"THE HYPER EXPERT COLLABORATIVE AI ASSISTANT"

Deliverable n4.1

Measurement scales and metrics for trustworthy AI

Version	1.0
Lead partner	CAT
Dissemination level	Public
Type of deliverable	Report
Contractual date	30/09/2024
Actual delivery date	30/09/2024

D4.1. Measurement scales and metrics for trustworthy AI

Deliverable description as per DoA.

Authors	**Organisations**
Charles FAGE	CAT
Elise DURNERIN	CAT
Joseph GARDETTE	CAT
Aline ROC	CAT
Alice MARANNE	CAT

Contributors	
Ana POP STEFANIJA	imec-SMIT, VUB
Willemien LAENENS	imec-SMIT, VUB
Jonne VAN BELLE	imec-SMIT, VUB

Revised by	
Matilde MENDES	SONAE

Table of Contents

List of Figures

List of Tables

1. Executive summary

In this first section (executive summary), we provide a short overview of the objectives, methodologies, outcomes and conclusions for the **Task T4.1** of the **Work Package 4 (WP4)**.

Within the PEER project, the **Work Package 4 (WP4)**, entitled **"Support: The Artificial Intelligence Acceptance (AIA) index aims at allowing the benchmarking of trustworthy AI application"** is focusing on **the development, the validation and the implementation of an evaluation and assessment framework for human-centric AI systems**, which we refer to as the AIA index. This composite index will allow to easily benchmark different AI systems from a human-centric perspective.

This deliverable presents the work on **Task T4.1 of WP: Transparent and reliable measurement scales for the evaluation of trustworthy AI.** This task lasted one year, and **aimed at identifying and documenting evaluation tools**, as the bases of the AIA index to be further developed.

1.1 Objectives of PEER's "Task 4.1"

As a first step before creating the AIA index itself, the first WP4 task (T4.1, M1-12) was to **identify transparent and reliable measurement scales for the evaluation of trustworthy AI**. These scales and metrics constitute the backbone of the AIA index and will be implemented in the Peac²h platform[1] already developed by CATIE (Centre Aquitain des Technologies de l'Information et Electroniques) and made available for the evaluation process.

The present deliverable (D4.1) presents the work conducted in T4.1, with **the definition of different measurement scales and metrics for trustworthy AI** as well as the **tools to measure it**. In this deliverable, we clearly **define the concepts of acceptance and trust and their underpinning notions**. We also collect different ways of measuring these notions in order to have a **set of tools to build the AIA index** during the Year 2 and 3 of the PEER project.

	Task	Deliverable
M1-M12	**T4.1 | Transparent and reliable measurement scales for the evaluation of trustworthy AI.**	**D4.1 | Measurement scales and metrics for trustworthy AI.**
M13-M36	T4.2 | The AI Acceptance index: Definition, design and prototype.	D4.2 | AIA index: definition and methodology.
M37-M48	T4.3 | Guidelines and recommendations to use the AI Acceptance index and the associated tools and protocols.	D4.3 | Guidelines AIA index implementation.

Table 1. Summary of tasks and deliverables of WP4.

[1] https://app.peac2h.io/

1.2 Methodology of PEER's "Task 4.1"

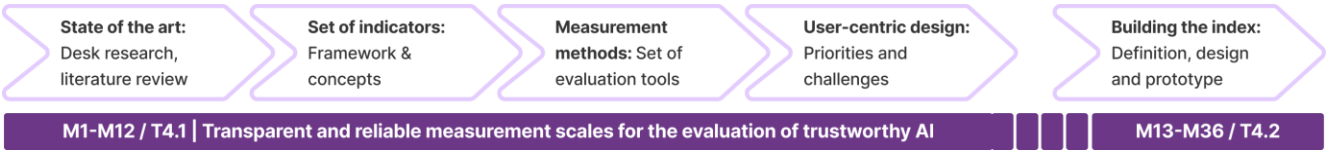


Figure 1. Summary of the steps in the first year of WP4 (M1-M12).

The task 4.1 (Figure 1) **was a preliminary step towards the development of the AIA index**.

First of all, we conduct a review of the current **state-of-the-art** (section 3.1) **to understand** more deeply **the notions of acceptance or trust toward AI systems**. We thoroughly examine a pluri-disciplinary scientific literature to build this knowledge base. To complement this desk-research, we gather information from secondary sources of inputs. These sources are mainly expert groups productions: scholar conferences, European Commission reports, *etc.*

The concepts of acceptance, trust and trustworthiness are detailed (definition, models) and understood in the notions that constitute them. As an example, acceptance is in fact the second step toward adoption, which starts by acceptability. This understanding allows us to identify and **build a set of indicators** underpinning these first concepts (section 3.1.3).

After that, we **identify the different measurement methods and associated tools** existing in the literature accounting for the indicators documented in the up-to-date state-of-the-art (section 3.2).

Then, we present the methodology we adopted for reviewing and **selecting a subset of existing evaluation tools** (section 3.3). Based on this methodology, we identified a **first set of 12 evaluation tools (*i.e.* questionnaires) for end-users which measure the different aspects underpinning acceptance and trust**. This set of questionnaires has been implemented in the online free Peac²h platform[2] to be further used by the project partners and beyond: industries, scholars, politics, *etc.*

Finally, in section 4, we present the **user-centric considerations** of the actual usage of the identified evaluation tools, in order to early **confront scientific knowledge with reality and actual needs on the field**. We present the results of consultations with use-case owners. The objective of these consultations is to **discuss which indicator(s) should be evaluated as a priority according to them**.

1.3 Outcomes

Accomplishing **Task T4.1: Transparent and reliable measurement scales for the evaluation of trustworthy AI**, led to **build a robust foundation of knowledge for this complex concept** involving, among others, **acceptance and trust** among users.

First, we present an up-to-date scientific **state-of-the-art on trustworthy AI** and its multiple components and notions related. Each notion and sub-notion are thoroughly defined, and associated scientific models are presented and commented. *This state-of-the-art is fully presented in this deliverable (section 3.1).*

Second, this comprehensive understanding and definition of trustworthy AI leads us to develop **a framework, to study trustworthy AI,** composed of 5 concepts, 10 sub-concepts, and 29 sub-sub-concepts. This framework

[2] https://app.peac2h.io/

describes every component involved in building trustworthy AI among users of AI systems. *This framework is presented in this deliverable (section 3.1.3).*

The state-of-the-art is detailed in **research sheets, each presenting a specific notion**. For example, we produce a research sheet on the evaluation of acceptance of AI systems. Contents of these research sheets are accessible to a non-scientist population and therefore will be disseminated across the PEER consortium, but also on the project dissemination platforms: website and social media platforms. *These research sheets are not presented in this deliverable, even though all their contents are based on the state-of-the-art.*

We conducted **an extensive review of the measurement methods** available and used within the scientific literature and expert groups (section 3.2). This review resulted in five different methods: **checklists** (*e.g.*, ALTAI checklist for human agency and oversight), **technical objective data** (*e.g.*, number of errors or response time), **behavioural measures** (*e.g.*, Situation Awareness Global Assessment Technique – SAGAT), **physiological measures** (*e.g.*, eye-tracking), and **surveys** (*e.g.*, the Artificial Intelligence Device Use Acceptance – AIDUA). Consequently, we present a total list of 60 different measures, including 46 tools, 18 post-usage surveys for end-users. These tools aim at covering the entirety of notions underpinning trustworthy AI, as identified in the state-of-the-art.

Then, we present **the methodology we developed to select a subset of these 46 evaluation tools** that can be operationalized within the AIA index to be developed (section 3.3). **This methodology account for the necessity to identify "transparent and reliable measurement scales"**. This methodology leads to the **identification of 12 surveys**, as the bases for the AIA index to be developed. These 12 evaluations tools are implemented in the Peac²h platform, available online and for free for the PEER consortium and outside for the society (industries, politics, scholars, individuals). *These tools and the links to their implementation in the Peac²h platform are presented in this deliverable (section 3.3.2).*

In order to design an AI Acceptance index which will be useful and accepted, we adopted a **user-centric approach** (section 4). In this vein, we present the results from semi-structured interviews with use-case owners. We present **a summary table of the notions to prioritize when evaluating the trustworthiness of an AI** from the use-case owners' point of view. In other words, which concepts underpinning acceptance and trust should be evaluated first. For instance, all three use-case owners identified measuring usability and user experience as highly important, whereas only the City of Amsterdam identified error management as important to measure, and therefore include in the AIA index to be developed. We also documented **the interpretations of this table, as a list of lessons learned**. Typically, the notion of understandability, which is a sub-notion of transparency, is a trade-off: sometimes, being over explanatory can lead to overburdening the user and thus impeding the adoption of the system. *This table and lessons learned are presented in this deliverable (section 4.1 and section 4.2).*

We develop our reflexion about the users of the AIA index. We plan to have an **index divided into two parts: one for evaluation by designers or technology providers, and one for evaluation by end-users** (section 4.3)**.** We indicate when in the system's lifecycle the AIA index could be used (section 4.4).

We also present **a set of 5 categories of characteristics potentially impacting the measures of trustworthy AI**, referred as mitigation variables (section 4.5). These characteristics can be human-related, system-related, task-related, environment-related, and human-AI relationship related.

1.4 Conclusion and perspectives

This year, **Task T4.1: Transparent and reliable measurement scales for the evaluation of trustworthy AI** is complete: 12 evaluation tools of trustworthy AI are identified and implemented in the online free Peac²h platform**.**

Accomplishing this task leads to the building of an in-depth **understanding of the complex concept of trustworthy AI, involving acceptance and trust, and their underpinning notions**. This extensive knowledge will be useful for scholars, industries and politics aiming at **better understanding, developing and promoting AI systems which better meet the needs and functioning of individuals**. Consequently, evaluations and designs of AI systems to be further performed, both within and outside the PEER project, will be both improved, and more specifically better trusted and therefore adopted.

Our user-centric approach paves the way for the development of the actual AIA index, based on the evaluation tools identified in Task 4.1. Through discussions with use-case owners and identification of mitigation variables (*i.e.*, characteristics that can modulate trust and adoption towards an AI system), we ensure that **the design of the AIA index will meet the needs of the field partners of the project**. This approach is rooted in the user-centred process of the PEER project, already initiated in the D2.1 (Pop Stefanija et *al.*, 2024) which identified **the socio-technical requirements for the development of the prototypes in the PEER project**.

Next year, **Task 4.2: The AI Acceptance index: Definition, design and prototype** will **articulate the present identified evaluation tools to produce a scoring for a given AI system**. The extensive knowledge produced in T4.1 will directly support the decision-making process of this design for the 2nd and 3rd years of the project.

The **transparent and reliable evaluation scales** identified here in T4.1 will be part of the **whole evaluation process of the different versions of prototypes** to be developed in the PEER project. As CATIE oversees the evaluation process (WP5), this **set of scales as well as other potential measurement methods** (behavioural, technical) presented in this deliverable will be directly **used to both support the design of the prototypes** (by giving scoring and feedback), and **design the AIA index in an iterative manner**, similarly to the prototypes.

2. Management T4.1

CATIE is leader of WP4, and therefore lead the Task 4.1 during the entire first year of the project.

In this section, we present how we worked on the management of task 4.1, including:

- An overview of the research process (section 2.1)
- The list of the actual outputs of the task (section 2.2)
- A presentation of the partners involved (section 2.3)
- A table to visualize the workplan (section 2.4)

2.1 Research process overview

In order to achieve the identification of the reliable and transparent measurement scales for the evaluation of trustworthy AI (T4.1) as the backbone of the AIA index, **CATIE** adopted **a hybrid approach involving both bibliography** and **User-Centred Design (UCD)**. It allowed gathering up-to-date **information from scientific literature and AI experts**, and ensuring the to-be-developed AIA index will be adapted to the needs of end-users, in the form of the use-cases for this project.

We present below a brief overview of the four main sources of information used to complete our work:

- A literature review (section 2.1.1)
- Additional resources (section 2.1.2)
- Workshops with use-cases (section 2.1.3)
- Interviews with use-case owners (section 2.1.4)

2.1.1 Scientific literature review

A **large range of scientific domains** have been covered to establish a recent basis of knowledge on the complex matter of trust and acceptance in AI. The variety of domains were as follows, with examples of scientific reviews where the papers were found:

- **Computer and Data Sciences:** ACM, Information Fusion, ACM Transactions on Interactive Intelligent Systems.
- **Human-Computer Interactions:** Computers in Human Behavior, Ergonomics, Proceedings of SIGCHI Conference on Human Factors in Computing Science, Human Factors.
- **Social Sciences:** EU Review of Applied Psychology, International Joint Conference on Neural Networks.
- **Organizational/Management Sciences:** EU Law Journal, Organization Science, Management Science, International Journal of Information Management, Information & Management.
- **Economics Science:** Journal of Marketing Science, Journal of Consumer Research.

2.1.2 Additional resources

To complement the recent scientific publications presented above, we also **consulted experts' inputs to account for the research conducted outside of the academic world**. To do so, we attended conferences provided by AI experts and collected documents produced by EU AI Experts. Below a non-exhaustive list of AI experts' inputs, we gathered:

- High-Level Expert Group on AI (AI HLEG) of the European Commission[3]
- France's technological research programme on trusted AI: Confiance.ia[4]
- FlexTech Industrial International Spring School on Human-AI Teaming (France)[5]

Finally, CATIE involved both its ***Human-Centred Systems*** (Human Factors and Interaction experts) and ***Algorithms and Data*** (Computer and Data Science experts) units to provide relevant and complementary insights on this complex domain.

Associating this bibliography research work with experts' knowledge allowed us to build a solid and comprehensive understanding of how trust and acceptance can be built, allowed, underpinned, or even undermined both on the user/human side and on the technical side.

2.1.3 Workshops with use-cases

The work conducted in WP4, and especially **Task T4.1**, was performed in **close collaboration with the first steps of WP2**. In this vein, we directly used the results provided by the workshops conducted by our partners from **Studies in Media, Innovation & Technology (imec-SMIT)** of the **Vrije Universiteit Brussel (VUB)** in the context of **Task T2.1 – D2.1**, namely **social and technical requirements workshops**.

In these workshops, our colleagues **gathered the needs and insights directly from the use-cases** across 3 different categories: **the actual usage of the technology** with steps and emotions felt by users throughout the tasks (without the PEER AI assistant), **the critical users' ethical values** and **options for actions to implement these values** into the design of the to-be-developed PEER prototype.

The information gathered in these **workshops was crucial to the identification of relevant and transparent evaluation tools**. This information gave us insights on **the spectrum of profiles of potential end-users** of AI systems to be evaluated, or **the important ethical values underpinning the building of trust and acceptance** amongst end-users.

2.1.4 Insights from use-case owners

To pursue the development of a user-centric AIA index, we discussed with all **use-case owners in semi-structured interviews** to complement information gathered through the **socio-technical requirements workshops (T2.1)**.

By **involving the PEER use-case owners early in the project**, we **ensure the AIA index relevance**: the index should address the **real-world issues** and **concerns related to AI trust and acceptance**.

2.2 Outputs produced

Aside the present deliverable (D4.1), **two other types of contents were developed** within the first year of **WP4**:

- Research sheets (section 2.2.1)
- Set of evaluation tools implemented on the Peac²h platform (section 2.2.2)

[3] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

[4] https://www.confiance.ai/

[5] https://www.flextechchair.org/FTSpringSchool2024/downloads-1.html

Both these types of outputs will be the object of **dissemination** inside and outside the PEER consortium during the project.

2.2.1 Research sheets

Research sheets are the first contents produced in the making of T4.1. **Each research sheet presents a definition, underpinning model and evaluation tools of a specific concept, sub-concept and sub-sub-concept related to acceptance and trust**. The exhaustive list of these concepts is presented in the section presenting the state of the art (section 3.1). This list includes the definition and modelling of the construct of trust, acceptance, but also the sub-factors as accountability, technical robustness, data management, *etc.* The research sheets are available in the PEER project SharePoint. *Each of these research sheets will be communication tool for Year 2 and Year 3 of the PEER project.*

2.2.2 Set of evaluation tools

The second type of contents produced in T4.1 is a **set of evaluation tools**. These tools **aim at measuring the concepts and sub-concepts related to trust and acceptance**. A total of **12 tools are selected** (section 3.3.2). Each tool takes the **form of questionnaire for end-users**. These tools have been implemented in the online free Peac²h platform (Figure 2) developed and owned by CATIE.

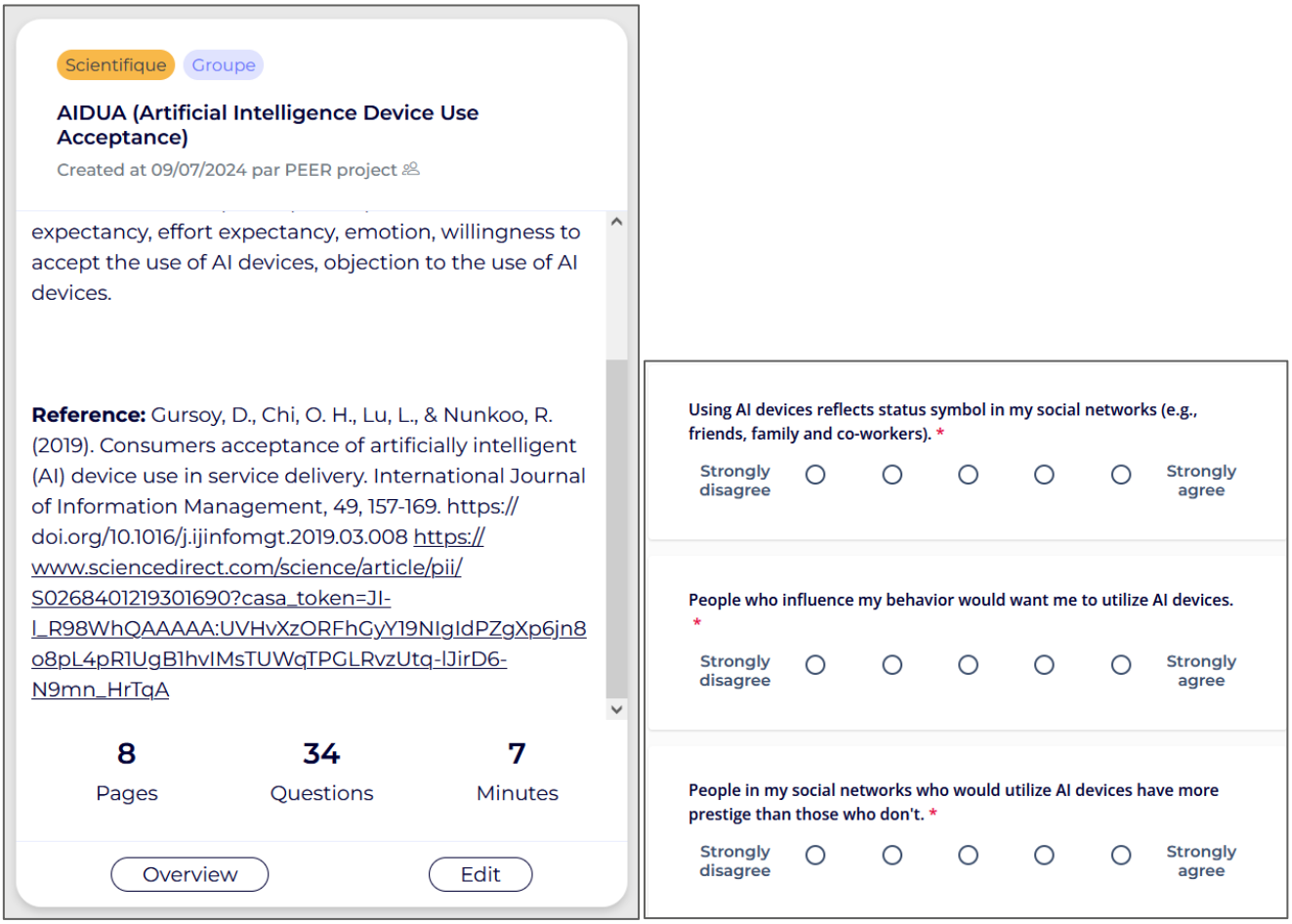


Figure 2: Thumbnail of the AIDUA evaluation tool on the Peac²h platform (left); Testing the 3 first items of the AIDUA evaluation tool (right).

2.3 Partners involved

Amongst the PEER consortium, and to be as transparent as possible with every partner, we held a **monthly meeting to account for the advancements on T4.1**. These meetings where the opportunity to also get **feedback from our partners** within the PEER consortium.

This first task T4.1 of WP4 involved two types of partners across the PEER consortium: SSH partners (section 2.3.1) and the use-cases (section 2.3.2).

2.3.1 SSH partners: imec-SMIT, VUB

As the most contributor of WP4 after CATIE, **imec-Studies in Media, Innovation and Technology (imec-SMIT)** of the **Vrije Universiteit of Brussels (VUB)** was involved in T4.1 since the beginning. Moreover, imec-SMIT, VUB organised the socio-technical requirements workshops (T2.1) in collaboration with CATIE, which represent the bases of the design of the AIA index. To conduct both T2.1 and T4.1 together, we held collaboration work meetings every two weeks during the entirety of Year 1. Imec-SMIT, VUB helped CATIE orientate the bibliography as well as the selection of evaluation tools and user-centred questions of the index.

2.3.2 Use-cases: Proditec, City of Amsterdam, SONAE

The PEER consortium involved 4 use-cases at its beginning (Oct. 2023), but only 3 use-cases were considered in this one-year long task, as the fourth use-case (*i.e.* Continental) was removed from the consortium. These use-cases are heterogeneous in terms of application domains, users and context of use.

- **Proditec (Pessac, France).** The context is improving a computer-vision based pill-sorting machine for the pharmaceutical industry. Users are **operators and supervisors**, in charge of setting the recipe of each type of pill.

- **City of Amsterdam (Amsterdam, Netherlands).** The context is a route-planner for individuals with motor impairments, using a wheelchair. The assistant is meant to assist users during the planning of their trip as well as during the actual trip across the city of Amsterdam. Users are **people with reduced mobility** moving across the City of Amsterdam.

- **SONAE (Porto, Portugal).** The context is the improvement of the customer application by providing grocery itinerary across the shop. Based on the actual grocery list of the user, the assistant will be able to provide an itinerary across the shop as well as a positioning indication of the products within the shop. Users will be **the actual customers of SONAE** using the customer application.

Proditec, City of Amsterdam and SONAE use-case owners were involved in the course of T4.1 to help CATIE select relevant evaluation tools and answer user-centred questions for the actual usage of the AIA index.

2.4 Workplan

	Year 1												Year 2		
	Q1			Q2			Q3			Q4			Q5		
	M1	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**	**M11**	**M12**	**M13**	**M14**	**M15**
Bibliography, State of the art															
Modeling AI acceptance and trust															
Selection of relevant factors															
Identification of evaluation tools															
Selection of evaluation tools															
Iterations with use cases															
Implementation of evaluation tools in Peac²h															
Writing deliverable															
Review of the deliverable															
D4.1												✓			

Table 2. Workplan of the first year of the WP4.

3. State-of-the-art and evaluation tools

The aim of the first year is to lay the foundations for the future design of the AIA index, through gathering **an initial set of transparent and reliable measurement scales for the evaluation of trustworthy AI**. To achieve this, we adopt a gradual process (Figure 3) to settle on a **framework** for the WP4, progressively building on inputs from our review of the literature in which we **explored various notions of the topic**. We settled on a scope covering **five major concepts**, which appear to be the main areas of interest: **acceptability, acceptance, adoption, trust and trustworthiness**. We also **uncovered many sub-concepts**, which we ended up organizing into **ten categories** in light of the literature review. In addition to the bibliography, we also conducted a **benchmark of the existing measurement methods**.

The key elements emerging from our process have been organized into four main sections to ensure the clarity of the D4.1 deliverable:

1. A state-of-the-art on the many notions related to trustworthy AI (section 3.1)
2. A list of the identified measurement methods (section 3.2)
3. The process and result of tools selection (section 3.3)

4. In the section 4, we present various user-centric insights to bear in mind in future WP4 tasks. We collected **feedback** directly from the PEER project use-case owners, and the literature review provided insights into **mitigation variables** (*i.e.*, characteristics that can modulate trust and adoption towards an AI system).

