iPROLEPSIS

D2.3 / The iPROLEPSIS trustworthy AI framework

Editor	Contractual delivery date	Actual delivery date
Apostolidis Georgios (AUTH)	December 2023	February 2024
Deliverable type	**Dissemination level**	**Version - date**
R - Document, report	PU – Public	1.0 - 06/02/2024

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency. Neither the European Union nor the European Health and Digital Executive Agency can be held responsible for them.

Deliverable ID

Project acronym	iPROLEPSIS
Project full title	Psoriatic arthritis inflammation explained through multi-source data analysis guiding a novel personalised digital care ecosystem
Grant Agreement ID	101095697

Deliverable number	D2.3
Deliverable title	The iPROLEPSIS trustworthy AI framework
Work package	WP2 - Knowledge mining, foundation and participatory design
Deliverable type	R - Document, report
Dissemination level	PU – Public
Version - date	1.0 - 06/02/2024

Contractual delivery date	December 2023
Actual delivery date	February 2024
Lead partner	AUTH
Editor	Apostolidis Georgios (AUTH)
Contributors	Eleni Vasileiou (AUTH); Vasilis Charisis (AUTH); Ioannis Drivas (DBC); Sotiris Michagiannis (DBC); Alex Bensenousi (AIN); Andreas Raptopoulos (WCS); Kosmas Dimitropoulos (CERTH), Silvia Reis (PLUX); Amalia Ntemou (INTRA)
Reviewed by	Sofia Balula Dias (FMH-ULISBOA) Nikos Melanitis (AIN)
Approved by	Leontios Hadjileontiadis (AUTH, Project Coordinator)
Keywords	ALTAI; Ethical AI; Inclusive AI; Responsible AI, Trustworthy AI

Document history

Version	Date	Contributors	Action / status
0.1	06/11/2023	AUTH	Document structure (table of contents) ready
0.2	21/12/2023	AUTH	Input in sections 2 and 4.
0.3	12/01/2024	DBC; AUTH	Input in section 5
0.4	16/01/2024	AUTH; AIN; WCS; CERTH; PLUX; INTRA	Input in section 3
0.5	23/01/2024	AUTH	Input in sections 1 and 6. Ready for internal review
0.6	31/01/2024	FMH-ULISBOA	Reviewed by [Reviewer 1 Full name] (Partner short name)
0.7	01/02/2024	AIN	Reviewed by [Reviewer 2 Full name] (Partner short name)
0.8	05/02/2024	AUTH	Document revised
0.9	05/02/2024	AUTH	Approved by the Project Coordinator
1.0	06/02/2024	AUTH	Submitted to the EC by the Project Coordinator

Contents

List of figures

List of tables

List of abbreviations

AI	Artificial intelligence
AIPM	AI prediction model
ALTAI	Assessment list for trustworthy artificial intelligence
DNN	Deep neural network
HLEG	High-level expert group
GDPR	General data protection regulation
ML	Machine learning
MLOps	Machine Learning Operations
MoSCoW	Must-have, Should-have, Could-have, and Won't-have, or Will not have right now or Wish-have
TAI	Trustworthy AI

Executive summary

This report introduces the iPROLEPSIS trustworthy artificial intelligence (TAI) framework, specifying a set of prioritised requirements, assessment scores, and an implementation workplan throughout the project's lifecycle. Developed through an innovative approach using responses from a diverse group of experts, the framework relies on questionnaires consolidating existing knowledge and best practices from a state-of-the-art landscape analysis. The latter consists of a thorough search of the notable literature review papers and the main high-level TAI frameworks. Moreover, the report offers a comprehensive list of recommended open-source software tools to support the technical aspects of a TAI system. Additionally, a business planning approach integrating TAI as a central component is outlined, along with a brief preview of key points from the upcoming EU AI ACT.

1 Introduction

The advent of modern AI and the outbreak of AI applications created the need for regulating the AI development and its delivery to the end-users. The concerns about the AI regard its vague nature and they are occasionally associated with threatening and unwanted situations. Thus, the attempts to specify the main properties of an AI system and standardise its lifecycle highlighted the ethical and trustworthy perspectives of its usage. In other words, the assessment of an AI-based solution should not only rely on quantifying accuracy-related measures, but also on ensuring ethical treatment and establishing trust. Although the predictive accuracy is imperative for an AI-based product that operates in real-life applications, there are other AI properties, e.g., robustness, privacy, explainability and transparency, which are not always given the sufficient focus.

Although there are quite a few recently proposed guidelines, recommendations, and principles, none of them regard a gold standard for providing a solid definition for the trustworthiness in AI (Jobin et. al, 2019; Kaur et. al, 2022). The main reason is that these proposals are high-level, i.e., they outline a generic abstract AI system regardless the field and scale of application. Automotive industry, insurance and financial sectors, education, agriculture, and medicine are a few examples of fields that AI will revolutionise, but AI trustworthiness should emphasise on different aspects based on the application. In addition, the high-level frameworks focus on the what should be achieved, but not on the how. Thus, there is not a unanimous procedure that should be followed from a practical point of view. However, academia and industry have started to investigate approaches and methods to determine practical guidelines (Li et. al, 2023; de Hond et. al, 2022). These efforts show some consensus that the different aspects of trustworthy AI (TAI) are dependent on each other while they should be assessed in different phases of system's lifecycle, i.e., data preparation, design, development, deployment, and maintenance. Also, some TAI-related issues have been raised regarding possible side effects, e.g., slower development, increased cost, and delayed time to market (Li et. al, 2023). Nonetheless, there are several efficient open-source software solutions covering the majority of the technical TAI aspects that can be used to increase the productive of an AI development team.

The aim of this deliverable is to consolidate the main existing concepts about the TAI and to adapt them to the scope of iPROLEPSIS project. The outcome is the iPROLEPSIS TAI framework that contains a prioritized list of requirements that are going to be implemented during the lifecycle of the project. To achieve this, in a first place, a study of the current state-of-the-art takes place to instil the existing knowledge and practical approaches to the iPROLEPSIS TAI concept. Next, we apply a novel methodology to form the iPROLEPSIS TAI framework. The methodology consists of four steps which are:

1. the extraction of a full set of TAI recommendations,
2. the prioritisation of these recommendations based on the responses from a multidisciplinary group of experts,
3. the design of the iPROLEPSIS TAI framework as a self-assessment checklist and corresponding evaluation metrics, and
4. the definition of the strategy of applying the framework during the project's lifecycle.

Furthermore, it is imperative to screen and present a series of open-source tools that efficiently solve TAI-related technical problems to aid and optimise the implementation of TAI principles within iPROLEPSIS project. Finally, the iPROLEPSIS TAI ecosystem might be commercially exploited. In that case, the trustworthiness of the AI may involve a significant aspect of the

business promotion and success. Thus, the TAI as a business component is also explored and documented.

1.1 Document scope

The scope of this deliverable is to develop a TAI framework adapted to the objectives and workplan of iPROLEPSIS project. Given that the global TAI landscape does not provide specific guidelines for forming such a framework, a novel approach is adopted that extracts the most relevant TAI aspects via asking a multidisciplinary cohort of experts. The responses were transformed into requirements which are going to be implemented and monitored across the lifecycle of the project. Moreover, a thorough analysis of the existing tools that can support the successful implementation of TAI is also presented as a recommendation to the AI engineers of the project. Finally, since the project has the long-term expectation to go to the market, a couple of business-related subjects are discussed to be taken into consideration for the exploitation of the project AI-related results.

The document will be used during the lifecycle of the project. On one hand, it will be utilised during the specification of technical requirements, including those related to TAI. On the other hand, it will be employed when assessing the compliance with TAI upon the delivery of a part or the entire iPROLEPSIS system version.

1.2 Document structure

D3.2 provides a review of TAI state-of-the-art landscape, a methodology to extract an iPROLEPSIS based TAI framework, an overview of the publicly available open-source software tools that can be used to establish trustworthiness in an AI system, and a brief analysis of relation between achieving and commercialising AI trustworthiness.

Apart from this introductory Section 1, the rest of the document is structured in additional five sections as follows:

- Section 2 disentangles the TAI landscape via presenting selected notable review papers and high-level frameworks proposed by significant global organisations.
- Section 3 presents the methodology to form the iPROLEPSIS TAI framework.
- Section 4 reviews the noteworthy open-source software tools that support the implementation of trustworthy AI systems.
- Section 5 reports a couple of subjects that are important for the business development of iPROLEPSIS ecosystem, i.e., the TAI implementation canvas and the EU AI ACT[1].

2 State-of-the-art of trustworthy AI

TAI is a recent field of study and several attempts to define its properties as well as best practices have been made by experts from academia, industry, and regulators. This section maps the landscape of TAI presenting a representative set of notable literature review papers, as well as the main high-level frameworks.

2.1 Selected literature reviews

Understanding the multifaceted dimensions of TAI necessitates a thorough examination of existing knowledge and perspectives. Through comprehensive analysis and synthesis of research findings, the selected papers investigate both the technical and the societal aspects

[1] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

that underpin TAI. Each review paper offers valuable insights to set the ground for understanding the evolving landscape of TAI. **Table 1** organises the selected review papers, whereas the key points of each are presented in the continuation of the section.

Table 1 Selected literature review papers in the field of TAI.

Paper Title	Journal name	Year of publishing	No. documents/ sources	Searching literature until
The global landscape of AI ethics guidelines	Nature Machine Intelligence	2019	84	April 2019
Trustworthy Artificial Intelligence: A Review	ACM Computing Surveys	2022	228	September 2021
Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review	npj Digital Medicine	2022	72	January 2021
Trustworthy AI: From principles to practices	ACM Computing Surveys	2023	394	May 2022

The review paper titled **"The global landscape of AI ethics guidelines" (Jobin et. al, 2019)** provides a comprehensive overview and analysis of various AI ethics guidelines established by different organizations worldwide. Key points and findings are:

- ***Proliferation of AI ethics guidelines***: The authors identified a wide range of AI ethics guidelines from diverse sources, including governmental institutions, international organizations, industry consortia, research groups, and academia.
- ***Diversity in content***: The guidelines significantly varied in their content, scope, and emphasis. They covered a broad range of ethical considerations, including fairness, transparency, accountability, privacy, safety, human rights, and societal impact.
- ***Geographical distribution***: The paper highlighted the global nature of these guidelines, showing that they emerged from various regions worldwide. Nonetheless, the majority of the guidelines' issuers originate from North America, Europe, and Japan.
- ***Differences in focus and emphasis***: The guidelines reflected the diverse priorities and values of the issuing organizations. For instance, some guidelines emphasised technical aspects, such as algorithmic transparency and bias mitigation, while others focused more on broader societal implications and ethical principles.
- ***Challenges of implementation***: Despite the proliferation of guidelines, the paper pointed out challenges related to their implementation. The authors highlighted the need for mechanisms to translate these guidelines into actionable practices and standards within the development and deployment of AI systems.
- **Call for harmonization and collaboration:** Given the diversity and fragmentation of these guidelines, the authors advocated for greater collaboration, harmonization, and consolidation efforts among stakeholders to create a more coherent and effective framework for AI ethics.

Overall, the paper provided a comprehensive landscape analysis of the existing AI ethics guidelines published until 2019, highlighting the diversity, global distribution, and the need for collaboration and harmonization among these guidelines to effectively address the ethical challenges posed by AI.

The review paper titled **"Trustworthy Artificial Intelligence: A Review" (Kaur et. al, 2022)** provides an overview of different approaches to handle AI risks and increase trust and acceptance of these systems. It also discusses existing strategies for validating and verifying AI systems and the current standardization efforts for TAI. At last, they provide a perspective of the recent advancements in TAI to offer possible future research directions. Key points and findings are:

- The ***main research questions*** tried to be answered are:
  - What are the requirements to make AI trustworthy?
  - What guidelines and policies are required to govern the working of AI systems?
  - Why is human involvement significant in this changing era of AI?
  - What aspects are essential to make AI decisions acceptable?
- The ***requirements to make AI systems trustworthy*** are making them lawful, ethical, and robust. This means that the AI development, deployment, and use should follow all applicable laws and regulations; respect and follow the humans' ethical principles and guidelines, such as fairness, explainability, accountability, privacy, and acceptance; and be technically robust and reliable.
- The implantation of the ***AI governance*** presents a significant gap between the research and practice. Thus, there is a need to establish policies and standards to bring guidelines and existing laws into practice.
- ***Human involvement*** is essential in AI lifecycle. AI systems are being applied in several critical applications, where the consequences of failures are hazardous. Thus, human involvement is needed to ensure safe, reliable, and TAI operation. The paper suggests a few significant activities that should be considered, i.e., develop efficient algorithms, set limits for performance, flag and correct errors raised by the system, override wrong decisions, and continuously improve the system's performance.
- To make ***AI decisions acceptable***, end-users should clearly understand its usability, performance, and limitations. A proper evaluation mechanism should be established to assess TAI requirements and any excessive expectations by end-users should be prevented.
- ***Future Directions***: The paper concludes by offering recommendations for future research directions, emphasizing the need for: 1) ***standardized guidelines and policies***, 2) ***multidisciplinary collaboration***, 3) ***expectation management***, and 4) ***measurement mechanisms*** to quantify the AI trustworthiness.

The review paper **"Trustworthy AI: From Principles to Practices" (Li et. al, 2023)** offers an extensive review and exploration of the transition from AI ethical principles to practical implementation and real-world applications, focusing on the concept of TAI beyond predictive accuracy. Key points are:

- The main TAI aspects extracted from the literature review are: 1) ***Robustness***, 2) ***Generalization***, 3) ***Explainability and Transparency***, 4) ***Reproducibility***, 5) ***Fairness***, 6) ***Privacy Protection***, and 7) ***Accountability***.
  - ***Robustness*** is the ability of an algorithm or system to deal with execution errors, erroneous inputs, or unseen data. Evaluation via robustness test[2] and mathematical verification.
  - **Generalization**: the capability to distil knowledge from limited training data to make accurate predictions regarding unseen data. Evaluation via existing datasets and benchmarks (Zhou et. al, 2022).
  - **Explainability and transparency**:

[2] http://www.exforsys.com/tutorials/testingtypes/monkey-testing.html

        - Explainability addresses to understand how an AI model makes decision. Evaluation via:
            - ***Subjective human evaluation.*** The methods of evaluation in this context include interviews, self-reports, questionnaires, and case studies that measure, e.g., user satisfaction, mental models, and trust.
            - ***Human–AI task performance.*** In tasks performed with human–AI collaboration, the collaborative performance is significantly affected by the human understanding of the AI collaborator and can be viewed as a reflection of the quality of explanation.
        - Transparency considers AI as a software system and seeks to disclose information regarding its entire lifecycle.
    - **Reproducibility** can be evaluation via checklists that have been widely adopted in ML conferences to assess the reproducibility of submissions.
    - **Fairness** of AI systems is defined as avoiding instilling or exacerbating social bias.
    - **Privacy protection** mainly refers to protecting against unauthorized use of the data that can directly or indirectly identify a person or household. Evaluation via data protection impact assessment (DPIA) if any data processing poses a risk to data privacy.
    - **Accountability** addresses the regulation on AI systems to follow all various aforementioned requirements. Information disclosure and transparency are the main mechanisms used to facilitate accountability. Evaluation via checklist-based assessments.
- The main AI system lifecycle stages extracted from the literature review are: 1) ***Data preparation***, 2) ***Algorithms Design***, 3) ***Development***, 4) ***Deployment***, and 5) ***Management***.
- A ***systemic framework*** that organises the multi-disciplinary and fragmented approaches is proposed. This framework focuses on the practical implementation of trustworthiness. Specifically, this framework relates the different aspects of TAI to the different stages of the AI lifecycle. **Figure 1** presents a look-up table that organises the TAI aspects in different stages of an AI system.
- The authors propose a new workflow for the AI industry called ***TrustAIOps***, which extends Machine Learning Operations (MLOps), aiming to impose the requirements of trustworthiness over the entire AI lifecycle. This new workflow contains the following properties:
    - **Close collaboration between inter-disciplinary roles**: Building TAI requires organizing different roles, such as ML researchers, software engineers, safety engineers, and legal experts.
    - **Aligned principles of trustworthiness**: The risk of untrustworthiness exists in every stage in the lifecycle of an AI system. Mitigating such risks requires that all stakeholders in the AI industry be aware of and aligned with unified trustworthy principles.
    - **Extensive management of artifacts**: An industrial AI system is built upon various artifacts such as data, code, models, configuration, product design, and operation manuals.
    - **Continuous feedback loops**: The feedback loops implemented through classical DevOps workflows should connect and iteratively improve the five stages of its lifecycle, i.e., data, algorithm, development, deployment, and management.

- A few challenges arise; thus, a prioritisation of the TAI aspects should take place to adapt to the ethical requirements of a specific AI system.
    - Side-effects of TAI aspects, e.g., slowed development, longer learning curves and increased cost to build the final system.
    - Trade-offs between TAI aspects, for instance:
        - Transparency vs. Privacy: Disclosing inappropriate information might increase potential risks (Akhtar & Mian, 2018).
        - Trust vs. Accuracy: Adversarial robustness increases the model's generalizability and reduces overfitting yet tends to negatively impact its overall accuracy (Zhang et. Al, 2019).
        - Adversarial robustness vs. Fairness: each other can be negatively affected during development (Xu et. al, 2021).
        - Fairness vs. Explainability: Studies have shown several cases where explanation can be unfair (Dodge et. al, 2019).

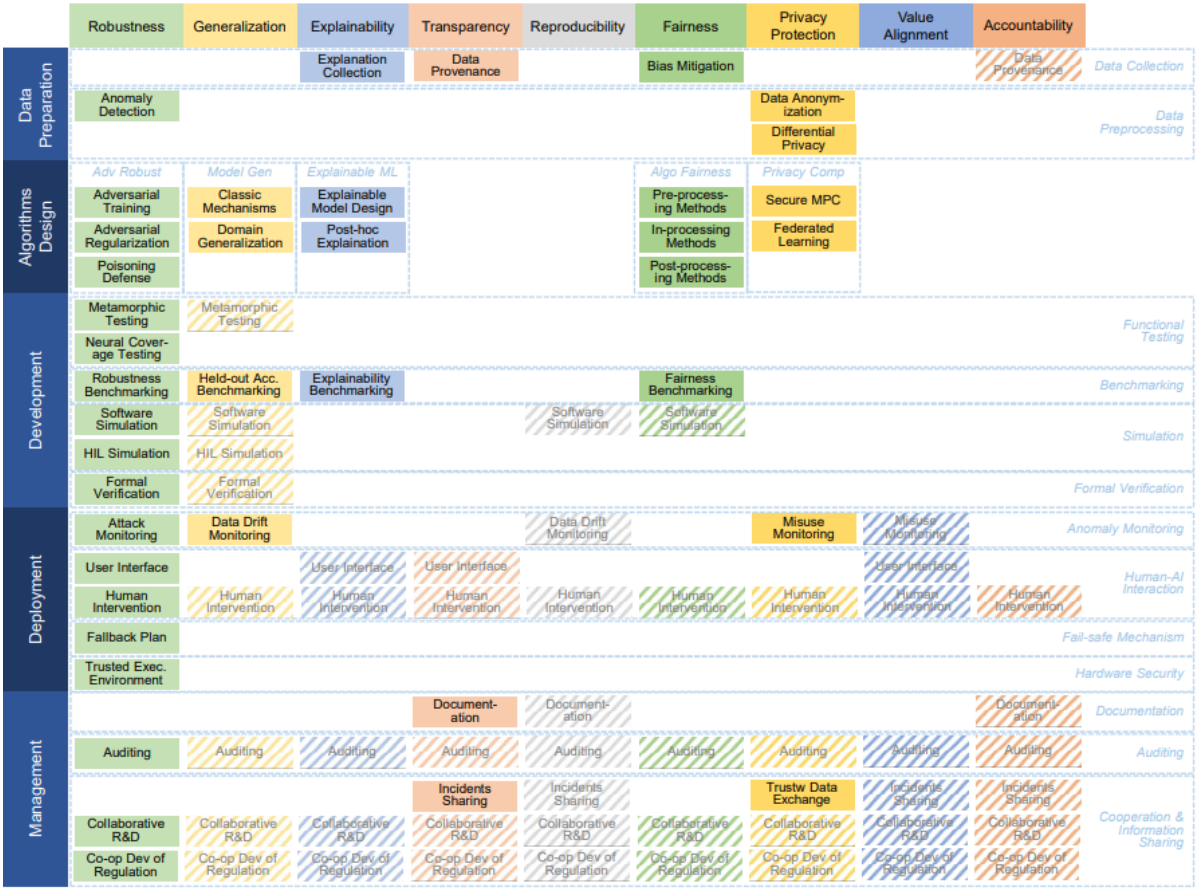


Figure 1 The trustworthy AI aspects and practices organised across the different stages of the life cycle of an AI system as extracted by (Li et. al, 2023).

The review paper **"Guidelines and Quality Criteria for Artificial Intelligence-Based Prediction Models in Healthcare: A Scoping Review" (de Hond et. al, 2022)** is the only work in the series of the selected papers presented in this section that focuses on the field of AI in healthcare. Specifically, it conducts a comprehensive scoping review on guidelines and quality criteria applicable to artificial intelligence (AI)-based prediction models in healthcare. The key points from the work are:

- The actionable guidelines and quality criteria identified in this work are summarised across the **six main phases of the AI prediction model (AIPM) construction**, i.e.,

1) ***Preparation, collection and checking the data***, 2) ***Development of an AIPM***, 3) ***Validation of an AIPM***, 4) ***Development of the software application***, 5) ***Impact assessment of the AIPM with software development***, and 6) ***Implementation and use in daily healthcare practice***.

