Artificial intelligence, Data and Robotics ecosystem

https://adra-e.eu/

Call: A human-centred and ethical development of digital and industrial technologies 2021
Topic: Horizon-CL4-2021-Human-01
Type of action: Coordination and Support actions
Grant agreement Nº: 101070336

WP Nº3:	ADR Awareness Centre
Deliverable Nº3.2:	Report on Selection of AI Trust Label
Lead partner:	University of Galway
Version Nº:	1
Date:	30/11/2023
Dissemination level[1]:	CO

[1] **PU:** Public**; CO:** Confidential, only for members of the consortium (including the Commission Services)

Document information	
Deliverable Nº and title:	D3.2 – Report on Selection of AI Trust Label
Version Nº:	1.0
Lead beneficiary:	University of Galway (UoG)
Author(s):	Fatemeh Ahmadi Zeleti (UoG)
Reviewers:	Emmanuel Vincent (INRIA), Edward Curry (UoG)
Submission date:	21/12/2023
Due date:	31/12/2023
Type[2]:	R
Dissemination level[3]:	CO

Document history			
Date	Version	Author(s)	Comments
30/11/2023	0.1	Fatemeh Ahmadi Zeleti	Draft deliverable finalized
12/12/2023	0.2	Reviewers	Review comments received and incorporated
13/12/2023	1.0	Fatemeh Ahmadi Zeleti	Format checked and V1 finalized

Disclaimer :

This document contains description of the Adra-e work and findings.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any responsibility for actions that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the Adra-e consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 28 Member States of the Union. It is based on the European Communities and the Member States cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors (http://europa.eu/).

Adra-e has received funding from the European Union's Horizon Europe under grant agreement 101070336.

[2] **R**: Report, **DEC:** Websites, patent filling, videos; **DEM:** Demonstrator, pilot, prototype; **OTHER:** Software Tools

[3] **PU:** Public**; CO:** Confidential, only for members of the consortium (including the Commission Services)

Document summary

This deliverable provides a detailed report on the activities and in particular the results from the work done to date on the selection of an AI Trust Label (T3.3 of WP3). For easy readability, this deliverable documents the progress and completed activities in two parts. Part A of this deliverable provides a summary of the work done. Part B provides a detailed analysis of the work done.

The aim of this task is to select and promote an informed-based AI trust label to support the consumers understanding the trustworthiness of the AI product and service. The focus here is on trustfulness and will be based on identifying, analysing and assessing labelling initiatives. In addition, T3.3 aims to identify indicators of trust for consumers through analysis of the assessed labelling initiatives, stakeholder engagement, and the input received from T3.2 regarding the trust concerns and adoption issues of European organisations.

This task will help improve public trust in and acceptance of AI products and services, support the awareness of and operationalization of the AI Act, support standardisation efforts towards a trust label for the EU, and improve international trade relations.

Finally, all relevant content or recommendations generated in this task will be pushed to a relevant standardisation body in Europe.

Table of Contents

List of Tables

List of Figures

PART A

1. Introduction

This deliverable provides a detailed report on the activities and in particular the results and offerings of work done to date on the selection of an AI Trust Label (T3.3 of WP3). For easy readability, this deliverable documents the progress and completed activities in two parts. Part A of this deliverable provides a summary of the work done. Part B provides a detailed analysis of the work done in a number of appendices.

The aim of this task is to select and promote an informed-based AI trust label that supports the consumers understanding the trustworthiness of AI products and services. The focus here is on trustfulness and will be based on identifying, analys ing and assessing labelling initiatives for the development and use of trustworthy AI systems and applications. In addition to the selection of an AI trust label, T3.3 aims to identify indicators of trust for consumers through analysis of the assessed labelling initiatives, stakeholder engagement, and the input received from T3.2 regarding the trust concerns and adoption issues of European organisations.

This task will help improve public trust in AI products and services. In addition, results will support the awareness of and operationalization of the AI Act and improve trust and acceptance of AI. Finally, all relevant content or recommendations generated in this task will be pushed to a relevant standardisation body in Europe.

Note: December 09, 2023, the Council presidency and the European Parliament’s negotiators have reached a provisional agreement on the proposal of the AI Act.

1.1 Overview

Artificial Intelligence (AI) has become increasingly central to innovations in many sectors in the last decade. It has also developed and achieved performance at unprecedented levels. Different communities raised concerns regarding AI creating and using decisions that are not justifiable, legitimate, or simply do not explain their behaviour clearly. Such issues are part of the extensive discussions about AI ethics across various communities developing or using AI, standardisation bodies and regulatory institutions, and civil society. In addition, citizens and other stakeholders also worry that AI can have unintended effects or even be used for malicious purposes. Companies are similarly concerned about legal uncertainties [3]. These concerns are potential barriers to the broader uptake of AI [1] and create the demand for transparency from the various stakeholders in AI, specifically those whose lives are affected by AI decisions.

To address these challenges, the European Commission (EC) unveiled the draft EU AI Act that sets out a regulatory framework for the trustworthy use of AI - *Trustworthiness is considered as the confidence of whether an AI system will act as intended when facing a given situation*. The Act primarily aims to regulate ’high-risk’ AI systems through minimum mandatory requirements. The Act seeks to:

- Facilitate the ‘development and uptake of AI’ with an ‘appropriate ethical and legal framework’ and
- Promote an ‘ecosystem of trust’ in Europe.

On April 27, 2023, the members of the European Parliament reached a provisional political agreement on the text of the AI Act (with more than 3,000 proposed amendments) to tackle generative AI head-on, subject to technical-level adjustments and a further committee vote in May. Although the text might still be subject to minor additions or technical amendments, the Act will ultimately regulate the use of AI across the EU to establish legal clarity on which uses of AI are prohibited, which uses are permitted subject to mandatory safeguards, and which uses are permitted with few or no restrictions. The AI Act will be finalised by the end of 2023.

At the same time, regulation should not hinder but support product and service innovation and the business environment. Both of these objectives are best achieved by increasing legal certainty and clarity throughout the regulation proposal to support the private sector and public administrations (primarily through the recently developed European Digital Innovation Hub AI-PACT to design, test and implement AI-based solutions and curtail the booming AI risks) to comply with the new obligations.

To support these objectives, fourteen EU Member States, including the Netherlands, Ireland, and France, have strongly advocated for the enhancement of self-certification that would support an innovation-friendly market for AI by "incentivising AI developers to proactively and systematically promote trustworthy AI" for the benefit of European citizens and economy. Self-certification tools could make visible which applications are based on secure, responsible, and ethical AI and data and, therefore, which applications to trust, thus empowering those affected to make an ethical choice.

1.2 Target group and Structure of the Deliverable

The main target audiences to this document include: the Project Officer, the Reviewer team formed by the EC, Adra-e consortium members, and ADRA members.

For easier readability, we structure this document in two main parts: PART A (sections) and PART B (Annexes).

In PART A (40 pages, main body of the deliverable), we briefly provide methodology, details of the completed activities and the future directions for T3.3. Every activity designed and completed in this part is aligned with the goals and objectives of T3.3 and follows a scientific methodology. Part A will not provide step by step and detailed analysis of the completed tasks. *Part A is mandatory to read as it provides a complete overview of how the task is designed and completed.*

In PART B, the step by step and detailed analysis of all the completed activities in T3.3 are annexed. By doing this, we aim to certify how and in what ways we have completed the activities presented in Part A. *Part B is not mandatory to read as it provides details of the analysis, however, reading this part is advised if the reader requires more in-depth information on the analysis.*

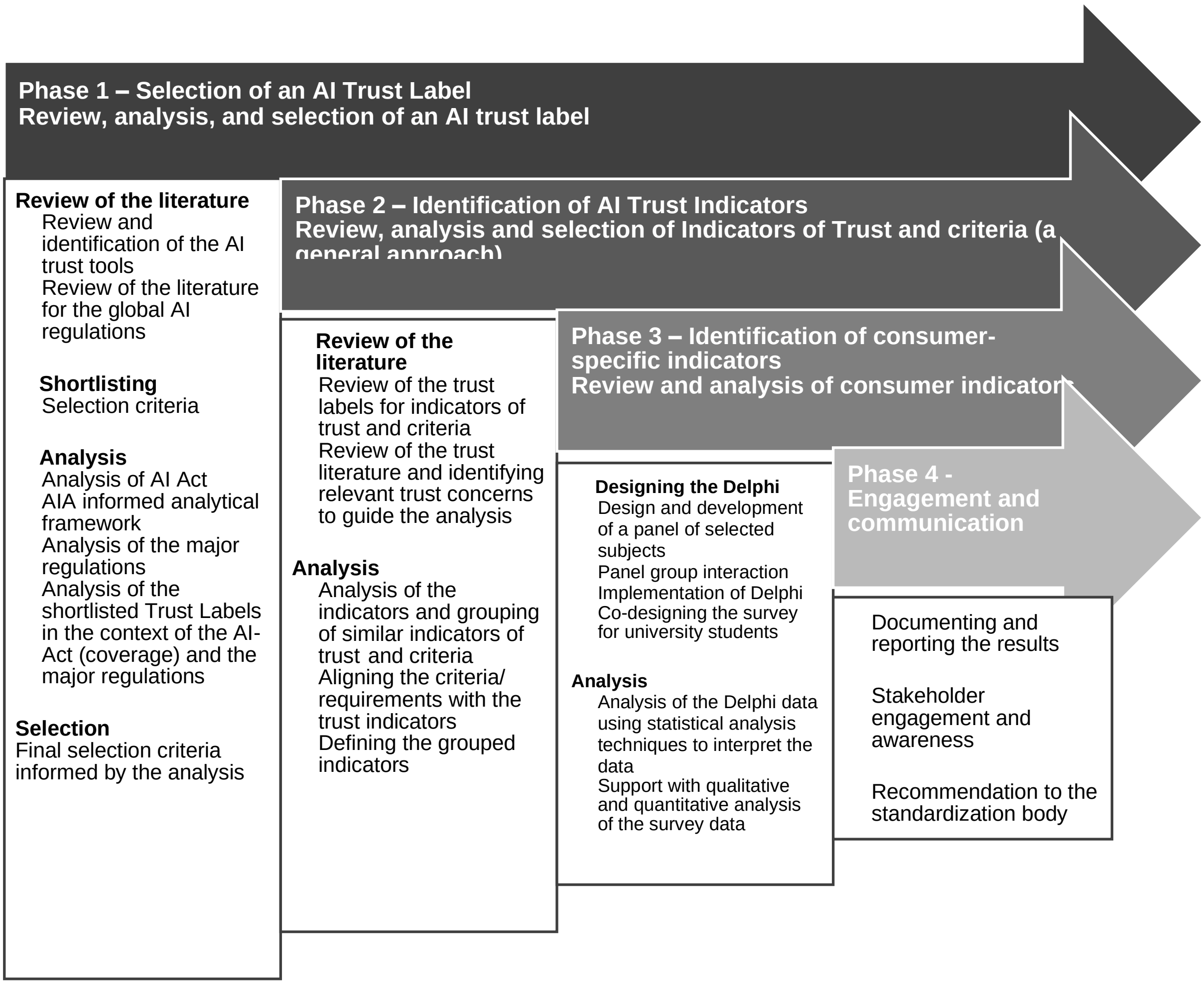
1.3 Methodological approach

This section provides a summary of the methodology employed to carry out and deliver T3.3 AI Trust Label with two main goals: 1) selection of an AI Trust Label and 2) selection of AI Trust Indicators. The methodological approach employed to select a trust label and a set of indicators for consumers of AI products is briefly outlined below.

To achieve the objectives, we followed a mixed method approach with a set of activities and methods that are structured in the following four phases and presented in Figure 1:

1) Selection of AI Trust Label
2) Identification of AI Trust Indicators
3) Identification of stakeholder-specific (consumer to be specific) trust indicators, and
4) Stakeholder engagement and communication

All the different methods and techniques used to complete this task lie within the design space presented in Figure 3.



Phase 1 – Selection of an AI Trust Label
Review, analysis, and selection of an AI trust label

Review of the literature
Review and identification of the AI trust tools
Review of the literature for the global AI regulations

Shortlisting
Selection criteria

Analysis
Analysis of AI Act
AIA informed analytical framework
Analysis of the major regulations
Analysis of the shortlisted Trust Labels in the context of the AI-Act (coverage) and the major regulations

Selection
Final selection criteria informed by the analysis

Phase 2 – Identification of AI Trust Indicators
Review, analysis and selection of Indicators of Trust and criteria (a general approach)

Review of the literature
Review of the trust labels for indicators of trust and criteria
Review of the trust literature and identifying relevant trust concerns to guide the analysis

Analysis
Analysis of the indicators and grouping of similar indicators of trust and criteria
Aligning the criteria/ requirements with the trust indicators
Defining the grouped indicators

Phase 3 – Identification of consumer-specific indicators
Review and analysis of consumer indicators

Designing the Delphi
Design and development of a panel of selected subjects
Panel group interaction
Implementation of Delphi
Co-designing the survey for university students

Analysis
Analysis of the Delphi data using statistical analysis techniques to interpret the data
Support with qualitative and quantitative analysis of the survey data

Phase 4 - Engagement and communication

Documenting and reporting the results

Stakeholder engagement and awareness

Recommendation to the standardization body

Figure 1. Methodological phases and activities

Phase 1 - Selection of the AI Trust Label (Completed)

- Review of the literature and existing reports **(Completed)**
  - Review of the relevant frameworks and trust literature 2. Review of the related literature
  - Review of a collection of initiatives (Label, Certification, Quality, Trust, and Kite marks, Rating Framework, Code of Conduct, Code of Ethics, Seal and Audit) 3.1.1 Review of the AI trust initiatives (self-regulatory initiatives)
  - Review of other major AI regulations 3.1.2 Review of the four major AI regulations

- Shortlisting the reviewed initiatives based on the following criteria **(Completed)** 3.1.3 Short-listing the reviewed initiatives
  - Initiatives of types 'label' and 'certification'
  - Initiatives that are cross-sectoral and horizontal
  - Initiatives that are developed and in use

- Analysis **(Completed)**
  - Analysis of AI Act requirements and development of a simple structure (tested the structure with AI Trust Label and the World Economic Forum AI

Label) A.4 Analysis of the AI Act mandatory requirements and development of an AI Act-informed analytical framework
        - Development of an AIA-informed Analytical Framework for analysis of the shortlisted initiatives for their coverage A.4 Analysis of the AI Act mandatory requirements and development of an AI Act-informed analytical framework
        - Inductive and directed approach to content analysis (initial taxonomy and coding categories are based on the analytical frame)
        - Deductive approach and conventional content analysis technique of open coding to form new taxonomy and categories (little flexibility in the restructuring of initial taxonomy due to the tight alignment with AI Act requirements)
        - Content analysis of major regulations and coverage analysis of the labels 3.1.4 Analysis and Selection of the Trust Label

    - Selecting a Trust Label **(Completed)** 3.1.4 Analysis and Selection of the Trust Label
        - Compatible and high coverage with the AI Act
        - Good coverage with other major regulations
        - Clear methodology
        - Value-based or value compliance, meaning that it gives flexibility to set target requirements for a value and it describes compliance with the specified values (for example, one product might better comply with privacy requirements, while the other might comply better with transparency criteria)
        - Applies to self-certification and third-party conformity
        - Industry-Academic Engagement
        - Active contributor community from multiple member states
        - Going EU-wide *(to be included in the label's next public release before the end of 2023)*
        - Involvement of 4 major industries for vertical requirements *(to be included in the label's next public release before the end of 2023)*
            - Finance
            - Defense
            - Mobility
            - Health

Phase 2 - Identification of Trust Indicators

- Review of the literature and related documents **(Partially completed)**
    - Review of the shortlisted initiatives and extraction of a list of trust indicators and criteria from the initiatives 3.2.1 Extracting indicators and criteria from the initiatives
    - Review of the Trust literature and identification of relevant trust concerns and indicators to guide the analysis 2. Review of the related literature
    - Review of the consumer protection requirements from other major regulations (AIA, Bill C-27, Bill of Right, UK AI Regulation and AI Regulation of Japan) to guide the analysis **(M19-M23)**

- Analysis of Trust Indicators **(Partially completed)**
    - Content analysis of the initiatives guided by the categories of the trust indicators identified from the initiatives (definition of the indicators, grouping of the similar indicators) B.1 Analysis stages (synthesis) of trust indicators extracted from the initiatives
    - Content analysis of the consumer protection requirements from other major regulations to guide indicators selection **(M19-M23)**

Phase 3 – Selection of consumer trust indicators (Started) (M18-M36)

- Developing the stakeholder group **(M18-M20)**
  - Identify the minimally sufficient number of subjects (informed by the resources available to UoG for this task) 1.3 Methodological approach
  - Identification of potential stakeholders from academia, company, public (aiming for different age groups), legal, students, educators (Table 1)
  - Geographic dispersion of the subjects
  - One to one discussion
  - Formal and written invitation
  - Group meetings scheduled if subject anonymity does not apply

The advantages of the stakeholder group approach are threefold:

- ✓ Collectively mediate AI's social and ethical controversies;
- ✓ Improve the quality of reflection on the topic;
- ✓ Strengthen the legitimacy of the proposals.

Table 1. Potential members of the stakeholder group

Stakeholder types	Country	Institution type and name if possible	Expected contribution	Membership status
Assistant professor	Italy	University of Bologna	Research-based, Human-centred AI, and Assessment and engineering of equitable, unbiased, impartial and trustworthy AI systems	Confirmed
AI project manager	Sweden	The Västernorrland Municipal Association	Public interest and trust concerns, municipalities needs	Pending
Senior legal researcher	Belgium	IMEC (CiTiP - KU Leuven)	Research-based legal expertise	Pending
Canada Research Chair in Governance and AI	Canada	Academia (Carleton University)	Governance aspects of AI solutions	Confirmed
College students	Ireland	Academia (University of Galway)	Student's perception of trustworthy AI and the value trust labels can offer	Confirmed
CEO	UK/ Germany	Association and technical-scientific organization (VDE group)	AI Trust Label standardization methodology and process	Pending
PhD students	Switzerland	Academia (University of Basel)	Research-based Psychology and Methodology	Confirmed
Research and Policy Analyst	Canada	Responsible AI Institution and	Responsible AI Certification program (methodology), AI	Pending

		Arizona State University	Governance, Law, and Policy	
University lecturers (junior level)	Ireland	Academia (University of Galway)	AI use in the classroom and the students perception	Pending
University professors	Ireland	Academia (University of Galway)	Philosophy and Ethics of social AI	Confirmed
EU project manager	The Netherlands	AI Lab (ICAI Netherlands)	Human-centred research in AI, AI-based methods and tools designed to create social impact and promote sustainable growth	Confirmed
Lawyer (in practice)	Greece	Law (Attorney At Law)	Data privacy l aw, Data privacy concerns of stakeholders	Confirmed
EU Project manager	Ireland	ADAPT Centre	AI Act knowledge, AI characterization	Confirmed
AI system expert	Ireland	Company (Galvia)	Company specific value sets, regulatory needs or criteria	Confirmed

- Designing the Delphi and the Survey **(M19-M25)**
    - Identifying the scope and the number of iterations
    - Subject anonymity decision
    - Designing the questionnaire based on the findings from the previous step (list of indicators and criteria and the rating/ranking)
    - Co-designing the survey for university students C.2 Student Survey

  This is presented in Figure 2.

- Implementation of Delphi
    - Iteration scheduling
    - Members interaction (one to one if anonymity applies)
    - Consensus checking

- Analysis and selection of the consumer trust indicators **(M21-M27)**
    - Stakeholder analysis of the initiatives C.1 Stakeholder analysis of the initiatives based on the AI Act
    - Analysis of the Delphi data per iteration specified above (disagreement and agreement)
    - Support from qualitative and quantitative analysis of the survey data from the student project. The aim is to use the survey data to support the data we collect during Delphi implementation.

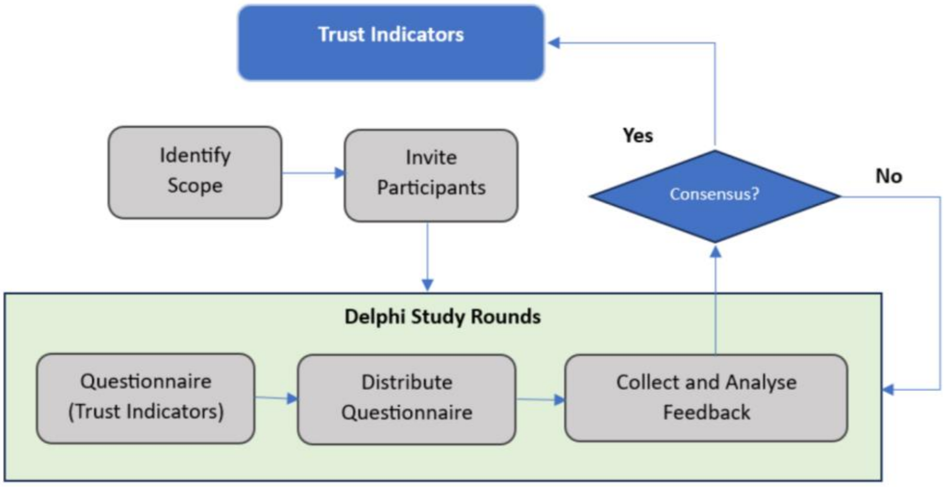


Figure 2. Delphi

Phase 4 - Stakeholder engagement, awareness, and communication of the results

- Stakeholder group formed and communicated in Phase 3
- Documenting the results from the phases
- Awareness Day and other possible channels
- ADR Awareness Center (resources on Trust Labels)
- Recommendation to the standardisation body

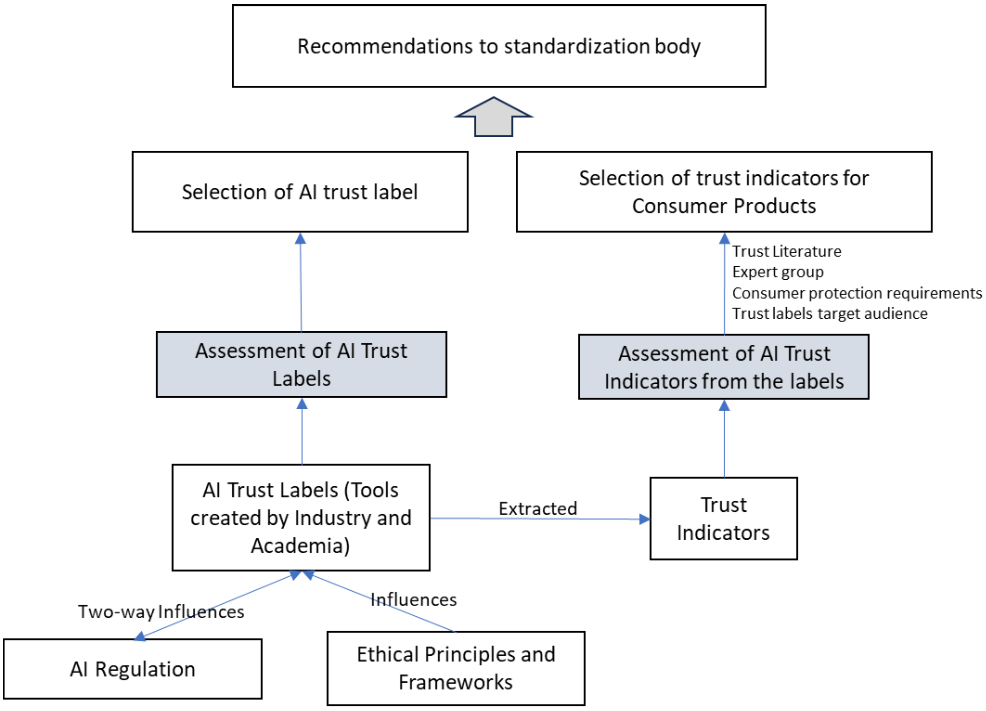

Recommendations to standardization body
Selection of AI trust label
Selection of trust indicators for Consumer Products
Trust Literature
Expert group
Consumer protection requirements
Trust labels target audience
Assessment of AI Trust Labels
Assessment of AI Trust Indicators from the labels
AI Trust Labels (Tools created by Industry and Academia)
Extracted
Trust Indicators
Two-way Influences
Influences
AI Regulation
Ethical Principles and Frameworks

Figure 3. Design space for T3.3

1.4 Timeline for completing Task 3.3

In Figure 4, we provide a complete timeline for the developed and described four phases to complete and deliver T3.3. Timelines for phases 3 and 4 are tentative subject to the confirmation of stakeholder group members and finalising the design space for Delphi (Figure 2). Possible updates to the timeline could be related to the availability of the group members, quality of the feedback provided, and the number of iterations required to develop consensus. We do not foresee any major update to the timeline caused by the above uncertainties.

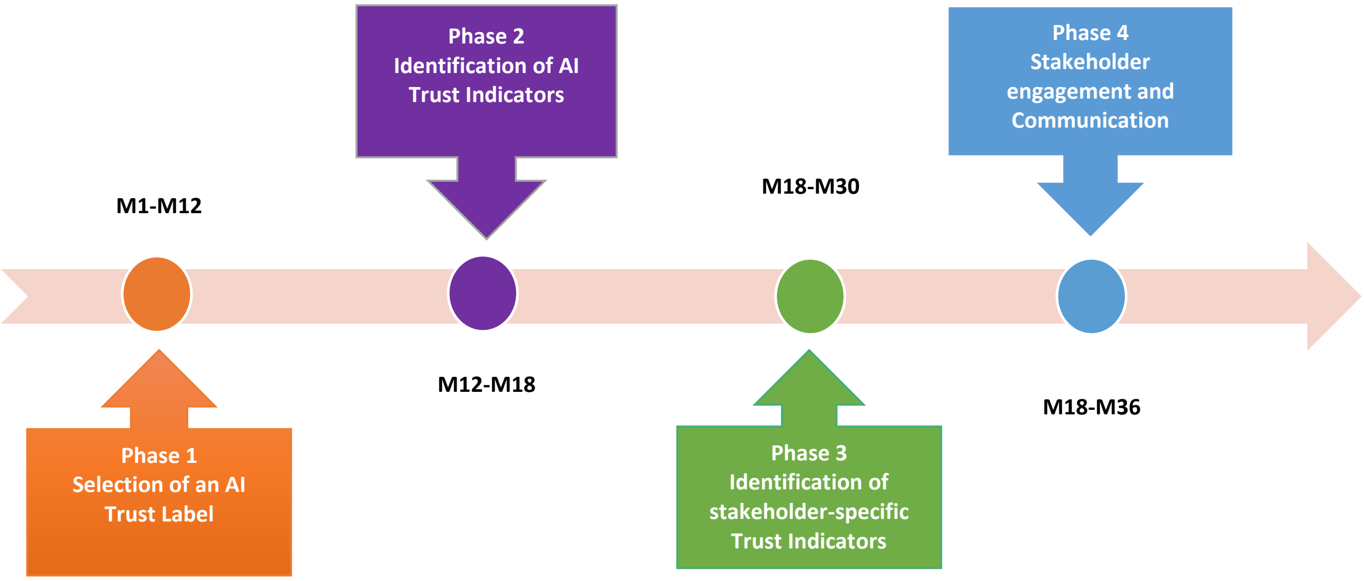

Phase 2
Identification of AI Trust Indicators
Phase 4
Stakeholder engagement and Communication
M1-M12
M18-M30
M12-M18
M18-M36
Phase 1
Selection of an AI Trust Label
Phase 3
Identification of stakeholder-specific Trust Indicators

Figure 4. T3.3 completion timeline

2. Review of the related literature

2.1 AI Trust literature

Both researchers and policy thinkers are wrestling with the previously referenced questions. As a result of this ongoing discussion, several high-minded guiding principles about AI design, development, and usage have been defined and publicly evaluated:

AI Act: The draft regulation aims to ensure that AI systems placed on the European market and used in the EU are safe and respect fundamental rights and EU values.

GDPR: The GDPR is an important component of EU privacy law and human rights law, in particular Article 8 of the Charter of Fundamental Rights of the European Union. The principles of GDPR are Lawfulness, Fairness, and Transparency; Purpose Limitation; Data Minimisation; Accuracy; Storage Limitations; Integrity and Confidentiality; and Accountability.

EU HLEG on AI's Ethics Guidelines for Trustworthy AI: On 8 April 2019, the High-Level Expert Group on AI presented Ethics Guidelines for Trustworthy Artificial Intelligence. According to the Guidelines, trustworthy AI should be: lawful - respecting all applicable laws and regulations, ethical - respecting ethical principles and values, and robust - both from a technical perspective while taking into account its social environment

GPAI's guidelines: General Purpose AI guideline is 100 pages of guidance for developers of cutting-edge general-purpose AI systems (GPAIS) and foundation models. This guideline suggests US and EU policymakers to implement the following strategies: Ensure that developers of GPAIS, foundation models, and generative AI adhere to appropriate AI risk management standards and guidance; Ensure that GPAIS, foundation models, and generative AI undergo sufficient pre-release evaluations to identify and mitigate risks of severe harm, including for open source or downloadable releases of models that cannot be made unavailable after release; Ensure that AI regulations and enforcement agencies provide sufficient oversight and penalties for non-compliance

IBM’s principles of trust and transparency: AI should augment human intelligence rather than replace it, trust is key to adoption, and data policies should be transparent.

The Asilomar’s AI principles: Drafted at the 2017 Asilomar Conference, these 23 principles cover research, ethics, and values in AI, in addition to long term issues. The principles have been signed by 1,273 researchers and 2,541 other interested parties, including Elon Musk and the late Stephen Hawking.

Partnership on AI (PAI): Eight tenets for an open and collaborative environment to discuss AI best practices, the social responsibility of companies delivering AI, AI explainability, and trust. Every partner that wants to join the PAI needs to sign onto these tenets.

The AI4PEOPLE principles and recommendations: Concrete recommendations for European policymakers to facilitate the advance of AI in Europe.

The World Economic Forum’s principles for ethical AI: Five principles that cover the purpose of AI, its fairness and intelligibility, data protection, the right for all to exploit AI for their wellbeing, as well as the opposition to autonomous weapons.

The Institute of Electrical and Electronics Engineers general principles: a set of principles that place AI within a human rights framework with references to wellbeing, accountability, corporate responsibility, value by design, and ethical AI.

Such principles are an important first step, but these conversations should be followed by concrete action to implement viable solutions. Literature highlights a few major and overarching elements that can gain trust of the consumers.

Explainability and transparency

Companies and users want AI systems that are transparent, explainable, ethical, and properly trained with appropriate data. Yet too often, commercially available AI systems are an opaque black box, offering users scarce visibility about the underlying data, processes, and logic that lead to the system's decisions. The most successful machine-learning approaches, such as those based on deep learning, are non-transparent and do not provide easy access into their decision-making. This makes explainability an outstanding challenge, although some attempts to demystify the technology are underway, including OpenScale from IBM.