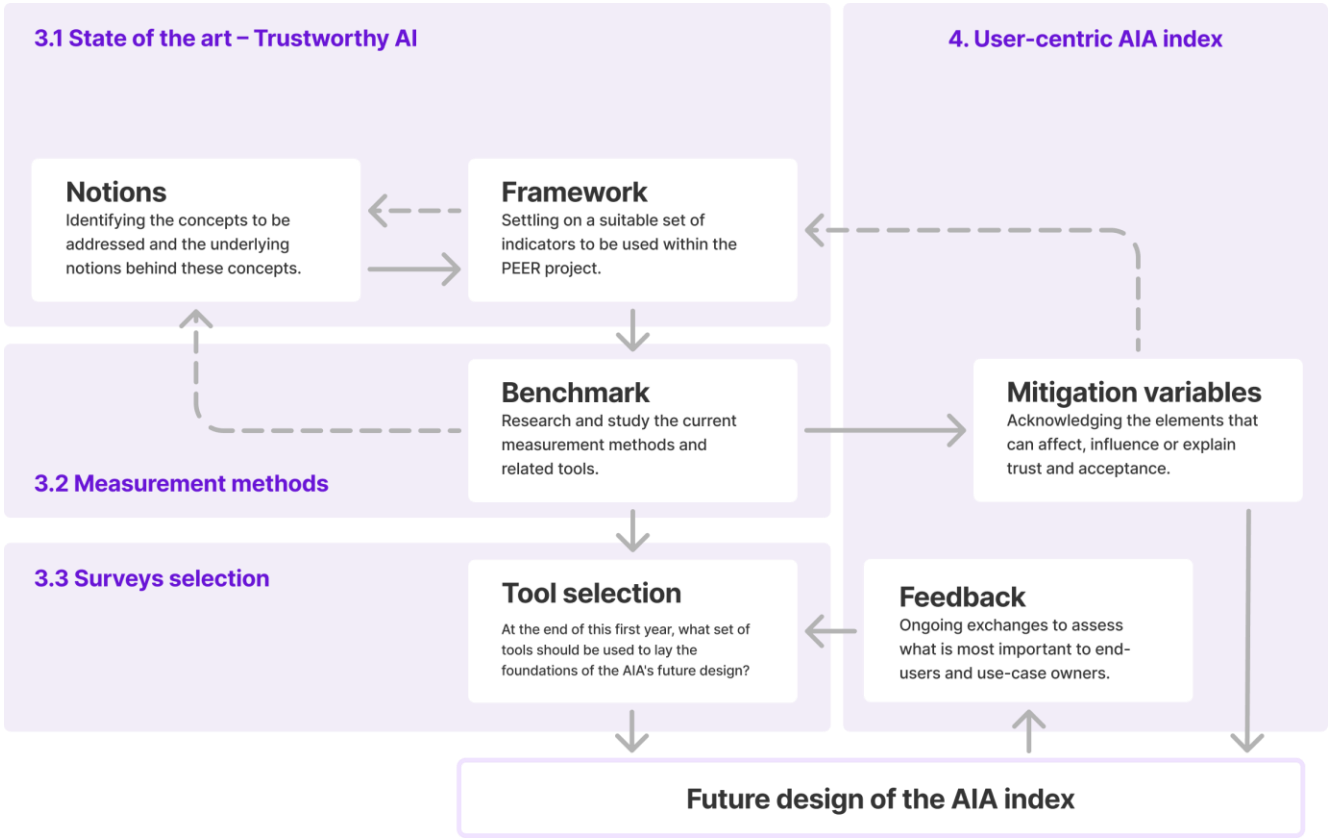
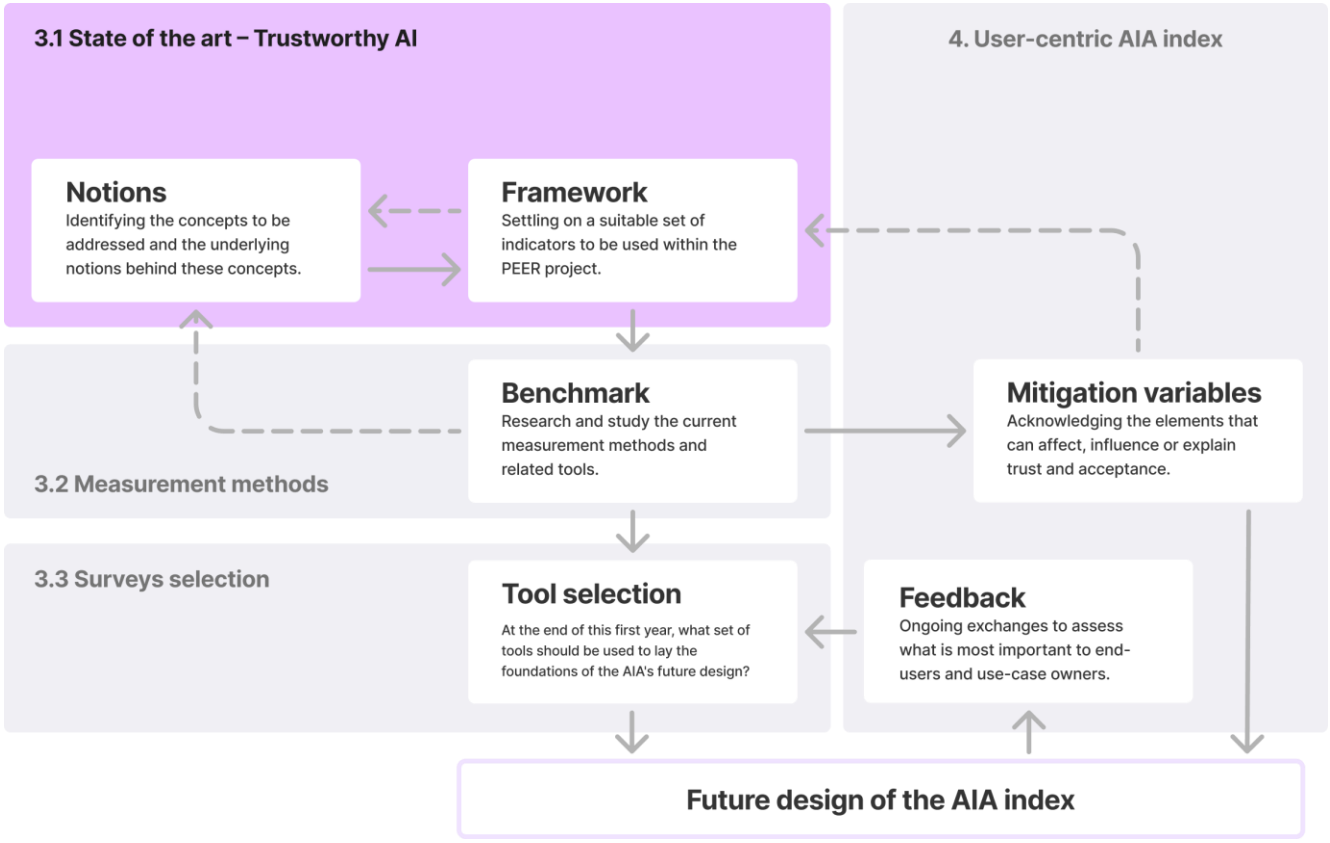


Figure 3. Process followed during the T4.1 of the PEER project.

3.1 State-of-the-art – Trustworthy AI



In this part, we define the concepts targeted in the PEER project proposal and the sub-concepts linked to them.

We settled on a framework comprising five core concepts (section 3.1.1) as well as many sub-concepts (*e.g.* transparency, accountability, collaboration) presented in the paragraphs below (section 3.1.2). Finally, we provide a summary of the different concepts, sub-concepts, and sub-sub concepts (section 3.1.3).

3.1.1 Acceptability, acceptance, adoption, trust and trustworthiness

At the launch of the PEER project, a variety of themes were considered, notably trust and acceptance. We undertook an in-depth exploration of these notions, their development, what they imply and how they are considered in different fields of research. To summarize what was found, we settled on a framework comprising five core concepts, namely:

1. Acceptability (section 3.1.1.1)
2. Acceptance (section 3.1.1.2)
3. Adoption (section 3.1.1.3)
4. Trust (section 3.1.1.4)
5. Trustworthiness (section 3.1.1.5)

The concept of acceptance is complex. In the literature, the terms **acceptability, acceptance** and **adoption** are used, sometimes interchangeably, or sometimes independently. In the **PEER project**, we make a **distinction between acceptability, acceptance, and adoption**. We are interested in **measuring acceptance of AI systems.**

According to Distler et *al.* (2018), Martin et *al.* (2015), Quiguer (2013), **technology acceptability** is one's **perception of a system before use**, while **technology acceptance** is one's **perception of the system after use**. According to Renaud & Van Biljon (2008), **technology adoption** is a **multi-phase process** starting with "*deciding*

to adopt (selecting, purchasing or committing to use it) and then achieving persistent use". In Karahanna et *al.* (1999), a **distinction** is made between "**pre-adoption and post-adoption** (continued use)". Thus, both phases of adoption, and sustained engagement (Doherty & Doherty, 2018) have distinctive characteristics (Nadal et *al.*, 2019) (Figure 4).

Figure 4. Continuum between acceptability, acceptance, and adoption (Rajaonah, 2010)

3.1.1.1 Acceptability

Bobillier-Chaumon (2016) distinguishes **two main complementary orientations of acceptability**, based on different theoretical and methodological paradigms:

1. Social acceptability
2. Practical acceptability

Social acceptability focuses on **the conditions** that make the **new technology and services acceptable (or not) to the user before its actual and effective use** (Terrade et *al.*, 2009). Social acceptability is also considering the **user profile**: age, gender, professional category, the social influence, culture, *etc.* Moreover, **social acceptability emerges not only at individual level, but also at the collective and organizational level**.

The second orientation in acceptability is the **practical acceptability**. This one is interested in making the **new technology more useful, usable and accessible** by the user (Brangier & Barcenilla, 2003). In practical acceptability, acceptance of the system depends on the **ergonomic qualities** of the system, its **ability to fit into a defined context**, and the **user experience** produced when interacting with the system (section 3.1.2.10).

3.1.1.2 Acceptance

The second stage towards adoption is acceptance. **After a first experience with the new technology, the user enters in the acceptance phase**. **This stage is the testing of technology in its context of use**. It makes it possible for the user to concretely **evaluate the advantages, disadvantages, and limits** of the new technology. This phase is important because it is during this stage that the user will **determine his or her interest in this technology in the context of his or her work**. It focuses on the conditions of accepting new practices (or the transformation of old ones) that are linked or induced by the use of the new technology. This more ecological approach makes it possible to assess acceptance in a context and its evolution as it is used. Bobillier-Chaumon (2016) proposed as part of acceptance, **the idea of situated acceptance** including:

- Personal: individual dimension
- Impersonal: organisational dimension
- Interpersonal: relational dimension (human-human)
- Transpersonal: professional and identity dimension

Igbaria & Tan, (1997) underline that acceptance follows a chronological sequence: **first impacting individuals, and then organisations**.

Acceptance goes beyond the intrinsic characteristics of a new technology. It is necessary to take into account the **context of use**, at the individual, collective, professional, organizational and cultural levels (Bobillier-Chaumon, 2003).

3.1.1.3 Adoption

If the acceptability and acceptance stages are successful, then the system has a good chance of being adopted by users and actually used. The final **adoption** is defined by Rogers (1995) as **the willingness of an individual or group to accept and use a new technology**. As trust into the new technology will evolve with the use, **the adoption may evolve too in the good way (adoption and use) or in the wrong way (rejection, or misuse)** (Rajaonah et *al.*, 2014).

This **evolution of use** can be due to **the system which can be well adapted or present limits** which during a prolonged use will become more and more constraining. **External factors can also be at the origin of the evolution of use**. The use of the new technology may become obsolete because the work and tasks to be performed have evolved or new technologies have replaced existing ones. "*It is the use (i.e., the conditions of use - collective, organisational - the user's project and experiences, the social system in which it is implemented) and not only the intrinsic characteristics of the technology that will determine its effects*" (Bobillier-Chaumon, 2003).

To sum up, we need to consider a **continuum of notions** ranging from **acceptability** to **adoption** (sections 3.1.1.1 to 3.1.1.3). Similarly, as described in the following sections, the concept of trust is complex and evolving over time and individuals.

3.1.1.4 Trust

In the **PEER project**, we make a **distinction between trust and trustworthiness,** and we are interested in **measuring trust of the user** (section 3.1.1.4) **and trustworthiness of AI systems** (section 3.1.1.5)**.** We thus present their definition and modelling to help grasp these notions.

Trust is a complex process which involves several components. When discussing trust, the literature (Fulmer & Gelfand, 2012; Mayer et *al.*, 1995; Pirson & Malhotra, 2011) makes a distinction between:

- The ***trustor***: the person who trusts.
- The referent of trust: what or whom the trustor trusts - the ***trustee***.
- The nature of trust: what are the risks, vulnerabilities or dependencies involved in trusting?

The notion of trust has been defined in a variety of ways in the scientific literature, reflecting the different perspectives of researchers and the contexts in which trust is studied. Some examples:

- The willingness of **a party [the trustor] to be vulnerable to the actions of another party** based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party [the trustee] (Mayer et *al.*, 1995).
- **A psychological state** comprising the **intention to accept vulnerability** [to another] **based upon positive expectations** of the intentions or behaviour of another (Rousseau et *al.*, 1998).
- **An attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability** (Lee & See, 2004).

However, there are common elements in the many definitions of trust:

- **Trust is subjective and personal**. It involves an *ascription*, a belief, or taking something for granted. It varies between individuals due to personal differences, such as personality traits, perceptions, past experiences, and cultural background.

- **Trust is conscious**. It involves regulation, intention and willingness, it is about accepting something and is sometimes defined as an ability.
- **Trust is not 100% certain**, it revolves around uncertainty, chances, predictability and confidence towards the future and users' expectations.
- **Trust requires real-world outcomes**, notably actual words, conducts, actions/behaviours or decisions. These outcomes are evaluated as either beneficial or detrimental to the user.
- **Trust requires actual stakes**. The trustor is an agent who will have to "act" in a meaningful situation, within a context of risk or vulnerability, irrespective of their ability to monitor or control the trustee party.
- **Trust is dynamic**. It can change over time based on interactions and experiences with the system.
- **Trust happens within a relationship**, although it is unclear whether it supposes dependence or interdependence. It can be **mutual** or **directional** - in which case it is directed to the trustor's goals, interests or judgment of importance.
- **Trust likely features a "human-like" trustee**. Many definitions of trust focus on human-human interactions and thus attribute human-like characteristics to the trustee, notably motives, intentions, and the faculty to have qualities such as goodness, benevolence, integrity or willingness not to exploit someone.

To sum up the context of the AIA index, we settle on the idea that trust is a subjective psychological state that influences how likely a person is to rely on the system, and that can be influenced by a variety of factors including past experiences, perceived reliability, and personal disposition. **Trust is based on attitude, positive expectations, and vulnerability** (Figure 5). **Trust is requested in collaborative exchanges characterized by uncertainty**.

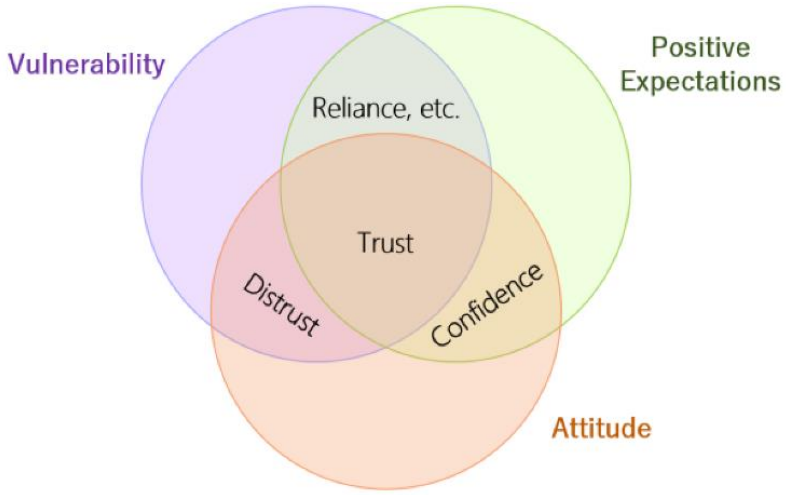


Figure 5. Constructs related to trust (Vereschak et al., 2021).

Trust is modulated by expectations of the user

When we talk about expectations, we need to consider the disconfirmation of these expectations. It is the gap between the user's initial expectations and the performance perceived during the use of the system (Oliver, 1993).

The intention to continue to use a system is based on 5 stages:

1. **A priori expectations** of the system, before it is used, based on information received.
2. **The user's projection** of how the system will be used.
3. **The difference between the user's initial expectations and the perceived performance** when the system is used.
4. **The level of satisfaction** (low or high) based on expectations.
5. **The intention to use** the system over time, based on the level of satisfaction.

There are **3 types of expectations** (Liao et *al.*, 2007):

1. What could happen (the imaginable).
2. What should happen (the expected).
3. What we would like to happen (the ideal).

Disconfirmations will be more or less important, depending on the stage and type of initial expectations.

Multiple and complementary models of trust

Several theoretical models have been developed to understand and explain the concept of trust. We decide to detail three of them: the model of Mayer et *al.*, (1995), one of the founding models of the literature, one from automation domain (Hoff & Bashir, 2015), and one from robotic domain (Schaefer et *al.*, 2016). We choose these models to have a little overview of the trust models in the literature across different application domains.

The model of Mayer et *al.*, (1995) (Figure 6) is one of the most influential and widely used models for understanding **trust in organisational relationships**. The model suggests that **trust develops over time** through **repeated interactions** between the parties**.** Trust is influenced by **perceptions of the three components (ability, integrity, benevolence)** and by the **individual's propensity to trust.** The **propensity to trust** is a **personality characteristic** that **reflects an individual's general tendency to trust others**. It varies from person to person and is **influenced by past experiences and personality traits**. Finally, in the model of Mayer et *al.*, (1995), when **trust is established**, it leads to a series of positive behaviours, such as **increased collaboration** and **relationship satisfaction**. **Trust also reduces the need for monitoring and control, facilitating more fluid and effective relationships.**

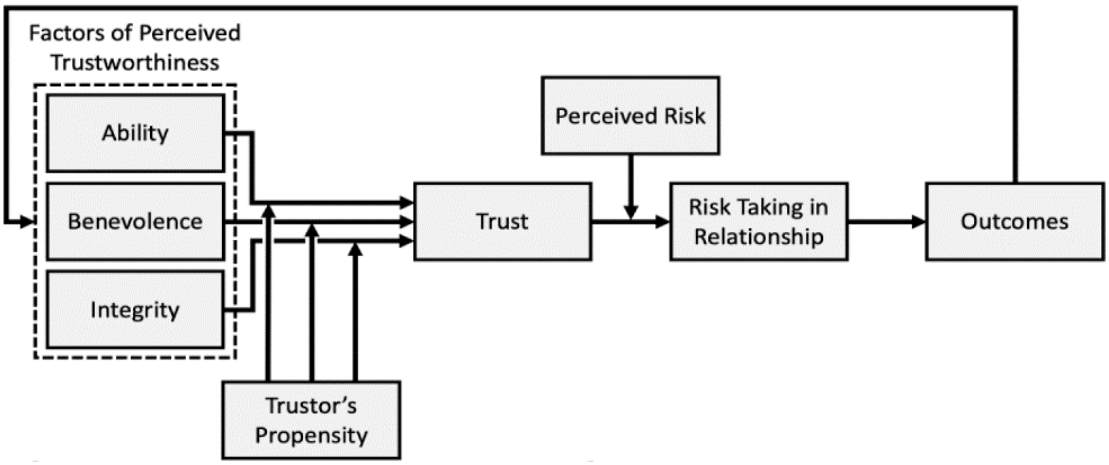


Figure 6. Mayer's organisation model of trust (Mayer et al., 1995).

In **automation domain**, Hoff & Bashir (2015) develop another model that accounts for the **different factors that influence trust in automation**. Automation can be defined as a **technology that performs tasks independently, without continuous input from a user**. Hoff & Bashir (2015) organised **the model in three-layered framework** for **conceptualised the trust variability**: **dispositional trust**, **situational trust**, and **learned trust** (Figure 7).

1. **Dispositional trust** refers to an individual's general tendency to trust automation, independent of context or specific systems.
2. **Situational trust** depends on the specific context of the interaction with the automated system.
3. **Learned trust** develops from past experiences with automation and can be divided into two categories. First, **initial learned trust** is influenced by the operator's preexisting knowledge, including attitudes and expectations towards the system, the system's reputation (see below), past experiences with similar technologies, and understanding of the system. Second, **dynamic learned trust** evolves based on the

system's performance during the interaction (reliability, validity, predictability, dependability, timing of error, difficulty of error, type of error, usefulness). Design features, such as appearance, ease of use, communication style, transparency, feedback, and level of control, play a crucial role in modifying perceptions of the system's performance. For example, an attractive and anthropomorphized interface can increase initial trust, while system transparency and accurate feedback can enhance dynamic trust.

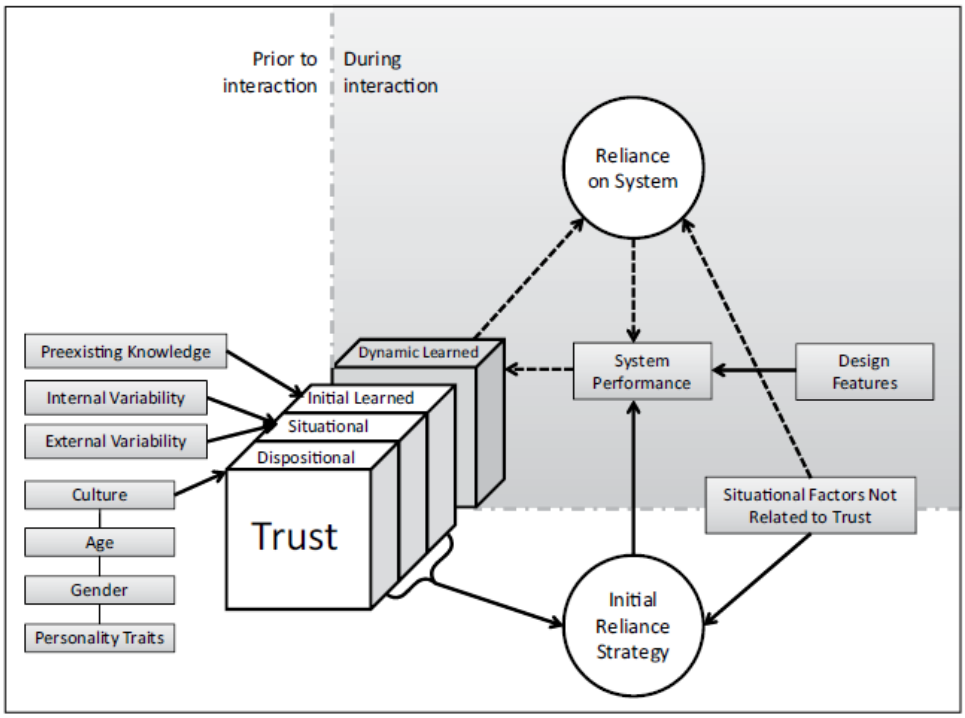


Figure 7. Model of factors that influence trust in automation (Hoff & Bashir, 2015).

As mentioned before, system's reputation influences initial learned trust. The **reputation** of a system is **built on the opinions of the various people who use this system**. Reputation is based on subjective and personal data and is an indicator of trust in a system. A better reputation can increase trust (Hendrikx et *al.*, 2015).

Hendrikx et *al.* (2015) have designed a **system reputation model** (Figure 8). The ***trustor*** is the agent who wants to interact with and trust the system: the ***trustee***. To decide whether or not to trust, **the *trustor* assesses the *trustee*'s reputation, based on previous interactions with him or her, if any**. If this is not the case, the *trustor* requests the opinion of one or more ***recommenders*** who have interacted with the *trustee*. The *recommender* provides a recommendation to the *trustor*, based on the relationship history with the *trustee*. With all the information gathered, the *trustor* can make his or her trust decision.