- Phase 1. ***Preparation, collection, and checking of the data***: This phase involves a series of preparatory activities that should take place before the AIPM development. The phase includes:
  - Clearly define the medical problem and context.
  - Take measures to ensure patient privacy, e.g., GDPR compliance.
  - Pre-specify and justify the sample size for the intended purpose.
  - Ensure representativeness, i.e., the real-world heterogeneity and diversity of the target population and the intended healthcare setting are sufficiently covered.
  - Ensure data quality via extensive assessment.
  - Data pre-processing, i.e., data augmentation, removing outliers, re-coding or transforming variables, standardization, and imputation of missing data.
  - Data standardization to facilitate interoperability and adoption in healthcare settings. Examples include SNOMED CT[3], ICD-10[4] and OPCS4[5].
- Phase 2. ***Development of the AIPM***: This phase involves the different tasks and practices should take place to produce an AIPM. This phase includes:
  - Model selection and interpretability.
  - Train the AIPM.
  - Internally validate the AIPM.
  - Measures to reduce risk of overfitting.
  - Measures to identify and prevent algorithmic bias.
  - Ensure the transparency of the modelling process via reporting all different aspects of the final AIPM, i.e., data, model architecture, configurations, and training scripts.
- Phase 3. ***Validation of the AIPM***: This phase involves the external evaluation of the AIPM to ensure the ***target performance*** and ***generalisability***.
- Phase 4. ***Development of the software application***: This phase involves the practices should be followed for the development of the "host" software application where the AIPM will be integrated. This phase includes:
  - Ensure interoperability to allow successful integration in existing digital infrastructure of hospitals and clinical care centres via following industry standards, e.g., ISO/IEC JTC 1/SC 42[6], IEEE 7000-2021[7], HL7 FHIR[8] and ISO/IEEE 11073-10418:2014[9].
  - Emphasise on human-AI interaction to ensure adoption, and effective and safe use in daily healthcare practice.
  - Facilitate software updating by not disrupting the relationship with the end-users, i.e., notifying them about the changes as well as allowing them to roll back to previous versions.

[3] https://www.snomed.org/
[4] https://icd.who.int/browse10/2019/en
[5] https://www.datadictionary.nhs.uk/supporting_information/opcs_classification_of_interventions_and_procedures.html
[6] https://www.iso.org/committee/6794475.html
[7] https://standards.ieee.org/ieee/7000/6781/
[8] https://hl7.org/fhir/
[9] https://www.iso.org/standard/61897.html

- Facilitate monitoring and auditing mechanisms to continuously trace and evaluate performance.
- Ensure security via applying cybersecurity best practices and standards.
- Test the software application using industry best practices and standards.

- Phase 5. ***Impact assessment of the AIPM with software***: This phase involves a series of studies that should take place to assess the final product. These studies are:
  - A feasibility study is performed to determine the clinical benefit of the AIPM for the intended healthcare setting.
  - A risk management study is performed to identify potential sources of risk, extreme situations, failures, accidental misuse, and manipulation of the AIPM as well as to specify monitoring, reporting and mitigation measures of the identified risks.
  - Impact study is performed where the effects on clinical outcomes and decision making are compared for a group exposed to the predictions of the AI versus a non-exposed control group receiving standard care. It is suggested that findings are communicated in standardised manner such as CONSORT (Liu et. al, 2020) and SPIRIT guidelines (Rivera et. al, 2020).
- Phase 6. ***Implementation and use in daily healthcare practice***: This phase involves the deployment of the produced system in the operational environment and the post deployment activities that should be continuously conducted.
  - Clinical implementation consists of all the steps that are necessary to deploy the AIPM in the healthcare environment outside of the clinical trial setting.
  - Maintenance and updating should follow practices to ensure improved predictive performance. It should be noticed that updating the AIPM may require recertification. Nevertheless, guidelines for updating AIPM without recertification, i.e., a change control plan, is proposed by USA Food and Drug Administration (Food and Drug Administration, 2019).
  - Education involves the training of end-users in the correct use of the AIPM.
  - Monitoring and auditing refer to the evaluation of AIPM throughout its lifecycle after the initial deployment.

2.2 High-level trustworthy AI frameworks

The aim of this section is to present the main high-level frameworks for TAI. The cornerstone document for EU and the foundation for the iPROLEPSIS TAI framework is the European Commission's "Ethics guidelines for trustworthy AI". Nevertheless, other frameworks proposed by non-EU leading entities, such as the Institute of Electrical and Electronics Engineers (IEEE), were searched and are presented to provide a complete outlook of the field.

2.2.1 Ethics guidelines for trustworthy AI by European Commission

The "Ethics Guidelines for Trustworthy AI by the European Commission"[10] refers to a set of guidelines outlined by the European Commission aimed at ensuring the development and deployment of AI that is lawful, ethical, and respects fundamental rights. These guidelines were established to provide a framework for creating AI systems that are trustworthy, accountable, and aligned with human values.

Released in April 2019 as part of the European Union's broader AI strategy, these guidelines were developed by the High-Level Expert Group on AI (AI HLEG), consisting of experts from

[10] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

various fields including academia, industry, and civil society. The document outlines seven key requirements for AI systems to be considered trustworthy:

1. **Human agency and oversight**: AI systems should support human autonomy and decision-making while allowing humans to intervene or oversee AI-generated outcomes.
2. **Technical robustness and safety**: AI systems should be reliable, secure, and resilient, ensuring their safety throughout their lifecycle.
3. **Privacy and data governance**: AI systems should respect privacy and data protection principles, ensuring transparency and providing control over personal data.
4. **Transparency**: The operation of AI systems should be transparent, allowing for traceability and providing explanations for their decisions and actions.
5. **Diversity, non-discrimination, and fairness**: AI systems should be developed and deployed in a way that ensures fairness, prevents discrimination, and promotes diversity and inclusivity.
6. **Societal and environmental well-being**: AI systems should be designed to enhance societal well-being, sustainability, and environmental responsibility.
7. **Accountability**: Mechanisms should be in place to ensure accountability for AI systems and their outcomes, including clear responsibilities and redress mechanisms.

2.2.2 The Responsible Machine Learning Principles by the Institute for Ethical AI & ML

The Institute for Ethical AI & ML[11] has established a set of "Responsible Machine Learning Principles" aimed at promoting ethical practices in the development and deployment of AI technologies. Specifically, eight principles were formulated to guide organizations, researchers, and developers in creating and implementing AI systems that are ethically sound and socially responsible. These principles covered various aspects of TAI, including:

1. **Human augmentation**: To assess the impact of incorrect predictions, as well as the developed AI systems should be reviewed by subject-domain-experts (human-in-the-loop review).
2. **Bias evaluation**: To build processes and methods to identify, document and monitor the inherent bias in the data, features, and inference results, as well as the subsequent implications of this bias.
3. **Explainability by justification**: To develop tools and processes to continuously improve transparency and explainability of AI models, where reasonable.
4. **Reproducible operations**: To develop the infrastructure required to allow for a reasonable level of reproducibility for the developed AI model.
5. **Displacement strategy**: To assess and develop mitigation plan concerning the risk for the workers of a business to be displaced due to AI related automation.
6. **Practical accuracy**: To develop processes to ensure the indented accuracy.
7. **Trust by privacy:** To build and communicate processes to protect and handle data with stakeholders who interact with the system either directly and/or indirectly.
8. **Data risk awareness:** To develop and continuously evolve processes and infrastructure to ensure data and model security.

These eight principles aimed to provide a foundational framework for organizations and practitioners working with AI/ML technologies, supporting them when designing, developing or maintaining systems that learn from data.

[11] https://ethical.institute/

2.2.3 IEEE's Ethically Aligned Design

IEEE's Ethically Aligned Design (IEEE, 2017) is a document that serves as a set of guidelines and recommendations for the ethical design and development of autonomous and intelligent systems. It was created by the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, which consists of experts from various disciplines, including technology, ethics, policy, and law. The goal of Ethically Aligned Design is to encourage the development and deployment of autonomous and intelligent systems that prioritize human well-being and align with ethical principles. The document provides extensive guidance across multiple domains and addresses a wide range of ethical considerations associated with the design, deployment, and use of these systems. The key principles on which the framework is based are:

1. **Human Rights**: To ensure that AI does not infringe on internationally recognised human rights. It is fundamental that people have the right to define access and provide informed consent with respect to the use of their personal digital data.
2. **Well-being**: To prioritize metrics of well-being when design and use AI systems. The latter can alter institutions and institutional relationships toward more human-centric structures resulting in increased individual and societal well-being.
3. **Accountability**: To ensure that AI designers and operators are responsible and accountable. AI systems should be subject to the applicable regimes of law.
4. **Transparency**: To ensure AI operates in a transparent manner. The logic and rules embedded in an AI system must be accessible to overseers, as well as audit trails should be generated including to support law decisions and allow third party verification.
5. **Awareness of misuse**: To minimize the risks of AI misuse via:
   a. educating the public on societal impacts of related technologies,
   b. attaining research and development leadership,
   c. supporting and promoting internationally recognised legal norms,
   d. developing workforce expertise in related technologies. and
   e. promoting regulations that ensure public safety and responsibility.

The IEEE Ethically Aligned Design aims to provide a comprehensive framework that promotes the ethical and responsible design, development, and deployment of autonomous and intelligent systems. It serves as a resource for technologists, policymakers, industry professionals, and other stakeholders to ensure that these systems are developed and used in ways that prioritize ethical considerations and societal well-being.

2.2.4 Montréal Declaration for a responsible development of artificial intelligence

The “Montréal Declaration for the Responsible Development of Artificial Intelligence”[12] is a significant document signed by various stakeholders, including AI researchers, policymakers, and industry leaders. This declaration aimed to establish ethical guidelines and principles for the responsible development and deployment of AI. The Montréal Declaration proposes ten key principles, namely:

1. **Principle of well-being**: The development and use of AI systems must permit the growth of the well-being of all sentient beings.
2. **Principle of respect for autonomy**: An AI system must be developed and used with respect for the autonomy of individuals and with the goal of increasing individual’s control over their live and their environment.
3. **Principle of protection of privacy and intimacy**: Privacy and intimacy must be protected from AI intrusion and data acquisition and archiving systems.

[12] https://montrealdeclaration-responsibleai.com/the-declaration/

4. **Principle of solidarity**: The development of an AI system must be compatible with maintaining the bonds of solidarity among people and generations.
5. **Principle of democratic participation**: An AI system must meet intelligibility, justifiability, and accessibility criteria, and must be subjected to democratic scrutiny, debate, and control.
6. **Principle of equity**: The development and use of an AI system must contribute to the creation of a just and equitable society.
7. **Principle of diversity inclusion**: The development and use of an AI system must be compatible with maintaining social and cultural diversity and must not restrict the scope of lifestyle choices or personal experiences.
8. **Principle of prudence**: Every person involved in AI development must exercise caution by anticipating, as far as possible, the adverse consequences of an AI system use and by taking the appropriate measures to avoid them.
9. **Principle of responsibility**: The development and use of an AI system must not contribute to lessening the responsibility of human beings when decisions are made.
10. **Principle of sustainable development**: The development and use of an AI system must be carried out to ensure a strong environmental sustainability of the planet.

The Montréal Declaration aims to develop an ethical framework for the development and deployment of AI, to guide the digital transition so everyone benefits from this technological revolution and open a national and international forum for discussion to collectively achieve equitable, inclusive, and ecologically sustainable AI development.

2.2.5 Summary of the high-level trustworthy AI frameworks

Four significant high-level TAI frameworks from key global organisations were presented showing the points of emphasis concerning TAI. Regardless the differences in the aspect organisation and terminology, the summarised common aspects include 1) the reliable and fair operation in all circumstances, i.e., ensuring technical robustness and preventing discrimination, 2) the establishment of communication mechanisms with the relative stakeholders, i.e., explaining the AI outcomes, notifying of misuse, and allowing auditing, and 3) the guarantee of social responsibility, i.e., enhancing well-being and protecting human rights.

3 iPROLEPSIS trustworthy AI methodology

The analysis of the state-of-the-art of TAI gave us great insights about the global ethical requirements associated with the design, the development, and the usage of an AI system, as well as the practical challenges faced by industry. It is worth noting that although different categorisation and terminology are used in different high-level frameworks, there is significant overlap on the main concepts. Specifically, the need for robustness, transparency, fairness, privacy, security, well-being, and accountability appears in different forms in every high-level analysed framework. Furthermore, the collection of recently selected review papers has been instrumental in providing insights into the practical aspects of what TAI is. The main outcome is that TAI does not prescribe a specific recipe that can be universally applied to any AI project. On the contrary, every AI project should adapt TAI principles to its objectives, scale, and field of application. To this end, a novel methodology was implemented to extract and prioritise TAI-related needs of iPROLEPSIS project. In brief, this novel methodology consists of four steps:

1. the extraction of a full set of TAI recommendations,

2. the prioritisation of the recommendations based on the responses from a multidisciplinary group of experts, i.e., people from the organisations of iPROLEPSIS consortium,
3. the design of the iPROLEPSIS TAI framework as a self-assessment checklist and corresponding evaluation metrics, and
4. the definition of the implementation strategy of the framework during the project's lifecycle.

3.1 The "untrustworthy" AI system

Firstly, the objective is to aggregate all possible recommendations that an ideal AI project should implement in order to fully satisfy all TAI high-level requirements/principles. Such a list of recommendations would consist of an excellent starting point to develop a TAI framework by either narrowing down and/or modifying these recommendations. The challenge is how such a list of recommendations can be collected? Instead of searching for all properties that an ideal, in terms of trustworthiness, AI system have, we inverse the problem and search for what misses from a completely "untrustworthy" AI system.

To extract the list of recommendations for this system, the ***Assessment List for Trustworthy Artificial Intelligence*** (ALTAI) (European Commission, 2020) is used. ALTAI was developed by the High-Level Expert Group on Artificial Intelligence set up by the European Commission to create a tool to assess whether the AI system that is being developed, deployed, procured, or used, complies with European Commission's Ethics Guidelines for Trustworthy AI. Specifically, a prototype web-based tool[13] that implements ALTAI was used. This tool contains all questions included in ALTAI and finally, when the questionnaire is completed, it provides a set of recommendations based on the responses. Thus, to define the "untrustworthy" AI system we negatively responded to all questions and 75 recommendations were resulted. These recommendations are what is needed for the "untrustworthy" AI system to be fully complaint with European Commission's Ethics Guidelines for Trustworthy AI. The full list of these 75 recommendations is presented in Appendix I. Also, needless to mention that both ALTAI questions and the full list of recommendations do not only support the European Commission's Guidelines for Trustworthy AI, but also cover in large the remaining high-level frameworks, i.e., the IEEE's Ethically Aligned Design, the Montréal Declaration for a responsible development of artificial intelligence and the Responsible Machine Learning Principles by the Institute for Ethical AI & ML.

3.2 Design of iPROLEPSIS specific ALTAI

The next step in designing the iPROLEPSIS TAI framework is to prioritise the extracted TAI recommendations. The purpose of the TAI framework is to support the iPROLEPSIS project to achieve its current contractual objectives, as well as its mid- and long-term impact expectations. Thus, the prioritisation aims at directing the focus and resources to the most crucial TAI concepts. Given that there is no existing theory that offers a method to adapt high-level TAI principles to specific AI needs, we approached this issue via asking a multidisciplinary cohort of experts for their perspectives. Specifically, experts from the organisations that belong to the iPROLEPSIS consortium, i.e., software engineers, data scientists, clinicians, project managers, and legal experts, were asked to assess the 75 extracted recommendation using the "Must-have, Should-have, Could-have, and Won't-have, or Will not have right now or Wish-have" (MoSCoW) prioritisation technique (Clegg & Barker, 1994).

[13] https://altai.insight-centre.org/

A questionnaire was set up using Google Forms[14] and the questions were organised in the seven categories of ALTAI, i.e., 1) human agency and oversight, 2) technical robustness and safety, 3) privacy and data governance, 4) diversity, non-discrimination and fairness, 6) societal and environmental well-being, and 7) accountability. Before the participants start responding to the questions, it is important for them to become familiarized with TAI. Thus, definitions for each ALTAI concepts were provided in the respective sections. Moreover, a few extraordinary technical aspects were also elaborated to aid non-technology experts to figure out their purpose and value. Nonetheless, if the responder did not feel confident or was unwilling to answer a specific recommendation, there was the option of "Cannot respond". The estimated time for completing the questionnaire was 30 minutes and it was anonymous. The responders were only asked to provide their organisation and their professional role among three options, namely:

- technology-related role: software developer/engineer, data scientist/engineer, AI researcher, technical project manager,
- healthcare-related role: clinician, healthcare researcher, clinical project manager, and
- humanities scientist.

Indicative screenshots from the pages of the questionnaire can be found in Appendix II.

The questionnaire was answered by 14 experts, six of them have a technology-related role, seven a healthcare-related role and one was humanities scientist. The responses were aggregated and analysed. The results of the analysis are shown in Figures **2**-**8** as heatmap tables, where each cell represents the percentage of the respective MoSCoW prioritisation (column) for the specific recommendation (row). Observing these results, a few general intriguing outcomes can be extracted:

1. There is a consensus among the experts that all "Privacy and Data Governance" recommendations ***must*** be satisfied in the iPROLEPSIS project. Moreover, regarding the "controversy" between privacy and transparency that has been found in state-of-the-art, privacy is considered more important for the iPROLEPSIS project.
2. The "Accountability" category consists of only a couple of recommendations that ***must*** be satisfied. Although accountability is considered the cornerstone of TAI in several publications, interestingly this argument cannot be confirmed in iPROLEPSIS case. The reason might be that the accountability aspects are beyond the scope of the project, as they concern the usage and management of an AI system in the operational environment, but iPROLEPSIS is a research and innovation action that focuses on developing and validating digital solutions for aiding the management and understanding of a complex disease, i.e. psoriatic arthritis.
3. The "Societal and Environmental Well-being" category consists of recommendations, the dominant priorities of which are either ***could*** or ***won't***, so this set of recommendations is not relevant to iPROLEPSIS.

Finally, the results from the analysis of the questionnaire responses provide solid basis for structuring the TAI framework of iPROLEPSIS project. Nonetheless, if the approach is applied to a greater scale better insights will be revealed. Particularly, if more responses are collected by experts who work in similar project, as well as other relevant stakeholder, e.g. patients, a general TAI framework for digital solutions in healthcare will be able to be designed.

[14] https://www.google.com/intl/en/forms/about/

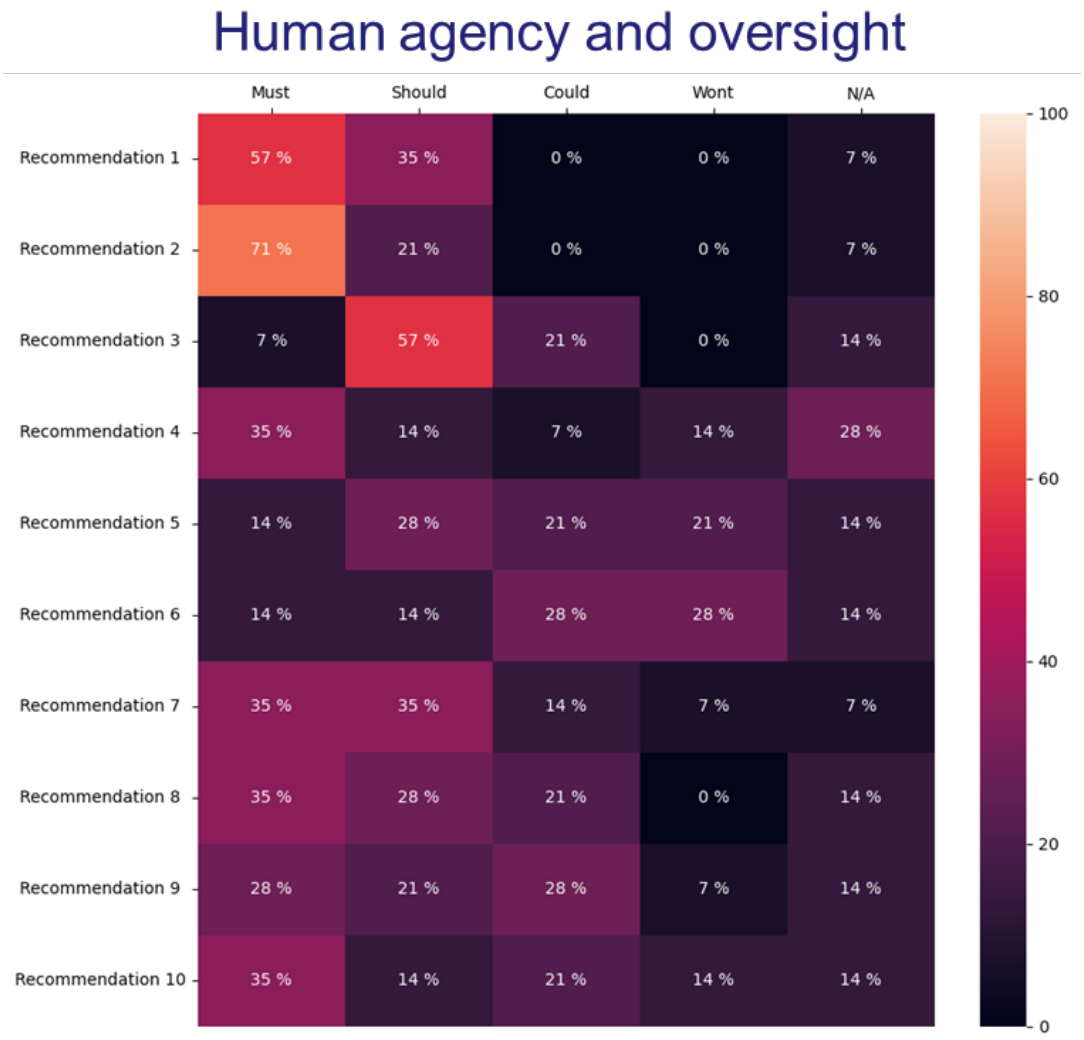


Figure 2 Questionnaire results for “Human Agency and Oversight” TAI category.

