



Workshop – Summary & Outcomes

Risk Management and Responsibility Assessment
for AI Systems

June 2022



Preliminaries and Background



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.



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?

Background

In order to determine how to manage risks effectively, the particular risks arising with AI systems need to be identified.

Risk
Assessment

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

Responsibility
Assessment

Who is accountable?

Risk
Management

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

Accountability
Framework

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

Risks of AI Systems

There are numerous real-life examples of how AI bears risks or can even cause physical or mental harm.

“We Teach AI Systems Everything, Including Our Biases“

– The New York Times (Nov 2019)

“This is the Stanford vaccine algorithm that left out frontline doctors“

– MIT Technology Review (Dec 2020)

“When Self-Driving Cars Can’t Help Themselves, Who Takes the Wheel?“

– The New York Times (Mar 2018)

“Vast data collection may be necessary for curtailing the spread of disease“

– MIT Sloan Management Review (May 2020)

Risks and Duties

Regulations and policy papers published by the EU indicate which objectives and core values should be maintained and reached in AI applications.



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

Risks and Duties

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

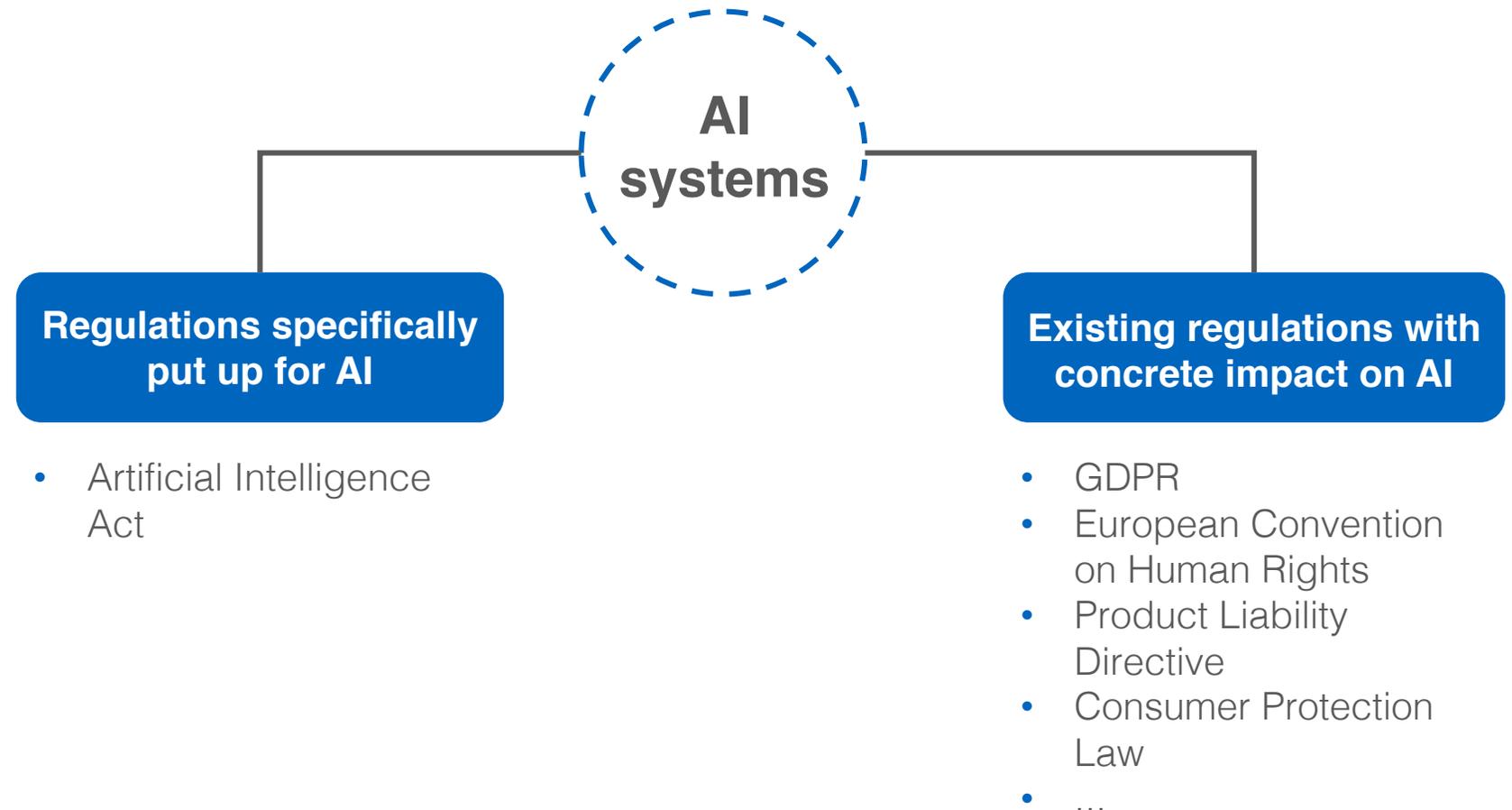
- Ethical norms

Robust

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

Risks and Duties: Lawful AI

While there are some directives that explicitly regulate AI, the majority of regulations that AI must adhere to is pre-existing and independent from a particular use case.



Risks and Duties: Ethical AI

Multiple studies and research groups, such as the AI4People network, have identified key principles for the ethical design of AI systems.