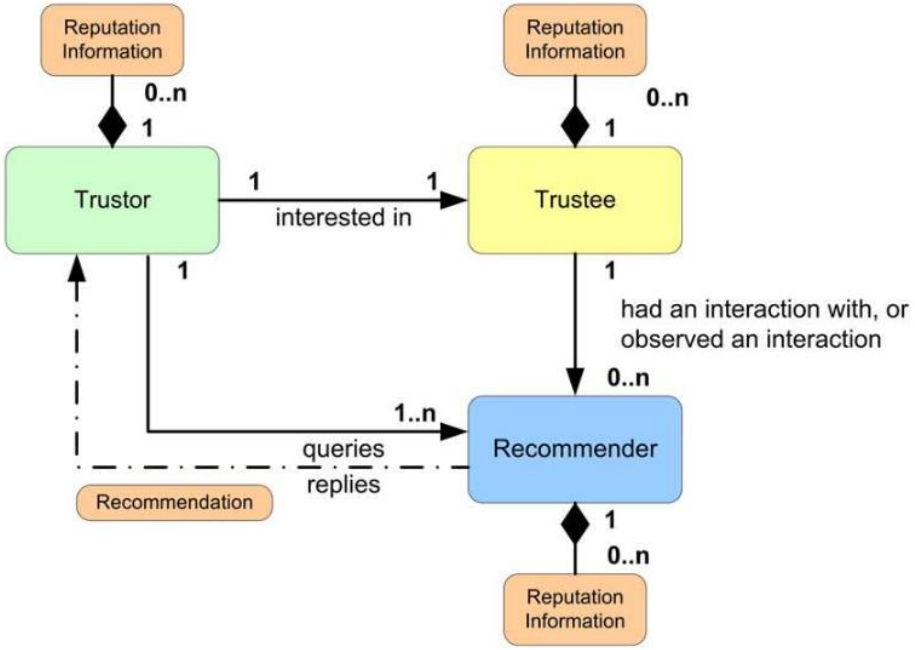


Figure 8. System reputation model (Hendrikx et al., 2015).

In **robotic domain**, the model of Schaefer et *al.*, (2016) shows **the different stages of the trust process** (Figure 9):

- **Development of trust**, which depends on human factors (traits, states, cognitive factors, and affective factors), system factors (characteristics and capabilities) and environmental factors (tasks and team-specific factors).
- **Confidence calibration**: the alignment between the trust that user place in the system and the actual capabilities of that system. Good confidence calibration means that the user's trust is proportional to the system's actual capabilities and performance.
- **Results based on trust** (e.*g.* dependency, compliance, complacency, general use).

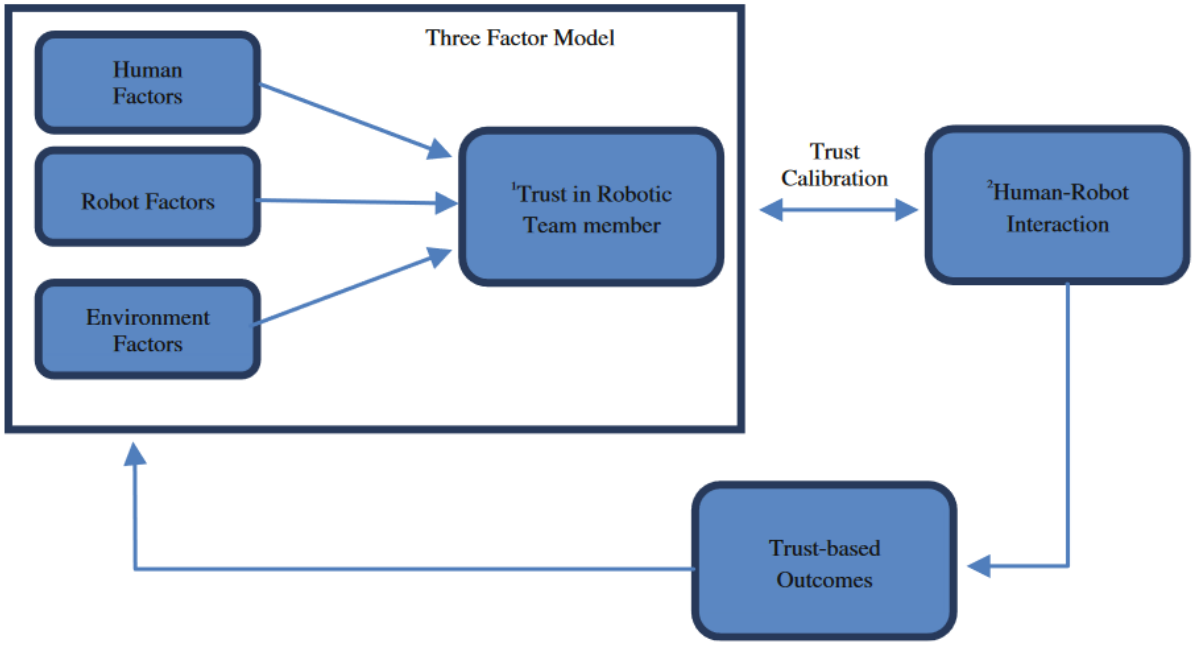


Figure 9. Trust process (Schaefer et al., 2016).

Finally, **trust can be directed towards only a part of the system**. In **human-automation trust** (Lee & Moray, 1992) there was a separation between **performance-based trust** (task execution), **process-based trust** (model integrity) and **purpose-based trust** (designer's intention). More recently, in **AI-human trust**, it was suggested that different perspectives can lead to different trust levels (Starke et *al.*, 2022). Typically, trust can be directed towards:

- **Physical stance**: trust in the reliability or robustness of the system ("Will it break down?").
- **Design stance**: trust the validity or accuracy of the model ("Does it behave in reasonable ways?") or trust in a specific prediction (*i.e.* sufficiently to take some actions based on it).
- **Intentional stance**: trust in the manifested motivation ("Why is the AI acting like this?").

Human-AI trust: can we really *trust* an AI?

Personifying the value of trust in inanimate systems is considered **unconstructive** by some academics who consider that trust requires the ability to make promises, to be capable of motivation, good will, remorse or pride (Starke et *al.*, 2022).

However, it has long been shown that **humans respond to technology socially** and that they tend to **attribute human qualities to inanimate systems**, with norms similar to human-human interactions (Madhavan & Wiegmann, 2007; Miller, 2019). Regardless of the trustee's nature, the human mental model of behaviour explanation relies on belief, desire or intention (De Graaf & Malle, 2017). **Thus, it can be considered reasonable to apply human concepts in human-AI relationship**.

Interestingly, there are **clear differences between interpersonal trust (human-human) and human-automation (or human-AI) trust** (Dzindolet et *al.*, 2003; Rempel et *al.*, 1985). For example, in **interpersonal relationships**, **trust starts with dependability and integrity** and **long-term trust** changes into **faith or benevolence**. However, **trust in automation** is considered to **begin with faith** and **then shift to dependability and predictability as interactions continue**. Additionally, users have **difficulty recognizing dysfunctional behaviour in an artificial agent**: they show the **same level of trust towards biased and unbiased agents** (Van Der Stigchel et *al.*, 2023). **Such attributes that humans project onto technologies regarding trust can be referred as trustworthiness**.

3.1.1.5 Trustworthiness

Trustworthiness is the **interpretation of the information provided by the system based on its characteristics** (Schaefer et *al.*, 2016). The factors of **perceived trustworthiness** are ability, integrity, and benevolence. If trustworthiness is based on a trustee's **ability**, it will depend on how well the trustee performs a task. Trustworthiness founded on a trustee's **integrity** does not depend on the trustee's actual performance, but on the extent to which the trustee's actions correspond to the trustor's values. The trustworthiness based on **benevolence** depends on how the trustee's actions match the trustor's goals and motivations (Lee & See, 2004).

Trustworthiness refers to the inherent qualities or attributes of the system that make it reliable and deserving of trust. It is an objective measure of how well the system performs its intended functions without errors, how transparent it is in its operations, and how well it aligns with the user's expectations and needs (Jacovi et *al.*, 2021).

Some authors propose **design recommendations for creating trustworthy automation**, operationalizing the definition presented above (Hoff & Bashir, 2015):

- **Appearance / anthropomorphism**: increase the anthropomorphism of automation and consider the different factors that influenced dispositional trust. Depending on the profile, anthropomorphism expectations are not the same.
- **Ease of use**: simplify interface, make automation ease of use, and improve feedback automation.

- **Communication style**: have a polite automated system.
- **Transparency / feedback**: provide users with accurate and ongoing feedback about reliability of the system and situation of the environment. In case of errors, provide additional explanations. Indeed, automation confusion is most likely to occur when the automation acts on its own without immediately preceding directions from the operator, the operator has gaps in knowledge of how the automation will work in different situations, weak feedback is provided on the activities of the automation and its future activities relative to the state of the world.
- **Level of control**: adapt the level of control according to user preferences and adapt the transparency according to the level of control.

3.1.2 Sub-concepts behind acceptability, acceptance, adoption, trust and trustworthiness

In the previous sections, we presented the five core concepts of the framework we adopted, namely: acceptability (section 3.1.1.1), acceptance (section 3.1.1.2), adoption (section 3.1.1.3), trust (section 3.1.1.4), trustworthiness (section 3.1.1.5). As seen, there are lot of sub-concepts underpinning acceptability, acceptance, adoption, trust and trustworthiness like reliability, transparency, robustness, collaboration, *etc.*

To categorise these sub-concepts, we follow the seven requirements for trustworthy AI (Assessment List for Trustworthy Artificial Intelligence - ALTAI) (Figure 10) set out by the High-Level Expert Group on AI (AI HLEG)[6] of the European Commission, and we **add new sub-concepts** that do not appear in these requirements, such as **collaboration**, **situation awareness**, **usability** and **user experience**.

Figure 10. Interrelationships of the seven requirements: all are of equal importance, support each other, and should be implemented and evaluated throughout the AI system's lifecycle.

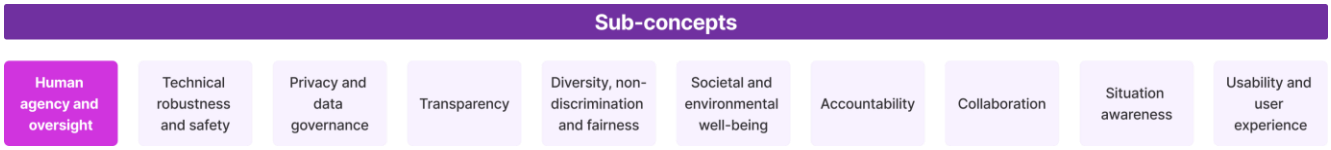
In the following sections, we define the different sub-concepts, and sub-sub-concepts, underpinning acceptability, acceptance, adoption, trust and trustworthiness:

- Human agency and oversight (section 3.1.2.1)
- Technical robustness and safety (section 3.1.2.2)
- Privacy and data governance (section 3.1.2.3)
- Transparency (section 3.1.2.4)
- Diversity, non-discrimination, and fairness (section 3.1.2.5)
- Societal and environmental friendliness (section 3.1.2.6)

[6] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

- Accountability (section 3.1.2.7)
- Collaboration (section 3.1.2.8)
- Situation awareness (section 3.1.2.9)
- Usability and user experience (section 3.1.2.10)

3.1.2.1 Human agency and oversight



In this section, we define the sub-concept of human agency and oversight, their related sub-sub-concepts, and describe a model that explains this sub-concept (HASO model).

Sub-concept	Sub-sub-concepts
Human agency and oversight	Fundamental rights
	Human agency
	Human oversight
	User autonomy

Human agency and oversight, including **fundamental rights**, **human agency**, and **human oversight**, mean AI systems should empower human beings, allowing them **to make informed decisions** and fostering their fundamental rights. At the same time, proper **oversight mechanisms** need to be ensured, which can be achieved through **human-in-the-loop**, **human-on-the-loop**, and **human-in-command approaches** (The High-Level Expert Group on AI, 2020).

Taking human agency and oversight in Artificial Intelligence in consideration is paramount to **ensure ethical, responsible, and safe deployment of new AI-based technologies.**

Human agency refers to the **capacity of individuals to make their own free choices and act independently** (even in a closed referent), while **oversight** involves the **supervision and regulation of AI systems** to align with societal values and norms. "***Agency*** *refers to the thoughts and actions taken by people that express their individual power*" (sociological definition). **Oversight** refers to "*systems or actions to control an activity and make sure that it is done correctly and legally*" (Cambridge Dictionary).

For example, an AI system offering human agency and oversight would offer different options for the same objective, just as a route planner. When using the Waze application, the user can accept or reject a novel route offered by the system during the itinerary.

Human agency and oversight have an impact on the other requirements:

- On ethical considerations: **human oversight is essential to ensure AI systems are used ethically**. This includes preventing biases, ensuring fairness, and safeguarding against discrimination. Human agency empowers individuals to question and challenge AI decisions that might negatively impact them.

- On accountability: **human oversight ensures accountability in AI operations**. By involving humans in the decision-making process, it is possible to trace responsibility for AI actions, thereby avoiding the *black box* problem where AI decisions are opaque and unexplainable for the user.

- On safety and reliability: **continuous human oversight helps in monitoring AI systems to prevent errors and unintended consequences**. This is particularly important in critical sectors like healthcare, finance, and transportation, where AI errors can have significant impacts.

- On control and adaptability: **human agency allows for the adaptation and reconfiguration of AI systems in response to changing environments and societal needs**. Human oversight ensures that AI can be controlled and redirected as necessary, maintaining alignment with human values.

- On transparency and trust: for AI-based systems to be widely accepted, there must be transparency in how decisions are made. **Human oversight (or perceived oversight) contributes to greater transparency, fostering trust between AI systems and their users**.

The **HASO (Human-Autonomy System Oversight) model** (Endsley, 2017), from **automation domain**, is a model that **describes the factors influencing human oversight and intervention in working with autonomous systems** (Figure 11). It considers different key factors: situation awareness, trust, workload, *etc.*

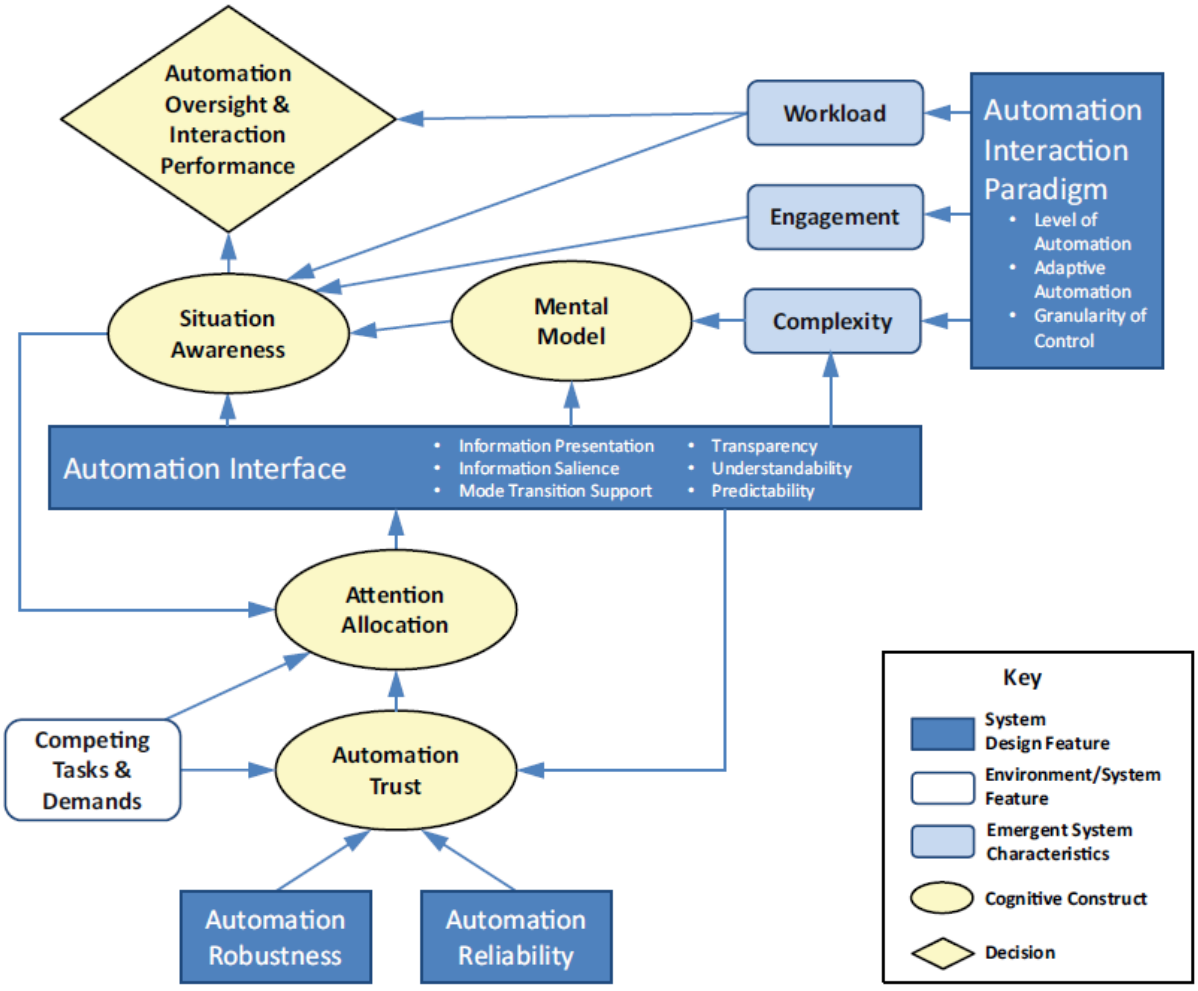


Figure 11. HASO model (Endsley, 2017).

Trust in automation can **decrease the level of monitoring by users**, leading to over-reliance on the system. Balancing trust and vigilance is essential to ensure that users remain attentive to system performance and are prepared to intervene. **Situation awareness is crucial** for users to effectively monitor and intervene in autonomous systems. As systems become more autonomous and reliable, **users tend to lose situation awareness**, and are less prepared to take over manual control when necessary. **It's important to maintain high**

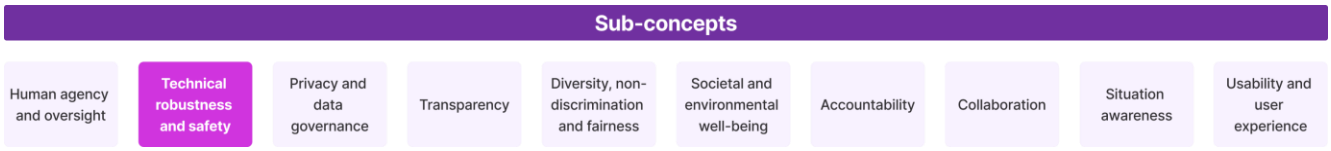
levels of situation awareness to recognize when the system is not performing correctly or when human intervention is needed. **Increased automation can lead to out-of-the-loop issues** where users become disengaged and have difficulty understanding the system's state, making it challenging to take over control effectively.

We need **effective oversight of AI**. AI will not be able to handle many unforeseen (unlearned) situations for the near future. **Synergistic human and AI team is critical to success**: overseeing what system is doing, intervening when needed, coordination and collaboration on functions. **Situation awareness is essential for autonomy, control and interaction**: understanding the state of the system, how it works, what needs to be done and how, how the state of the system affects the operator's tasks and vice-versa, how objectives are being met (Endsley, 2023).

HASO model provides **design features** about the automation interface and the automation interaction paradigm.

- **For automation interface, effective presentation of information is critical.** The interface should provide clear, understandable, and predictable information about the system's state and actions. Salient cues should be used to highlight mode transitions and important system states to enhance situation awareness.
- **For the automation interaction paradigm, different levels of automation affect human engagement and workload differently.** Intermediate level of automation can help maintain operator engagement and situation awareness. To have an adaptive automation and introducing periods of manual control can help maintain user engagement and improve performance. The level of detail in control actions affects workload and situation awareness. Less granular control can reduce workload but may also decrease situation awareness.

3.1.2.2 Technical robustness and safety



In this section, we define the sub-concept of technical robustness and safety, and the sub-sub-concepts related to it.

Sub-concept	Sub-sub-concepts
Technical robustness and safety	Accuracy
	Performance
	Reliability
	Reproducibility
	Safety
	Security

AI-based systems need to be **resilient** to attack and **secure**. They need to be **safe**, ensuring a fall-back plan in case something goes wrong, as well as being **accurate**, **reliable**, and **reproducible**. That is the only way to ensure that also unintentional harm can be minimized and prevented (The High-Level Expert Group on AI, 2020).

In the realm of AI, ensuring **robustness** and **accuracy** is paramount, especially in **safety-critical** applications. For instance, Tesla's full self-driving mechanism mistaking the moon for a yellow traffic light, or autopilot being fooled by stickers on the ground, highlights the susceptibility of AI-based systems to errors and external attacks. Such vulnerabilities can lead to undesirable behaviour and decreased performance. Therefore, it is crucial to design reliable systems for safe integration, particularly in areas like medical diagnosis. Extensive research

focuses on developing and testing robust AI-based systems, addressing all phases of the machine learning pipeline from data collection to model prediction. These methodologies are applied across various tasks, including image classification, object detection, and text classification (Tocchetti et *al.*, 2022).

The **robustness** of an AI system includes **technical robustness** and **robustness from social perspective**. For an AI system to be **technically robust**, it must be **adapted to a context**. For an AI system to be **socially robust**, it must **take into account the context and environment in which it functions** (The High-Level Expert Group on AI, 2019). **AI robustness** refers to the **capability** of AI systems **to handle errors during model training or inference**. A **model** is deemed **robust if it consistently produces accurate predictions**, even when input variables or assumptions are altered unexpectedly. To assess robustness, models should be tested against input variations, such as introducing noise to the test data and modifying its intensity. **High robustness signifies that the model will maintain strong performance** with new data and various noise sources (Wei & Liu, 2024).

Accuracy is used to **assess the predictive ability of the AI model**. An AI model is trained on one set of data and then tested on another set of data unknown to the system. **Accuracy is evaluated on unknown data to assess the generalisation ability of the model.** Accuracy is defined as the ratio of correct classifications to the total number of classifications made by an AI. **Accuracy represents the number of values predicted correctly** (Sanderson et *al.*, 2023). In AI literature, in order to capture complex data relationship, model such as deep neural networks require a high number of parameters which would result in the model being opaque (Figure 12). **Finding the right balance between accuracy and interpretability for the AI system design is a challenge** (Arrieta et *al.*, 2019).

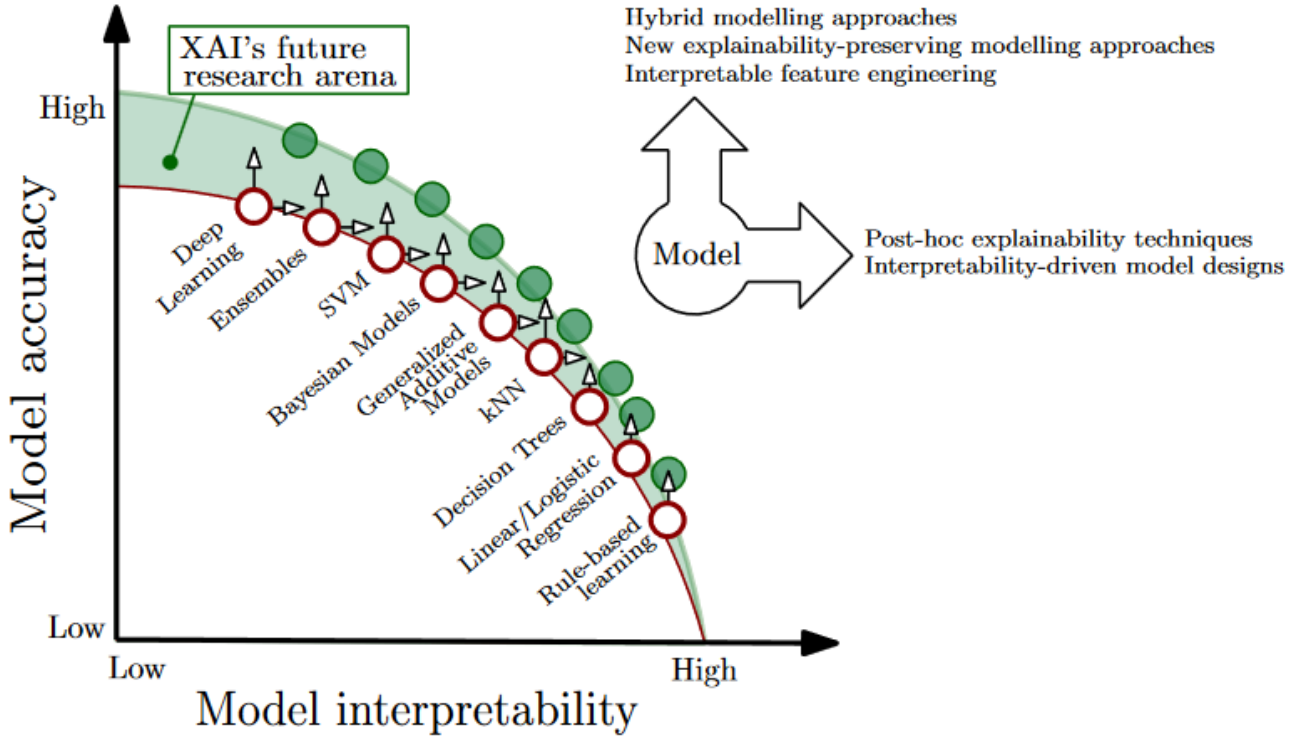


Figure 12. Trade-off between model interpretability and accuracy (Arrieta et al., 2019).