Bias Awareness and Mitigation

Bias detection and mitigation are also fundamental in achieving trust in AI. Bias can be introduced through training data, when it is not balanced and inclusive enough, but it can also be injected in the AI model in many other ways. Moreover, among the many notions of fairness, it is important to choose the most appropriate given the specific application context. It is also important to help developers become aware of what is available and can be used in current AI systems because of the abundance of bias metrics, notions of fairness, and bias mitigation and detection algorithms. The global community of data scientists and developers can and should continue to improve upon these capabilities in a collaborative way. To that end, IBM has made available to the open-source community a toolkit called "AI Fairness 360" to help developers and data scientists check for and mitigate bias in AI models using bias-handling solutions, and supporting them with guidelines, datasets, tutorials, metrics, and algorithms.

Trusting AI Producers

Trust in the technology should be complemented by trust in those producing the technology. Yet such trust can only be gained if companies are transparent about their data usage policies and the design choices made while designing and developing new products. If data are needed to help AI make better decisions, it is important that the human providing the data is aware of how his/her data are handled, where they are stored, and how they are used. Regulations such as the General Data Protection Regulation (GDPR) in Europe provide some fundamental rights over personal data. Besides performance and accuracy, bias definition and detection and mitigation methods should also be communicated clearly, and explainability capabilities described and made accessible to all users.

The good news is that the industry is beginning to offer such solutions. For instance, research scientists and engineers at IBM have released a set of trust and transparency capabilities for AI, designed around three core principles: explainability, fairness, and traceability. These software tools provide explainability and bias detection on AI models in real time, detecting potentially unfair outcomes and automatically recommending data to add to the model to help mitigate bias.

Imagine an insurance company searching for unintended bias in their AI-powered claim fraud detection process. Using these tools, the company could flag discrepancies between normal and actual approval rates, identify any bias affecting the decision, and highlight factors that may influence why a claim was denied. The toolkit also shows a measure of the confidence that the system has in a recommendation and the factors behind that confidence level. The system would also automatically recommend adding certain kinds of data to help reduce instances of bias moving forward.

Additionally, businesses operating in heavily regulated industries often require extensive information about the decision processes of their AI systems. The ability to track the accuracy, performance, and fairness of their applications, and of recording this information, can provide that level of detail for compliance, accountability, or customer-service purposes. To this end, IBM has proposed the idea of an "AI factsheet", where developers should record all design decisions and performance properties of the developed AI system, from the bias handling algorithms, to the training datasets, to the explainability tools, etc. Also, to help developers and designers think about these issues, IBM has released a booklet, called "Everyday Ethics for Artificial Intelligence", to raise the awareness of developers and designers on these topics and help them to think and find solutions to trust-related capabilities in their everyday job.

2.2 Trust Building Factors

To ensure high quality of systems and build trust it is advised to enable their auditing, that is enabling testing and monitoring them. Different frameworks and certifications are suggested for making the audition process easier: standardised mechanisms and metrics for the AI products trustworthiness, FactSheets, governance frameworks, full-stack supply chain [6] or Social Impact Statement for Algorithms (Drobotowicz, 2020). Drobotowicz (2020) highlights a need to have clear information about **who is accountable** for the service. Before such a piece of information is published, however, responsible organisations need to select **accountable people** for system operations. Moreover, it is important to provide an **accessible avenue of redress**. Floridi et al. (2018) suggests an **AI watchdog** to ensure the auditing of allegedly unfair or inequitable uses of AI; A guided process for registering a complaint akin to making a Freedom of Information request; and the development of liability insurance mechanisms, which would be required as an obligatory accompaniment of specific classes of AI offerings in EU and other markets.

Another way to mitigate the risks of automated AI systems is to **monitor it and keep control over it**. In detail, that could mean having humans decide on which decisions should be taken automatically or by humans (Floridi et al., 2018) (IEEE, 2017), especially leaving decisions that can affect people's lives to experts. Such a solution is being called **human-in-command** (European Commission, 2020). Another option would be to have the **human-in-the-loop** approach (Drobotowicz, 2020), which contains monitoring a system's operations and intervening when needed. Not giving too much autonomy to AI can decrease risk perception and lead to improving users' attitude to it (Rzepka & Berger, 2018).

Looking at AI systems purely from the user perspective, it is advised that they can be in control over the system and data. They should always be able to **invoke needed AI services**, but also **revert, change or disable them** (Amershi et al., 2019), or **ask for human interaction instead of automated one** (European Commission, 2020). Moreover, users should also be provided **privacy** (IBEC, 2022). From the user perspective, they should **have control of their own data and be always able to access it** (IEEE, 2017) (Amershi et al., 2019). That relates to **consent**, that is asking users beforehand to allow the use of their data (Floridi et al., 2018). The other privacy measures are from an organisation's perspective: it should **restrict the amount and age of data held** (IEEE, 2017), do not sense data from personal spaces or any intimate thoughts or emotions, and anonymise personal profiles.

An important but often forgotten factor impacting user trust on the AI tool is its **interface**. Some of its features can be **derived from the transparency and user control need** (Yang & Wibowo, 2022) and as Zerilli et al. (2022) claim: **transparency can modulate trust** in AI and help users to **understand underlying algorithms** and give the potential for better control.

Moreover, transparent-driven interface design helps users to **distinguish their interaction with human and AI** (Floridi et al., 2018) (Mylrea & Robinson, 2023). Kosan et al. (2023) add the need to enable **user feedback** and keep on adapting and **personalising the tool based on users' actions and feedback**. Lastly, predictable outcomes and behaviour can also help in growing trust in the system. Another important trustworthiness factor of AI systems is how **users perceive**

its benefit and efficiency for a specific task (Schaefer et al., 2016). Research shows different benefits such as **well-being** of as many people as possible and **societal and environmental well-being** (Floridi et al., 2018). Furthermore, it is important that any AI systems follow **human rights** (European Union, 2020; IBEC, 2022), **reduce social inequities** (Kelly et al., 2023) and **maintain bonds of solidarity between people** (Floridi et al., 2018). Moreover, AI systems also should be **compatible with cultural diversity, social norms and values** (Angelo et al., 2022) (Ulnicane, 2022); they cannot impose any lifestyle choices on society.

Citizens need to know **when AI is used**, **how and for what purpose**, as well as **what data is used** and **why they receive specific results**. Citizens' needs and concerns, as well as ethical requirements, ought to be addressed in the design and development of trustworthy AI services. Those are, for example, **mitigating discrimination risks**, **providing citizens with control over their data** and having a **person involved in AI processes**. Designers and developers of trustworthy public sector AI services should aim to understand citizens and guarantee that their needs and concerns are met, through the transparent service and the positive experience of using the service (Drobotowicz, 2020).

All the above trust requirements could be divided into two parts: **Information Transparency** and **Principles**. The former presents the information needed on different stages of the AI system and the priority of the information. The latter presents requirements during service design, development and operation stages.

2.3 What is the basis of trustworthiness?

The attributes of an AI system which constitute the basis of trustworthiness are the following (Yang & Wibowo, 2022).

Ability (this is system oriented) refers to the capabilities of the AI system regarding its output or the function it provides to the user: what the AI can do? overall **performance**, **fairness**, **robustness**, **improvability** of the system algorithm with new input.

Intention/goodwill refers to the degree of goodwill behind the creation of the technology: f or what was the AI developed? intended use (e.g., social good) and intended compliance (e.g., privacy-preserving)

Process integrity (this is outcome oriented) refers to whether the operational or decision process of the system is appropriate to achieve the users' goal: how the AI works? context- and user-dependent decision, aligning user's needs/goals with known decision processes.

The above three attributes determine the **level of trust that users should have in an ideal world**. However, these attributes need to be **communicated through trustworthiness prompts**, and then the prompts are judged through a plurality of cognitive processes, both of which introduce noise. A trustworthiness prompt is any information within a system that can prompt, or contribute to, users' trust judgments. Prompts are like indicators in the sense that they provide specific information to the users on the system they aim to use.

This is presented in Figure 5.

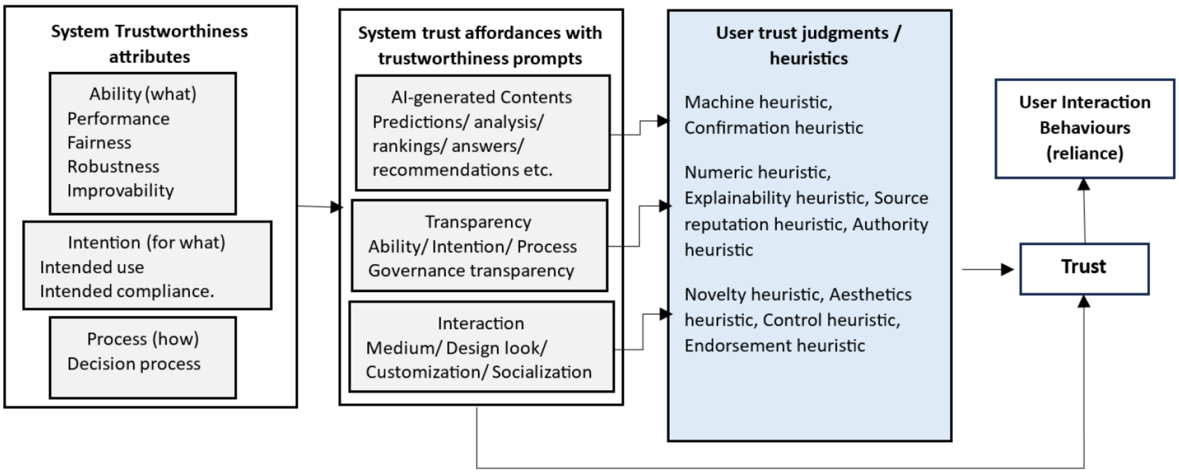

System Trustworthiness attributes
Ability (what)
Performance
Fairness
Robustness
Improvability
Intention (for what)
Intended use
Intended compliance.
Process (how)
Decision process
System trust affordances with trustworthiness prompts
AI-generated Contents
Predictions/ analysis/ rankings/ answers/ recommendations etc.
Transparency
Ability/ Intention/ Process
Governance transparency
Interaction
Medium/ Design look/
Customization/ Socialization
User trust judgments / heuristics
Machine heuristic, Confirmation heuristic
Numeric heuristic, Explainability heuristic, Source reputation heuristic, Authority heuristic
Novelty heuristic, Aesthetics heuristic, Control heuristic, Endorsement heuristic
User Interaction Behaviours (reliance)
Trust

Figure 5. What makes the AI system trustworthy?

2.4 Trust affordances of AI systems

Trust affordances of AI systems can be challenging to identify. Literature (Liao & Sundar, 2022)suggests three types of common affordances of AI systems: **AI-generated content**, **transparency**, and **interaction**. All these affordances will prompt users to trust the system.

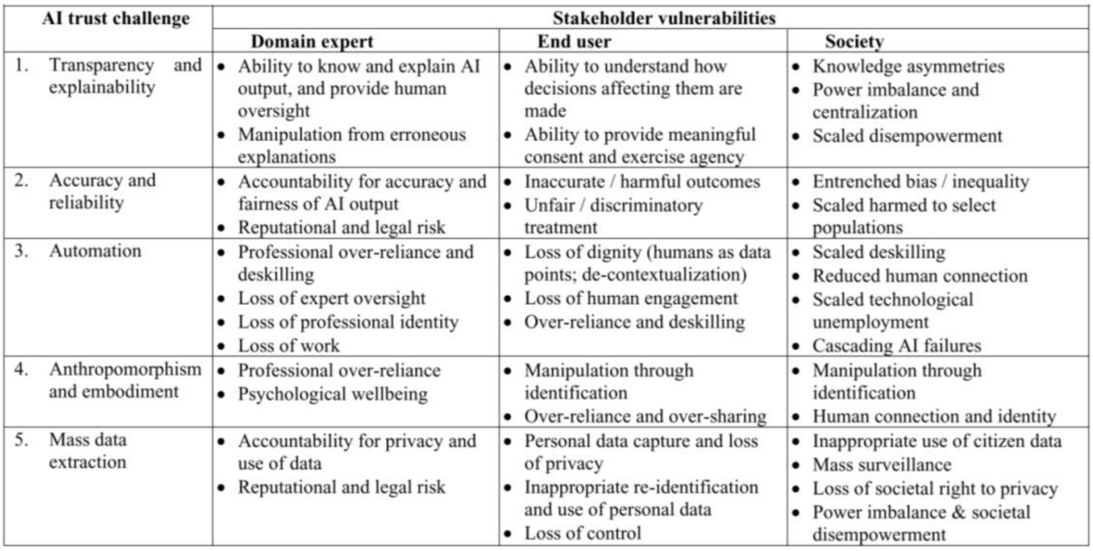
AI-generated content concerns refers to displays of the model output or the functional support provided by the AI system. Depending on the type of model, displays can take the form of a predicted class label, a score, a list of suggestions, generated texts or images, etc. These displays can serve as direct trust worthiness cues for users to assess the ability attributes of the AI model. In some cases the design, e.g., under what circumstances AI assistance is provided or not, can also cue users' judgment of the intention benevolence of the model. These outputs can serve as direct trustworthiness prompts for users to assess the ability attributes of the AI system.

Transparency means outputs allowing a better understanding of the model, broadly defined, including its behaviours, processes, development, and so on. We single out transparency as a unique affordance of AI systems given the increasing industry emphasis on providing transparency. Transparency regarding system performance, fairness, robustness, and improvability, etc. **Governance structures** to ensure trustworthy AI, such as *internal reviews, testing, independent and government oversight*, and so on. Communicating the process and outcomes of such governance structures should also be considered a form of transparency.

Interaction means outputs that suggest how users can interact with the system, beyond the content of the output, for which we consider both **perceptual affordances** (e.g., medium, using a visualisation, and design look) and **action affordances** (e.g., customization of the system, socialisation possibilities with other people).

2.5 AI trust matrix

Literature (Lockey et al., 2021) suggests an AI trust concept matrix with five trust indicators and vulnerabilities each indicator creates for the stakeholders (Figure 6).



AI trust challenge	Stakeholder vulnerabilities		
	Domain expert	End user	Society
1. Transparency and explainability	• Ability to know and explain AI output, and provide human oversight • Manipulation from erroneous explanations	• Ability to understand how decisions affecting them are made • Ability to provide meaningful consent and exercise agency	• Knowledge asymmetries • Power imbalance and centralization • Scaled disempowerment
2. Accuracy and reliability	• Accountability for accuracy and fairness of AI output • Reputational and legal risk	• Inaccurate / harmful outcomes • Unfair / discriminatory treatment	• Entrenched bias / inequality • Scaled harmed to select populations
3. Automation	• Professional over-reliance and deskilling • Loss of expert oversight • Loss of professional identity • Loss of work	• Loss of dignity (humans as data points; de-contextualization) • Loss of human engagement • Over-reliance and deskilling	• Scaled deskilling • Reduced human connection • Scaled technological unemployment • Cascading AI failures
4. Anthropomorphism and embodiment	• Professional over-reliance • Psychological wellbeing	• Manipulation through identification • Over-reliance and over-sharing	• Manipulation through identification • Human connection and identity
5. Mass data extraction	• Accountability for privacy and use of data • Reputational and legal risk	• Personal data capture and loss of privacy • Inappropriate re-identification and use of personal data • Loss of control	• Inappropriate use of citizen data • Mass surveillance • Loss of societal right to privacy • Power imbalance & societal disempowerment

Figure 6. AI Trust Concept Matrix

2.6 AI Trust Conceptual Frameworks

The first set of principles to promote safe and beneficial AI development was proposed in 2017 at the “Asilomar Conference on Beneficial AI”. Of the 23 resulting principles, 13 are on ethics and values, including safety, failure transparency, judicial transparency, and responsibility [20]. The same year, the Montreal Declaration for a Responsible Development of Artificial Intelligence was announced to stimulate public debate and encourage a progressive and inclusive orientation to the development of AI. Its main objectives include Developing an ethical framework for the development and deployment of AI; Guiding the digital transition to enable everyone to benefit from the technology and Opening a national and international forum for discussion to collectively achieve equitable, inclusive, and ecologically sustainable AI development [21]. The two most recent developments are the European Commission’s High-Level Expert Group on AI [4] [22][23] and The Global Index for Responsible AI [26]. The first developed a framework to guide the European AI community in developing and using “trustworthy AI” (i.e., AI that is lawful, ethical, and robust). The framework includes Human Agency and Oversight, Technical Robustness and Safety, Privacy and Data Governance, Transparency, Diversity, Societal and Environmental wellbeing, and Accountability. The guidelines emphasise four principles: respect for human autonomy, prevention of harm, fairness, and explainability. The Global Index for Responsible AI establishes and provides a set of benchmark indicators to be used by all 120 participating countries to assess compliance with responsible AI practices. Indicators are grouped into three broad groups: Preconditions for Responsible AI, Responsible AI Governance, and Responsible AI Capacities [25].

IEEE is also a pioneer of trusted AI. IEEE published two versions of ethical guidelines for intelligent systems, IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems and Ethically Aligned Design: A Vision for Prioritising Human Wellbeing with Artificial Intelligence and Autonomous Systems, which are high-level general principles for autonomous and intelligent design [20].

A recent research article published in Telecommunications Policy journal [4] proposed an emerging AI system ethical regime that includes six areas, namely transparency, accountability, fairness, privacy, reliability, and human control, as well as 27 requirements. This article breaks down the transparency into five different kinds of information to the public: Documentation, Notification, Traceability, Reproducibility, and Explainability. Fairness considers whether people are treated equally in a decision-making process and is broken down into four different kinds of information to the public: Bias prevention, Data representativeness, Inclusive benefit distribution, Inclusiveness in

Design. Accountability can be understood as "a relationship in which a decision-maker is asked to report on their activities, and likely involving sanctions in the case of misconduct". Accountability is broken down into five different kinds of information to the public: Verification and validation, Assessment, Auditability, Appealability, and Liability and legal responsibility. Privacy is a widely recognized right both in international human rights law and national legislation in almost all democratic countries and is broken down into four different kinds of information to the public: Consent, Data minimization, Data agency, and Anonymisation. An AI system is expected to be reliable, meaning that proper arrangements should be put in place to ensure the ongoing correct functioning for the purpose it was created and to avoid unintended harm to people, regardless of whether it is caused by design or manufacturing faults, malfunctioning, external threats, or misuse. Reliability can be broken down into the following requirements: Safety, Security, Resilience, and Predictability. The Human Control indicator entails that humans retain the ultimate say on AI operations and outputs to safeguard human dignity and autonomy, preventing people from being subjected to completely automated and unsupervised processes. This principle can be broken down into the following more specific requirements: Human oversight, Human review, and Opt-outs.

Other examples include AI4People [25], which developed a Unified Framework of Principles for AI in Society and presented a synthesis of five ethical principles that should undergird its development and adoption. These principles include Right to Transparency; Right to Human Determination; Identification Obligation; Fairness Obligation; Assessment and Accountability Obligation; Accuracy, Reliability, and Validity obligation; Data Quality Obligation; Public Safety Obligation; Cybersecurity Obligation; Prohibition on Secret Profiling; Prohibition on Unitary Scoring; Termination Obligation. In addition, AI4People offers 20 concrete recommendations—to assess, develop, incentivise, and support good AI—which in some cases, may be undertaken directly by national or supranational policymakers. Finally, The Public Voice [27] proposed guidelines that aim to improve the design and use of AI, maximise the benefits of AI, protect human rights, and minimise risks and threats associated with AI. They claim that the guidelines should be incorporated into ethical standards, adopted in national law and international agreements, and built into the design of systems.

A summary of this section is presented in Table 2.

Table 2. AI Trust framework based on the literature

Frameworks	Scope and Requirements
First set of principles to promote safe and beneficial AI development at the "Asilomar Conference on Beneficial AI"	Safety, failure transparency, judicial transparency, and responsibility
Montreal Declaration for a Responsible Development of Artificial Intelligence	Well-being, Autonomy, Privacy and intimacy, Solidarity, Democracy, Equity, Inclusion, Prudence, Responsibility and Accountability, Sustainable development
European Commission's High-Level Expert Group on AI	Human Agency and Oversight, Technical Robustness and Safety, Privacy and Data Governance, Transparency, Diversity, Societal and Environmental well-being, and Accountability
The Global Index for Responsible AI	Preconditions for Responsible AI, Responsible AI Governance, and Responsible AI Capacities
IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems and Ethically Aligned Design	Human Wellbeing and Autonomous Systems
AI system ethical regime	Transparency, accountability, fairness, privacy, reliability, and human control

AI4People - Unified Framework of Principles for AI in Society	Transparency; Human Determination; Identification; Fairness; Assessment and Accountability; Accuracy, Reliability, and Validity; Data Quality; Public Safety; Cybersecurity; Prohibition on Secret Profiling; Prohibition on Unitary Scoring; Termination
The Public Voice	Design and use of AI, human rights, risks and threats associated with AI, ethical standards, national law and international agreements

3. Four Phases to Complete T3.3 – Achievements and Future Directions

3.1 *Phase 1 – Selection of the AI Trust Label* (Completed)

This section outlines and elaborates phase 1 - selection of the AI Trust Label for T3.3.

3.1.1 Review of the AI trust initiatives (self-regulatory initiatives)

AI self-regulation mechanisms include using AI trust labels, implementing a code of conduct, and voluntary adoption of AI standardisation processes.

There are two broad categories of self-regulatory mechanisms [32]:

1) *Labelling, certification schemes*, and other initiatives that define a certain standard for AI applications and outline a set of criteria against which that standard is assessed, generally through an audit process. This consists of a broad category of initiatives that contains the following types of mechanisms and tools: *labels and certification schemes; kite marks, trust marks and quality marks; and seals*; and

2) *Codes of conduct* can be characterised as statements that set out and define specific requirements or principles that should be followed by organisations developing or procuring AI applications to ensure the safe and ethical development and use of these systems. These generally do not define measurable criteria or include an audit process. This category includes *codes of conduct and ethics,* often used interchangeably in the literature.

We found a range of labelling and certification initiatives, codes of conduct, and a selection of additional self-regulatory mechanisms proposed or implemented in the context of AI applications. List of the initiatives and their types are presented in Table 3 with detailed analysis in PART B.

Table 3. Types of AI Trust initiatives

Types	Initiatives Names
Labels	The AI Trust label, Gender Equality European & International Standard – Artificial Intelligence, Mandatory Labelling Scheme, Open Ethics Label, Swiss Digital Trust Label
Certificates	Certification System for AI Applications (Fraunhofer Institute), Certificate of Fairness for AI Systems, Certification Mechanism for AI Tools and Services, Certification of Responsible Limits on Facial Recognition, Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS), EU Certification for 'Trusted AI' Products, Malta's National AI Certification Framework, Responsible Artificial Intelligence Certification Beta (RAII), SECure: A Social and Environmental Certificate for AI Systems, The Certification as a Mechanism for Control of Artificial Intelligence in Europe, Turing Stamp
Quality, Trust, and Kite marks	Certification System for AI Applications (Fraunhofer Institute), Foundation for Responsible Robotics (FRR) Quality Mark for (AI-Based) Robotics, Fair Artificial Intelligence in Education (FairAIEd) Trust Mark, OPACITY trust mark, Kite Mark for AI
Rating Framework	CEN-CENELEC 'Road Map on Artificial Intelligence (AI)'
Code of Conduct	Ethical Use of Artificial Intelligence in Canadian Financial Services, A Guide to Good Practice for Digital and Data-Driven Health Technologies, Algo.Rules, Clinical AI Governance, Guiding Principles for AI Ethics, IFC

	Technological Code of Conduct, Oxford-Munich Code of Conduct, Partnership on AI Tenets
Code of Ethics	AI2ES 'Code of Ethics', BMW Group's Code of Ethics for the Use of AI, Bosch Code of Ethics for AI, Capgemini's Code of Ethics for AI, Continental Code of Ethics for AI, Ethical and Professional Guidance on Data Science: A Guide for Members, iCIMS Code of Ethics
Seal and Audit	D-seal, O'Neil Risk Consulting & Algorithmic Auditing Seal, Z-Inspection

3.1.2 Review of the four major AI regulations

Governments in other parts of the world are also developing new legislation to ensure the responsible deployment of AI in the workplace. In this section, we present an overview of the three internationally influential regulatory frameworks: the *Artificial Intelligence and Data Act (AIDA) of Canada*, the AI *Bill of Rights of the United States*, and the *AI Regulation of Japan*.

AIDA is the first federal law in Canada to regulate the development and use of AI systems. The Bill of Rights is a Blueprint and non-binding document to help guide the design, use, and deployment of automated systems. Japan's AI regulatory policy is a non-binding guideline to maximise AI's positive impact on society rather than suppress it out of overestimated risks. While the EU AI Act, AIDA, and the AI Bill of Rights set out a centralised approach to regulating AI across all sectors, the AI Regulation in Japan sets out a de-centralised and sectorial approach to AI regulation leveraging the experience and expertise of existing regulators to issue guidance and highlight the relevant regulatory requirements applicable to the businesses they regulate. This requires that businesses take appropriate measures and disclose information about risks [30].

A summary of these regulations is presented in Table 4 and with detailed analysis in PART B.

Table 4. The four internationally influential regulatory frameworks

Frameworks	Approach	Regulated AI Systems	Focus and responsibilities	Requirements
Artificial Intelligence and Data Act of Canada (Binding)	Centralised	General and high-impact systems	To provide measures to identify, assess and mitigate the risks of harm or biased output; Measures to monitor compliance with the mitigation measures and the effectiveness of those mitigation measures; Publish on a public website a plain-language description of the system; Notify the Minister of Industry if the use of the	Anonymised data, Assessment of high-impact system, Measures related to risks, Monitoring of mitigation measures, keeping general records, Additional records, publication of description related to making the system available for use, Publication of description related to managing operation of the system, and Notification of material harm

			system results or is likely to result in material harm	
AI Bill of Rights (Non-binding)	Centralised	High-risks systems	To guarantee the protection of civil rights, civil liberties, and privacy of American citizens; guide the design, use, and deployment of automated systems; and actualise democratic values and principles in the technological design process of AI systems	Safe and effective systems, Algorithmic discrimination protections, Data privacy, Notice and Explanation, Human Alternatives, consideration, and feedback
AI Regulation of Japan (Non-binding)	De- centralised	All AI systems (Sectorial approach)	Focuses on risk-based, agile, and multistakeholder process to guide the improvement and implementation of AI policy in a wide range of industries through industry-specific Acts.	Ownership/intellectual property rights regarding AI, Competition, Data protection under the Act on Protection of Personal Information, Regulation/Government intervention, and Civil liability
EU AI Act (Binding)	Centralised	High-risks systems	A risk-based approach to regulate all automated technology rather than specific areas of concern to guide the use of AI in both the private and public sectors. The approach defines three risk categories: unacceptable risk applications,	Risk management and testing, Data and data governance, Technical documentation, Record Keeping, Transparency and human oversight, Accuracy, robustness, and cybersecurity

			high-risk applications, and applications not explicitly banned.	

3.1.3 Short-listing the reviewed initiatives

The aim of short-listing is to shorten the list of existing initiatives to a list that is more relevant to the goal of T3.3. After reviewing the relevant AI trust self-regulatory initiatives, we used the following criteria to short-list initiatives for in-depth analysis.

Short-listing criteria include :

- Initiatives of types 'label' and 'certification'
  - According to the categories presented in section 3.1.1 Review of the AI trust initiatives (self-regulatory initiatives), literature defined two broad categories of initiatives. The scope and the goal of T3.3 is to focus on the first category which is label and certification.
- Initiatives that are cross-sectoral and horizontal
  - As can be seen from the Table presented in section 3.1.1 Review of the AI trust initiatives (self-regulation initiatives), the scope and focus of initiatives varies. Some initiatives focus on a specific sector or industry (sector-specific) while others provide horizontal requirements that are relevant to various sectors (cross-sectoral). The aim of T3.3 is to select a label that is good to multiple sectors and provides horizontal requirements.
- Initiatives that are developed and in use
  - Some initiatives are still in development phase , finalisation phase, or completed but not evaluated and not in-use. For short-listing, we aim to select initiatives that are either evaluated or evaluated and in-use .

The list of short-listed initiatives is presented in Table 5.

Table 5. Short-listed initiatives for a more in-depth analysis

Name	Type	Sector	Scope	Short Description
The AI Trust Label (VDE)	Label	Generic	Germany	Inspired by the EU energy-efficiency label. It shows a rating of an AI system's ethical characteristics based on six ethical values.
Mandatory Labelling Scheme	Label	Generic	Germany	The German Data Ethics Commission recommended the introduction of a mandatory labelling scheme for algorithmic systems of enhanced criticality, with the view that this would oblige operators to make it clear whether, when and to what extent algorithmic systems are being used.
Open Ethics Label	Label	Data & decision technology	Europe	The Open Ethics label aims to strengthen users' trust in AI systems by encouraging and supporting AI transparency. For the consumer, this label provides information to enable better decision-making; for software developers, the label is a type of disclosure tool to provide information about their product.

Certification System for AI Applications (Fraunhofer Institute)	Certificate and quality mark	Generic	Germany	The AI certification (Fraunhofer Institute) consists of a certification system and quality mark to signal the technical reliability of an AI system and responsible usage from an ethical and legal perspective. Furthermore, it aims to facilitate comparison between different products and help promote open competition in AI.
Z-Inspection	Audit process	Generic	Europe	Z-Inspection is an audit process that assesses whether an AI system is trustworthy. The process is based on applied ethics and uses the definition of trustworthy AI put forward by the European Commission's AI HLEG. The process is designed to be applied to a variety of sectors in which AI systems could be used, such as business, healthcare and the public sector
Swiss Digital Trust Label	Trust mark	Generic	Swiss	Denotes the trustworthiness of a digital service in clear, visual, and plain, non-technical language for consumers.
Malta's National AI Certification Framework	Certificate and audit	AI sector	Malta	The certification aims to build trust and transparency by providing valuable information about AI in their marketplace to signal that their AI systems have been developed ethically, transparently and in a socially responsible manner.
EU Certification for 'Trusted AI' Products	Certificate	Generic	EU	Certification for trustworthy AI applications, where products are tested for resilience, safety and absence of prejudice, discrimination, or bias.
Responsible Artificial Intelligence Institute Certification Beta (RAII)	Certificate	Generic	Global	The RAI Certification Beta is an independent certification programme for the responsible and trusted use of AI systems. The certification aims to increase trust among end users by signalling that the AI system was built following specific standards.
Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)	Certificate	Generic	Global	ECPAIS consists of a certification system that aims to signal to stakeholders in different sectors whether an AI system is safe, ethical, and trustworthy. Ultimately the certification aims to promote responsible innovation in AI systems.
Certificate of Fairness for AI Systems	Certificate	Generic	United Kingdom	A certificate of fairness for AI systems alongside a kite mark type scheme to display it, with criteria to be defined at the industry level. The certification helps build an AI that avoids discrimination against women and ethnic minorities.

3.1.4 Analysis and Selection of the Trust Label

To select a trust label for this task, we analysed the short-listed initiatives in the context of the AI Act as well as the major and influential international regulations presented in 3.1.2 Review of the four major AI regulations.

