Fraunhofer IKS

The European Artificial Intelligence Act
Overview and Recommendations for Compliance

Fraunhofer IKS

L. Heidemann, B. Herd, J. Kelly, N. Mata, W. Tsai, S. Zafar, A. Zamanian

13.05.2024

Abstract

The European Union's AI Act establishes ethical guidelines and a regulatory framework for the development, deployment, and use of Artificial Intelligence (AI) systems in the European Union. In this whitepaper, we provide an overview of the key provisions outlined in the EU AI Act, addressing risk-classification, stakeholder considerations, and requirements, with a particular focus on safety-critical and therefore high-risk AI systems. In the scope of the EU AI Act, AI systems are categorized based on risk levels, ranging from minimal- to unacceptable- risk, each with its corresponding set of regulatory obligations. High-risk AI systems are subject to rigorous requirements spanning risk management, data governance, transparency, and human oversight. This whitepaper delves into the specifics of Articles 9 to 15 of the EU AI Act, covering the requirements for high-risk AI systems while also identifying gaps with respect to existing safety standards. We propose a framework for bridging these gaps by deriving concrete requirements from the EU AI Act inspired by contract-based design. We leverage our expertise in trustworthy AI and safety to develop a framework for deriving argumentation trees for generic properties of Machine Learning (ML) systems. We demonstrate this framework on three practical use cases across various sectors. Our work illustrates how AI systems in safety-critical domains such as automotive, industrial automation, and healthcare can meet regulatory standards while upholding ethical principles. The EU AI Act represents a significant step towards fostering trust, accountability, and responsible innovation in AI technologies. By following systematic verification processes for EU AI Act requirements, stakeholders can navigate the complex AI landscape with confidence, ensuring the ethical development and deployment of AI systems while safeguarding human interests and values.

Contents

List of Figures

List of Tables

1 The European Artificial Intelligence Act: An Overview

The EU AI Act represents a landmark in the regulation of AI systems [2]. In line with the 2018 European Strategy for AI presented by the European Commission, the AI Act positions Europe in the competitive global landscape of AI. By providing a comprehensive framework governing the development, deployment, and use of AI systems in the European Union (EU), the AI Act facilitates the free movement of AI products and services across member states. The AI Act works towards safeguarding fundamental human rights, while fostering the development of human-centric and trustworthy AI technologies.