The **performance** of an AI system can be measured by looking at **how effective it is when faced with new unknown data**, compared with the data on which it has been trained and tested (Wing, 2021).

Reliability is the ability of a **system to function as expected**, without failure, in a given context and for a given period. An AI system is said to be reliable if it behaves as expected, even for novel inputs on which it has not been trained or tested earlier (The Confiance.ai program, 2022). **Technical reliability** refers to the **consistency**

or reproducibility of measurements: do repeated measurements give the same result, are the results stable. **The smaller the variations in results are, the higher the reliability is** (Bruton et *al.*, 2000).

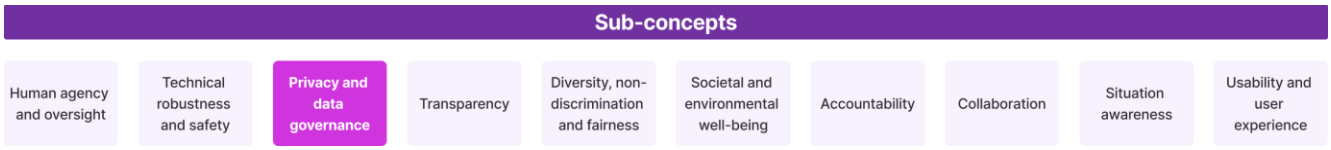
Reproducibility indicates **whether an AI experiment shows the same behaviour when repeated under the same conditions**. This reproducibility of computational procedures, and therefore of data, is essential to ensure that the AI system works properly. This verification makes it possible to detect, analyse and reduce potential risks to the AI system and improve its reliability (Li et *al.*, 2023).

The **safety** and **security** of the AI system must be considered to avoid the system causing damage to others, and to **protect the system against attacks**, such as data poisoning (The High-Level Expert Group on AI, 2019).

For example, a robust and safe AI system would offer the same (or better) performances when aggregating new data and offer features to protect its users. The Waze application asks the user to confirm if an element is still on the road and remind the user to avoid using his or her smartphone while driving.

AI has limited reliability and robustness. An AI system has perceptual limitations, it continues to struggle with reliable and accurate object recognitions in noisy environments. It only capable in situations that are covered by its training, learning "lag" (brittleness). It has hidden biases from using a limited set of training data, or from biases within that data itself (Endsley, 2023). AI cannot use reason to understand cause and effect, it cannot predict future events, simulate the effects of potential, reflect on past actions, or learn when to generalize to new situations (no model of causation) (Pearl & Mackenzie, 2018).

3.1.2.3 Privacy and data governance



In this section, we define the sub-concept of privacy and data governance.

Sub-concept	Sub-sub-concepts
Privacy and data governance	Privacy
	Data protection

Privacy and data governance mean besides **ensuring full respect for privacy and data protection**, adequate data governance mechanisms must also be ensured, taking into account the **quality and integrity of the data**, and ensuring **legitimised access to data** (The High-Level Expert Group on AI, 2020).

For example, an AI system ensuring privacy and data governance would provide its users with the information it collects, how it stores this data, and let its users which data it can collect.

There is a **design trade-off between** extent of **privacy** by being in control over personal data vs. degree of **smartness** provided by a smart system or service (Figure 13). Tricky trade-off, because a system can be smarter the more data it has available. People should control the trade-off between the benefits received and the data provided (Streitz, 2019).

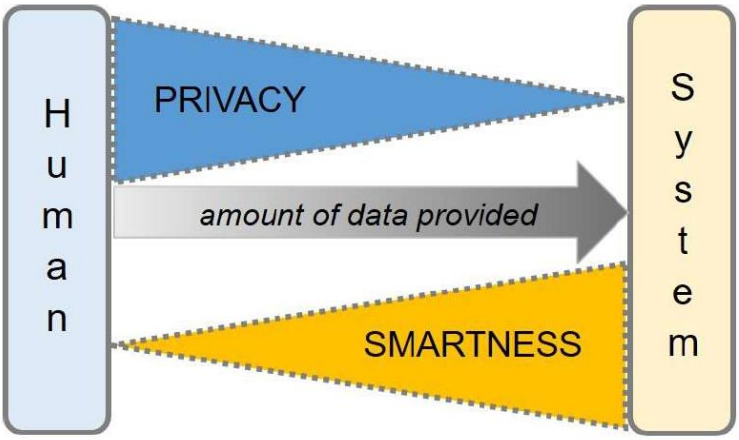
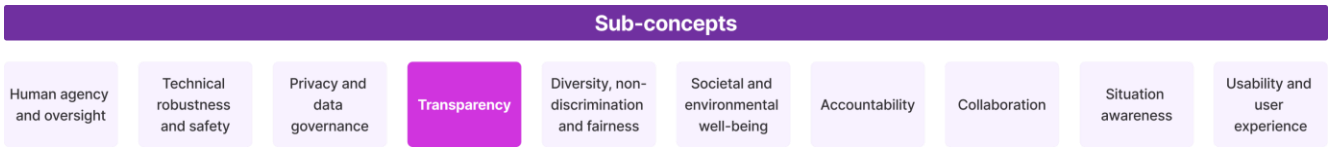


Figure 13. Design trade-off between privacy by control over personal data vs. degree of smartness provided by a smart system or service (Streitz, 2019).

3.1.2.4 Transparency



In this section, we define the sub-concept of transparency, and the sub-sub-concepts related to it.

Sub-concept	Sub-sub-concepts
Transparency	Traceability
	Explainability
	Understanding
	Interpretability
	Intelligibility
	Communication

Transparency means the data, system and AI business models should be transparent. **Traceability** mechanisms can help achieving this. Moreover, AI systems and their decisions should be **explained** in a manner adapted to the stakeholder concerned. Humans need to be aware that they are **interacting** with an AI system and must be informed of the system's capabilities and limitations (The High-Level Expert Group on AI, 2020).

The **concept of transparency** here encompasses and is underpinned by the concepts of **traceability**, **explainability**, **interpretability**, **intelligibility**, and **communication**, as documented within the scientific literature. However, the concept of transparency can also be understood and cover other concepts of information disclosure, such as non-discrimination and fairness, responsibility and accountability, privacy and data governance. **The benefits of transparency** are to **reduce negative effects of out-of-the-loop**, and **improve performance**, **oversight**, **situation awareness** and **trust calibration**. The **goals of transparency** are **understandability**, **predictability**, **system reliability** and **system robustness**. **Transparency is a key mechanism for supporting situation awareness and shared situation awareness in human-AI teams** (Endsley, 2023).

The **transparency** of a system refers to the **system's ability to make its processes, decisions and actions understandable and accessible to users** (Mohseni et *al.*, 2021).

The **traceability** of a system is **the ability to access all or part of the system's information, throughout its lifecycle, by means of recorded identifications** (Olsen & Borit, 2013). Traceability enables system decisions to be traced through its various components and processing stages (history, modifications, updates, *etc.*).

Explainability is the **system's ability to provide the user with precise and understandable explanations of how it works, what was its reasoning and its decisions**. Explainable AI helps the user to understand why it has taken this decision, or why it recommends a specific action (Mohseni et *al.*, 2021).

Interpretability is the **ability to support user understanding and comprehension of the model decision-making process and predictions**. An interpretable AI is inherently human-interpretable models due to their low complexity of machine learning algorithms (Mohseni et *al.*, 2021).

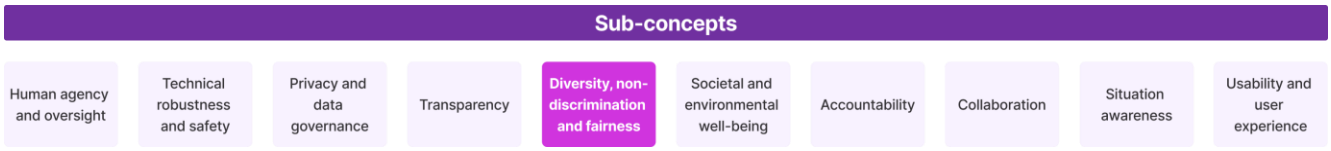
Intelligibility is the ability to be understood by users. An intelligible system implies that **users understand the internal processes, results and decision-making mechanisms** (Lim et *al.*, 2009).

Communication is a process based on **the exchange of information and meaning**. It is through communication that **interaction is established between two entities** (Taylor-Powell, 1998).

For example, a transparent AI system would provide explanations to its choices, the possibilities it considered, a list of its own modifications and their contents, etc.

When it comes to understanding an **explicable output** or an **interpretable algorithm**, users need to form **a mental model** for what the AI system does and what it is doing over time. AI makes it more difficult to develop and maintain an accurate mental model. When humans create mental models, they usually need answers explaining why the AI system did something. This includes causal descriptions (causes and effects), contrast (why A and not B), as well as contextual information and operationally relevant descriptions of how the system will perform in different circumstances.

3.1.2.5 Diversity, non-discrimination, and fairness



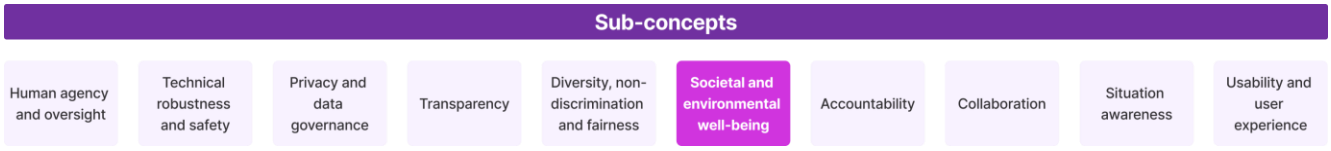
In this section, we define the sub-concept of diversity, non-discrimination, and fairness.

Sub-concept	Sub-sub-concepts
Diversity, non-discrimination, and fairness	Accessibility

Diversity, non-discrimination, and fairness mean **unfair bias must be avoided**, as it could have multiple negative implications, from the marginalization of vulnerable groups to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be **accessible** to all, regardless of any disability with **universal design**, and involve relevant **stakeholders throughout their entire life circle** (The High-Level Expert Group on AI, 2020).

For example, an AI system ensuring diversity, non-discrimination and fairness would offer multiple different ways of interactions (e.g. vocal commands, eye-tracking), or would not require expensive and high-quality technological platforms to operate.

3.1.2.6 Societal and environmental well-being



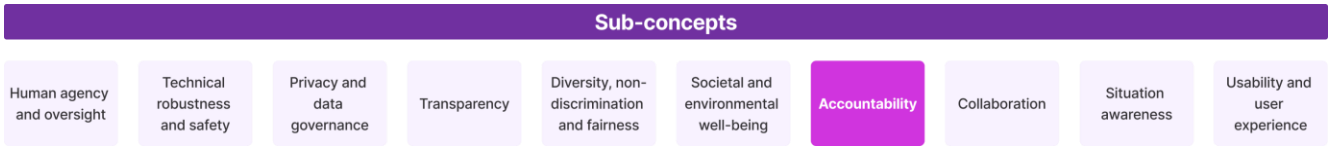
In this section, we define the sub-concept of societal and environmental well-being.

Sub-concept	Sub-sub-concepts
Societal and environmental well-being	Social and societal impact
	Sustainability
	Environmental friendliness

Societal and environmental well-being means AI systems should **benefit all human beings**, including future generations. It must hence be ensured that they are **sustainable** and **environmentally friendly**. Moreover, they should take into account the environment, including other living beings, and their **social and societal impact** should be carefully considered (The High-Level Expert Group on AI, 2020).

For example, an AI system ensuring societal and environmental well-being would promote responsible user behaviours or use renewable energies to operate.

3.1.2.7 Accountability



In this section, we define the sub-concept of accountability, and the sub-sub-concepts related to it.

Sub-concept	Sub-sub-concepts
Accountability	Auditability
	Responsibility
	Error management
	Risk

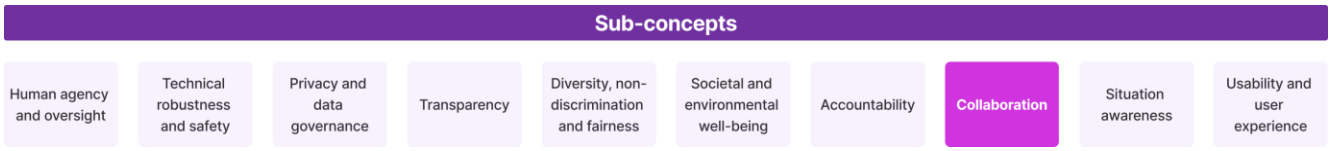
Accountability means mechanisms should be put in place to **ensure responsibility and accountability for AI systems and their outcomes**. **Auditability**, which enables the **assessment of algorithms, data and design processes** plays a key role therein, especially in critical applications. Moreover, adequate an accessible **redress** should be ensured (The High-Level Expert Group on AI, 2020).

Accountability is the "***relationship between an actor and a forum**, in which the **actor has an obligation to explain and justify his or her conduct**, the forum can pose questions and pass judgement, and the actor may face consequences*" (Bovens, 2007). In corporate governance, "*effectiveness involves the **accountability of organizational decision-makers and the legitimacy of decisions** about their economic and non-economic goals and values*" (Aguilera et *al.*, 2008).

For example, an accountable AI system would easily provide all elements for an audit by a public organisation.

Control enables actors to achieve desired and avoid undesired outcomes they are **held accountable for**. **Accountability motivates actors** to act in **alignment with the goals** of a superordinate social system and is thereby a mechanism for **managerial/organizational control**. **Misalignment** results from **control without accountability** or **accountability without control**. These two forms of misalignment are often connected when actors with control transfer accountability to actors without control. Accountability and control may be **aligned across actors** or even **across organizations**. **Sharing of accountability and control** is possible but requires handling of human and exploitation problems (Grote, 2023).

3.1.2.8 Collaboration



In this section, we define the sub-concept of collaboration.

Collaboration is "*a process by which parties who see different aspects of a problem can constructively explore their differences and seek solutions that go beyond their own limited vision of what is possible*" (Wood & Gray, 1991). Although some authors use the terms collaboration, cooperation and coordination interchangeably, Taylor-Powell (1998) proposes to **make a distinction** between communication, contribution, coordination, cooperation and collaboration: **5-C model** (Figure 14) (Shah, 2010). **These five concepts reflect increasing degrees of stakeholder involvement in a joint process**. The establishment of a concept with a high degree of involvement requires the establishment of concepts with a lower degree of involvement. In this way, each concept in the 5-C model supports the implementation of the higher-level concept. **Collaboration is therefore a long-term process that depends on the associated processes of communication, contribution, coordination, and cooperation** (Louvet, 2019).

Communication is a process based on the exchange of information and meaning. It is through communication that interaction is established between two entities. **Contribution** is a communication process, in which the parties provide resources to each other, to achieve their objectives. **Coordination** is a contribution process, in which the sharing of resources is aimed at effectiveness and efficiency. **Cooperation** is a process of coordination, in which resources are shared to achieve a common goal (Taylor-Powell, 1998).

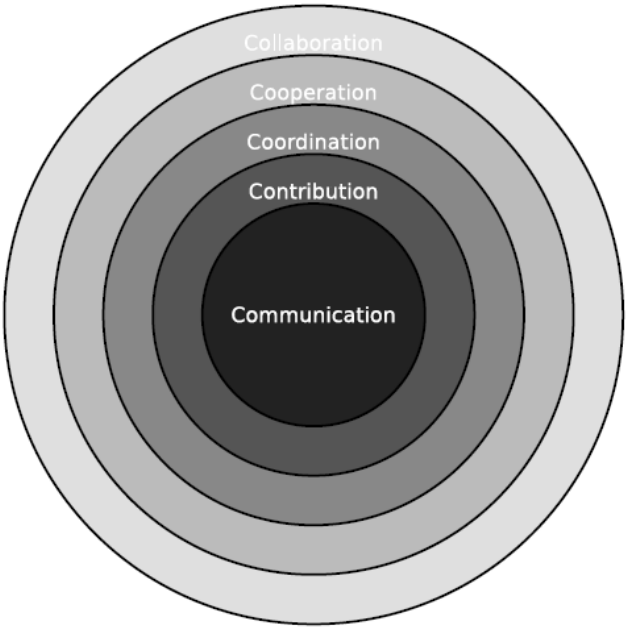


Figure 14. 5-C model (Shah, 2010).

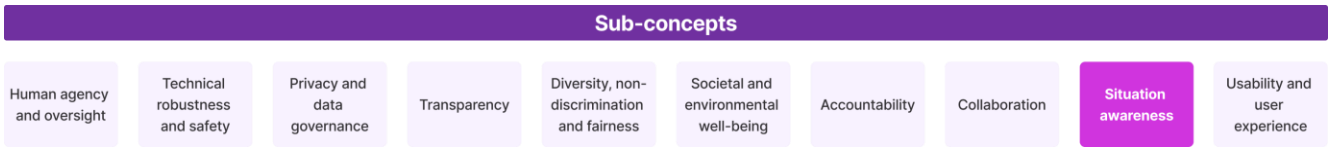
Human-system collaboration can be studied as an interaction. "*Interactions are reciprocal events requiring at least two objects and two actions. They occur when these objects and events influence each other*" (Wagner, 1994). Any exchange that is not one-way between at least two entities and that has an influence on these entities is therefore perceived as an interaction (Louvet, 2019).

Today's systems are increasingly advanced and autonomous. Human-system interaction is no longer limited to delegating repetitive tasks to the system. Interaction takes on more complex forms in which the system can interpret the user's actions and take initiatives to help it achieve its goals (Louvet, 2019).

For example, an AI system which provides collaboration offers the user opportunities to interact, take into account his or her feedback and react accordingly; it works with the user but not in place of him or her.

Collaboration and trust are thus closely linked. Indeed, good collaboration requires a high level of trust, and vice versa, high level of trust requires good collaboration (Neerincx et *al.*, 2006).

3.1.2.9 Situation awareness



In this section, we define the sub-concept of situation awareness.

Situation Awareness (SA) involves perceiving critical elements in the situation, understanding their meaning and projecting into the future (Endsley, 1995) (Figure 15). SA is a critical **dimension of performance**, particularly **in high-risk situations**. Information perceived in the environment is integrated into a mental model aimed at **understanding the situation**. Before executing a response, the user projects how the situation will evolve, since the effect of his or her action is not instantaneous. For users to be able **to anticipate successfully**, they must **have easy access to** information, in particular **key indicators**, be able to project themselves mentally, identify patterns and compare them with experience.

For example, an AI system providing situation awareness of its users allow them to develop and maintain a clear mental model when functioning. It encompasses providing elements helping the understanding of its functioning by users, projections of future states of the system, etc.

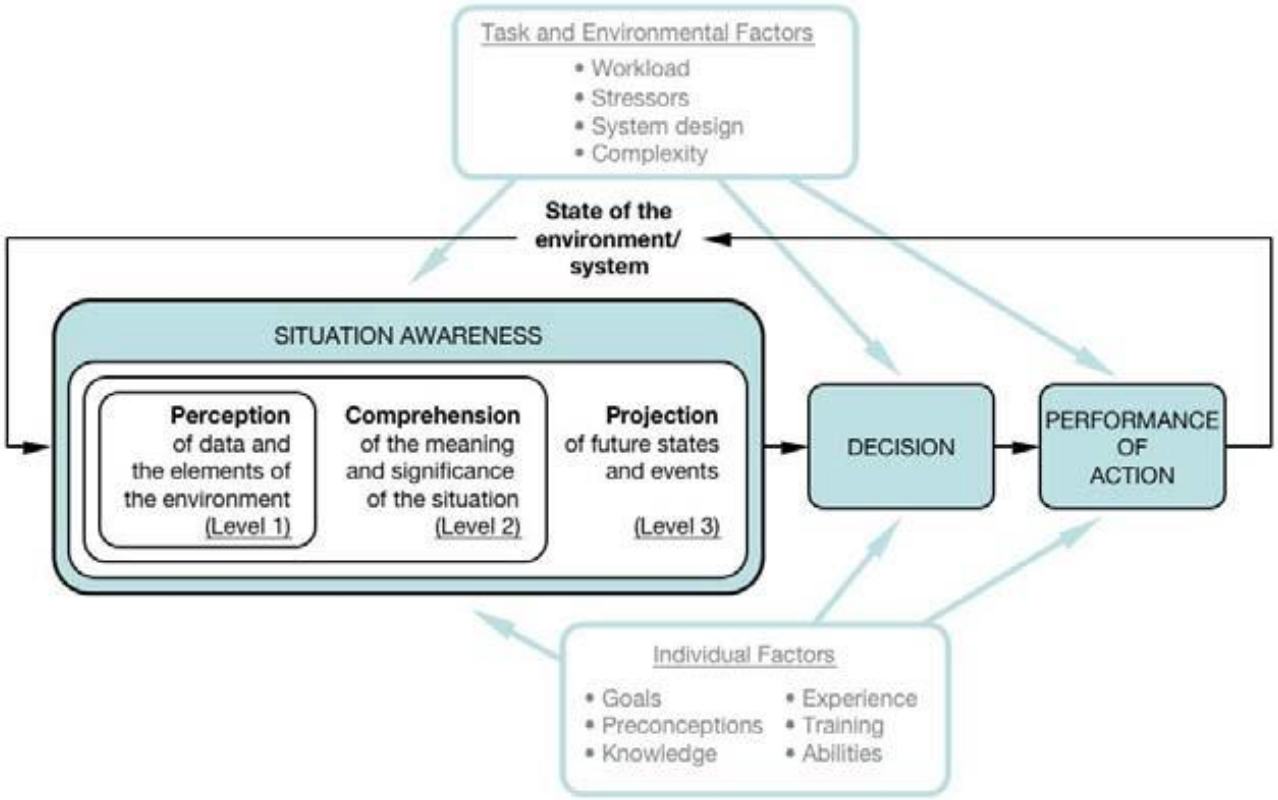


Figure 15. Model of Situation Awareness in dynamic decision making (Endsley, 1995).

Several factors can cause a deterioration or loss of SA: **the Endsley Situation Awareness Demons** (SAD) (Stratmann & Boll, 2016).

- **SAD1 attention tunnelling (SA level 1):** good SA is dependent on switching attention among multiple data streams. Locking in on certain data sources and excluding others is attention tunnelling.

- **SAD2 requisite memory trap (SA level 2):** the working memory processes and holds chunks of data to support SA level 2. The working memory is a limited resource. Systems that rely on robust memory do not support the user.

- **SAD3 Workload, Anxiety, Fatigue, and Other Stressors (WAFOS) (SA level 1 and 2):** stress and anxiety are likely issues in the warning environment. WAFOS taxes attention and working memory.

Workload is the relationship between the time required to perform a task and the time available to perform the task (Sperandio, 1978). In addition to workload, the user's mental workload can be assessed. Mental workload refers to the relationship between the level of resources required to perform a task and the user's actual ability to make these resources available (Moray, 1979).

- **SAD4 data overload (SA level 1):** there is more data available than can be processed by the human "bandwidth".

- **SAD5 misplaced salience (SA level 1):** salience is the compelling aspect of a piece of data, which often dependent on how it is presented.

- **SAD6 complexity creep (SA level 1, 2 and 3):** complexity slows down the perception of information and it undermines the understanding and the projection of information.

- **SAD7 errant mental models (SA level 2 and 3):** wrong mental model may result in poor understanding and projection, so incorrect interpretation of data.

Mental models are internal representations that can support prediction, explanation and add further understanding to potential interactions with objects, people or across tasks. While accurate mental models can improve team operations, incomplete or inaccurate mental models can lead to disastrous consequences such as the Three Mile Island accident. Mental models are often constructed around an individual's previous experience with a system or similar system (Schaefer et *al.*, 2019).