Analysis - The rationale behind this approach of analysis is to make sure that the selected AI trust label has:

- High coverage and is compatible with the AI Act mandatory requirements (Articles 9 to 15)
  - A label that promotes an ecosystem of trust and an innovation-friendly market for AI in Europe
  - A label that supports the application of AI Act mandatory requirements in the context of non-high-risk AI systems, therefore improves trust of the consumers
  - A label that has the potential to become widely used by the Member States
  - A label that provides opportunities for developers of non-high-risk AI systems to harness the label to voluntarily address AI Act mandatory requirements within and beyond Europe (internationalisation)

- High coverage and is compatible with the influential international AI regulations
  - A label that promotes an ecosystem of trust and an innovation-friendly market for AI in Europe and beyond
  - A label that supports compliance with AI regulations within and beyond Europe
  - A label that facilitates international trade relations and supports interaction between international AI regulations
  - A label that supports developers of non-high-risk AI systems in their standardisation and internationalisation (developers will be confident that the label they are using is compatible with other AI regulations and this provides opportunities for them to market their products outside Europe)

Analysis of the short-listed initiatives in the context of the AI Act, AIDA, AI Bill of Rights, and the AI Regulation of Japan is presented in Part B.

Selection - The rationale presented above, and the in-depth analysis presented in Part B strongly supports the selection of the AI Trust Label for T3.3.

AI Trust Label developed by VDE is the label we selected as part of T3.3 of WP3. The most important selection criteria include:

- Compatible and high coverage with the AI Act
- Good coverage with other major regulations
- Clear methodology
- Value-based or value compliance meaning that it gives flexibility to set target requirements for a value and it describes compliance with the specified values (for example, one product might better comply with privacy requirements, while the other might comply better with transparency criteria)
- Applies to self-certification and third-party conformity
- Industry-academic engagement
- Active contributor community from multiple member states
- Going EU-wide *(to be included in the label's next public release before the end of 2023)*
- Involvement of 4 major industries for vertical requirements *(to be included in the label's next public release before the end of 2023)*
  - Finance
  - Defense
  - Mobility
  - Health

3.2 Phase 2 –Identification of Trust Indicators (Partially Completed)

This section outlines and elaborates phase 2 – identification of trust indicators for T3.3. In Phase 2, we focus on identifying general trust indicators that apply or may apply to a wide range of stakeholders (e.g., government organisations, developers, etc.). In this phase, our intention is not to focus on indicators that are relevant to a specific stakeholder group but to identify trust indicators based on a general logic.

3.2.1 Extracting indicators and criteria from the initiatives

After a comprehensive review and analysis of the short-listed AI trust initiatives presented in 3.1.1 Review of the AI trust initiatives (self-regulatory initiatives), we identified and defined a list of indicators and briefly presented them in Figure 7 -10. A complete analysis is provided in Annex B. Phase 2 – Selection of the Indicators of Trust.

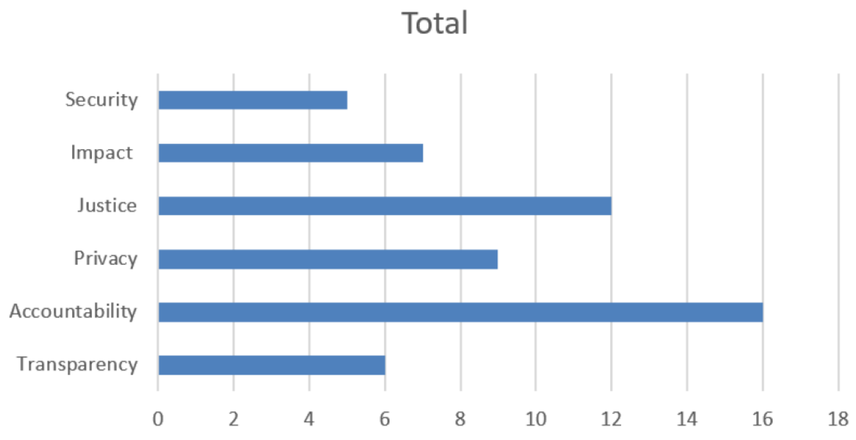


Figure 7. Indicators from the initiatives - counted

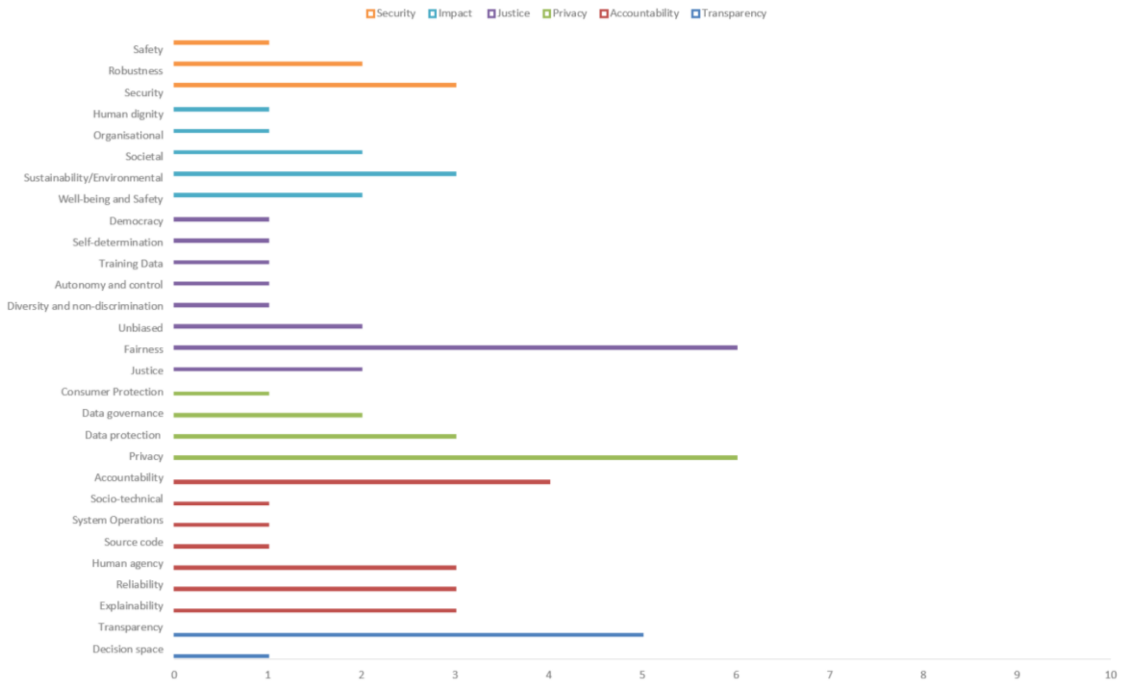
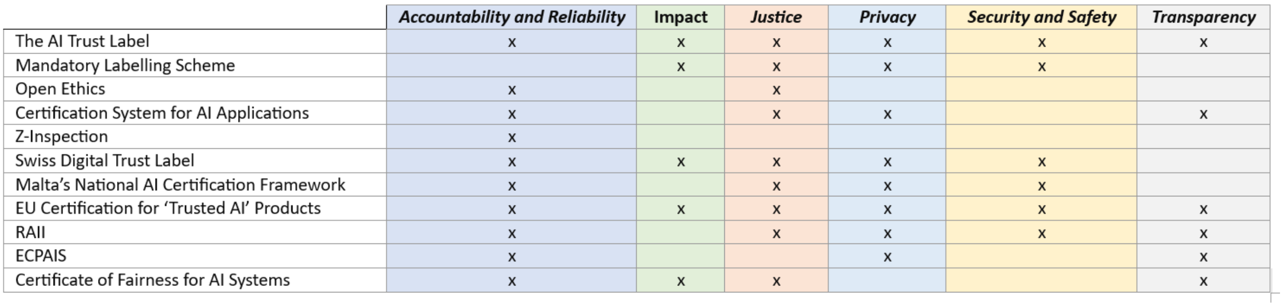
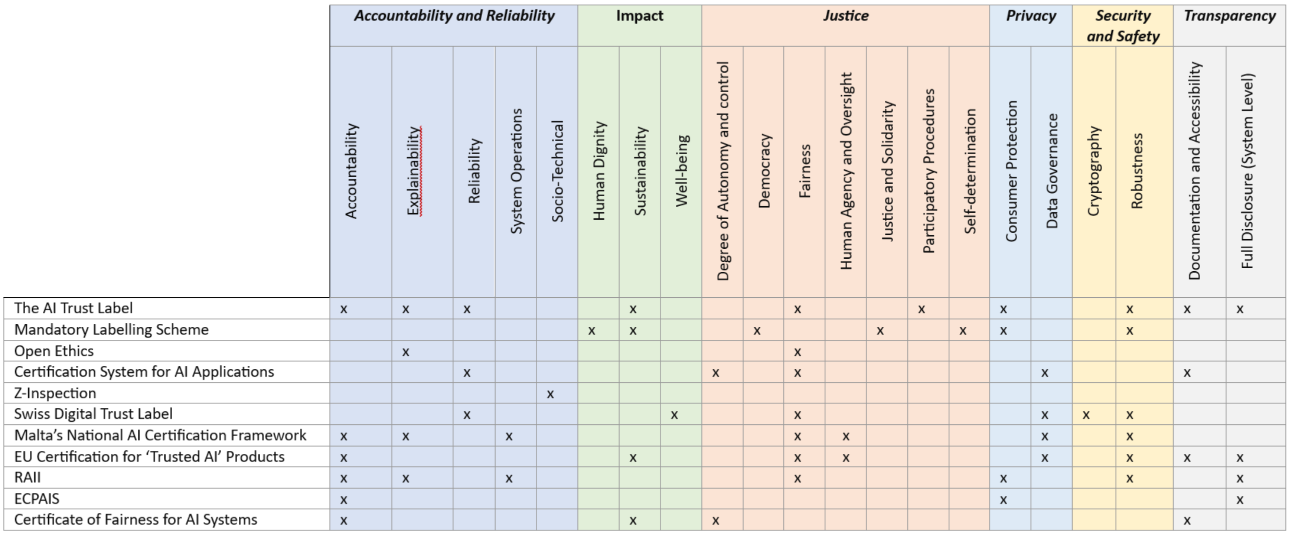


Figure 8. Requirements from the initiatives - counted



	Accountability and Reliability	Impact	Justice	Privacy	Security and Safety	Transparency
The AI Trust Label	x	x	x	x	x	x
Mandatory Labelling Scheme		x	x	x	x	
Open Ethics	x		x			
Certification System for AI Applications	x		x	x		x
Z-Inspection	x					
Swiss Digital Trust Label	x	x	x	x	x	
Malta's National AI Certification Framework	x		x	x	x	
EU Certification for 'Trusted AI' Products	x	x	x	x	x	x
RAII	x		x	x	x	x
ECPAIS	x			x		x
Certificate of Fairness for AI Systems	x	x	x			x

Figure 9. Trust indicators covered in each trust initiative



	Accountability and Reliability					Impact			Justice							Privacy		Security and Safety		Transparency	
	Accountability	Explainability	Reliability	System Operations	Socio-Technical	Human Dignity	Sustainability	Well-being	Degree of Autonomy and control	Democracy	Fairness	Human Agency and Oversight	Justice and Solidarity	Participatory Procedures	Self-determination	Consumer Protection	Data Governance	Cryptography	Robustness	Documentation and Accessibility	Full Disclosure (System Level)
The AI Trust Label	x	x	x				x				x			x		x			x	x	x
Mandatory Labelling Scheme						x	x			x			x		x	x			x		
Open Ethics		x									x										
Certification System for AI Applications			x						x		x						x			x	
Z-Inspection					x																
Swiss Digital Trust Label			x					x			x						x	x	x		
Malta's National AI Certification Framework	x	x		x							x	x					x		x		
EU Certification for 'Trusted AI' Products	x						x				x	x					x		x	x	x
RAII	x	x		x							x					x			x		x
ECPAIS	x															x					x
Certificate of Fairness for AI Systems	x						x		x											x	

Figure 10. Trust indicators and requirements covered in each trust initiative

3.2.2 Review and analysis of the consumer protection requirements

In this part, we will rely on other major regulations (AIA, Bill C-27, Bill of Right, UK AI Regulation and AI Regulation of Japan) to guide the analysis (M19-M23).

3.3 Phase 3 – Selection of consumer trust indicators (Ongoing) (M18-M36)

This phase starts in M18 and we expect it to continue towards the end of the project where we provide trust label recommendations to the EC and the standardisation body in Europe. There is a back-and-forth relationship and connection between phase 3 and phase 4. Stakeholder engagement will provide input into phase 3 activities where we collect, analyse, and internally validate the trust indicators for the consumers of the AI systems.

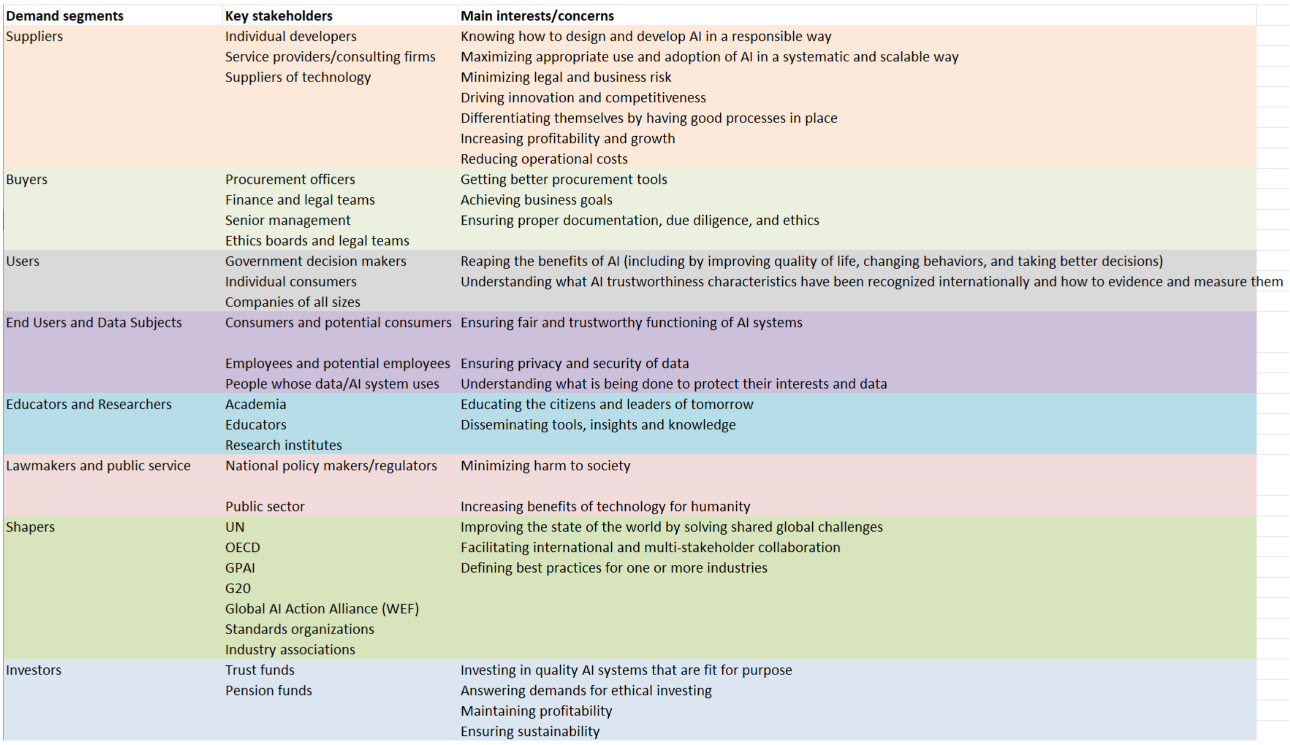
So far, we have made the following efforts to complete this task.

1) We looked at the initiatives to understand how they perceive the trust label and what value they expect it to offer. See Table 6.

Table 6. Stakeholder's perception and expected value from the trust label

Stakeholders	**Value of Trust Label for Stakeholder Group**
All Stakeholders	Sets a clear bar for global best practices to implement AI responsibly, providing certainty, direction, and actionable next steps.
Senior Executives & Executive Review Boards	Gives confidence that the products and services they are deploying are fit for purpose, legally compliant, of an appropriate quality, and scalable.
Compliance Officers	Enables involvement at the design and development phases, thereby avoiding costly and difficult compliance decisions later in the AI system lifecycle
Procurement Officers	Provides processes to procure trustworthy AI systems, enabling an organisation to deliver quality AI products and services while reducing liability and risk.
Regulators	Enables compliance with established regulations and alignment with proposed regulatory approaches.
Investors	Provides assurance that AI systems are built on recognized global best practices.
Consumers	Gives comfort that rights, privacy, and civil liberties are protected.

2) We have also reviewed the literature to understand how literature identified key stakeholders and what their trust concerns or interests are. The summary of the literature review is presented in Figure 11.



Demand segments	Key stakeholders	Main interests/concerns
Suppliers	Individual developers Service providers/consulting firms Suppliers of technology	Knowing how to design and develop AI in a responsible way Maximizing appropriate use and adoption of AI in a systematic and scalable way Minimizing legal and business risk Driving innovation and competitiveness Differentiating themselves by having good processes in place Increasing profitability and growth Reducing operational costs
Buyers	Procurement officers Finance and legal teams Senior management Ethics boards and legal teams	Getting better procurement tools Achieving business goals Ensuring proper documentation, due diligence, and ethics
Users	Government decision makers Individual consumers Companies of all sizes	Reaping the benefits of AI (including by improving quality of life, changing behaviors, and taking better decisions) Understanding what AI trustworthiness characteristics have been recognized internationally and how to evidence and measure them
End Users and Data Subjects	Consumers and potential consumers Employees and potential employees People whose data/AI system uses	Ensuring fair and trustworthy functioning of AI systems Ensuring privacy and security of data Understanding what is being done to protect their interests and data
Educators and Researchers	Academia Educators Research institutes	Educating the citizens and leaders of tomorrow Disseminating tools, insights and knowledge
Lawmakers and public service	National policy makers/regulators Public sector	Minimizing harm to society Increasing benefits of technology for humanity
Shapers	UN OECD GPAI G20 Global AI Action Alliance (WEF) Standards organizations Industry associations	Improving the state of the world by solving shared global challenges Facilitating international and multi-stakeholder collaboration Defining best practices for one or more industries
Investors	Trust funds Pension funds	Investing in quality AI systems that are fit for purpose Answering demands for ethical investing Maintaining profitability Ensuring sustainability

Figure 11. Demand segments, key stakeholders and their interest - from the literature

3) We completed the stakeholder analysis of all the short-listed initiatives to understand what trust indicators and requirements are relevant to a particular stakeholder. Detailed analysis is presented in Annex B. Phase 2 – Selection of the Indicators of Trust
4) We have further completed an analysis of the AI Act requirements considering the results of the analysis from step 3. This is presented in Annex B. Phase 2 – Selection of the Indicators of Trust
5) We further rely on the findings from the previous steps to highlight public facing indicators. Detailed analysis is presented in B.3 Public facing trust indicators

3.4 Phase 4 –Stakeholder Engagement (M19-M36)

This task will start in M19. Some work has been done with regard to forming a stakeholder activity group where we aim to implement the Delphi method to finalise consumer trust indicators. The list of potential members is provided in 1.3 Methodological approach.

In addition to the stakeholder activity group, other stakeholder engagement events such as the ADR Awareness Day will be organised as part of WP3 to support AI Trust Label awareness.

3. Opportunities, Implications, and Barriers

We have used related literature to support the outcome of T3.3. This allows us to provide sufficient grounds for the outcome generated in T3.3.

3.1 Opportunities afforded by the AI Trust label in the context of the European AI trust ecosystem

Trust as the foundation of AI and uptake of AI solutions – Trust in AI is fostered by a clear regulatory mechanism [25], evaluation by auditors [43], data governance [25], and AI risk management [49]. Completed activities in this task show that the *AI trust label has a great potential to simplify and operationalise AI regulations, constitute an environment for voluntary third-party auditing, and provide straightforward and easy-to-understand data governance and risk management requirements in compliance with European AI Regulations*. Additionally, the aim of the AI Act is to harmonise regulation for all AI systems (low-medium-high risk systems). Therefore, its scope must remain highly general to suit the way different consumers want a particular AI system or application to be automated [50]. As an example, a study on Japanese citizens [50] reveals that using AI to provide public services for "parental support" and "waste collection" has some effects on citizens. However, these application domains and the AI systems used in these domains are not qualified as high-risk according to the Act [6]. As a result, the AI Act can be considered as the first step towards regulating a particular AI system at the sectoral level and trustworthiness of such systems. Here, based on our works carried out in this task, we argue that the *trust label can support the application of AI Act mandatory requirements in the context of non-high-risk AI systems to improve public trust in the AI system and the developing institutions and incentivise the public to use automated services*. These opportunities are aligned with Europe's AI agenda on *trust as the foundation of AI and uptake of AI*, outlining that *trust is a prerequisite for the uptake of AI*, particularly by consumers [51].

Ethics and competitiveness – This is similar to the concept of "responsible competitiveness" highlighted in [19] and [43] to simultaneously promote ethical and trustworthy AI [7] in European businesses. On the one hand, while AI provides companies with a unique and enduring competitive advantage, what we observed from the completed activities in T3.3 is that the *AI trust label is a great instrument to ensure AI Act compliance, in particular for the medium and low-risk systems (voluntarily to give more confidence in delivering trustworthy AI solutions) and enable responsible competition among the AI system developers and businesses*. However, *even when adopted voluntarily, label requirements and the labelling process can significantly influence fostering and engineering citizens' trust* [50]. Lack of trust in AI systems and applications has always been and still is one of the most important obstacles [19],[51],[25] that has the tendency to slow down the adoption and uptake of AI products, and thus their potential benefit [50]. On the other hand, a *strong alignment between the AI Act and self-regulatory initiatives such as labels can generate consumer trust in the product and the trustworthiness of the object of trust* (i.e. auditing). According to [50], the level of trust will increase with the enhancement of trustworthiness of AI systems and applications. However, this requires that this enhancement in trustworthiness is *perceived and noticed in the population*. 1) The role of external institutions' "impartiality and independence" from the under assessment providing organisation [Art. 33(4) AI Act [6]] and 2) the willingness of these organisations to subject the system to be audited against trustworthiness of the systems [6][50] are substantial to the quality of judgement of the population with regard to system reliability or trustworthiness. This implies that AI system developers and providers shall "safeguard the independence, objectivity and impartiality of their activities" [Art. 33(5) AI Act [6]].

T3.3 shows that less than half of the labels involve external audit organisations in the auditing process (Table 12). For ethical and responsible competitiveness, this task suggests that the auditing process involves external organisations to verify the conformity of the medium and low-risk systems.

Our finding is aligned with Europe's AI agenda on ethics and competitiveness as complementary [51].

Value-based AI approach – Since AI is understood to have significant societal impacts [52], building trust is considered essential. The preferred European AI approach is grounded in European values, fundamental rights, human dignity, and privacy protection [20]. Outcomes from the completed work in T3.3 confirm that *all the studied labels are grounded in European values and are based on frameworks and principles related to fundamental rights, human dignity, and data protection and privacy* (Table 11). This suggests that the AI trust label should be considered *a common approach to support Europe's agenda and help avoid regulatory uncertainty on ethical and trustworthy aspects of AI that may arise when the AI Act becomes fully implementable*. This is aligned with Europe's strong sense of seeking a European value-based AI approach [51].

Europe being a global leader in trustworthy AI – Europe is expected and perceived to provide a *unique contribution to the global AI debate and a strong AI regulatory framework that sets the global standard* [20],[53]. The *strong attachment of the AI trust label to European values and principles, data protection and privacy regulations, and their ability to promote responsible competitiveness among system developers and companies will contribute to harmonising AI regulations in Europe and beyond*. Outcomes from the completed work in T3.3 are aligned with Europe's strong sense of being a global leader in trustworthy AI. In this regard, Europe states that it is "well positioned to exercise global leadership in building alliances around shared values", "well placed to lead this debate on the global stage" and can "be the champion of an approach to AI that benefits people and society as a whole" [6], [19], [25].

3.2 Opportunities afforded by the AI Trust label for international uptake of AI trust labels and global regulatory standards

AI research reveals concern over increasing global competition, often called an AI race in the literature. Subsequently, the race to AI brings forward a race to AI regulation [54]. As a result, a new playground for global regulatory competition emerges that requires an explicit effort towards setting global regulatory standards for AI that can potentially lead towards harmonising and standardising AI regulation. Indeed, international organisations such as the OECD, the Council of Europe and UNESCO have joined the "regulatory playground", bringing together a diversity of countries to work towards a consensus on trust and ethical considerations raised by AI [54].

On the one hand, outcomes from this task suggest that *AI Trust mechanisms can create the right environment to harness the benefits of AI and, at the same time, get regulation right by enabling effective communication and interaction between the AI regulatory bodies and the regulated companies so that innovators can thrive and the risks posed by AI can be communicated and addressed to improve the uptake of AI systems* [59]. Here, we would also like to add that the AI trust label could be considered as both an innovation testbed and a regulatory sandbox to facilitate and support AI innovation. This is achieved through *reducing the organisational resources necessary to develop and get the AI innovation out to the market and access to finance*. In addition, the AI trust label allows integrating and safeguarding consumer-protection and also helps regulators determine when to regulate a given market or technology. On the other hand, there is a large amount of effort and interest in making the label more efficient by securing the possibility of an international uptake. This task echoes the importance of *global uptake of the self-regulatory mechanisms and supports that their international acceptance and application will contribute to the safe and genuinely positive applications of AI ethics and assure the proper development of trustworthy AI systems and solutions to be used across borders for the benefit of society and economy*. In addition, the global uptake of self-regulatory mechanisms can impact global value chains and international trade [60]. Specifically, the EU and USA will benefit significantly from the international uptake of AI trust initiatives. They are partners firmly committed to driving digital transformation and cooperating on new technologies based on their shared democratic values, including respect for human rights [61][62]. For example, eBay (a digital platform that deployed AI) provides an unprecedented opportunity for small businesses to go global. In the US, 97% of small businesses on eBay export, compared to just 4 percent of offline peers, trade internationally, and large numbers of their customers are from the EU [63]. International application of the trust initiatives acts as a shared language of trust between these economies.

3.3 Significant barriers to leveraging opportunities from the trust label

The first barrier we have noted is the *lack of a unitary framework, standard, and structure for developing and documenting AI trust initiatives, for example, what systems could be regulated and how auditing should be done*. This can form a more critical problem for governments, institutions, researchers, and regulators wishing to adopt or assess the outcome of AI regulation and self-regulatory mechanisms. Governments that seek to fund SMEs and entrepreneurs to develop and deploy AI products and services, will need, for example, *a clear definition of what they mean by AI, what AI systems are being developed, and the risk of developing and deploying such a system.* This "uncertainty is even more problematic when governments instead aim to adopt measures imposing obligations upon those developing or deploying AI, mainly if non-compliance entails a risk of being sanctioned" [54].

The second barrier we have noted relates to the *lack of a unitary definition of stakeholders and standard stakeholder groups and personalisation of AI trust self-regulatory mechanisms* [55]. The different types of stakeholders have been the target of the trust initiatives (see Table 4). On the one hand, trust and trustworthiness can be recognised differently by different stakeholder groups. *Requirements, documentation, and explanations should, amongst others, be understandable for the individual user rather than generic*. *The aim should be to communicate "strong confidence level" and "high information value", and "personalised to the explainee"* [56].

Literature [56] shows five broad stakeholder groups; AI developers, AI managers, AI regulators, AI users or consumers, and individuals affected by AI-based decisions. On the other hand, more general AI objectives manifest differently for various stakeholder groups. As an example, developers of AI systems focus on improving the performance and functionality of the algorithm through debugging and verification of the system based on the "structured engineering approach" and informed by cause analysis instead of trial and error [57]; there are AI Regulators who need explanations to be able to test and certify the system; there are AI managers who need explanations to supervise and control the algorithm, its usage and assure its compliance; there are AI users who are interested in explainability features to understand and compare the AI system's reasoning with their own reasoning, to analyse its validity and reliability, or to determine influential factors for a specific prediction (e.g., doctors); there are individuals affected by AI-based decisions (e.g., patients) caused by AI users (e.g., doctors) or even by autonomous decisions, who may have an interest in explainability to evaluate the fairness of a given AI system and its decisions. In addition, as Meske et al. [56] highlighted, members between different and within the same groups can also have various backgrounds regarding training, experience, and demographic characteristics. This can lead to different needs for AI system labels. Thus, based on stakeholder groups' needs and in combination with their role and task-related interest in transparency, the *trust label needs to be personalised* [58], [55].

3.4 Grounds for the selected AI trust label

In addition to the selection criteria used to select the VDE AI Trust Label, there are other important grounds for selecting the AI Trust Label for this task. In Table 7, we present the summary of the rationale and provide strong grounds for the selected AI Trust Label (all based on the output of the analysis and supported by the literature). Later in Table 7, we provide three main barriers we have faced during the analysis and completion of this task, and offer recommendations to the standardisation body.