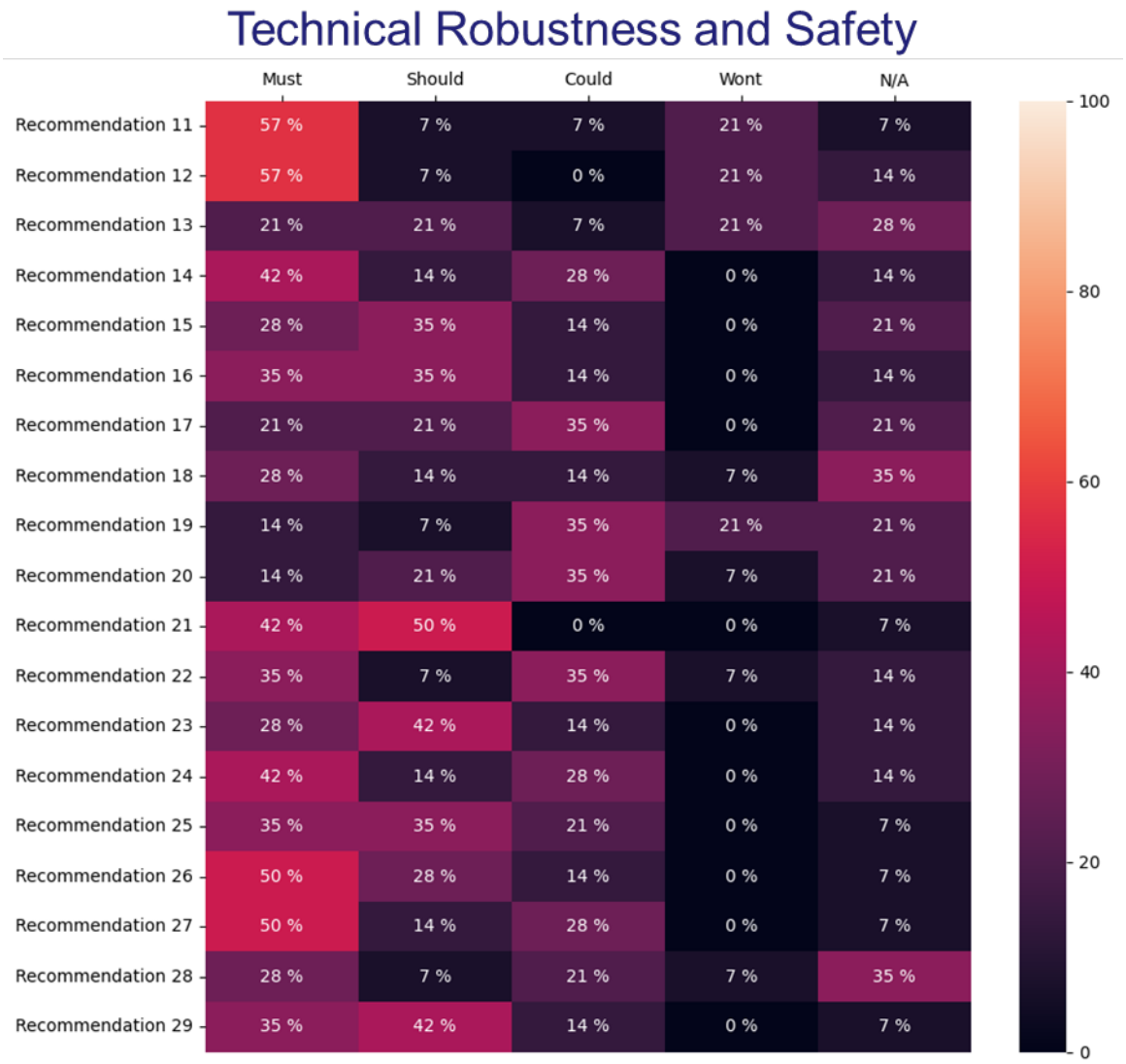


Figure 3 Questionnaire results for “Technical Robustness and Safety” TAI category.

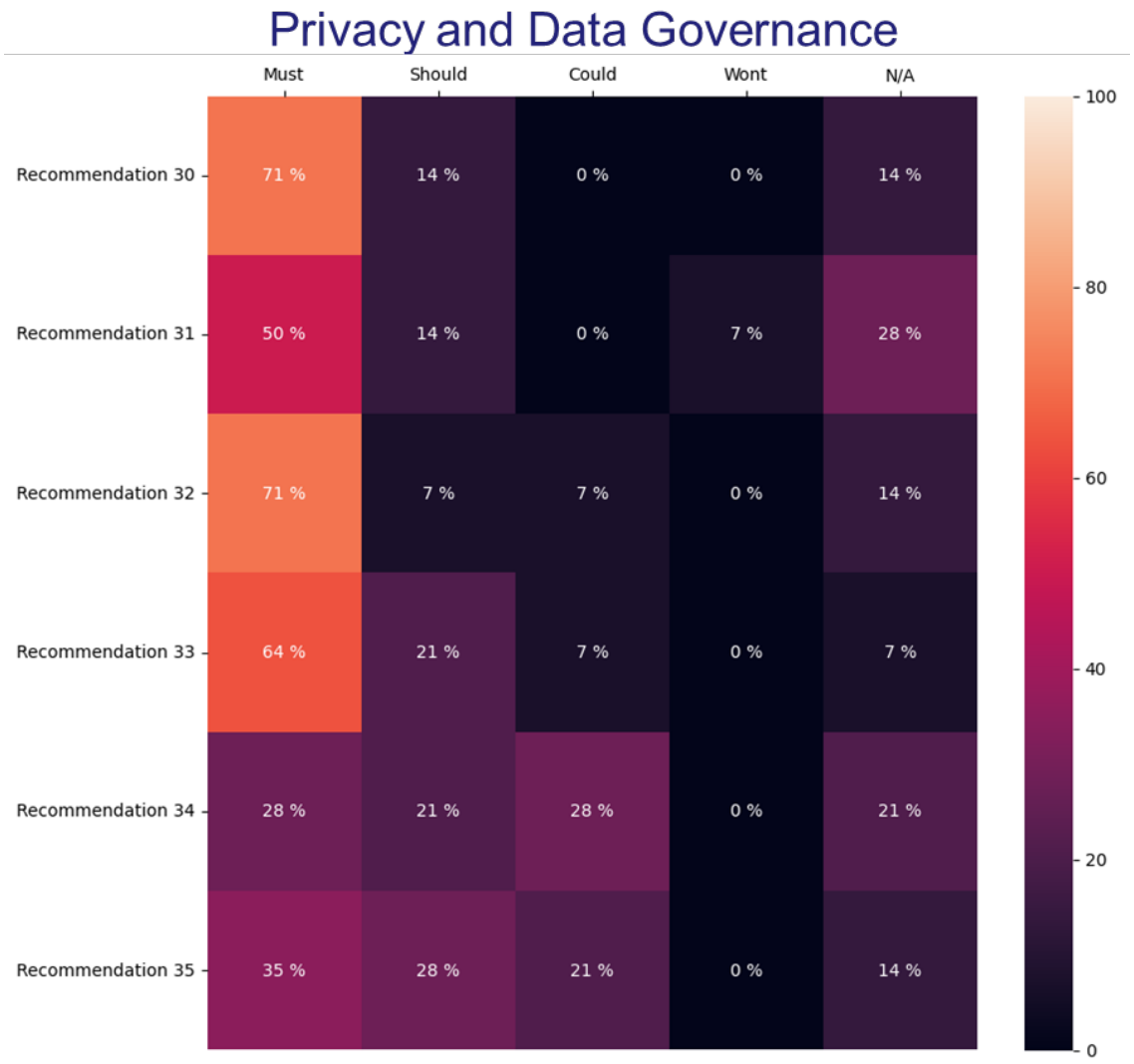


Figure 4 Questionnaire results for “Privacy and Data Governance” TAI category.

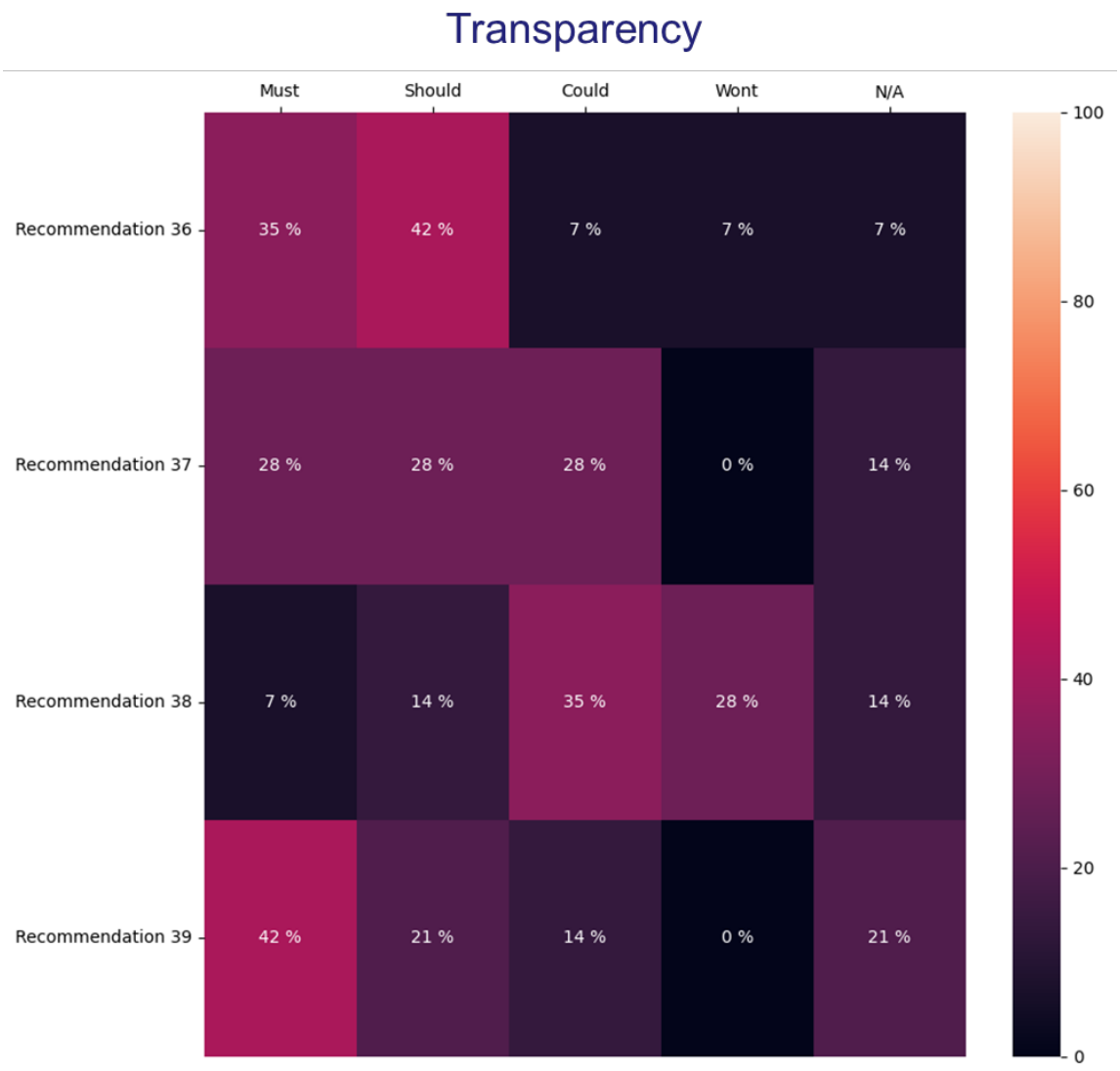


Figure 5 Questionnaire results for “Transparency” TAI category.

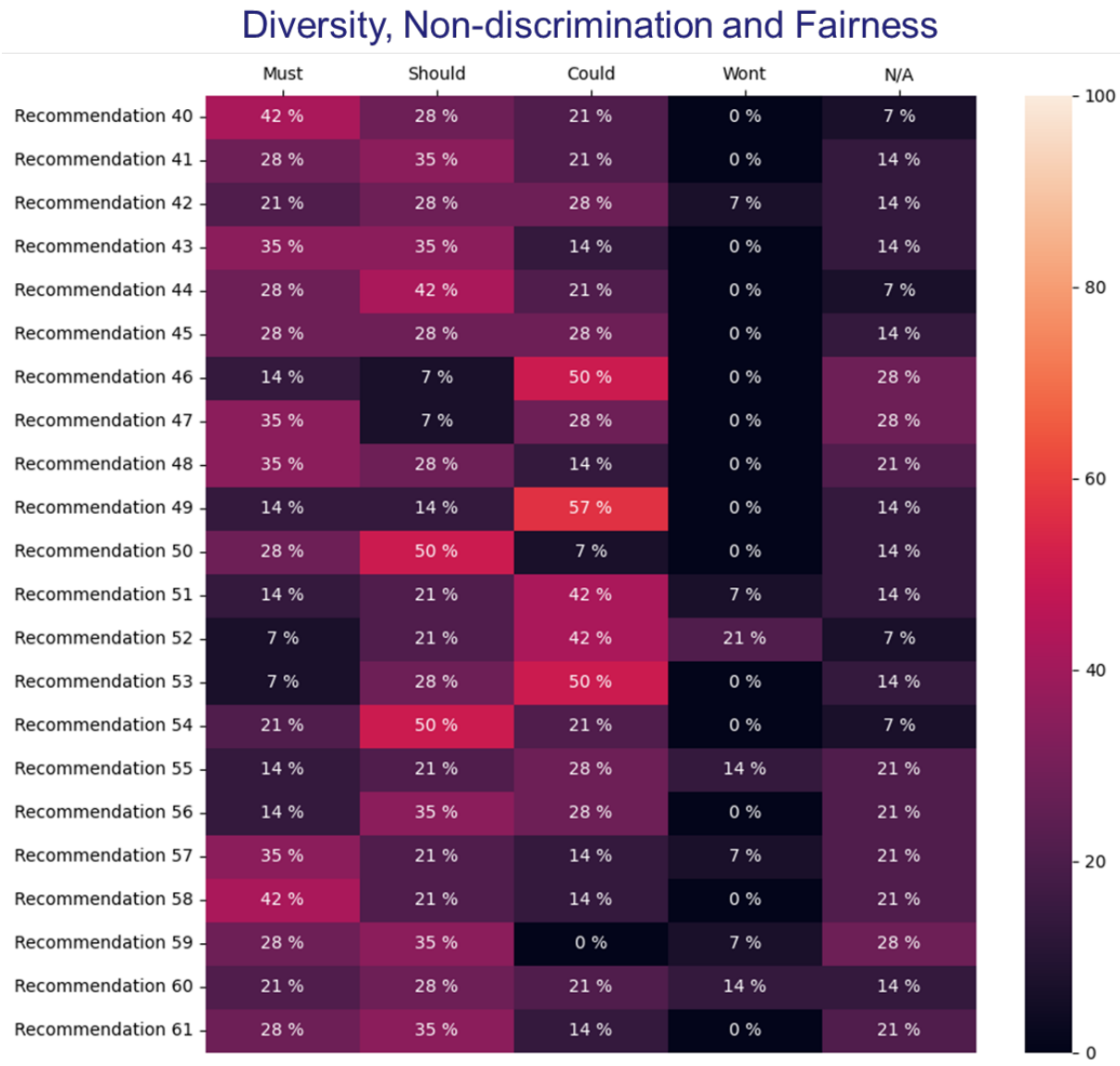


Figure 6 Questionnaire results for “Diversity, Non-discrimination and Fairness” TAI category.

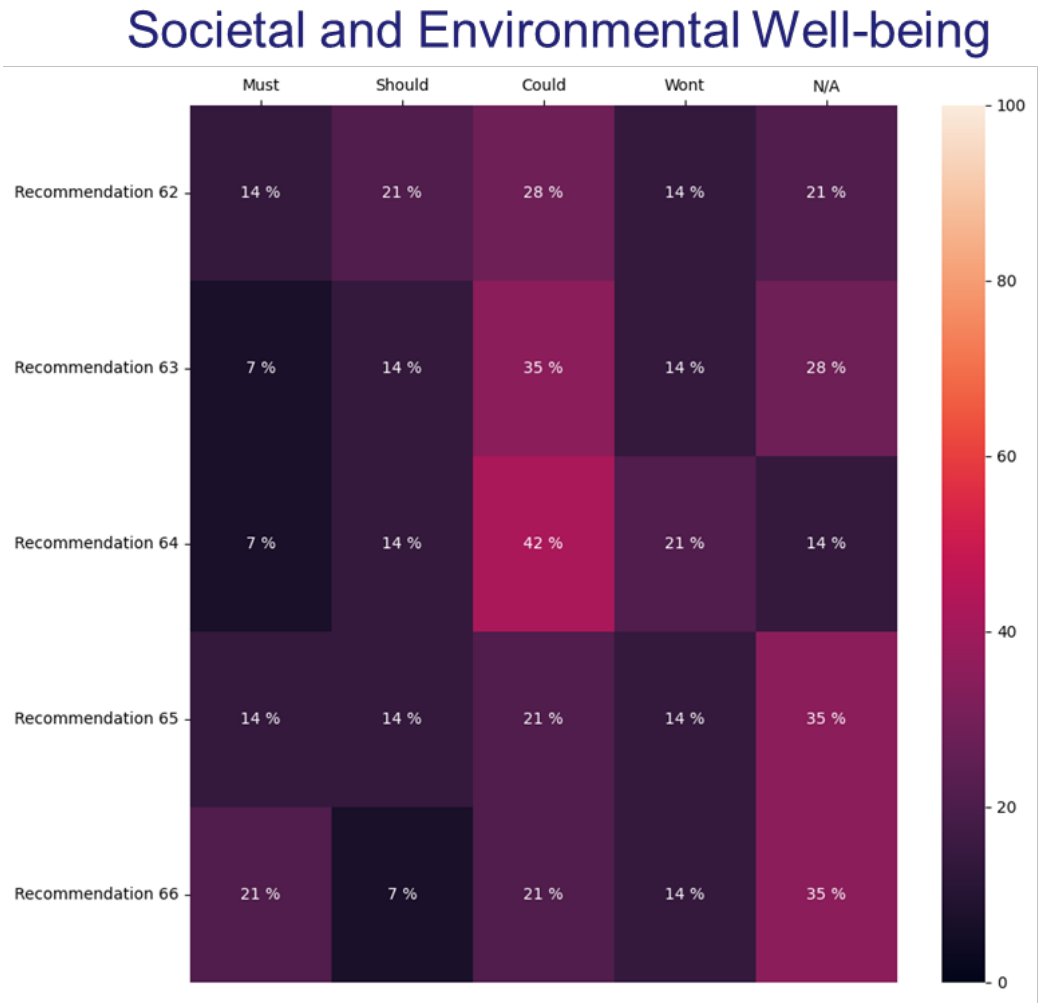


Figure 7 Questionnaire results for “Social and Environmental Well-being” TAI category.