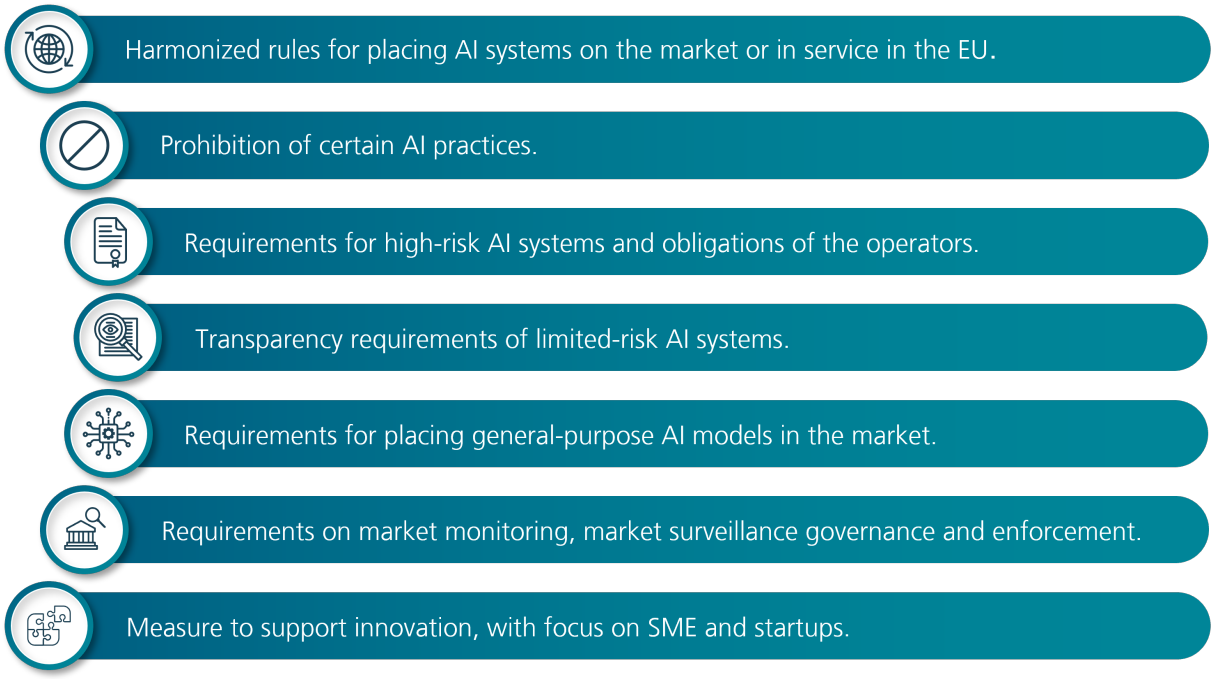

Figure 1: Overview of the EU AI Act Contributions.

Central to the interpretation of the legislation is the classification of AI systems based on their risk level. Depending on the level of risk they pose, AI systems must adhere to rigorous requirements to ensure safety, transparency, and human rights needs are met. These requirements extend beyond the AI products themselves to encompass all stakeholders involved in the AI value chain. The Act covers the prohibition of certain AI practices, requirements on high-risk AI systems, as well as transparency requirements for certain AI systems which pose limited risk (Figure 1). As organizations navigate the evolving landscape of the EU AI Act, they must strike a balance between innovation and regulatory compliance. Adapting to the Act's requirements will necessitate adjustments throughout the AI value chain, with implications for future standards and regulations in the field. This whitepaper presents an overview of the EU AI Act and proposes a systematic methodology for approaching compliance to the regulation across the value chain, particularly for safety-critical AI systems. Chapter 2 provides an overview of the EU AI Act and the risk-based classification, while Chapter 3 dives into the requirements for High-Risk AI Systems. Chapter 4 discusses the gap in high-level requirements and existing standards, and Chapter 5 addresses this gap by presenting a verification framework for AI systems. Chapter 6 presents three case studies tailored to specific industries, showcasing the practical effectiveness of the methodology.

1.1 Stakeholders and Value Chain Considerations

To understand how the EU AI Act will impact individuals or organizations, it is important to understand their roles within the Act's framework. The regulation applies to providers, deployers, importers, distributors, product manufacturers, and authorized representatives of providers of AI systems. The requirements laid out focus on four key stakeholders: providers, deployers, importers, and distributors (Figure 2), of AI systems. The AI Act does not discuss obligations of end-users. Rather, the regulation focuses on the stakeholders involved in the placing on the market and distribution of AI systems.

