Project Acronym: STAR
Grant Agreement number: 956573 (H2020-ICT-2020-1 – Research and Innovation Action)
Project Full Title: Safe and Trusted Human Centric Artificial Intelligence in Future Manufacturing Lines
Project Coordinator: INTRASOFT International

Funded by the Horizon 2020 Framework Programme of the European Union

DELIVERABLE

D7.6 - Safety and Security Certification Programme for AI Services in Manufacturing-Initial version

Dissemination level	PU -Public
Type of Document	Other
Contractual date of delivery	30/03/2023
Deliverable Leader	SIE, INTRA
Status - version, date	Final - V1.0, 31/10/2023
WP / Task responsible	WP7
Keywords:	Artificial Intelligence, Safety and Security, Auditing, Certification, Manufacturing, Explainable Artificial Intelligence, AI Fairness, Data Quality, Data Security, Data Certification, AI Act, GDPR, Regulations

Executive Summary

STAR is research and developing trusted Artificial Intelligence (AI) solutions for production lines and industrial use cases. The project complements its scientific developments and research prototypes with added value assets that help manufacturers and providers of industrial solutions to successful implement, deploy and operate trusted AI solutions. These complementary assets include the means for assessing, benchmarking and improve the trustworthiness of AI systems and solutions. In this direction, this deliverable introduces a practical framework for auditing the trustworthiness of AI systems. The framework serves a dual objective: On the one hand it enables manufacturers and providers of industrial automation solutions to benchmark the level of trustworthiness of their solution, while on the other it provides them with practical suggestions for improving this trustworthiness.

In the scope of the present deliverable, the auditing framework is introduced as the following collection of artifacts:

- A series of questions, structured as a self-evaluation form, which enable manufacturers to provide information about the trustworthiness of their AI systems. The questions cover different aspects of AI trust, including technical/technological, organizational, ethical, and social aspects.
- A guide for processing the various questions and producing an evaluation/auditing outcome for the AI system. The auditing outcome consists of a quantitative score and of a set of guidelines for improving the trustworthiness of the AI system. Note that the auditing framework is configurable in terms of the factors that influence the trustworthiness scores. Stakeholders can adjust the scoring mechanism in ways that enable them to prioritize the factors that matter the most in the scope of their industrial processes.

The framework deals with important areas of trustworthiness, including data quality, data security, AI model fairness, data bias, explainability, transparency, user interfaces, training and more. A core part of the framework focuses on the security and reliability of data assets, which is a core element of AI trustworthiness.

The presented auditing framework is linked to the STAR research and technologies. Specifically, the STAR technologies can be used to improve the trustworthiness of AI systems as part of the implementation of the improvement guidelines that are provided as part of the system evaluation. For instance, STAR's explainable AI (XAI) technologies can be used to improve the transparency and interpretability of an AI system, while STAR's data provenance solutions can be used to boost data reliability. Nevertheless, the project's scientific and technological development do not address all the trustworthiness aspects of the auditing framework.

This is the first version of the deliverable on data certification and AI trustworthiness auditing. As part of the project's workplan, the framework will be used and validated by the industrial partners of the STAR consortium, who will provide feedback for fine-tuning the framework. The fine-tuned edition of the auditing framework will be described in deliverable D7.7 i.e., the next and final version of the present deliverable. This final version will also present guidelines for the evolution of the auditing framework to a certification suite. The exploitation plan of the project will seek to pursue the evolution of this framework to a certification programme

in collaboration with some certification organization, industrial association, or SDO (Standards Development Organization).

Deliverable Leader:	SIE,
Contributors:	John Soldatos (INTRA)
Reviewers:	R2M
Approved by:	Charalampos Ipektsidis, John Soldatos (INTRA)

Document History			
Version	**Date**	**Contributor(s)**	**Description**
0.1	25/09/2023	INTRA-LU	Table of Contents; Fine-Tuning of the Structure
0.2	29/09/2023	INTRA-LU	Driving Requirements and Brief Analysis of the State of the Art (Section 2)
0.3	05/10/2023	INTRA-LU	Description of the Auditing Framework, including a scoring guide and a feedback provision guide (Section 3)
0.4	09/10/2023	INTRA-LU	Documentation of links to STAR research components (Section 4)
0.5	10/10/2023	INTRA-LU	Conclusions, Executive Summary
0.6	16/10/2023	INTRA-LU	Final version sent for review
0.7	19/10/2023	R2M	Comments provided by reviewers
1.0	31/10/2023	INTRA	QA and creation of the submitted version

Table of Contents

List of Figures

List of Tables

Definitions, Acronyms and Abbreviations

Acronym/ Abbreviation	Title
AI	Artificial Intelligence
AIOTI	Alliance for the IoT Innovation
ALTAI	The Assessment List for Trustworthy Artificial Intelligence
AMR	Automated Mobile Robots
BDVA	Big Data Value Association
DLT	Distributed Ledger Technology
EFFRA	European Factories of the Future Research Association
FaMS	Fatigue Monitoring System
FRAIA	Fundamental Rights and Algorithms Impact Assessment
GDPR	General Data Protection Regulation
HLEG	High Level Expert Group
ICT	Information and Communications Technology
LIME	Local Interpretable Model-Agnostic Explanations
MFA	Multi-Factor Authentication
PII	Personally Identifiable Information
RAME	Risk Assessment and Mitigation Engine
RBAC	Role-Based Access Control
SDO	Standards Development Organization
SHAP	Shapley Additive Explanations
SPM	Security Policies Manager
SR	Simulated Reality
SSL	Secure Sockets Layer
TLS	Transport Layer Security
WP	Work Package
XAI	Explainable Artificial Intelligence

1 Introduction

1.1 Purpose and Scope

The STAR project researches and develops technologies for trustworthy artificial intelligence (AI) in production lines and relevant industrial use cases. The STAR technologies help manufacturers in ensuring the trustworthiness of their AI deployments. In-line with the STAR architecture that is detailed in deliverable D2.7 of the project, the functionalities of the project are clustered into three main domains:

- The cyber-security domain, which comprises functionalities that boost the cyber-resilience of AI system against cyber attacks, including for example data breaches, AI poisoning, and AI evasion attacks.
- The human robot collaboration domain, which comprises functionalities that boost the trusted interactions between humans and robots or other types of AI systems.
- The safety domain, which includes functionalities that safeguard the safety of workers and of the industrial processes that they engage.

These technologies enable developers and deployers of industrial solutions to increase the trustworthiness and resilience of their AI-based industrial solutions. Moreover, based on the development of these solutions, the STAR partners have gain experienced on the different elements and dimensions of an industrial AI system's trustworthiness, which include technical/technological, organizational/managerial, and social/ethical aspects. Based on this experience, STAR is herewith providing a framework for auditing the security, safety, and overall trustworthiness of AI systems for industrial use cases. This framework could serve as basis for validating, scoring, characterizing, and ultimately certifying an AI system against its compliance to important trustworthiness mandates, as well as against its

In this context, this deliverable introduces the AI secure/safety and trustworthiness framework that will evolve to a Safety and Security Certification Programme for AI Services in Manufacturing.

1.2 Methodology

The development of an auditing framework for security, safety and trustworthiness of AI systems in manufacturing was a challenging process for the following factors:

- The novelty and volatility of the AI area, since new innovations are constantly developed, while the regulatory frameworks for AI services are in their infancy.
- The need to balance between theorical concepts and their practical applications. The framework had to be based on a sound theoretical basis, yet being practical for manufacturers and providers of industrial solutions to use.
- The breadth of the framework that should comprise technical, organizational and social aspects.

To address these challenges, the methodology of the project considered the following inputs:

- The state of the art in AI technology for manufacturing lines, including some of STAR's innovative developments that extend the state of the art.

- Existing reports on trustworthy AI and AI regulations, including documents produced in the scope of the AI Act, Europe's AI strategy and the work of HLEG (High Level Expert Group) on AI.
- STAR's key deliverables in different aspects of trustworthy AI in manufacturing, including explainability, transparency, data security and data quality aspects.

These inputs have driven the specification of the first version of the STAR's auditing framework, which will be provided to STAR partners and other stakeholders for comments, feedback and validation. This feedback will be accordingly used in order to evolve the STAR auditing framework to a security, safety and trustworthiness certification programme. Figure 1 outlines the main methodological steps that were employed to produce this version of the deliverable, as well as its updated and final version that is due at the end of the project.

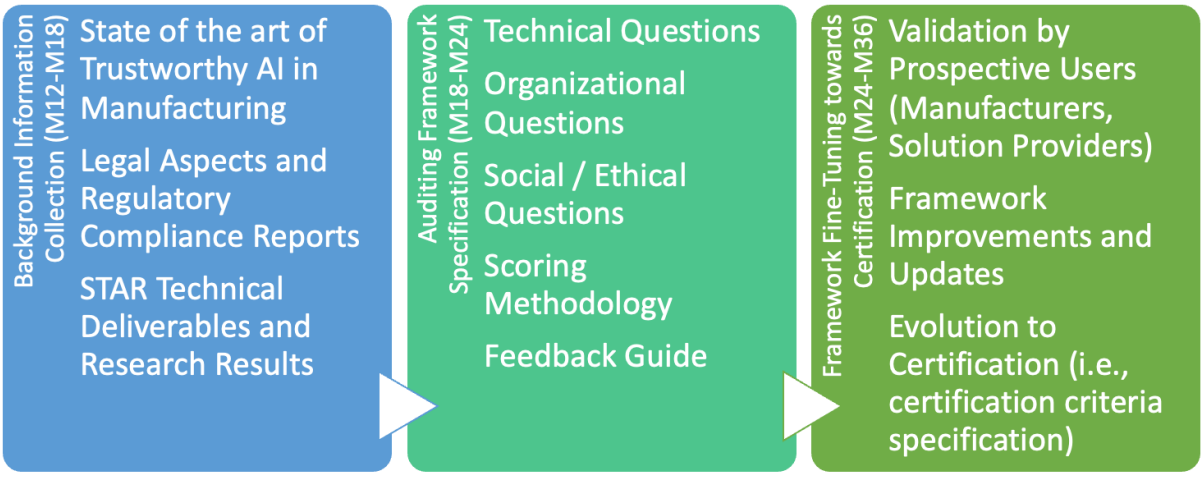

Figure 1: Auditing Framework Specification Methodology: Main Activities and Their Spread over the Project's Lifetime

1.3 Structure of the Document

The structure of the document is as follows:

- Section 2, following this introductory section, discusses the motivation and rationale behind the proposed AI trustworthiness auditing framework, which also includes data auditing and certification aspects. It also presents a brief overview of related works on data and AI trustworthiness auditing.
- Section 3 presents the AI trustworthiness framework in terms of the questions about the AI system and the guide for scoring the answers to the questions and for providing feedback to users of the framework.
- Section 4 presents how STAR technical developments can be used to aid the trustworthiness of AI systems in general and for increasing their trustworthiness score in-line with the presented framework.
- Section 5 provides a future outlook about the framework, including follow-up validation steps and its evolution to a certification programme for data and AI trustworthiness.