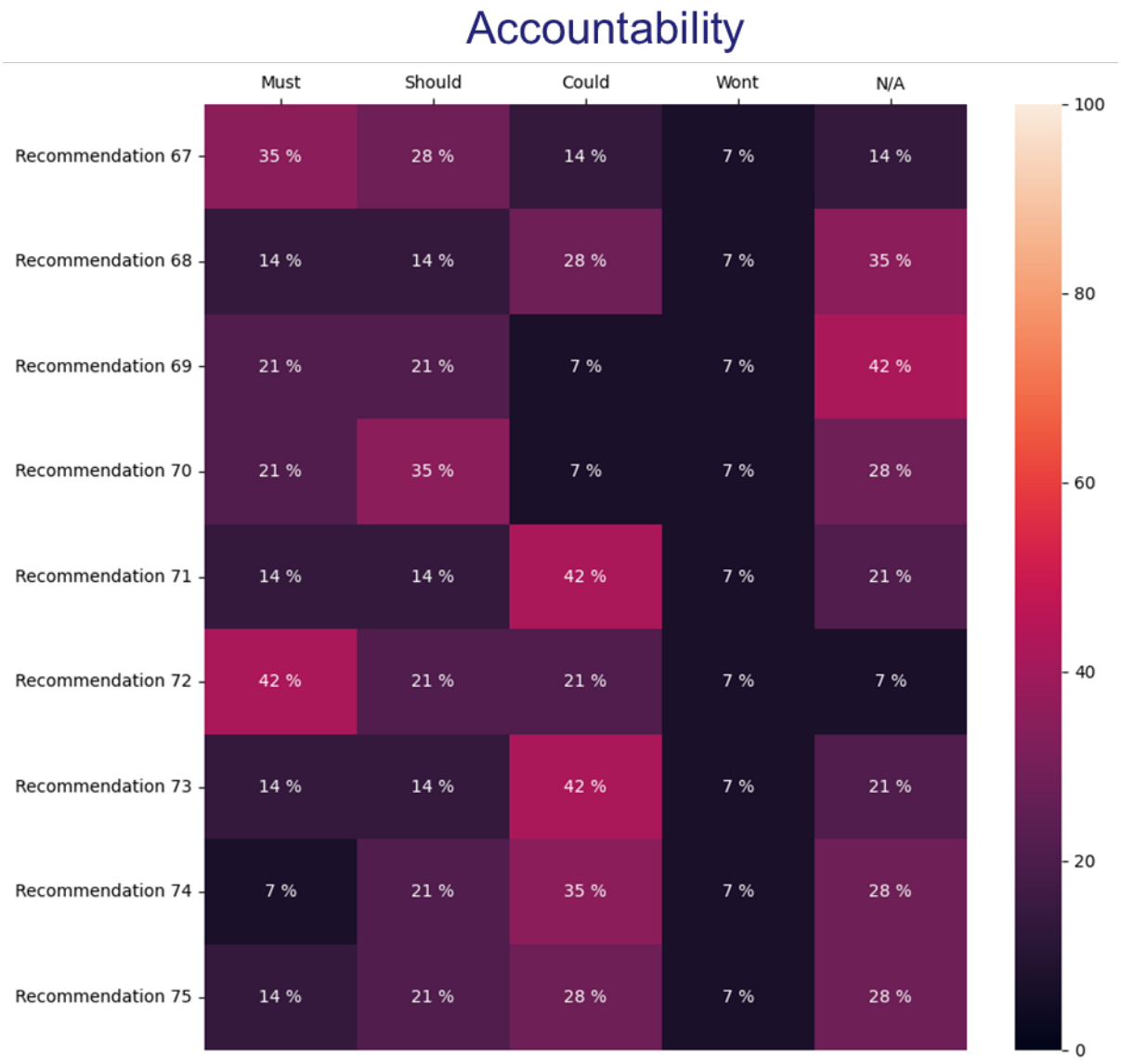


Figure 8 Questionnaire results for "Accountability" TAI category.

3.3 The iPROLEPSIS trustworthy AI framework

In this section, the iPROLEPSIS TAI framework is defined. The TAI framework consists of a set of TAI-related requirements with the respective priority and a method to quantify the degree of compliance. The TAI-related requirements are the recommendations described in the previous section with the priority to be set as the most dominant one from the questionnaire results. If there are more than one priority with the same percentage for a requirement, the most critical one is selected, i.e., between "must have" and "should have" the former will be selected. The final set of requirements organised by priority and TAI category are presented in Tables **2**-**4**. Also, a unique ID is attributed to every requirement for reference purposes. The ID is structured using a TAI category short name[15] at the prefix and the first letter of priority type at the suffix.

[15] HUM: "Human Agency and Oversight"; TEC: "Technical Robustness and Safety"; PRI: "Privacy and Data Governance"; TRA: "Transparence"; DIV: "Diversity, Non-discrimination and Fairness"; SOC: "Societal and Environmental Well-being"; ACC: "Accountability"

Table 2 "Must do" requirements of TAI iPROLEPSIS framework.

ID	Must have
HUM01M	Incorporate a process where end-users and/or subjects are adequately made aware that an AI-system influenced the decision, content, advice or outcome.
HUM02M	Ensure that the end-users or subjects are adequately informed that they are interacting with an AI system.
HUM03M	Put in place any procedure to avoid that the system inadvertently affects human autonomy.
HUM04M	Take measures to mitigate the risk of manipulation, including providing clear information about ownership and aims of the system, avoiding unjustified surveillance, and preserving autonomy and mental health of users.
HUM05M	Give specific training to humans (human-in-the-loop, human-on-the-loop, human-in-command) on how to exercise oversight.
HUM06M	Establish detection and response mechanisms in case the AI system generates undesirable adverse effects for the end-user or subject.
HUM07M	Deploy a "stop button" or procedure to safely abort an operation when needed.
TEC01M	Put in place measures to ensure the integrity, robustness and overall security of the AI system against potential attacks over its lifecycle.
TEC02M	Inform users as soon as possible if some new threats are detected.
TEC03M	Identify the possible threats to the AI system (design faults, technical faults, environmental threats) and the possible resulting consequences.
TEC04M	Assess the dependency of critical system's decisions on its stable and reliable behaviour.
TEC05M	Consider whether the AI system's operation can invalidate the data or assumptions it was trained on, and how this might lead to adversarial effects (e.g., biased estimators, echo chambers).
TEC06M	Put in place a well-defined process to monitor if the AI system is meeting the goals of the intended applications.
TEC07M	Test whether specific contexts or conditions need to be taken into account to ensure reproducibility.
TEC08M	Put in place verification and validation methods and documentation (e.g., logging) to evaluate and ensure different aspects of the system's reliability and reproducibility.
TEC09M	Clearly document and operationalize processes for the testing and verification of the reliability and reproducibility of the AI system.
PRI01M	Take measures to consider the impact of the AI system on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection.
PRI02M	Consider establishing mechanisms that allow flagging issues related to privacy or data protection concerning the AI system.

ID	Must have
PRI03M	When relevant, implement the right to withdraw consent, the right to object and the right to be forgotten in the AI system.
PRI04M	Consider the privacy and data protection implications of data collected, generated or processed over the course of the AI system's lifecycle.
PRI05M	Consider the privacy and data protection implications of the AI system's non-personal training-data or other processed non-personal data.
PRI06M	Whenever possible and relevant, align the AI-system with relevant standards (e.g., ISO, IEEE) or widely adopted protocols for (daily) data management and governance.
TRA01M	Consider explaining the decision adopted or suggested by the AI system to its end users.
TRA02M	In case of interactive AI system, consider communicating to users that they are interacting with a machine.
DIV01M	Consider establishing a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design.
DIV02M	Research and use publicly available technical tools, that are state-of-the-art, to improve your understanding of the data, model and performance.
DIV03M	Consider diversity and representativeness of end-users and or subjects in the data.
DIV04M	Depending on the use case, ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system.
DIV05M	You should establish clear steps and ways of communicating on how and to whom such issues can be raised.
DIV06M	You should ensure that Universal Design principles are taken into account during every step of the planning and development process, if applicable.
DIV07M	You should take the impact of the AI system on the potential end-users and/or subjects into account.
ACC01M	Designing a system in a way that can be audited later, results in a more modular and robust system architecture. Thus, it is highly recommended to ensure modularity, traceability of the control and data flow and suitable logging mechanisms.
ACC02M	If AI systems are increasingly used for decision support or for taking decisions themselves, it has to be made sure these systems are fair in their impact on people's lives, that they are in line with values that should not be compromised and able to act accordingly, and that suitable accountability processes can ensure this. Consequently, all conflicts of values, or trade-offs should be well documented and explained

Table 3 "Should do" requirements of TAI iPROLEPSIS framework.

ID	Should have
HUM01S	Put in place procedures to avoid that end users over-rely on the AI system.
HUM02S	Take measures to deal with the possible negative consequences for end-users or subjects in case they develop attachment. In particular, provide means for the user to have control of the interactions.
TEC01S	Define risk, risk metrics and risk levels of the AI system in each specific use case.
TEC02S	Put in place a series of steps to monitor and document the AI system's accuracy.
TEC03S	Put in place processes to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated.
TEC04S	Put in place a proper procedure for handling the cases where the AI system yields results with a low confidence score.
TRA01S	Consider adopting measures to continuously assess the quality of the input data to the AI system.
DIV01S	Consider diversity and representativeness of end-users and/or subjects in the data.
DIV02S	Test for specific target groups or problematic use cases.
DIV03S	Assess and put in place processes to test and monitor for potential biases during the entire lifecycle of the AI system (e.g., biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness).
DIV04S	Your definition of fairness should be commonly used and should be implemented in any phase of the process of setting up the AI system.
DIV05S	Establish mechanisms to ensure fairness in your AI system.
DIV06S	You should assess whether the AI system's user interface is usable by those with special needs or disabilities or those at risk of exclusion.
DIV07S	You should assess whether the team involved in building the AI system engaged with the possible target end-users and/or subjects.
DIV08S	You should assess whether there could be groups who might be disproportionately affected by the outcomes of the system.
DIV09S	You should assess the risk of the possible unfairness of the system onto the end-user's or subject's communities.
ACC01S	AI systems should be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm. Consequently, developers and deployers should receive appropriate training about the legal framework that applies for the deployed systems.

Table 4 "Could do" requirements of TAI iPROLEPSIS framework.

ID	Could have
HUM01C	Take measures to minimize the risk of addiction by involving experts from other disciplines such as psychology and social work.
TEC01C	Assess the risk of possible malicious use, misuse or inappropriate use of the AI system.
TEC02C	Plan fault tolerance via, e.g., a duplicated system or another parallel system (AI-based or "conventional").
TEC03C	Develop a mechanism to evaluate when the AI system has been changed enough to merit a new review of its technical robustness and safety. Develop a mechanism to evaluate when the AI system has been changed enough to merit a new review of its technical robustness and safety.
TRA01C	Consider continuously surveying the users to ask them whether they understand the decision(s) of the AI system.
DIV01C	Put in place educational and awareness initiatives to help AI designers and AI developers be more aware of the possible bias they can inject in designing and developing the AI system.
DIV02C	Identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the (end)-users.
DIV03C	Consider other definitions of fairness before choosing one.
DIV04C	Consult with the impacted communities about the correct definition of fairness, such as representatives of elderly persons or persons with disabilities.
DIV05C	Ensure a quantitative analysis or metrics to measure and test the applied definition of fairness.
DIV06C	You should ensure that the AI system corresponds to the variety of preferences and abilities in society.
SOC01C	Consider the potential positive and negative impacts of your AI system on the environment and establish mechanisms to evaluate this impact.
SOC02C	Define measures to reduce the environmental impact of your AI system's lifecycle and participate in competitions for the development of AI solutions that tackle this problem.
SOC03C	Inform and consult with the impacted workers and their representatives but also involve other stakeholders. Implement communication, education, and training at operational and management level.
ACC01C	A useful non-technical method to ensure the implementation of trustworthy AI is to include various stakeholders, e.g. assembled in an "ethical review board" to monitor and assist the development process.
ACC02C	Involving third parties to report on vulnerabilities and risks does help to identify and mitigate potential pitfalls.
ACC03C	A risk management process should always include new findings since initial assumptions about the likelihood of occurrence for a specific risk might be faulty and thus, the quantitative risk analysis was not correct and should be revised with the new findings.

ID	Could have
ACC04C	Acknowledging that redress is needed when incorrect predictions can cause adverse impacts to individuals is key to ensure trust. Particular attention should be paid to vulnerable persons or groups.

To quantify the degree of compliance, a series of weighted-average-based metrics is proposed. Specifically, the global TAI ($gtai$) score is defined as

$$gtai = w_{must} \cdot \frac{n_{must}}{N_{must}} + w_{should} \cdot \frac{n_{should}}{N_{should}} + w_{could} \cdot \frac{n_{could}}{N_{could}},$$

where $w_{(.)}$, $n_{(.)}$, and $N_{(.)}$ are the weight value, the number of the satisfied requirements and the total number of requirements for each priority type, respectively. An intuitive selection of values for the weights is: $w_{must} = 0.8$, $w_{should} = 0.15$, and $w_{could} = 0.05$. For example, if all "must have" requirements are only satisfied, i.e., $n_{must} = N_{must}$, $n_{should} = 0$, $n_{could} = 0$, $gtai$ score will be 0.8.

These weights are introduced for the first time in the iPROLEPSIS TAI framework. Although there is no previous analysis to absolutely justify the selection of such weights, the main directions seen in the TAI responses of the iPROLEPSIS stakeholders guided the initial weights setting. During the course of the project, these weights will be adjusted, accordingly, towards the maximization of the degree of compliance with the TAI principles.

Except for general TAI, the same formula can be used to assess the compliance of each TAI category. Thus, the local TAI ($ltai$) scores are defined as

$$ltai^{category} = w_{must} \cdot \frac{n_{must}^{category}}{N_{must}^{category}} + w_{should} \cdot \frac{n_{should}^{category}}{N_{should}^{category}} + w_{could} \cdot \frac{n_{could}^{category}}{N_{could}^{category}},$$

where $category \in \{HUM, TEC, PRI, TRA, DIV, SOC, ACC\}$[16], while the $n_{(.)}^{category}$ and $N_{(.)}^{category}$ are the number of the satisfied and the total number of requirements for each TAI category, respectively.

The value of these scores is that they provide a straightforward means to monitor the TAI evolution across the lifecycle of the AI system.

3.4 The pathway towards a trustworthy iPROLEPSIS AI system

The defined TAI framework will be used during the iPROLEPSIS project to ensure that the delivered AI system meets the required level of trustworthiness. Two general activities of the iPROLEPSIS project are associated with the TAI framework, i.e., the specification of system requirements, and the delivery of a stable system version. During the system specification, apart from the user requirements, the TAI requirements will be also included in the analysis and respective technical requirements will be extracted. Moreover, the TAI framework will be assessed after the development of every version of the system and the outcomes will be presented in the respective deliverable. **Figure 9** illustrates the planned journey of TAI frameworks during the lifecycle of project. Specifically, in M23, M34 and M48, three different

[16] HUM: "Human Agency and Oversight"; TEC: "Technical Robustness and Safety"; PRI: "Privacy and Data Governance"; TRA: "Transparence"; DIV: "Diversity, Non-discrimination and Fairness"; SOC: "Societal and Environmental Well-being"; ACC: "Accountability"

versions of iPROLEPSIS ecosystem applications will be delivered and the TAI framework will be applied to assess the degree of trustworthiness of each version. Moreover, in M32 an updated version of the technical specifications will be defined so the TAI requirements with the specified priority should be included in that deliverable.

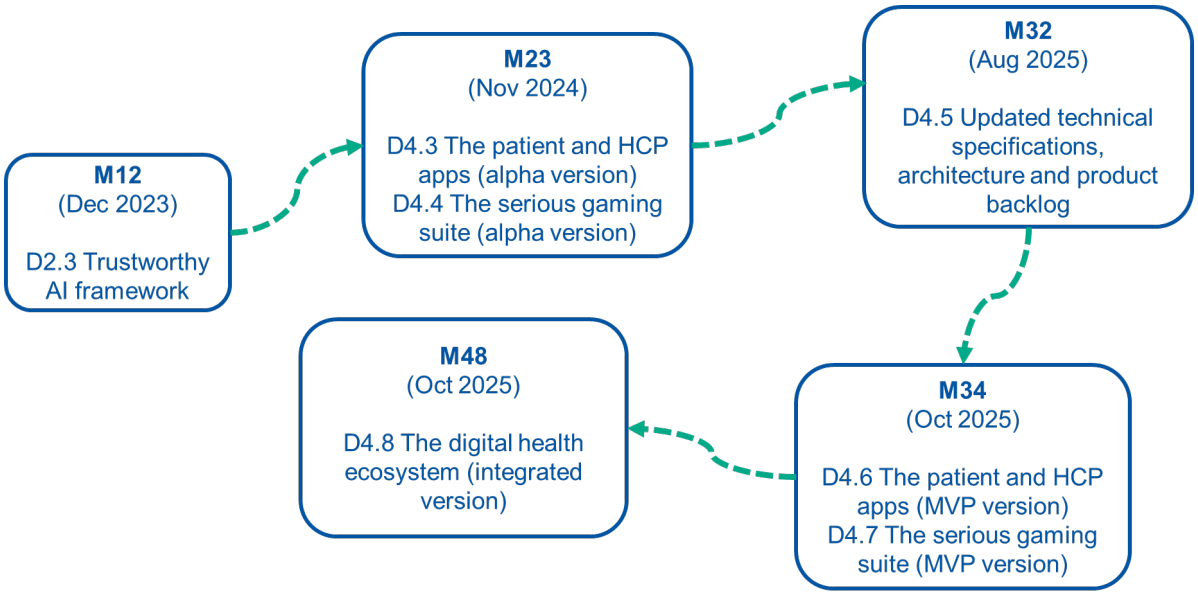


Figure 9 The iPROLEPSIS TAI framework journey.

4 Trustworthy AI technical implementation

This section provides orientation through the assimilation of trustworthiness in real-life AI systems. Particularly, Section 2, i.e., the presentation of the general landscape of TAI, is extended focusing on implementation concepts and tools. Firstly, the main practical aspects and algorithmic approaches are presented. Next, a comprehensive list of existing publicly available software tools that implement the various aspects of TAI are described.

4.1 Practical aspects and algorithms

The main practical aspects and algorithmic approaches of TAI are briefly presented in the following list and described in detail in the continuation of the section.

1. **Explainability and transparency**: Encouraging transparency in ML models and systems to ensures that their operations and decision-making processes are understandable and explainable to stakeholders.
2. **Safety and robustness**: Prioritizing the safety and robustness of ML systems to protect against potential threats, vulnerabilities, and adversarial attacks.
3. **Fairness**: Striving for fairness in ML algorithms by mitigating biases and ensuring accountability for the decisions and outcomes generated by these systems.
4. **Accountability and reproducibility:** Accountability and reproducibility are foundational principles that ensure transparency and reliability throughout AI systems' development, deployment, and use.
5. **Privacy and data governance**: Upholding privacy standards and robust data governance practices to safeguard individuals' data and ensure compliance with relevant regulations.

6. **Human-centric design**: Emphasizing a human-centric approach in ML development, considering the impact on end-users and society as a whole.
7. **Ethical considerations**: Integrating ethical considerations into the entire lifecycle of ML systems, from design and development to deployment and ongoing monitoring.

These principles aimed to provide a foundational framework for organizations and practitioners working with ML technologies, guiding them in the responsible and ethical creation and utilization of machine learning models and systems. It is worth mentioning that the organisation of aspects is derived by "The TAILOR Handbook of Trustworthy AI"[17]. TAILOR is a network of research excellence centres funded by EU and regards an exceptional scientific and technical reference point of TAI.

4.1.1 Explainability and transparency

Explainability and transparency (Li et al., 2023) in AI/ML models and systems are detrimental to ensuring that their operations and decision-making processes are understandable and explainable to the stakeholders. Explainability is understanding how an AI model makes decisions, while transparency encompasses the disclosure of information throughout the AI system's lifecycle.

Explanations are tailored to the specific context and requirements (Nauta et al., 2023; Li et al., 2023). Firstly, explanations can be categorized based on the **type of task** they are intended for, adapting to the needs of that task. Additionally, customization based on the **kind of data** processed by the AI system is essential for tailoring explanations effectively. Further customization is required for **end users**, whether non-experts, regulators, or researchers, indicating the need for final user explanations. Another dimension involves deciding whether the explanation should understand the entire logic of the AI model (**global**) or focus on a specific case (**local**). Additionally, in scenarios with **decentralized nodes**, explanations may require information not directly available on-site.

Regarding the dimensions of explanations, the approaches can vary (Guidotti et. al, 2018; Phillips et. Al, 2020). One dimension involves directly designing interpretable models, like linear regression or decision trees, which prioritize comprehensibility but may sacrifice some performance (**Explanation by Design**). Another approach, **Post Hoc Explanation**, addresses the challenge of explaining complex models, such as deep neural networks (DNNs), by analysing inputs, intermediate results, and outputs. Further, **local explanations** provide reasons for a specific outcome for a particular instance without explaining the entire AI logic. In contrast, **global explanations** offer a way to interpret any possible decision of a black box model by approximating its behaviour with a transparent model. Another dimension of explainability involves considering whether the explanations are **model-specific** or **model-agnostic**.

4.1.2 Safety and robustness

AI systems need to prioritize **security**, ensuring they **reliably** fulfil their intended purpose over time while minimizing unintentional or unexpected harm to humans or other valuable elements (European Commission, 2019). Safety extends to how the system stops its operation and the consequences of such events. The term **robustness** underlines the need for AI systems to maintain safety and functionality under challenging conditions, including unforeseen errors, unseen data, damage, or manipulation (Li et al., 2023). Concerns about the risks associated with AI are growing, particularly as humans are progressively excluded from the decision-making processes of intelligent machines.

[17] http://tailor.isti.cnr.it/handbookTAI/TAILOR_project.html

Ensuring the robustness and safety of AI systems involves a diverse approach (Hanif et. al, 2018). **Reliability**, focusing on consistent behaviour aligned with the system's intended design, instils confidence in its safety by adhering to programmed specifications over time. **Security** measures safeguard against threats and unauthorized access, maintaining the system's integrity. Understanding **distribution shifts** (Szegedy et. al, 2013) is crucial to AI systems trained on static models, ensuring adaptability to changes in data distribution in dynamic environments. Guarding against **adversarial attacks** (Goodfellow et. al, 2014), i.e., manipulations of input features inducing system failure, and **Data poisoning** (Schwarzschild et. al, 2021), where adversaries manipulate datasets to cause misclassification, is vital for diverse machine learning applications. Additionally, implementing **fallback plans** for autonomous systems contributes to the overall safety and robustness of the AI systems. Finally, **robustness testing** assesses the system's ability to function reliably under challenging conditions, guaranteeing integrity despite adversities, such as adversarial interventions or implementer errors.