The addition of an intelligent agent, such as an AI, can change the relationships within a team. Knowing the task, roles and capabilities of the team is essential. Any gap, on the part of one of the team members or the AI, between expectations and actual behaviour, can lead to a breakdown in trust. So, we need:

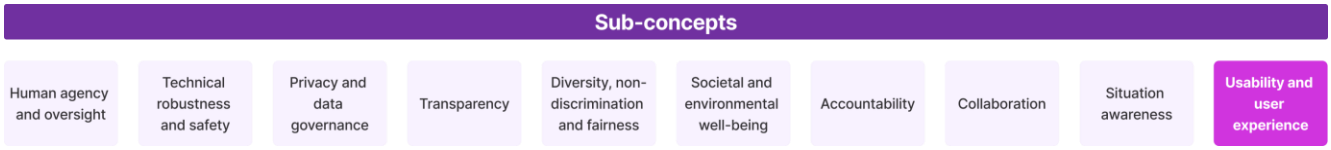
- Training to improve mental models: frequent training on how system works, capabilities, changes.
- Explainable AI: often backwards looking, focused on why (mental model), may be done in low workload periods, pre-mission, post-mission, hard to do in time demanding scenarios.
- Real-time display transparency: real-time support integrated with operator displays, what it is doing and will do (SA), make obvious so don't need to rely on mental models.

Therefore, trust depends on developing appropriate mental models and maintaining shared situational awareness (Schaefer et *al.*, 2019).

- **SAD8 out-of-the-loop syndrome (SA level 1):** for example, automated systems that do not involve the human until there is a problem.

If a system has a high degree of automation, the user who manages it risks finding himself or herself out-of-the-loop. His or her SA will then be degraded, and consequently his or her trust too. It is important to keep the human-in-the-loop and to maintain its SA, particularly in the event of automation failure. If a problem occurs and the user is out-of-the-loop, *i.e.* is not fully aware of the situation, he or she will be unable to diagnose the problem and intervene in time. The level of automation must be balanced against the risk of reduced SA, and must be adapted to the user's capabilities, limitations, and SA. Indeed, people are increasingly unable to perform when they need to take over for automation, because of increases in cognitive workload, reduction of manual skills, and less understanding of what is happening. This increases catastrophic failures (Habib, 2019).

3.1.2.10 Usability and user experience



In this section, we define the sub-concepts of usability and user experience, and the sub-sub-concepts related to it.

Sub-concept	Sub-sub-concepts
Usability and user experience	Effectiveness
	Efficiency
	Satisfaction

Usability

The **usability of a system** is the **degree to which a user can perform his or her task with effectiveness, efficiency, and satisfaction in a specified context of use** . This means that the system's functions are adapted to the user's characteristics, to minimise the gap between human functioning and the system, and to ensure that the system is easy to use. Usability assesses the ease of interaction between the user and the system (Brangier & Barcenilla, 2003).

A system is effective if **users can achieve the objectives**, they have set themself, and it can be considered an **efficient system** if to achieve such objectives, users **consume a minimum of resources**.

Satisfaction is a subjective evaluation of the system. A user is satisfied with the system he or she is using if it is **pleasant to use** (Brangier et *al.*, 2010). But that's not all. For example, according to Seddon (1997) and Rai et *al.* (2022), **perceived usefulness** and **perceived ease of use** of the system are determinants of satisfaction. Furthermore, according to Oliver (1980, 1981), the first impression (satisfaction or dissatisfaction) after an initial use of the system determines whether it will be used for a long time or not. In this sense, long-term use of a system is the result of a satisfactory first experience (McKinney et *al.*, 2002 ; Patterson et *al.*, 1997).

McKinney et *al.* (2002) identified **eight variables influencing user satisfaction**:

1. Relevance of the information
2. Timeliness of the information
3. Reliability of the source
4. Perceived usefulness
5. Accessibility
6. User-friendliness
7. Navigation
8. System interactivity

In some research fields, the concept of usability has been extended over the years to the broader notion **user experience**, to expend beyond efficiency/effectiveness and better account for the non-pragmatic aspects.

User experience

The **User Experience (UX)** (Figure 16) regroups the "***user's perceptions and responses that result from the use and/or anticipated use of a system, product or service***" (ISO 9241-11, 2018) although the definition favoured by the academics (Lallemand et *al.*, 2015) is to describe the user experience as "***a consequence of a user's internal state, the characteristics of the designed system and the context within which the interaction occurs***" (Hassenzahl & Tractinsky, 2006).

User experience is divided into two categories (Hassenzahl & Tractinsky, 2006 ; Thüring & Mahlke, 2007):

1. The perception of **instrumental qualities** (pragmatic)
2. The perception of **non-instrumental qualities** (hedonic)

The **instrumental qualities (pragmatic)** are close to the notion of usability introduced earlier in this section. It includes **effectiveness** (how accurate is the AI), **efficiency** (how effectiveness relates to the effort and time invested), **learnability** (how easy is it for users to operate the system the first times they encounter the AI) and **memorability** (how easy it is for users to re-establish proficiency after a period of non-use).

The **non-instrumental qualities (hedonic)** include psychological well-being of users, *e.g.* induced by **aesthetic, identification,** or **motivating stimulations**.

The experience of instrumental and non-instrumental qualities leads to **emotional reactions**, which shapes the overall **appraisal** of the system.

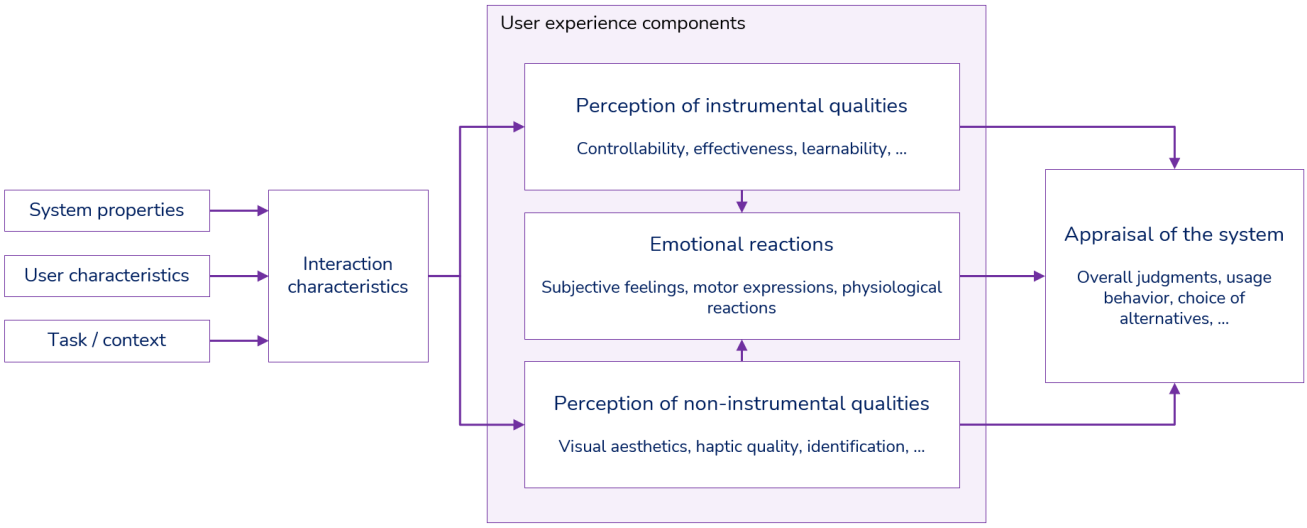


Figure 16. Components of User Experience Model (adapted from Thüring & Mahlke, 2007).

Additionally, **User eXperience (UX)** has multiple facets depending on the time span, with different internal processes taking place in different times (Figure 17) (Karapanos et *al.*, 2010; Roto et *al.*, 2011). This includes **anticipated UX** (imagining experience before usage), **momentary UX** (experiencing during usage), **episodic UX** (reflecting on an experience after usage) and **cumulative UX** (recollecting multiple periods of use over time). Single momentary experiences influence the overall cumulative user experience (Forlizzi & Battarbee, 2004).

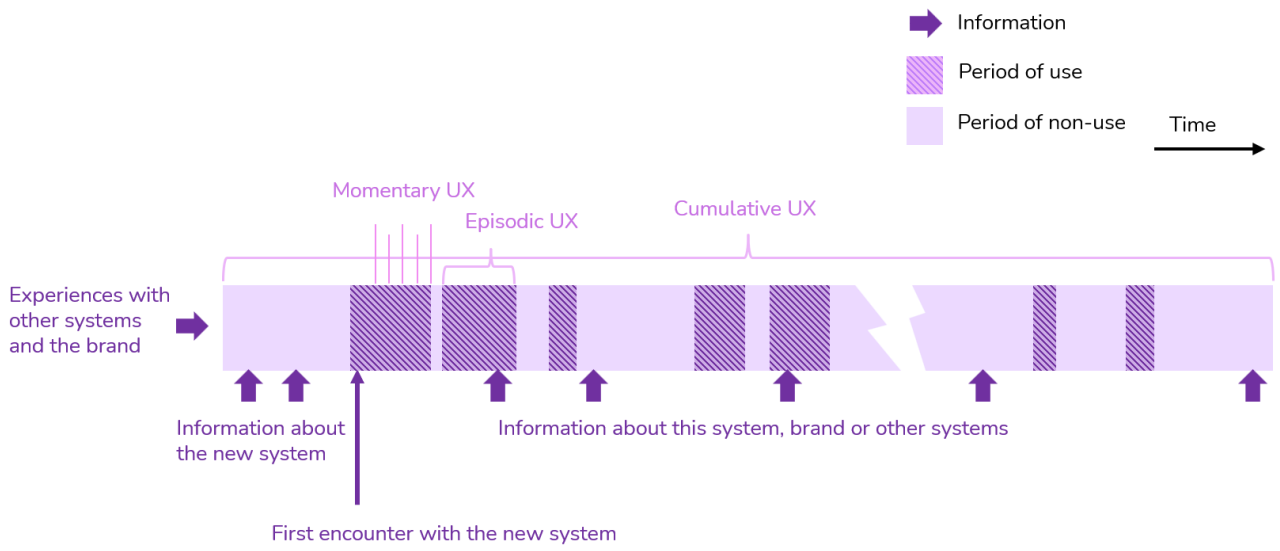


Figure 17. UX over time with periods of use and non-use (adapted from Roto et al., 2011).

For example, an AI system usable and offering a good user experience presents an interface and functioning allowing the user to accomplish his goal task in the easiest, fastest and most satisfying way.

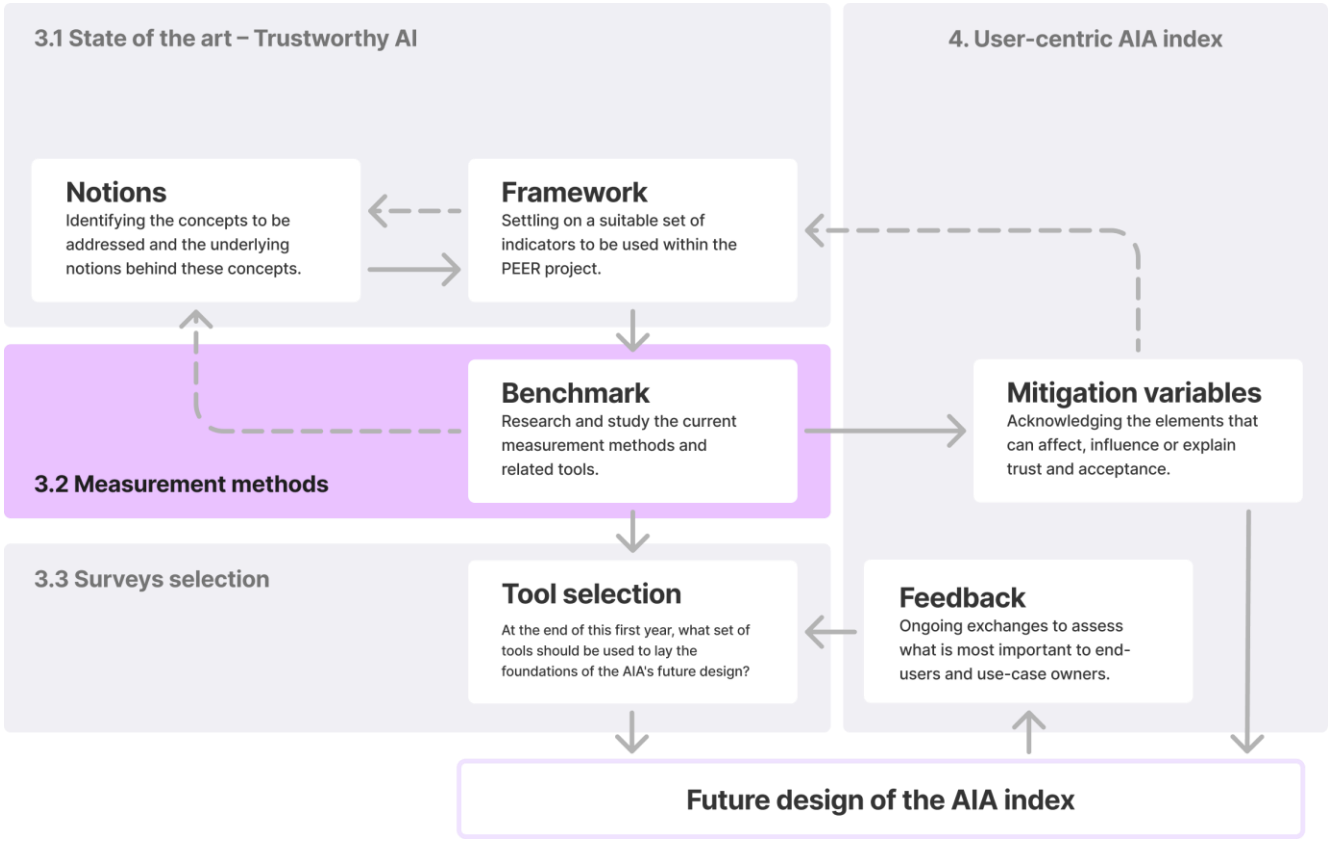
3.1.3 Summary of the different concepts, sub-concepts, and sub-sub concepts

Based on the literature review presented before, we settled on a framework to **list the notions that we should focus on for the PEER project**. This framework includes 5 concepts, 10 sub-concepts, and 29 sub-sub-concepts that might be needed in out next steps to provide an evaluation of AI-based systems (Table 3).

Concepts	Sub-concepts	Sub-sub-concepts
Acceptability Acceptance Adoption Trust Trustworthiness	Human agency and oversight	Fundamental rights
		Human agency
		Human oversight
		User autonomy
	Technical robustness and safety	Accuracy
		Performance
		Reliability
		Reproducibility
		Safety
		Security
	Privacy and data governance	Privacy
		Data protection
	Transparency	Traceability
		Explainability
		Understanding
		Interpretability
		Intelligibility
		Communication
	Diversity, non-discrimination, and fairness	Accessibility
	Societal and environmental well-being	Social and societal impact
		Sustainability
		Environmental friendliness
	Accountability	Auditability
		Responsibility
		Error management
		Risk
	Collaboration	
	Situation awareness	
	Usability and user experience	Effectiveness
		Efficiency
		Satisfaction

Table 3. Summary of the different concepts, sub-concepts and sub-sub-concepts related to acceptance and trust.

3.2 Measurement methods



In the following sections, we present the methodology (section 3.2.1) to find the different methods of measurement to assess the different concepts, sub-concepts and sub-sub-concepts related to acceptance and trust:

- Checklists (section 3.2.2)
- Technical objective data (section 3.2.3)
- Behavioural measures (section 3.2.4)
- Physiological measures (section 3.2.5)
- Surveys (section 3.2.6)

This state-of-art of measurement methods is the preliminary phase in the construction of the AIA index. Over the next two years of the project, we will develop the complete methodology to build the AIA index.

3.2.1 Methodology

Data collection methods are usually separated in two categories:

1. Primary data collection (*i.e.* direct collection of new data)
2. Secondary data collection (published sources, books, technical records, public records, *etc.*)

The AIA index is part of the first one. Primary data collection methods include *e.g.* questionnaires, interviews, observations, experimental methods, diaries, process analysis, checklists, *etc.*

We benchmarked the **existing methods, metrics, and tools to measure acceptance-related and trust-related constructs**. Based on the literature review, notably the surveys and meta-analysis of Kohn et *al.* (2021) and Vereschak et *al.* (2021), we gathered a list of 60 ways of measurement, with 46 tools.

We found many different approaches, ranging from single-item questionnaires to large in-depth scales and technical observations. Overall, since acceptance or trust are latent constructs that may not always be directly observable, they require indirect indicators that can take three forms:

- **Predictors** in the process or design (*e.g.* checklists).
- **Correlates**, measured through physiological measures or with observable behaviours (*e.g.* reliance towards the AI suggestions).
- **Direct reports** of the users' subjective perceptions (*e.g.* reported trust in a questionnaire or interview).

A **multimodal approach** that combines these distinct types of measurements can ensure a comprehensive understanding and help mitigate the limitations of each method. Additionally, different tools can be better suited to **different purposes and/or evaluators**. Typically, checklists for designers (and technology providers) and surveys for end-users.

3.2.2 Checklists

Checklists are **methodological tools that enable system designers and technology providers to evaluate the system**. In the PEER project, checklists could be used by the designers and technology providers of the AI systems (WP3). This kind of tools support our idea to have an AIA index that incorporate design-related assessments, and end-user perceptions assessments.

With the ALTAI, the AI HLEG has developed 19 **checklists to guide people who designed the AI system**. Many of the items in these checklists should be used for **assessment during the design phase**, to ensure that the choices made are appropriate. However, these items should also be used to **assess existing AI-based systems** and ensure that their monitoring and disclosure are maintained over time (The High-Level Expert Group on AI, 2020).

The 19 **ALTAI checklists**:

- Human agency and oversight

1. Human agency and autonomy
2. Human oversight

- Technical robustness and safety

3. Resilience to attack and security
4. General safety
5. Accuracy
6. Reliability, fall-back, and reproducibility

- Privacy and data governance

7. Privacy
8. Data governance

- Transparency

9. Traceability
10. Explainability
11. Communication

- Diversity, discrimination, and fairness

12. Avoidance unfair bias
13. Accessibility and universal design
14. Stakeholder participation

- Societal and environmental well-being

15. Environmental well-being
16. Impact on work and skills
17. Impact on society at large or democracy

- Accountability

18. Auditability
19. Risk management

The NASA has also developed some checklists to guide designers in the conception of an AI systems (McLarney et *al.*, 2021).

The 6 **NASA checklists**:

1. Scientifical and technical robustness
2. Security and safety
3. Explainability and transparency
4. Fairness
5. Human-centric and societally beneficial
6. Accountability

In addition of the explanation satisfaction scale for end-users (section 3.2.6.3), Hoffman et *al.*, (2018) develop the **Explanation Goodness Checklist** which is for XAI designers.

Special attention is paid to data protection in the context of AI system. Since 2018, the European Union regulates data protection with the **GDPR (General Data Protection Regulation).** AI system's designers have to make privacy a first-order design objective, not a subsequent add-on, by employing **Privacy Enhancing Technology (PET)** and **privacy-by-design as competitive advantage** (Streitz, 2019).

In complement to the checklists related to "accountability", some techniques can be used by designers and technology providers in the design phase to anticipate the risks. With automation and AI, there is a **transfer of control, or authority, between human and system**. This **transfer includes intent, rules, authorities and other contextual information**. For example, with Waze, there is an intent to transfer the authority to the machine. Waze gives orders / requests actions from the user: choose the itinerary. There are some rules in the program, and some contextual information, that allow to increase the trustworthy.

This **transfer needs to be anticipated in the design phase**, because most **accidents** are **caused by errors of interpretation of information by either the human or the system**. The **STPA (Systems Theoretic Process Analysis),** based on the STAMP (Systems Theoretic Accident Model and Process), is a **hazard analysis technique** (Leveson, 2014). **STPA is an iterative process that consists of identifying the hazardous situations that can lead to an accident, and how these accidents can occur**. The aim is to **determine what mitigations can be put in place at the system design stage**.

There are multiple effects of increasing automation:

- Mix of qualitative overload and quantitative underload for human operators.
- Human operators as stop gap for not yet automated functions.
- Loss of human knowledge.
- **Misfit between accountability and control**

That's why it is important to **keep human-in-the-loop** and use the **method KOMPASS** (Grote et *al.*, 2000). It allows to create a more **holistic and shared design mindset** among technology developers; foster **systematic consideration of work design principles** already in early phases of technology development; facilitate processes of **continuous technology-work co-constitution.**

3.2.3 Technical objective data

The performance of an AI system can be assessed by measuring the **number of hits, errors, misses, false alarms** (Hoffman et *al.*, 2023), and its **response time** (Wing, 2021).

3.2.4 Behavioural measures

During **Human-In-The-Loop (HITL) simulations**, the evaluator can collect direct measures of behaviour/response that may be indicators of the trust like the **decision time** of the user to make an action and the **response time** of the user to respond to a stimulus (Yuksel et *al.*, 2017). **In these simulations, human agency and oversight can be assessed.**

To have a **successful human-AI collaboration**, one of the key elements to be considered is **situational awareness** (section 3.1.2.9). In other words, the user must be aware of the functioning of the system, its current internal state, and be able to forecast the next states of the system. The system, or collaborative agent, must **take the context into account before giving information to the user**. Indeed, if the agent interrupts the user too frequently when he or she is performing tasks for which he or she does not wish to be distracted, the user's work will be slowed down and degraded, leading to an alteration in situational awareness (Louvet, 2019).

Moreover, **the performance of the system affects the performance of the user**, even when the system just makes recommendations. **If the recommendation of the system is correct, the human performance is better**. In contrast, **if the recommendation of the system is incorrect, the human performance is worse**. Users are not independent cross-checkers of system recommendations. **They include system inputs into their decision process**.

Based on the study by Amiel et *al.* (2004), **the user's response** to information given by the system can be considered as another indicator of collaboration. Following an initiative from the system, **if the user responds to the information given, the collaboration can be qualified as good**. **If the information is not processed, collaboration can be described as poor**.

The two indicators - the **system's understanding of the context** and the **user's response** - can be used to **classify initiatives as having good or poor collaboration**. To obtain an **overall measure of collaboration**, a **ratio can be calculated**:

Number of initiatives with good collaboration / Number of total initiatives

According to Cai & Lin (2010), **this ratio corresponds to the trust percentage, which indicates the probability that the user trusts the system**.

Trust % = Number of initiatives accepted by the user / Total number of initiatives provided by the system

In addition, to **assess the situational awareness of the user**, the **Situation Awareness Global Assessment Technique (SAGAT)** (Endsley, 1988) can be used. It consists of asking to the user some questions during the experiment to recall information to access the user's awareness. The user's responses are compared with the actual situation at the time to assess the correctness of his or her SA. It should be noted that the user may have the impression that his or her SA is good when in fact it is not.

The **SAT (Situation Awareness agent-based Transparency) model** (Chen et *al.*, 2014) can also be used. It comes from robotics domain, was developed by the US Army Research Laboratory Robotics Collaborative Technology Alliance (RCTA) and based on Endsley's Situation Awareness model. **The SAT model focuses on the transparency of the requirements needed to understand the task parameters, logic and expected outcomes.** The three levels indicate what is happening and what the system is trying to achieve (SAT level 1), the reasoning process for the system's decision (SAT level 2), and what the user should expect to happen in the future (SAT level 3). **Identifying the correct SAT level for mediating transparency can suggest a trust calibration process** (Schaefer et *al.*, 2019).

3.2.5 Physiological measures

During HITL simulations, the user can be equipped with sensors to measure physiological data, which can be indicators of the trust:

- **Electrodermal activity** (Akash et *al.*, 2018) expressing affective processes such as emotional arousal.
- **Eye-tracking** (Hergeth et *al.*, 2016): if the end-user looks more at the visual area containing information about the automation's process (secondary visual area) than at his or her primary visual, he or she is perceived to have less trust in the automation.
- **Heart rate change and variability** (Waytz et *al.*, 2014) expressing emotional and mental arousal like workload and stress. If a user works with a teammate he or she can trust, his or her workload and stress should decrease.

- **Neural measures** (De Visser et *al.*, 2018) with EEG (ElectroEncephaloGram), fMRI (functional Magnetic Resonance Imaging) and fNIRS (functional Near-InfraRed Spectroscopy) to determine the location and degree of brain activity associated with attitudes and behaviours related to trust.