Table 7. Grounds and justification for the selected AI Trust Label

Grounds	Selected AI Trust Label	Brief reasoning
Grounded in European principles and values	✓	The selected label is grounded in European values and are based on frameworks and principles related to fundamental rights, human dignity, and data protection and privacy
Supports Europe's agenda and helps avoid regulatory uncertainty (via label standardisation effort)	✓	The selected label can be considered a common/standard approach to support Europe's agenda and help avoid regulatory uncertainty on ethical and trustworthy aspects of AI that may arise when the AI Act becomes fully implementable
Constitute an environment for voluntary third-party auditing	✓	Organisations using the label to self-certify their system are voluntarily complying with the AI Act
Provides straightforward and easy-to-understand requirements	✓	The label uses the VCIO model with the bottom-up approach that makes it easy to understand and use requirements. While, understanding the Act may not be easy.
AI Act compliance	✓	The selected AI Trust label is highly compatible with the European AI Act. However, this does not suggest that all the mandatory requirements are completely covered.
Supports the application of AI Act mandatory requirements in the context of non-high-risk AI systems	✓	The selected AI Trust label is highly compatible with the European AI Act and can be used to self-regulate non-high- risk systems. This suggests that when organisations are labelling non- high- risk systems, they are voluntarily complying with the mandatory AI Act requirements. This contributes to Europe's vision to harmonise regulation for all AI systems (low-medium-high risk systems).
Contributes to generating and improving public trust on AI (products and services) significantly, and incentivises the public to use these products and services	✓	The label can be considered a great instrument to ensure AI Act compliance, in particular for the medium and low-risk systems (voluntarily to give more confidence in delivering trustworthy AI solutions) and enable responsible competition among the AI system developers and businesses. However, even when adopted voluntarily, label requirements and the labelling process can significantly influence fostering and engineering citizens' trust as trustworthiness is better perceived and noticed by the public and consumers. This will improve the quality of the consumer's judgement regarding system reliability or trustworthiness.
Facilitates responsible competition	✓	The selected label enables responsible competition among the AI system developers and

		businesses as adopting the AI Trust Label voluntarily gives more confidence in developing and delivering trustworthy AI solutions that public/consumers are more eager to use.
Contributes to setting, harmonising and standardising AI regulation at the global level	✓	The selected label is compatible with other influential global regulatory standards for AI (Canada, Japan, USA). This can potentially lead towards harmonising and standardising AI regulation. This is very much aligned with the international organisations such as the OECD, the Council of Europe and UNESCO that are aiming to bring together a diversity of countries to work towards a consensus on trustworthy AI.
Stakeholder-informed	✓	The AI Trust Label brings together standardization bodies and industries in the label development and implementation process, therefore the label could be an effective tool to facilitate communication and interaction between different entities so that trust concerns and risks posed by AI can be communicated and addressed to improve the uptake of AI systems. We recommend regular stakeholder-engagement activities to provide regulators with the grounds to determine when to regulate a given market or technology.
Improves efficiency of organizational/ developers resources	✓	Using the label can significantly contributes to reducing the organizational resources necessary to develop and getting the AI innovation out to the market and access to finance. Before adopting the label, organizations must allocate resources to 1) understand the regulatory environment around AI systems, 2) find out what needs to be regulated, and 3) find out how to develop innovative solutions in a sufficient amount of time and enter the market before competitors. AI Act compatible label can save a lot of organisation resources and budget and help them to develop a solution faster than before.
Securing the possibility of an international acceptance and uptake and therefore, supports the international trade relations	✓	The selected label has a great potential to be used at the international level for two main reasons: 1) it is highly compatible with the AI regulations in the USA, Canada, and Japan; 2) industries are joining the AI Trust Label consortium from other countries beyond Europe. This results in supporting Europe with international trade relations.
Working with European standardis ation authorities and communities	TBC (in the new release)	The AI Trust Label consortium is working with CEN-CENEC towards standardisation of the label. *Note: At this moment, we are unable to validate this claim.*
Standard definition of stakeholders	X	Based on the barriers we faced, it is recommended that the AI trust label provides unitary definition of stakeholders and standard stakeholder groups for the trust label. Based on

		what we have learned and observed in this task, trust concerns and the skillset of stakeholders forming a group must match.
Personalis ed labels with the most communicable label visualis ation	X	Based on the barriers we faced, it is recommended that the AI trust label in its next versions provides a personalised label with requirements relevant to each stakeholder group (as formed in the item above). This includes stakeholder group-informed label visualisation and scoring methodology. Trust and trustworthiness can be recognized differently by different stakeholder groups. Requirements, documentation, and explanations should, amongst others, be understandable for the individual user rather than generic. The aim should be to communicate "strong confidence level" and "high information value", and "personalised to the explainee"
Contribute to standardis ing the framework, standard, and structure for developing and documenting AI trust initiatives	X	Based on the barriers we faced, it is recommended that the AI trust label provides a unitary framework, standard, and structure for developing and documenting AI trust initiatives.

4. Key recommendations to European Regulatory Bodies

In the following sections, we provide a list of recommendations to the European decision-makers and regulators. Recommendations are grounded in our analysis of the initiatives.

Recommendation 1: Increase global alignment on and promote an interoperable approach to AI regulation – Given the cross-border nature of the digital economy (e.g., international trade between the EU and USA to facilitate the adoption, use, and interoperability of AI technologies across the two nations), AI regulatory initiatives and frameworks should ideally operate across nations and regions. In this regard, the EU needs to support the general vision and offer its networks, resources, and expertise to reinforce formal and informal coordination among the different self-regulatory initiatives to move towards one self-regulation mechanism in Europe.

Recommendation 2: The design of the label should support the awareness of the AI Act and be aligned with the ongoing AI standards, in particular, the harmonised standards developed by CEN_CENELEC JTC21 – Self-regulatory mechanisms in general and the investigated initiatives, in particular, are effective ways to communicate the AI Act in Europe and beyond. A well-designed AI Label can support the increased awareness of the AI Act. The label should communicate the relevance of the technical requirements to users of AI products, services, and processes. Moreover, AI trust labels should be aligned with the ongoing AI standards, in particular the harmonised standards developed by CEN_CENELEC JTC21 to foster widespread adoption by users and facilitate their development by industry and innovation stakeholders and ensure that they meet relevant market and societal needs [64].

Recommendation 3: Use a suitable visualisation to maximise communication for the target user – The label should present the most relevant information to the consumers to enhance trust and confidence in using and adopting the AI system. The label visualisation may take the form of a Trust Fact Table, Trust score, or Trust Seal. The label could be designed using the progressive disclosure design pattern to initially convey a general message, with details and complexity handled in the background.

Recommendation 4: Develop a unitary framework, standard, and structure for developing and documenting AI trust initiatives and the definition of stakeholders and stakeholder groups – Different foundations, frameworks, and structures were used in all the studied labels. No two labels followed a single and standard way of documentation. This may develop some uncertainty in perceiving AI trust labels for different purposes (e .g ., research), in particular for self-regulation. This study recommends institutions to work towards a standard to develop and document AI trust labels.

Recommendation 5: Develop a unitary definition of targeted stakeholders and stakeholder groups – The different types of stakeholders have been the target of the trust initiatives. This study shows that trust and trustworthiness can be perceived differently by the various AI stakeholders. Requirements, documentation, and explanations should, amongst others, be understandable for the individual user, easy to access, role-specific rather than generic, with a conveyed strong confidence level and high information value and personalised to the explainee. In this regard, this study recommends institutions to work towards standard stakeholder groups and the definition of such groups.

Recommendation 6: Take a risk-based approach to the auditing process – This study recommends a more pragmatic approach to conformity assessment and the notion that high-risk AI systems should be required to go under thorough auditing by a trained and accredited auditor outside the system provider's organisation. We have identified auditing aspects that lead to no specific recommendations, such as who should choose the external auditor, how can self-assessment be combined with third-party auditing, whether organisations should develop or use AI tools to conduct periodic audits of their governance processes, and when should the next audit be.

PART B

Annex A. Phase 1 – Detailed Analysis for the selection of the AI Trust Label

A.1. Detailed analysis (content analysis) of the four Influential AI Regulations

The AI Act

Against the political context of the European Commission for the 2019-2024, the Commission puts forward legislation for a coordinated European approach on the human and ethical implications of AI. Based on this agenda, on 19 February 2020 the Commission published the White Paper on *AI - A European approach to excellence and trust* [28]. The White Paper sets out policy options on how to achieve the twin objective of *promoting the uptake of AI* and of *addressing the risks associated with certain uses of such technology*.

To address the second objective for the development of an ecosystem of trust, the Commission proposed a horizontal legal framework and harmonised rules for the trustworthy use of AI [5], [10] based on recommendations by an appointed high-level expert group, a preceding white paper, and public stakeholder consultation [10]. The AI Act applies to both private, public, as well as extraterritorial providers whose AI tools are used in the EU [10] and its main objective is to regulate 'high-risk' AI systems through minimum mandatory requirements (Articles 9 to 15 of the Act) to give people the confidence to embrace AI systems and solutions while encouraging companies to develop them [5].

In addition to the above overarching objective, Article 8 to 15 aim to achieve the following specific objectives: 1) to ensure that AI systems placed on the Union market and used are safe and respect existing law on fundamental rights and Union values; 2) to ensure legal certainty to facilitate investment and innovation in AI; 3) to enhance governance and effective enforcement of existing law on fundamental rights and safety requirements applicable to AI systems; and 4) to facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation [5].

High-risk AI Systems - While not all AI systems harbour potential harm for individuals, there are various applications of AI systems that may cause direct or indirect harms. The aim for AI Act is to mitigate risks caused by these systems [29].

The Act identifies two main categories of high-risk AI systems: 1) AI systems intended to be used as safety component of products that are subject to third party ex-ante conformity assessment and 2) other stand-alone AI systems with mainly fundamental rights implications that are explicitly listed in Annex III [3].

We looked at the Annex III [5] to have a sense of what AI systems are explicitly classify high-risk by the European Commission. A limited number of high-risk AI systems referred to in Article 6(2) include: Biometric identification and categorisation of natural persons, Management and operation of critical infrastructure, Education and vocational training, Employment, workers management and access to self-employment, Access to and enjoyment of essential private services and public services and benefits, Law enforcement, Migration, asylum and border control management, and Administration of justice and democratic processes. These eight AI systems will have to comply with the mandatory requirements for trustworthy AI and follow conformity assessment procedures before those systems can be placed on the Union market [5].

The Commission may expand the list of high-risk applications within certain pre-defined areas, by applying a set of criteria and risk assessment methodology [5]. AI systems can be added to the list when and if there is evidence that they pose a similar or even higher risk to the health, safety, or fundamental rights [10].

AI Act mandatory requirements for high-risk systems - Under the Act, systems considered 'high-risk' are permitted on the European market subject to compliance with mandatory requirements relating to data and data governance, documentation and recording keeping, transparency and provision of information to users, human oversight, robustness, accuracy and security, as well as ex-ante conformity assessment. These requirements are briefly presented below.

Article 9 - Risk management and testing: The known and foreseeable risks associated with the AI system should be identified, evaluated, and documented systematically within a risk management system;

Article 10 - Data and data governance: Training, validation, and testing data are subject to appropriate data governance practices. These practices concern data collection and pre-processing, data assumptions, data availability and quantity, and the examination of possible biases, among other things. All data sets must be relevant, representative, error-free, and complete. In addition, datasets should consider the characteristics of the individuals and the geographical, behavioural, or functional settings for which the AI system will be used;

Article 11 – Technical documentation: The technical documentation of a high-risk AI system shall be drawn up before that system is placed on the market or put into service and kept up-to-date;

Article 12 – Record Keeping: High-risk AI systems shall be designed and developed with capabilities to automatically record events ('logs') while the high-risk AI systems are operating;

Articles 13 and 14 - Transparency and human oversight: The AI system must be sufficiently transparent such that users can interpret and use the AI system's output appropriately. Providers must specify measures that allow a human operator to fully understand the capabilities and limitations of the AI system, interpret the AI system's output correctly; and

Article 15 - Accuracy, robustness, and cybersecurity: High-risk AI systems must achieve an appropriate level of accuracy, robustness, and cybersecurity for their intended purpose.

The AI system must be accompanied by several sources of documentation, including an instruction manual (Article 13), technical documentation (Annex IV), a risk management system (Article 9), and a quality management system (Article 17). The instruction manual specifies the AI system's capabilities and limitations. This includes its intended purpose, the expected level of accuracy, robustness, and cybersecurity, known and foreseeable circumstances that impact these performance metrics, specifications for the input data, and interpretability measures. The technical documentation describes the system components and their development in detail, including conceptional decisions. The risk management system describes the risks associated with using the system and how they are mitigated. The quality management system documents compliance with all regulations of the AI Act.

Furthermore, the Act introduces codes of conduct that can either relate to 1) non-high-risk AI systems or 2) all types of AI systems.

Detailed analysis of the rest of the regulations is presented in the table 8.

A.2 Other three international AI regulations

AI Bill of Rights of the United States, AI and Data Act of Canada (AIDA), AI Regulation of Japan

Table 8. Contextual analysis of the international AI regulations

Document	Year	Supporting Act and Regulation	Purpose	Approach	Defining AI system	Remit	Prohibited Activities	System of focus	Application Scope	Non compliance penalties
Artificial Intelligence and Data Act (AIDA) - the first federal law in Canada regulating the creation and use of AI systems and would create penalties for non-compliance.	November 2020	Bill C-27: Digital Charter Implementation Act Consumer Privacy Protection Act, Personal Information and Data Protection Tribunal Act Other Acts: Personal Information Protection and Electronic Documents Act, Consumer Privacy Protection Act	Regulate international and inter provincial trade and commerce in AI systems by establishing common requirements applicable across Canada for the design, development and use of those systems, prohibit certain conduct in relation to AI systems that may result in serious harm to	Holistic and hard-law-based	a technological system that, autonomously or partly autonomously, processes data related to human activities through the use of a genetic algorithm, a neural network, machine learning or another technique in order to generate content or make decisions, recommenda	Measures to identify, assess and mitigate the risks of harm or biased output that could result from the use of the system; Measures to monitor compliance with the mitigation measures and the effectiveness of those mitigation measures; Where the system is	Processing or use of unlawfully obtained personal information in AI system; An AI system resulting in serious physical or psychological harm or substantial damage to property; An AI system defraudin	The AIDA imposes regulatory requirements for both AI systems generally and those AI systems specifically referred to as "high-impact systems."	focused on organizations carrying out a "regulated activity," which means (a) processing or making available for use any data relating to human activities for the purpose of designing, developing or using an AI system, or (b) designing, developing or making available for use an AI system	Maximum fine of the greater of $10 million and 3 per cent of gross global revenues

			individuals or harm to their interests. The AIDA defines "harm" as (a) physical or psychological harm to an individual, (b) damage to an individual's property, or (c) economic loss to an individual		tions or predictions	made available for use or an organization is managing the operation of the system, publish on a public website a plain-language description of the system that includes prescribed content; Notify the Minister of Industry (or other designated Minister) if use of the system results or is likely to result in material harm	g the public and causing substantial economic loss		or managing its operations.	

Document	Year	Supporting Act and Regulation	Purpose	Approach	Defining AI system	Remit	Prohibited Activities	System of focus	Application Scope	Non compliance penalties
USA Bill of Right (BoR) - A Blueprint and non-binding	October 2022	Executive Order 13960 on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government, National Institute of Standards and Technology’s AI Risk Management Framework, Algorithmic Accountability Act, which was reintroduced in the Senate in an amended form earlier in 2022	The purpose of the Bill of Right is to help guide the design, use, and deployment of automated systems to protect the American Public by implement the followings: Safe and effective systems: You should be protected from unsafe or ineffective systems. Algorithmic discrimination protection: You should not face discrimination by algorithms	At the moment-> Sector-specific and soft-law-based (promote appropriate AI governance through nonbinding guidance) Later when the Algorithmic Accountability Act or similar bill is adopted in Congress -> Holistic and hard-law-based Agencies are to take a risk-based approach and	Any system, software, or process that uses computation as whole or part of a system to determine outcomes, make or aid decisions, inform policy implementation, collect data or observations , or otherwise interact with individuals and/or communities ; Any automated systems that have the potential to meaningfully impact the American	Protections that should be applied with respect to all automated systems that have the potential to meaningfully impact individuals' or communities' exercise of rights (Civil rights, civil liberties, and privacy), equal opportunities, or access to critical resources/ services; To guide the design,	Automated systems should not be designed with an intent reasonably foreseeable possibility of endangering your safety or the safety of your community; You should not face discrimination by algorithms and systems should be used and designed in an equitable way;	Apply to automated systems that have the potential to meaningfully impact the American public’s rights, opportunities, or access to critical resources or services, generally excluding many industrial and/or operational applications of AI. The Blueprint expands for use of AI in Lending, Human Resources, surveillance and other	To support the development of policies and practices that protect civil rights and promote democratic values in the building, deployment, and governance of automated systems. It depends significantly on the context in which automated systems are being utilized. Future sector-specific guidance will likely be necessary and important for guiding the use of automated systems in	NA (non-binding)

			and systems should be used and designed in an equitable way. Data privacy: you should be protected from abusive data practices via built-in protections and you should have agency over how data about you is used. Notice and explanation: you should know that an automated system is being used and understand how and why it contributes to outcomes	determine which risks are acceptable while considering potential benefits, and they think “it is not necessary to mitigate every foreseeable risk” and do not favor prescriptive regulations	public’s rights, opportunities, or access to critical resources or services;	use, and deployment of automated systems; To actualizing democratic values and principles (protect civil rights, civil liberties, and privacy) in the technological design process of AI systems	protected from abusive data practices via built-in protections and you should have agency over how data about you is used; You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you; You should be able to opt out, where	areas (which would also find a counterpart in the ‘high-risk’ use case framework of the forthcoming EU AI Act)	certain settings such as AI systems used as part of school building security or automated health diagnostic systems	

			that impact you. Alternative options: you should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter.				appropriate, and have access to a person who can quickly consider and remedy problems you encounter			
Document	**Year**	**Supporting Act and Regulation**	**Purpose**	**Approach**	**Defining AI system**	**Remit**	**Prohibited Activities**	**System of focus**	**Application Scope**	**Non compliance penalties**
Japan's AI regulatory policy (non-binding guidelines) The Japan AI Regulations is a combination of sector-specific guidelines	July 2021	AI Governance guidelines in Japan Digital Platform Transparency Act Financial Instrument and	Japan has developed and revised AI-related regulations with the goal of maximizing AI's positive impact on society, rather than suppressing it out of	Sector-specific and soft-law-based (promote appropriate AI governance through nonbinding guidance)		To realize the four Social Principles of Human-Centric AI (human dignity, diversity and inclusion, and sustainabili		No specific system - a number of existing laws are applicable to AI including the Constitution and laws pertaining to contracts, torts, certain economic	Japan has no regulations that generally constrain the use of AI. Regulations face difficulties in keeping up with the speed and complexity of AI innovation. However, The	NA

		Exchange Act The Act on the Protection of Personal Information Product Liability Act Unfair Competition Prevention Act Copyright Act Road traffic Act and Road Transport Vehicle Act Machine Learning Quality Management Guideline	overestimated risks. The emphasis is on a risk-based, agile, and multistakeholder process, rather than a one-size-fits-all obligation or prohibition. Japan took the approach of respecting companies’ voluntary governance and providing nonbinding guidelines to support it, while imposing transparency obligations on some large digital platforms. AI regulations			ty) through AI. Providing nonbinding guidance to support or guide companies’ voluntary efforts for AI governance Industries develop their own self-generated guidelines to govern their activities		statutes, intellectual property, personal data, privacy and the criminal code	operator may be held liable for tort or product liability if an accident occurs due to AI systems. “legally-binding horizontal requirements for AI systems are deemed unnecessary at the moment.” Several guidance have been published AI systems and protection of Data: Governance Guidelines for Implementation of AI Principles, Guidebook on Corporate Governance for Privacy in	

			is classified into two categories: Regulation on AI: 1) Regulations to manage risks associated with AI (risk-based approach). 2) Regulation for AI: Regulatory reform to promote the implementation of AI (soft-law approach).						Digital Transformation, Guidebook for Utilization of Camera Images, Contract Guidelines on Utilization of AI and Data Voluntary guidelines by businesses: Fujitsu published a practice guide , Sony Group AI Ethics Guidelines, NEC Group AI and Human Rights Principles Voluntary guidelines by research institutes: Machine Learning Quality Management Guideline,	

									University of Tokyo developed the Risk Chain Model to structure risk factors for AI	

A.3 Other major regulations – Requirements analysis

Table 9. Detailed analysis of the requirements of the international AI regulations

AIDA	
Requirement	**Description - Requirements**
Anonymized data	A person who carries out any regulated activity and who processes or makes available for use anonymized data in the course of that activity must, in accordance with the regulations, establish measures with respect to (a) the manner in which data is anonymized; (b) the use or management of anonymized data.
Assessment — high-impact system	A person who is responsible for an artificial intelligence system must, in accordance with the regulations, assess whether it is a high-impact system
Measures related to risks	A person who is responsible for a high-impact system must, in accordance with the regulations, establish measures to identify, assess and mitigate the risks of harm or biased output that could result from the use of the system
Monitoring of mitigation measures	A person who is responsible for a high-impact system must, in accordance with the regulations, establish measures to monitor compliance with the mitigation measures they are required to establish and the effectiveness of those mitigationmeasures
Keeping general records	A person who carries out any regulated activity must, in accordance with the regulations, keep records describing in general terms, as the case may be, (a) the measures they establish; (b) the reasons supporting their assessment
Additional records	The person must, in accordance with the regulations, keep any other records in respect of the requirements
Publication of description — making system available for use	A person who makes available for use a high-impact system must, in the time and manner that may be prescribed by regulation, publish on a publicly available website a plain-language description of the system that includes an explanation of (a) how the system is intended to be used; (b) the types of content that it is intended to generate and the decisions, recommendations or predictions that it is intended to make; (c) the mitigation measures established in respect of it; and (d) any other information that may be prescribed by regulation.

Publication of description — managing operation of system	A person who manages the operation of a high-impact system must, in the time and manner that may be prescribed by regulation, publish on a publicly available website a plain-language description of the system that includes an explanation of (a) how the system is used; (b) the types of content that it generates and the decisions, recommendations or predictions that it makes; (c) the mitigation measures established in respect of it; and (d) any other information that may be prescribed by regulation.
Notification of material harm	A person who is responsible for a high-impact system must, in accordance with the regulations and as soon as feasible, notify the Minister if the use of the system results or is likely to result in material harm
The AI Bill of Rights	
Requirement	**Description - Requirements**
Safe and effective systems	Protect the public from harm in a proactive and ongoing manner: ○ Consultation - Automated systems should be developed with consultation from diverse communities, stakeholders, and domain experts to identify concerns, risks, and potential impacts of the system. ○ Testing - Systems should undergo pre-deployment testing. This testing should follow domain-specific best practices, when available, for ensuring the technology will work in its real-world context. ○ Risk identification and mitigation - Before deployment, and in a proactive and ongoing manner, potential risks of the automated system should be identified and mitigated. Outcomes of these protective measures should include the possibility of not deploying the system or removing a system from use; ○ Ongoing monitoring: Automated systems should have ongoing monitoring procedures, including recalibration procedures, in place to ensure that their performance does not fall below an acceptable level over time, based on changing real-world conditions or deployment contexts, post-deployment modification, or unexpected conditions. ○ Clear organizational oversight – Entities responsible for the development or use of automated systems should lay out clear governance structures and procedures. This includes clearly-stated governance procedures before deploying the system, as well as responsibility of specific individuals or entities to oversee ongoing assessment and mitigation. Avoid in appropriate, low quality, and irrelevant data use: ○ Relevant and high quality data - Data used as part of any automated system's creation, evaluation, or deployment should be relevant, of high quality, and tailored to the task at hand. Relevancy should be established based on research-backed demonstration of the causal influence of the data to the specific use case or justified more generally based on a reasonable expectation of usefulness in the domain and/or for the system design or ongoing development. ○ Derived data sources tracked and reviewed carefully - Data that is derived from other data through the use of algorithms, such as data derived or inferred from prior model outputs, should be identified and tracked ○ Data refuse limits in sensitive domains - Data from some domains, including criminal justice data and data indicating adverse outcomes in domains such as finance, employment, and housing, is especially sensitive, and in some cases its reuse is limited by law. Accordingly, such data should be subject to extra oversight to ensure safety and efficacy. Demonstrate the safety and effectiveness of the system

	○ Independent evaluation - Automated systems should be designed to allow for independent evaluation (e.g., via application programming interfaces). Independent evaluators, such as researchers, journalists, ethics review boards, inspectors general, and third-party auditors, should be given access to the system and samples of associated data, in a manner consistent with privacy, security, law, or regulation (including, e.g., intellectual property law), in order to perform such evaluations. ○ Reporting - Entities responsible for the development or use of automated systems should provide regularly-updated reports
Algorithmic discrimination protections	Protect the public from algorithmic discrimination in a proactive and ongoing manner ○ Proactive assessment of equality in design - Those responsible for the development, use, or oversight of automated systems should conduct proactive equity assessments in the design phase of the technology research and development or during its acquisition to review potential input data, associated historical context, accessibility for people with disabilities, and societal goals to identify potential discrimination and effects on equity resulting from the introduction of the technology. ○ Representative and robust data - Any data used as part of system development or assessment should be representative of local communities based on the planned deployment setting and should be reviewed for bias based on the historical and societal context of the data. Such data should be sufficiently robust to identify and help to mitigate biases and potential harms. ○ Guarding against proxies - In many cases, attributes that are highly correlated with demographic features, known as proxies, can contribute to algorithmic discrimination. In cases where use of the demographic features themselves would lead to illegal algorithmic discrimination, reliance on such proxies in decision-making (such as that facilitated by an algorithm) may also be prohibited by law. Proactive testing should be performed to identify proxies by testing for correlation between demographic information and attributes in any data used as part of system design, development, or use. I ○ Ensure accessibility during design, development, and deployment - Systems should be designed, developed, and deployed by organizations in ways that ensure accessibility to people with disabilities. ○ Disparity assessment - Automated systems should be tested using a broad set of measures to assess whether the system components, both in pre-deployment testing and in-context deployment, produce disparities. ○ Disparity mitigation - When a disparity assessment identifies a disparity against an assessed group, it may be appropriate to take steps to mitigate or eliminate the disparity. ○ Ongoing monitoring and mitigation - Automated systems should be regularly monitored to assess algorithmic discrimination that might arise from unforeseen interactions of the system with inequities not accounted for during the pre-deployment testing, changes to the system after deployment, or changes to the context of use or associated data. Monitoring and disparity assessment should be performed by the entity deploying or using the automated system to examine whether the system has led to algorithmic discrimination when deployed. Demonstrate that the system protects against algorithmic discrimination

	○ Independent evaluation - entities should allow independent evaluation of potential algorithmic discrimination caused by automated systems they use or oversee. In the case of public sector uses, these independent evaluations should be made public unless law enforcement or national security restrictions prevent doing so. Care should be taken to balance individual privacy with evaluation data access needs; in many cases, policy-based and/or technological innovations and controls allow access to such data without compromising privacy ○ Reporting - Entities responsible for the development or use of automated systems should provide reporting of an appropriately designed algorithmic impact assessment
Data privacy	Protect privacy by design and by default ○ Privacy by design and by default - Automated systems should be designed and built with privacy protected by default. Privacy risks should be assessed throughout the development life cycle, including privacy risks from reidentification, and appropriate technical and policy mitigation measures should be implemented. ○ Data collection and use-case scope limits - Data collection should be limited in scope, with specific, narrow identified goals, to avoid "mission creep." Anticipated data collection should be determined to be strictly necessary to the identified goals and should be minimized as much as possible. ○ Risk identification and mitigation - Entities that collect, use, share, or store sensitive data should attempt to proactively identify harms and seek to manage them so as to avoid, mitigate, and respond appropriately to identified risks. ○ Privacy-preserving security - Entities creating, using, or governing automated systems should follow privacy and security best practices designed to ensure data and metadata do not leak beyond the specific consented use case. Best practices could include using privacy-enhancing cryptography or other types of privacy-enhancing technologies or fine-grained permissions and access control mechanisms, along with conventional system security protocols. Protect the public from unchecked surveillance ○ Highlighted oversight of surveillance - Surveillance or monitoring systems should be subject to heightened oversight that includes at a minimum assessment of potential harms during design (before deployment) and in an ongoing manner, to ensure that the American public's rights, opportunities, and access are protected. ○ Limited and proportionate surveillance - Surveillance should be avoided unless it is strictly necessary to achieve a legitimate purpose and it is proportionate to the need. Designers, developers, and deployers of surveillance systems should use the least invasive means of monitoring available and restrict monitoring to the minimum number of subjects possible. ○ Scope limits on surveillance to protect rights and democracy values - Civil liberties and civil rights must not be limited by the threat of surveillance or harassment facilitated or aided by an automated system. Surveillance systems should not be used to monitor the exercise of democratic rights, such as voting, privacy, peaceful assembly, speech, or association, in a way that limits the exercise of civil rights or civil liberties. Provide the public with mechanisms for appropriate and meaningful consent, access, and control over their data

	○ Use-specific consent - Consent practices should not allow for abusive surveillance practices. Where data collectors or automated systems seek consent, they should seek it for specific, narrow use contexts, for specific time durations, and for use by specific entities. ○ Brief and direct consent requests - When seeking consent from users short, plain language consent requests should be used so that users understand for what use contexts, time span, and entities they are providing data and metadata consent. User experience research should be performed to ensure these consent requests meet performance standards for readability and comprehension. ○ Data access and correction - People whose data is collected, used, shared, or stored by automated systems should be able to access data and metadata about themselves, know who has access to this data, and be able to correct it if necessary. ○ Consent withdrawal and data deletion - Entities should allow (to the extent legally permissible) withdrawal of data access consent, resulting in the deletion of user data, metadata, and the timely removal of their data from any systems (e.g., machine learning models) derived from that data ○ Automated system support - Entities designing, developing, and deploying automated systems should establish and maintain the capabilities that will allow individuals to use their own automated systems to help them make consent, access, and control decisions in a complex data ecosystem Demonstrate that data privacy and user control are protected ○ Independent evaluation - entities should allow independent evaluation of the claims made regarding data policies. These independent evaluations should be made public whenever possible. Care will need to be taken to balance individual privacy with evaluation data access needs. ○ Reporting - When members of the public wish to know what data about them is being used in a system, the entity responsible for the development of the system should respond quickly with a report on the data it has collected or stored about them. Such a report should be machine-readable, understandable by most users, and include, to the greatest extent allowable under law, any data and metadata about them or collected from them, when and how their data and metadata were collected, the specific ways that data or metadata are being used, who has access to their data and metadata, and what time limitations apply to these data. In cases where a user login is not available, identity verification may need to be performed before providing such a report to ensure user privacy. Additionally, summary reporting should be proactively made public with general information about how peoples' data and metadata is used, accessed, and stored. Extra protection for data related to sensitive domains (health, employment, education, criminal justice, and personal finance) See Appendix A Provide enhanced protections for data related to the sensitive domains ○ Necessary functions only – Sensitive data should only be used for functions strictly necessary for that domain or for functions that are required for administrative reasons (e.g., school attendance records), unless consent is acquired, if appropriate, and the additional expectations in this section are met. Consent for non necessary functions should be

	optional, i.e., should not be required, incentivized, or coerced in order to receive opportunities or access to services. In cases where data is provided to an entity (e.g., health insurance company) in order to facilitate payment for such a need, that data should only be used for that purpose o Ethical review and use prohibitions- Any use of sensitive data or decision process based in part on sensitive data that might limit rights, opportunities, or access, whether the decision is automated or not, should go through a thorough ethical review and monitoring, both in advance and by periodic review (e.g., via an independent ethics committee or similarly robust process). o Data quality - In sensitive domains, entities should be especially careful to maintain the quality of data to avoid adverse consequences arising from decision-making based on flawed or inaccurate data. Such care is necessary in a fragmented, complex data ecosystem and for datasets that have limited access such as for fraud prevention and law enforcement. o Limit access to sensitive data and derived data - Sensitive data and derived data should not be sold, shared, or made public as part of data brokerage or other agreements. o Reporting - In addition to the reporting on data privacy (as listed above for non-sensitive data), entities developing technologies related to a sensitive domain and those collecting, using, storing, or sharing sensitive data should, whenever appropriate, regularly provide public reports describing: any data security lapses or breaches that resulted in sensitive data leaks; the number, type, and outcomes of ethical pre-reviews undertaken; a description of any data sold, shared, or made public, and how that data was assessed to determine it did not present a sensitive data risk; and ongoing risk identification and management procedures, and any mitigation added based on these procedures. Reporting should be provided in a clear and machine-readable manner.
Notice and Explanation	Provide clear, timely, understandable, and accessible notice of the use and explanations o Generally accessible plain language documentation - The entity responsible for using the automated system should ensure that documentation describing the overall system (including any human components) is public and easy to find. The documentation should describe, in plain language, how the system works and how any automated component is used to determine an action or decision. It should also include expectations about reporting described throughout this framework, such as the algorithmic impact assessments described as part of Algorithmic Discrimination Protections. o Accountable - Notices should clearly identify the entity responsible for designing each component of the system and the entity using it o Timely and up-to-date - Users should receive notice of the use of automated systems in advance of using or while being impacted by the technology. An explanation should be available with the decision itself, or soon thereafter. Notice should be kept up-to-date and people impacted by the system should be notified of use case or key functionality changes. o Brief and clear - Notices and explanations should be assessed, such as by research on users' experiences, including user testing, to ensure that the people using or impacted by the automated system are able to easily find notices and explanations, read them quickly, and understand and act on them.