4.1.3 Fairness

The development and deployment of AI systems raise critical concerns related to fairness, equity, and justice. Designing AI systems committed to these principles is essential to foster trust, prevent biases, and uphold the rights of individuals (European Commission, 2019). Since many AI systems rely on data-driven models, the link between data and functionality must be carefully investigated. In instances where training data carry inherent biases, the algorithms derived from it can perpetuate and amplify those biases, impacting the fairness and inclusivity of predictions. As AI's influence expands across diverse sectors, developing fair and inclusive systems is imperative. The presence of biases in data takes various forms, including measurement bias, omitted variable bias, representation bias, and aggregation bias (Mehrabi et al., 2022). These biases can adversely affect machine learning algorithms if not addressed, leading to skewed outcomes. Moreover, algorithms may exhibit biased behaviour independent of data biases due to design choices and historical and temporal biases. The consequences of biased algorithms extend to real-world systems, influencing user decisions. Therefore, taking a comprehensive approach to mitigating biases in AI is essential for promoting fairness and inclusivity in its applications.

Methods **addressing biases** in machine learning algorithms can be categorized into three main groups (Mehrabi et al., 2022). **Pre-processing** techniques aim to alter the data to eliminate inherent discrimination. If the algorithm has the capability to modify the training data, pre-processing can be employed. This approach is model agnostic, since preprocessing is independent of the AI/ML models and the developed algorithms. **In-processing** techniques focus on adapting the learning procedure of the AI/ML models and algorithms to eliminate discrimination during the model training process. In-processing can be applied by incorporating changes into the optimization objective or imposing constraints. **Post-processing** techniques are employed when the algorithm is restricted from modifying the training data or learning algorithm, e.g. DNNs. In this approach, labels assigned by the black-box model are reassigned based on a function during the post-processing phase to remove unfair decision paths.

4.1.4 Accountability and reproducibility

Accountability and reproducibility are foundational principles that ensure transparency and reliability throughout the development, deployment, and use of AI systems (European Commission, 2019). **Accountability**, closely tied with the principle of fairness, demands the establishment of mechanisms to observe and analyse AI systems and their outcomes, emphasizing responsibility before and after the system's entire lifecycle. **Reproducibility** (Erik Gundersen, 2021) involves the ability of independent investigators to derive the same

conclusions from an experiment by following the documentation provided by the original investigators. Achieving reproducibility is challenging due to the complexity of new ML methods, large datasets, and the use of advanced computational resources. The lack of access to training data, code, and specifications of models contributes to difficulties in replicating experiments. **Traceability** (European Commission, 2020) is defined as the need to maintain clear documentation of the data and processes involved in the entire lifecycle of an AI model, ensuring accountability and performance tracking in practice. Choices made during the development process may result in diverse behaviours and functionality, particularly for learning-based approaches, given the dependency on training data and the complexity of methods. AI models based on learning are dynamic systems, and their performance in real-world conditions may differ from their training data.

4.1.5 Privacy and data governance

Ethical considerations surrounding personal data protection form the foundation of the General Data Protection Regulation (GDPR) (European Union Agency for Fundamental Rights, 2019) emphasizing privacy by design principles (Article 5). Balancing between **data utility** and **individual privacy** is crucial for responsible AI development. Implementing privacy-preserving techniques, such as **homomorphic encryption** (Hardy et al., 2017) and **federated learning** (McMahan et al., 2016), could enable effective data analysis without compromising individual privacy. Moreover, advanced **anonymization strategies** beyond pseudonymization, such as **Differential Privacy** (Dwork & Roth, 2014) and **k-anonymity** (Samarati & Sweeney, 1998), can be exploited to safeguard user identities and sensitive information further. Additionally, identifying and mitigating **privacy attacks**, such as attacks on anonymization schemes and re-identification attacks, is vital for ensuring privacy protection; therefore, developing robust anonymization schemes and conducting risk assessments is crucial.

4.2 Recommended tools

Industry and academia have been produced a series of open-source tools which can be used to resolve various TAI-related issues. In this section, a recommended set of these tools is presented grouped by technical aspect, i.e., fairness and bias mitigation, interpretability and explainability, privacy, and robustness. The selection of the tools presented in this section was made based on their popularity at this moment in time (2023). It is certain that additional tools will be introduced in the future due to the active nature of the field; hence, a regular search for new and updated tools should be made by the responsible AI developer.

4.2.1 Fairness and bias mitigation tools

Aequitas' Bias & Fairness Audit Toolkit by *Center for Data Science and Public Policy - University of Chicago*[18] (Saleiro et al., 2018)

The Aequitas Bias & Fairness Audit Toolkit is an open-source bias audit toolkit for data scientists, machine learning researchers, and policymakers to audit machine learning models for discrimination and bias, and to make informed and equitable decisions around developing and deploying predictive tools. Specifically, it offers a range of fairness metrics, including disparate impact, false positive rates, false negative rates, predictive parity, equalized odds, and other measures to assess disparities in model performance among different groups. The toolkit offers a collection of bias detection algorithms and fairness mitigation techniques that can be applied to machine learning models to reduce or mitigate biases. Includes pre-

[18] http://aequitas.dssg.io/upload.html

processing and post-processing algorithms to address biases at different stages of the machine learning pipeline.

AI Fairness 360 by *Linux Foundation AI & Data*[19] (Bellamy et al., 2018)
The AI Fairness 360 (AIF360) toolkit is an extensible open-source library containing techniques developed by the research community to help detect and mitigate bias in machine learning models throughout the AI application lifecycle. AIF360 aims to promote fairness, transparency, and accountability in AI systems by providing a comprehensive set of tools and resources for evaluating, mitigating, and monitoring biases in machine learning models.

FAT Forensics by Thales and the University of Bristol[20] (Sokol et al., 2022)
FAT Forensics implements the state-of-the-art fairness, accountability, and transparency algorithms for the three main components of any data modelling pipeline: data (raw data and features), predictive models and model predictions. We envisage two main use cases for the package, each supported by distinct features implemented to support it: an interactive research mode aimed at researchers who may want to use it for an exploratory analysis and a deployment mode aimed at practitioners who may want to use it for monitoring FAT aspects of a predictive system. FAT Forensics is intended to contribute to the TAI deployment by enabling users to scrutinize and validate machine learning models, thereby enhancing trust and confidence in these systems, especially in domains where transparency, accountability, and regulatory compliance are critical.

Fairlearn by *Microsoft*[21] (Weerts et al., 2023)
Fairlean is an open-source toolkit for assessing and improving fairness in machine learning products. Fairlearn is a Python package that empowers developers of artificial intelligence systems to assess their system's fairness and mitigate any observed unfairness issues. Apart from fairness metrics and bias mitigation algorithms, Fairlearn is highly customisable and can be easily integrated with popular machine learning libraries such as scikit-learn, enabling seamless integration into existing machine learning pipelines and workflows.

4.2.2 Interpretability and explainability

LIME[22] (Ribeiro et al., 2016)
LIME (Local Interpretable Model-agnostic Explanations) is a well-known Python library that provides a framework for explaining the predictions of machine learning models, mostly for black-box ones (model-agnostic). It can explain the predictions of various types of models, including deep neural networks, random forests, support vector machines, among others. It is designed to generate local explanations that can help interpret the decisions made by a model on individual predictions. The individual predictions can be originated by both text classifiers, classifiers that act on tables (arrays of numerical or categorical data) and images. LIME is compatible with different machine learning libraries, including scikit-learn, TensorFlow, Keras, and XGBoost, making it easy to integrate into existing machine learning workflows.

SHAP *by Lab of AI for bioMedical Sciences, University of Washington and Microsoft Research*[23] (Lundberg & Lee, 2017)
SHAP (SHapley Additive exPlanations) is another popular Python library that provides explainability. A SHAP is based on a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the Shapley values from game theory to assign a unique value to each feature, indicating its

[19] https://github.com/Trusted-AI/AIF360
[20] https://github.com/fat-forensics/fat-forensics
[21] https://fairlearn.org/
[22] https://github.com/marcotcr/lime
[23] https://github.com/shap/shap

contribution to the prediction. Specifically, it calculates the contribution of each feature by considering all possible combinations of features and their contributions to the prediction outcome, providing a fair attribution value to each feature. SHAP is model-agnostic, meaning it can be applied to a wide range of machine learning models, including tree-based models (such as decision trees, random forests), linear models, and neural networks. Moreover, SHAP can generate both local explanations (explaining a specific prediction) and global explanations (summarizing feature importance across all predictions) to understand the model's behaviour at different levels.

AI Explainability 360 by *Linux Foundation AI & Data*[24] (Arya et al., 2019)
The AI Explainability 360 (AIX360) toolkit is an open-source library that supports interpretability and explainability of datasets and machine learning models. The AIX360 Python package includes a comprehensive set of algorithms that cover different dimensions of explanations along with proxy explainability metrics. The AIX360 toolkit supports tabular, text, images, and time series data. AIX360 offers a collection of algorithms that provide explanations for machine learning model predictions. Moreover, the explanations can be both local and global ones. Local explanations focus on explaining individual predictions, whereas global explanations aim to provide an overall understanding of the model's behaviour across the dataset. Model-Agnostic Explanations: AIX360 is designed to be algorithm/model-agnostic, meaning it can be applied to a wide range of machine learning models, including those based on decision trees, neural networks, support vector machines, among others.

Alibi by *Seldon Technologies Limited*[25] (Klaise et al., 2021)
Alibi is an open-source Python library aimed at machine learning model inspection and interpretation. Alibi offers a range of both model-agnostic and model-specific explanation techniques. These methods aim to provide insights into how machine learning models arrive at their predictions or classifications. Moreover, the toolkit includes various explanation algorithms, such as Anchors, Counterfactual Explanations, Contrastive Explanations, Integrated Gradients, Kernel SHAP, among others. These algorithms aid AI developers to generate human-understandable explanations for model predictions. Furthermore, Alibi provides tools for debugging machine learning models and detecting biases and anomalies, i.e., outliers in data. Finally, it is compatible with popular machine learning libraries such as scikit-learn, TensorFlow, and Pytorch.

XAI - An eXplainability toolbox for machine learning by *The Institute for Ethical AI & ML*[26]
XAI is a machine learning python library that contains various tools that enable for analysis and evaluation of data and models. The XAI library developed based on the eight principles for Responsible Machine Learning (see section 2.2.2). The XAI library is designed using the 3-steps of explainable machine learning, which involve 1) data analysis, 2) model evaluation, and 3) production monitoring (**Figure 10**).

[24] https://github.com/Trusted-AI/AIX360
[25] https://github.com/SeldonIO/alibi
[26] https://github.com/EthicalML/XAI

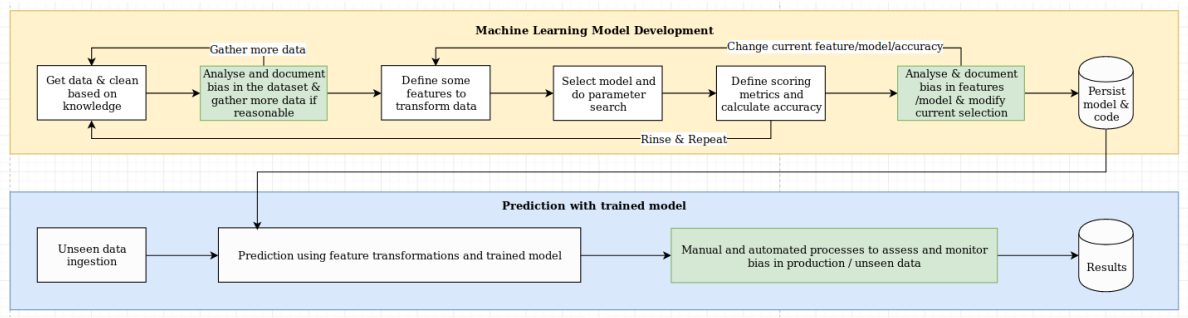


Figure 10 Visual overview of the machine learning model lifecycle incorporating the 3-steps design of XAI toolbox. The XAI toolbox involvement is indicated the green boxes[26].

InterpretML by *Microsoft*[27] (Nori et al., 2019)

InterpretML is an open-source Python library that incorporates state-of-the-art machine learning interpretability techniques in one suite. InterpretML offers a variety of model-agnostic and model-specific interpretability algorithms. The architecture of InterpretML API is shown in **Figure 11**. These algorithms help in generating human-understandable explanations for the predictions made by machine learning models. It provides capabilities to generate both global explanations (overall model behaviour) and local explanations (explanation for individual predictions).

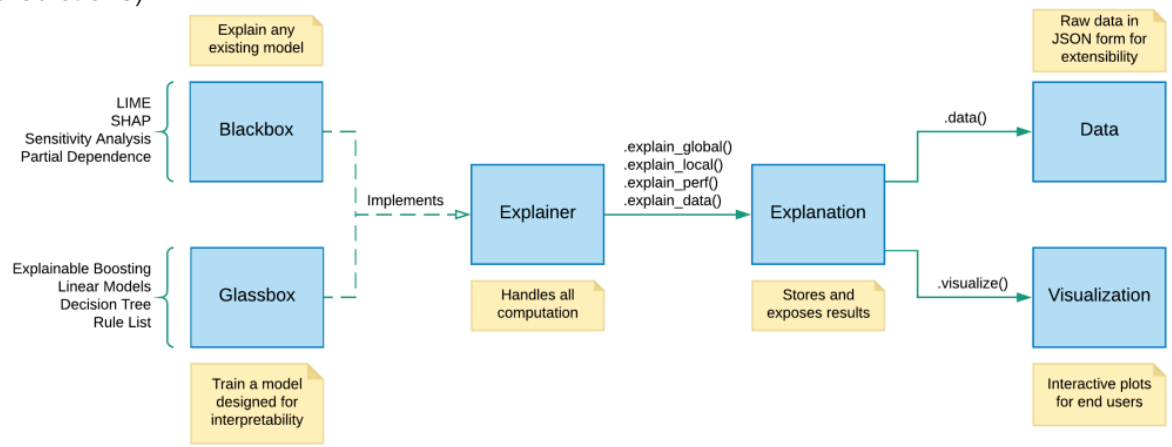


Figure 11 InterpretML API architecture (Nori et al., 2019).

Captum by *PyTorch Foundation*[28] (Kokhlikyan et al., 2020)

Captum is a model interpretability and understanding Python library the supports PyTorch models. Captum provides state-of-the-art algorithms such as Integrated Gradients, Testing with Concept Activaton Vectors (TCAV), Tracin influence functions, just to name a few, that provide researchers and developers with an easy way to understand which features, training examples or concepts contribute to a models' predictions and in general what and how the model learns. Moreover, Captum supports adversarial attacks and minimal input perturbation capabilities that can be used both for generating counterfactual explanations and adversarial perturbations. The full stack of the algorithms provided by Captum are shown in **Figure 12**. Also, it allows researchers to quickly benchmark their work against other existing algorithms available in the library. Finally, it has quick integration for models built with domain-specific libraries such as torchvision and torchtext.

[27] https://interpret.ml/
[28] https://github.com/pytorch/captum

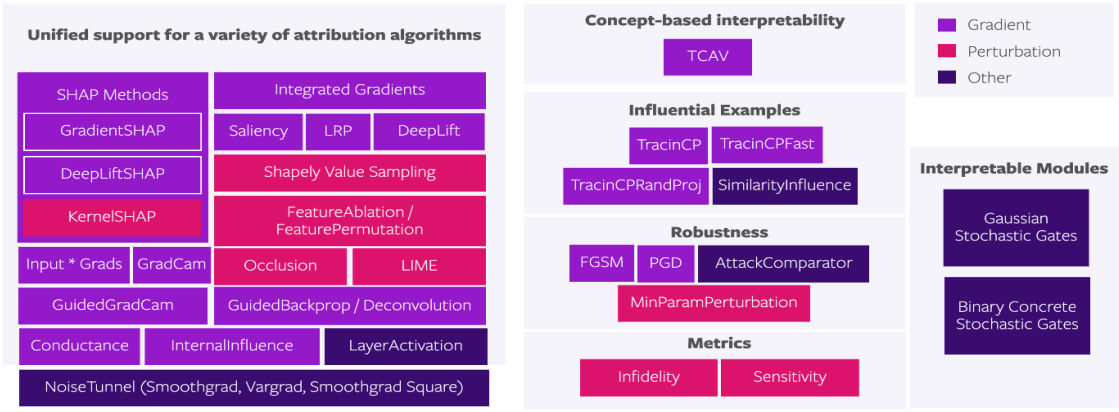


Figure 12 Full stack of algorithms and features provided by Captum library (Kokhlikyan et al., 2020).

4.2.3 Privacy

PySyft by *OpenMined*[29] (Ziller et. al, 2021)

PySyft is an open-source Python library developed by OpenMined that focuses on privacy-preserving machine learning (ML) that can be integrated with deep learning frameworks such as PyTorch. PySyft decouples private data from model training, exploiting **federated learning**, **differential privacy**, and **encrypted computation**. The framework facilitates federated learning, allowing the training of models across decentralized devices without compromising raw data. Models are trained locally on individual devices, and only model updates (gradients) are shared and aggregated. Moreover, by employing homomorphic encryption, PySyft enables secure computations on encrypted data, ensuring the confidentiality of sensitive information. PySyft is part of the broader OpenMined initiative for developing tools and technologies for privacy-preserving, decentralized, and secure artificial intelligence (AI).

TensorFlow Privacy by *Google*[30] (Abadi et al., 2016)

The TensorFlow Privacy library is an open-source Python library that provides a suite of features within the TensorFlow ecosystem to integrate privacy-preserving techniques into machine learning models with a primary focus on **differential privacy** (Abadi et al., 2016). This library encompasses implementations of widely used TensorFlow optimizers designed for training ML models with differential privacy. The core modification, differentially private stochastic gradient descent (DP-SGD), alters the conventional stochastic gradient descent (SGD) algorithm, introducing two critical adjustments to enhance privacy assurances. Firstly, it limits the sensitivity of each gradient by restricting the influence of individual training points on gradient computations and model parameters through gradient clipping. Secondly, it introduces random noise into the clipped gradients, making it statistically impossible to recognise the presence or absence of a specific data point in the training dataset. The main objective is to enable ML developers using standard TensorFlow APIs to train privacy-preserving models with minimal code modifications. The differentially private optimizers are compatible with high-level APIs such as Keras. Moreover, the library offers differentially private implementations of Keras models and is consistent with training in a federated context.

4.2.4 Robustness

Alibi Detect by *Seldon Technologies Limited*[31] (Klaise et al., 2020)

Alibi Detect is an open-source Python library focusing on **outlier**, **adversarial**, and **drift** detection. The library is designed to cover diverse data modalities, including tabular data, text,

[29] https://github.com/OpenMined/PySyft
[30] https://github.com/tensorflow/privacy
[31] https://github.com/SeldonIO/alibi-detect

images, and time series, with support for both TensorFlow and PyTorch backends specifically for drift detection. Outlier detection in terms of responsible AI refers to the identification and handling of data points that deviate significantly from the training data distribution, leading to overconfident predictions which are not reliable and cannot be used in production, while the objective of the drift detector is to identify when the distribution of the requests for the deployed model starts to diverge from the training data and the model should be retrained. Outlier detection is addressed by Alibi Detect through unsupervised off-the-shelf detectors, making it adaptable to various problem settings where labelled outlier data are unavailable. Moreover, Alibi Detect incorporates drift detectors, differentiating between covariate and label shifts, to identify when the deployed model's performance is compromised due to changes in underlying data distributions. The library also provides functionalities for detecting potential malicious data drift.

Deequ by *Amazon Web Services Labs*[32] (Schelter et al., 2018)

Deequ is an open-source library built on top of Apache Spark developed by Amazon that focuses on data quality verification and unit testing for data to ensure the quality of large datasets and identify potential issues early in data processing pipelines. PyDeequ, a Python API for Deequ, enables seamless integration with Python environments. The four main components of Deequ (**Figure 13**) allow scalable data quality assessment. **Metrics Computation** relies on Analyzers to analyse each dataset column, providing a foundational module for profiling and validating data at scale. **Constraint Suggestion** allows users to specify rules for Analyzers, generating a set of constraints for a Verification Suite. The **Constraint Verification** component performs data validation against user-defined constraints. Finally, the **Metrics Repository** enables the persistence and tracking of Deequ runs over time.