1 **Beneficence**

Promoting well-being, preserving dignity and sustaining the planet

2 **Non-maleficence**

Ensuring privacy, security and “capability caution” (upper limit of future AI capabilities)

3 **Autonomy**

Striking a balance between the decision-making power we retain for ourselves and which we delegate to AI

4 **Justice**

Creating benefits that are (or could be) shared, preserving solidarity

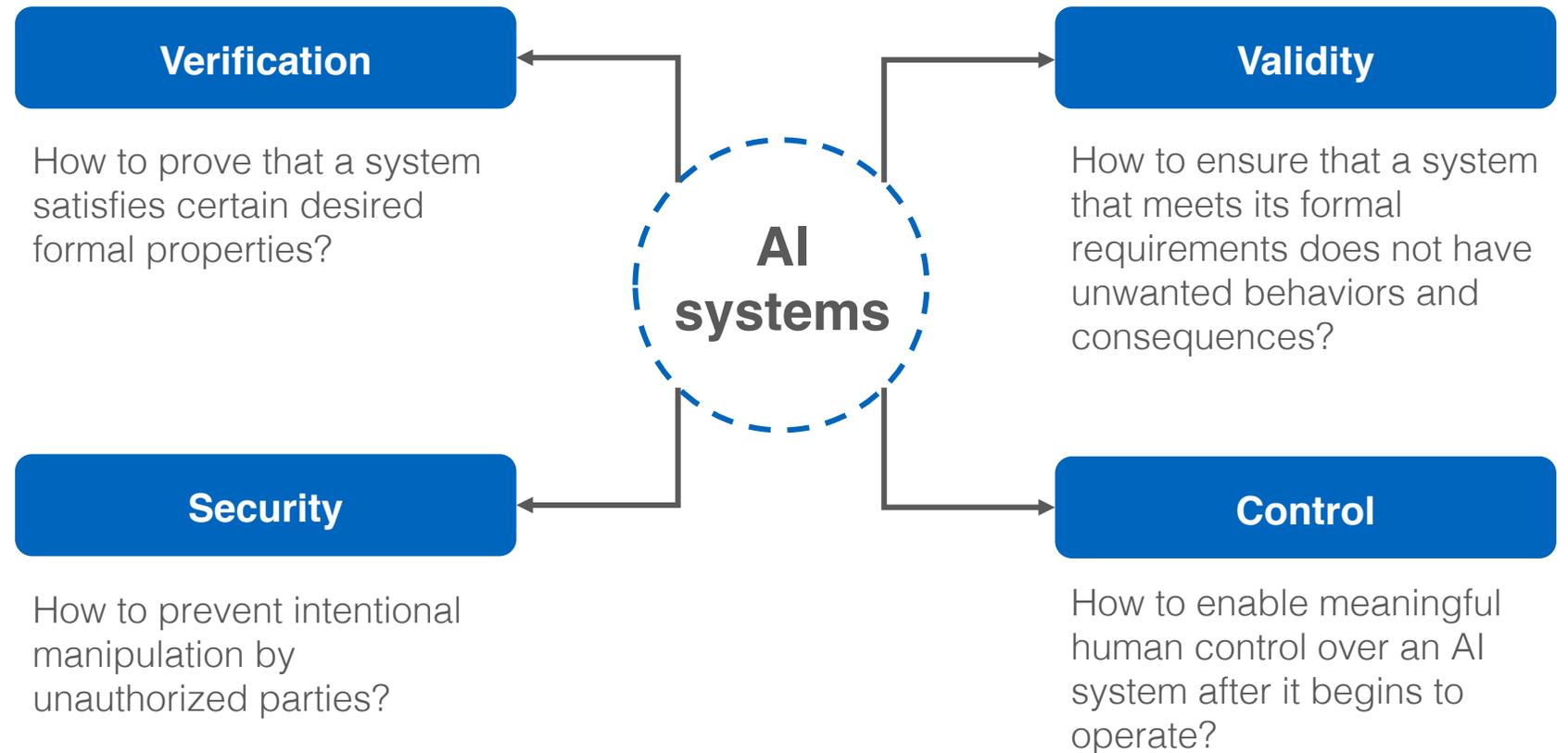
5 **Explicability**

Enabling the other principles through intelligibility and accountability



Risks and Duties: Robust AI

Robustness of AI systems regarding technical problems and social value alignment is fundamental to ensure safety and functionality.



Background

In order to determine how to manage risks effectively, the particular risks arising with AI systems need to be identified.

Risk
Assessment

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

Responsibility
Assessment

Who is accountable?

Risk
Management

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

Accountability
Framework

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

Implications and Stakeholders

Risks arising from technical specifications of AI applications unfold their implications on organizations (particularly the AI provider) and the society.

Technical AI risks

e.g., lack of explainability, robustness, accuracy, safety, quality, monitoring, testing, ...

Organizational implications

- Finance
- Reputation
- Safety and security
- Business operation

Societal implications

- Physical or mental harm
- Human oversight and agency
- Transparency and explainability
- Discrimination and fairness
- Privacy and data governance
- Safety and security

Background

This workshop will focus on risk management and responsibility assessment for AI systems to ultimately determine accountabilities.

Risk
Assessment

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

Responsibility
Assessment

Who is accountable?

Risk
Management

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

**Accountability
Framework**

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

Risk Management Strategies

Various strategies of how to approach risk management have been identified in previous literature.