	Provide explanations as to how and why a decision was made or an action was taken by an automated system ○ Tailored to the purpose - Explanations should be tailored to the specific purpose for which the user is expected to use the explanation, and should clearly state that purpose. An informational explanation might differ from an explanation provided to allow for the possibility of recourse, an appeal, or one provided in the context of a dispute or contestation process. ○ Tailored to the target of the explanation - Explanations should be targeted to specific audiences and clearly state that audience. An explanation provided to the subject of a decision might differ from one provided to an advocate, or to a domain expert or decision maker. Tailoring should be assessed (e.g., via user experience research). ○ Tailored to the level of risk – An assessment should be done to determine the level of risk of the automated system. In settings where the consequences are high as determined by a risk assessment, or extensive oversight is expected (e.g., in criminal justice or some public sector settings), explanatory mechanisms should be built into the system design so that the system's full behavior can be explained in advance (i.e., only fully transparent models should be used), rather than as an after-the-decision interpretation. In other settings, the extent of explanation provided should be tailored to the risk level. ○ Valid - The explanation provided by a system should accurately reflect the factors and the influences that led to a particular decision, and should be meaningful for the particular customization based on purpose, target, and level of risk. Demonstrate protections for notice and explanation ○ Reporting - Summary reporting should document the determinations made based on the above considerations, including: the responsible entities for accountability purposes; the goal and use cases for the system, identified users, and impacted populations; the assessment of notice clarity and timeliness; the assessment of the explanation's validity and accessibility; the assessment of the level of risk; and the account and assessment of how explanations are tailored, including to the purpose, the recipient of the explanation, and the level of risk. Individualized profile information should be made readily available to the greatest extent possible that includes explanations for any system impacts or inferences. Reporting should be provided in a clear plain language and machine-readable manner.
Human Alternatives, consideration, and feedback	Provide a mechanism to conveniently opt from automated systems in favour of a human alternative, where appropriate ○ Brief, clear, accessible notice and instructions – Those impacted by an automated system should be given a brief, clear notice that they are entitled to opt-out, along with clear instructions for how to opt-out. Instructions should be provided in an accessible form and should be easily findable by those impacted by the automated system. ○ Human alternatives provided when appropriate - In many scenarios, there is a reasonable expectation of human involvement in attaining rights, opportunities, or access. When automated systems make up part of the attainment process, alternative timely human-driven processes should be provided. The use of a human alternative should be triggered by an opt-out process. ○ Timely and not burdensome human alternative - Opting out should be timely and not unreasonably burdensome in both the process of requesting to opt-out and the human-driven alternative provided.

	Provide timely human consideration and remedy by a fallback and escalation system in the events that an automated system fails, produces errors, or you would like to appeal or contest its impacts on you ○ Proportionate – The availability of human consideration and fallback, along with associated training and safeguards against human bias, should be proportionate to the potential of the automated system to meaningfully impact rights, opportunities, or access. ○ Accessible – Mechanisms for human consideration and fallback, whether in-person, on paper, by phone, or otherwise provided, should be easy to find and use. ○ Convenient – Mechanisms for human consideration and fallback should not be unreasonably burdensome as compared to the automated system's equivalent. ○ Equitable – Consideration should be given to ensuring outcomes of the fallback and escalation system are equitable when compared to those of the automated system and such that the fallback and escalation system provides equitable access to underserved communities ○ Timely – Human consideration and fallback are only useful if they are conducted and concluded in a timely manner. The determination of what is timely should be made relative to the specific automated system, and the review system should be staffed and regularly assessed to ensure it is providing timely consideration and fallback. ○ Effective – The organizational structure surrounding processes for consideration and fallback should be designed so that if the human decision-maker charged with reassessing a decision determines that it should be overruled, the new decision will be effectively enacted. ○ Maintained – The human consideration and fallback process and any associated automated processes should be maintained and supported as long as the relevant automated system continues to be in use Institute training, assessment, and oversight to oversight to combat automation bias and ensure any human-based components of a system are effective ○ Training and assessment – Anyone administering, interacting with, or interpreting the outputs of an automated system should receive training in that system, including how to properly interpret outputs of a system in light of its intended purpose and in how to mitigate the effects of automation bias. ○ Oversight – Human-based systems have the potential for bias, including automation bias, as well as other concerns that may limit their effectiveness. The results of assessments of the efficacy and potential bias of such human-based systems should be overseen by governance structures that have the potential to update the operation of the human-based system in order to mitigate these effects. Implement additional human oversight and safeguard for automated systems related to sensitive domains ○ Narrowly scoped data and inferences – Human oversight should ensure that automated systems in sensitive domains are narrowly scoped to address a defined goal, justifying each included data item or attribute as relevant to the specific use case. ○ Tailored to the situation – Human oversight should ensure that automated systems in sensitive domains are tailored to the specific use case and real-world deployment scenario, and evaluation testing should show that the system is safe and effective for that specific situation

	○ Human consideration before any high-risk decision – Automated systems, where they are used in sensitive domains, may play a role in directly providing information or otherwise providing positive outcomes to impacted people. However, automated systems should not be allowed to directly intervene in high-risk situations, such as sentencing decisions or medical care, without human consideration. ○ Meaningful access to examine the system - Designers, developers, and deployers of automated systems should consider limited waivers of confidentiality (including those related to trade secrets) where necessary in order to provide meaningful oversight of systems used in sensitive domains, incorporating measures to protect intellectual property and trade secrets from unwarranted disclosure as appropriate. Demonstrate access to human alternatives, consideration, and fallback ○ Reporting - Reporting should include an assessment of timeliness and the extent of additional burden for human alternatives, aggregate statistics about who chooses the human alternative, along with the results of the assessment about brevity, clarity, and accessibility of notice and opt-out instructions. Reporting on the accessibility, timeliness, and effectiveness of human consideration and fallback should be made public at regular intervals for as long as the system is in use.
AI Regulation in Japan	
Requirement	**Description - Requirements**
Ownership/intellectual property rights regarding AI	Learning Stage ○ Raw data ○ Training Data ○ Program of learning ○ Learned model ○ Learned parameters ○ Inference program Usage Stage ○ AI product is in the presence of creative contribution or creative intent by humans -> The AI user who has made the creative contribution (e.g., using a digital camera as a tool to produce a photograph as a work or the input to the system) is basically recognized as the right holder of the AI product under the Copyright Act and the Patent Act. This applies to the training data, AI program ○ AI product is in the absence of creative contribution by or creative intent of human -> In such a case the AI product should not be regarded as a work or an invention and should not be protected under the Copyright Act and Patent Act. ○ Misleading AI-created content -> the right in and to an AI product vary greatly depending on whether human creative contribution is admitted in the AI product production process.

Competition	○ Algorithm cartels - > The cartel activity using algorithms can be dealt with under current antitrust laws in many cases, it is necessary to contribute to monitor changes in technology, trends in their use, and related cases for autonomous machine. ○ Data aggregation and anti-competition effect -> When analysing the anti-competition effect resulting from the aggregation of data, certain factors must be taken into consideration. Factors are 1) whether there is an alternative method to obtain such data; 2) economic analysis on the usage of data, 3) correlation with AI. ○ Enforcement against digital-related vertical restrains -> Carefully watching the digital platforms in Japan for horizontal restrictions (i.e. cartels) and vertical restrictions (i.e. abuse of superior bargaining position).
Data protection under the Act on Protection of Personal Information	○ Collection of personal data ->consent from the data subject is not required upon collection of the personal data from such data subjects (except from sensitive personal data). However, the purpose of use must either be disclosed or notified to the data subject prior to collection and proper collection of personal data is required. ○ Use of personal data -> The use of personal data by the business is limited to the purpose of use disclosed or notified to the data subject. ○ Transfer of personal data -> Under the Act on Protection of Personal Information, if a business is transferring personal data to a third party, such business must obtain the prior consent of the data subject, unless such transfer falls under exception specified under the Act.
Regulation/Government intervention (covers regulations with respect to AI, big data, and deep learning)	○ Special law on automated driving -> Japan aims for Level 4 automated driving on express highways for private cars by 2025. The Road Transport Vehicle Act and the Road Traffic Act ate the two most relevant regulations supporting the use of AI in automated driving. ○ Special laws on AI development and utilization of data -> In Japan laws have been enacted and amended to further promote AI development and data utilization. The Act on Anonymously Processed Medical Information to Contribute to Research and Development in Medical Field came into force in May 2018. Universities and research institutes can utilize patients' medical information in a more flexible manner. The Telecommunication Business Act effective April 2022 is to place cyber security measures on IoT devices. The Platform Transparency Act becomes effective on Feb 1 2021. ○ Guidelines for AI -> The government is publishing various guidelines to facilitate the utilization of AI technology and big data. Guides on the Contract (Contract guidelines for AI and Data), Government guidelines for Implementing AI Principles for AI businesses.
Civil liability	○ AI and civil liability ○ Liability of AI users ○ Liability of AI manufacturers

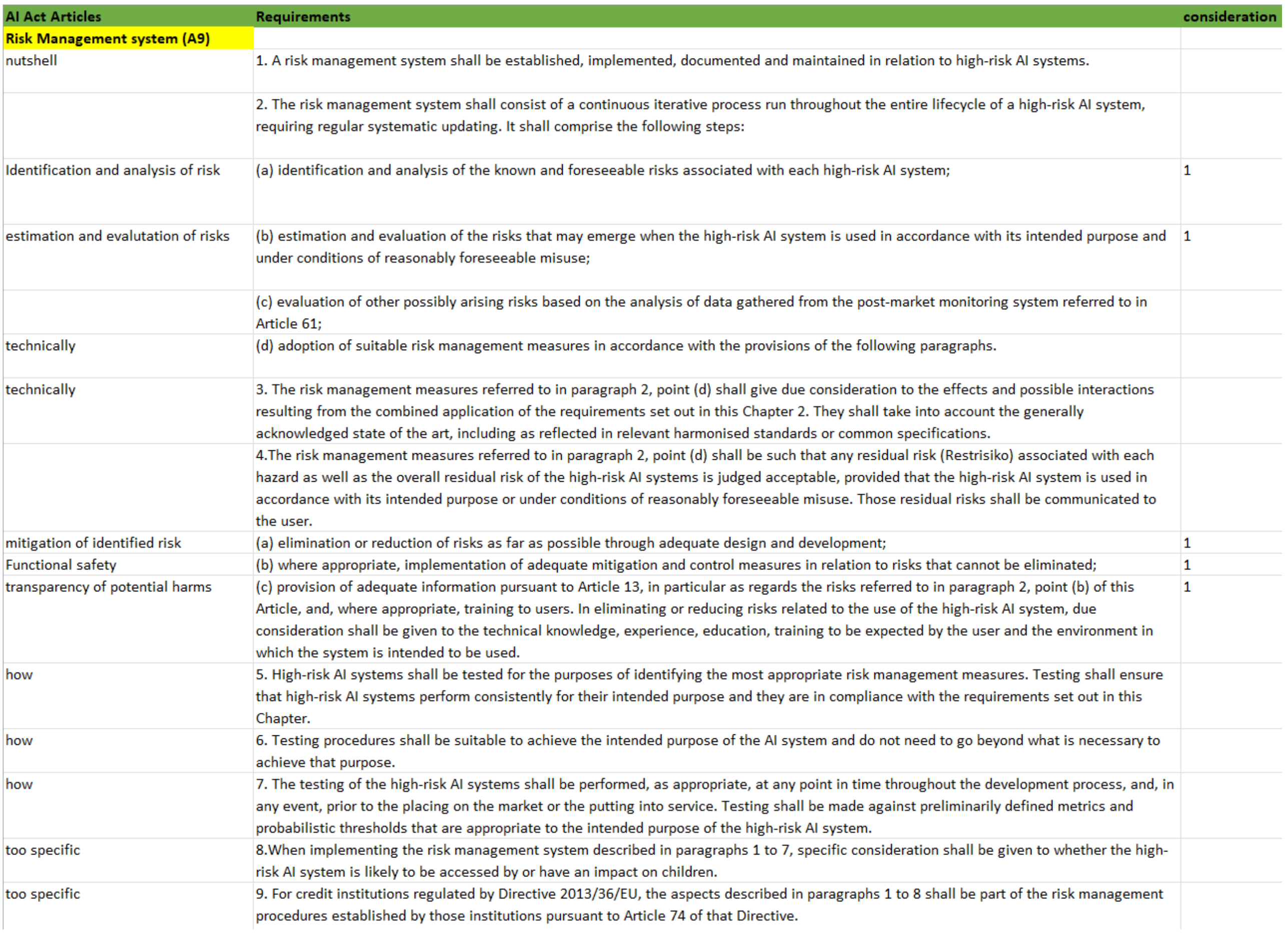
A.4 Analysis of the AI Act mandatory requirements and development of an AI Act-informed analytical framework

The basis for our analysis in this phase is the AI Act and its mandatory requirements. This is to ensure that the label we select is highly compatible with the AI Act requirements. Therefore, we briefly analysed the mandatory requirements from the AI Act and use the analysis to develop an analytical framework to assist us with the analysis of the trust initiatives.

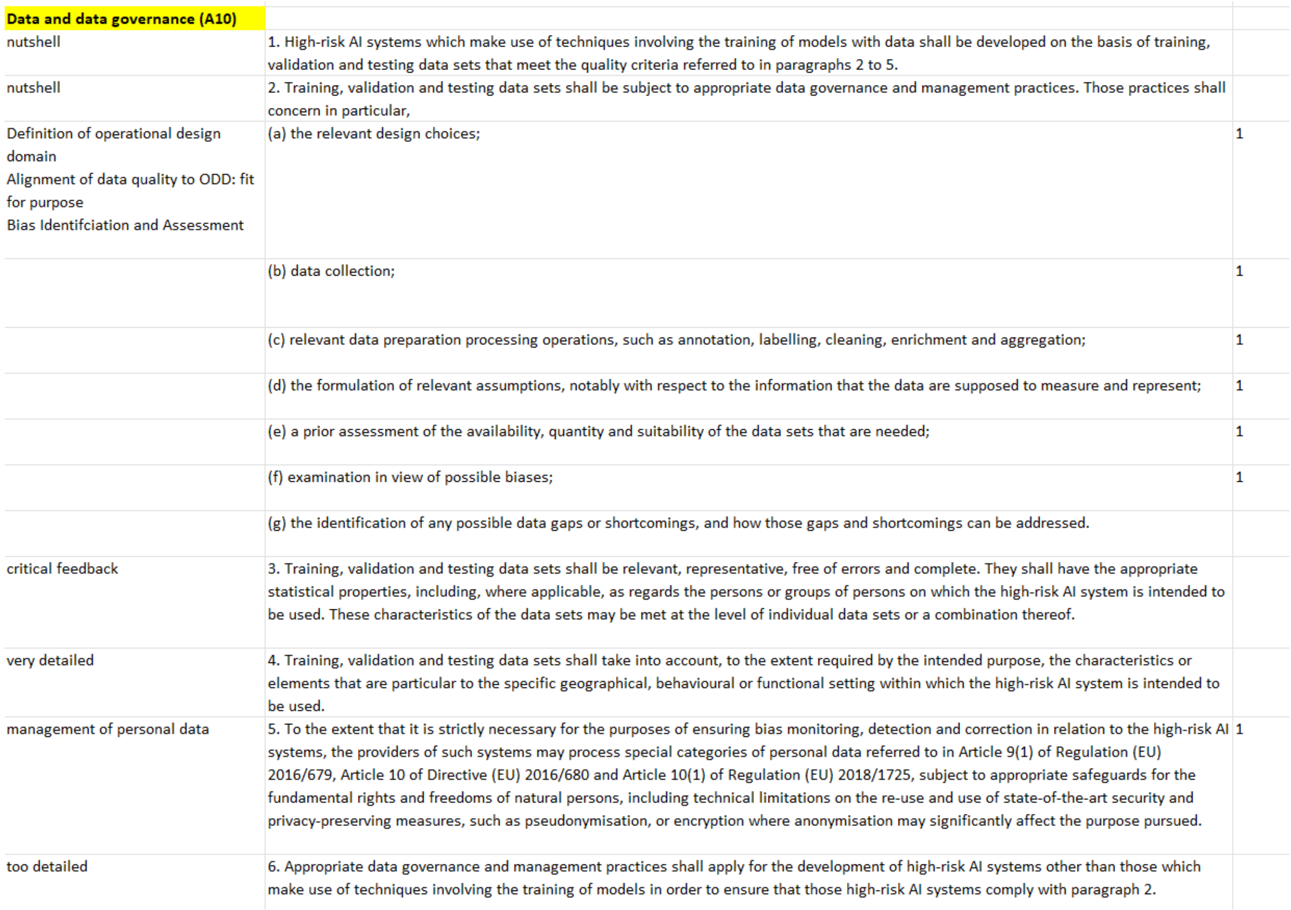
Brief content analysis of the AI Act mandatory requirements – Figure 12 presents the screenshot of the analysis excel sheet. After careful analysis of the requirements by two partners in the task, we found that some of the requirements are either too specific/detailed or too broad, some are in nutshell and some are statements. What we were looking was the list of specific requirements that are definable and measurable. The last column of the sheet, we identified these items by stating 1 and left the rest in blank. These have formed our analytical framework for analysis of the trust initiatives in the context of the AI Act.

To test the emerged framework, we used two initiatives 1. The VDE AI Trust Label, and 2. The World Economic Forum trust initiative. The aim of this testing was to find out whether or not the elements of the framework is capable of analysing these labels.

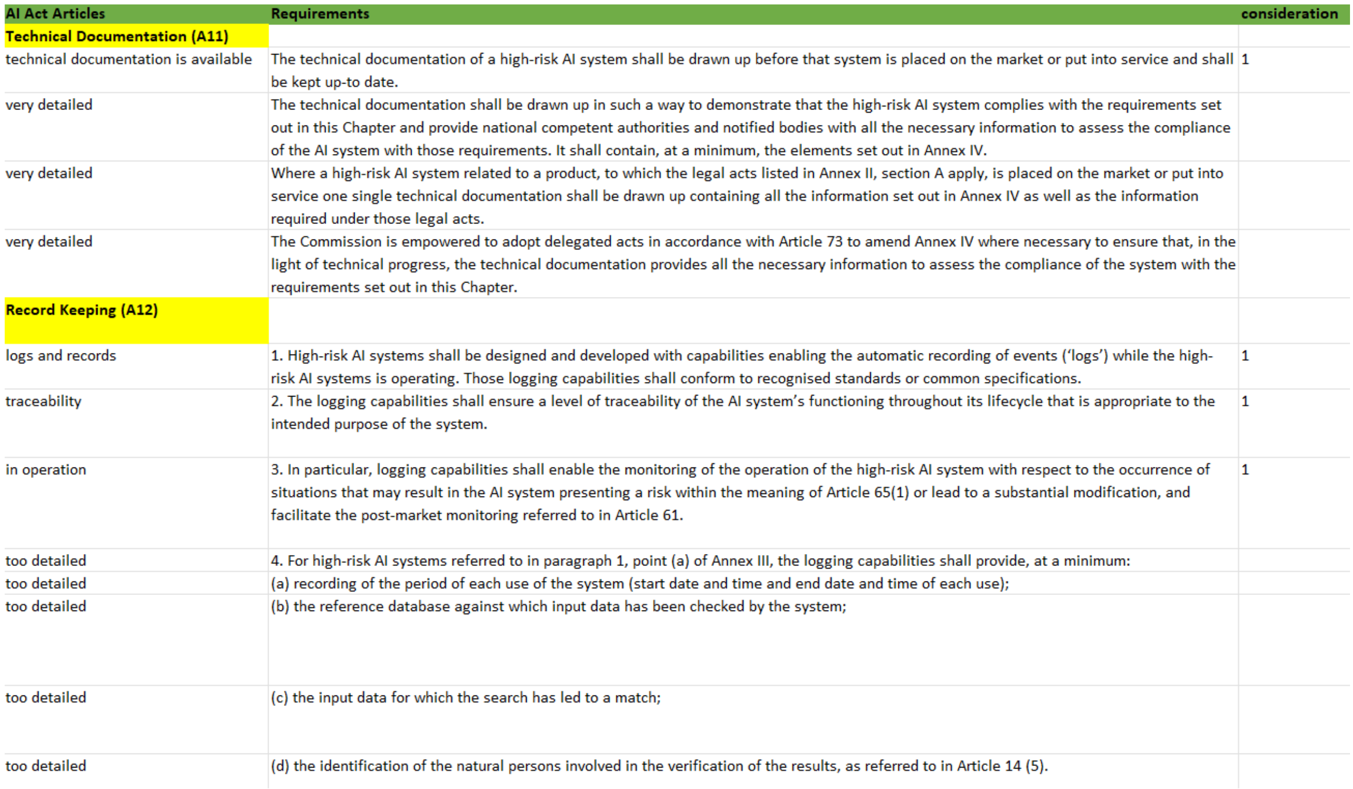
The analytical framework emerged from this is presented in Figure 13.



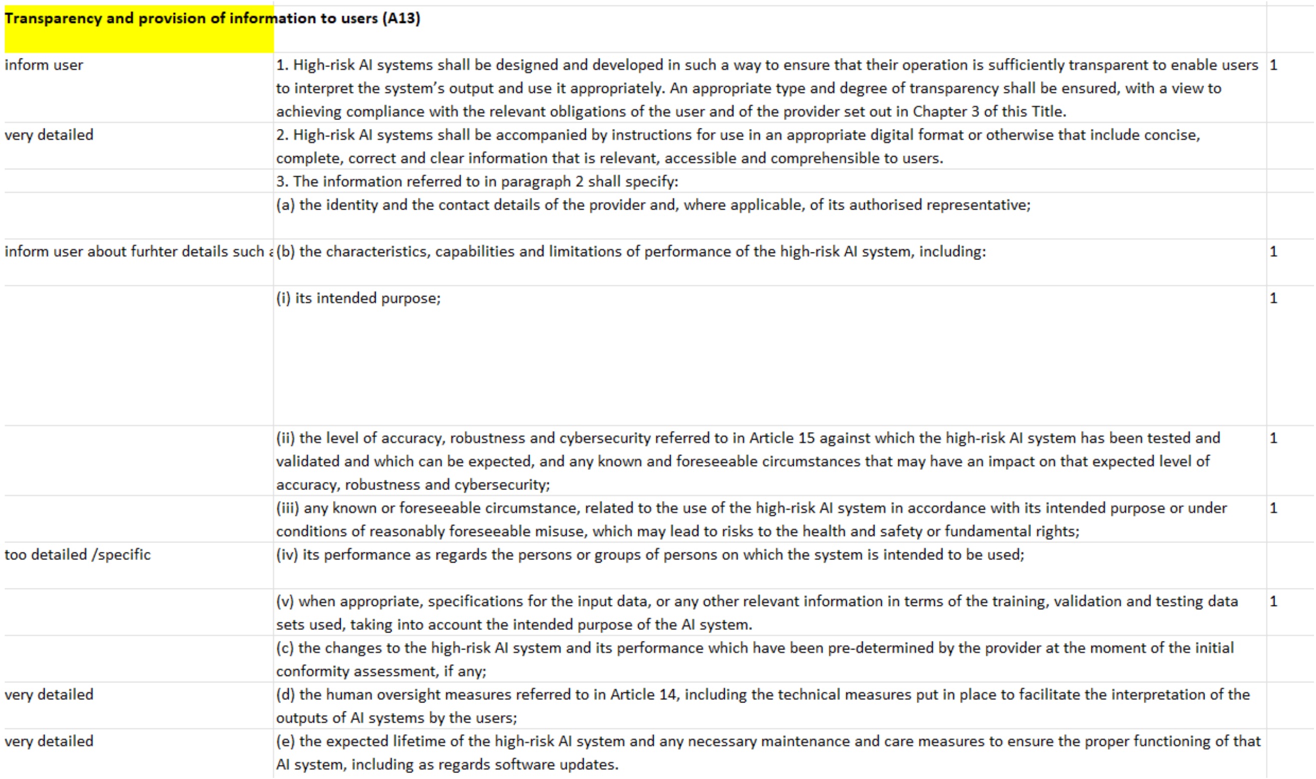
AI Act Articles	Requirements	consideration
Risk Management system (A9)		
nutshell	1. A risk management system shall be established, implemented, documented and maintained in relation to high-risk AI systems.	
	2. The risk management system shall consist of a continuous iterative process run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic updating. It shall comprise the following steps:	
Identification and analysis of risk	(a) identification and analysis of the known and foreseeable risks associated with each high-risk AI system;	1
estimation and evalutation of risks	(b) estimation and evaluation of the risks that may emerge when the high-risk AI system is used in accordance with its intended purpose and under conditions of reasonably foreseeable misuse;	1
	(c) evaluation of other possibly arising risks based on the analysis of data gathered from the post-market monitoring system referred to in Article 61;	
technically	(d) adoption of suitable risk management measures in accordance with the provisions of the following paragraphs.	
technically	3. The risk management measures referred to in paragraph 2, point (d) shall give due consideration to the effects and possible interactions resulting from the combined application of the requirements set out in this Chapter 2. They shall take into account the generally acknowledged state of the art, including as reflected in relevant harmonised standards or common specifications.	
	4.The risk management measures referred to in paragraph 2, point (d) shall be such that any residual risk (Restrisiko) associated with each hazard as well as the overall residual risk of the high-risk AI systems is judged acceptable, provided that the high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse. Those residual risks shall be communicated to the user.	
mitigation of identified risk	(a) elimination or reduction of risks as far as possible through adequate design and development;	1
Functional safety	(b) where appropriate, implementation of adequate mitigation and control measures in relation to risks that cannot be eliminated;	1
transparency of potential harms	(c) provision of adequate information pursuant to Article 13, in particular as regards the risks referred to in paragraph 2, point (b) of this Article, and, where appropriate, training to users. In eliminating or reducing risks related to the use of the high-risk AI system, due consideration shall be given to the technical knowledge, experience, education, training to be expected by the user and the environment in which the system is intended to be used.	1
how	5. High-risk AI systems shall be tested for the purposes of identifying the most appropriate risk management measures. Testing shall ensure that high-risk AI systems perform consistently for their intended purpose and they are in compliance with the requirements set out in this Chapter.	
how	6. Testing procedures shall be suitable to achieve the intended purpose of the AI system and do not need to go beyond what is necessary to achieve that purpose.	
how	7. The testing of the high-risk AI systems shall be performed, as appropriate, at any point in time throughout the development process, and, in any event, prior to the placing on the market or the putting into service. Testing shall be made against preliminarily defined metrics and probabilistic thresholds that are appropriate to the intended purpose of the high-risk AI system.	
too specific	8.When implementing the risk management system described in paragraphs 1 to 7, specific consideration shall be given to whether the high-risk AI system is likely to be accessed by or have an impact on children.	
too specific	9. For credit institutions regulated by Directive 2013/36/EU, the aspects described in paragraphs 1 to 8 shall be part of the risk management procedures established by those institutions pursuant to Article 74 of that Directive.	



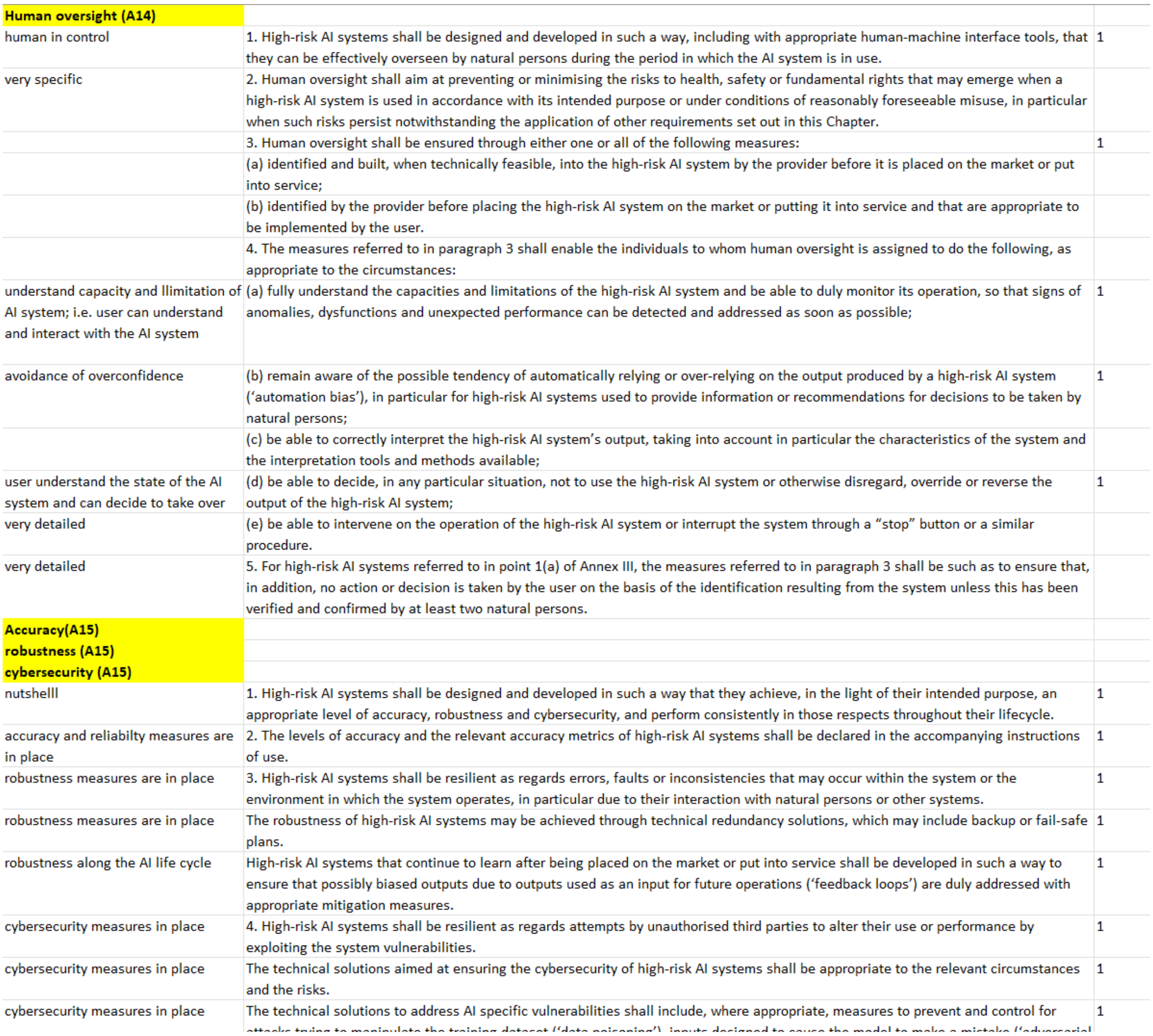
Data and data governance (A10)		
nutshell	1. High-risk AI systems which make use of techniques involving the training of models with data shall be developed on the basis of training, validation and testing data sets that meet the quality criteria referred to in paragraphs 2 to 5.	
nutshell	2. Training, validation and testing data sets shall be subject to appropriate data governance and management practices. Those practices shall concern in particular,	
Definition of operational design domain Alignment of data quality to ODD: fit for purpose Bias Identfciation and Assessment	(a) the relevant design choices;	1
	(b) data collection;	1
	(c) relevant data preparation processing operations, such as annotation, labelling, cleaning, enrichment and aggregation;	1
	(d) the formulation of relevant assumptions, notably with respect to the information that the data are supposed to measure and represent;	1
	(e) a prior assessment of the availability, quantity and suitability of the data sets that are needed;	1
	(f) examination in view of possible biases;	1
	(g) the identification of any possible data gaps or shortcomings, and how those gaps and shortcomings can be addressed.	
critical feedback	3. Training, validation and testing data sets shall be relevant, representative, free of errors and complete. They shall have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons on which the high-risk AI system is intended to be used. These characteristics of the data sets may be met at the level of individual data sets or a combination thereof.	
very detailed	4. Training, validation and testing data sets shall take into account, to the extent required by the intended purpose, the characteristics or elements that are particular to the specific geographical, behavioural or functional setting within which the high-risk AI system is intended to be used.	
management of personal data	5. To the extent that it is strictly necessary for the purposes of ensuring bias monitoring, detection and correction in relation to the high-risk AI systems, the providers of such systems may process special categories of personal data referred to in Article 9(1) of Regulation (EU) 2016/679, Article 10 of Directive (EU) 2016/680 and Article 10(1) of Regulation (EU) 2018/1725, subject to appropriate safeguards for the fundamental rights and freedoms of natural persons, including technical limitations on the re-use and use of state-of-the-art security and privacy-preserving measures, such as pseudonymisation, or encryption where anonymisation may significantly affect the purpose pursued.	1
too detailed	6. Appropriate data governance and management practices shall apply for the development of high-risk AI systems other than those which make use of techniques involving the training of models in order to ensure that those high-risk AI systems comply with paragraph 2.	