2 Rationale and Needs for Industrial Data Certification

2.1 The Rationale

As industrial organizations accelerate their digital transformation, they are developing and deploying a proliferating number of AI systems. In the scope of Industry 5.0 compliant use cases these cases must be trustworthy. The trustworthiness of these systems comprises multiple aspects, spanning technical performance, industrial system performance, social performance, ethics, regulatory compliance and more. Nowadays, and in the light of the absence of broadly applicable laws and regulations about AI, AI developers, integrators and deployers lack a set of best practices and guidelines on how to ensure the safety, security, and trustworthiness of their AI systems. This makes them prone to mistakes and omissions that weaken the trustworthiness of their systems including for example data quality, data security, system safety, and data fairness issues. These mistakes and omissions can in several cases be unintended, which indicates the challenge of developing, deploying and operating trustworthy AI systems in industrial use cases.

Motivated by the above-listed challenges, STAR is providing an auditing framework that can support developers and deployers of AI solutions for production lines in:

- Assessing the trustworthiness of their AI systems, in a quantitative and qualitative fashion, considering technical, organizational and social/ethical criteria.
- Obtain actionable feedback for improving the trustworthiness of their systems.

The tangible impact of the framework once used by industrial organizations will be reflected on their enhanced capability to produce trusted, safe and secure AI systems for production lines and industrial use cases.

2.2 Auditing Framework and Certification Scheme Requirements and Specifications

The production of an auditing framework that can add tangible value to the AI development, deployment and operation processes of industrial organizations must balance comprehensiveness, simplicity and completeness, which can be quite contradictory requirements in a practical auditing setting. Overall, the auditing framework is driven by the following requirements:

- **Simplicity and Understandability**: The framework should be quite simple, easy to understand and easy for industrial organizations to use. While training and education sessions can improve the ability of stakeholders to use the framework, its use should not be destined for expert users only, as it should be usable by manufacturing organizations of all sizes (including SMEs).
- **Completeness and Multi-disciplinarity**: Ensuring security, safety and trustworthiness of AI systems is not a matter of employing and deploying the right technology. Rather organizational, ethical, social, and process-related aspects must be considered as well. The framework must therefore cover all these aspects based on a multi-disciplinary approach.

- **Data Reliability Auditing**: The development, deployment and operation of trustworthy AI systems hinges on the availability of trusted and reliable data. For instance, reliable and quality data are a key prerequisite for training trusted and unbiased AI models. Nevertheless, data in industrial environments tend to be inherently unreliable for a variety of reasons including for example environmental noise and interference, faulty sensors, security attacks against IoT devices, and many more [Soldatos21]. It is therefore imperative for the framework to cover the reliability of industrial data, which is in-line with one of the main goals of the auditing framework description in the STAR DoA. Note that the term data in the context of STAR does not only refer to raw or processed data sets. It also refers to metadata of the AI models, as well as to AI outcomes that must be trusted as well.
- **Support for Scoring and Quantification**: The framework shall not be limited to providing a qualitative assessment of security, safety and trustworthiness of AI systems. Rather it should also offer a quantitative indication of a system's trustworthiness based on proper metrics and scoring mechanisms.
- **Data Auditing and Certification**: As illustrated in other STAR deliverables (e.g., WP3 deliverables) and the DoA (Description of Action), industrial data are inherently unreliable, which can create trustworthiness and security issues for AI systems and models. Hence, the auditing framework must foster the auditing and certification of industrial data to ensure the availability of trusted data for training and developing AI models, but also for executing AI models and algorithms.
- **Training and Education**: The proper use of the auditing framework requires that its users understand its structure and operation. In this direction, the framework must be accompanied with proper training resources and materials that will help the AI reskilling/upskilling of workers to a level that they can use the framework. This requirement can be addressed in conjunction with the training services of the project in WP7, also considering the training resources (e.g., training catalogue, AI Ethics course) that are being integrated in the training section of the STAR marketplace.
- **Feedback Provision**: The framework shall facilitate the provision of feedback for improving the security, safety and trustworthiness of an AI system for industrial use cases. This is essential for fulfilling one of the main objectives of the framework, which is to help developers and deployers to improve the security, safety and trustworthiness of their AI systems, while at the same time boosting a continuous improvement discipline for AI trustworthiness.

To address these requirements in the specified framework, the project has considered related assessment concepts in the AI and Industry 5.0 space, including the initiatives and reports that are briefly summarized in the following paragraph.

2.3 State of the Art Insights in Framework for Assessing Security and Trustworthiness of AI systems

The development of the STAR AI trustworthiness auding framework has been driven by existing definitions about the different dimensions of trustworthiness in the context of AI systems, as well as by on-going regulatory developments and guidelines for the development and operation of trustworthy AI system. Specifically, several research works (e.g., [Liu23]) identify six (6) main dimensions for trustworthy AI, including: (i) Safety & Robustness, (ii) Nondiscrimination & Fairness; (iii) Explainability; (iv) Privacy; (v) Accountability & Auditability;

and (vi) Environmental well-being. These dimensions are also considered in the AI Act[1], as well as in assessment guidelines specified by the HLEG[2] and EU member states[3]. Hence, they have been integrated in the auditing framework i.e., the auditing framework that is presented in Section 3 assesses whether an AI system properly addresses the above-listed aspects of trustworthiness. However, we have also integrated other aspects (e.g., data quality as illustrated in [Byabazaire20]), given that the above-presented list of trustworthiness aspects is not complete.

Apart from general research works that specify the different elements of trustworthiness, there are also many papers that focus on solutions for reinforcing trustworthiness along these dimensions. For instance, there is a host of works on the development of Explainable AI (XAI) systems and other systems that foster explainability (e.g., [Elkhawaga23], [Gilpin18]), as well as on the development of systems that exhibit fairness and operate in an unbiased way (e.g., [Lee21], [Bellamy19], [Dwork12]).

One of the main value propositions of our framework is that it provides quantitative calculations of trustworthiness, which do not exist in the literature to the best of our knowledge. However, there are several research works that define and compute metrics for specific dimensions of trustworthiness. For instance, in [Litman23] there are definitions and calculation of quantitative values for explainability. Similarly, in [Makridis23] the authors calculate explainability metrics and link them directly to the trustworthiness of AI systems. There are also works that quantify the quality of the data (e.g., [Karkouch16], [Fatouros23]) that are used in AI systems/models. Nevertheless, there is a lack of frameworks that combine all these different metrics and their scores in a unified and consistent framework for assessing/auditing an AI system in terms of its trustworthiness. This deliverable covers this gap based on the project's auditing framework that is presented in the next section.

[1] EU Parliament, AI Act https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-regulation-on-artificial-intelligence?sid=7101

[2] AI HLEG The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment https://op.europa.eu/en/publication-detail/-/publication/73552fcd-f7c2-11ea-991b-01aa75ed71a1

[3] Government of the Netherlands, Fundamental Rights Algorithm impact assessment https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms

3 AI Trustworthiness Auditing Framework Specification

3.1 Framework Overview

Figure 2 illustrates the structure, the main components and basic operation of the auditing framework. Specifically, the auditing framework consists of the following components:

- A Self-Assessment / Self-Evaluation Form for the trustworthiness of AI systems. It consists of a set of questions that relate to different aspects of the trustworthiness of an AI system.
- An AI Trustworthiness Evaluation Guide, which provides the means for processing the information of the self-evaluation form. It comprises rules for scoring the trustworthiness of an AI system, based on the information of the self-evaluation form about the system.
- Two main outputs of the AI Auditing process, which are produced based on the processing of the self-evaluation questions in-line with the rules and guidelines of the evaluation guide.

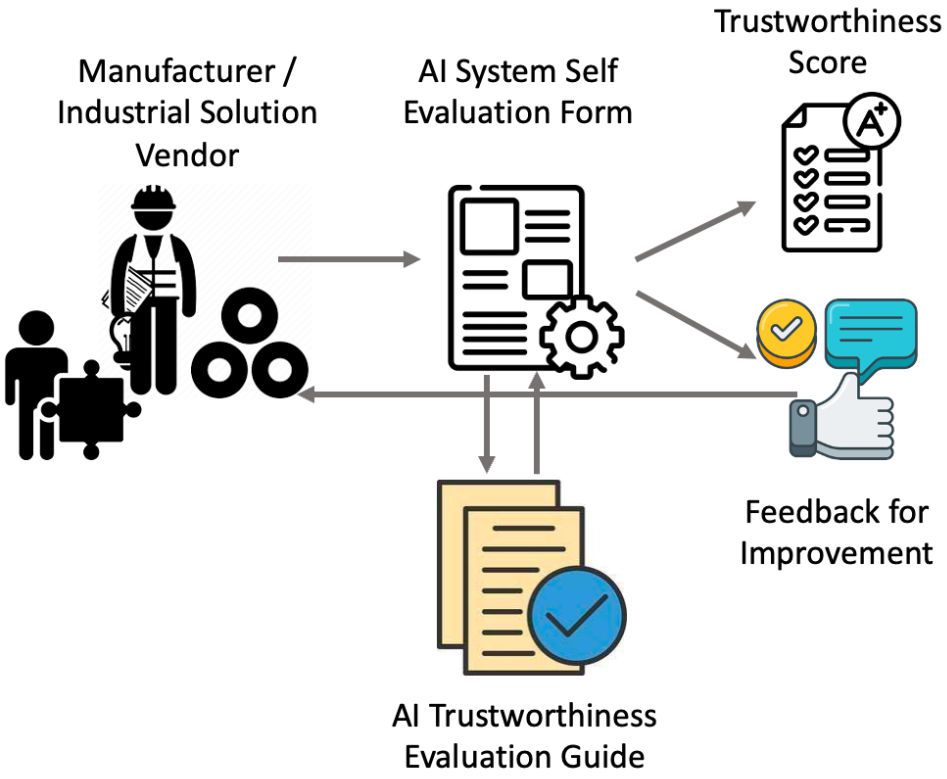

Figure 2: AI Trustworthiness Auditing Framework

Considering the above listed components, the AI trustworthiness auditing process involves the following steps:

- **Supply of Information about the AI system**: The business owner of the AI system (e.g., manufacturing worker, production manager) or the developer/integrator of the AI system provides information about the system. The information is provided in the

form of answers to a specific set of multiple-choice questions based on a proper self-evaluation form/questionnaire. The completion of this step requires that the user of the framework has a good understanding of the AI system and of AI technology in general.