Proactive Strategies

- **Avoidance**
e.g. non-use of risk-prone component
- **Deterrence**
e.g. signs, threats of dismissal, prosecutions, substantial fines
- **Prevention**
e.g. quality software, designed and documented procedures, staff training, assigned responsibilities
- **Redundancy**
e.g. multiple, parallel evaluations with cross-checking of results

Reactive Strategies

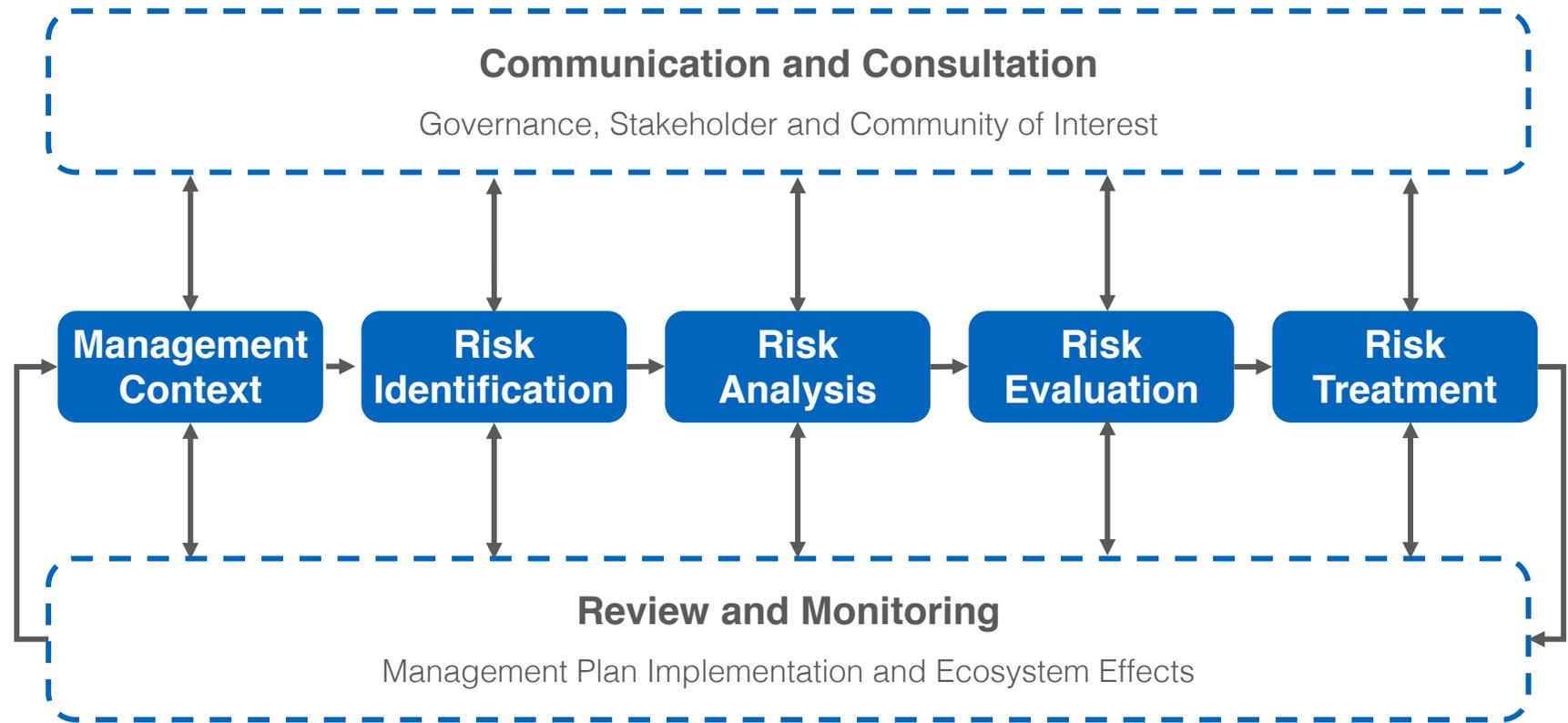
- **Detection**
e.g. exception definitions, software-versioning, logging and time-stamping
- **Reduction/mitigation**
e.g. contingent measures to compensate for harm
- **Recovery**
e.g. designed and documented fallback procedures, staff training, assigned responsibilities
- **Insurance**
e.g. maintenance contracts with suppliers, policies with insurance companies

Non-Reactive Strategies

- **Tolerance/self-insurance**
where assessment of the contingent costs concludes that they are bearable
- **Graceful degradation**
e.g. a pre-funded compensation fund, combined with suspension or cancellation of processing when unexpected harm arises
- **Graceless degradation**
e.g. preparedness to liquidate or disestablish the organization when relatively very large unexpected harm arises

Risk Management Process

General risk management processes are already defined and standardized, for example, according to the ISO 31000 – Risk Management.





Workshop Methodology





Workshop Agenda

The goal of this workshop was to gather insights from practice on the use of and requirements for good AI risk management.