AI Act Articles	Requirements	consideration
Technical Documentation (A11)		
technical documentation is available	The technical documentation of a high-risk AI system shall be drawn up before that system is placed on the market or put into service and shall be kept up-to date.	1
very detailed	The technical documentation shall be drawn up in such a way to demonstrate that the high-risk AI system complies with the requirements set out in this Chapter and provide national competent authorities and notified bodies with all the necessary information to assess the compliance of the AI system with those requirements. It shall contain, at a minimum, the elements set out in Annex IV.	
very detailed	Where a high-risk AI system related to a product, to which the legal acts listed in Annex II, section A apply, is placed on the market or put into service one single technical documentation shall be drawn up containing all the information set out in Annex IV as well as the information required under those legal acts.	
very detailed	The Commission is empowered to adopt delegated acts in accordance with Article 73 to amend Annex IV where necessary to ensure that, in the light of technical progress, the technical documentation provides all the necessary information to assess the compliance of the system with the requirements set out in this Chapter.	
Record Keeping (A12)		
logs and records	1. High-risk AI systems shall be designed and developed with capabilities enabling the automatic recording of events ('logs') while the high-risk AI systems is operating. Those logging capabilities shall conform to recognised standards or common specifications.	1
traceability	2. The logging capabilities shall ensure a level of traceability of the AI system's functioning throughout its lifecycle that is appropriate to the intended purpose of the system.	1
in operation	3. In particular, logging capabilities shall enable the monitoring of the operation of the high-risk AI system with respect to the occurrence of situations that may result in the AI system presenting a risk within the meaning of Article 65(1) or lead to a substantial modification, and facilitate the post-market monitoring referred to in Article 61.	1
too detailed	4. For high-risk AI systems referred to in paragraph 1, point (a) of Annex III, the logging capabilities shall provide, at a minimum:	
too detailed	(a) recording of the period of each use of the system (start date and time and end date and time of each use);	
too detailed	(b) the reference database against which input data has been checked by the system;	
too detailed	(c) the input data for which the search has led to a match;	
too detailed	(d) the identification of the natural persons involved in the verification of the results, as referred to in Article 14 (5).	



Transparency and provision of information to users (A13)		
inform user	1. High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured, with a view to achieving compliance with the relevant obligations of the user and of the provider set out in Chapter 3 of this Title.	1
very detailed	2. High-risk AI systems shall be accompanied by instructions for use in an appropriate digital format or otherwise that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to users.	
	3. The information referred to in paragraph 2 shall specify:	
	(a) the identity and the contact details of the provider and, where applicable, of its authorised representative;	
inform user about furhter details such a	(b) the characteristics, capabilities and limitations of performance of the high-risk AI system, including:	1
	(i) its intended purpose;	1
	(ii) the level of accuracy, robustness and cybersecurity referred to in Article 15 against which the high-risk AI system has been tested and validated and which can be expected, and any known and foreseeable circumstances that may have an impact on that expected level of accuracy, robustness and cybersecurity;	1
	(iii) any known or foreseeable circumstance, related to the use of the high-risk AI system in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, which may lead to risks to the health and safety or fundamental rights;	1
too detailed /specific	(iv) its performance as regards the persons or groups of persons on which the system is intended to be used;	
	(v) when appropriate, specifications for the input data, or any other relevant information in terms of the training, validation and testing data sets used, taking into account the intended purpose of the AI system.	1
	(c) the changes to the high-risk AI system and its performance which have been pre-determined by the provider at the moment of the initial conformity assessment, if any;	
very detailed	(d) the human oversight measures referred to in Article 14, including the technical measures put in place to facilitate the interpretation of the outputs of AI systems by the users;	
very detailed	(e) the expected lifetime of the high-risk AI system and any necessary maintenance and care measures to ensure the proper functioning of that AI system, including as regards software updates.	



Human oversight (A14)		
human in control	1. High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use.	1
very specific	2. Human oversight shall aim at preventing or minimising the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, in particular when such risks persist notwithstanding the application of other requirements set out in this Chapter.	
	3. Human oversight shall be ensured through either one or all of the following measures:	1
	(a) identified and built, when technically feasible, into the high-risk AI system by the provider before it is placed on the market or put into service;	
	(b) identified by the provider before placing the high-risk AI system on the market or putting it into service and that are appropriate to be implemented by the user.	
	4. The measures referred to in paragraph 3 shall enable the individuals to whom human oversight is assigned to do the following, as appropriate to the circumstances:	
understand capacity and llimitation of AI system; i.e. user can understand and interact with the AI system	(a) fully understand the capacities and limitations of the high-risk AI system and be able to duly monitor its operation, so that signs of anomalies, dysfunctions and unexpected performance can be detected and addressed as soon as possible;	1
avoidance of overconfidence	(b) remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system ('automation bias'), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;	1
	(c) be able to correctly interpret the high-risk AI system's output, taking into account in particular the characteristics of the system and the interpretation tools and methods available;	
user understand the state of the AI system and can decide to take over	(d) be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system;	1
very detailed	(e) be able to intervene on the operation of the high-risk AI system or interrupt the system through a "stop" button or a similar procedure.	
very detailed	5. For high-risk AI systems referred to in point 1(a) of Annex III, the measures referred to in paragraph 3 shall be such as to ensure that, in addition, no action or decision is taken by the user on the basis of the identification resulting from the system unless this has been verified and confirmed by at least two natural persons.	
Accuracy(A15)		
robustness (A15)		
cybersecurity (A15)		
nutshelll	1. High-risk AI systems shall be designed and developed in such a way that they achieve, in the light of their intended purpose, an appropriate level of accuracy, robustness and cybersecurity, and perform consistently in those respects throughout their lifecycle.	1
accuracy and reliabilty measures are in place	2. The levels of accuracy and the relevant accuracy metrics of high-risk AI systems shall be declared in the accompanying instructions of use.	1
robustness measures are in place	3. High-risk AI systems shall be resilient as regards errors, faults or inconsistencies that may occur within the system or the environment in which the system operates, in particular due to their interaction with natural persons or other systems.	1
robustness measures are in place	The robustness of high-risk AI systems may be achieved through technical redundancy solutions, which may include backup or fail-safe plans.	1
robustness along the AI life cycle	High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way to ensure that possibly biased outputs due to outputs used as an input for future operations ('feedback loops') are duly addressed with appropriate mitigation measures.	1
cybersecurity measures in place	4. High-risk AI systems shall be resilient as regards attempts by unauthorised third parties to alter their use or performance by exploiting the system vulnerabilities.	1
cybersecurity measures in place	The technical solutions aimed at ensuring the cybersecurity of high-risk AI systems shall be appropriate to the relevant circumstances and the risks.	1
cybersecurity measures in place	The technical solutions to address AI specific vulnerabilities shall include, where appropriate, measures to prevent and control for attacks trying to manipulate the training dataset ('data poisoning'), inputs designed to cause the model to make a mistake ('adversarial	1

Figure 12. AI Act Article 9 to 15 brief content analysis

AI Act-informed Analytical framework is shown in Figure 13.

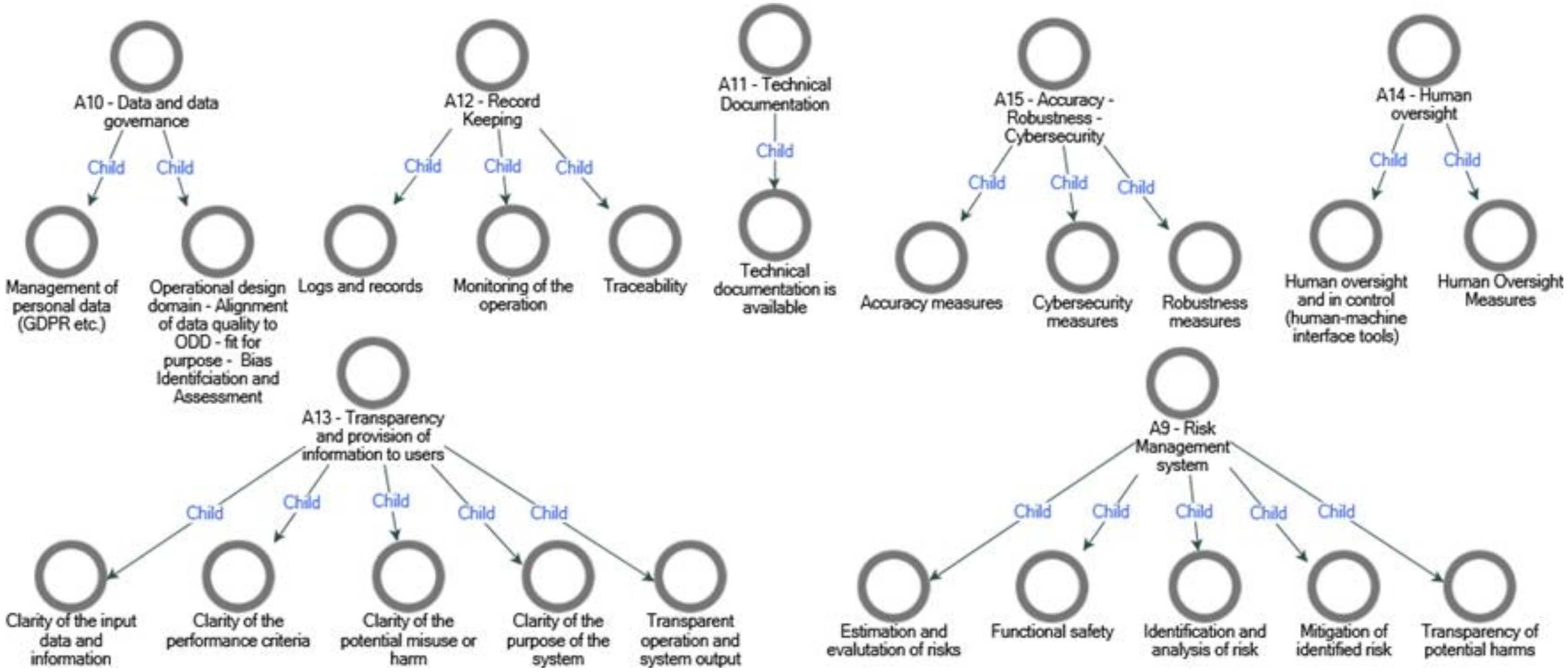

A10 - Data and data governance
Child
Child
Management of personal data (GDPR etc.)
Operational design domain - Alignment of data quality to ODD - fit for purpose - Bias Identifciation and Assessment
A12 - Record Keeping
Child
Child
Child
Logs and records
Monitoring of the operation
Traceability
A11 - Technical Documentation
Child
Technical documentation is available
A15 - Accuracy - Robustness - Cybersecurity
Child
Child
Child
Accuracy measures
Cybersecurity measures
Robustness measures
A14 - Human oversight
Child
Child
Human oversight and in control (human-machine interface tools)
Human Oversight Measures
A13 - Transparency and provision of information to users
Child
Child
Child
Child
Child
Clarity of the input data and information
Clarity of the performance criteria
Clarity of the potential misuse or harm
Clarity of the purpose of the system
Transparent operation and system output
A9 - Risk Management system
Child
Child
Child
Child
Child
Estimation and evalutation of risks
Functional safety
Identification and analysis of risk
Mitigation of identified risk
Transparency of potential harms

Figure 13. Analytical framework informed by the AI Act

A.5 Descriptions of the short-listed AI Trust Labels

Table 10. Description of the shortlisted AI Trust self-regulatory initiatives

Name	Type	Sector	Scope	Short Description
The AI Trust Label (VDE)	Label	Generic	Germany	Inspired by the EU energy-efficiency label. It shows a rating of an AI system's ethical characteristics based on six ethical values.
Mandatory Labelling Scheme	Label	Generic	Germany	The German Data Ethics Commission recommended the introduction of a mandatory labelling scheme for algorithmic systems of enhanced criticality, with the view that this would oblige operators to make it clear whether, when and to what extent algorithmic systems are being used.
Open Ethics Label	Label	Data & decision technology	Europe	The Open Ethics label aims to strengthen users' trust in AI systems by encouraging and supporting AI transparency. For the consumer, this label provides information to enable better decision-making; for software developers, the label is a type of disclosure tool to provide information about their product.
Certification System for AI Applications (Fraunhofer Institute)	Certificate and quality mark	Generic	Germany	The AI certification (Fraunhofer Institute) consists of a certification system and quality mark to signal the technical reliability of an AI system and responsible usage from an ethical and legal perspective. Furthermore, it aims to facilitate comparison between different products and help promote open competition in AI.
Z-Inspection	Audit process	Generic	Europe	Z-Inspection is an audit process that assesses whether an AI system is trustworthy. The process is based on applied ethics and uses the definition of trustworthy AI put forward by the European Commission's AI HLEG. The process is designed to be applied to a variety of sectors in which AI systems could be used, such as business, healthcare and the public sector
Swiss Digital Trust Label	Trust mark	Generic	Swiss	Denotes the trustworthiness of a digital service in clear, visual, and plain, non-technical language for consumers.
Malta's National AI Certification Framework	Certificate and audit	AI sector	Malta	The certification aims to build trust and transparency for key by providing valuable information about AI in their marketplace to signal that their AI systems have been developed ethically, transparently and in a socially responsible manner.
EU Certification for 'Trusted AI' Products	Certificate	Generic	EU	Certification for trustworthy AI applications, where products are tested for resilience, safety and absence of prejudice, discrimination, or bias.
Responsible Artificial Intelligence Institute Certification Beta (RAII)	Certificate	Generic	Global	The RAI Certification Beta is an independent certification programme for the responsible and trusted use of AI systems. The certification aims to increase trust among end users by signalling that the AI system was built following specific standards.

Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)	Certificate	Generic	Global	ECPAIS consists of a certification system that aims to signal to stakeholders in different sectors whether an AI system is safe, ethical, and trustworthy. Ultimately the certification aims to promote responsible innovation in AI systems.
Certificate of Fairness for AI Systems [46]	Certificate	Generic	United Kingdom	A certificate of fairness for AI systems alongside a kite mark type scheme to display it, with criteria to be defined at the industry level. The certification helps build an AI that avoids algorithms discriminating against women and ethnic minorities.

A.6 *Detailed analysis of the short-listed AI Trust initiatives*

Here is the analysis of the short-listed initiatives according to *Foundations, Functions, Target Users, Stages of Development, Transparency Mechanisms, and Audit Structure*.

Foundation – Some initiatives, including the AI ethics label, Mandatory Labelling, and Swiss Digital Trust Label, have adopted the most cited general ethical frameworks, values, and principles. Other initiatives, such as the Certificate of Fairness for AI Systems, are based on existing frameworks such as Data Ethics Framework and Trustworthy AI Framework [34]. Yet, other initiatives, including the Certificate of Fairness for AI systems and The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS), focus on algorithmic discrimination [35]. We also observed that regional norms and Western democratic values were the foundation of some initiatives (e.g. Z-Inspection). The RAI Certification is grounded in the Organisation for Economic Co-operation and Development (OECD) AI principles, which incorporate human rights objectives, good technology practices, and an emphasis on accountability and oversight. The Trusted AI product is based on the framework for achieving Trustworthy AI and fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (EU Charter). Malta has published its own ethical AI Framework to systems developers and users with a platform to showcase AI systems and applications that followed the framework (Table 11).

Functions –All the initiatives in this article share a common goal: to foster more ethical AI systems by providing a benchmark to evaluate them [41]. All the initiatives aim to 1) increase users trust in AI systems and applications through providing easy to understand information on the quality of the system and its trustworthiness; 2) increase competition among developers and companies by providing them with tools to use to transparently compare different AI systems; and 3) help AI system developers to better understand how they can comply with the standards for AI [29]. They provide information, either offering guidance to technology designers and companies about which ethical considerations must be considered or informing consumers on what to look for when choosing AI products and services [41] (Table 11).

Target Users – There are different target users for the existing initiatives. However, they target two broad groups: companies developing the AI systems and users or consumers who buy, use, or deploy the AI systems. First, labels allow companies to create and develop AI systems and solutions that uphold good practices and meet the regulatory requirements of the AI ecosystem. Secondly, labels allow users of AI systems to make informed choices and decisions that are align with the consumer values [37]. The full potential of a label will be unleashed if it is widely used by AI systems developers and gains recognition and acceptance among end users (Table 11).

Stages of development - The initiatives discussed in this section span different stages of development, from early-stage proposals to fully operational initiatives. However, many are yet to gain widespread acceptance and use (Table 11).

Transparency mechanisms - The initiatives shortlisted and analysed in this paper address trust and transparency through three core elements: 1) allowing consumers and the public to know when the AI system is being used (Transparent operation), 2) providing sufficient information on the label to support human decision making, and 3) signalling to consumers and users that the AI system has gone through some degree of human oversight (Table 11).

Audit structure – We studied the labels through an internal audit lens designed to be used by stakeholders tasked with the safe and trustworthy development and delivery of the AI system. This is shown in Table 5. Results show that there was no clear structure, framework, or model setting out the audit process for the stakeholders adopting the labels. Moreover, results show that AI trust self-regulatory initiatives can and should address several elements. This includes 1) the regulatory scope that the label is aiming at, 2) the type of the system being audited, 3) the label's approach to self-regulation, 4) the method of assessment, 5) the label visualisation or form of the label that best suits the stakeholders it is addressing (see Section 3 for the three possible forms for the presentation of the trust label and certificate for AI systems), and 6) the auditor and reassessment schedule (Table 12).

Table 11. Detailed analysis of the shortlisted initiatives

Name	Foundations	Functions	Target users	Stages	Key requirements	Transparency
The AI Trust Label	Global ethical principles and guidelines [32], [35]	Incorporating values into algorithmic decision-making and measuring the fulfilment of values	End consumers, companies and government organisations	Proposed (conceptual) (Proposed in 2020)	Transparency; Accountability; Privacy; Justice; Reliability; Environmental sustainability	Transparent operation, Human intervention and oversight
Mandatory Labelling Scheme	Basic values, rights and freedoms enshrined in the German Constitution and in the Charter of Fundamental Rights of the European Union [38]	Data and algorithmic considerations	Algorithmic systems of enhanced criticality	Proposed (conceptual) (Proposed in 2019)	Whether, when and to what extent algorithmic systems are being used	Human oversight
Open Ethics Label	Pre-existing frameworks, such as Data Ethics Framework and Trustworthy AI Framework [39]	Informing consumers	Developers and product owners of an AI system	Operational (Developed in 2017)	How the system uses data, How the system processes the data, What decisions the system makes	Transparency of system performance criteria, Transparent operation
Certification System for AI Applications	IT, philosophy, and law [40]	Responsible AI promotion and adoption in finance, health care, HR, and procurement	Developers, providers, users, consumers	In progress (Developed in 2020)	Fairness; Transparency; Autonomy and control; Data protection; Security; Reliability	Human oversight, Transparent operation
Z-Inspection	Regional norms and Western democratic values [34]	Offering guidance to technology designers and companies	Developers, users, the public	Proposed (conceptual) (Proposed in 2020)	Human agency and oversight; Technical robustness and safety; Privacy and data governance; Transparency; Diversity, non-discrimination, and fairness; Societal and	

					environmental wellbeing; Accountability; Assessment of EU democratic values; Avoiding concentration of power	
Swiss Digital Trust Label	Ethical values and principles related to data protection law, privacy law, consumer protection law and competition law [41]	To signal to the consumer their commitment to and good practices around digital security and data handling	Consumers	Operational	Security; Data protection; Reliability; Fair user interaction	Transparency of system performance criteria, Transparent operation
Malta's National AI Certification Framework	Malta's ethical AI Framework, Towards Trustworthy AI [42]	To create the conditions for AI to springboard from Malta to the world	Practitioners and companies	Proposed (conceptual) (Proposed in 2019)	Human agency; Privacy and data governance; Explainability and transparency; Well-being Accountability; Fairness and being unbiased; Performance and safety	Human intervention, Human oversight, Transparent operation
EU Certification for 'Trusted AI' Products	Charter of Fundamental Rights of the European Union (EU Charter) and in relevant international human rights law [43]	To offer guidance on fostering and securing ethical and robust AI and operationalise ethical principles in sociotechnical systems	Public	In progress (It was in the pilot phase in 2019)	Resilience; Safety; Absence of prejudice, discrimination, or bias	Human oversight, Transparency of system performance criteria
Responsible Artificial Intelligence Institute Certification Beta	Grounded in OECD AI principles (human rights objectives, good technology practices, and an emphasis on accountability and oversight) [44]	To assess the data, model, and contextual deployment of the system as these are all factors in the efficacy, fairness, or usefulness of the system	Organisations, Senior executives, compliance officers, procurement officers, regulators,	In progress (Developed in 2019. beta version launched in 2021)	Robustness; Accountability; Bias and fairness; Data quality; Explainability; Interpretability	Human oversight

			investors, consumers, trusted integrators			
Ethics Certification Program for Autonomous and Intelligent Systems	Transparency, Accountability, and Privacy (TAP) framework [45]	Recommendation for CTA/CTT developers and users to promote ethical integrity in the design, development, implementation, operation, maintenance, retirement, and regulatory processes within this emerging domain	Cities, and public and private organisations in diverse sectors	Operational (Developed in 2018; criteria launched in 2020)	Transparency. Accountability; Reduction in algorithmic bias in autonomous and intelligent systems	Human oversight, Transparent operation
Certificate of Fairness for AI Systems	United Nation's Universal Declaration of Human Rights [46]	Recommendation to developers	Women	Proposed (conceptual) (Proposed in 2019)	Criteria to be defined at industry level.	Human oversight, Transparent operation

Table 12. Audit structure of the shortlisted initiatives

Name	**Regulatory Scope**	**Types of systems**	**Approach**	**Method of assessment**	**Visualisation**	**Auditor & Reassessment**
The AI Trust Label (VDE)	AI System/Operational level	AI systems (independent of the risk posed)	Risk and Value-based	Value-Criteria-Indicators-Observable Model	Trust Score	Self-assessed, Third-party; Regular assessment
Mandatory Labelling Scheme	Data and Algorithmic systems	AI systems	Risk-based	Criticality pyramid and risk-adapted regulatory system	No info found	Self-assessed; When needed
Open Ethics Label	Training data, Algorithm, Decision Space	Medium and low-risk	Risk-based	Open Ethics Transparency Protocol	Trust Seal	Self-assessed; When needed
Certification System for AI Applications (Fraunhofer Institute)	AI systems	Certain application areas and risk classes	Risk-based	No info found	No info found	Third-party (Neutral and accredited inspectors); Assessment as required
Z-Inspection	AI System/ System level	AI systems	Holistic and analytical	Set up, Assess, and Resolve	Trust Score	Self-assessed; Assessment as required
Swiss Digital Trust Label	General digital services	AI systems	Practice-based	No info found	Trust Score	No info; Assessment as required
Malta's National AI Certification Framework	AI-based solutions	AI systems	Risk and Value-based	AI-specific Control Objectives and Evaluation Criteria	No info found	Third-party; Assessment as required
EU Certification for 'Trusted AI' Products	Consumer-facing AI systems	AI systems	Risk-based	Trustworthy AI assessment list	No info found	Self-assessed; Assessment as required
Responsible Artificial Intelligence Institute Certification Beta (RAII)	AI System/ System level [44]	AI systems	Risk-based	A set of 89 questions, response indicators, and evidence requirements	Trust Seal	Third-party; Annual assessment

Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)	Autonomous and Intelligent Systems [45]	High-risk	Risk-based and Technology-informed	Risk-based conformity assessment satisfactory criteria	Trust Seal (Level-based)	Self-assessed, Third party; Assessment as required
Certificate of Fairness for AI Systems	System algorithm [46]	AI systems	Value-based	Algorithm Impact Assessments	Textual	Third-party; Assessment as required

A.7 Analysis of the short-listed labels in the context of the AI Act

Using the analytical framework, in Table 13, we show what requirements from the AI Act are covered by each trust label, and we further show where the strength of the label is with regards to their alignment with the AI Act using the coloured blocks both across the labels and the Articles.

Risk Management - nine trust initiatives address the Risk Management requirements of the AI Act to some extent. For example, as shown in Figure 14 and Table 13, Mandatory Labelling Scheme, AI certification, Malta's AI Certification, and RAI Certificate addressed at least three of the five requirements of the AI Act under Article 9.

Data and Data Governance - nine initiatives also address the Data Governance requirements of the AI Act to some degree. For example, as shown in Figure 14 and Table 13, AI certification (Fraunhofer Institute) and Trusted AI Product initiatives address at least five data governance requirements of the AI Act. However, the coverage of these requirements by the remaining initiatives is low as they address only one of the seven data governance requirements in the act.

Technical Documentation - eight AI trust initiatives address the Documentation requirements of the AI Act to some degree. However, as shown in Figure 14 and Table 13, we found no evidence showing that the Open Ethics, Trusted AI Product, and ECPAIS cover the documentation requirements of the AI Act regarding the provision of detailed information to end users and the information on the system and its purpose for authorities to assess system compliance. The remaining initiatives address the importance of documenting and providing users and regulatory assessors with information on the technical and non-technical aspects of the AI system and application.

Record Keeping - the Ethics Label and the Swiss Digital Trust Label initiatives have relatively greater coverage of the AI Act requirements in the areas of record keeping, monitoring, and traceability. This is followed by Malta's AI Certification, Z-Inspect, and AI certification (Fraunhofer Institute). However, as shown in Figure 14 and Table 13, we found no evidence that the remaining initiatives address record-keeping requirements.

Transparency - as seen from Figure 14 and Table 13, the AI certification, the Open Ethics, and the RAI Certificate cover at least three of the five AI Act transparency requirements under this article. However, these frameworks define and approach transparency differently. The transparency principle is fundamental to AI's ethical, safe, and responsible use. The notion of transparency in these frameworks also enables sustained awareness about inclusion and accountability in developing and adopting AI systems. For example, Malta has developed a certification framework to provide a standard mechanism to establish transparency and build trust amongst stakeholders of AI systems [54]. However, the transparency requirements of Malta's AI Certification framework are categorised under the System Explicability to allow users to understand and challenge the AI system's operation.

Human Oversight - as shown in Figure 14 and Table 13, the human oversight component of ECPAIS, RAI Certificate, Fairness for AI Systems, and Ethics Label ensures that an AI system does not undermine human autonomy or cause other adverse effects. Human-in-the-Loop and Human-in-Command are two mechanisms employed in the four frameworks. However, the Swiss Digital Trust Label, Open Ethics Label, and Z-Inspect Label appear not to cover human oversight requirements in the AI Act.

Accuracy, Robustness, and Cybersecurity - as seen in Figure 14, Malta's AI Certification, RAI Certificate, Swiss Digital Trust Label, and Ethics Label are found to be aligned with several requirements under Article 15. The remaining initiatives either do not comply with Article 15 or comply with a minimal number of requirements under this Article. Despite the significance of system accuracy, robustness, and cybersecurity measures, results presented in Table 8 show that this area is not well addressed due to accuracy, robustness, and cybersecurity being fragmented and implemented at different levels. According to Malta's AI Certification, the implementation level and definition of accuracy and security are required in the AI system and use case context.