- **Scoring of the system's Trustworthiness**: The supplied answers to the various multiple-choice questions are properly analysed and a trustworthiness scored is computing. The computation is based on the scoring guide of the Auditing framework.
- **Provision of feedback for improvement**: The AI trustworthiness evaluation guide is leveraged in order to provide the user of the auditing framework with information and feedback for improving the trustworthiness score of their system. This is aimed at boosting a continuous improvement discipline for the trustworthiness of the AI system. The improvement feedback must be crafted by an expert on the topic of AI trust, following a proper assessment of the supplied answers.

3.2 AI System Trustworthiness Evaluation

3.2.1 Areas of Consideration

The following questions provide the means for collecting information about the trustworthiness of an AI system. They can be structured in the so-called self-evaluation form and cover different aspects of AI trust, including technical/technological, organizational/management and social/ethical aspects. For instance, the establishment of audit trail mechanism and the employment of explainable AI systems fall in the realm of technical and technological measures, while users' training fall in the realm of organizational measures. Likewise, the establishment of ethical committees falls in the realm of ethics related measures.

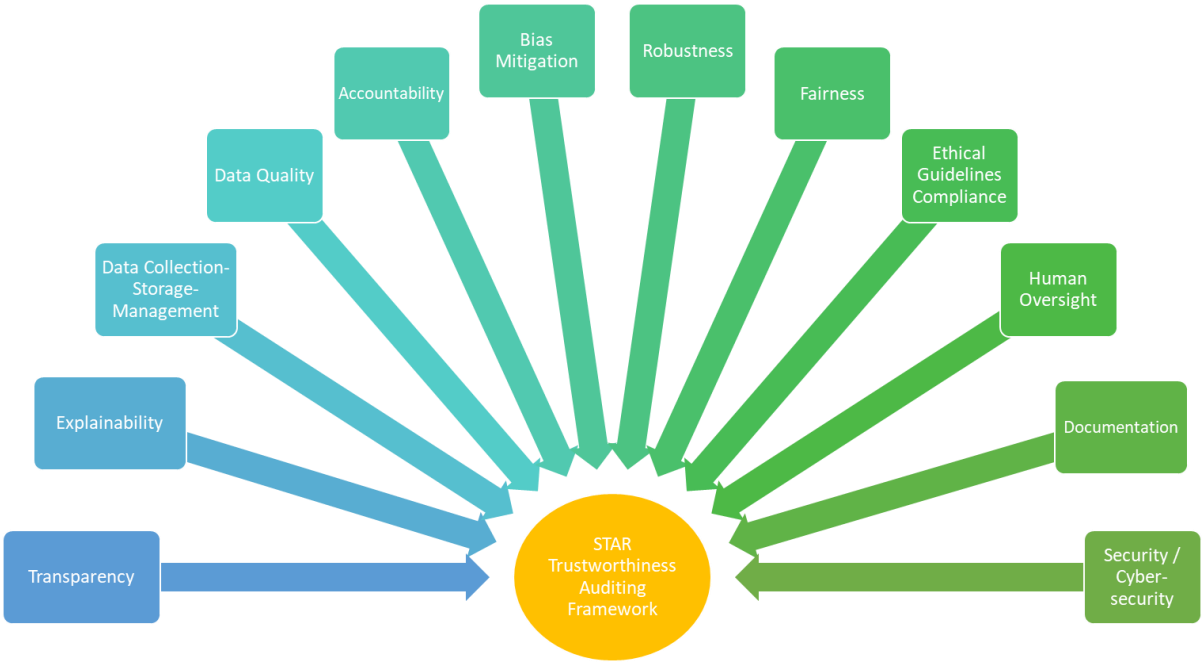

Figure 3: Different AI Trustworthiness Aspects Considered by the STAR Auditing Framework

Figure 3 illustrates the different aspects that are considered by the STAR trustworthiness auditing framework, which include transparency, explainability, fairness, bias mitigation, robustness, ethical guidelines compliance, documentation, human oversight, data collection, data storage, data management and data quality aspects. As evident from the presented list, several aspects have to be with the trustworthiness of the data that are used for developing, training and executing AI systems.

3.2.2 Trustworthiness Auditing Questions

1. How does your AI system ensure the transparency of AI models and algorithms? Select all the options that apply.

A. The system employs XAI techniques (e.g., LIME, SHAP) to interpret decision-making processes and make them more understandable to humans.
B. The system uses feature importance analysis to identify which factors or features the AI model relies on the most when making decisions.
C. The system is accompanied by comprehensive documentation of the AI system's design, architecture, algorithms, and data sources.
D. The system is accompanied by visual representations of the AI model's internal processes in order to explain complex models to non-technical stakeholders.
E. The system possesses User-Friendly Interfaces that provide insights into the AI system's behaviour and allow users to interact with the system while understanding its decision-making process.
F. The system comes with auditing tools and dashboards that allow for real-time monitoring of the AI system's performance, including model accuracy and fairness metrics.
G. The system provides information about the sources and quality of training data, including any potential biases in the data.
H. The system complies with applicable and emerging regulations, such as the GDPR, the AI Act and industry-specific standards.
I. The system is subject to regular external by third-party auditors or experts that assess the AI system's transparency and compliance with best practices.
J. Stakeholders engaging with the system are properly trained on transparency principles and practices.

2. How does your AI system ensure the explainability of AI models and algorithms? Select all the options that apply.

A. The system operates based on algorithms that are inherently interpretable (e.g., decision trees, linear models, rule-based systems).
B. The system leverages specialized XAI techniques and tools in order to explain complex AI models that operate as black-boxes.
C. The system employs feature importance analysis i.e., it can present the importance of individual features or variables in the model's decision-making process.
D. The system provides explanations on a per-instance basis, which explain why the AI system makes a specific decision for a given input.
E. The system provides visualizations that illustrate how the model processes data and arrives at conclusions.
F. The system provides natural language explanations.

G. The system provides sensitivity analysis that demonstrate how changes in input data affect the model's output.
H. The system supports counterfactual explanations that demonstrate how a slight change in input data would have led to a different result.
I. The system comes with interactive interfaces that allow users to explore and experiment with the AI system's decision-making process.
J. The system is accompanied by educational materials and resources that help users understand AI concepts and interpret model outputs.
K. The system has a feedback mechanism that allows users to provide feedback on the quality and clarity of explanations.

3. How does your system collect data within the AI system to ensure privacy, security, and integrity? Select all the options that apply.

A. The system collects data based on the data minimization principle i.e., it collects only the data necessary for the AI system's intended purpose. No sensitive or personal information is collected that is not directly relevant to the operation of the AI system.
B. Data collection is based on informed consent i.e., personal data collection takes place only after obtaining informed consent from individuals in order to ensure that they understand how their data will be used and for what purposes.
C. Data collection anonymizes or pseudonymizes data whenever possible. This includes removal or encryption of personally identifiable information (PII) to protect individual identities.
D. During data collection the system uses encryption techniques (e.g., SSL/TLS) when transmitting data over networks to prevent interception and eavesdropping.
E. The AI system ensures data quality during data collection by validating, cleaning, and sanitizing incoming data to reduce errors and inaccuracies.

4. How does your system store data within the AI system to ensure privacy, security, and integrity? Select all the options that apply.

A. Data at rest are encrypted using strong encryption methods to protect it from unauthorized access in storage.
B. The system implements access control policies to limit who can decrypt and access the data.
C. The system implements role-based access control (RBAC) and least privilege principles to restrict access to data to only those who need it for their specific roles.
D. Data is regularly backed up and the backup copies are also encrypted and stored securely.
E. Data retention policies have been developed and enforced to determine how long data is stored, while data that is no longer needed are deleted.
F. The system implements robust logging and monitoring systems to track who accesses the data and what changes are made.
G. Access to data is monitored continuously to identify potentially malicious and/or suspicious activities.

5. How does your system manage data within the AI system to ensure privacy, security, and integrity? Select all the options that apply.

A. The system classifies the various data assets based on their sensitivity and importance, while applying appropriate security measures to each classification level.
B. Data access and usage are regularly audited to ensure compliance with privacy and security policies.
C. When sharing data masking techniques are used to replace sensitive information with fictional or obfuscated data.
D. When sharing data with third parties or between systems, secure methods such as secure APIs and encrypted file transfers are used.
E. There are established ethical guidelines for data handling and use, to ensure the behaviour of the AI system aligns with ethical principles and regulations.
F. Users are educated about data privacy and security best practices, as part of measures to promote a culture of security within the organization.
G. There is a comprehensive incident response plan to address data breaches or security incidents promptly.
H. The system adheres to applicable data protection regulations (i.e., GDPR) and relevant industry-specific standards, while data policies and procedures have been updated to meet compliance requirements.

6. What measures do you implement to ensure the accountability of the AI system's decisions i.e., to attribute these decisions to specific algorithms or components? Select all the options that apply.

A. The system maintains audit trails i.e., detailed records of its activities, including data inputs, model parameters, and decision outputs.
B. The system supports model versioning i.e., it keeps track of different versions of AI models, along with the changes made to each version.
C. The system supports algorithm logging i.e., it logs the specific algorithms and techniques used in the AI system.
D. The system offers data provenance and traceability functionalities through documenting the origin and history of data assets, including its sources, transformations, and of any pre-processing.
E. The system records the explanations and interpretations generated by the AI system for specific decisions, including information about the reasons behind the choices of the system.
F. The system enables associate every action or decision made by the AI system with a timestamp. allowing for temporal tracking and analysis.
G. The system maintains user interaction logs i.e., records of interactions between users or operators and the AI system.
H. The system maintains error and exception logs that record cases where the AI system diverges from expected behaviour.
I. Ethics Committee Reports are available i.e., reports that document the decisions and recommendations of ethics committees responsible for overseeing AI system behaviour.
J. The system documents the process of training data annotation, including the actors involved, the annotations provided, and any guidelines provided to human annotators.