Welcome

Part I: Survey and Discussion

Part II: Prototyping

- Introduction from TUM
- Background on AI risks and implications
- Background on AI risk management approaches

- Assessment of currently used risk management approaches
- Requirements for good risk management techniques

- Prototyping of risk management techniques in small groups
- Wrap-up in panel

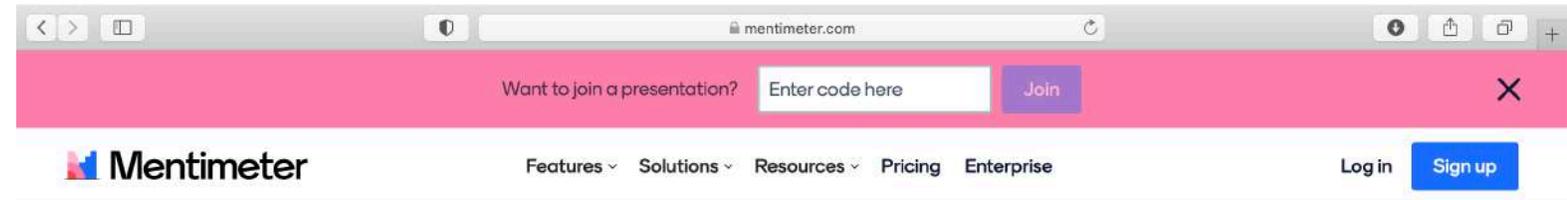
Participant Background

In total, 16 participants brought a great variety and diversity to the discussions, polls and exercises during the workshop.



Workshop Tools – Part I

Mentimeter, an online tool for interactive polls and word clouds, was used for the data collection in the form of surveys during part I.



Engage your audience & eliminate awkward silences

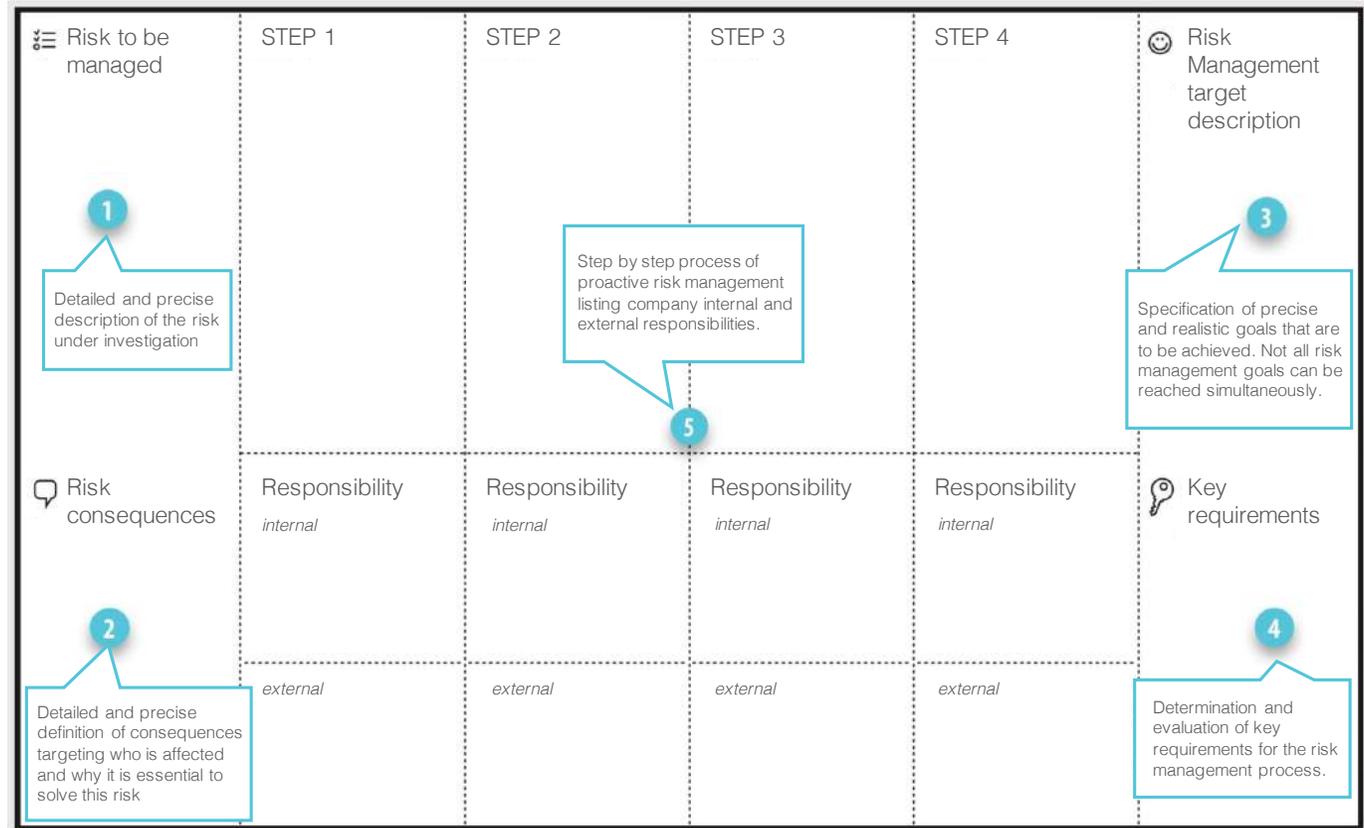
Our easy-to-build presentations, interactive Polls, Quizzes, and Word Clouds mean more participation and less stress.

[Sign up](#)



Workshop Tools – Part II

A Prototype Canvas¹ was used during the prototyping session in Part II to help participants structure their ideas regarding risk management methods.