Table 13. AI Act requirements coverage by the short-listed AI trust initiatives

	Article 9					*Article 10*							*Article 11*	*Article 12*			*Article 13*					*Article 14*					*Article 15*						
																											Accuracy	Cybersecurity			Robustness		
	Risk estimation and evaluation	Functional safety	Identification & analysis of risk	Risk mitigation	Transparency of potential harm	Management of personal data	Bias examination	Data collection	Data preparation	Design choice	Assumptions formulation	Prior dataset assessment	Technical Documentation is available	Logs and Records	Monitoring of the operation	Traceability	Clarity of the input data	Clarity of the performance criteria	Clarity of the possible misuse/harm	Clarity of the system purpose	Transparent operation and system output	Human oversight and in control	Ability to decide to take over	Avoidance of automation bias	Understanding system capacity and monitoring operations	Understanding the state of the AI system	Accuracy metrics are declared	Measures to prevent risks	Measures to prevent & control for attack	Unauthorised attempt resilience	Technical redundancy solutions	Measures to prevent biased output	Resilience: error, fault, inconsistencies
The AI Trust Label			x		x	x	x	x		x			x	x	x	x					x	x	x		x	x			x	x			x
Mandatory Labelling Scheme	x	x	x	x		x							x						x			x					x						
Open Ethics	x						x										x	x			x												x
Certification System for AI Applications	x		x		x	x	x		x	x		x	x			x	x		x	x	x	x		x				x	x				
Z-Inspection	x	x											x			x				x													
Swiss Digital Trust Label					x	x							x	x	x	x		x			x						x				x		x
Malta's National AI Certification Framework		x	x		x	x	x					x	x		x	x					x	x	x				x		x		x		x
EU Certification for 'Trusted AI' Products		x	x			x	x	x	x			x					x	x				x					x						x
RAII	x	x		x			x		x		x	x	x				x		x		x	x	x	x	x	x		x	x	x			
ECPAIS																	x				x	x	x	x	x	x							
Certificate of Fairness for AI Systems										x			x								x	x	x	x	x	x							

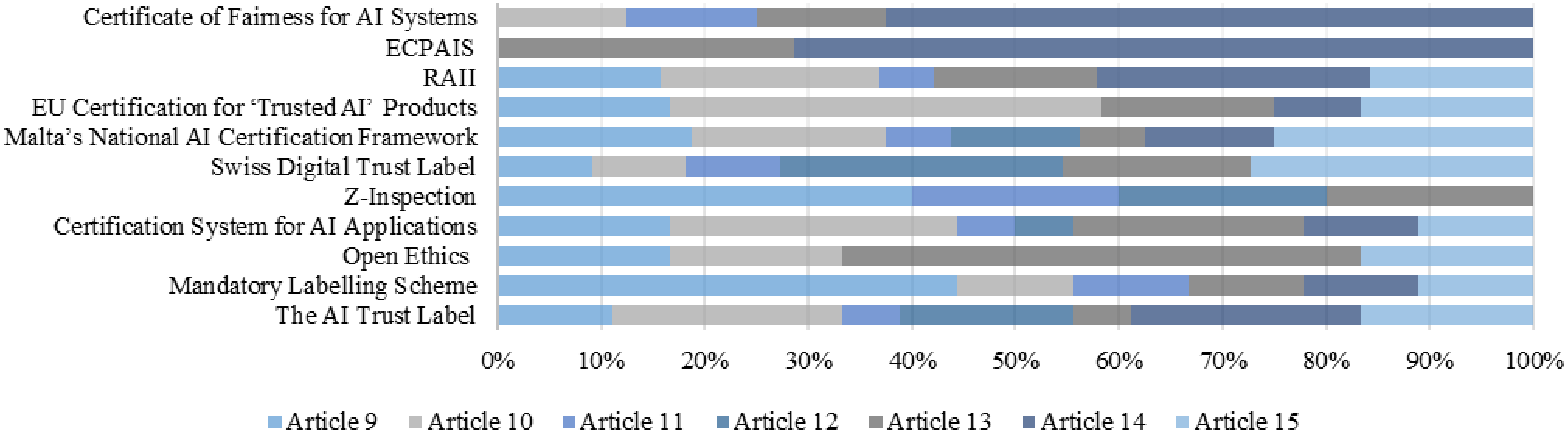


Figure 14. AI Act requirements coverage by each AI self-regulatory initiatives

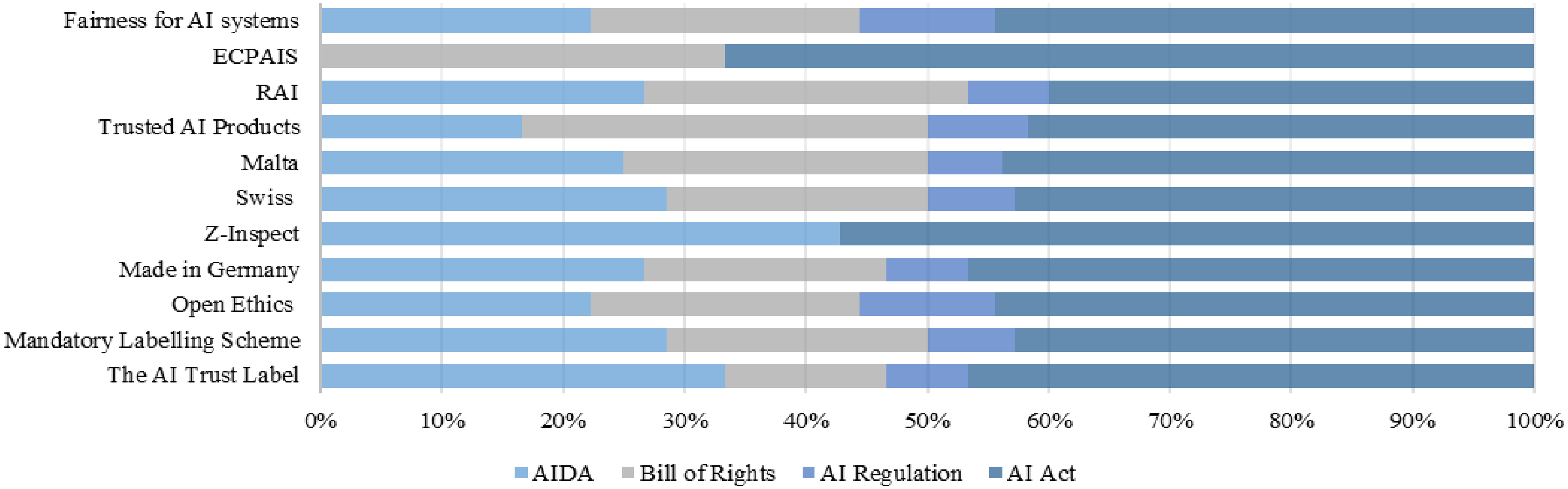
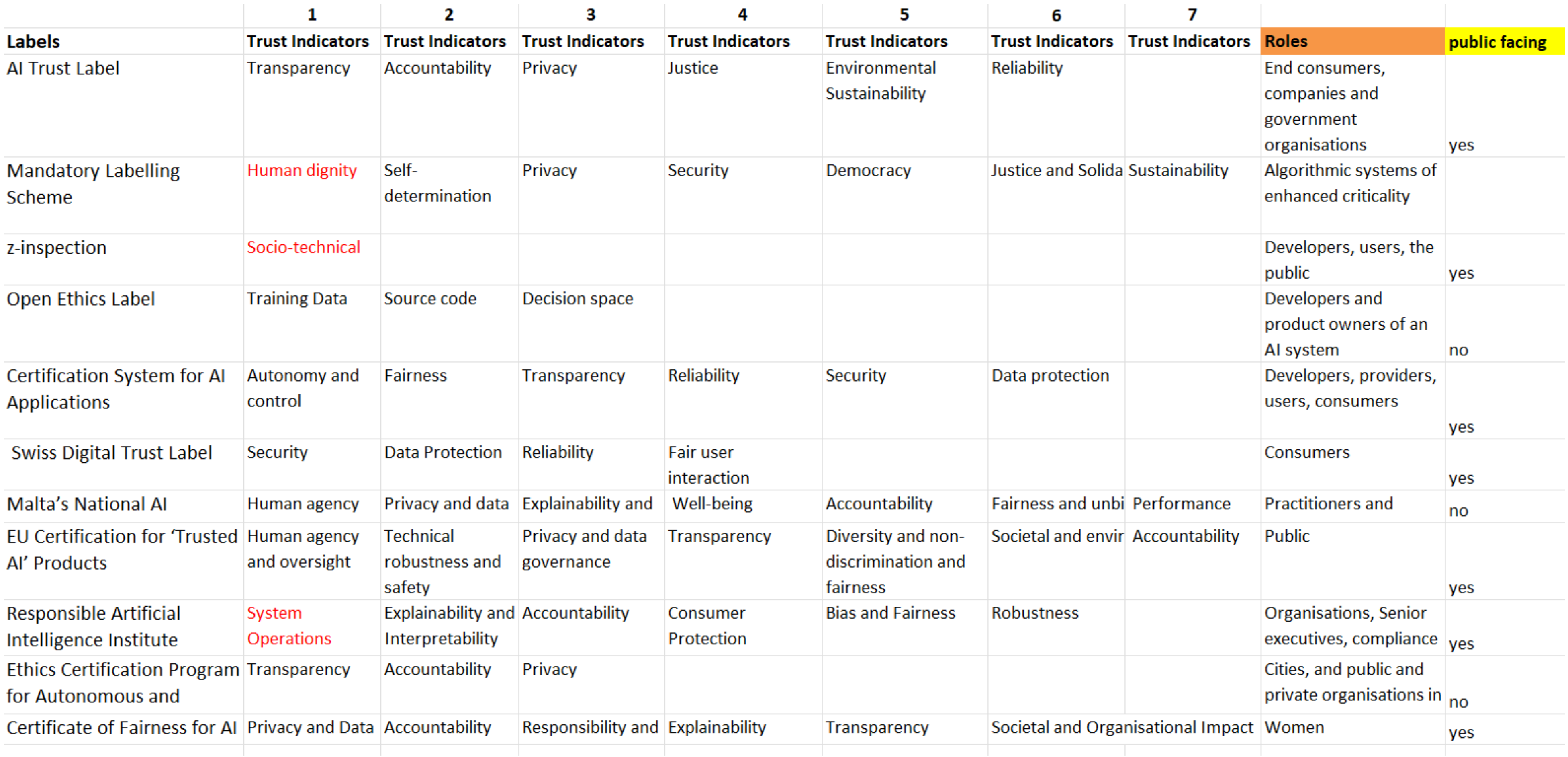


Figure 15. Self-regulatory initiatives coverage of the elements of the AI Act, the Bill of Rights, AIDA, and the AI Regulation in Japan

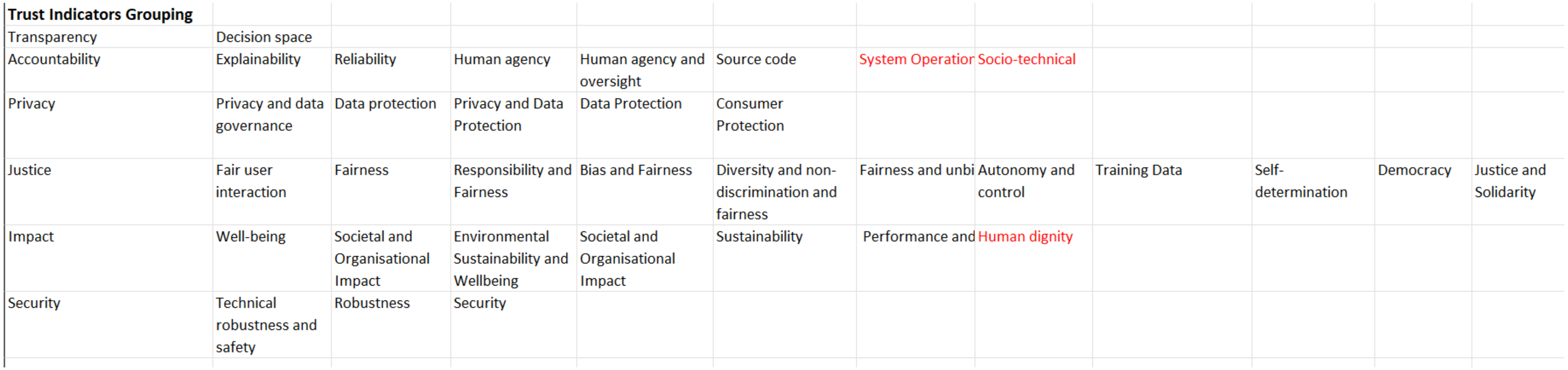
Annex B. Phase 2 – Selection of the Indicators of Trust

B.1 *Analysis stages (synthesis) of trust indicators extracted from the initiatives*



	1	2	3	4	5	6	7		
Labels	**Trust Indicators**	**Trust Indicators**	**Trust Indicators**	**Trust Indicators**	**Trust Indicators**	**Trust Indicators**	**Trust Indicators**	**Roles**	**public facing**
AI Trust Label	Transparency	Accountability	Privacy	Justice	Environmental Sustainability	Reliability		End consumers, companies and government organisations	yes
Mandatory Labelling Scheme	Human dignity	Self-determination	Privacy	Security	Democracy	Justice and Solida	Sustainability	Algorithmic systems of enhanced criticality	
z-inspection	Socio-technical							Developers, users, the public	yes
Open Ethics Label	Training Data	Source code	Decision space					Developers and product owners of an AI system	no
Certification System for AI Applications	Autonomy and control	Fairness	Transparency	Reliability	Security	Data protection		Developers, providers, users, consumers	yes
Swiss Digital Trust Label	Security	Data Protection	Reliability	Fair user interaction				Consumers	yes
Malta's National AI	Human agency	Privacy and data	Explainability and	Well-being	Accountability	Fairness and unbi	Performance	Practitioners and	no
EU Certification for 'Trusted AI' Products	Human agency and oversight	Technical robustness and safety	Privacy and data governance	Transparency	Diversity and non-discrimination and fairness	Societal and envir	Accountability	Public	yes
Responsible Artificial Intelligence Institute	System Operations	Explainability and Interpretability	Accountability	Consumer Protection	Bias and Fairness	Robustness		Organisations, Senior executives, compliance	yes
Ethics Certification Program for Autonomous and	Transparency	Accountability	Privacy					Cities, and public and private organisations in	no
Certificate of Fairness for AI	Privacy and Data	Accountability	Responsibility and	Explainability	Transparency	Societal and Organisational Impact		Women	yes

Figure 16. First synthesis of indicators and criteria from the labels – 1st round of analysis of extracted indicators from the labels – screenshot of the raw data from the excel



Trust Indicators Grouping											
Transparency	Decision space										
Accountability	Explainability	Reliability	Human agency	Human agency and oversight	Source code	System Operatior	Socio-technical				
Privacy	Privacy and data governance	Data protection	Privacy and Data Protection	Data Protection	Consumer Protection						
Justice	Fair user interaction	Fairness	Responsibility and Fairness	Bias and Fairness	Diversity and non-discrimination and fairness	Fairness and unbi	Autonomy and control	Training Data	Self-determination	Democracy	Justice and Solidarity
Impact	Well-being	Societal and Organisational Impact	Environmental Sustainability and Wellbeing	Societal and Organisational Impact	Sustainability	Performance and	Human dignity				
Security	Technical robustness and safety	Robustness	Security								

Figure 17. Second synthesis of criteria – 2nd round of analysis of extracted indicators from the labels – screenshot of the raw data from the excel

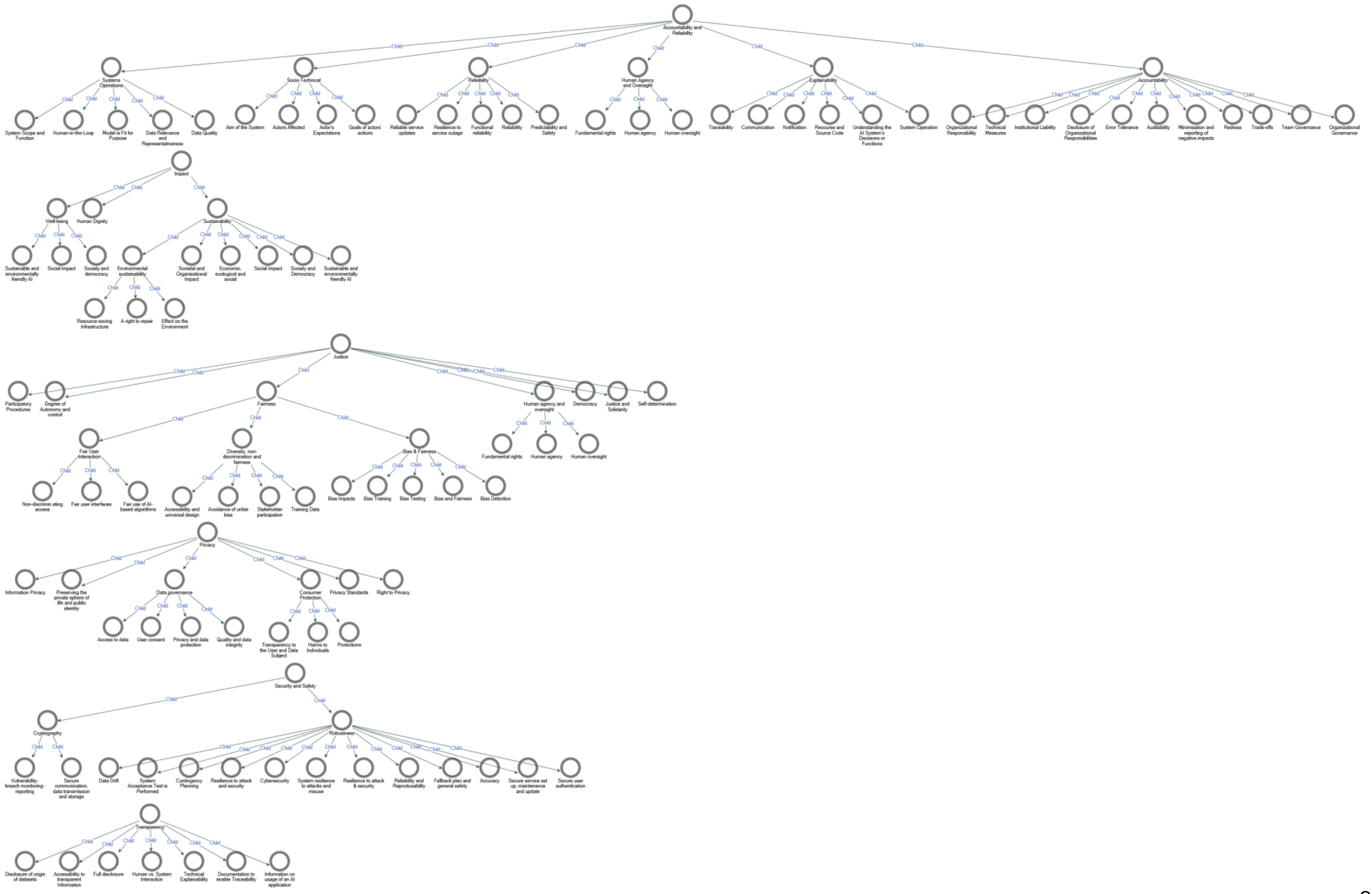

Accountability and Reliability
Child
Systems Operations
System Scope and Function
Human-in-the-Loop
Model is Fit for Purpose
Data Relevance and Representativeness
Data Quality
Socio-Technical
Aim of the System
Actors Affected
Actor's Expectations
Goals of actors' actions
Reliability
Reliable service updates
Resilience to service outage
Functional reliability
Reliability
Predictability and Safety
Human Agency and Oversight
Fundamental rights
Human agency
Human oversight
Explainability
Traceability
Communication
Notification
Recourse and Source Code
Understanding the AI System's Decisions or Functions
System Operation
Accountability
Organizational Responsibility
Technical Measures
Institutional Liability
Disclosure of Organizational Responsibilities
Error Tolerance
Auditability
Minimisation and reporting of negative impacts
Redress
Trade-offs
Team Governance
Organizational Governance
Impact
Well-being
Human Dignity
Sustainable and environmentally friendly AI
Social impact
Society and democracy
Sustainability
Environmental sustainability
Societal and Organisational Impact
Economic, ecological and social
Social impact
Society and Democracy
Sustainable and environmentally friendly AI
Resource-saving Infrastructure
A right to repair
Effect on the Environment
Justice
Participatory Procedures
Degree of Autonomy and control
Fairness
Human agency and oversight
Democracy
Justice and Solidarity
Self-determination
Fair User Interaction
Diversity, non-discrimination and fairness
Bias & Fairness
Fundamental rights
Human agency
Human oversight
Non-discrimin ating access
Fair user interfaces
Fair use of AI-based algorithms
Accessibility and universal design
Avoidance of unfair bias
Stakeholder participation
Training Data
Bias Impacts
Bias Training
Bias Testing
Bias and Fairness
Bias Detection
Privacy
Information Privacy
Preserving the private sphere of life and public identity
Data governance
Consumer Protection
Privacy Standards
Right to Privacy
Access to data
User consent
Privacy and data protection
Quality and data integrity
Transparency to the User and Data Subject
Harms to Individuals
Protections
Security and Safety
Cryptography
Robustness
Vulnerability-breach monitoring-reporting
Secure communication, data transmission and storage
Data Drift
System Acceptance Test is Performed
Contingency Planning
Resilience to attack and security
Cybersecurity
System resilience to attacks and misuse
Resilience to attack & security
Reliability and Reprodusability
Fallback plan and general safety
Accuracy
Secure service set up, maintenance and update
Secure user authentication
Transparency
Disclosure of origin of datasets
Accessibility to transparent Information
Full disclosure
Human vs. System Interaction
Technical Explainability
Documentation to enable Traceability
Information on usage of an AI application

Figure 18. Second synthesis of criteria - NVivo qualitative analysis generated

Table 14. Third synthesis - 3rd round of analysis that includes requirements associated with each grouped indicators and criteria from the labels

Trust Indicators	Criteria	Requirements
Accountability and Reliability	Accountability	Auditability
		Disclosure of Organizational Responsibilities
		Error Tolerance
		Institutional Liability
		Minimisation and reporting of negative impacts
		Organizational Governance
		Organizational Responsibility
		Redress
		Team Governance
		Technical Measures
		Trade-offs
	Explainability	Communication
		Notification
		Recourse and Source Code
		System Operation
		Traceability
		Understanding the AI System’s Decisions or Functions
	Reliability	Functional reliability
		Predictability and Safety
		Reliable service updates
		Resilience to service outage
	Socio-Technical	Actors Affected
		Actor's Expectations
		Aim of the System
		Goals of actors’ actions
	Systems Operations	Data Quality Data Relevance and Representativeness Human-in-the-Loop Model is Fit for Purpose

		System Scope and Function
Impact	Human Dignity	-
	Sustainability	Economic, ecological and social Environmental sustainability (A right to repair, Effect on the Environment, Resource-saving Infrastructure) Social impact Societal and Organisational Impact Society and Democracy Sustainable and environmentally friendly AI
	Well-being	Social impact Society and democracy Sustainable and environmentally friendly AI
Justice	Degree of Autonomy and control	-
	Democracy	-
	Fairness	Bias & Fairness (Bias and Fairness, Bias Detection, Bias Impacts, Bias Testing, Bias Training) Diversity, non-discrimination and fairness (Accessibility and universal design, Avoidance of unfair bias. Stakeholder Participation, Training Data) Fair User Interaction (Fair use of AI-based algorithms, Fair user interfaces, non-discriminating access)
	Human agency and oversight	Fundamental rights Human agency Human oversight
	Justice and Solidarity	
	Participatory Procedures	
	Self-determination	
Privacy	Consumer Protection	Harms to Individuals Information Privacy Preserving the private sphere of life and public identity Privacy Standards Protections Right to Privacy
	Data governance	Access to data Privacy and data protection Quality and data integrity User consent

Security and Safety	Cryptography	Secure communication, data transmission and storage Vulnerability-breach monitoring-reporting
	Robustness	Accuracy Contingency Planning Cybersecurity Data Drift Fallback plan and general safety Reliability and Reproducibility Resilience to attack & security Resilience to attack and security Secure service set up, maintenance and update Secure user authentication System Acceptance Test is Performed System resilience to attacks and misuse
Transparency	Documentation and Accessibility	Accessibility to transparent Information Information on usage of an AI application Documentation to enable Traceability
	Full Disclosure (System Level)	Human vs. System Interaction Disclosure of origin of datasets Technical Explainability Transparency to the User and Data Subject

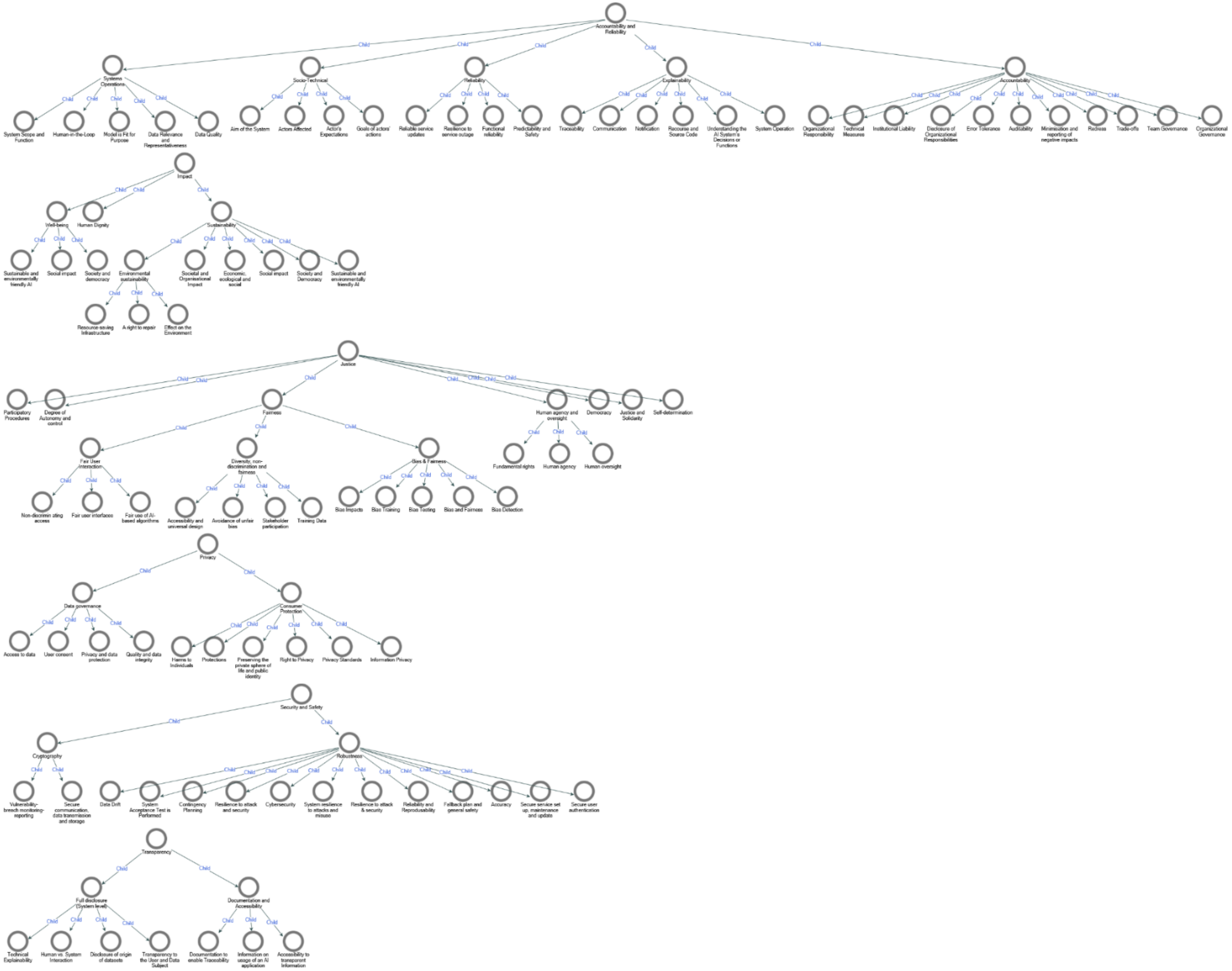

Accountability and Reliability
Child
Systems Operations
System Scope and Function
Human-in-the-Loop
Model is Fit for Purpose
Data Relevance and Representativeness
Data Quality
Socio-Technical
Aim of the System
Actors Affected
Actor's Expectations
Goals of actors' actions
Reliability
Reliable service updates
Resilience to service outage
Functional reliability
Predictability and Safety
Explainability
Traceability
Communication
Notification
Recourse and Source Code
Understanding the AI System's Decisions or Functions
System Operation
Accountability
Organizational Responsibility
Technical Measures
Institutional Liability
Disclosure of Organizational Responsibilities
Error Tolerance
Auditability
Minimisation and reporting of negative impacts
Redress
Trade-offs
Team Governance
Organizational Governance
Impact
Well-being
Human Dignity
Sustainability
Sustainable and environmentally friendly AI
Social impact
Society and democracy
Environmental sustainability
Societal and Organisational Impact
Economic, ecological and social
Social impact
Society and Democracy
Sustainable and environmentally friendly AI
Resource-saving Infrastructure
A right to repair
Effect on the Environment
Justice
Participatory Procedures
Degree of Autonomy and control
Fairness
Human agency and oversight
Democracy
Justice and Solidarity
Self-determination
Fair User Interaction
Diversity, non-discrimination and fairness
Bias & Fairness
Fundamental rights
Human agency
Human oversight
Non-discrimin ating access
Fair user interfaces
Fair use of AI-based algorithms
Accessibility and universal design
Avoidance of unfair bias
Stakeholder participation
Training Data
Bias Impacts
Bias Training
Bias Testing
Bias and Fairness
Bias Detection
Privacy
Data governance
Consumer Protection
Access to data
User consent
Privacy and data protection
Quality and data integrity
Harms to Individuals
Protections
Preserving the private sphere of life and public identity
Right to Privacy
Privacy Standards
Information Privacy
Security and Safety
Cryptography
Robustness
Vulnerability-breach monitoring-reporting
Secure communication, data transmission and storage
Data Drift
System Acceptance Test is Performed
Contingency Planning
Resilience to attack and security
Cybersecurity
System resilience to attacks and misuse
Resilience to attack & security
Reliability and Reproducability
Fallback plan and general safety
Accuracy
Secure service set up, maintenance and update
Secure user authentication
Transparency
Full disclosure (System level)
Documentation and Accessibility
Technical Explainability
Human vs. System Interaction
Disclosure of origin of datasets
Transparency to the User and Data Subject
Documentation to enable Traceability
Information on usage of an AI application
Accessibility to transparent Information

Figure 19. Third synthesis of criteria - NVivo qualitative analysis generated

Table 15. Trust indicators - a general logic – results of the synthesis of criteria

Trust Indicators	Criteria	Requirements	Labels
Accountability	Accountability	Auditability	Malta's National AI Certificate; EU Certificate for 'Trusted AI Product'
		Disclosure of Organizational Responsibilities	AI Trust Label
		Error Tolerance	AI Trust Label
		Institutional Liability	AI Trust Label
		Minimisation and reporting of negative impacts	Malta's National AI Certificate; EU Certificate for 'Trusted AI Product'
		Organizational Governance	ECPAIS, RAII
		Organizational Responsibility	AI Trust Label
		Redress	Malta's National AI Certificate; EU Certificate for 'Trusted AI Product'
		Team Governance	Certificate of Fairness for AI Systems; RAII
		Technical Measures	AI Trust Label
		Trade-offs	EU Certificate for 'Trusted AI Product'
	Explainability	Communication	Malta's National AI Certificate; RAII
		Notification	RAII
		Recourse and Source Code	Open Ethics Label; RAII
		System Operation	Malta's National AI Certificate
		Traceability	Malta's National AI Certificate
		Understanding the AI System's Decisions or Functions	Malta's National AI Certificate; RAII
	Socio-Technical	Actors Affected	Z-Inspection
		Actor's Expectations	
		Aim of the System	
		Goals of actors' actions	
	Systems Operations	Data Quality	RAII
		Data Relevance and Representativeness	RAII
		Human-in-the-Loop	RAII
		Model is Fit for Purpose	RAII