K. The system keeps track of feedback and correction logs i.e. feedback from users or experts, and documents corrective actions taken to address the received feedback.
L. The system comes with model validation reports i.e., records of model validation processes such as testing, validation datasets, and evaluation metrics used to assess model performance.
M. The system offers security incident reports that provide information about security incidents, breaches, or attempts to compromise the AI system's integrity, along with responses and mitigation efforts.
N. The system supports change management processes which ensure that any changes made to the AI system's configuration, code, or parameters, along with the rationale for these changes.
O. There are training and certification records for the personnel involved in AI system development and operation, including information their roles and responsibilities.

7. What measures do you implement to identify and mitigate AI bias situations? Select all the options that apply.

A. Collection of diverse and representative training data to reduce bias during AI system training and development.
B. Careful annotation of data based on structured guidelines to avoid stereotypes and biases.
C. Generation of synthetic data to increase the diversity of the datasets used for the system's training.
D. Augmentation of possible underrepresented groups or data regions towards balancing the dataset.
E. Conduct of subgroup analysis to identify bias against specific demographic groups.
F. Use of feature selection mechanism to remove potentially biased features and/or creation of new features to counteract biases.
G. Standardization and normalization of data to mitigate the influence of outliers.
H. Adjusting the importance of data samples or features to give more weight to underrepresented groups.
I. Implementation of fairness-aware machine learning algorithms (e.g., adversarial training) that consider fairness constraints during training.
J. Addition of fairness-related regularization terms to the objective function to penalize biased predictions.
K. Analysis of the model's sensitivity to different features or groups to detect and correct bias.
L. Use of metrics like disparate impact, equal opportunity, and calibration to assess the fairness of AI systems.
M. Application of algorithms that adjust the predictions or decisions post-training to reduce bias.
N. Specification and use classification thresholds to achieve fairness (e.g., equal false-positive rates for different groups).
O. Bias auditing by external organizations or experts.
P. Support for explanations for decisions to allow for external scrutiny.
Q. Collection of user feedback to identify and address bias in AI systems.
R. Promotion of diversity in AI development teams to reduce the risk of unintentional bias.

S. Education and training about bias, fairness, and ethics to AI developers and other stakeholders.

8. What measures do you implement to ensure the robustness of the AI system? Select all the options that apply.

A. The system employs adversarial training i.e., AI models are trained on data that includes adversarial examples in order to improve their resistance to attacks.
B. The training dataset is augmented with diverse and challenging examples to expose the model to a wider range of scenarios.
C. Predictions and decisions from multiple models are combined to reduce the impact of errors and increase robustness.
D. Features are carefully selected or engineered to make the model more resilient to variations and adversarial input.
E. The system uses data preprocessing techniques to remove noise and irrelevant information that might make the model more susceptible to adversarial inputs.
F. The system uses loss functions that are less sensitive to adversarial inputs, such as robust variants of cross-entropy loss.
G. The system employs mechanisms that detect when the input data is out of the model's training distribution, which mitigates the impact of adversarial inputs.
H. The system employs explainability to gain insights into model decisions and identify potential issues or adversarial attacks.
I. The system is subject to security audits towards identifying vulnerabilities and potential attack vectors.
J. The system monitors AI system behaviour and performance towards respond to any issues or adversarial attacks nearly in real time.
K. The system is deployed in a secure environment and access to the model and data is restricted.

9. What measures do you implement to ensure the fairness of the AI system? Select all the options that apply.

A. Model development is driven by clear and measurable fairness metrics, such as equal opportunity, demographic parity, and predictive parity.
B. The system has incorporated fairness constraints during model training to ensure that the model's output adheres to fairness objectives.
C. The system implements fairness-aware machine learning algorithms that reduce disparate impact and enhance fairness in AI decisions.
D. The system’s model(s) are trained using adversarial networks to make them resistant to adversarial attacks and to improve fairness.
E. Human reviewers and subject matter experts engage into the model development and evaluation processes.
F. AI system outputs are continually monitored for fairness and corrective actions are taken if needed.
G. There is diversity in AI development teams to bring a wide range of perspectives and reduce the risk of unintentional bias.

H. There are mechanisms for users to report and provide feedback on potential fairness issues.

10. What measures do you implement to ensure compliance with ethical standards and guidelines in manufacturing? Select all the options that apply.

A. The system has been developed in-line with a set of ethical AI development principles that align with your manufacturing organization's values and industry standards.
B. The system has been developed and deployed in-line with the comprehensive code of conduct that outlines the organization's ethical principles for AI development and use.
C. The development and operation of the system is overseen in terms of ethical and responsible AI principles by an AI ethics committee or advisory board.
D. There are measures to hold individuals and teams accountable for the ethical use of AI.
E. The system is developed, deployed and operated in ways that are up-to-date with relevant laws and regulations governing AI and manufacturing.
F. Employees, stakeholders, and end-users are educated and trained on AI ethics, privacy, and responsible use of the AI system.
G. Any AI components and technologies used in the system meet ethical standards, including labour practices and environmental responsibility.
H. The system's behaviour and performance are regularly monitored to detect and rectify ethical issues.
I. There are mechanisms for individuals to report ethical concerns and violations related to AI systems without fear of retaliation.
J. The development and operation of the system considers environmental ethics into AI strategies, towards reducing the environmental impact of AI.

11. What measures do you implement to ensure the quality of data of the AI system? Select all the options that apply.

A. There are clear and well-defined guidelines for data collection.
B. There are precise and consistent data annotation standards, including clear instructions to human annotators.
C. There are data cleaning processes in place that remove duplicate records, correct inaccuracies, and handle missing data.
D. Data are verified the data for accuracy and reliability based on proper checks that identify anomalies or errors.
E. The system keeps track of different versions of datasets to maintain a history of changes and updates.
F. There are measures for identifying and handling outliers in the data.
G. There are data quality metrics defined and regularly measure and monitor data quality against these metrics.
H. There system employs bias mitigation measures, especially for sensitive attributes.
I. The system documents metadata of the various datasets, including data source, collection methods, and any pre-processing steps.
J. There are data retention and data disposal policies in place for efficient and secure data management.

K. Data are backed up regularly data to prevent data loss due to accidental deletions or technical issues.

12. What measures do you implement to ensure the cyber-security of the AI system? Select all the options that apply.

A. The system encrypts data at rest and in transit using strong encryption algorithms.
B. The system implements robust access controls to restrict who can access the AI system and what actions they can perform.
C. The system employs the principle of least privilege, ensuring that users and processes have the minimum level of access necessary.
D. The system employs strong authentication methods, such as multi-factor authentication (MFA).
E. The system remains up-to-date with respect to security patches and updates.
F. The system is integrated with intrusion detection and prevention systems that monitor network traffic and detect and block suspicious activities.
G. There are regular security audits and vulnerability assessments associated with the systems and the infrastructure that supports its operation.
H. Firewalls and network segmentation are used to isolate the AI system from other parts of the network.
I. There is a comprehensive incident response plan in place that outlines how to detect, respond to, and recover from cybersecurity incidents against the AI system.
J. The users of the system are trained on security best practices such as how to identify and report phishing and other social engineering attacks.
K. The system incorporates security considerations from the early stages of its development in-line with "security by design" approaches.
L. There are regular security processes in place, including penetration testing, vulnerability scanning, and code reviews.

13. **What measures do you implement for human oversight and intervention when necessary to ensure that AI decisions align with human values and intentions? Select all the options that apply.**

A. There are clear policies and guidelines for human oversight, outlining when and how human intervention is required during the operation of the AI system.
B. The system provides explanations for its decisions in a human-understandable manner.
C. There are triggers or thresholds that prompt human intervention when certain conditions are met or when certain risks are materialised.
D. The system is designed to include humans in the decision-making process.
E. The system incorporates redundant systems and safety checks that require human approval before certain actions are taken.
F. Human oversight is employed to review and correct potential biases.
G. There are feedback mechanisms for end-users to report concerns or disputes, which can trigger human review and intervention.
H. There are established review panels or teams consisting of experts and stakeholders that periodically evaluate AI decisions and make necessary adjustments.

I. Users can customize AI behaviour within certain limits, enabling them to align the system with their values and intentions.
J. The system adheres to ethical AI frameworks and principles (e.g., IEEE 7000).

14. **What measures do you implement to provide comprehensive documentation for the AI system? Select all the options that apply.**

A. The purpose, the scope, and key objectives of the AI system are properly documented.
B. There are visual representations of the AI system's architecture, including components, data flows, and interactions.
C. There is adequate documentation about the algorithms, models, and techniques used in the AI system.
D. There is documentation of all data sources used by the systems, including information about their types, formats, and how they are accessed or collected.
E. There is documentation about all data preprocessing steps, including data cleaning, normalization, and feature engineering.
F. There is documentation about the AI models training process, including hyperparameters, training data, and validation procedures.
G. There is documentation about the evaluation metrics used to assess model performance.
H. There documentation for the APIs used to interact with the AI system, including input and output formats.
I. The documentation of the system includes external libraries, frameworks, and services used in the AI system, including information about their versions and licenses.
J. There is detailed documentation about how the AI system complies with relevant regulations and ethical guidelines, including the GDPR, the AI Act and the guidelines of the HLEG.
K. There is adequate documentation of the measures taken to protect user data and ensure data privacy (e.g., encryption and access controls).
L. There is version control for the system's documentation.
M. The documentation is accessible to all relevant stakeholders, including developers, users, and compliance officers.
N. The documentation includes references to the external resources, research papers, and documents that influenced the AI system's design.
O. The system includes legal disclaimers, terms of use, and licensing information.

3.3 AI Trustworthiness Evaluation Guide

3.3.1 Scoring Guide

Based on the answers to the above-listed questions, it is possible to calculate a trustworthiness score. In this direction, it is assumed that the more measures an organization takes regarding an AI system, the greater the trustworthiness of the system is. Table 1 presents the scoring guide in-line with this approach, including the upper and lower margins of the trustworthiness score for each question and for the auditing framework ("scorecard") as a whole.