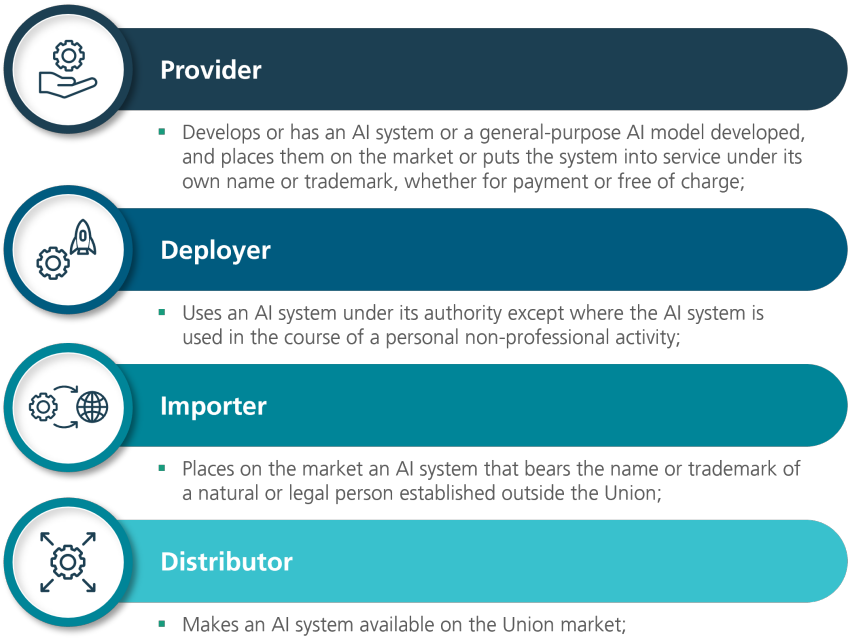

Figure 2: Key stakeholders of AI systems defined by the EU AI Act.

According to Article 25 of the EU AI Act, any deployer, distributor, importer, or other third party is considered a provider of an AI system if any of the following conditions hold:

- They place on the market or put into service a high-risk AI system under their name or trademark;
- They modify the intended purpose of a high-risk AI system already placed on the market or put into service;
- They make a substantial modification to the high-risk AI system.

This article has substantial implications for stakeholders involved in value chains for high-risk AI systems. If any of the above conditions hold, these stakeholders will need to ensure compliance to the requirements for providers of high-risk AI systems are met. Not only will this affect stakeholders in the EU, but any third party located outside of the EU wishing to place their systems on the EU market will need to demonstrate compliance. The importance of establishing value chain accountability for AI governance is highlighted in [3], and this whitepaper discusses and extends the methodology presented in [4] to address this.

1.2 Risk-based classification of AI-systems

Central to interpreting the EU AI Act's requirements is the risk-based classification approach for AI systems (Figure 3). The Act defines requirements on AI systems which are proportional to the level of risk that the

systems pose. To ease interpretation of the Act requirements, we follow the categories outlined in Figure 3. While only unacceptable- and high- risk AI systems are explicitly named in the AI Act, additional transparency requirements are defined for certain systems posing limited-risk. Any system not falling under the previous three categories is assumed to pose minimal or no risk and can follow a voluntary code of conduct.

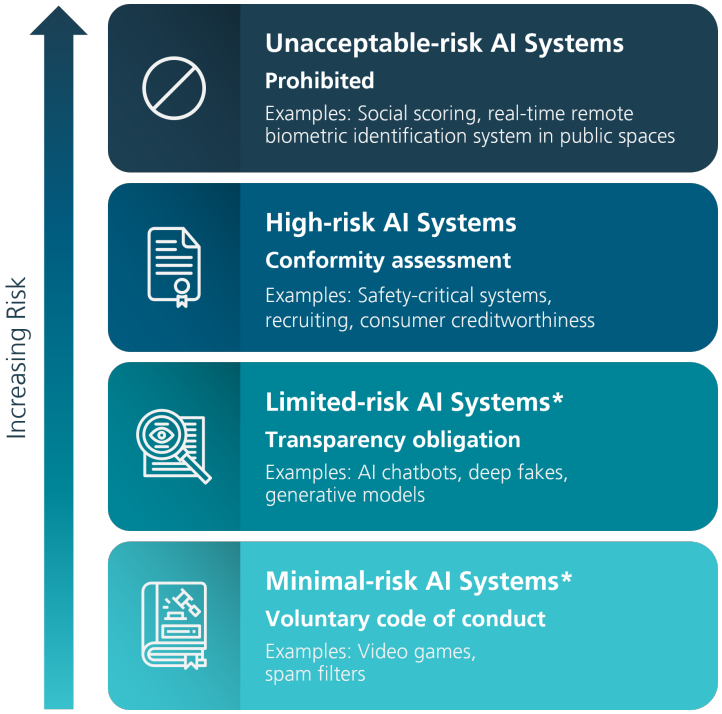

Figure 3: Overview of the risk-based classification useful for intepreting the EU AI Act. Categories marked with a * are not explicitly mentioned in the Act, however are useful for the interpretation of the requirements.