		System Scope and Function	Malta’s National AI Certificate; RAII
Reliability	Functional reliability		Swiss Digital Trust Label
	Predictability and Safety		AI Trust Label
	Reliable service updates		Swiss Digital Trust Label
	Resilience to service outage		Swiss Digital Trust Label
Impact	Humiliation, Attachment and Empathy		Mandatory Labelling Scheme; Malta’s National AI Certificate
	Human Rights	Human Rights Impact Assessment	Swiss Digital Trust Label
	Environment	A right to repair	AI Trust Label
		Effect on the Environment	
		Resource-saving Infrastructure	
		Environmentally friendly AI	Malta’s National AI Certificate; EU Certificate for ‘Trusted AI Product’
	Social		Certificate of Fairness for AI Systems; Mandatory Labelling Scheme; Malta’s National AI Certificate; EU Certificate for ‘Trusted AI Product’
	Societal		EU Certificate for ‘Trusted AI Product’
	Workplace		Certificate of Fairness for AI Systems
	Democratic Process		Mandatory Labelling Scheme; EU Certificate for ‘Trusted AI Product’
Fairness or Non-discrimination	Autonomy and Self-determination	-	
	Fairness or non-discrimination	Bias (Bias and Fairness, Bias Detection, Bias Impacts, Bias Testing, Bias Training)	EU certification for trusted AI products, RAII; AI Trust Label
		Diversity, non-discrimination and fairness (Accessibility and universal design, Avoidance of unfair bias. Stakeholder Participation, Training Data)	Malta’s National AI Certificate; Open Ethics Label
		Fair User Interaction (Fair use of AI-based algorithms, Fair user interfaces, non-discriminating access)	Swiss Digital Trust Label
	Human agency		Malta’s National AI Certificate

Human agency and oversight	Human oversight		EU Certification for trusted AI products; Malta’s National AI Certificate
Privacy and Data Governance	Consumer Protection	Harms to Individuals	RAII
		Information Privacy	AI Trust Label
		Privacy Standards	AI Trust Label
		Right to Privacy	Mandatory Labelling Scheme
		Preserving the private sphere of life and public identity	ECPAIS
	Data governance	Access to data	EU Certification for trusted AI products; Malta’s National AI Certificate
		Privacy and data protection	Certification System for AI Applications, EU Certification for trusted AI products; Malta’s National AI Certificate
		Quality and data integrity	EU Certification for trusted AI products; Malta’s National AI Certificate
		User consent	Swiss Digital Trust Label
Technical Robustness and Safety	Cryptography	Secure communication, data transmission and storage	Swiss Digital Trust Label
		Vulnerability-breach monitoring-reporting	Swiss Digital Trust Label
	Robustness	Accuracy	Malta’s National AI Certificate
		Contingency Planning	RAII
		Cybersecurity	AI Trust Label, Mandatory Labelling Scheme
		Data Drift	RAII
		Fallback plan and general safety	EU Certification for Trusted AI Products; Malta’s National AI Certificate
		Reliability and Reproducibility	Malta’s National AI Certificate
		Resilience to attack & security	EU Certification for Trusted AI Products; Malta’s National AI Certificate
		Secure service set up, maintenance and update	Swiss Digital Trust Label
		Secure user authentication	Swiss Digital Trust Label
		System Acceptance Test is Performed	RAII
Transparency	Documentation and Accessibility	Accessibility to transparent Information	AI Trust Label; Certificate of Fairness for AI Systems

		Information on usage of an AI application	Certification System for AI Applications
		Documentation to enable Traceability	Certification System for AI Applications; EU Certification for Trusted AI Products;
	Full Disclosure (System Level)	Human vs. System Interaction	EU Certification for Trusted AI Products;
		Disclosure of origin of datasets	AI Trust Label
		Technical Explainability	EU Certification for Trusted AI Products;
		Transparency to the User and Data Subject	RAII

B.2 Detailed definitions of the requirements

Table 16. Definitions of the trust indicators and requirements associated with each indicators

Requirements	**Labels**	**Definition/Description**
Auditability	Malta's National AI Certificate; EU Certificate for 'Trusted AI Product'	Auditability entails the enablement of the assessment of algorithms, data and design processes. This does not necessarily imply that information about business models and intellectual property related to the AI
Disclosure of Organizational Responsibilities	AI Trust Label	This entails if there is an institutionalised opportunity to provide anonymous information to relevant parties, responsibilities defined with respect to third parties (affected persons/users), if responsibilities for possible damage and liability cases documented, and if there a comprehensive logging of the design process
Error Tolerance	AI Trust Label	There is a culture of dealing openly with mistakes within organisations
Institutional Liability	AI Trust Label	This entails the availability of appropriate monetary means, an insurance policy and/or other forms of compensation in case of liability

Minimisation and reporting of negative impacts	Malta's National AI Certificate; EU Certificate for 'Trusted AI Product'	ability to report on actions or decisions that contribute to a certain system outcome, and to respond to the consequences of such an outcome. Identifying, assessing, documenting and minimising the potential negative impacts of AI systems
Organizational Governance	ECPAIS, RAII	Holding organizations and people behind entities in the integrated system of CTA/CTT to account through the fulfillment of ethical obligations (as set forth in this report) for their roles and decisions that impact the inputs, process, outputs, and ecosystem outcomes.
Organizational Responsibility	AI Trust Label	Assignment of internal organisational responsibility
Redress	Malta's National AI Certificate; EU Certificate for 'Trusted AI Product'	When unjust adverse impact occurs, accessible mechanisms should be foreseen that ensure adequate redress.
Team Governance	Certificate of Fairness for AI Systems; RAII	Independent review processes and ongoing monitoring of an AI system throughout its lifecycle.
Technical Measures	AI Trust Label	Methods for complexity reduction of technical functions, to ensure internal traceability. Systems with a learning component to monitor system interaction with their environment.
Trade-offs	EU Certificate for 'Trusted AI Product'	This entails that relevant interests and values implicated by the AI system should be identified and that, if conflict arises, trade-offs should be explicitly acknowledged and evaluated in terms of their risk to ethical principles, including fundamental rights.
Communication	Malta's National AI Certificate; RAII	The extent to which people are appropriately informed about the inputs and outputs of the AI system. Communicate to end-users that they are interacting with an AI system rather than a human (e.g. by way of a label or disclaimer).
Notification	RAII	The processes, if any, that are in place to notify a person when an automated decision has been made about them
Recourse and Source Code	Open Ethics Label; RAII	The mechanisms available to end users to appeal the AI system's decisions and/or outputs. And, access to source code allows programmers and skilled users to contribute to the solution. Providing information about algorithmic choices is a way to evaluate privacy and security risks associated with the application logic.
System Operation	Malta's National AI Certificate	Assessing and explaining the operation of the AI system to ensure that end-users and other affected individuals can understand the operation of the AI system, e.g., model interoperability, training and testing data
Traceability	Malta's National AI Certificate	This entails measures to ensure traceability, including Design and development, rogramming methods or how the model is built, Training methods, including which input data is collected and

		selected and how, and Scenarios or cases used to test and validate e.., detail on data, Outcomes of the algorithmic system, Outcomes or decisions that could be made by or based on the algorithm.
Understanding the AI System's Decisions or Functions	Malta's National AI Certificate; RAII	The extent to which the organization documents, reviews, and/or publishes additional system information
Actors Affected	Z-Inspection	Usage scenarios are a useful tool to describe the aim of the system, the actors, their expectations, the goals of actors' actions, the technology, and the context. socio-technical scenarios can also be used to broaden stakeholder understanding of one's own role in understanding technology, as well as awareness of stakeholder interdependence.
Actor's Expectations		
Aim of the System		
Goals of actors' actions		
Data Quality	RAII	The strength of the AI system's performance and accuracy alongside the types of data it uses.
Data Relevance and Representativeness	RAII	The extent to which an AI system is used within or outside an organization and how many people it affects.
Human-in-the-Loop	RAII	The extent of staff interaction with an AI system's decision-making process.
Model is Fit for Purpose	RAII	The sector/industry in which the AI system operates and that sector/industry's associated risk level alongside what the AI system is programmed to do
System Scope and Function	Malta's National AI Certificate; RAII	This entails that the purpose for which the AI system is deployed in a particular area and the contexts, use cases, and limitations of the AI system is clear
	Swiss Digital Trust Label	The service shall provide its users with an extensive, easy-to-access, easy-to-understand description of its functionalities, and shall operate in strict accordance with this description.
	AI Trust Label	Predictability and safety as robustness and resilience (AI applications are considered reliable when they perform in intended ways as well as when they do not possess vulnerabilities to external attackers. Reliability is akin to the concept of predictability, meaning that systems can prevent manipulation of various kinds. AI security problems arise when AI applications have software vulnerabilities, when they are not resilient against cyberattacks, or when the integrity and confidentiality of personal data are being compromised. AI applications, no different from any other intricate pieces of software, have security vulnerabilities. In most cases, we are talking about data poisoning attacks, adversarial examples or the exploitation of other flaws in the design of autonomous systems)

	Swiss Digital Trust Label	The service provider shall publish, in a way that is easy to access and understand for the user, the defined support period and the need for that support period.
	Swiss Digital Trust Label	Disaster recovery, business continuity and data backup and restore policies and procedures shall be in place and regularly tested to ensure ongoing availability of the service and associated data.
	Mandatory Labelling Scheme; Malta's National AI Certificate	Assess whether the AI system encourages humans to develop attachment and empathy towards the system. AI systems can sometimes be deployed to shape and influence human behaviour through mechanisms that may be difficult to detect, since they may harness sub-conscious processes, including various forms of unfair manipulation, deception, herding and conditioning, all of which may threaten individual autonomy. The overall principle of user autonomy must be central to the system's functionality. Key to this is the right not to be subject to a decision based solely on automated processing when this produces legal effects on users or similarly significantly affects them
Human Rights Impact Assessment	Swiss Digital Trust Label	Conduct human rights impact assessment, identifying and documenting potential trade-offs between different principles and rights.
A right to repair	AI Trust Label	This entails the disposal of obsolete IT hardware, used to run AI applications. In this context, a right to repair can improve the situation.
Effect on the Environment		This includes the extent to which AI systems have positive or negative effects on the environment
Resource-saving Infrastructure		Within the field of AI, this includes setting up resource-saving infrastructures for information technology, primarily through building power-efficient data centres as well as developing less power consuming machine learning models. So far, the more computational resources AI models have at their disposal and the more training data they process, the more powerful and accurate the systems are. Increase in computation, however, means an increase in energy consumption, which brings with it increased carbon footprints. In this field, certification processes are especially useful for end-users to evaluate the carbon footprint of a given AI application. An important criterion to arrive at environment-friendly AI applications is the transparency regarding power consumption and the provision of sustainability data in general.
Environmentally friendly AI	Malta's National AI Certificate; EU Certificate for 'Trusted AI Product'	Sustainable and environmentally friendly AI to ensure negative environmental impacts of AI development and use are minimised, e.g. the amount of data used by the data centres
	Certificate of Fairness for AI Systems; Mandatory Labelling Scheme; Malta's National AI Certificate;	This entails that the indirect negative social impacts of AI development and use are minimised. The AIA needs to highlight the impact on the workforce as well as society / community as a whole.

	EU Certificate for ‘Trusted AI Product’	
	EU Certificate for ‘Trusted AI Product’	This entails that the impact of the system should also be assessed from a societal perspective, taking into account its effect on institutions, democracy and society at large.
	Certificate of Fairness for AI Systems	This entails that the AIA needs to highlight the impact on the workforce as well as society / community as a whole. For example, it needs to demonstrate how the system augments human capabilities and how the algorithm does not become policy, thus removing human autonomy in wider decision making.
	Mandatory Labelling Scheme; EU Certificate for ‘Trusted AI Product’	This entails that the use of AI systems should be given careful consideration particularly in situations relating to the democratic process, including not only political decision-making but also electoral contexts.
-		This justifies the freedom of individuals to make autonomous decisions. Users should be able to make informed autonomous decisions regarding AI systems. Self-determination is a fundamental expression of freedom, and encompasses the notion of informational self-determination. The term “digital self- determination” can be used to express the idea of a human being a self-determined player in a data society.
Bias (Bias and Fairness, Bias Detection, Bias Impacts, Bias Testing, Bias Training)	EU certification for trusted AI products, RAII; AI Trust Label	The bias dimension assesses whether the AI system was designed in a manner that promotes fairness and avoids bias. The degree to which the organization has put mitigation processes in place to combat unintended bias and similar issues and, the organization and development team have engaged with bias and fairness issues, such as by conducting research, situating the system in its historical and cultural context, hiring team members with relevant expertise, and providing opportunities for workers displaced by the system, is considered. The assessment also reviews any bias training that the organization has provided to the AI system’s
Diversity, non-discrimination and fairness (Accessibility and universal design, Avoidance of unfair bias. Stakeholder Participation, Training Data)	Malta’s National AI Certificate; Open Ethics Label	This entails the extent to which AI system assesses and verifies the accommodation of a wide range of individual preferences and abilities
Fair User Interaction (Fair use of AI-based	Swiss Digital Trust Label	This entails that the system shall provide a non-discriminating access to all its potential users to interact with the system

algorithms, Fair user interfaces, non-discriminating access)		
	Malta’s National AI Certificate	Human agency Ensure appropriate level of human engagement with AI
	EU Certification for trusted AI products; Malta’s National AI Certificate	Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects.
Harms to Individuals	RAII	The degree to which the AI system could harm its users
Information Privacy	AI Trust Label	Informational privacy as data is being used for specific purposes, after explicit consent, and with a right to delete or rectify
Privacy Standards	AI Trust Label	The extent to which privacy standards are integrated into data processing itself, meaning privacy by design
Right to Privacy	Mandatory Labelling Scheme	The right to privacy is intended to preserve an individual’s freedom and the integrity of his or her personal identity. Potential threats to privacy include the wholesale collection and evaluation of data about even the most intimate of topics.
Preserving the private sphere of life and public identity	ECPAIS	Preserving the private sphere of life and public identity of an entity (individual, group, community) to be free from unacceptable intrusion or invasion, and upholding the entity’s dignity.
Access to data	EU Certification for trusted AI products; Malta’s National AI Certificate	In any given organisation that handles individuals’ data (whether someone is a user of the system or not), data protocols governing data access should be put in place. These protocols should outline who can access data and under which circumstances. Only duly qualified personnel with the competence and need to access individual’s data should be allowed to do so.
Privacy and data protection	Certification System for AI Applications, EU Certification for trusted AI products; Malta’s National AI Certificate	Privacy and data protection Ensure protection of individuals’ privacy rights, including compliance with all relevant data processing laws
Quality and data integrity	EU Certification for trusted AI products;	Quality and data integrity Ensure quality and integrity of data used in AI design, development and training

	Malta's National AI Certificate	
User consent	Swiss Digital Trust Label	This entails that users shall be informed about the purpose of the processing and the legal basis for processing of their personal data in clear and plain language.
Secure communication, data transmission and storage	Swiss Digital Trust Label	This entails that the system shall apply best practice cryptography to data in transit and at rest, ensuring that the cryptography is reviewed and evaluated, delivers the required functions for all transmitted data/sensitive and applicable data at rest, and is appropriate to the properties of the technology, risk, and usage. All data in transit/rest over open communication lines such as the internet must be encrypted.
Vulnerability-breach monitoring-reporting	Swiss Digital Trust Label	The service provider shall continually monitor, identify, and rectify security vulnerabilities and/or breaches, and shall provide a public point of contact as part of a vulnerability disclosure policy so that security researchers and others are able to report issues. Critical security vulnerabilities shall be communicated to relevant authorities within 72 hours if not corrected, and the impacted users shall be timely and adequately informed. Personal data breaches shall be communicated to relevant authorities and impacted data subjects within 72 hours.
Accuracy	Malta's National AI Certificate	Accuracy Ensure accuracy of AI system's outputs
Contingency Planning	RAII	The extent to which the organization is prepared for adversarial attacks, load inputs, and other edge cases and extreme scenarios.
Cybersecurity	AI Trust Label, Mandatory Labelling Scheme	It is traditionally understood to include three aims concerning IT systems: confidentiality, integrity and availability (entails compliance with stringent requirements, e. g. in relation to human/machine interaction or system resilience to attacks and misuse).
Data Drift	RAII	The organization's processes and procedures for combatting the degradation of the AI system's performance due to changing data and variable relationships.
Fallback plan and general safety	EU Certification for Trusted AI Products; Malta's National AI Certificate	AI systems should have safeguards that enable a fallback plan in case of problems. Fallback plan and general safety Ensure the AI system is developed and used safely.
Reliability and Reproducibility	Malta's National AI Certificate	Reliability and reproducibility Ensure reliability of the AI system
Resilience to attack & security	EU Certification for Trusted AI Products; Malta's National AI Certificate	Resilience to attack and security Mitigate the AI system's vulnerabilities

Secure service set up, maintenance and update	Swiss Digital Trust Label	Guidance for secure installation, configuration, and updates shall be in place and updated for each release if necessary. Guidance shall be available in a manner that is easy to access and understand. Any major changes shall lead to a communication to the users in an easy-to-understand format. All software components shall be updatable in a secure manner, and verification of security updates shall be in place.
Secure user authentication	Swiss Digital Trust Label	This entails that the system shall be subject to a state of art password policy for secure authentication
System Acceptance Test is Performed	RAII	The extent to which the AI system has been exposed to and tested across several edge cases.
Accessibility to transparent Information	AI Trust Label; Certificate of Fairness for AI Systems	This entails that the information listed above be easily accessible to any person subject to the algorithm.
Information on usage of an AI application	Certification System for AI Applications	This includes understanding what purpose the application has, what it does, what the potential risks are (also in terms of other audit areas, for example, reliability, security, and fairness), and who the target group of the application is.
Documentation to enable Traceability	Certification System for AI Applications; EU Certification for Trusted AI Products;	This entails that the system information should be documented to the best possible standard to allow for traceability and an increase in transparency. This also applies to the decisions made by the AI system. This enables identification of the reasons why an AI-decision was erroneous which, in turn, could help prevent future mistakes. Traceability facilitates auditability as well as explainability.
Human vs. System Interaction	EU Certification for Trusted AI Products;	AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system. This entails that AI systems must be identifiable as such. In addition, the option to decide against this interaction in favour of human interaction should be provided where needed to ensure compliance with fundamental rights.
Disclosure of origin of datasets	AI Trust Label	The origin of the dataset or the data used to train the model should be disclosed
Technical Explainability	EU Certification for Trusted AI Products;	Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system). Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings.

Transparency to the User and Data Subject	RAII	The degree to which AI system users are informed that AI is assisting with decisions

B.3 Public facing trust indicators

Table 17. Public facing trust indicators - according to the trust initiatives

Labels	Indicators of trust	Requirements	Public facing	Other users

Ethics Label for AI Systems (VDE)	Transparency;
	Accountability;
	Privacy;
	Justice;
	Environmental Sustainability
		Transparency as explainability and interpretability;
		Accountability refers to questions of assigning responsibility;
		Justice with aspects of algorithmic fairness and inclusion;
		Environmental Sustainability
		Privacy to safeguard an individual's private sphere
			yes	companies and government organisations

Mandatory Labelling Scheme	Human dignity;
	Self-determination;
	Privacy;
	Security;
	Democracy;
	Justice and Solidarity;
	Sustainability
			Specific product only (Algorithmic systems of enhanced criticality)	Specific product only (Algorithmic systems of enhanced criticality)

z-inspection	Socio-technical	Socio-technical (AI domain and Usage, Frameworks/ regulations/ laws, Evidence-base)	Yes	Developers

Open Ethics Label	Training Data;
	Source code;
	Decision space
		Training Data (proprietary, limited access, open, rule-based);
		Source code (proprietary source, open source);
		Decision space (restricted and unrestricted);
			No	Developers and product owners of an AI system

Certification System for AI Applications	Autonomy and control;
	Fairness;
	Transparency;
	Reliability;
	Security;
	Data protection
		Autonomy and control (Are autonomous, effective usage of the AI possible?);
		Fairness (Does the AI treat all persons concerned fairly?);
		Transparency (Are the AI functions and the decisions made by the AI comprehensible?);
		Reliability (Does the AI work reliably and is it robust?);
			yes	Developers, providers

		Security (Is the AI protected against attacks, accidents, and errors?); Data protection (Does the AI protect privacy and other sensitive information?)		

Swiss Digital Trust Label	Security;
	Data Protection;
	Reliability;
	Fair user interaction
		Security (Secure communication, data transmission and storage, Secure user authentication, Secure service set up, maintenance and update, Vulnerability/ breach monitoring/ reporting);
		Data Protection (User consent, Data retention and data processing);
		Reliability (Reliable service updates, Resilience to service outage, Functional reliability, Accountability);
		Fair user interaction (Non-discriminating access, Fair user interfaces, Fair use of AI-based algorithms)
			Yes	No

Malta's National AI Certification Framework	Human agency;
	Privacy and data governance;
	Explainability and transparency;
	Well-being;
	Accountability;
	Fairness and unbiased;
	Performance and safety
		Human agency (fundamental rights, human agency, human oversight);
		Privacy and data governance (privacy and data protection, quality and data integrity, access to data);
		Explainability and transparency (traceability, Explainability, Communication);
		Well-being (Sustainability and environmentally friendly AI, Social impact, society and democracy);
		Accountability (Auditability, Redress, minimization and reporting of negative impacts);
		Fairness and unbiased (Avoidance of unfair bias, Accessibility and universal design, Stakeholder participation);
		Performance and safety (Accuracy, Reliability and responsibility, Resilience to attach and security, Fallback plan and general safety)
			No	Practitioners and companies

EU Certification for 'Trusted AI' Products	Human agency and oversight, Technical robustness and safety, Privacy and data governance,	Human agency and oversight (fundamental rights, human agency and human oversight); Technical robustness and safety (resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility);	Yes (only)	no

	Transparency, Diversity/non-discrimination and fairness, Societal and environmental wellbeing, Accountability	Privacy and data governance (respect for privacy, quality and integrity of data, and access to data); Transparency (traceability, explainability and communication); Diversity/non-discrimination and fairness (avoidance of unfair bias, accessibility and universal design, and stakeholder participation); Societal and environmental wellbeing (sustainability and environmental friendliness, social impact, society and democracy); Accountability (auditability, minimisation and reporting of negative impact, trade-offs and redress)		

Responsible Artificial Intelligence Institute Certification Beta (RAII)	System Operations;
	Explainability and Interpretability;
	Accountability;
	Consumer Protection;
	Bias and Fairness;
	Robustness
		System Operations (System Scope and Function, Human-in-the-Loop, Model is Fit for Purpose, Data Relevance and Representativeness, Data Quality);
		Explainability and Interpretability (Communication about the Outcome, Notification, Recourse, Understanding the AI System's Decisions or Functions);
		Accountability (Organizational Governance, Team Governance);
		Consumer Protection (Transparency to the User and Data Subject, Harm to Individuals, Protections);
		Bias and Fairness (Bias Impacts, Bias Training, Bias Testing);
		Robustness (Data Drift, System Acceptance Test Performed, Contingency Planning)
			yes	Organisations, Senior executives, compliance officers, procurement officers, regulators, investors, consumers, trusted integrators

Ethics Certification Program for Autonomous and Intelligent Systems	Transparency;
	Accountability;
	Privacy (TAP is used as the framework)
		System Operations (Confidence in the total ecosystem behavior, degree of autonomy, Accessible and fair control and feedback)
		System Design (Ethical architecture, design, development, and sunset; Clarity of CTA/CTT concepts of operations)
		Compliance/Legal (Suitable and sufficient CTA/CTT organizational governance and oversight, ethical integrity)
			No	Cities, and public and private organisations in diverse sectors

Certificate of Fairness for AI Systems	Privacy and Data Protection; Accountability;
	Responsibility and Fairness;
	Explainability;
	Transparency;
	Societal and Organisational Impact
		Privacy and Data Protection (appropriate and secure sourcing, handling and use of data on the correct legal basis):
		Accountability (Person responsible at every step of the process);
		Responsibility and Fairness (values embedded in the machine, depending on the degree of machine autonomy;
		Explainability (description and explanation of the key decision making processes);
		Transparency (Information easily accessible to any person subject to the algorithm ; Performance metrics and accountability must be prominent in any information provided to the individual data subject);
		Societal and Organisational Impact (impact on the workforce as well as society / community)
			Yes	no

Annex C. Phase 3 – Stakeholder specific Indicators

C.1 Stakeholder analysis of the initiatives based on the AI Act

Table 18. Initiative's stakeholder analysis – Stakeholders as identified in the initiatives

	Public	Developers and product owners of AI systems	Practitioners and Companies	Developers, providers and system users	Senior executives, Compliance officers, procurement officers, regulators, investors, trusted integrators	Companies, Governments	Consumers	Developers, users, public	System users- Algorithmic systems of enhanced criticality
Labels Analysis acc. to the AI Acts Articles									
A10 - Data and data governance									
Management of personal data (GDPR etc.)						x			
Operational design domain - Alignment of data quality to ODD - fit for purpose - Bias Identification and Assessment									
Bias examination	X			x	x				
Data Collection	X					x			
Data preparation processing operations	X				x				
Design Choice				x		x			
Formulation of relevant assumptions					x				
Prior assessment of the dataset	X				x				
A11 - Technical Documentation									
Technical documentation is available			x	x	x	x		x	
A12 - Record Keeping									
Logs and records						x			
Monitoring of the operation						x			
Traceability				x		x			
A13 - Transparency and provision of information to users									
Clarity of the input data and information				x	x	x			
Clarity of the performance criteria	X								

Clarity of the potential misuse or harm				x	x				
Clarity of the purpose of the system				x				x	
Transparent operation and system output				x	x	x	x		
A14 - Human oversight									
Human oversight and in control (human-machine interface tools)	x		x	x	x	x			
Human Oversight Measures									
Ability to decide to take over in particular circumstances						x			
Avoidance of automation bias				x					
Understanding system capacity and monitor operations						x			
Understanding the State of the AI System						x			
A15 - Accuracy - Robustness - Cybersecurity									
Accuracy measures									
Accuracy metrics are declared in the accompanying instructions of use	x								
Cybersecurity measures									
Appropriate measures to relevant circumstances and risks					x	x			
Measures to prevent and control for attacks					x	x			
Resilient as regards attempts by unauthorised third parties					x	x			
Robustness measures									
Availability of technical redundancy solutions (e.g. backups)									
Availability of measures to prevent biased output to be used as system input									
Resilient as regards errors, faults or inconsistencies	x	x	x			x			
A9 - Risk Management system									
Estimation and evaluation of risks				x	x			x	
Functional safety	x		x		x				
Identification and analysis of risk	x			x				x	
Mitigation of identified risk					x	x			
Transparency of potential harms			x	x		x			

Table 19. Initiative's stakeholder analysis – Stakeholders grouped based on their nature

	Consumers	Developers	Provider (Governments)	Senior executives, Compliance officers, procurement officers, regulators, investors, trusted integrators
Labels Analysis acc. to the AI Acts Articles				

A10 - Data and data governance				
Management of personal data (GDPR etc.)	X	x	x	
Operational design domain - Alignment of data quality to ODD - fit for purpose - Bias Identification and Assessment				
Bias examination	X	x		x
Data Collection	X	x	x	
Data preparation processing operations	X			x
Design Choice	X	x	x	
Formulation of relevant assumptions	X			x
Prior assessment of the dataset	X			x
A11 - Technical Documentation				
Technical documentation is available	X	x	x	x
A12 - Record Keeping				
Logs and records		x	x	
Monitoring of the operation		x	x	
Traceability	X	x	x	
A13 - Transparency and provision of information to users				
Clarity of the input data and information	X	x	x	x
Clarity of the performance criteria	X			
Clarity of the potential misuse or harm	X	x		x
Clarity of the purpose of the system	X	x		
Transparent operation and system output	X	x	x	x
A14 - Human oversight				
Human oversight and in control (human-machine interface tools)	X	x	x	x
Human Oversight Measures				
Ability to decide to take over in particular circumstances		x	x	
Avoidance of automation bias	X	x		
Undrestanding system capacity and monitor operations		x	x	
Undrestanding the State of the AI System		x	x	
A15 - Accuracy - Robustness - Cybersecurity				
Accuracy measures				
Accuracy metrics are declared in the accompanying instructions of use	X			
Cybersecurity measures				
Appropriate measures to revevant circumstances and risks	X	x	x	x
Measures to prevent and control for attacks	X	x	x	x
Resilient as regards attempts by unauthorised third parties	X	x	x	x
Robustness measures				

Availability of technical redundancy solutions (e.g. backups)				
Availability of mesures to prevent biased output to be used as system input				
Resilient as regards errors, faults or inconsistencies	X	x	x	
A9 - Risk Management system				
Estimation and evalutation of risks	X	x		x
Functional safety	X	x		x
Identification and analysis of risk	X	x		
Mitigation of identified risk	X	x	x	x
Transparency of potential harms	X	x	x	

C.2 Student Survey

Student survey has been designed and developed.

Survey Draft

Introduction:

Welcome to our survey on AI trust labels developed by the Adra-e project and how they can address modern risks associated with AI advancements. In today's fast-paced digital world, AI technology is all around us, from healthcare to finance, education, and more. With the increasing power and use of AI in products, services, and applications, we need ways to assess and trust these systems. That's where "AI trust labels" come in.

Think of AI trust labels like food nutrition labels. Just as a food label helps you make choices about what you eat, AI trust labels provide information about AI systems, giving you transparency and guidance when you use them. The trust labels developed by the Adra-e project will be certified by the European Commission.

In this survey, we want to know what you think about AI trust labels and how much you trust them. We're also interested in your opinions about how AI trust labels could benefit society in the long run. Your input is important, and it will help us understand what university students like you think about AI trust labels. Thank you for participating!

Questions:

1. **What year of study are you in?**

2. **What is your field of study?**

3. **On average, how often do you use AI technologies? (ex: ChatGPT)**

A. Daily
B. Weekly
C. Never

4. **Rate your understanding of AI.**

A. I have in-depth knowledge of AI; I study AI systems and could implement them.

B. I am somewhat knowledgeable about AI,
C. I have very little understanding of AI

5. Rate your understanding on AI trust labels from (1-10), 1 being very poor, 10 being excellent.

1 2 3 4 5 6 7 8 9 10

6. Have you encountered an AI trust label before?

A. Yes
B. No
C. I do not know

7. On a scale from 1-10, how concerned are you about AI technology? 1 being not concerned at all, 10 being extremely concerned.

1 2 3 4 5 6 7 8 9 10

8. On a scale from 1-10, how concerned are you that an AI-based product/service may be using your data? 1 being not concerned at all, 10 being extremely concerned.

1 2 3 4 5 6 7 8 9 10

9. Would you be more likely to use a product with a trust label developed by the EU?

A. yes
B. no

10. Would you feel comfortable using regulated AI technology? Example: Chat GPT is not yet regulated.

Strongly yes Yes Somewhat No Strongly not

11. Select the 3 most important things that should be on a trust label. (tick boxes)

A. Criteria or Standards that ensure the safety of a product/service
B. Link to Verification
C. Trustworthy Content
D. Contact information/Chatbot available/FAQ page (customer support)
E. Reputable logos and images bearing the EU seal of approval
F. Certified Data Privacy Standards

12. Would you like to see more education and awareness initiatives on AI trust labels in your university curriculum or elsewhere?

A. Yes

B. No

13. What are your main concerns about AI technology? (typed response)

Student Feedback:

- **The survey was very clear, Question 7 could be a bit clearer as to what we would like in an answer. How concerned they are in a general sense. Similar to how Q8 was specific and clear.**

Annex D. Phase 4 – AI Trust Label and Stakeholder Engagement

This task will start in M19. Some work has been done with regard to forming an stakeholder's activity group where we aim to implement Delphi method to finalize consumer trust indicators. The list of potential members is provided in 1.3 Methodological approach.