¹ adapted from the original 'Prototype Canvas' for product design and customer benefit satisfaction

Source: own version of the 'Prototype Canvas' taken from designbetterbusiness.tools

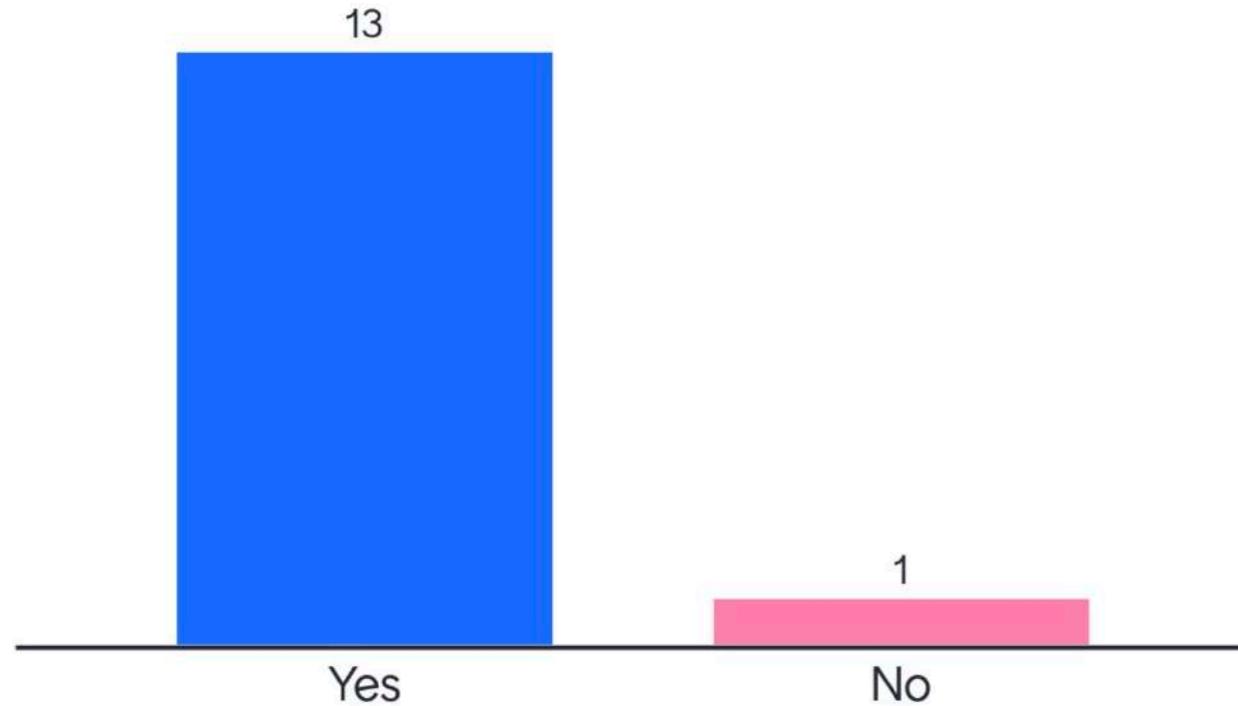


Outcomes – Part I



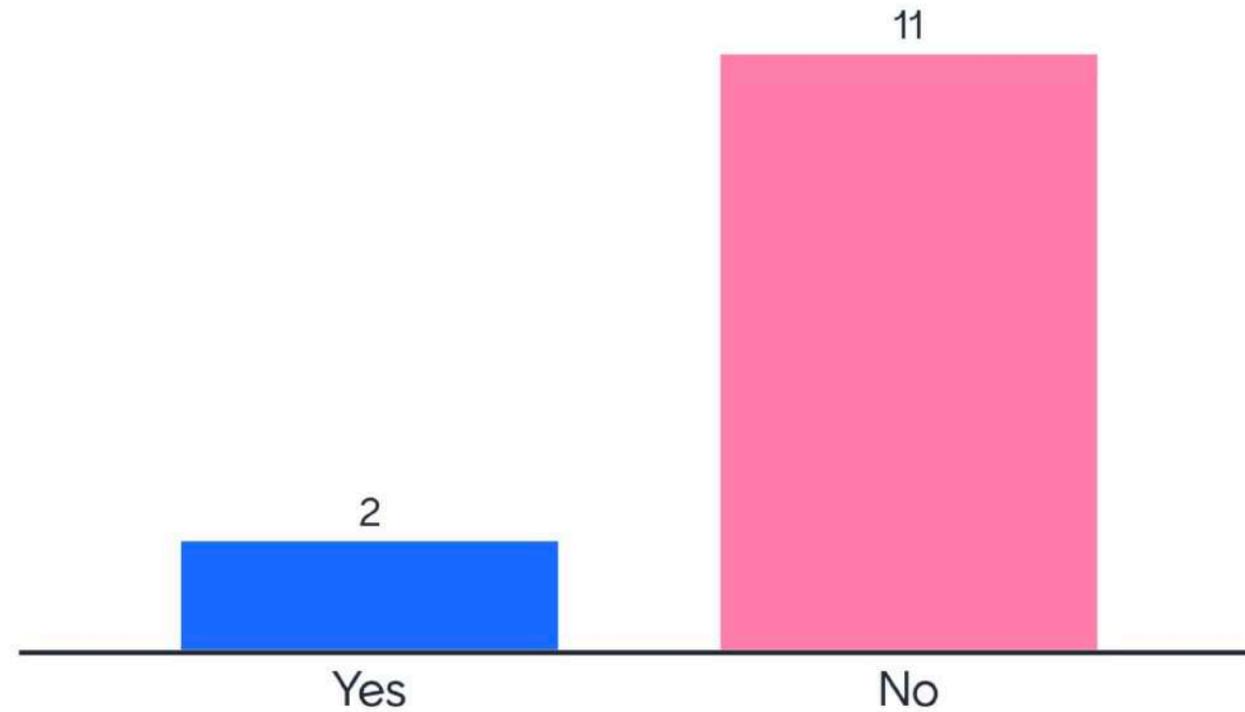
Survey

We asked the participants whether they perceive the risks of AI in their everyday work.