3.2.6 Surveys

As part of the T4.1, one very common measurement method that can be listed and collected is the survey approach (quantitative questionnaires). The surveys presented below are used to **evaluate a system after it has been used**. They must be **completed by the end-users**. Some tools we found were originally **developed for automation domain**, then adapted to **robotics** and, more recently, to **AI**.

We present surveys to measure:

- Acceptance (section 3.2.6.1)
- Trust (section 3.2.6.2)
- Transparency and explainability (section 3.2.6.3)
- Usability and user experience (section 3.2.6.4)

3.2.6.1 Acceptance

AIDUA (Artificial Intelligence Device Use Acceptance) model (Gursoy et *al.*, 2019) is a perfect example of the domain evolution. It results from the evolution of the **TAM (Technology Acceptance Model)** (Davis et *al.*, 1989) and **UTAUT (Unified Theory of Acceptance and Use of Technology)** (Venkatesh et *al.*, 2003)**,** developed **to explain and predict technology acceptance and usage behaviours of the users**.

Lu et *al.*, (2019) argue that some of the concept of acceptance models, that link perceived usefulness and perceived ease of us, are **not applicable to the context of intention to use AI devices**. Authors have developed a **scale to measure users' willingness to use an AI device**. They identified major predictors: **performance expectancy, effort expectancy, hedonic motivation, anthropomorphism, social influence, facilitating condition, and emotion**. To build this scale, they used UTAUT scale (Venkatesh et *al.*, 2012), SRIW (Service Robot Integration Willingness) scale (Lu et *al.*, 2019), and Wirtz et *al.*, (2018) qualitative study.

In the **AIDUA** model, three factors are identified as critical constructs: social influence, hedonic motivation, and anthropomorphism. **Social influence** refers to the degree that a user's social group (*e.g.* family, friends, *etc.*) believes that using AI devices is relevant and congruent with group norms. Social influence has a significant impact on users' assessment of the costs and benefits associated with AI device use. **The stronger is the social influence, the higher is the benefit perceptions and the lower is the cost perceptions** (Gursoy et *al.*, 2019). **Hedonic motivation** refers to the perceived fun or pleasure an individual expects to receive from using AI devices. Hedonic motivation appears as the main predictor of technology adoption behaviour (Venkatesh et *al.*, 2012). **Anthropomorphism** refers to the level of human characteristics of an object, such as human appearance, self-awareness, and emotion (Kim & McGill, 2018).

3.2.6.2 Trust

Lots of scales to measure user trust have been developed in different domains. These surveys assess the user trust in general. But some focus more on a few sub-concepts of trust.

- In organisation domain:
    - Measures of trust, trustworthiness, and performance appraisal perceptions (Mayer & Davis, 1999) focusing on **employee trust for top management**.

- In HMI domain:
    - Human-Computer Trust Scale (HCTS) (Gulati et *al.*, 2019) focusing on **risk perception**, **competency**, **reciprocity**, **benevolence**, and **general trust**.
    - The effect of anthropomorphism on investment decision-making with robo-advisor chatbots (Morana et *al.*, 2020).

- In automation domain:
    - Checklist for trust between people and automation (Jian et *al.*, 2000) assessing **trust** and **distrust**.
    - Trust in automation scale (Körber et *al.*, 2015) focusing more on the **reliability and intentions of developers**.
    - TOAST: Trust of Automated Systems Test (Wojton et *al.*, 2020) focusing on **understanding** and **performance**.

- In robotics domain:
    - Trust perception scale-HRI (Human Robot Interaction) (Schaefer, 2016), available in long and short versions, focusing on **collaboration**.

- In AI domain:
    - Different facets of trust (Ashoori & Weisz, 2019) focusing on **decision-making process**.
    - Agent and system evaluation (Weitz et *al.*, 2021), to evaluate the **trust in a virtual agent and a speech recognition system**. It is coupled with Jian et *al.*, (2000) scale.
    - Trust scale for XAI (Hoffman et *al.*, 2023) which is specific to the **XAI context** and includes items specific to **decision-making**.
    - AI Literacy Scale (AILS) (Wang et *al.*, 2023), to determine **user competence in using AI technology**.

3.2.6.3 Transparency and explainability

Transparency can be evaluated by end-users with the instrument for measuring user's perception of transparency in recommender systems (Hellmann et *al.*, 2022). This survey covers the aspect of **data** and more specifically the **type of data used to generate recommendations**.

To measure more specifically **explainability of an AI from user's perspective**, Hoffman et *al.* (2023) developed the explanation satisfaction scale. Moreover, during an evaluation, if the participant asks for explanations of how the system works, the **evaluator can use** the curiosity checklist to understand why the participant has asked for explanations. Holzinger et *al.*, (2020) introduce the System Causability Scale (SCS) to **measure the quality of explanations**.

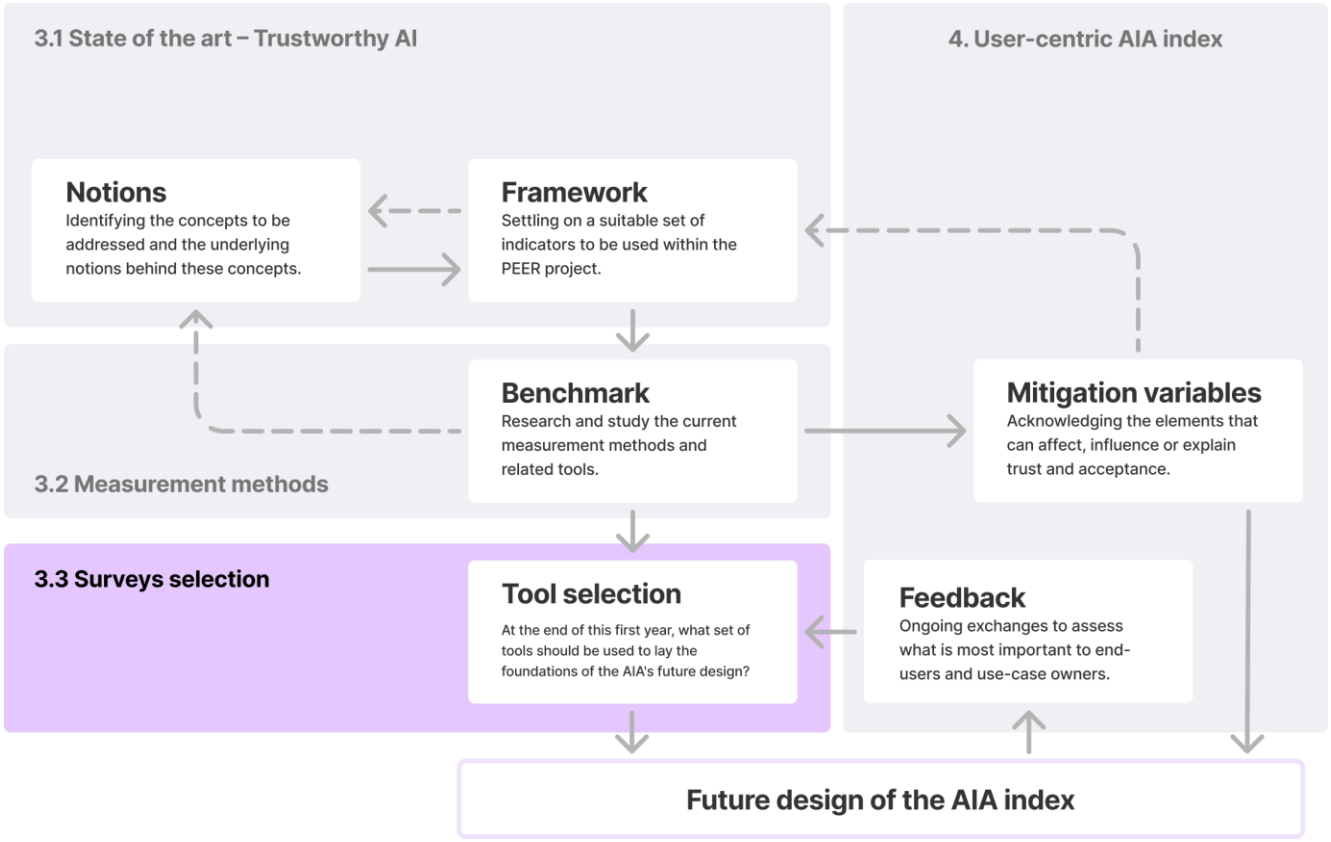
3.2.6.4 Usability and user experience

The SUS (System Usability Scale) (Brooke, 1996) allows to **measure the usability of a system**. Baumgartner et *al.* (2021) develop the Hybrid-SUS (H-SUS) by replacing the 'strongly disagree' and 'strongly agree' at each end of the Likert scale with illustrations. The SUS is one of the first scales to measure perceived usability. This scale is based on the ISO 9241-11 usability standard and is designed to be quick and dirty, *i.e.* quick to fill in and easy to understand. Stetson & Tullis, (2004) showed that the SUS is the most sensitive scale, *i.e.* it can differentiate perceived usability between several systems.

The User Experience Questionnaire (UEQ) (Laugwitz et *al.*, 2008) can also be used. It comprises 26 items, divided into 6 sub-scales: **attractiveness, perspicuity, efficiency, dependability, stimulation and novelty**. Perspicuity,

efficiency and dependability refer to pragmatic aspects, stimulation and novelty to hedonic aspects, and attractiveness to the overall attractiveness of the system.

3.3 Surveys selection



The **first year of the WP4 (T4.1)** called for the **selection of a first set of transparent and reliable measurement scales for the evaluation of trustworthy AI**.

3.3.1 Selection methodology

In the benchmark conducted on the many **methods of measuring** acceptance and trust (and related sub-concepts), we end up with 18 surveys (section 3.2.6). As these scales will be the bases of the AIA index to be developed in Year 2 and Year 3 of the project, we decided to reduce this set by selecting the most reliable and relevant evaluation scales.

In the following paragraphs, we present the selection process and outcomes, including the choice of **methodology** (section 3.3.1.1), our **scoring** approach (section 3.3.1.2), and the **results** and findings (section 3.3.1.3) that led to the final selection (section 3.3.1.4). **A total of 12 surveys are selected and implemented on the Peac²h platform**. The complete list is presented in the section 3.3.2.

3.3.1.1 Defining a selection method

Based on our benchmark, we settled on a **framework** that includes 5 concepts, 10 sub-concepts and 29 sub-sub-concepts (*e.g.* user autonomy, interpretability or error management) (Table 3) – to which we might need to add in the future the concepts identified during the back-and-forth with the use-cases.

To ensure a **comprehensive coverage of all concepts**, the surveys we found were **scored** to rate how well they measure one or multiple concepts.

3.3.1.2 Scoring of each survey

In our benchmark, we reviewed **59 subscales** from 18 surveys. All surveys were gathered in an excel document and scored for each concept / sub-concept / sub-sub-concept using the following system:

- 0 when the subscale does not measure the concept.
- 1 when the concept is measured partially or moderately appropriately.
- 2 when the concept is measured appropriately.

3.3.1.3 Findings of the scoring

Some notions are well measured (lot of subscales): trust, explainability, understanding, reliability, performance, usability and user experience. The least-measured notions, typically when there is only one subscale, were: acceptability, reproducibility, safety, error management, risk, human agency. These notions are linked to the **human factors' domains**. It is therefore logical to have many **tools for assessing these concepts from the point of view of the end-user**.

Notions for which no direct measurement was found are acceptance, traceability, auditability, fundamental rights, human oversight, user autonomy, privacy, data protection, accessibility, sustainability, environmental friendliness. These notions have less connection with the end user's experience. It makes sense to not have evaluation tools from the end-user's point of view. **These notions can be assessed from a technical point of view using checklists** (section 3.2.2 Checklists).

3.3.1.4 A first selection of surveys

For each notion, we selected one tool using the following rule:

- If there is no tool with a "2" on the notion, we select tools with a "1". It is the case just for "safety" and "security" notions.
- If there is no tool with either a "2" or a "1", no tool is selected.
- If there is only one tool with a "2" (or "1") on the notion, then the tool is selected.
- If there are multiple tools with a "2" (or "1") on the notion, we select the one with overall largest number of "2" (or "1") (*i.e.* the most complete).

3.3.2 List of surveys selected

In the following table, we present a set of evaluation tools that addresses assessment of users' trust and/or acceptance. The table presents:

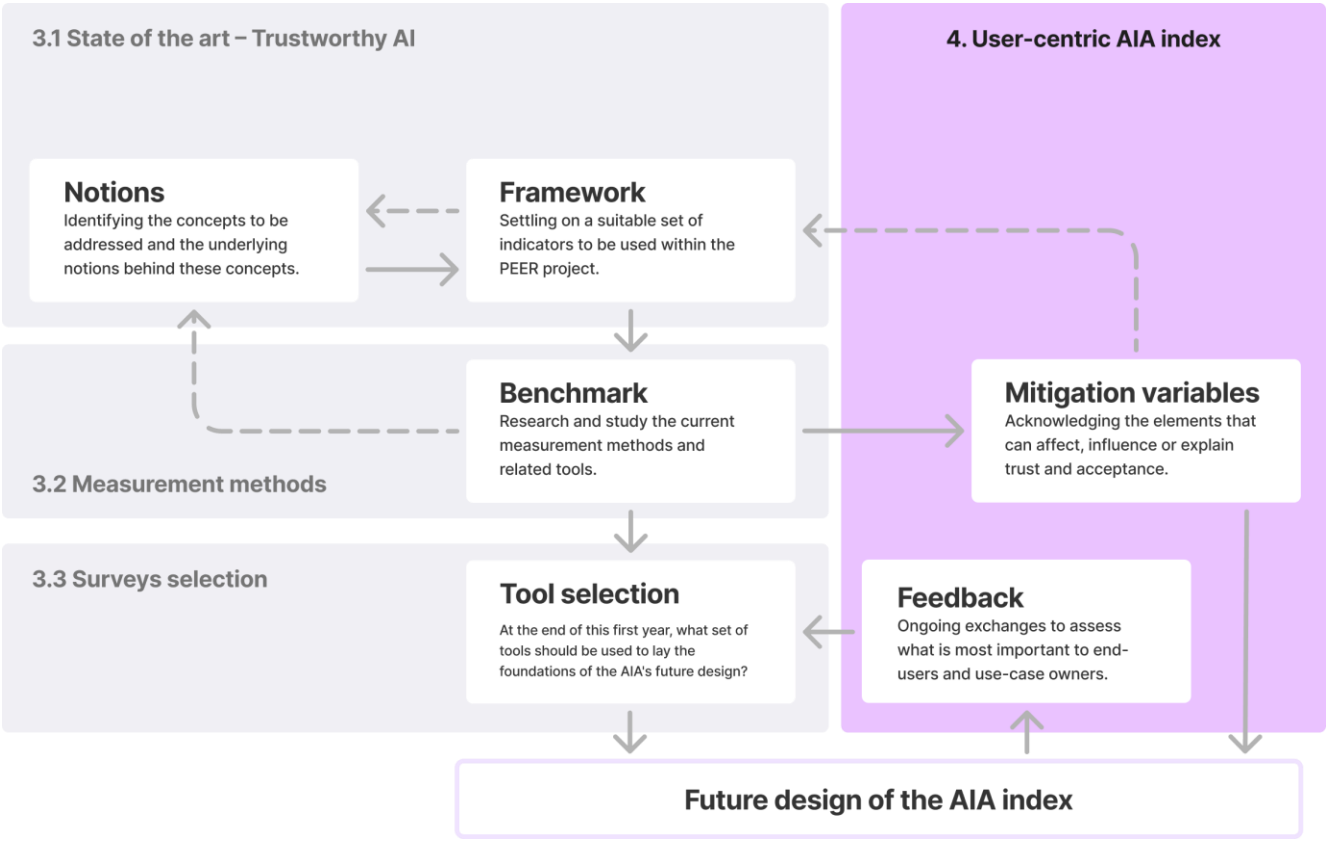
- The **name** of the questionnaire, preceded by an asterisk when this is a scientific validated scale.
- The **domain** from which the questionnaire comes (*e.g.*, business organisation, robotics, *etc.*).
- The **format** of the questionnaire, *i.e.*, the number of questions presented to the respondents (items), and the format of the questions (Likert scale or open question). A Likert scale is a gradation, generally comprising five or seven response options (points), describing the degree of agreement to be qualified: usually from "Strongly disagree" to "Strongly agree".
- The **sub-scales** of the questionnaire. Question sets can yield multiple sub-scores.
- The **link** to access the questionnaire on the Peac²h platform.

For each questionnaire, a score can be calculated. Usually, it is the average of the answers to the questions.

Name	Domain	Format	Sub-scale	Link on Peac²h
* Measures of Trust, Trustworthiness, and Performance Appraisal Perceptions (Mayer & Davis, 1999)	Organisation	41 items 5 points Likert scale		https://app.peac2h.io/surveys/4762/test_protocol
* User Experience Questionnaire (UEQ; Laugwitz, Schrepp & Held, 2008)	UX	26 items 7 points Likert scale	Attractiveness Perspicuity Efficiency Dependability Novelty	https://app.peac2h.io/surveys/4737/test_protocol
* Trust perception scale HRI (Human-Robot Interaction) (short and long version; Schaefer, 2016)	Robotics	Short: 14 items Long: 40 items 11 points Likert scale		Short version: https://app.peac2h.io/surveys/4780/test_protocol Long version: https://app.peac2h.io/surveys/4759/test_protocol
Trust in Automation (TiA; Körber, 2019)	Automation	19 items 5 points Likert scale	Reliability / Competence Understandability / Predictability Propensity to Trust Intention of Developers Familiarity Trust in Automation	https://app.peac2h.io/surveys/4738/test_protocol
Different facets of trust (Ashoori & Weisz, 2019)	AI	14 items 4 points Likert scale	Overall trustworthiness Reliability Technical competence Understandability Personal attachment	https://app.peac2h.io/surveys/4778/test_protocol
* Artificial Intelligence Device Use Acceptance (AIDUA ; Gursoy, Chi, Lu & Nunkoo, 2019)	AI	34 items 5 points Likert scale	Social influence Hedonic motivation Anthropomorphism Performance expectancy Emotion Willingness to accept the use of AI devices Objection to the use of AI devices	https://app.peac2h.io/surveys/4779/test_protocol
* Human-Computer Trust Scale (HCTS; Gulati, Sousa & Lamas, 2019)	HMI	16 items 7 points Likert scale	Risk perception Competency Reciprocity Benevolence General trust	https://app.peac2h.io/surveys/4760/test_protocol
The effect of anthropomorphism on investment decision-making with robo-advisor chatbots (Morana, Gnewuch, Jung & Granig, 2020)	HMI	24 items 7 points Likert scale	Anthropomorphism Social presence Trusting beliefs Disposition to trust in technology	https://app.peac2h.io/surveys/4756/test_protocol
Agent and system evaluation (Weitz, Schiller, Schlagowski, Huber & André, 2021)	AI	22 items 7 points Likert scale + open form questionnaires		https://app.peac2h.io/surveys/4758/test_protocol
* Users' Perception of Transparency in Recommender Systems (Hellman, Bocanegra & Ziegler, 2022)	AI	13 items 5 points Likert scale	Input Output Functionality Interaction	https://app.peac2h.io/surveys/4763/test_protocol
* XAI trust scale (Hoffman, Mueller, Klein & Litman, 2023)	AI	8 items 5 points Likert scale		https://app.peac2h.io/surveys/4739/test_protocol
* AI Literacy Scale (AILS; Wang, Rau & Yuan, 2023)	AI	12 items 7 points Likert scale		https://app.peac2h.io/surveys/4740/test_protocol

Table 4. Recapitulative table of selected evaluation tools for end-users.
** Tools preceded by an asterisk have scientific validated scales.*

4. User-centric AIA index



Now we have a defined framework relating to acceptance and trust (Table 3) and a set of evaluation tools (Table 4), we can move deeper into the reflections around the AIA index.

The design of the AIA index raises several essential questions to ensure that **the AIA index will be robust, usable, useful and actually used**.

- How comprehensive is the list of indicators identified in the literature?
- Who would use the AIA index and what for?
- How to engage stakeholders in building and using the index?
- What are the elements impacting the notions identified in the precedent part?

Answering these three questions requires the **involvement of the use-case owners** in defining and prioritizing the indicators to be used (section 4.1) and there are lessons to be learned from these interviews (section 4.2). At the same time, we also intend to promote this same user-centred design approach **within the measurement tool itself**. We plan to have an **index divided into two parts: one for evaluation by designers or technology providers, and one for evaluation by end-users** (section 4.3)**.** We indicate when in the system's lifecycle the AIA index could be used (section 4.4). Furthermore, accounting for users involves the acknowledgement of usage, context and profiles: the mitigation variables (section 4.5).

4.1 Meeting the needs of the heterogeneous use-cases

To pursue the development of a user-centric AIA index, we discussed with all use-case owners in semi-structured interviews to complement information gathered through the **socio-technical requirements workshops (T2.1)**.

We opened the discussion by reminding the objectives of our task within **WP4**, as we prior documented the factors potentially underpinning the building of trust and acceptance towards an AI technology (*e.g.* transparency, risk management, robustness, *etc.*). We also recalled the context of the meeting and the current state of the whole project: workshops **T2.1**, advancements of **T4.1**, alignment meetings with **WP3**. **The semi-structured interviews were conducted in three steps.**

First, we asked very **wide and open questions** about the AIA index itself, and we collected/discussed their answers without interfering or suggesting any elements which did not come from them. The questions were as follows:

- **What are the objectives purposes of the AIA index?**
- **Who are the users of the AIA index?**
- **What type of information/data should the AIA index produce?**
- **In which context will the AIA index be used?**

Second, we asked the use-case owners about the factors underpinning trust and acceptance (Table 3) towards AI that **they considered mandatory to assess regarding their own use-case**. In other words, we asked what they wanted to know about the end-users' perceptions from using their to-be-developed AI assistant.

Finally, we presented the exhaustive list of factors created based on our literature review work and discussed the list with them. This final step was more of **an open and two-way discussion about the factors they previously mentioned**, the ones they do not want to measure, or assuring what they will/can do with the measurements they thought critically.

The exchanges were highly instructive. The table below (Table 5) summarizes all the key factors identified and their estimated importance (to date) relative to the current needs of the interviewees. For each use-case, importance in the table was coded as **"1"** when the notion was mentioned by the use-case owner as **very important** spontaneously, as **"0,5"** when it was considered **somewhat important**, but mentioned as secondary or needed a cue in the interview, as **"0"** when explicitly considered less important or **not important** and not applicable (**"n.a."**) when the notion was **not mentioned** in the discussion.

Notion			How important is it?	City of Amsterdam	Proditec	SONAE
Human agency and oversight	Self-efficacy, perceived mastery	perceived	important	n.a.	1	n.a.
	Guidance level, quality of guidance	perceived	important	n.a.	1	n.a.
	Reliance	empirical	very important	1	n.a.	1
	Independence, user autonomy	perceived	very important	1	1	1
Technical robustness and safety	Accuracy, performance	empirical	very important	n.a.	1	1
		perceived	secondary	n.a.	0,5	0,5
	Quality of the results	perceived	important	1	n.a.	n.a.
	Repeatability, traceability, reliability/robustness	perceived	secondary	n.a.	0,5	n.a.
		empirical	important	n.a.	1	n.a.
	Security, safety	perceived	secondary	n.a.	0	0
Privacy and data governance	Data governance, privacy	perceived	secondary	0	n.a.	0,5
Transparency	Understanding of the AI model / result	perceived	very important	n.a.	1	1
	Transparency, explainability, interpretability	perceived	very important	0,5	1	0,5
Diversity, non-discrimination, and fairness	Fairness, accessibility	perceived	secondary	n.a.	0	0
Societal and environmental well-being	Social and societal impact, social cohesion, inclusivity	perceived	secondary	0,5	0	0
	Environmental friendliness	perceived	secondary	n.a.	0	n.a.
Accountability	Error management	empirical	important	1	n.a.	1
	Frequency/criticality of risk-taking by the user	empirical	important	n.a.	1	n.a.
	Risk	perceived	secondary	n.a.	n.a.	0
Collaboration	Need adequation, Personalization	perceived	very important	1	n.a.	1
Usability and user experience	Efficiency and effectiveness, task duration	perceived	very important	0,5	0,5	1
		empirical	important	n.a.	1	n.a.
	Simplicity, intuitiveness, learnability, user-friendliness of the overall interface	perceived	very important	1	1	1
	Overall Satisfaction and UX	perceived	very important	1	1	1
	Emotions (stress, doubt, ...)	perceived	important	n.a.	1	n.a.
Summary	Distrust, mistrust, defiance, scepticism	perceived	secondary	n.a.	0,5	n.a.
	Need prioritization	perceived	important	n.a.	n.a.	1

Table 5. Summary of the semi-structured interviews with the use-cases.