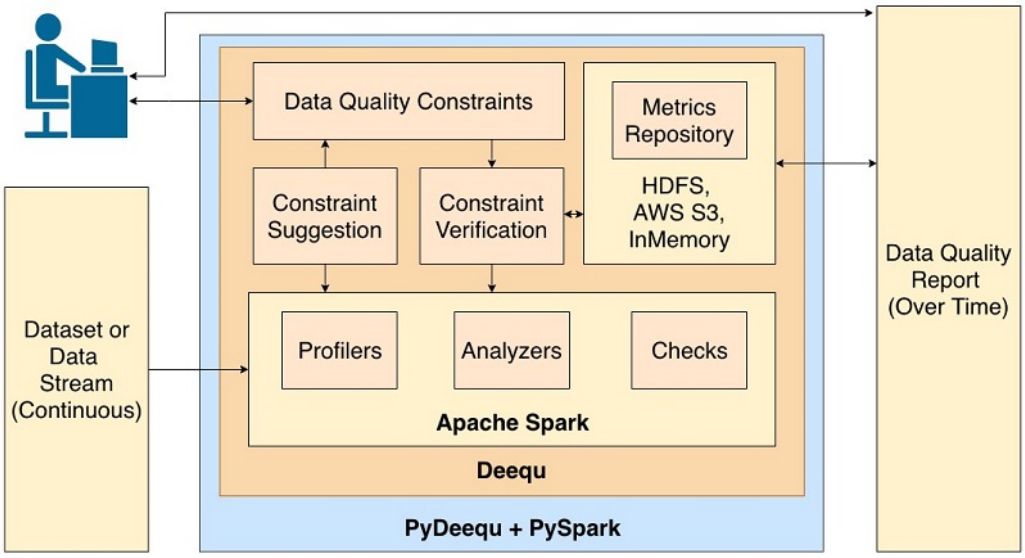


Figure 13 PyDeequ architecture[33].

PyOD[34] (Zhao et al., 2019)

Python Outlier Detection (PyOD) library is an open-source Python toolbox for outlier detection in multivariate datasets. PyOD offers a unified and user-friendly interface across many outlier

[32] https://github.com/awslabs/deequ
[33] https://github.com/awslabs/python-deequ/blob/master/imgs/pydeequ_architecture.jpg?raw=true
[34] https://github.com/yzhao062/pyod

detection algorithms, including traditional ensembles and neural network-based approaches. It supports diverse data types, including numerical and categorical features. Efficiency and high performance are highlighted by incorporating numba and joblib for JIT compilation and parallel processing. Moreover, PyOD achieves fast training and prediction through the SUOD (Scalable Unsupervised Outlier Detection) framework (Zhao et al., 2020).

CleverHans by *CleverHans Lab*[35] (Papernot et al., 2016)
CleverHans is an open-source Python library that focuses on benchmarking machine learning systems' vulnerability to adversarial examples. Adversarial machine learning involves studying and mitigating vulnerabilities in machine learning models, particularly when they are exposed to intentionally crafted inputs designed to deceive or mislead the model, known as adversarial examples. CleverHans facilitates the generation of such examples and provides tools for evaluating the robustness of machine learning models against adversarial attacks and implementing defence mechanisms. CleverHans is compatible with multiple machine-learning frameworks, such as Tensorflow and PyTorch.

5 Trustworthy AI as a business component

Beyond the development and implementation of AI solutions, a company that develops and commercialise AI-based products and services should incorporate the AI trustworthiness into the business planning procedure. Especially the small and medium-sized companies deal with significant challenges when it comes to the implementation of the ethical and trustworthy AI principles (Baker-Brunnbauer, 2021). Regardless of the inconvenience and tensions that the TAI concept might bring to the AI companies, TAI can also become an opportunity for building more robust businesses. If TAI has an integral part in business development, other than AI development business operations, e.g., marketing, sales, legal, and customer support will be significantly benefited. Baker-Brunnbauer has proposed the Trustworthy AI Implementation (TAII) Canvas (Baker-Brunnbauer, 2022), a tool inspired by the well-known Business Model Canvas. Moreover, the upcoming EU regulation of AI regards a challenging concern for the AI-related business, thus the main points currently available are briefly presented. The purpose of this section is to enrich the TAI framework iPROLEPSIS project with business development concepts to support exploitation strategy in WP6.

5.1 Trustworthy AI and business planning

The TAII Framework Canvas (Figure 14) consists of 12 sectors which represent different stages and concepts of an AI company. The filling in of the canvas should take place following specific sequence of steps starting from the company values and ending to the certification. As with similar tools, its completion can be interactive involving various stakeholders by using design thinking approaches.

The 12 steps for filling in the TAII Framework Canvas are the following:

1. **Company Values**: Used for the development of the company's Business Model as well as to shape the AI System Brief Overview.
2. **Business Model**: Provides a holistic picture of how the organization creates and captures value.
3. **AI System Brief Overview**: Describes the purpose, use case, and the used input data of the AI system.
4. **Stakeholder**: Includes all internal and external involved people, groups, departments, companies, organizations, institutions, clusters etc.

[35] https://github.com/cleverhans-lab/cleverhans

5. **Justice**: Considers existing regulations and standards for the specific AI system. Safeguard a high standard of transparency, respect for democratic values, and legitimacy.
6. **Risk**: Assessment of the AI system's ethical impact potential harm, and the affected human groups including the unintended results of the AI system.
7. **Common Good**: Analyses the dependencies of the 17 Sustainable Development Goals and the Universal Declaration of Human Rights[36].
8. **Ethics**: Generates the core list of ethical requirements. Human agency and oversight; robustness and safety; privacy and data governance; transparency; diversity; non-discrimination and fairness; societal and environmental well-being; accountability.
9. **Translation**: Transfer and translation of the ethical principles and requirements for the AI system's ecosystem.
10. **Merge**: Consolidation of the assessed input factors. Definition of the current state, visualization of the dependencies and planning of the next tasks for improvement.
11. **Execution**: Test, implementation, and verification of the results.
12. **Certification**: Safety assessment of the AI system based on legal regulations or taken actively into account by the company to provide transparency.

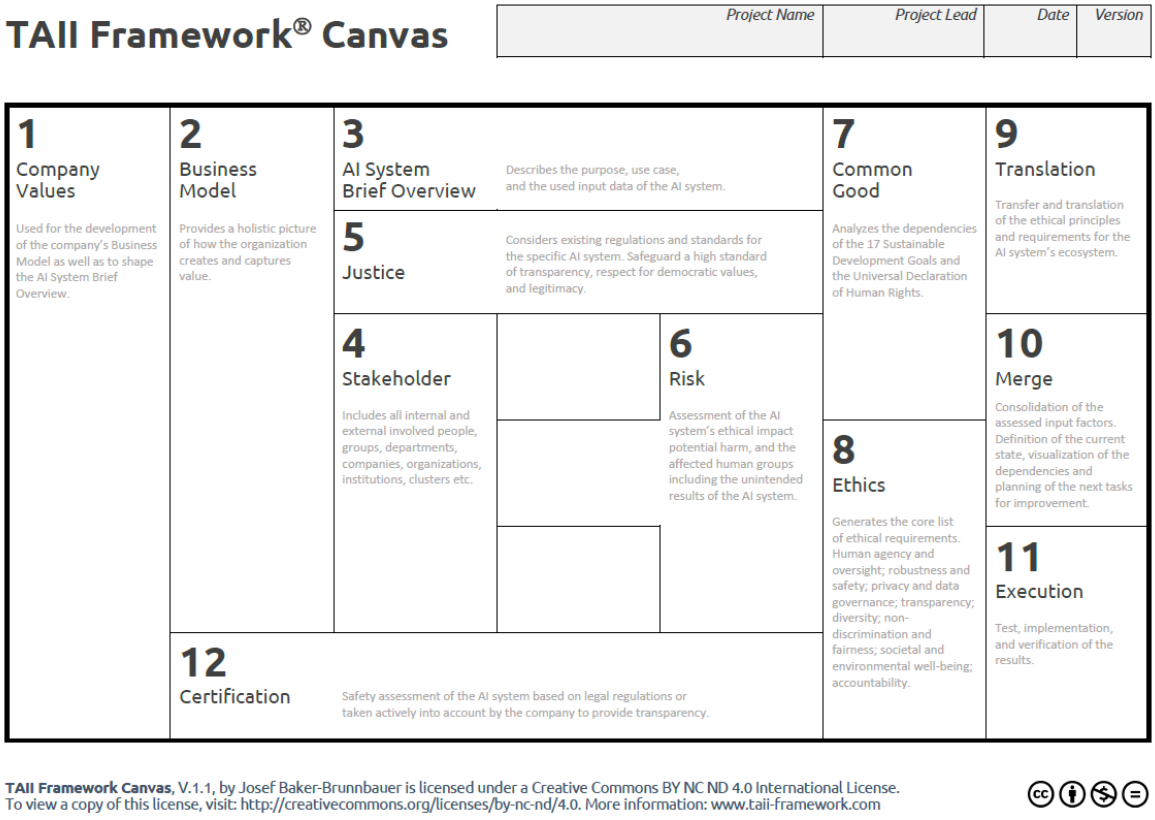


Figure 14 The Trustworthy AI Implementation (TAII) Framework Canvas.

5.2 Trustworthy AI and EU AI Act

In terms of creation of an AI-related regulatory framework, there is currently a draft proposal of the AI ACT that is expected to be finalized in year 2024. The goal of the AI ACT is the implementation of specific measures and principles, highlighting the roadmap to a TAI, by

[36] https://sdgs.un.org/goals

proposing at the same time a robust legal framework for the use and embodiment of AI in a more accurate, homocentric, and social - friendly way.

Although a two-year period will be given to the AI developers and users to comply to the provisions of the Regulation, the use of the AI technology in the iPROLEPSIS framework and its development/function during the following years, renders its by design compliance with the AI related legislation a factor of outmost importance and an absolute obligation. Taking into consideration the forementioned factors, the proposed AI ACT provides an insight into the eventually adopted legislation and the measures that need to be implemented in order to ensure compliance with it.

In article 6[37] and those that follow, the classification of an AI system is regulated, and certain requirements and procedures are envisaged in order for the trustworthiness of a high-risk AI system to be ensured. The establishment of a risk management system and data governance and management practices is envisaged, the maintenance of technical documentation and operation logs is foreseen, and the active involvement of humans during the use of the AI systems is noted, through the provision of the necessary information to the users and the possibility of humans to oversee its operation. The principles of accuracy, robustness and respect to security are also highlighted, with references to the measures to be implemented in order to ensure compliance with them.

Following this, obligations to the providers, the users, and other parties related to the use or distribution of the AI technology, such as importers, manufacturers are set out. In article 16 of the proposal, it is envisaged that providers of high-risk AI systems shall:

1. ensure that their high-risk AI systems are compliant with the requirements set out in the articles of the proposal,
2. have a quality management system in place,
3. draw-up the technical documentation of the high-risk AI system,
4. when under their control, keep the logs automatically generated by their high-risk AI systems,
5. ensure that the high-risk AI system undergoes the relevant conformity assessment procedure, prior to its placing on the market or putting into service,
6. comply with the registration obligations referred to in Article 51 of the proposal,
7. take the necessary corrective actions, if the high-risk AI system is not in conformity with the requirements set out in Chapter 2 of the proposal,
8. inform the national competent authorities of the Member States in which they made the AI system available or put it into service and, where applicable, the notified body of the non-compliance and of any corrective actions taken,
9. affix the CE marking to their high-risk AI systems to indicate the conformity with this Regulation in accordance with Article 49 of the proposal,
10. upon request of a national competent authority, demonstrate the conformity of the high-risk AI system with the requirements set out in Chapter 2 of the proposal.

6 Conclusions

The key takeaways from D2.3 are:

- The definition of the TAI framework of the iPROLEPSIS project, i.e., a set of prioritised requirements, some scores to assess different versions of the delivered AI ecosystem, and a workplan on how to be implemented during project's lifecycle.

[37] https://www.euaiact.com/article/6

- This TAI framework was developed following a novel approach using questionnaire responses from a multidisciplinary cohort of experts. The questionnaires were formed consolidating existing knowledge and best practices instilled from the analysis of the state-of-the-art landscape.
- An extensive set of recommended open-source software tools is provided. These tools can support the implementation of various technical aspects of a TAI system.
- A description of a business planning approach that involves TAI as the main component is presented, along with a brief presentation of the main points of the forthcoming EU AI ACT.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016, October). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security (pp. 308-318).

Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. Ieee Access, 6, 14410-14430.

Arya, V., Bellamy, R. K., Chen, P. Y., Dhurandhar, A., Hind, M., Hoffman, S. C., ... & Zhang, Y. (2019). One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. arXiv preprint arXiv:1909.03012.

Baker-Brunnbauer, J. (2021). Management perspective of ethics in artificial intelligence. AI and Ethics, 1(2), 173-181.

Baker-Brunnbauer, J. (2022). Trustworthy Artificial Intelligence Implementation: Introduction to the TAII Framework. Springer Nature.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943.

Cihon, P. (2019). Standards for AI governance: international standards to enable global coordination in AI research & development. Future of Humanity Institute. University of Oxford.

Clegg, D., & Barker, R. (1994). Case method fast-track: a RAD approach. Addison-Wesley Longman Publishing Co., Inc..

de Hond, A. A., Leeuwenberg, A. M., Hooft, L., Kant, I. M., Nijman, S. W., van Os, H. J., ... & Moons, K. G. (2022). Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ digital medicine, 5(1), 2.

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K., & Dugan, C. (2019, March). Explaining models: an empirical study of how explanations impact fairness judgment. In Proceedings of the 24th international conference on intelligent user interfaces (pp. 275-285).

Dwork, C., & Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. Found. Trends Theor. Comput. Sci., 9, 211-407.

Erik Gundersen O. (2021). The fundamental principles of reproducibility. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 379(2197), 20200210. https://doi.org/10.1098/rsta.2020.0210

European Commission, Directorate-General for Communications Networks, Content and Technology, (2019). Ethics guidelines for trustworthy AI, Publications Office. https://data.europa.eu/doi/10.2759/346720

European Commission, Directorate-General for Communications Networks, Content and Technology, (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment, Publications Office. https://data.europa.eu/doi/10.2759/002360

European Union Agency for Fundamental Rights, (2019). The General Data Protection Regulation: one year on civil society: awareness, opportunities and challenges, Publications Office. https://data.europa.eu/doi/10.2811/538633

Food and Drug Administration. (2019). Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD).

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and machines, 30(1), 99-120.

Hanif, M. A., Khalid, F., Putra, R. V. W., Rehman, S., & Shafique, M. (2018, July). Robust machine learning systems: Reliability and security for deep neural networks. In 2018 IEEE 24th international symposium on on-line testing and robust system design (IOLTS) (pp. 257-260). IEEE.

Hardy, S., Henecka, W., Ivey-Law, H., Nock, R., Patrini, G., Smith, G., & Thorne, B. (2017). Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. arXiv preprint arXiv:1711.10677.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature machine intelligence, 1(9), 389-399.

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: a review. ACM Computing Surveys (CSUR), 55(2), 1-38.

Klaise, J., Van Looveren, A., Cox, C., Vacanti, G., & Coca, A. (2020). Monitoring and explainability of models in production. arXiv preprint arXiv:2007.06299.

Klaise, J., Van Looveren, A., Vacanti, G., & Coca, A. (2021). Alibi explain: Algorithms for explaining machine learning models. The Journal of Machine Learning Research, 22(1), 8194-8200.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., ... & Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for pytorch. arXiv preprint arXiv:2009.07896.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., ... & Zhou, B. (2023). Trustworthy AI: From principles to practices. ACM Computing Surveys, 55(9), 1-46.

Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., Denniston, A. K., Ashrafian, H., ... & Yau, C. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. The Lancet Digital Health, 2(10), e537-e548.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). PMLR.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), 1-35.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Science and engineering ethics, 26(4), 2141-2168.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., ... & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. ACM Computing Surveys, 55(13s), 1-42.

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223.

Papernot, N., Faghri, F., Carlini, N., Goodfellow, I.J., Feinman, R., Kurakin, A., Xie, C., Sharma, Y., Brown, T.B., Roy, A., Matyasko, A., Behzadan, V., Hambardzumyan, K., Zhang, Z., Juang, Y., Li, Z., Sheatsley, R., Garg, A., Uesato, J., Gierke, W., Dong, Y., Berthelot, D., Hendricks, P.N., Rauber, J., Long, R., & Mcdaniel, P. (2016). Technical Report on the CleverHans v2.1.0 Adversarial Examples Library. arXiv: Learning.

Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., & Przybocki, M. A. (2020). Four principles of explainable artificial intelligence. Gaithersburg, Maryland, 18.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Rivera, S. C., Liu, X., Chan, A. W., Denniston, A. K., Calvert, M. J., Ashrafian, H., ... & Yau, C. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. The Lancet Digital Health, 2(10), e549-e560.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., ... & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577.

Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.

Schelter, S., Lange, D., Schmidt, P., Celikel, M., Biessmann, F., & Grafberger, A. (2018). Automating large-scale data quality verification. Proceedings of the VLDB Endowment, 11(12), 1781-1794.

Schwarzschild, A., Goldblum, M., Gupta, A., Dickerson, J. P., & Goldstein, T. (2021, July). Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In International Conference on Machine Learning (pp. 9389-9398). PMLR.

Sokol, K., Hepburn, A., Poyiadzi, R., Clifford, M., Santos-Rodriguez, R., & Flach, P. (2022). Fat forensics: A python toolbox for implementing and deploying fairness, accountability and transparency algorithms in predictive systems. arXiv preprint arXiv:2209.03805.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2017). Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, version 2. Accessed on: Jan. 11, 2024. [Online]. Available: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

Weerts, H., Dudík, M., Edgar, R., Jalali, A., Lutz, R., & Madaio, M. (2023). Fairlearn: Assessing and Improving Fairness of AI Systems. arXiv preprint arXiv:2303.16626.

Xu, H., Liu, X., Li, Y., Jain, A., & Tang, J. (2021, July). To be robust or to be fair: Towards fairness in adversarial training. In International conference on machine learning (pp. 11492-11501). PMLR.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., & Jordan, M. (2019, May). Theoretically principled trade-off between robustness and accuracy. In International conference on machine learning (pp. 7472-7482). PMLR.

Zhao, Y., Nasrullah, Z., & Li, Z. (2019). PyOD: A Python Toolbox for Scalable Outlier Detection. J. Mach. Learn. Res., 20, 96:1-96:7.

Zhao, Y., Ding, X., Yang, J., & Bai, H. (2020). SUOD: toward scalable unsupervised outlier detection. arXiv preprint arXiv:2002.03222.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2022). Domain generalization: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Ziller, A., Trask, A., Lopardo, A., Szymkow, B., Wagner, B., Bluemke, E., Nounahon, J., Passerat-Palmbach, J., Prakash, K., Rose, N., Ryffel, T., Reza, Z.N., & Kaissis, G. (2021). PySyft: A Library for Easy Federated Learning.

Appendix I List of recommendations for the "untrustworthy" AI system

The 75 recommendations that extracted using the prototype web-based ALTAI tool are grouped into the seven categories of the European Commission's "Ethics guidelines for trustworthy AI" are shown in Table 5.

Table 5 The recommendations that should be satisfied by the "untrustworthy" AI system to become complaint with ethics guidelines for trustworthy AI.