Table 1: Scoring Guide for the STAR Trustworthiness Auditing Framework

Question	Scoring Range (points)
Q1: How does your AI system ensure the transparency of AI models and algorithms? Select all the options that apply.	0-10
Q2: How does your AI system ensure the explainability of AI models and algorithms? Select all the options that apply.	0-11
Q3: How does your system collect data within the AI system to ensure privacy, security, and integrity? Select all the options that apply.	0-5
Q4: How does your system store data within the AI system to ensure privacy, security, and integrity? Select all the options that apply.	0-7
Q5: How does your system manage data within the AI system to ensure privacy, security, and integrity? Select all the options that apply.	0-8
Q6: What measures do you implement to ensure the accountability of the AI system's decisions i.e., to attribute these decisions to specific algorithms or components? Select all the options that apply.	0-15
Q7: What measures do you implement to identify and mitigate AI bias situations? Select all the options that apply.	0-19
Q8: What measures do you implement to ensure the robustness of the AI system? Select all the options that apply.	0-11
Q9: What measures do you implement to ensure the fairness of the AI system? Select all the options that apply.	0-8
Q10: What measures do you implement to ensure compliance with ethical standards and guidelines in manufacturing? Select all the options that apply.	0-10
Q11: What measures do you implement to ensure the quality of data of the AI system? Select all the options that apply.	0-11
Q12: What measures do you implement to ensure the cyber-security of the AI system? Select all the options that apply.	0-12
Q13: What measures do you implement for human oversight and intervention when necessary to ensure that AI decisions align with human values and intentions? Select all the options that apply.	0-10
Q14: What measures do you implement to provide comprehensive documentation for the AI system? Select all the options that apply.	0-15
TOTAL	0-152

The presented approach offers the following advantages:

- It is very simple and easy to understand.
- It can score different systems automatically based on clear and unambiguous scoring rules.

However, it also suffers from the following problems and potential inaccuracies:

- It assumes that each of the listed measures in each one of the questions is of equal importance to any other measure. This might not be the case as some measures can be clearly more important than others.
- It weights the various questions differently based on the number of candidate questions, which can also be problematic as the importance of the various

trustworthiness dimensions does not necessarily depend on the number of ways that can be used to ensure that a dimension is properly and completely covered.

- It includes some similar or overlapping measures across different questions, which can lead to double benefit for implementing a measure or similarly double penalization for not implementing it. Such overlapping measures exist despite our efforts to eliminate overlaps and attribute certain measures to only one question. Nevertheless, users of the auditing framework can eliminate overlaps by deleting certain measures from instances of the framework.

To address these limitations, it is possible to configure the scoring process as illustrated in the following paragraph.

3.3.2 Scoring Configurability Options

The presented scoring process can be configured in the different ways, which overcome some of its identified limitations:

- **Weighting measures**: This approach assigns different weights to different measures of the auditing framework. In this way, the relevant importance of different measures can be considered in the computation of the trustworthiness score.
- **Weighting questions**: This approach assigns different weights to different questions of the auditing framework. It can be used to consider the relevant importance of the various questions in the computation of the trustworthiness score.
- **Weighting questions and measures**: In the scope of this approach different weights are assigned to different measures and to different questions. This provides greater flexibility in the scoring process, as the relevant importance of both individual measures and of questions (i.e., trustworthiness dimensions) can be considered.

To weigh the different measures, a weight shall be applied to each measure of every question. Assuming that W_{ij} is the weight of the j-th measure of the i-th question, the trustworthiness score (TS) will be calculated as:

$$TS = (\sum_{i=1}^{n} \sum_{j} Wij) / m,$$

where n is the total number of questions and m the total number of measures across all questions. In the above-listed formular, the score is normalized i.e., divided by the total number of measures. To this end, it is assumed that the weight W_{ij} falls in the interval [0,1].

Similarly, to weigh the different questions, a weight is applied to the score of each individual question i.e., QSi for the i-th question. Hence, the trustworthiness score (TS) will be calculated as:

$$TS = (\sum_{i=1}^{n} Wi * QSi) / n),$$

where n is the total number of questions and the QS_i is calculated as the normalized sum of possible measures that are implemented for the AI system under audit. The score is also

normalized i.e., divided by the total number of questions, while it is also assumed that the weight W_{ij} falls in the interval [0,1].

By setting different values to the weights W_{ij} and W_i in the above listed formulas it is possible to factor different measures and questions differently. Likewise, it is possible to assign weights to both measures within questions and to entire questions in order to enhance configurability.

No matter the scoring formula, the trustworthiness score will provide a measure of a system's trustworthiness, which can be generally interpreted as follows: (i) Higher scores indicate the system employs several trustworthiness measures and it likely to be robust, reliable, secure and trusted; (ii) Lower scores indicate that few trustworthiness related measures are implemented and hence there is a need for improving the robustness, the reliability, the cyber-security and other aspects of the AI system. It is also possible to set specific thresholds in order to define different classes of trustworthiness (e.g., "platinum", "gold", "silver", "bronze") so as to help organizations understand how they compare with the "best in the class" in AI trustworthiness for manufacturing systems.

3.3.3 Feedback Guide

Along with the trustworthiness score, the auditing framework is destined to provide information about how to evolve and improve the trustworthiness of the AI system under audit. In this direction, the person or team that performs the trustworthiness auditing is expected to provide guidelines for improvement for each one of the audited aspects. The feedback can be verbal or written and could be provided at different levels of detail. The following tables provide guidelines for formulating the feedback for each one of the questions of the framework. Note that a detailed and verbose feedback guide for each one of the measures is beyond the scope of this deliverable. Nevertheless, relevant insights are planned to be provided as part of a training tutorial that will accompany the next version of this deliverable and that will be made accessible through the STAR market platform.

Table 2: Trustworthiness Auditing Feedback Guidelines for Question 1

Q1: How does your AI system ensure the transparency of AI models and algorithms? **Feedback Guidelines**: Please provide feedback about the different measures of the question, prompting the AI system stakeholders to focus on the measures that they have not implemented yet. Regarding XAI Techniques, explain that XAI techniques, like LIME [Ribeiro16] and SHAP, help make AI decision-making processes more understandable to humans. Emphasize that this fosters trust in the AI system, as users and stakeholders can see why and how decisions are made. Moreover, highlight XAI's significance in ensuring accountability, given that it allows for the identification and correction of potential biases or errors in the AI model. Provide feedback about the important of feature importance analysis when it comes building transparent AI systems. Specifically, describe how feature importance analysis helps identify the factors or features the AI model relies on, which is crucial for understanding the decision-making logic. Also, stress the importance of this measure in making AI models transparent and interpretable, while explaining how they aid in regulatory compliance and fairness assessment.

In terms of documentation, discuss the importance of comprehensive documentation as a cornerstone of transparency. Documentation is key for providing a clear understanding of the AI system's design, data sources, and algorithms. Explain that this documentation serves as a reference for developers, auditors, and regulators, ensuring accountability and compliance with best practices. Discuss the importance of visual representations that help bridge the gap between technical and non-technical stakeholders by making complex AI models understandable. Explain also how these visual aids promote user trust and effective communication, enabling users to make informed decisions based on AI outputs. Refer to the importance of user-friendly interfaces that provide insights into the AI system's behaviour, in ways that enhance user understanding and trust. Explain that this transparency measure empowers users to interact with the system confidently, knowing how and why it makes decisions. In terms of auditing tools and dashboards, explain that real-time monitoring through auditing tools and dashboards is essential for maintaining AI system performance and fairness. Discuss how this allows for quick identification and mitigation of issues, reinforcing accountability and compliance with fairness standards. In terms of training data, stress the importance of disclosing training data sources and potential biases to ensure fairness and ethical use of AI systems. Discuss how this transparency measure helps users and regulators assess potential risks and biases in AI decision-making. Provide feedback about the important of regulatory compliance. Highlight that complying with applicable and emerging regulations and industry-specific standards is not only a legal requirement but also a commitment to responsible AI use. Indicate that this measure safeguards against legal risks and supports trust-building with users and stakeholders. In terms of external audits by third-party, discuss why they are essential for independent verification of transparency and compliance with best practices. Such audits can enhance credibility and provide assurance to users and regulators that the AI system meets established standards. As far as stakeholder training is concern, please present how proper stakeholder training on transparency principles and practices fosters responsible and ethical AI use.

Table 3: Trustworthiness Auditing Feedback Guidelines for Question 2

Q2: How does your AI system ensure the explainability of AI models and algorithms? **Feedback Guidelines**: Please provide feedback about the different measures of the question, prompting the AI system stakeholders to focus on the measures that they have not implemented yet. Regarding interpretable algorithms explain that using inherently interpretable algorithms (e.g., decision trees, linear models) simplifies the understanding of the AI system's decision-making process. In terms of XAI techniques, discuss how specialized XAI techniques are crucial for explaining complex black-box models (e.g., deep neural networks) towards increasing trust and accountability. Moreover, stress the significance of XAI in making AI systems more interpretable for both stakeholders and regulatory bodies.

In terms of feature importance analysis, you are expected to explain that feature importance analysis helps users understand which factors influence AI model decisions. Also, explain that this information can be valuable for addressing bias, improving user trust, and compliance. Highlight the importance of per-instance explanations when it comes to providing specific, context-aware reasons for AI decisions. Discuss how this measure helps users trust the system by clarifying why certain decisions are made for individual inputs. Present the importance of visualizations in making the AI model's decision-making process more understandable. Explain why this is important for non-technical users and discuss why it enhances transparency and trust as well, as it unveils the inner workings of the AI system. Discuss natural language explanations and how they help bridge the gap between technical and non-technical users. Explain that this fosters better user comprehension and trust. Provide feedback on the importance of sensitivity analysis, describing how it can demonstrate the impact of input changes on model outputs. Explain how that this helps users assess the robustness and reliability of AI decisions. Provide feedback on the importance of counterfactual explanations. Explain that they show how slight changes in inputs could lead to different outcomes. Discuss why this fosters a deeper understanding of the AI model's behaviour. In terms of interactive interfaces, explain that they empower users to explore and experiment with the AI system's decision-making process. Discuss how this makes AI more accessible and user-friendly. Your feedback about educational materials should emphasize how they help users understand AI concepts and interpret model outputs. Also, discuss the important of feedback mechanisms, explaining that a feedback mechanism allows users to provide input on the quality and clarity of explanations, improving user satisfaction and trust. Discuss how this promotes a continuous improvement loop for AI explainability.