Survey

We asked the participants whether they use methodologies or technical tools to cope with these risks.



Survey

We asked the participants what they think is working with current risk management tools and methodologies.



Survey

We asked the participants what they think is *not* working with current risk management tools and methodologies.



Discussion

We asked participants about their personal experiences on AI risks and their perception during daily practice.

If you perceive risks:

How do you **cope with AI risks?**

If you don't perceive risks:

Why do you think you **don't perceive these risks of AI** in your everyday work?

Why do you think you **don't need to cope with the risks of AI?**

Discussion

Some quotes from the discussion session on AI risks and their perception during daily practice.



Bias

Bias is a very important issue and risk. External opinions or expert advisors might help address and avoid them.

Focus on People

Handling AI risks must serve people, not things. The focus should be on the interests of people and society, also considering 'the bigger picture', in order to avert harm.

Diversity

In order to adequately address risks of AI applications, various stakeholder perspectives must be sufficiently involved. Therefore, a diverse team from different cultures and disciplines is preferable.

Urgency & Accessibility

Many companies do not perceive the urgency of coping with AI risks. In addition, smaller companies often do not have sufficient resources to develop their own strategies and concepts.

Unintended Consequences

It is not easy but at the same time important to know what else the system could be used for. Investigating one's own product regarding its deficiencies and in terms of unintended use, e.g., through workshops, is needed.

Multidimensionality

There is a multitude of risks that come with AI applications and all of them impact humans. We should look at all the risks in total and consider them altogether to be able to grasp their impacts.

Survey

We asked the participants which problems or challenges they come across (in their everyday work) in coping with these risks.



Survey

We asked the participants what they require for a good AI risk management tool.



Discussion

We asked participants about their personal experiences on which problems are perceived with managing AI risks and what needs to be done to do better.

What are (from your experience or in your opinion)
requirements for a good AI risk management approach?

Discussion

Some quotes from the discussion session on how to deal with AI related risks and currently existing risk management tools and methodologies.



Education & Explainability

Explainability of AI products is one of the most important points regarding accountability. To achieve this, stakeholders need to be educated on how to use and supervise AI applications.

Coverage

It needs to be identified if all possible risks are covered. The 'unknown unknown' is a big issue for accountability.

Standardization

AI ethic assessments are scattered and methods are not complete. Currently there is no standardization. Therefore, the development of more comprehensive, end-to-end methodologies should be the focus of the next years.

Specification vs. Generalization

You can't be generic and specific at the same time, as different systems have different characteristics or features, e.g., different sectors have different risks. Balancing usefulness and detail is therefore very important, although it might be difficult to reach.

Extendibility

New fields always have new requirements and demand adaptations or changes. There will be new risks in the future, so methods cannot be static but need to be adaptable to upcoming aspects.

'One size fits all'

A 'one size fits all' approach is not desirable, and hardly achievable, for AI risk management. A generic model to avoid common mistakes and context-aware add-ons to be enacted for addressing specific issues seems more practical.



Outcomes – Part II



Prototyping Results – Group 1 (1)

The participants were asked to prototype in 2 break-out groups an example process for managing AI risks **proactively**.

	STEP 1	STEP 2	STEP 3	STEP 4	
<p> Risk to be managed</p> <p>Fairness (i.e. sectors: health, education, emerging technologies)</p>	<p>Conceptualization / Justification</p> <p>Understanding problem to solve Identifying the target group</p>	<p>(Design) Assimilating data</p> <p>Sources for the data - where it comes from, how was it collected. Representability of the population. Adequate application.</p>	<p>Evaluations & Analysis of the data</p> <p>Sampling Evaluation of data set to ensure it's diverse & inclusive & targeted properly</p>	<p>User Test</p> <p>Diverse group Good sample Considering your target group External (not within the team)</p> <p>Implement feedbacks</p> <p>Set up for working for multiple iterations</p>	<p> Risk Management target description</p> <p>Equity, Inclusion, Consideration of the exact demographic of your target population</p>
<p> Risk consequences</p> <p>Discrimination linked to lack of fairness for specific groups (too much or not enough specific data training)</p>	<p>Responsibility <i>internal</i></p> <p>Top Management, Product Owner</p> <p><i>external</i></p> <p>Governments (e.g., EU level)</p>	<p>Responsibility <i>internal</i></p> <p>Data scientists, Data Owner, Data Management Team</p> <p><i>external</i></p> <p>Data owner (if external)</p>	<p>Responsibility <i>internal</i></p> <p>Data Scientists, Engineers</p> <p><i>external</i></p> <p>Auditors, Regulators</p>	<p>Responsibility <i>internal</i></p> <p>Product Owner, UX Team, Engineers, Data Scientists</p> <p><i>external</i></p> <p>Participants, Regulators</p>	<p> Key requirements</p> <p>Clear definition of the target group with adaptability of the dataset</p>

Prototyping Results – Group 2 (1)

The participants were asked to prototype in 2 break-out groups an example process for managing AI risks **proactively**.