Building on this feedback, the first version of the index will incorporate constructs and ideas both from existing literature and from these initial insights gathered from use-case owners. **Further discussions will also be essential. In fact, this table reflects the point of view of the use-case owners. This vision should be compared with the vision of the end-users, which is not necessarily represented in the table above.**

4.2 Lessons learned from meeting the use-case owners

Some guidelines for the prioritization of indicators were outlined (section 4.1), but we also identified key points to remember from our interviews with the use-case owners, grouped into five main takeaways:

- Lesson learned #1: "Understanding" is a trade-off (section 4.2.1)
- Lesson learned #2: On the importance of empirical data (section 4.2.2)
- Lesson learned #3: One step at a time! (section 4.2.3)
- Lesson learned #4: A simple score that makes sense? (section 4.2.4)
- Lesson learned #5: Filling the gaps of unaddressed notions (section 4.2.5)

4.2.1 Lesson learned #1: "Understanding" is a trade-off

Transparency, interpretability and explainability are system-centred notions which are very important when building trustworthy AI. Their user-centred pendant is **understandability** - *how well you can understand what the AI does and why*. A typical question mentioned during our discussions was: "Do the users make choices by chance?". However, actual **understanding** is more intricate, as users should not be **overwhelmed** or **overburdened** by the explanation provided. Some use-case owners expressed **concern that transparency goals could have a negative impact on the user experience**. It might not be suitable for all use-cases to foster real-time awareness of the *how's* and *why's* during the human-AI interaction.

4.2.2 Lesson learned #2: On the importance of empirical data

Although initially focused on creating an index leaning towards **questionnaires and subjective data**, the interviews revealed that the use-case owners consider **empirical data** especially important. They tended to report more **many objective factors** (*e.g.* algorithm robustness indicators) and to report less the perceived qualities of the system (*e.g.* actual robustness perceived by the user) notably regarding:

- Reliance (*i.e.* "do they actually take the route or not?"; "do they pick the product we suggested?")
- Accuracy, performance
- Repeatability, traceability, reliability/robustness
- Error management ("how well the system could actually react to unplanned obstacles")
- Efficiency and effectiveness, task duration

Although the importance and **usefulness** of the human-centred counterpart of these notions is not yet clear from the current perspective of the stakeholders, we still intend to delve deeper into the aspects and measures of the **perceived quality** of AI-based systems *e.g.* perceived performance, perceived robustness, *etc.*

4.2.3 Lesson learned #3: One step at a time!

Apart from the **empirical data** mentioned above (section 4.2.2), the most important notions were:

- Perceived independence, user autonomy
- Understanding of the AI model / result
- Transparency, explainability, interpretability
- Need adequation, personalization
- Efficiency and effectiveness, task duration
- Simplicity, intuitiveness, learnability, user-friendliness of the overall interface
- Overall satisfaction and UX

This highlights the importance of global aspects (user interface, interaction design) which involve the entire system and ultimately extend beyond the assessment of the AI part alone.

On the contrary, as primarily reported by the use-case owners, some aspects (*e.g.* fairness, accessibility, social/societal impact, *etc.*) **are not a primary focus in the first MVPs** (Minimum Viable Product) although these are clearly considered **long-term goals** for the stakeholders of each use-case.

It's interesting to note that **"risk" is regarded by use-case owners as secondary**. However, we insist on the importance of adapting risk management to suit the use-case. The PEER use-cases face very distinct risks and will not need to evaluate their "risk management" the same way.

4.2.4 Lesson learned #4: A simple score that makes sense?

The use-case owners report that one of their primary questions to answer regarding **the overall AI-based system** will be: **"is it good?"**. They require a comparison to a benchmark for simple scores. When discussing how we could make sense of an overall "AI trust and acceptance score", another question that arises is **"what is the most important for the user?"**.

To obtain a score that makes sense, it was mentioned that we would need **weights** (*e.g.* the users values error management at 20% and values privacy at 15%), then use the different ratings to build a satisfaction score.

It was also mentioned that the **temporality of measurements** matters a lot. Typically, the (perceived) quality of the results might be different when received (*e.g.* "how good, acceptable, understandable the routes are") and after use (*e.g.* "post-hoc, how satisfied is the user with the trade-offs?").

4.2.5 Lesson learned #5: Filling the gaps of unaddressed notions

Unaddressed notions could affect the ability of the AIA index to meet stakeholders' expectations. Based on the interview series, we identified unaddressed notions - providing insight into potential adjustments to ensure the AIA index's comprehensive coverage and ultimate success.

Based on the discussions with the use-cases, the key concepts that need to be considered in future steps are:

- Self-efficacy, perceived mastery
- Guidance level, quality of guidance
- Perceived quality of the results
- Need adequation, personalization
- Simplicity, intuitiveness, learnability, user-friendliness of the overall interface
- Emotions (stress, doubt, *etc.*)

Other concepts, less important but still worth of consideration, are:

- Perceived traceability
- Social cohesion, inclusivity
- Perceived risk

These five lessons (sections 4.2.1 to 4.2.5) will help inform the coming design of the AIA index during the next years of the PEER project.

4.3 An index both for end-users and technology providers

A critical aspect of the AIA index design involves understanding who the users of the index **could be** (section 4.3.1), who they **will be** (section 4.3.2) and **who will answer to the AIA index** (section 4.3.3).

4.3.1 Who are the *users* of an AI-based system?

When studying AI-based systems, the notions of trust, trustworthiness, acceptability, acceptance or adoption can be worth studying at different levels and with different stakeholders. However, within the PEER project, it is suitable to settle on a specific typology of user to build the AIA index (T4.2, T4.3). This raises questions: what is the definition of a user and who will be our priority for the upcoming tasks?

In this section, we present the different types of users that we will find in the PEER project use-cases.

Many authors (Glomsrud et *al.*, 2019; Ras et *al.*, 2018; Schoenherr et *al.*, 2023; Vereschak et *al.*, 2021) consider that trust and trustworthiness are characterised by different opinions and goals according to **who is the stakeholder** (*i.e.* what is their role and their relationship with the AI): end-user, technical specialist, owner, (in)direct lay users, regulators, assurance, expert users, developers, researchers, *etc.*). These different stakeholders will **not pay attention to the same thing**, as shown in the citations below.

The end-user: "*The person that ultimately uses or is intended to ultimately use the AI system. This could either be a **consumer** or a **professional** within a public or private organisation. The end-user stands in contrast to users who support or maintain the product, such as system administrators, database administrators, information technology experts, software professionals and computer technicians*" (Schoenherr et *al.*, 2023).

The owner: "*The owner of the software application in which the DNN is embedded. The owner is usually an entity that acquires the application for possible commercial, practical or personal use. For example, an owner can be an **organization** (e.g. a hospital or a car manufacturer) that purchases the application for end-users [e.g. employees (doctors) or clients (car buyers)], but the owner can also be a **consumer** that purchases the application for personal use*" (Ras et *al.*, 2018).

The indirect user: "*Would patients still listen to the doctor if they had known beforehand the doctor is assisted by an AI for diagnosis assessment? Would **citizens** be upset to the same extent about a new bus schedule if it had been created manually instead of with the help of an AI? [...] Automated vehicles research starts to focus on studying trust of indirect stakeholders such as **pedestrians**, because they are also affected by the decisions of direct users*" (Vereschak et *al.*, 2021).

The data subject: *"The data subject is **the entity whose information is being processed** by the application or the entity which is directly affected by the application outcome. An outcome is the output of the application in the context of the use case. Sometimes the data subject is the same entity as the end-user, for example in the case that the application is meant for personal use. The data subject is mostly concerned with the ethical and moral aspects that result from the actionable outcomes"* (Ras et *al.*, 2018).

The developer or researcher: *"Developers, engineers and researchers learn, or verify, improve or make the system comply to requirements"* (Glomsrud et *al.*, 2019).

Finally, **the stakeholder:** *"Stakeholders can include scientists (ethicists, historians, etc.), authors, governments, insurance agencies, etc. They are people or organizations without a direct connection to either the development, use or outcome of the application and who **can reasonably claim an interest in the process**, for instance when its use runs counter to particular values they protect. Governmental and non-governmental organizations may put forward legitimate information requests regarding the operations and consequences of DNNs. Stakeholders are often interested in the ethical and legal concerns raised in any phase of the process"* (Ras et *al.*, 2018).

4.3.2 Which users of the AI-based systems are targeted for the AIA index evaluation?

Among these different types of users (section 4.3.1), **the users targeted by the AIA index will be the end-users** as part of the PEER project (and outside the project itself).

Within each user profile, there are different levels of expertise (Mohseni et *al.*, 2021). On the one hand, **AI novices** (someone unfamiliar with the operation and use of an AI) can be mostly interested in privacy awareness, transparency, AI usefulness or Human-AI performance. On the other hand, **data experts and AI experts** might be more interested in model visualization, information about the training data, model performance and interpretability, task performance, fidelity of the explanation or model trustworthiness.

In the context of the AIA index development, we focus on end-users who can be AI novices.

4.3.3 Who will answer to the AIA index?

Applying **human-centred design** for the AI systems in the different use-cases will require an in-depth consideration of users' needs within the design process - and is not limited to satisfaction, success and performance indicators after implementation. **The AIA index could help to achieve this process by incorporating design-related assessments, independently of the end-user perceptions (that also need to be measured)**. In our interview sessions with the use-case owners, the idea of using a **two-facets evaluation** was positively received: **combining assessments from both designer and technology providers (technical partners) and evaluations from end-users**.

Additionally, it was mentioned that it would be interesting to **check if there is a match** between users' perspective and designer's perspective regarding the same notions.

4.4 Outcomes of the AIA index: for which purpose

During our interviews with the use-case owners (section 4.1 and section 4.2), we discussed different scenarios where the AIA index could be used (and be useful). We describe below **four potential applications for the AIA index**, across various phases of the development and deployment of AI-based system.

1. The AIA index could be used in the **design phase** of AI-based systems. It could provide indicators to guide the iterative design of the different use-cases. It has been mentioned that designers and technology providers could/should use this index to identify and understand users' needs. This would help to incorporate these needs into the systems' functionalities based on the users' priorities.

2. **Beyond the design phase**, some use-cases mention it will function as a tool to assess the overall value of the AI system and identify what needs improvements in future projects (*e.g.* when assessing societal impact).

3. The AIA index can also be of **interest for the end-users** as a way to promote AI transparency.

4. Furthermore, the AIA index can be **used as a Key Performance Indicator (KPI)** to track the evolution of the AI-based systems across different versions of the prototype, allowing for longitudinal analysis over time. Some use-cases mentioned that these KPI can be used to demonstrate improvements in user satisfaction or trust. That way, the AIA index will help engage stakeholders or marketing teams.

4.5 Mitigation variables

Apart from the main targeted notions of the AIA index (*i.e.*, trust, acceptance, *etc.*), it is important to consider certain variables that could influence or distort these measurements. These additional factors, known as

mitigation variables, could account for (and control) potential biases and confounding factors, *i.e.* variables that may unintentionally cause systematic errors or distortions and could lead to inaccurate results or non-representative results. Ignoring them could lead to biased outcomes and diminish the impact of the AIA index.

There are many **mitigation variables** in the literature. To provide a structured list of the identified variables we use a model from human-computer interaction (presented in section 3.1.2.1), highlighting different broad categories (**interaction characteristics**) influencing the **user experience**:

- Human-related variables (section 4.5.1)
- System-related variables (section 4.5.2)
- Task-related variables (section 4.5.3)
- Environment-related variables (section 4.5.4)

We also use an additional category useful for the current topic:

- Factors specific to the human-AI relationship (section 4.5.5)

Based on categories presenting the trust-influencing factors *from* Schaefer (2016), we developed an enriched classification of mitigation variables (Figure 18).

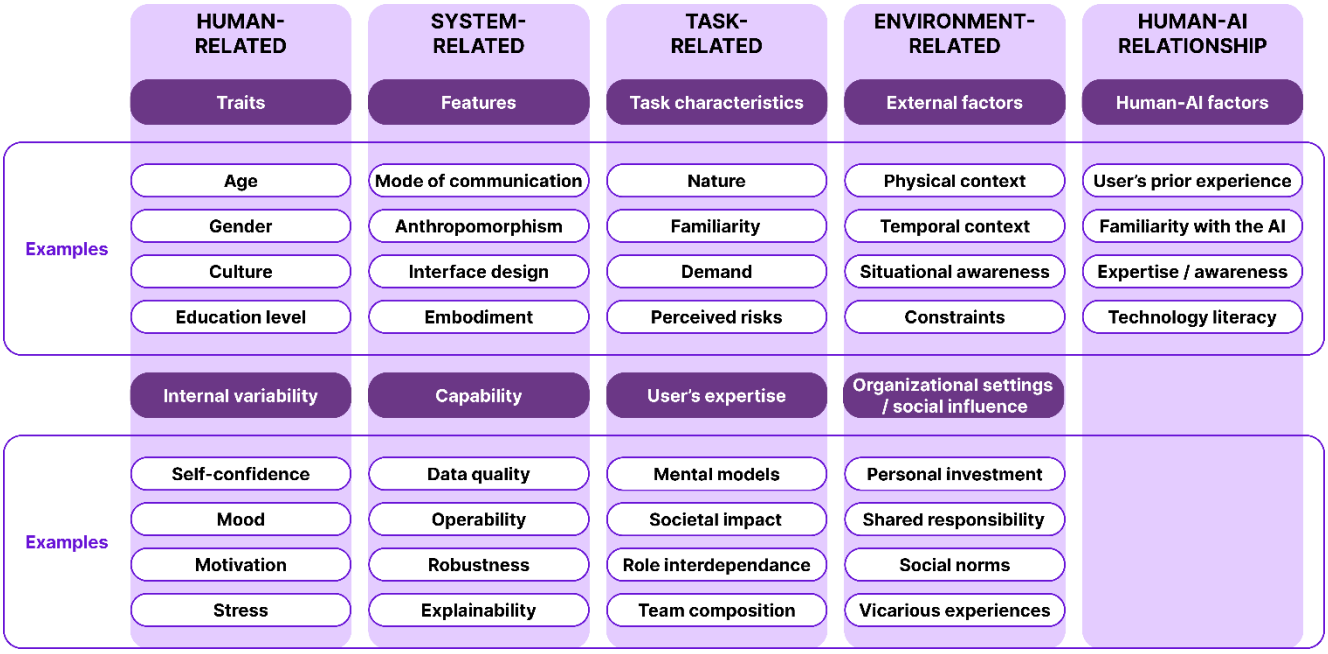


Figure 18. Enriched classification of mitigation variables.

4.5.1 Human-related factors

The AIA index, by design, will be seeking to consider intersectional factors (gender, ethnicity, age, socioeconomic status, disability). These factors are part of the **human-related factors** that directly affect the human abilities and attitudes towards AI. Indeed, multiple factors can have an influence on human-AI trust – some of them listed in a review by Hoff & Bashir (2015).

User traits are relatively stable attributes of individuals (*e.g.* **demographics**) that can influence their interaction with the AI and/or their answers in the AIA index questionnaires. This might include education level, gender, ethnicity, financial situation, countries, religions, handicap, age and generational cohorts as well as **cognitive abilities** (memory, attention, *etc.*). In the review by Hoff & Bashir (2015), the most significant factors reported were **culture** and **age** which affect **technology adoption and comfort level**. Although consistent gender differences have not always emerged, it is suggested that men and women may respond differently to automated systems based on communication style and presentation of information.

Additionally, personality traits, such as extraversion, emotional stability, and intuitive tendencies, can influence an individual's propensity to trust automated systems. Another example of user traits would be **attachment style**. A study by Gillath et *al.* (2011) suggests that attachment anxiety predicts less trust and that exposure to attachment security cues results in increased trust.

Language/vocabulary could also be an important factor to consider. For example, there are some constructs related to trust (*e.g.* trust, reliance, confidence) that have different definitions in English but translate the same in French ("confiance"). It will be critical to ensure that both the questions asked, and the results reported are meaningful and convey the right information. Other traits, such as overall anxiety, trust propensity as well as mastery confidence could be determinant predictors.

Internal variability refers to context-dependent characteristics of the operator, such as self-confidence, subject matter expertise, mood, and attentional capacity. For instance, greater self-confidence and specific expertise can reduce reliance on automation – as well as extrinsic and intrinsic **cognitive load** (Hoff & Bashir, 2015). The authors also identified that important factors were motivation, stress, sleep loss, boredom and other **attention-related variables**. As seen in the section 3.1.2.9, situation awareness is an important part of trust and many of these states are SAD[7] (Situation Awareness Demons): factors can cause a deterioration or loss of SA.

4.5.2 System-related factors

System-related factors include the features, functionalities, usability and accessibility of the overall AI-based system (algorithmics, interface, interaction design, *etc.*).

Complementary to the human-centred assessment planned in the AIA index, some significant progress is being made in the measurement of **system-centric data** related to trustworthiness, typically within the Confiance.ai program [see *e.g.* Awadid et *al.* (2024)]. Different ways of measurements were notably identified for:

- Data quality assessment (completeness, correctness, diversity, representativeness)
- Operability assessment (accuracy, precision, recall, F1_score, specificity, ROC Curve)
- Dependability assessment (availability, reliability, repeatability and reproducibility)
- Robustness assessment and monitoring (dataset corruption, metamorphism, time series, adversarial attacks, amplification, *etc.*)
- Explainability assessment (interpretability, monotonicity, sensitivity, explanation fidelity, usefulness and faithfulness of the interpretation)

However, these factors mostly related to trustworthiness and not to actual trust, acceptance or adoption – for which other system-related factors might have a significant influence, as described below.

In the domains of human-computer interactions and human-automation trust, many mitigating factors have been identified and could be determinant as well in human-AI trust or acceptance. This includes **system aesthetic**, anthropomorphism and **ease-of-use** (Hoff & Bashir, 2015) but also the "politeness" or "communication etiquette" of the system (Spain & Madhavan, 2009).

Performance, **false alarms** and **misses** negatively influence the notions focused by the AIA index – especially when bad performance occur early in the course of an interaction (Hoff & Bashir, 2015).

Feedback mechanisms and **system responsiveness** also seem to play a role in trust development. For example, Glikson & Woolley (2020) showed that immediacy behaviours led to better results for both cognitive-based and emotional-based trust. Additionally, transparency-related feedback (regarding reliability, explanation, errors,

[7] Situation Awareness Deamons are the situations where a user experiences a deteriorated understanding of the current functioning of the system, or its next steps.

predictability, *etc.*) could be an essential factor when measuring the notions focused by the AIA index (Hoff & Bashir, 2015).

The **physical appearance** and **interface design** of the AI-based system are also significant factors, notably the AI representation as robotic, virtual or embedded. **Embodiment** – and especially *tangibility* - could be an important factor in the development of AI-human trust (Glikson & Woolley, 2020). Typically, the **anthropomorphism** of the AI design and is thought to increase AI-human trust (De Visser et *al.*, 2012; Waytz et *al.*, 2014).

4.5.3 Task-related factors

In any human-system relationship, the interaction outcomes are highly influenced by the **task characteristics**, notably its nature, familiarity, demand, complexity, difficulty, perceived risks and benefits, as well as the users' level of control (Hoff & Bashir, 2015). For example, higher workloads can moderate the positive relationship between trust and reliance on automation. This probably applies to AI-based systems. Typically, the AI assistant can be trusted for a specific task and not for another (Starke et *al.*, 2022). The **users' expertise on the matter** (domain-specific knowledge) could also be an important mitigating variable.

4.5.4 Environment-related factors

Factors such as user experience, trust or acceptance are related to the **environment** and its **uncertainty**. Thus, they can be affected by external factors which are important in defining the potential risk/benefit balance of the task. This includes the **physical context** (comfort, physical well-being, familiarity, perceived risk, *etc.*), the **temporal context** (time pressure or constraints) as well as **situational awareness**.

Environment-related factors also include **organizational settings** and **social influences**. For example, the degree of **personal investment** or **shared responsibility** could be an important factor (Hoff & Bashir, 2015). **Social norms** for (or against) AI also influence one's attitude towards AI, as well as **vicarious experiences** in cases where the user witnesses another person's success or failure with the evaluated AI.
Finally, it has been shown that human-AI attitudes are influenced by the users' trust in the **AI stakeholders** such as owners, designers, developers, and overall, the innovating firm and its communication (Hengstler et *al.*, 2016). This can be explained by the fact that human-human trust is known to develop and co-evolve **across levels**: person, group, organization and intergroups (Currall & Inkpen, 2006).

4.5.5 Human-AI relationship factors

Human-AI relationship factors are good predictors for the indicators measured in the AIA index. This encompasses the **user's prior experience** with AI, their level of trust in the technology, their **familiarity** with AI functionalities, their AI expertise/awareness (domain-specific knowledge), their **technology literacy** (tech-savviness), and their overall attitude toward AI (ethics, evaluation, *etc.*). This list is not exhaustive and will be developed and studied in the Task 4.2 during Year 2 and Year 3 of the PEER project.

5. Conclusion and perspectives

5.1 Conclusion

To engage the building of the AIA index, we conducted a review on the concepts of acceptance and trust. Based on our research on these two notions, **we developed a framework** composed of 5 concepts (acceptability, acceptance, adoption, trust, and trustworthiness), 10 sub-concepts (human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity; non-discrimination and fairness; societal and environmental well-being; accountability; collaboration; situation awareness; usability and user experience), and 29 sub-sub-concepts (*e.g.* user autonomy, interpretability or error management). **All these concepts were defined in the state-of-the-art** (section 3.1).

An initial list of measurement tools has been gathered. 60 measurement methods, including 46 tools, 18 of which are surveys, were found in the literature to measure the different concepts, sub-concepts and sub-sub-concepts related to acceptance and trust (section 3.2). However, on one hand, for certain concepts (acceptance, traceability, auditability, fundamental rights, human oversight, user autonomy, privacy, data protection, accessibility, sustainability, environmental friendliness), we have not found any end-users evaluation tools. On the other hand, we did find evaluation tools for designers (*i.e.* checklists).

To ensure that only the most reliable and relevant group of surveys were implemented on the Peac²h platform, we decided to limit our sample to a reduced number (section 3.3). We examined 59 subscales from 18 surveys, to arrive at a selection of **12 tools implemented on the Peac²h platform**.

With the aim of **designing a user-centred index**, with one part aimed at designers and technology providers and another at end-users, the **use-case owners were consulted** to determine which notion(s) should be evaluated as a priority according to them (section 4). Reliability, user autonomy, accuracy / performance (empirical), understanding of the AI model, transparency, explainability, interpretability, collaboration, personalization, efficiency, effectiveness, user-friendliness, and satisfaction, appear to be very important. This vision needs be compared with the vision of the end-users.

5.2 Next steps

This selection of evaluation tools (task T4.1) was the preliminary phase in the building of the AIA index, which is planned in task T4.2. Over the next two years of the project, we will develop the complete methodology to define, design, prototype and test the AIA index. This means designing the tools and assessment processes associated to the different measurement scales and provide a common framework for the evaluation and assessment of AI systems.

The next steps, as the index further grows, will also be **to ensure the usability, efficiency, and adoption of the AIA index**. This will ensure that **the data collected through it is relevant and reliable**. We expect to see very positive results in terms of index quality and usability from the actions planned over the next few months. Indeed, these evaluations will enable us to adjust the index in order to improve it and make it as suitable as possible for the PEER project's use-cases.

To build an index that is **useful**, we need to ensure the **index covers the entire evaluation of an AI system in a comprehensive way**, *i.e.* with varied measurements in terms of **temporality** (both a posteriori and a priori

measures), **of data type** (with a quantitative/qualitative balance) or with **assessment methods** (expert evaluation, subjective data collection from the users, *etc.*).

The AIA index should aim for **a complete, composite measure** that correctly represents the broad notions involved (*e.g.*, transparency, acceptability). This will require **assembling standardized scores and metrics aggregated from a list of indicators** and **carefully selecting and weighting them.** These will be extracted either from the selected tools (Table 4) or by considering methods to add missing notions (*e.g.* self-built tools). Notably, the indicators used in the AIA index will need to be (as much as possible):

- **Sound**, based on robust definitions and frameworks.
- **Relevant** substantively to the targeted notion (*i.e.* should not be "proxies").
- **Universal**, measurable on different AIs and permit meaningful comparisons.
- **Verifiable** with accessible, available data.
- **Trackable**, allowing for future measures and permit monitoring over time.
- **Actionable**, providing insights that lead to specific actions or improvements.