Table 4: Trustworthiness Auditing Feedback Guidelines for Question 3

Q3: How does your system collect data within the AI system to ensure privacy, security, and integrity? **Feedback Guidelines**: Please provide feedback about the different measures of the question, prompting the AI system stakeholders to focus on the measures that they have not implemented yet. Your feedback about data minimization shall explain that data minimization is crucial for privacy and security by collecting only the necessary data. This reduces the potential impact of data breaches. Discuss how this measure aligns with privacy regulations and ethical principles. In case you provide feedback about informed consent, please describe how obtaining informed consent is essential for respecting individuals' privacy and autonomy. Discuss why and how this measure fosters trust and transparency, as users are aware of how their data will be used. In terms of anonymization and pseudonymization, stress the importance of anonymizing or pseudonymizing data to protect individual identities. This is crucial for data privacy and security.

Explain that this reduces the risk of data leaks and unauthorized access while allowing data to be useful for AI system functionality. Your feedback for data encryption should emphasize the need for data encryption during transmission in order to safeguard data from interception and eavesdropping. Explain why this is important for ensuring data confidentiality and security, notably when data is transferred over networks. Finally, when giving feedback about data quality assurance, explain how data quality assurance, (e.g., validation, cleaning, and sanitization) is vital for reducing errors and inaccuracies during the data collection process.

Table 5: Trustworthiness Auditing Feedback Guidelines for Question 4

Q4: How does your system store data within the AI system to ensure privacy, security, and integrity? **Feedback Guidelines**: Please provide feedback about the different measures of the question, prompting the AI system stakeholders to focus on the measures that they have not implemented yet. If relevant provide feedback about data encryption at rest, stressing the importance of encrypting data at rest using strong encryption methods to protect sensitive information from unauthorized access. Explain that this measure safeguards data confidentiality and helps meet compliance requirements. In case you provide feedback about access control policies, discuss how access control limits who can decrypt and access data i.e., it ensures that only authorized personnel can view or modify data. Explain why this measure is important for protecting data from unauthorized access, leaks, or tampering. Your feedback about RBAC shall discuss how RBAC and least privilege principles restrict data access to only those who need it for their specific roles. Explain that this significantly reduces the risk of data breaches. Furthermore, stress the importance of adhering to the principle of least privilege for limiting potential exposure of sensitive information. In terms of data backups discuss that regular data backups, along with encryption and secure storage, protect against data loss and facilitate data recovery in case of failures or disasters. Also, stress that this measure boosts data availability and business continuity. With respect to data retention policies, consider explaining how these policies help managing data lifecycle and compliance with privacy regulations. Discuss how they can also minimize the risk of holding unnecessary or outdated data. In terms of logging and monitoring, explain that robust logging and monitoring systems track data access and changes. Discuss how this provides a trail of accountability and stress how it helps detecting and investigating security incidents. Finally, when it comes to continuous monitoring measures, outline their importance for promptly identify potentially malicious or suspicious activities. Explain that this measure is crucial for early threat detection and rapid incident response.

Table 6: Trustworthiness Auditing Feedback Guidelines for Question 5

Q5: How does your system manage data within the AI system to ensure privacy, security, and integrity?
Feedback Guidelines: Please provide feedback about the different measures of the question, prompting the AI system stakeholders to focus on the measures that they have not implemented yet. In terms of data classification explain how it helps identifying the sensitivity and importance of different data assets. Discuss how data classification facilitates the proper allocation of security resources and the protection of sensitive data. Any feedback about data auditing shall describe how regular data access and usage audits ensure compliance with privacy and security policies. For instance, explain how data auditing helps detecting unauthorized or suspicious activities. Moreover, highlight that auditing is crucial for maintaining data integrity and identifying potential breaches. In case you provide feedback about data masking, you must stress the importance of data masking to replace sensitive information with fictional or obfuscated data when sharing data. Discuss why and how this measure safeguards sensitive data during sharing and collaboration. In terms of secure data sharing, it is recommended that you give feedback on the importance of secure APIs and encrypted file transfers to protect data in transit and maintain data confidentiality. Highlight the importance of this measure towards preventing data interception and breaches during data exchange. Give feedback about the importance of established ethical guidelines for data handling. Explain how they ensure that the AI system's behaviour aligns with ethical principles and regulatory requirements. Discuss how this measure prevents data misuse and promotes responsible AI use. Your user education related feedback should explain that educating users about data privacy and security best practices fosters a culture of security and reduces the risk of data breaches. Discuss the importance of having employees that can make informed decisions regarding data security. In terms of incident response plans, consider highlighting the importance of a comprehensive incident response plan for addressing data breaches or security incidents. Indicate that this measure helps minimizing damage, recover data, and maintain user trust. Finally, any feedback on regulatory compliance must stress that adhering to applicable data protection regulations and industry-specific standards is essential for legal compliance and data privacy. Discuss how this ensures that data policies and procedures meet compliance requirements.

Table 7: Trustworthiness Auditing Feedback Guidelines for Question 6

Q6: What measures do you implement to ensure the accountability of the AI system's decisions i.e., to attribute these decisions to specific algorithms or components?
Feedback Guidelines: Please provide feedback about the different measures of the question, prompting the AI system stakeholders to focus on the measures that they have not implemented yet. In case of feedback about audit trails, please explain that maintaining audit trails provides transparency and accountability by recording all activities. Discuss also how it helps reconstructing

events and assess system behaviour. Refer to audit trails-based processes like incident investigations, compliance verification, and decision accountability.

In terms of model versioning, please describe how model versioning allows tracking changes in AI models. Stress that this measure helps reverting to previous versions and assessing model performance over time.

If relevant provide feedback about algorithm logging, focusing on algorithm logging techniques that record the specific algorithms and techniques used in the AI system. Discuss why this is critical for transparency and reproducibility.

Any feedback on data provenance and traceability, should stress the importance of data provenance and traceability for understanding the origin and history of data assets. Discuss why this boosts accountability, while also explaining its importance for other trustworthiness dimensions like data quality assessment and bias detection.

Your feedback might also ask for explanations recording. Explain how recording explanations and interpretations for specific decisions helps users understand why AI decisions are made. Discuss why and how this boosts accountability in AI decision-making.

Refer to the importance of timestamps: Associating actions and decisions with timestamps enables temporal tracking, analysis, and accountability.

In terms of user interaction logs, you have to stress that maintaining user interaction logs boosts processes like user support, accountability, and system improvement. User interaction logs provide invaluable information about how users engage with the system.

In addition to user interaction logs, you may have to explain the importance of error and exception Logs. Discuss their importance for identifying deviations from expected behaviour, for enabling timely responses, and for boosting accountability.

Ethics Committee Reports is another measure that you may consider in your feedback. Explain that Ethics Committee Reports document the oversight and recommendations of ethics committees. Also discuss the role of ethics committees in overseeing AI system behaviour.

Regarding training data annotation records, it is suggested that you describe how documenting the data annotation process, actors involved, and annotation guidelines ensures transparency and accountability in data preparation.

Likewise, in terms of feedback and correction logs, you had better highlight the value of tracking feedback and corrective actions in improving the AI system. Discuss why this measure boosts the establishment of a continuous improvement discipline.

Regarding model validation reports, your feedback shall be focused on how the reports can be used to document the process of assessing model performance. Explain why and how this helps to evaluate model accuracy and fairness. Moreover, describe the role of this measure in supporting compliance verification and decision accountability.

In terms of security incident reports, consider stressing that these reports provide crucial information about incidents, breaches, and responses. Discuss how this measure helps mitigate security risks and protect system integrity.

Your feedback about change management processes must emphasize how change management ensures accountability for changes made to the AI system's configuration, code, or parameters. Finally, any feedback about training and certification records shall explain that maintaining records of personnel involved in AI system development and operation ensures accountability for system performance. In this direction, it is also important to stress the significance of personnel training and certification in upholding ethical and responsible AI use.

Table 8: Trustworthiness Auditing Feedback Guidelines for Question 7

Q7: What measures do you implement to identify and mitigate AI bias situations? **Feedback Guidelines**: Please provide feedback about the different measures of the question, prompting the AI system stakeholders to focus on the measures that they have not implemented yet. Your feedback might refer to the importance of diverse and representative training data. In this case you will have to explain why representative training data is essential to reduce bias in AI systems by ensuring that the data reflects the real-world population. You may also want to refer to how this helps avoiding underrepresentation and overrepresentation of certain groups. Explain the merits of careful data annotation based on structured guidelines, which is crucial for avoiding stereotypes and biases in training data. In terms of synthetic data generation, it is important to explain that generating synthetic data increases dataset diversity, which is key to mitigating biases and improving model performance. In case you give feedback about data augmentation, please highlight that data augmentation for underrepresented groups balances the dataset and reduces the risk of bias by ensuring equal representation. If you give feedback about subgroup analysis, you will have to stress the importance of subgroup analysis for identifying bias against specific demographic groups. You may also underline that this measure is essential for fairness assessment. In terms of feature selection and engineering you had better explain how feature engineering help identifies and mitigate potentially biased features. If you refer to data standardization and normalization, you will have to describe how data standardization and normalization mitigates the influence of outliers. Stress the importance of these techniques during data preprocessing processes. In terms of weighted data importance, you can explain that adjusting the importance of data samples or features is vital for giving more weight to underrepresented groups. Discuss the important of fairness-aware ML algorithms and their ability to consider fairness constraints during training. Discuss how these algorithms address bias proactively. In terms of fairness-related regularization terms, you shall explain that adding fairness-related regularization terms to the objective function penalizes biased predictions. This encourages fairness.

