Wieringa, M. (2020). What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20),* New York, NY, USA.