Human agency and oversight	
1	Incorporate a process where end-users and/or subjects are adequately made aware that an AI-system influenced the decision, content, advice or outcome.
2	Ensure that the end-users or subjects are adequately informed that they are interacting with an AI system.
3	Put in place procedures to avoid that end users over-rely on the AI system.
4	Put in place any procedure to avoid that the system inadvertently affects human autonomy.
5	Take measures to deal with the possible negative consequences for end-users or subjects in case they develop attachment. In particular, provide means for the user to have control of the interactions.
6	Take measures to minimize the risk of addiction by involving experts from other disciplines such as psychology and social work.
7	Take measures to mitigate the risk of manipulation, including providing clear information about ownership and aims of the system, avoiding unjustified surveillance, and preserving autonomy and mental health of users.
8	Give specific training to humans (human-in-the-loop, human-on-the-loop, human-in-command) on how to exercise oversight.
9	Establish detection and response mechanisms in case the AI system generates undesirable adverse effects for the end-user or subject.
10	Deploy a "stop button" or procedure to safely abort an operation when needed.
Technical robustness and safety	
11	Assess potential forms of attacks to which the AI system could be vulnerable.
12	Put in place measures to ensure the integrity, robustness and overall security of the AI system against potential attacks over its lifecycle.
13	Red-team/pentest the system
14	Inform users as soon as possible if some new threats are detected.
15	Define risk, risk metrics and risk levels of the AI system in each specific use case.
16	Identify the possible threats to the AI system (design faults, technical faults, environmental threats) and the possible resulting consequences.
17	Assess the risk of possible malicious use, misuse or inappropriate use of the AI system.
18	Assess the dependency of critical system's decisions on its stable and reliable behaviour.
19	Plan fault tolerance via, e.g., a duplicated system or another parallel system (AI-based or "conventional").
20	Develop a mechanism to evaluate when the AI system has been changed enough to merit a new review of its technical robustness and safety. Develop a mechanism to evaluate when the

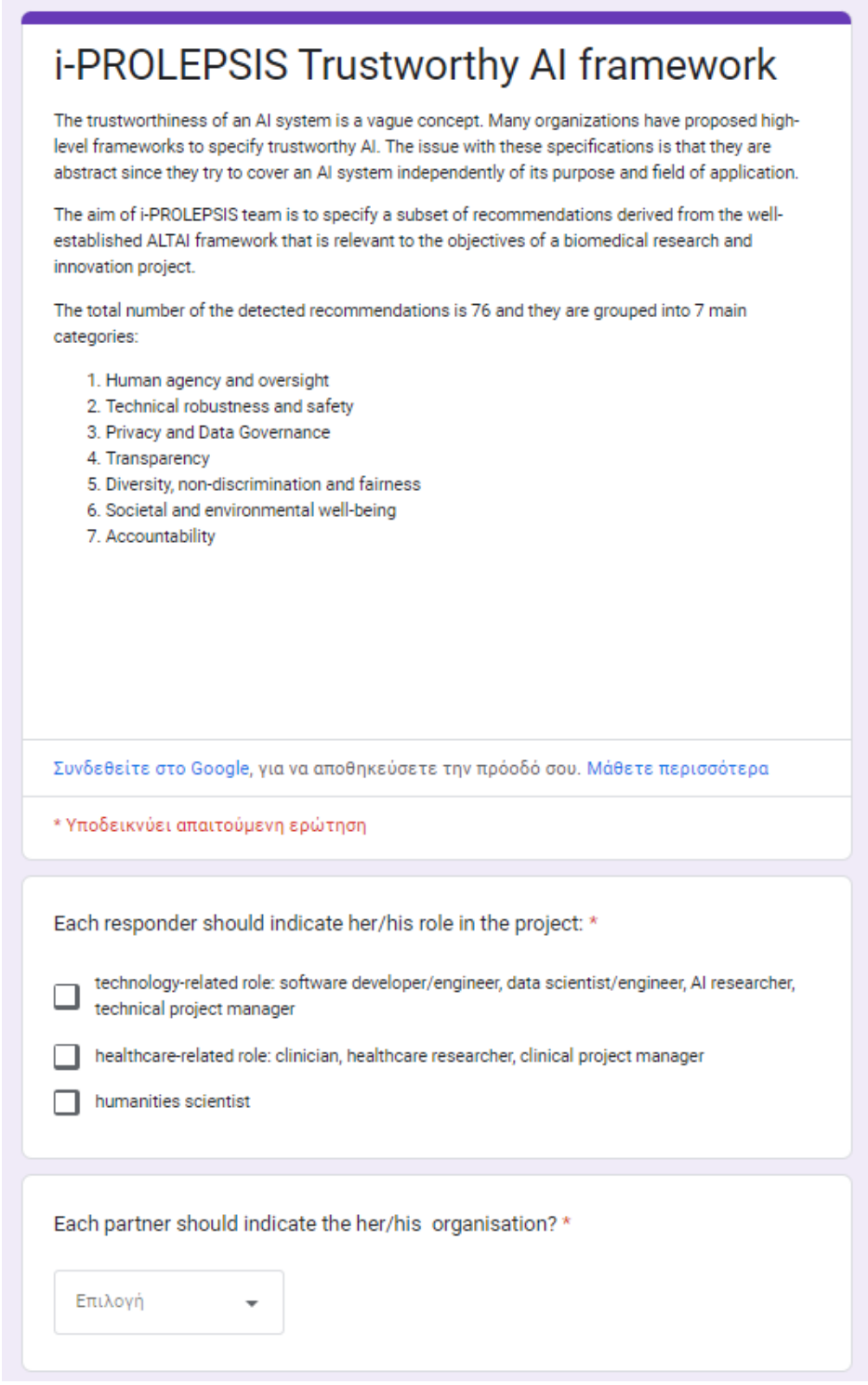
	AI system has been changed enough to merit a new review of its technical robustness and safety.
21	Put in place a series of steps to monitor and document the AI system's accuracy.
22	Consider whether the AI system's operation can invalidate the data or assumptions it was trained on, and how this might lead to adversarial effects (e.g. biased estimators, echo chambers etc.)
23	Put in place processes to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated.
24	Put in place a well-defined process to monitor if the AI system is meeting the goals of the intended applications.
25	Test whether specific contexts or conditions need to be taken into account to ensure reproducibility.
26	Put in place verification and validation methods and documentation (e.g. logging) to evaluate and ensure different aspects of the system's reliability and reproducibility.
27	Clearly document and operationalize processes for the testing and verification of the reliability and reproducibility of the AI system.
28	Define tested failsafe fallback plans to address AI system errors of whatever origin and put governance procedures in place to trigger them.
29	Put in place a proper procedure for handling the cases where the AI system yields results with a low confidence score.
Privacy and Data Governance	
30	Take measures to consider the impact of the AI system on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection.
31	Consider establishing mechanisms that allow flagging issues related to privacy or data protection concerning the AI system.
32	When relevant, implement the right to withdraw consent, the right to object and the right to be forgotten in the AI system.
33	Consider the privacy and data protection implications of data collected, generated or processed over the course of the AI system's lifecycle.
34	Consider the privacy and data protection implications of the AI system's non-personal training-data or other processed non-personal data.
35	Whenever possible and relevant, align the AI-system with relevant standards (e.g. ISO, IEEE) or widely adopted protocols for (daily) data management and governance.
Transparency	
36	Consider adopting measures to continuously assess the quality of the input data to the AI system.
37	Consider explaining the decision adopted or suggested by the AI system to its end users.
38	Consider continuously surveying the users to ask them whether they understand the decision(s) of the AI system.
39	In case of interactive AI system, consider communicating to users that they are interacting with a machine.
Diversity, non-discrimination and fairness	
40	Consider establishing a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design.
41	Consider diversity and representativeness of end-users and/or subjects in the data.
42	Test for specific target groups or problematic use cases.

43	Research and use publicly available technical tools, that are state-of-the-art, to improve your understanding of the data, model and performance.
44	Assess and put in place processes to test and monitor for potential biases during the entire lifecycle of the AI system (e.g. biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness).
45	Consider diversity and representativeness of end-users and or subjects in the data.
46	Put in place educational and awareness initiatives to help AI designers and AI developers be more aware of the possible bias they can inject in designing and developing the AI system.
47	Depending on the use case, ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system.
48	You should establish clear steps and ways of communicating on how and to whom such issues can be raised.
49	Identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the (end)-users.
50	Your definition of fairness should be commonly used and should be implemented in any phase of the process of setting up the AI system.
51	Consider other definitions of fairness before choosing one.
52	Consult with the impacted communities about the correct definition of fairness, such as representatives of elderly persons or persons with disabilities.
53	Ensure a quantitative analysis or metrics to measure and test the applied definition of fairness.
54	Establish mechanisms to ensure fairness in your AI system.
55	You should ensure that the AI system corresponds to the variety of preferences and abilities in society.
56	You should assess whether the AI system's user interface is usable by those with special needs or disabilities or those at risk of exclusion.
57	You should ensure that Universal Design principles are taken into account during every step of the planning and development process, if applicable.
58	You should take the impact of the AI system on the potential end-users and/or subjects into account.
59	You should assess whether the team involved in building the AI system engaged with the possible target end-users and/or subjects.
60	You should assess whether there could be groups who might be disproportionately affected by the outcomes of the system.
61	You should assess the risk of the possible unfairness of the system onto the end-user's or subject's communities.
Scietal and environmental well-being	
62	Consider the potential positive and negative impacts of your AI system on the environment and establish mechanisms to evaluate this impact.
63	Define measures to reduce the environmental impact of your AI system’s lifecycle and participate in competitions for the development of AI solutions that tackle this problem.
64	Inform and consult with the impacted workers and their representatives but also involve other stakeholders. Implement communication, education, and training at operational and management level.
65	Take measures to ensure that the work impacts of the AI system are well understood on the basis of an analysis of the work processes and the whole socio-technical system.
66	Provide training opportunities and materials for re- and up-skilling measures.

Accountability	
67	Designing a system in a way that can be audited later, results in a more modular and robust system architecture. Thus, it is highly recommended to ensure modularity, traceability of the control and data flow and suitable logging mechanisms.
68	To facilitate 3rd party auditing can contribute to generate trust in the technology and the product itself. Additionally, it is a strong indication of applying due care in the development and adhering to best practices and industrial standards.
69	To foresee 3rd party auditing or guidance can help with both, qualitative and quantitative risk analysis. In addition, it can contribute to generate trust in the technology and the product itself.
70	AI systems should be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm. Consequently, developers and deployers should receive appropriate training about the legal framework that applies for the deployed systems.
71	A useful non-technical method to ensure the implementation of trustworthy AI is to include various stakeholders, e.g. assembled in an “ethical review board” to monitor and assist the development process.
72	If AI systems are increasingly used for decision support or for taking decisions themselves, it has to be made sure these systems are fair in their impact on people’s lives, that they are in line with values that should not be compromised and able to act accordingly, and that suitable accountability processes can ensure this. Consequently, all conflicts of values, or trade-offs should be well documented and explained
73	Involving third parties to report on vulnerabilities and risks does help to identify and mitigate potential pitfalls.
74	A risk management process should always include new findings since initial assumptions about the likelihood of occurrence for a specific risk might be faulty and thus, the quantitative risk analysis was not correct and should be revised with the new findings.
75	Acknowledging that redress is needed when incorrect predictions can cause adverse impacts to individuals is key to ensure trust. Particular attention should be paid to vulnerable persons or groups.

Appendix II iPROLEPSIS TAI questionnaire

The questionnaire can be accessed via https://forms.gle/3M1fWFJu6rLHcChn8. Screenshots from the pages/sections of the questionnaire are presented in Figures 15 - 23.



i-PROLEPSIS Trustworthy AI framework

The trustworthiness of an AI system is a vague concept. Many organizations have proposed high-level frameworks to specify trustworthy AI. The issue with these specifications is that they are abstract since they try to cover an AI system independently of its purpose and field of application.

The aim of i-PROLEPSIS team is to specify a subset of recommendations derived from the well-established ALTAI framework that is relevant to the objectives of a biomedical research and innovation project.

The total number of the detected recommendations is 76 and they are grouped into 7 main categories:

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and Data Governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental well-being
7. Accountability

Συνδεθείτε στο Google, για να αποθηκεύσετε την πρόοδό σου. Μάθετε περισσότερα

* Υποδεικνύει απαιτούμενη ερώτηση

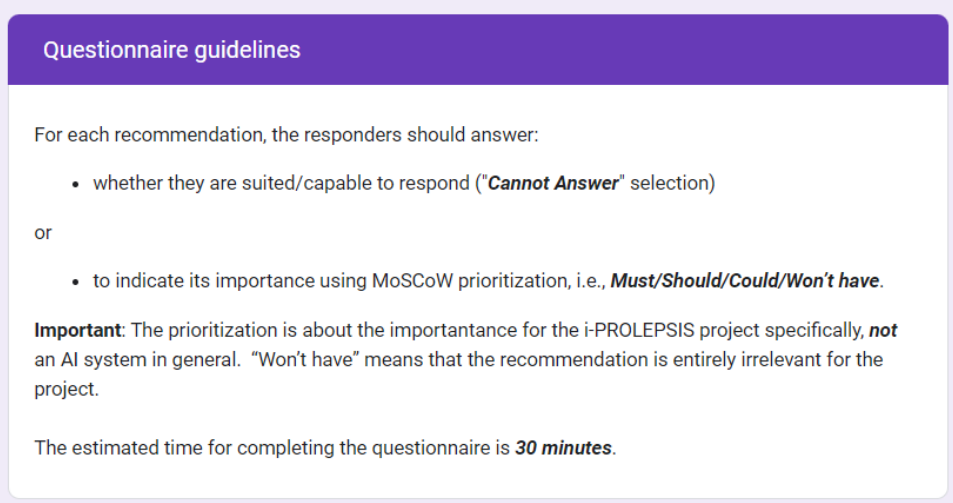
Each responder should indicate her/his role in the project: *

- ☐ technology-related role: software developer/engineer, data scientist/engineer, AI researcher, technical project manager
- ☐ healthcare-related role: clinician, healthcare researcher, clinical project manager
- ☐ humanities scientist

Each partner should indicate the her/his organisation? *

Επιλογή

Figure 15 Questionnaire's "Introduction" page.



Questionnaire guidelines

For each recommendation, the responders should answer:

- whether they are suited/capable to respond ("***Cannot Answer***" selection)

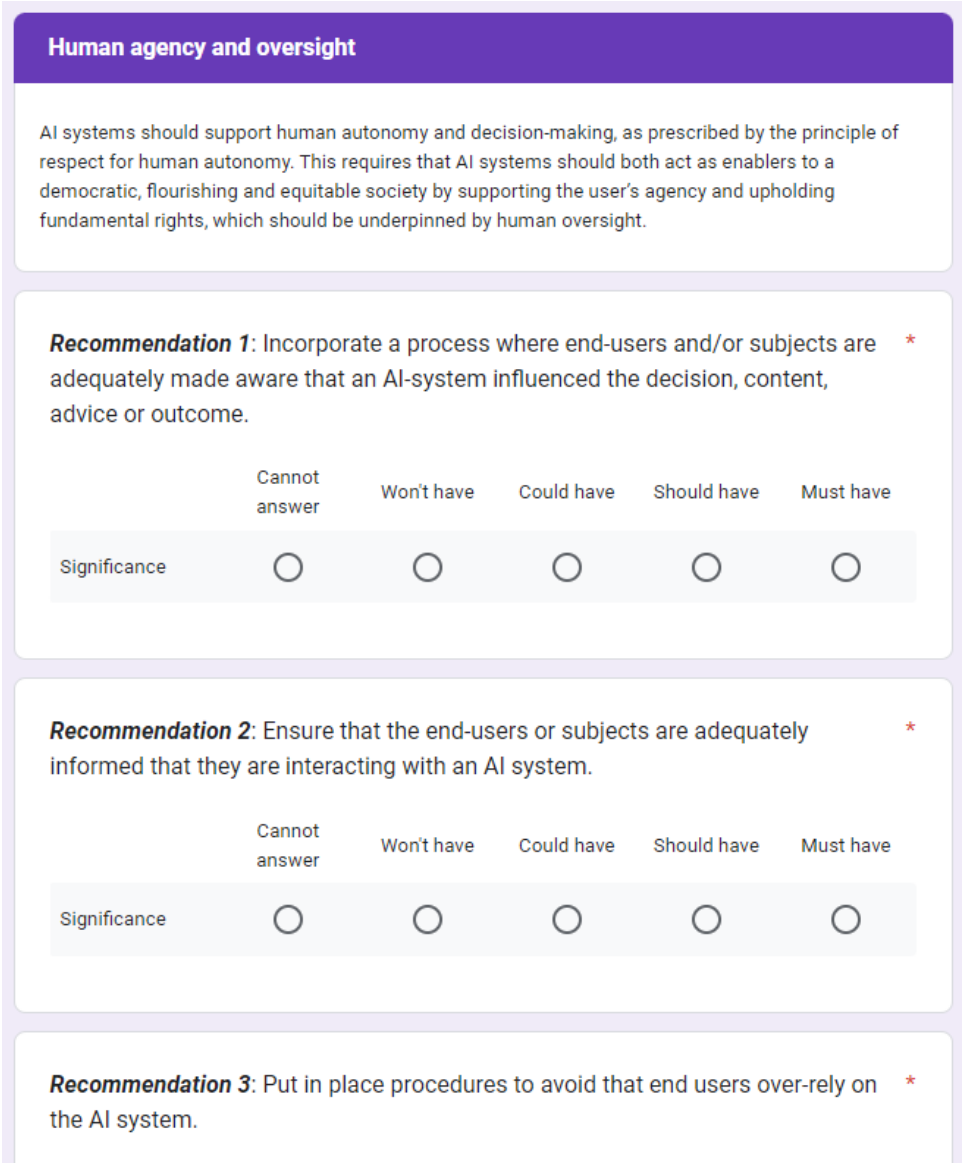
or

- to indicate its importance using MoSCoW prioritization, i.e., ***Must/Should/Could/Won't have***.

Important: The prioritization is about the importantance for the i-PROLEPSIS project specifically, ***not*** an AI system in general. "Won't have" means that the recommendation is entirely irrelevant for the project.

The estimated time for completing the questionnaire is ***30 minutes***.

Figure 16 Questionnaire's "Guidelines" page.



Human agency and oversight

AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user's agency and upholding fundamental rights, which should be underpinned by human oversight.

Recommendation 1: Incorporate a process where end-users and/or subjects are adequately made aware that an AI-system influenced the decision, content, advice or outcome. *

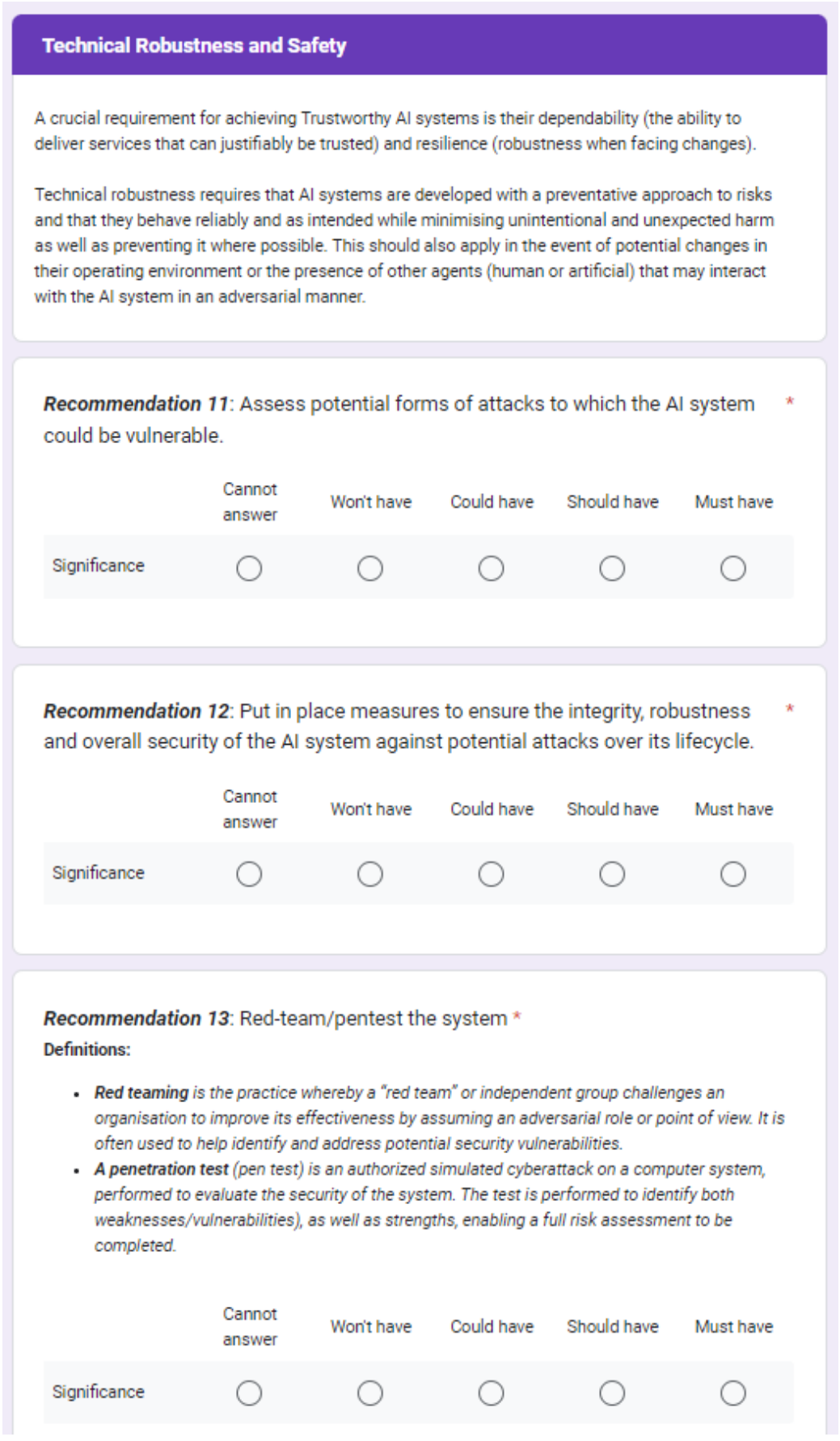
	Cannot answer	Won't have	Could have	Should have	Must have
Significance	○	○	○	○	○

Recommendation 2: Ensure that the end-users or subjects are adequately informed that they are interacting with an AI system. *

	Cannot answer	Won't have	Could have	Should have	Must have
Significance	○	○	○	○	○

Recommendation 3: Put in place procedures to avoid that end users over-rely on the AI system. *

Figure 17 First part of questionnaire's "Human Agency and Oversight" page.



Technical Robustness and Safety

A crucial requirement for achieving Trustworthy AI systems is their dependability (the ability to deliver services that can justifiably be trusted) and resilience (robustness when facing changes).

Technical robustness requires that AI systems are developed with a preventative approach to risks and that they behave reliably and as intended while minimising unintentional and unexpected harm as well as preventing it where possible. This should also apply in the event of potential changes in their operating environment or the presence of other agents (human or artificial) that may interact with the AI system in an adversarial manner.