 Risk to be managed Unanticipated Human Impact <ul style="list-style-type: none"> economical impact (organization and individual) undeliberate human rights unintended uses 	STEP 1 Problem Definition <ul style="list-style-type: none"> understanding the problem define the target group in terms of user and data subjects define use scenarios 	STEP 2 Improvement <ul style="list-style-type: none"> train the development team on risks and harm awareness enactments based on learnings, e.g., data collection according to target group and problem definition 	STEP 3 Use Manual Creation <p>explanation of use to the user based on the target group defined in Step 1</p>	STEP 4 Testing & Evaluation <p>of use description for anticipation scenarios</p>	 Risk Management target description <ul style="list-style-type: none"> mental/physical harm discrimination can lead to harm disclosure of data can lead to harm → prevention of harm	
	 Risk consequences <ul style="list-style-type: none"> discrimination mental/physical harm security economical impact safety 	Responsibility <i>internal</i> top management design engineers	Responsibility <i>internal</i> developers, data collection team head of department	Responsibility <i>internal</i> team consisting of participants of Step 1 and 2	Responsibility <i>internal</i> quality control team	 Key requirements <ul style="list-style-type: none"> diversity because it can prevent discrimination, different social backgrounds data access control
		<i>external</i> ethics board legislator	<i>external</i> legislator	<i>external</i> end user	<i>external</i> end user (maybe through feedback loops)	

Prototyping Results – Group 1 (2)

The participants were asked to prototype in 2 break-out groups an example process for managing AI risks **reactively**.

 Risk to be managed Fairness (i.e. sectors: health, education, emerging technologies)	STEP 1 Checking inputs from the system & users <ul style="list-style-type: none"> Incident Identification & analysis of the problem statement Feedback loop that is working - success reporting & feedback from users 	STEP 2 Acknowledge (problem) <ul style="list-style-type: none"> Communicate (internally and externally) 	STEP 3 Take Responsibility <ul style="list-style-type: none"> Explain Be Transparent Educate (internally and externally) Learn something from it 	STEP 4 Find a Solution <ul style="list-style-type: none"> Pull it from the market Research Solutions Review Assumptions & adapt Pay money (fines) & remedy 	 Risk Management target description <ul style="list-style-type: none"> Equity Inclusion Consideration of the exact demographic of your target population
 Risk consequences Discrimination linked to lack of fairness for specific groups (too much or not enough specific data training)	Responsibility <i>internal</i> Product or Middle Management, Customer Service, Data Management, System Operator	Responsibility <i>internal</i> Top Management, Decision makers, Ethics Boards (internal or external)	Responsibility <i>internal</i> Top Management, Decision makers, Ethics Boards, Business Unit	Responsibility <i>internal</i> Developing companies, Legal Departments	 Key requirements <ul style="list-style-type: none"> Measure if there is any drifts or gaps since the product has been completed Feedback system for any issues
	<i>external</i> Regulators, Audit and Insurance Parties	<i>external</i> Media (correctly & rightly), Regulators	<i>external</i> Regulators (e.g., through updating law)	<i>external</i> Regulators, Judiciary bodies	

Prototyping Results – Group 2 (2)

The participants were asked to prototype in 2 break-out groups an example process for managing AI risks **reactively**.

<p> Risk to be managed</p> <p>Unanticipated Human Impact</p> <ul style="list-style-type: none"> economical impact (organization and individual) undeliberate human rights unintended uses 	<p>STEP 1</p> <p>Harm Identification & Assessment</p> <ul style="list-style-type: none"> find out where, why and how the harm has occurred 	<p>STEP 2</p> <p>Harm Reduction (improvement of user/actual person (inter-) action)</p> <ul style="list-style-type: none"> extend user manual to contain this scenario informing the user to make a different decision 	<p>STEP 3</p> <p>Harm Prevention (improvement of system)</p> <ul style="list-style-type: none"> train that wrong decision does not happen again 	<p>STEP 4</p> <p>Feedback to and Testing of Design Process</p> <ul style="list-style-type: none"> improve data collection improve "online learning" process general improvement of product testing 	<p> Risk Management target description</p> <ul style="list-style-type: none"> mental/physical harm discrimination can lead to harm disclosure of data can lead to harm <p>→ prevention of harm</p>
<p> Risk consequences</p> <ul style="list-style-type: none"> discrimination mental/physical harm security economical impact safety 	<p>Responsibility</p> <p><i>internal</i></p> <p>top management & account team, design team & developers, testing engineers</p> <p><i>external</i></p> <p>governments (e.g., TÜV), user</p>	<p>Responsibility</p> <p><i>internal</i></p> <p>top management & account team, data scientists, user manual responsible actors</p> <p><i>external</i></p> <p>depends on the type of system: legal (e.g., GDPR) or user involvement</p>	<p>Responsibility</p> <p><i>internal</i></p> <p>top management & account team, data scientists, quality control team</p> <p><i>external</i></p>	<p>Responsibility</p> <p><i>internal</i></p> <p>top management & account team, quality insurance, data collection team, data scientists</p> <p><i>external</i></p> <p>user involvement</p>	<p> Key requirements</p> <ul style="list-style-type: none"> diversity because it can prevent discrimination, different social backgrounds data access control



Summary



Risk Management Process

The prototyping task revealed common steps in the risk management processes developed by the two groups.