If relevant, discuss also model sensitivity analysis. Highlight that analysing the model's sensitivity to different features or groups helps detect and correct bias, ensuring fair AI behaviour. In the case of feedback about fairness metrics, you must describe how fairness metrics like disparate impact, equal opportunity, and calibration assess the fairness of AI systems. This allows for objective evaluation. If you give feedback about post-training adjustment, make sure you explain that post-training adjustment algorithms reduce bias in predictions or decisions. Discuss the merits of classification thresholds in achieving fairness, such as equal false-positive rates for different groups. In terms of external bias auditing, please stress the importance of external organizations or experts conducting bias audits, providing independent scrutiny to identify and address bias. As far as explanations are concerned, describe how supporting explanations allows for external scrutiny and transparency, enabling users and experts to understand AI decision-making. In terms of user Feedback collection, please explain that collecting user feedback helps identify and address bias in AI systems. This promotes continuous improvement and fairness. Explain why ensuring diversity in AI Development Teams reduces unintentional bias, as diverse teams bring diverse perspectives. Finally, stress the importance of bias education and training. Explain that educating AI developers and stakeholders about bias, fairness, and ethics is essential for fostering awareness and responsible AI practices.

Table 9: Trustworthiness Auditing Feedback Guidelines for Question 8

Q8: What measures do you implement to ensure the robustness of the AI system? **Feedback Guidelines**: Please provide feedback about the different measures of the question, prompting the AI system stakeholders to focus on the measures that they have not implemented yet. In case you suggest adversarial training, please explain that adversarial training improves the system's resistance to attacks by exposing it to adversarial examples during training. If you feedback refers to diverse and challenging examples, make sure that you describe how augmenting the training dataset with diverse and challenging examples exposes the model to a wider range of scenarios. In case you recommend ensemble learning, please Highlight that combining predictions and decisions from multiple models reduces the impact of errors and enhances robustness by leveraging diverse perspectives. In terms of feature engineering, please explain that careful feature selection and engineering make the model more resilient to variations and adversarial input.

If your feedback includes data preprocessing aspects, please describe how data preprocessing techniques remove noise and irrelevant information, reducing susceptibility to adversarial inputs. You may also want to suggest the use of robust loss functions, as means to reduce sensitivity to adversarial inputs (e.g., robust variants of cross-entropy). In terms of out-of-distribution detection, please highlight the use of mechanisms that detect when input data is out of the model's training distribution. This can mitigate the impact of adversarial inputs. As part of this question explainability related feedback can be provided. Please stress the importance of explainability for gaining insights into model decisions and identifying issues or adversarial attacks. In terms of security audits, please discuss that subjecting the system to security audits identifies vulnerabilities and potential attack vectors. You may also want to stress the importance of real-time monitoring, which allows for swift responses to issues or adversarial attacks. Finally, you may opt to include in your feedback insights about how deploying the system in a secure environment with restricted access to the model and data safeguards against unauthorized access and tampering.

Table 10: Trustworthiness Auditing Feedback Guidelines for Question 9

Q9: What measures do you implement to ensure the fairness of the AI system? **Feedback Guidelines**: Please provide feedback about the different measures of the question, prompting the AI system stakeholders to focus on the measures that they have not implemented yet. You may want to provide feedback about clear fairness metrics. Explain that model development driven by clear and measurable fairness metrics (e.g., equal opportunity, demographic parity) ensures that fairness is a primary objective. Discuss fairness constraints in model training. Describe how incorporating fairness constraints during model training enforces adherence to fairness objectives. This can reduce the risk of biased decisions. In terms of fairness-aware algorithms, please highlight that the use of such algorithms reduces disparate impact and enhances fairness in AI decisions. Illustrate also that these algorithms are designed to produce equitable results. In case you refer to adversarial networks for fairness, explain that training models using adversarial networks makes them resistant to adversarial attacks and improves their fairness. You may also want to stress the importance of involving human reviewers and subject matter experts in model development and evaluation. This involvement provides domain-specific insights and ensures fairness. Your feedback may also describe how continual monitoring of AI system outputs for fairness ensures that corrective actions are taken if bias or unfairness is detected. Explain that this measure maintains fairness over time and promotes responsiveness.

In terms of diversity in development teams, it is crucial to highlight the importance of diversity in AI development teams, as diverse teams bring a wide range of perspectives. This is a key for enhancing sensitivity to bias and fairness issues. Finally, discuss user feedback mechanisms and how they empower users to contribute to fairness improvement. Consider also looking at the feedback guide for the question on bias mitigation, as the bias mitigation measures are relevant to fairness as well.

Table 11: Trustworthiness Auditing Feedback Guidelines for Question 10

Q10: What measures do you implement to ensure compliance with ethical standards and guidelines in manufacturing? **Feedback Guidelines**: Please provide feedback about the different measures of the question, prompting the AI system stakeholders to focus on the measures that they have not implemented yet. In terms of ethical AI development principles, please explain that aligning AI development with ethical principles ensures that the system reflects your manufacturing organization's values and industry standards. In case your feedback addresses the need for a code of conduct, please describe how a comprehensive code of conduct outlines ethical principles for AI development and use. Recommendations about AI Ethics Committees must highlight how such committees oversee the system's development and operation in terms of ethical and responsible AI principles. In terms of accountability measures you must explain that measures to hold individuals and teams accountable for the ethical use of AI promote responsible behaviour and adherence to ethical standards. In terms of legal compliance, it is important to stress that developing, deploying, and operating the system in line with relevant laws and regulations ensures legal compliance and ethical behaviour. Your feedback about education and training shall explain that educating employees, stakeholders, and end-users on AI ethics, privacy, and responsible use of the AI system builds awareness and fosters ethical behaviour. You may also want to describe how ensuring that AI components and technologies meet ethical standards (e.g., labour practices, environmental responsibility) demonstrates commitment to ethical values. In terms of regular monitoring, you will have to explain that regular monitoring of the system's behaviour and performance detects and rectifies ethical issues. You may also want to highlight the importance of mechanisms for individuals to report ethical concerns and violations related to AI systems without fear of retaliation. Finally, in case your feedback refers to environmental ethics, it is crucial to explain that considering environmental ethics in AI strategies reduces the environmental impact of AI. This is a key for contributing to ambitious sustainability goals.

Table 12: Trustworthiness Auditing Feedback Guidelines for Question 11

Q11: What measures do you implement to ensure the quality of data of the AI system?
Feedback Guidelines: Please provide feedback about the different measures of the question, prompting the AI system stakeholders to focus on the measures that they have not implemented yet. Refer to the importance of data collection guidelines. Explain that clear and well-defined data collection guidelines ensure that data is collected in a structured and consistent manner. In terms of precise annotation standards you had better describe how precise and consistent data annotation standards, along with clear instructions to human annotators, ensure that labeled data is accurate and reliable. Your feedback may also want to address the merits of data cleaning processes in terms of removing duplicate records, correct inaccuracies, and handle missing data. Likewise, you may provide feedback on data verification focusing on how data verification processes identify anomalies or errors. This is key to ensuring that data is accurate and reliable. In case your feedback refers to version control for datasets, it is important to stress the value of maintaining different versions of datasets to track changes and updates. This boosts reproducibility and historical reference. In terms of outlier identification, you had better describe measures for identifying and handling outliers in the data. This ensures that extreme values do not unduly influence AI model performance. Your feedback may also refer to data quality metrics that provide objective measures of data quality and boost data quality assessment processes. In terms of bias mitigation and its impact on data quality, it is crucial to highlight the importance of employing bias mitigation measures, especially for sensitive attributes. You may also want to explain why and how documenting metadata (e.g., data sources, collection methods) is a key for transparency, data understanding and accountability. In terms of data retention and disposal policies, please describe how these policies ensure efficient and secure data management by retaining data only as long as necessary. Finally, your feedback shall consider data backup processes. Explain that regular data backup prevents data loss due to accidental deletions or technical issues, ensuring data availability and integrity.

Table 13: Trustworthiness Auditing Feedback Guidelines for Question 12

Q12: What measures do you implement to ensure the cyber-security of the AI system?

Feedback Guidelines: Please provide feedback about the different measures of the question, prompting the AI system stakeholders to focus on the measures that they have not implemented yet. Your feedback shall explain the cyber-security measures that are important for ensuring the secure operation of the AI system. It should emphasize on the measures that are not implemented yet. For instance, in terms of data encryption, you should underline that encrypting data at rest and in transit safeguards data from unauthorized access. Likewise, when it comes to access controls, you will have to explain how robust access controls restrict who can access the AI system and what actions they can perform. Access control is essential to preventing unauthorized usage. You may also want to explain the principle of least privilege. Highlight that applying the principle of least privilege minimizes access rights for users and processes i.e., reducing the attack surface. In terms of strong authentication, explain that strong authentication methods, such as MFA, verify the identity of users, enhancing security. You should also provide feedback about standard and popular cyber-security processes like patch management, intrusion detection, intrusion prevention, creation of incident response plans, penetration testing, vulnerability scanning, code reviews, as well as firewalls and network segmentation. These measures are not peculiar to AI deployments. Rather they concern all the digital infrastructures of the organization. In terms of security audits and vulnerability assessments, it is good to explain that regular security audits and vulnerability assessments can be used to identify weaknesses in systems and infrastructure. Finally, it is important to provide feedback about user training and security by design. Specifically, you can describe how training users on security best practices enhances the human element of cybersecurity. Likewise, incorporating security considerations from the early stages of development, in-line with "security by design" approaches, builds security into the system's foundation.

Table 14: Trustworthiness Auditing Feedback Guidelines for Question 13

Q13: What measures do you implement for human oversight and intervention when necessary to ensure that AI decisions align with human values and intentions? **Feedback Guidelines**: Please provide feedback about the different measures of the question, prompting the AI system stakeholders to focus on the measures that they have not implemented yet. In terms of policies and guidelines, you had better explain that clear policies and guidelines for human oversight outline when and how human intervention is required. This measure can define the boundaries of AI decision-making. In the case of feedback about decisions' explanations please explain why it is important for the system to provide explanations for its decisions in a human-understandable manner. Stress that this measure helps users comprehend and trust AI outcomes. With respect to triggers for human intervention, please highlight that such triggers or thresholds ensure that humans remain in control of critical decisions. Hence, this measure safeguards against unexpected AI actions.

In terms of human involvement in decision-making please explain that complex or sensitive decisions can greatly benefit from human judgment. This measure maintains accountability and ethical decision-making. If your feedback concerns safety checks, make sure to explain that incorporating redundant systems and safety checks requiring human approval before certain actions are taken, reduces the risk of unintended outcomes. You may also want to comment on measures for bias review and correction. Discuss how human oversight is employed to review and correct potential biases, preventing discriminatory or unfair AI behaviour. Feedback Mechanisms are also important for human oversight. Highlight the presence of feedback mechanisms for end-users to report concerns or disputes, which can trigger human review and intervention. Along with feedback mechanism, you may also want to refer to review Panels and expert teams Stress the importance of established review panels or teams consisting of experts and stakeholders that periodically evaluate AI decisions and make necessary adjustments. Moreover, you may wish to explain that users can customize AI behaviour within certain limits, enabling them to align the system with their values and intentions. Finally, illustrate the importance of considering ethical frameworks. Discuss that systems which adhere to ethical AI frameworks and principles, are more likely to promoting responsible and ethical use of AI.