Recommendation 11: Assess potential forms of attacks to which the AI system could be vulnerable. *

	Cannot answer	Won't have	Could have	Should have	Must have
Significance	◯	◯	◯	◯	◯

Recommendation 12: Put in place measures to ensure the integrity, robustness and overall security of the AI system against potential attacks over its lifecycle. *

	Cannot answer	Won't have	Could have	Should have	Must have
Significance	◯	◯	◯	◯	◯

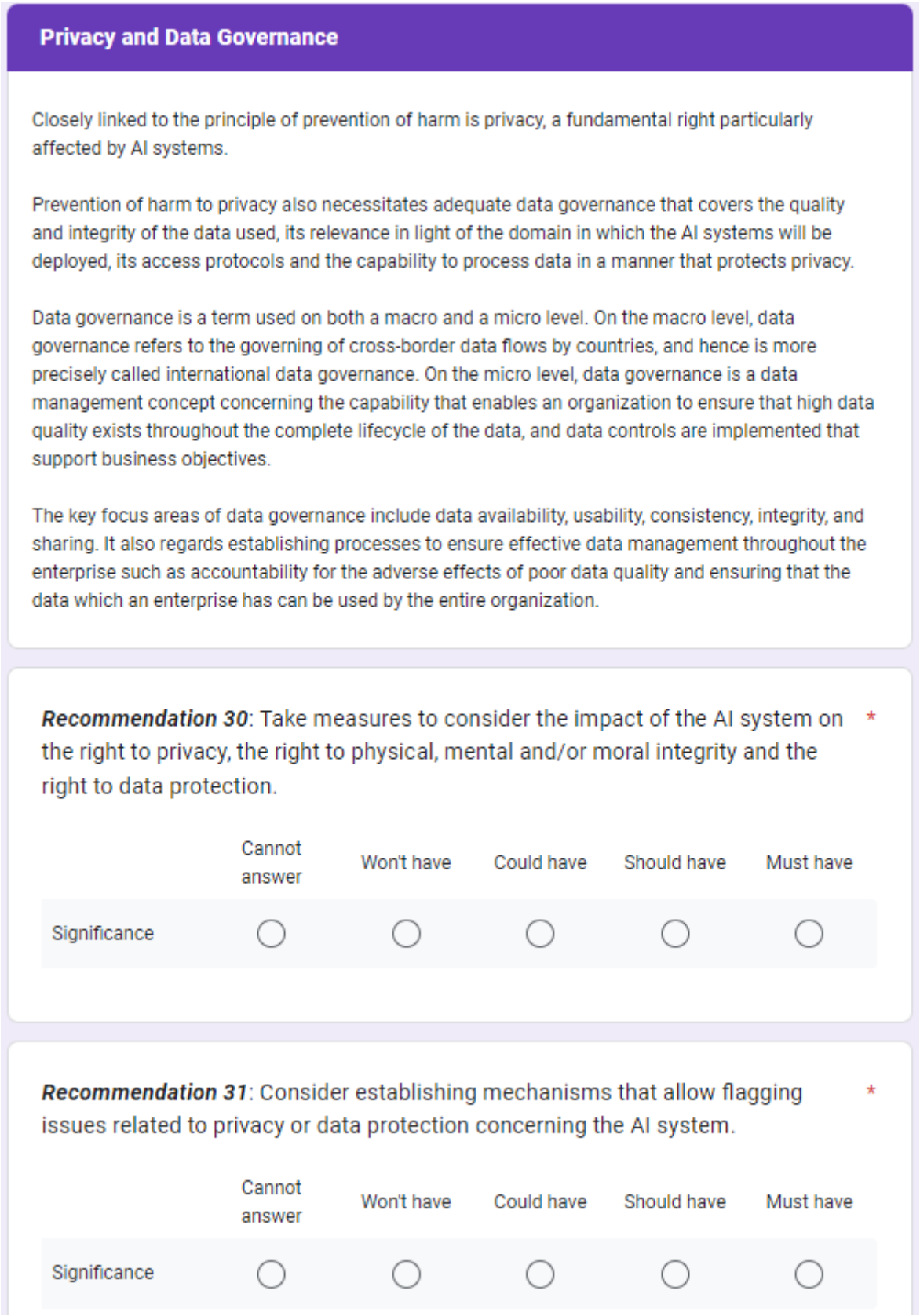
Recommendation 13: Red-team/pentest the system *

Definitions:

- ***Red teaming*** *is the practice whereby a "red team" or independent group challenges an organisation to improve its effectiveness by assuming an adversarial role or point of view. It is often used to help identify and address potential security vulnerabilities.*
- ***A penetration test*** *(pen test) is an authorized simulated cyberattack on a computer system, performed to evaluate the security of the system. The test is performed to identify both weaknesses/vulnerabilities), as well as strengths, enabling a full risk assessment to be completed.*

	Cannot answer	Won't have	Could have	Should have	Must have
Significance	◯	◯	◯	◯	◯

Figure 18 First part of questionnaire's "Technical Robustness and Safety" page. In Recommendation 13 the involved technical terms are elaborated.



Privacy and Data Governance

Closely linked to the principle of prevention of harm is privacy, a fundamental right particularly affected by AI systems.

Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy.

Data governance is a term used on both a macro and a micro level. On the macro level, data governance refers to the governing of cross-border data flows by countries, and hence is more precisely called international data governance. On the micro level, data governance is a data management concept concerning the capability that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data, and data controls are implemented that support business objectives.

The key focus areas of data governance include data availability, usability, consistency, integrity, and sharing. It also regards establishing processes to ensure effective data management throughout the enterprise such as accountability for the adverse effects of poor data quality and ensuring that the data which an enterprise has can be used by the entire organization.

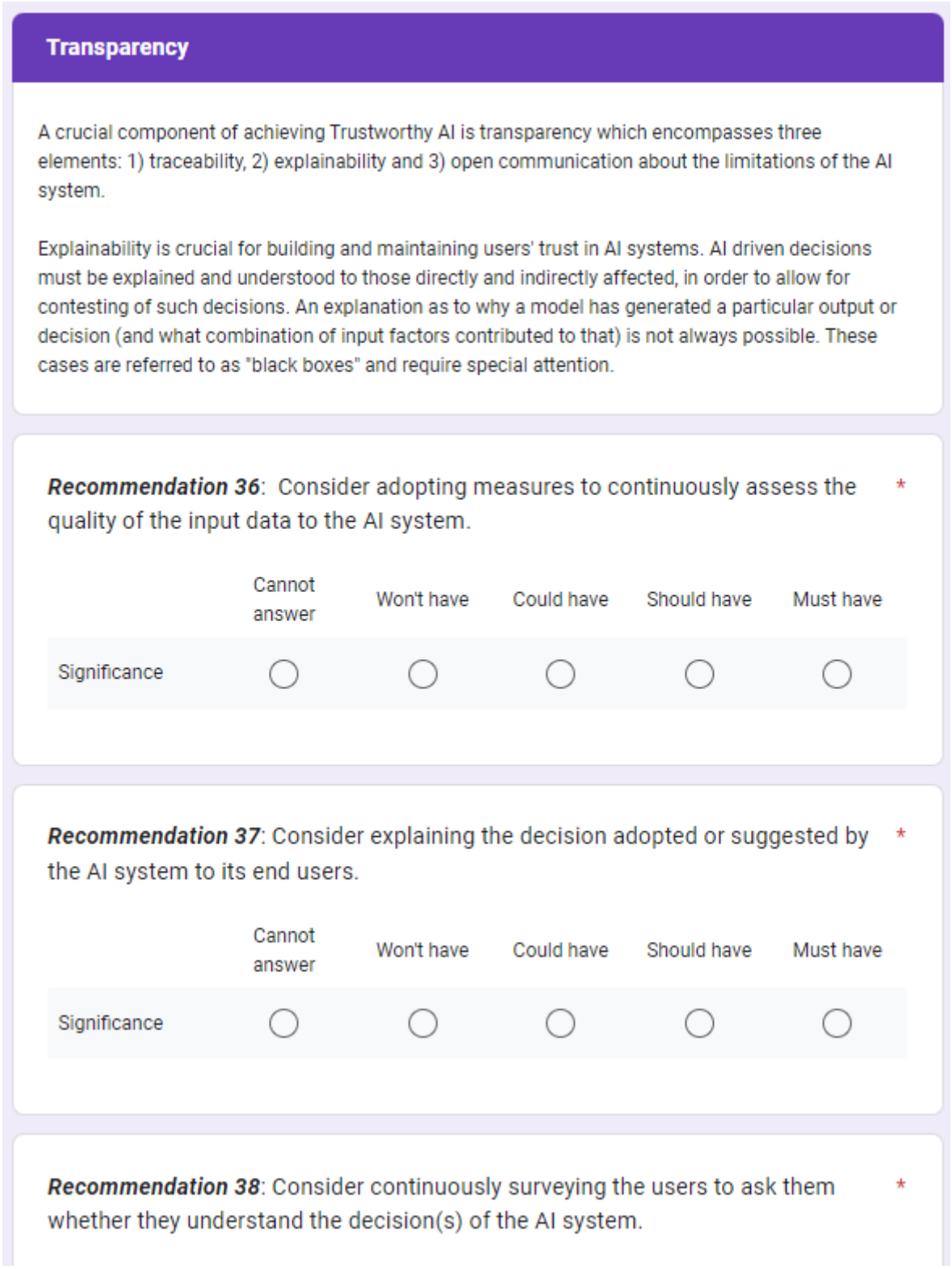
Recommendation 30: Take measures to consider the impact of the AI system on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection. *

	Cannot answer	Won't have	Could have	Should have	Must have
Significance	◯	◯	◯	◯	◯

Recommendation 31: Consider establishing mechanisms that allow flagging issues related to privacy or data protection concerning the AI system. *

	Cannot answer	Won't have	Could have	Should have	Must have
Significance	◯	◯	◯	◯	◯

Figure 19 First part of questionnaire's "Privacy and Data Governance" page.



Transparency

A crucial component of achieving Trustworthy AI is transparency which encompasses three elements: 1) traceability, 2) explainability and 3) open communication about the limitations of the AI system.

Explainability is crucial for building and maintaining users' trust in AI systems. AI driven decisions must be explained and understood to those directly and indirectly affected, in order to allow for contesting of such decisions. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as "black boxes" and require special attention.

Recommendation 36: Consider adopting measures to continuously assess the quality of the input data to the AI system. *

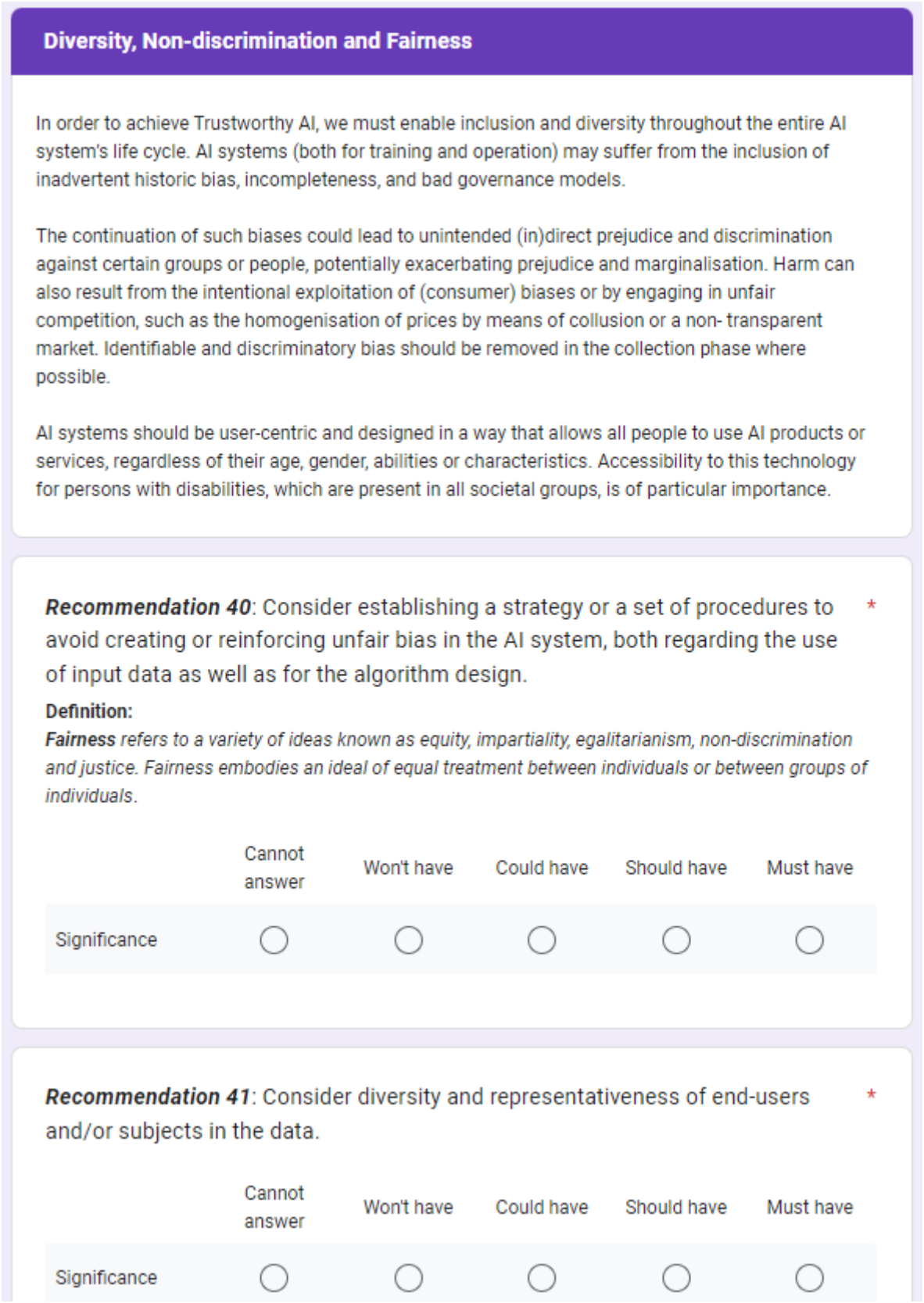
	Cannot answer	Won't have	Could have	Should have	Must have
Significance	◯	◯	◯	◯	◯

Recommendation 37: Consider explaining the decision adopted or suggested by the AI system to its end users. *

	Cannot answer	Won't have	Could have	Should have	Must have
Significance	◯	◯	◯	◯	◯

Recommendation 38: Consider continuously surveying the users to ask them whether they understand the decision(s) of the AI system. *

Figure 20 First part of questionnaire's "Transparency" page.



Diversity, Non-discrimination and Fairness

In order to achieve Trustworthy AI, we must enable inclusion and diversity throughout the entire AI system's life cycle. AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness, and bad governance models.

The continuation of such biases could lead to unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalisation. Harm can also result from the intentional exploitation of (consumer) biases or by engaging in unfair competition, such as the homogenisation of prices by means of collusion or a non- transparent market. Identifiable and discriminatory bias should be removed in the collection phase where possible.

AI systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for persons with disabilities, which are present in all societal groups, is of particular importance.

Recommendation 40: Consider establishing a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design. *

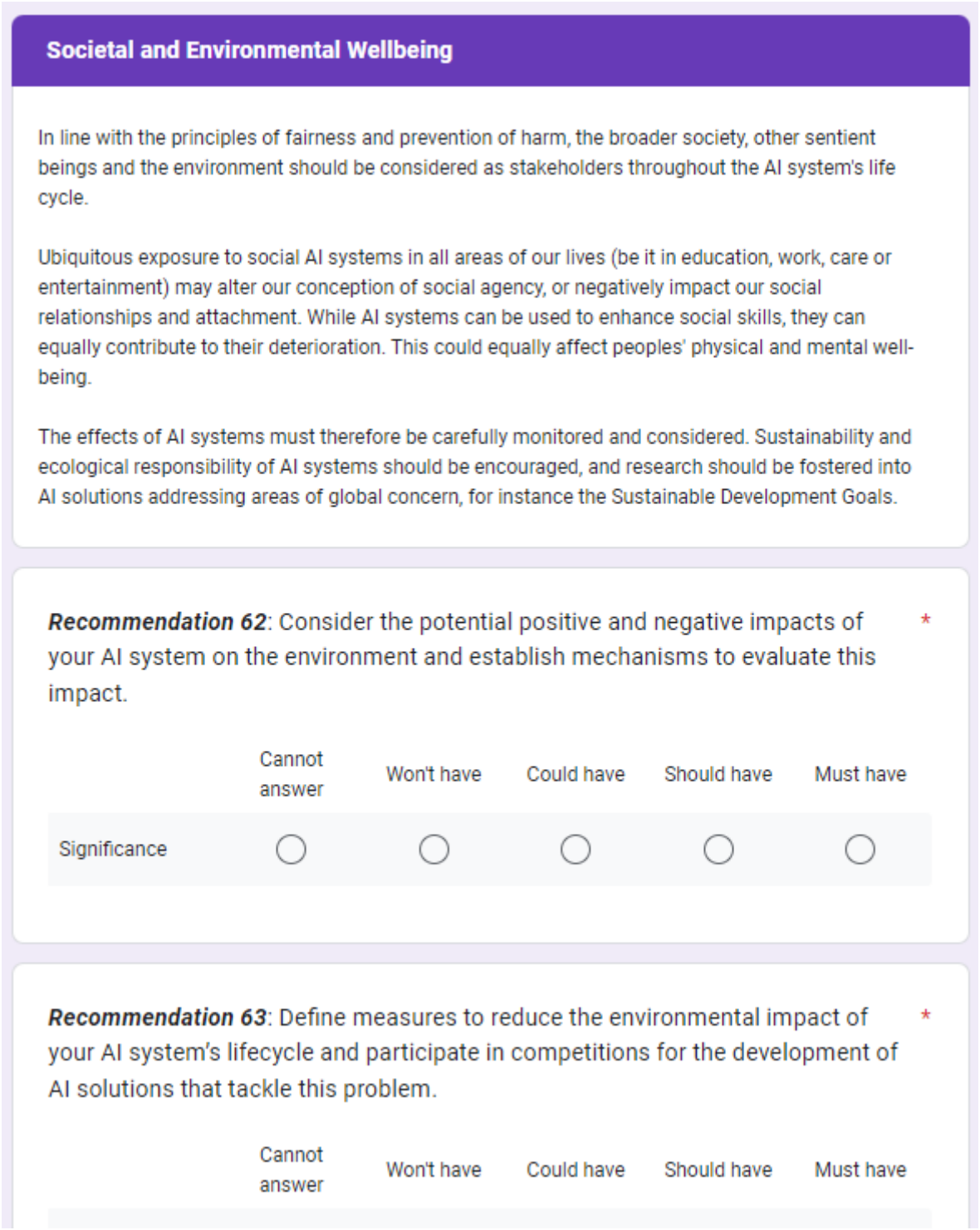
Definition:
Fairness *refers to a variety of ideas known as equity, impartiality, egalitarianism, non-discrimination and justice. Fairness embodies an ideal of equal treatment between individuals or between groups of individuals.*

	Cannot answer	Won't have	Could have	Should have	Must have
Significance	○	○	○	○	○

Recommendation 41: Consider diversity and representativeness of end-users and/or subjects in the data. *

	Cannot answer	Won't have	Could have	Should have	Must have
Significance	○	○	○	○	○

Figure 21 First part of questionnaire's "Diversity, Non-discrimination and Fairness" page.



Societal and Environmental Wellbeing

In line with the principles of fairness and prevention of harm, the broader society, other sentient beings and the environment should be considered as stakeholders throughout the AI system's life cycle.

Ubiquitous exposure to social AI systems in all areas of our lives (be it in education, work, care or entertainment) may alter our conception of social agency, or negatively impact our social relationships and attachment. While AI systems can be used to enhance social skills, they can equally contribute to their deterioration. This could equally affect peoples' physical and mental well-being.

The effects of AI systems must therefore be carefully monitored and considered. Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern, for instance the Sustainable Development Goals.

Recommendation 62: Consider the potential positive and negative impacts of your AI system on the environment and establish mechanisms to evaluate this impact. *

	Cannot answer	Won't have	Could have	Should have	Must have
Significance	◯	◯	◯	◯	◯

Recommendation 63: Define measures to reduce the environmental impact of your AI system's lifecycle and participate in competitions for the development of AI solutions that tackle this problem. *

	Cannot answer	Won't have	Could have	Should have	Must have

Figure 22 First part of questionnaire's "Societal and Environmental Well-being" page.

Accountability

The principle of accountability necessitates that mechanisms be put in place to ensure responsibility for the development, deployment and/or use of AI systems.

This topic is closely related to risk management, identifying and mitigating risks in a transparent way that can be explained to and audited by third parties. When unjust or adverse impacts occur, accessible mechanisms for accountability should be in place that ensure an adequate possibility of redress.

Accountability refers to the idea that one is responsible for his or her action – and as a corollary their consequences – and must be able to explain their aims, motivations, and reasons. Accountability has several dimensions. Accountability is sometimes required by law. For example, the General Data Protection Regulation ("GDPR") requires organisations that process personal data to ensure security measures are in place to prevent from data breaches and report if these fail. But accountability might also express an ethical standard, and fall short of legal consequences. Some tech firms that do not to invest in facial recognition technology in spite of the absence of a ban or technological moratorium might do so out of ethical accountability considerations.

Recommendation 67: Designing a system in a way that can be audited later, results in a more modular and robust system architecture. Thus, it is highly recommended to ensure modularity, traceability of the control and data flow and suitable logging mechanisms. *

	Cannot answer	Won't have	Could have	Should have	Must have
Significance	◯	◯	◯	◯	◯

Recommendation 68: To facilitate 3rd party auditing can contribute to generate trust in the technology and the product itself. Additionally, it is a strong indication of applying due care in the development and adhering to best practices and industrial standards. *

Figure 23 First part of questionnaire's "Accountability" page.