1

Problem Analysis

Conceptualization
/Justification

Problem Definition

Checking Inputs from
the system & user

Harm identification &
assessment

2

Reaction Planning

(Design) Assimilating
data

Improvement

(Acknowledge)
problem

Harm reduction

3

Reaction Execution

Evaluations & Analysis
of the data

Use manual creation

Take responsibility

Harm prevention

4

Outcome Testing

User test

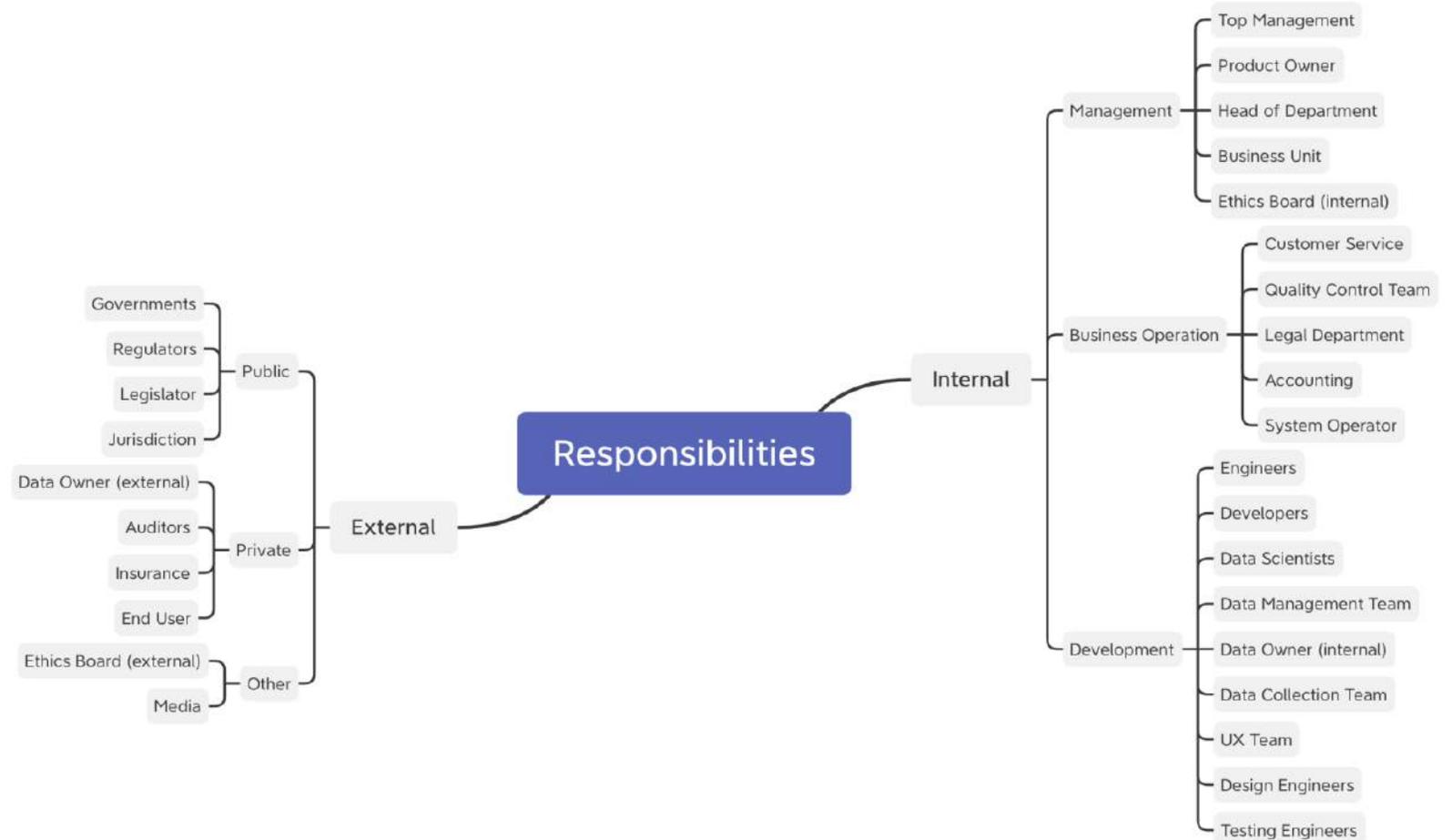
Testing & Evaluation

Find a solution

Feedback to &
testing of design
process

Responsibility Map

Managing risks of AI involves various stakeholders from inside or outside the responsible organization.



Risk Management Requirements

A topic-frequency analysis revealed that a holistic basis with sector adaptability and a focus on understandability and human impact is desirable.



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

Risk Management Content

Specific content was mentioned during the workshop to be helpful for using risk management methodologies in practice.

Risk management methodologies should...

- ... provide a clear accountability distribution per stakeholder
- ... explicitly consider and elaborate on impacts on humans (i.e., groups, individuals or society)
- ... suggest tested methodologies for risk assessment and management
- ... offer communication tools for internal and external use
- ... propose training opportunities, especially for unintended consequences and AI ethics in general

Stay connected!

We are happy to see you again.



Stay connected through our websites ieai.sot.tum.de and mos.ed.tum.de/en/ftm/, subscribe to our newsletter or follow us on twitter, LinkedIn and YouTube.