AI systems which are rated with **Unacceptable-Risk** are prohibited under the provisions of the Act. These are AI systems that pose clear threats to human safety or rights. Examples of such systems include:

- Subliminal or purposefully manipulative or deceptive techniques to modify or distort an individual's behavior.
- Exploitation of vulnerabilities due to age, disability, or social or economic situation.
- Biometric categorization systems (except for labeling or filtering of lawfully acquired biometric datasets, or for law enforcement purposes).

High-risk AI systems must abide by the requirements for high-risk AI systems laid out in the Act. These are grouped into two categories:

- Safety components/products that are already subject to EU safety legislation (listed in Annex II of the Act) and are required to undergo a third-party conformity assessment.
- AI systems expressly designated as high-risk by the European Commission (EC) (Annex III of the Act), e.g.:
  - Migration, asylum, and border control management systems.
  - Systems used in the administration of justice and democratic processes.

Limited-risk AI systems need to comply with transparency obligations defined in the regulation, while

AI systems whose risk is deemed **minimal** can follow a voluntary code of conduct defined in the Act. These include any other AI system that is not categorized as limited-/high-/unacceptable-risk. We dive further into the requirements for high-risk AI systems.

2 Requirements for high-risk AI-systems

Articles 9-15 of the EU AI Act outline obligations for high-risk AI systems. In general, the requirements are high-level, and concrete guidelines for compliance are not provided. Rather, it is the role of the provider to demonstrate that these requirements are met. The following sections provide a look into each of the articles, outlining relevant implications for providers of high-risk AI systems.

2.1 Article 9: Risk management system

Article 9 states that a risk management system shall be established, implemented, documented, and maintained throughout the lifecycle of the high-risk AI system. This risk management lifecycle shall include the following stages:

- **Risk Identification and Analysis:** Known and foreseeable risks to health, safety, and fundamental rights.
- **Risk Mitigation Measures:** Elimination, reduction, and mitigation measures such that the overall residual risk of AI system is less than or equal to acceptable risk.
- **Testing:** Including real-world testing to identify appropriate risk mitigation measures and compliance with requirements before market deployment.
- **Post-market Monitoring and Update:** Evaluation of additional risk by analyzing post-market monitoring data (more details in Article 72 of the EU AI Act).