Table 15: Trustworthiness Auditing Feedback Guidelines for Question 14

Q14: What measures do you implement to provide comprehensive documentation for the AI system? **Feedback Guidelines**: Please provide feedback about the different measures of the question, prompting the AI system stakeholders to focus on the measures that they have not implemented yet. Explain the importance of the different types of feedback and documentation. Explain which types of documentation are important for different stakeholders' groups. For instance, API documentation is primarily important for developers, while the documentation of audit trails and explanations are also important for regulatory authorities and other auditors of the system.

4 STAR Technologies and Developments Linked to the Auditing Framework

STAR researches and develops several technologies that are aimed to support the development, deployment and operation of trustworthy AI. Using these technologies, manufacturers and industrial automation solution providers can meet some of their trustworthy AI goals. In terms of the presented auditing framework, STAR's technologies can be used to increase the trustworthiness score of AI systems for production lines. Nevertheless, STAR technologies support only a fraction of the trustworthiness measures that are considered and listed in the questions of the auditing framework. This is due to the following main reasons:

- STAR research does not cover the full range of AI trustworthiness topics. For instance, the project does not research solutions for AI fairness and bias mitigation, beyond standards solutions.
- Many of the presented measures are implemented as part of the established infrastructures and processes of an organizations. For instance, many of the listed cyber-security, data management and data storage mechanisms are implemented for all IT projects rather than in the scope of the implementation of a trusted AI system.
- STAR's use cases leverage some of the readily available infrastructures and processes of the partners in order to implement trusted AI use cases. This means that they do not rely on the STAR prototypes for the implementation of the full range of measures that support AI trustworthiness.

Table 16 illustrates how some of the main results of STAR can be used to implement selected measures mentioned in the auditing framework. Note that the list and terminology of STAR components is directly derived from the initial list of STAR exploitable assets in deliverable D8.7. A detailed technical description of these components is however provided in other STAR deliverables, as well as in the STAR book [Soldatos21].

Table 16: Overview of how STAR Results Map to Measures of the Auditing Framework

STAR Result	**Auditing Framework Relevance (Examples)**
Blockchain-based Data Provenance and Traceability Solution	The component can be used to implement provenance as an accountability measure (Q6).
AI Cyber-Defence tool	The component can be used to implement AI cyber-security measures (Q12)
Risk Assessment and Mitigation Engine (RAME)	The component can be used to implement cyber-security measures (Q12), as well as data management measures (Q5)
Security Policies Manager (SPM)	The component can be used to implement cyber-security measures and policies (Q12).
Simulated Reality (SR)	The component can be used to implement data augmentation strategies for bias mitigation (Q7) and data collection (Q3).
Active Learning (AL)	The component facilitates the implementation of AI models that foster human-AI collaboration and is therefore suitable for human oversight measures (Q13).

Production Processes Knowledge Base (PPKB)	This component facilitates access to domain knowledge about processes that can foster explainability and interpretability (Q2).
Natural Language Processing (NLP)	This component enables human interaction with the AI system and is relevant to the human oversight measures (Q13).
Feedback Module	This is one more component that enables human interaction with the AI system and is relevant to the human oversight measures (Q13).
AMR Safety	This component fosters safe interactions between humans and automated robots, which supports the implementation of human oversight measures (Q13).
Human Centred Digital Twin	This component fosters the implementation of human oversight measures (Q13).
Fatigue Monitoring System (FaMS)	This component fosters the implementation of human oversight measures (Q13).
STAR Training Services (D7.5)	The STAR training programs and services support users' education and upskilling. As such they can support the education/training related measures for different questions such as cyber-security (Q12) and data management (Q5).

STAR is finalizing a FRAIA tool, which can be used as part of a broader risk management tool, legal and ethical compliance tool, and a decision-making support tool for the STAR technologies and use cases. Fundamental Rights and Algorithms Impact Assessment can be used to implement ethical directives and regulatory compliance measures, such as the measures listed in Q10.

Overall, STAR provides technologies that support many different trustworthiness measures of the presented auditing framework.

5 Conclusion

STAR develops trusted AI systems and services, which are aimed at boosting the trustworthiness of AI systems for industrial use cases. STAR's technologies are very appropriate for boosting different aspects of AI trustworthiness such as data provenance and traceability, explainable AI, and AI cybersecurity. Nevertheless, AI trustworthiness includes many other technological, organizational, and social aspects beyond STAR's technical results. This deliverable has presented a broader view of AI trustworthiness for production lines and manufacturing use cases. It has introduced a framework for auditing the trustworthiness of AI systems, considering a wide array of parameters such as transparency, robustness, accountability, cyber-security, ethical and regulatory compliance, data quality, data management, data collection, bias mitigation, fairness, system documentation and more. These parameters have been integrated, structured, and organized in the context of an auditing framework that can help manufacturers and providers of industrial solutions in a dual manner:

- It enables them to assess the trustworthiness of their AI systems in a quite objective and quantitative way. The framework is able to produce a trustworthiness score that can be used to indicate and level of trustworthiness of an AI solution, including how it compares to other AI systems in terms of trustworthiness.
- It provides them with feedback and ideas for improving the trustworthiness of their systems. In this direction, the framework includes a feedback provision guide, which auditors can use to provide users of the framework with ideas for improvement.

A simple trustworthiness scoring mechanism has been suggested, which has some possible limitations. The latter has been discussed and can be overcome based on a proper configuration of the framework. In this context, the presented auditing framework is primarily aiming at helping organizations to create an AI trustworthiness culture beyond individual measures. Organizations can later configure the framework in order to prioritize certain measures and give less emphasis on others.

This is first version of the trustworthiness auditing framework of the project, which will be improved in the scope of the next and final version of the deliverable (i.e., D7.7). To this end, we aim to solicit feedback about the framework from STAR partners and other experts in AI in manufacturing, including feedback from the industrial pilots of the project. This feedback will be considered in enhancing and fine-tuning the auditing framework. Moreover, we plan to provide complementary assets that will boost the sustainability and wider use of the framework. Such assets include a mini training tutorial about the different questions of the framework and the trustworthiness improvement measures that they suggest, as well as a plan for potential standardization of this auditing tool. Also, the tool will be made available on-line in the scope of the STAR marketplace, which will boost its dissemination and wider use by AI and manufacturing communities. Finally, the final version of the deliverable will provide an outlook for the evolution of the presented auding framework to a certification program. This is likely to require partnership with some certification organizations, given that data and AI certification does not fall within the main activities of the STAR partners i.e., the STAR consortium does not include a certification organization.

References

[AI HLEG20] 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

[Bellamy19] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, A Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, Vol. 63, 4/5 (2019), 4--1. https://arxiv.org/abs/1810.01943.

[Byabazaire20] J. Byabazaire, G. O'Hare and D. Delaney, "Data quality and trust: Review of challenges and opportunities for data sharing in iot", Electronics, vol. 9, no. 12, pp. 2083, 2020.

[Dwork12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference. ACM, 214--226.

[EUAI23] EU Parliament, AI Act, https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-regulation-on-artificial-intelligence?sid=7101

[Elkhawaga23] Elkhawaga, G.; Elzeki, O.; Abuelkheir, M.; Reichert, M. Evaluating Explainable Artificial Intelligence Methods Based on Feature Elimination: A Functionality-Grounded Approach. Electronics 2023, 12, 1670. https://doi.org/10.3390/electronics12071670

[Fatouros23] G. Fatouros, G. Makridis, A. Mavrogiorgou, J. Soldatos, M. Filippakis and D. Kyriazis, "Comprehensive Architecture for Data Quality Assessment in Industrial IoT," 2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT), Pafos, Cyprus, 2023, pp. 512-517, doi: 10.1109/DCOSS-IoT58021.2023.00085.

[Gilpin18] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning", 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), pp. 80-89, 2018.

[Karkouch16] A. Karkouch, H. Mousannif, H. Al Moatassime and T. Noel, "Data quality in internet of things: A state-of-the-art survey", Journal of Network and Computer Applications, vol. 73, pp. 57-81, 2016.

[Lee21] Michelle Seng Ah Lee and Jatinder Singh. 2021. Risk Identification Questionnaire for Detecting Unintended Bias in the Machine Learning Development Lifecycle. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21). Association for Computing Machinery, New York, NY, USA, 704–714. https://doi.org/10.1145/3461702.3462572

[Litman23] J. Litman, Measures for explainable ai: Explanation goodness user satisfaction mental models curiosity trust and human-ai performance, 2023.

[Liu23] Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Yunhao Liu, Anil Jain, and Jiliang Tang. 2022. Trustworthy AI: A Computational Perspective. ACM Trans. Intell. Syst. Technol. 14, 1, Article 4 (February 2023), 59 pages. https://doi.org/10.1145/3546872

[Makridis23] G. Makridis et al., "Towards a Unified Multidimensional Explainability Metric: Evaluating Trustworthiness in AI Models," 2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT), Pafos, Cyprus, 2023, pp. 504-511, doi: 10.1109/DCOSS-IoT58021.2023.00084.

[Mazumder22] M. Mazumder, C. Banbury, X. Yao, B. Karlaš, W. G. Rojas, S. Diamos, G. Diamos, L. He, D. Kiela, D. Jurado et al., "Dataperf: Benchmarks for data-centric ai development", arXiv preprint, 2022.

[NLGov22] Government of the Netherlands, Fundamental Rights Algorithm impact assessment, https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms

[Ribeiro16] M. T. Ribeiro, S. Singh and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier", Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144, 2016.

[Soldatos21] John Soldatos (ed.), Dimosthenis Kyriazis (ed.) (2021), "Trusted Artificial Intelligence in Manufacturing: A Review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production", Boston-Delft: now publishers, http://dx.doi.org/10.1561/9781680838770

[Vilone21] Giulia Vilone, Luca Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence", Information Fusion, Volume 76, 2021, Pages 89-106, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2021.05.009.