For some constructs, we only have expert evaluations and need to find other ways of assessing the construct (*e.g.* traceability). The quality of the AIA index will also rely on it being **as reliable as possible** (*e.g.* with test-retest consistency) and as **valid as possible** (*i.e.* accurate, credible). This will be achieved by aiming at a reasonable content **validity** (*i.e.* covering the multiple notions in a comprehensive manner) and a reasonable **construct validity** (*i.e.* ensuring we measure what we intend to measure).

Additionally, to build an index that is **usable**, we need to use appropriate methods (and/or composition of methods) notably regarding **the ease of administratio**n, ease of providing instructions, ease of analysis, ease of result communication, *etc.* Part of the task will include providing a common platform to centralise all the end-user's evaluation while guaranteeing GDPR, ethical and privacy management.

To support real **use**, **adoption and meaning** of the AIA index, we will also need to look at identifying a **benchmark**, or some baseline requirements to be able to state whether a score is "high" or "low".

The **AIA index** will be **used during the evaluation process** (WP5) to ensure end-users trust and acceptance of the AI system. We will provide, manage and organise a common evaluation plan and assessment process in all the use-cases/pilots (WP5), as well as provide guidelines and policy recommendations for the usage, implementation and interpretation of the AIA index.

During the evaluation process, the **AIA index itself will be tested** in order to **improve it and provide guidelines and recommendation for using it.**

This document reports on the first steps towards our goal **of creating the most comprehensive and viable AI trust and acceptance index possible**. Through this document, we have taken stock of the major work that has been carried out, in order to lay the foundations for what we will be building over the next few years. We can now begin to evaluate our productions to further refine the reliability and the accuracy of our index. Our future productions will continue to take a user-centric approach, with the aim of making the index as relevant as possible to the PEER project and beyond. This index will be equally useful to a broad range of the society to study, evaluate, compare and promote AI systems: scholars, industries, politics, public services.

6. Bibliography

Aguilera, R. V., Filatotchev, I., Gospel, H., & Jackson, G. (2008). An organizational approach to comparative corporate governance : Costs, contingencies, and complementarities. *Organization science*, *19*(3), 475-492.

Akash, K., Hu, W.-L., Jain, N., & Reid, T. (2018). A classification model for sensing human trust in machines using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *8*(4), 1-20.

Amiel, V., Bigot, L. L., Terrier, P., Cellier, J.-M., & Babin, L.-M. (2004). *Deux indicateurs de la collaboration de l'utilisateur en dialogue homme-machine : Une étude expérimentale*. 17-19.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). *Explainable Artificial Intelligence (XAI) : Concepts, taxonomies, opportunities and challenges toward responsible AI*. http://arxiv.org/abs/1910.10045

Ashoori, M., & Weisz, J. D. (2019). In AI we trust ? Factors that influence trustworthiness of AI-infused decision-making processes. *arXiv preprint arXiv:1912.02675*.

Awadid, A., Amokrane-Ferka, K., Sohier, H., Mattioli, J., Adjed, F., Gonzalez, M., & Khalfaoui, S. (2024). AI Systems Trustworthiness Assessment : State of the Art. *Workshop on Model-based System Engineering and Artificial Intelligence-MBSE-AI Integration 2024*.

Baumgartner, J., Ruettgers, N., Hasler, A., Sonderegger, A., & Sauer, J. (2021). Questionnaire experience and the hybrid System Usability Scale : Using a novel concept to evaluate a new instrument. *International Journal of Human-Computer Studies*, *147*, 102575.

Bobillier-Chaumon, M.-É. (2003). Évolutions techniques et mutations du travail : Émergence de nouveaux modèles d'activité. *Le travail humain*, *66*(2), 161-192. https://doi.org/10.3917/th.662.0161

Bobillier-Chaumon, M.-E. (2016). L'acceptation située des technologies dans et par l'activité : Premiers étayages pour une clinique de l'usage. *Psychologie du Travail et des Organisations*, *22*(1), 4-21. https://doi.org/10.1016/j.pto.2016.01.001

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

Brangier, É., & Barcenilla, J. (2003). *Concevoir un produit facile à utiliser* (Editions d'organisation.).

Brangier, É., Hammes-Adelé, S., & Bastien, J.-M. C. (2010). Analyse critique des approches de l'acceptation des technologies : De l'utilisabilité à la symbiose humain-technologie-organisation. *European Review of Applied Psychology*, *60*(2), 129-146. https://doi.org/10.1016/j.erap.2009.11.002

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, *189*(194), 4-7.

Bruton, A., Conway, J. H., & Holgate, S. T. (2000). Reliability : What is it, and how is it measured? *Physiotherapy*, *86*(2), 94-99.

Cai, H., & Lin, Y. (2010). Tuning trust using cognitive cues for better human-machine collaboration. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(28), 2437-2441.

Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). *Situation Awareness-Based Agent Transparency:* Defense Technical Information Center. https://doi.org/10.21236/ADA600351

Currall, S. C., & Inkpen, A. C. (2006). On the complexity of organizational trust : A multi-level co-evolutionary perspective and guidelines for future research. *Handbook of trust research*, 235-246.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology : A comparison of two theoretical models. *Management Science*, *35*(8), 982-1003. https://doi.org/10.1287/mnsc.35.8.982

De Graaf, M. M., & Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should too). *2017 AAAI Fall Symposium Series*.

De Visser, E. J., Beatty, P. J., Estepp, J. R., Kohn, S., Abubshait, A., Fedota, J. R., & McDonald, C. G. (2018). Learning from the slips of others : Neural correlates of trust in automated agents. *Frontiers in human neuroscience*, *12*, 309.

De Visser, E. J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., & Parasuraman, R. (2012). The world is not enough : Trust in cognitive agents. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 263-267.

Distler, V., Lallemand, C., & Bellet, T. (2018). Acceptability and Acceptance of Autonomous Mobility on Demand : The Impact of an Immersive Experience. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1-10. https://doi.org/10.1145/3173574.3174186

Doherty, K., & Doherty, G. (2018). Engagement in HCI : Conception, theory and measurement. *ACM Computing Surveys*, *51*(5), 1-39. https://doi.org/10.1145/3234149

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International journal of human-computer studies*, *58*(6), 697-718.

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*, 789-795. https://doi.org/10.1109/NAECON.1988.195097

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32-64. https://doi.org/10.1518/001872095779049543

Endsley, M. R. (2017). From here to autonomy : Lessons learned from human–automation research. *Human factors*, *59*(1), 5-27.

Endsley, M. R. (2023). Supporting Human-AI Teams : Transparency, explainability, and situation awareness. *Computers in Human Behavior*, *140*, 107574.

Forlizzi, J., & Battarbee, K. (2004). Understanding experience in interactive systems. *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*, 261-268.

Fulmer, C. A., & Gelfand, M. J. (2012). At what level (and in whom) we trust : Trust across multiple organizational levels. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1873149

Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., & Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior*, *115*, 106607.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence : Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660. https://doi.org/10.5465/annals.2018.0057

Glomsrud, J. A., Ødegårdstuen, A., Clair, A. L. S., & Smogeli, Ø. (2019). Trustworthy versus explainable AI in autonomous vessels. *Proceedings of the International Seminar on Safety and Security of Autonomous Vessels (ISSAV) and European STAMP Workshop and Conference (ESWC)*, *37*.

Grote, G. (2023). Shaping the development and use of Artificial Intelligence : How human factors and ergonomics expertise can become more pertinent. *Ergonomics*, *66*(11), 1702-1710. https://doi.org/10.1080/00140139.2023.2278408

Grote, G., Ryser, C., Wäler, T., Windischer, A., & Weik, S. (2000). KOMPASS: A method for complementary function allocation in automated work systems. *International Journal of Human-Computer Studies*, *52*(2), 267-287.

Gulati, S., Sousa, S., & Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, *38*(10), 1004-1015.

Gursoy, D., Chi, O. H., Lu, L., & Nunkoo, R. (2019). Consumers acceptance of artificially intelligent (AI) device use in service delivery. *International Journal of Information Management*, *49*, 157-169. https://doi.org/10.1016/j.ijinfomgt.2019.03.008

Habib, L. (2019). *Niveaux d'automatisation adaptables pour une coopération homme-robots* [Université Polytechnique]. http://www.theses.fr/2019UPHF0021

Hassenzahl, M., & Tractinsky, N. (2006). User experience-a research agenda. *Behaviour & information technology*, *25*(2), 91-97.

Hellmann, M., Hernandez-Bocanegra, D. C., & Ziegler, J. (2022). Development of an instrument for measuring users' perception of transparency in recommender systems. *system*, *12*, 7.

Hendrikx, F., Bubendorfer, K., & Chard, R. (2015). Reputation systems : A survey and taxonomy. *Journal of Parallel and Distributed Computing*, *75*, 184-197. https://doi.org/10.1016/j.jpdc.2014.08.004

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, *105*, 105-120.

Hoff, K. A., & Bashir, M. (2015). Trust in automation : Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407-434. https://doi.org/10.1177/0018720814547570

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv Preprint arXiv:1812.04608*.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI : Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, *5*, 1096257. https://doi.org/10.3389/fcomp.2023.1096257

Holzinger, A., Carrington, A., & Müller, H. (2020). Measuring the quality of explanations : The system causability scale (SCS) comparing human and machine explanations. *KI-Künstliche Intelligenz*, *34*(2), 193-198.

Igbaria, M., & Tan, M. (1997). The consequences of information technology acceptance on subsequent individual performance. *Information & Management*, *32*(3), 113-121. https://doi.org/10.1016/S0378-7206(97)00006-2

ISO 9241-11 : Ergonomics of human-system interaction—Part 11 : Usability : Definitions and concepts. (2018).

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence : Prerequisites, causes and goals of human trust in AI. *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 624-635.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, *4*(1), 53-71.

Karahanna, E., Straub, D. W., & Chervany, N. L. (1999). Information technology adoption across time : A cross-sectional comparison of pre-adoption and post-adoption beliefs. *MIS Quarterly*, *23*(2), 183. https://doi.org/10.2307/249751

Karapanos, E., Zimmerman, J., Forlizzi, J., & Martens, J.-B. (2010). Measuring the dynamics of remembered experience over time. *Interacting with computers*, *22*(5), 328-335.

Kim, H.-Y., & McGill, A. L. (2018). Minions for the rich? Financial status changes how consumers see products with anthropomorphic features. *Journal of Consumer Research*, *45*(2), 429-450. https://doi.org/10.1093/jcr/ucy006

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of trust in automation : A narrative review and reference guide. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.604977

Körber, M., Gold, C., Goncalves, J., & Bengler, K. (2015). Vertrauen in Automation-Messung, Auswirkung und Einflüsse. *TÜV SÜD Akademie GmbH*, *7*.

Lallemand, C., Gronier, G., & Koenig, V. (2015). User experience : A concept without consensus ? Exploring practitioners' perspectives through an international survey. *Computers in human behavior*, *43*, 35-48.

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and Evaluation of a User Experience Questionnaire. In A. Holzinger (Éd.), *HCI and Usability for Education and Work* (p. 63-76). Springer. https://doi.org/10.1007/978-3-540-89350-9_6

Lee, J. D., & See, K. A. (2004). Trust in automation : Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243-1270.

Leveson, N. G. (2014). *Extending the human controller methodology in systems-Theoretic Process Analysis (STPA)*. Massachusetts Institute of Technology.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy AI : From principles to practices. *ACM Computing Surveys*, *55*(9), 177:1-177:46. https://doi.org/10.1145/3555803

Liao, C., Chen, J.-L., & Yen, D. C. (2007). Theory of planning behavior (TPB) and customer satisfaction in the continued use of e-service : An integrated model. *Computers in Human Behavior*, *23*(6), 2804-2822. https://doi.org/10.1016/j.chb.2006.05.006

Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. *Proceedings of the SIGCHI conference on human factors in computing systems*, 2119-2128.

Louvet, J.-B. (2019). *Collaboration humain-machine à l'aide de motifs dialogiques pour la réalisation d'une tâche complexe : Application à la recherche d'information*. Normandie Université.

Lu, L., Cai, R., & Gursoy, D. (2019). Developing and validating a service robot integration willingness scale. *International Journal of Hospitality Management*, *80*, 36-51. https://doi.org/10.1016/j.ijhm.2019.01.005

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust : An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277-301.

Martin, N., Erhel, S., Jamet, É., & Rouxel, G. (2015). Quels liens entre expérience utilisateur et accdeptabilité? *Proceedings of the 27th Conference on l'Interaction Homme-Machine*, 1-6. https://doi.org/10.1145/2820619.2825015

Mayer, R. C., & Davis, J. H. (1999). The Effect of the Performance Appraisal System on Trust for Management : A Field Quasi-Experiment. *Journal of Applied Psychology*, *84*(1), 123.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709. https://doi.org/10.2307/258792

McKinney, V., Yoon, K., & Zahedi, F. "Mariam". (2002). The measurement of web-customer satisfaction : An expectation and disconfirmation approach. *Information Systems Research*, *13*(3), 296-315. https://doi.org/10.1287/isre.13.3.296.76

McLarney, E., Gawdiak, Y., Oza, N., Mattmann, C., Garcia, M., Maskey, M., Tashakkor, S., Meza, D., Sprague, J., & Hestnes, P. (2021). *NASA framework for the ethical use of artificial intelligence (AI)*.

Miller, T. (2019). Explanation in artificial intelligence : Insights from the social sciences. *Artificial Intelligence*, *267*, 1-38.

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary Ssurvey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems*, *11*(3-4), 1-45. https://doi.org/10.1145/3387166

Morana, S., Gnewuch, U., Jung, D., & Granig, C. (2020). The Effect of Anthropomorphism on Investment Decision-Making with Robo-Advisor Chatbots. *ECIS*.

Moray, N. (1979). Models and measures of mental workload. In N. Moray (Éd.), *Mental Workload : Its Theory and Measurement* (Moray, N., Vol. 8, p. 13-21). Springer US. https://doi.org/10.1007/978-1-4757-0884-4_2

Nadal, C., Doherty, G., & Sas, C. (2019, mai 4). Technology acceptability, acceptance and adoption—Definitions and measurement. *2019 CHI Conference on Human Factors in Computing Systems*. 2019 CHI Conference on Human Factors in Computing Systems, GBR. https://eprints.lancs.ac.uk/id/eprint/131906/

Neerincx, M. A., Lindenberg, J., Smets, N., Grant, T., Bos, A., Olmedo-Soler, A., Brauer, U., & Wolff, M. (2006). Cognitive engineering for long duration missions : Human-machine collaboration on the Moon and Mars. *2nd IEEE International Conference on Space Mission Challenges for Information Technology (SMC-IT'06)*, 7 pp.

Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of marketing research*, *17*(4), 460-469.

Oliver, R. L. (1981). Measurement and evaluation of satisfaction processes in retail settings. *Journal of Retailing*, *57*(3), 25-48.

Oliver, R. L. (1993). Cognitive, affective, and attribute bases of the satisfaction response. *Journal of Consumer Research*, *20*(3), 418-430. https://doi.org/10.1086/209358

Olsen, P., & Borit, M. (2013). How to define traceability. *Trends in Food Science & Technology*, *29*(2), 142-150. https://doi.org/10.1016/j.tifs.2012.10.003

Patterson, P. G., Johnson, L. W., & Spreng, R. A. (1997). Modeling the determinants of customer satisfaction for business-to-business professional services. *Journal of the Academy of Marketing Science*, *25*(1), 4-17. https://doi.org/10.1007/BF02894505

Pearl, J., & Mackenzie, D. (2018). AI can't reason why. *Wall Street J.*

Pirson, M., & Malhotra, D. (2011). Foundations of organizational trust : What matters to different stakeholders? *Organization Science*, *22*(4), 1087-1104.

P. Stefanija, Ana, V. Belle, Jonne, Laenens, Willemien, and Heyman, Rob (2024). Social and Technical Requirements (Deliverable 2.1), EU-Horizon PEER project.

Quiguer, S. (2013). *Acceptabilité, acceptation et appropriation des systèmes de transport intelligents : Elaboration d'un canevas de co-conception multidimensionnelle orientée par l'activité* [Psychologie]. Université Rennes 2.

Rai, A., Lang, S. S., & Welker, R. B. (2022). Assessing the validity of IS success models : An empirical test and theoretical analysis. *Information Systems Research*, *13*(1), 50-69.

Rajaonah, B., Castelli, J. C., Ravenel, J.-B., Osmont, A., Cabrol, P., & Fur, G. L. (2014). *Acceptability of security scanners at airports : A French opinion survey*.

Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation methods in deep learning : Users, values, concerns and challenges. *Explainable and interpretable models in computer vision and machine learning*, 19-36.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of personality and social psychology*, *49*(1), 95.

Renaud, K., & Van Biljon, J. (2008). *Predicting technology acceptance and adoption by the elderly : A qualitative study*. 210-219.

Rogers, E. M. (1995). Diffusion of Innovations : Modifications of a model for telecommunications. *Die diffusion von innovationen in der telekommunikation*, 25-38.

Roto, V., Law, E.-C., Vermeeren, A. P., & Hoonhout, J. (2011). *User experience white paper : Bringing clarity to the concept of user experience*.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Introduction to special topic forum : Not so different after all : A cross-discipline view of trust. *The Academy of Management Review*, *23*(3), 393-404.

Sanderson, C., Douglas, D., & Lu, Q. (2023). Implementing Responsible AI : Tensions and Trade-Offs Between Ethics Aspects. *2023 International Joint Conference on Neural Networks (IJCNN)*, 1-7. https://doi.org/10.1109/IJCNN54540.2023.10191274

Schaefer, K. E. (2016). Measuring trust in human robot interactions : Development of the "trust perception scale-HRI". In R. Mittu, D. Sofge, A. Wagner, & W. F. Lawless (Éds.), *Robust Intelligence and Trust in Autonomous Systems* (p. 191-218). Springer US. https://doi.org/10.1007/978-1-4899-7668-0_10

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation : Implications for understanding autonomy in future systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 377-400. https://doi.org/10.1177/0018720816634228

Schaefer, K. E., Hill, S. G., & Jentsch, F. G. (2019). Trust in human-autonomy teaming : A review of trust research from the US army research laboratory robotics collaborative technology alliance. In J. Chen (Éd.), *Advances in Human Factors in Robots and Unmanned Systems* (Vol. 784, p. 102-114). Springer International Publishing. https://doi.org/10.1007/978-3-319-94346-6_10

Schoenherr, J. R., Abbas, R., Michael, K., Rivas, P., & Anderson, T. D. (2023). Designing AI using a human-centered approach : Explainability and accuracy toward trustworthiness. *IEEE Transactions on Technology and Society*, *4*(1), 9-23.

Seddon, P. B. (1997). A respecification and extension of the DeLone and McLean model of IS success. *Information Systems Research*, *8*(3), 240-253. https://doi.org/10.1287/isre.8.3.240

Shah, C. (2010). *A framework for supporting user-centric collaborative information seeking*. The University of North Carolina at Chapel Hill.

Spain, R. D., & Madhavan, P. (2009). The role of automation etiquette and pedigree in trust and dependence. *Proceedings of the human factors and ergonomics society annual meeting*, *53*(4), 339-343.

Sperandio, J.-C. (1978). The regulation of working methods as a function of work-load among air traffic controllers. *Ergonomics*, *21*(3), 195-202. https://doi.org/10.1080/00140137808931713

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2022). Fairness perceptions of algorithmic decision-making : A systematic review of the empirical literature. *Big Data & Society*, *9*(2), 20539517221115189.

Stetson, J. N., & Tullis, T. S. (2004). A comparison of questionnaires for assessing website usability. *UPA Presentation*.

Stratmann, T. C., & Boll, S. (2016). Demon hunt—The role of Endsley's demons of situation awareness in maritime accidents. In C. Bogdan, J. Gulliksen, S. Sauer, P. Forbrig, M. Winckler, C. Johnson, P. Palanque, R. Bernhaupt, & F. Kis (Éds.), *Human-Centered and Error-Resilient Systems Development* (Vol. 9856, p. 203-212). Springer International Publishing. https://doi.org/10.1007/978-3-319-44902-9_13

Streitz, N. (2019). Beyond 'smart-only'cities : Redefining the 'smart-everything'paradigm. *Journal of Ambient Intelligence and Humanized Computing*, *10*(2), 791-812.

Taylor-Powell, E. (1998). *Evaluating collaboratives : Reaching the potential* (Numéro 8). University of Wisconsin--Extension.

Terrade, F., Pasquier, H., Reerinck-Boulanger, J., Guingouain, G., & Somat, A. (2009). L'acceptabilité sociale : La prise en compte des déterminants sociaux dans l'analyse de l'acceptabilité des systèmes technologiques. *Le travail humain*, *72*(4), 383-395. https://doi.org/10.3917/th.724.0383

The Confiance.ai program. (2022). *Towards the engineering of trustworthy AI applications for critical systems*.

The High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI*.

The High-Level Expert Group on AI. (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*.

Thüring, M., & Mahlke, S. (2007). Usability, aesthetics and emotions in human–technology interaction. *International journal of psychology*, *42*(4), 253-264.

Tocchetti, A., Corti, L., Balayn, A., Yurrita, M., Lippmann, P., Brambilla, M., & Yang, J. (2022). *A.I. Robustness : A Human-Centered Perspective on Technological Challenges and Opportunities*. https://doi.org/10.48550/ARXIV.2210.08906

Van Der Stigchel, B., Van den Bosch, K., Van Diggelen, J., & Haselager, P. (2023). Intelligent decision support in medical triage : Are people robust to biased advice? *Journal of Public Health*, *45*(3), 689-696.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology : Toward a unified view. *MIS Quarterly*, *27*(3), 425-478. https://doi.org/10.2307/30036540

Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology : Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, *36*(1), 157. https://doi.org/10.2307/41410412

Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1-39. https://doi.org/10.1145/3476068

Wagner, E. D. (1994). In support of a functional definition of interaction. *American Journal of Distance Education*, *8*(2), 6-29.

Wang, B., Rau, P.-L. P., & Yuan, T. (2023). Measuring user competence in using artificial intelligence : Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*, *42*(9), 1324-1337. https://doi.org/10.1080/0144929X.2022.2072768

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine : Anthropomorphism increases trust in an autonomous vehicle. *Journal of experimental social psychology*, *52*, 113-117.

Wei, W., & Liu, L. (2024). Trustworthy Distributed AI Systems : Robustness, Privacy, and Governance. *ACM Computing Surveys*, 3645102. https://doi.org/10.1145/3645102

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2021). "Let me explain!" : Exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces*, *15*(2), 87-98.

Wing, J. M. (2021). Trustworthy AI. *Communications of the ACM*, *64*(10), 64-71. https://doi.org/10.1145/3448248

Wirtz, J., Patterson, P. G., Kunz, W. H., Gruber, T., Lu, V. N., Paluch, S., & Martins, A. (2018). Brave new world : Service robots in the frontline. *Journal of Service Management*, *29*(5), 907-931. https://doi.org/10.1108/JOSM-04-2018-0119

Wojton, H. M., Porter, D., T. Lane, S., Bieber, C., & Madhavan, P. (2020). Initial validation of the trust of automated systems test (TOAST). *The Journal of social psychology*, *160*(6), 735-750.

Wood, D. J., & Gray, B. (1991). Toward a comprehensive theory of collaboration. *The Journal of Applied Behavioral Science*, *27*(2), 139-162. https://doi.org/10.1177/0021886391272001

Yuksel, B. F., Collisson, P., & Czerwinski, M. (2017). Brains or beauty : How to engender trust in user-agent interactions. *ACM Transactions on Internet Technology (TOIT)*, *17*(1), 1-20.

PEER will focus on how to systematically put the user at the centre of the entire AI design, development, deployment, and evaluation pipeline, allowing for truly mixed human-AI initiatives on complex sequential decision-making problems. The central idea is to enable a two-way communication flow with enhanced feedback loops between users and AI, leading to improved human-AI collaboration, mutual learning and reasoning, and thus increased user trust and acceptance. As an interdisciplinary project between social sciences and artificial intelligence, PEER will facilitate novel ways of engagement by end-users with AI in the design phase; will create novel AI planning methods for sequential settings which support bidirectional conversation and collaboration between users and AI; will develop an AI acceptance index for the evaluation of AI systems from a human-centric perspective; and will conduct an integration and evaluation of these novel approaches in several real-world use cases.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101120406.