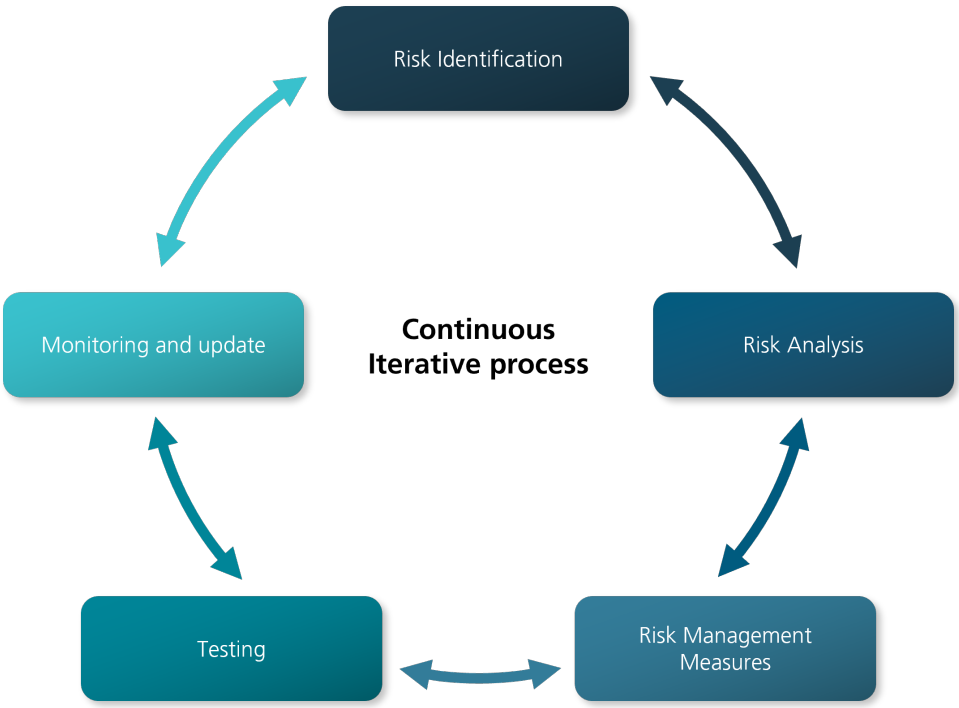

Figure 4: Overview of risk management lifecycle for high-risk AI systems.

2.2 Article 10: Data and data governance

Article 10 states that training, validation, and testing data sets shall meet quality criteria, where applicable. Figure 5 provides an overview of this article.

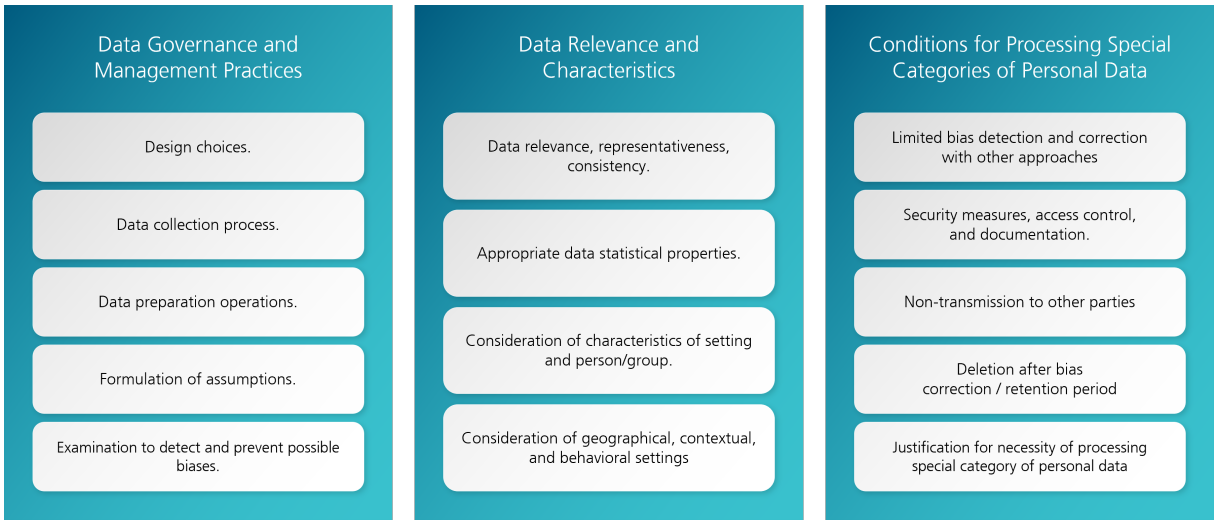

Figure 5: Quality criteria for Data discussed in the EU AI Act.

2.3 Article 11: Technical documentation

Article 11 lays down guidelines for technical documentation of AI systems. The article states that the technical documentation of a high-risk AI system shall be drawn up before that system is placed on the market or put into service and shall be kept up-to-date. This documentation enables national competent authorities or notified bodies to assess the compliance of system to the requirements in the EU AI Act. It provides clear and comprehensive information to assessing bodies. Examples of what the documentation should contain include, for example:

- A general description of the AI system.
- A detailed description of the elements of the AI system and of the process for its development.
- Detailed information about the monitoring, functioning, and control of the AI system.
- A description of the appropriateness of the performance metrics for the specific AI system.

2.4 Article 12: Record keeping

Article 12 mandates the automatic recording of events ('logs') over the duration of the lifetime of the system. These logs should facilitate the post-market monitoring to evaluate compliance with EU AI Act requirements, and the identification of risk posed by the high-risk AI products at the national level. The article lays down minimum logging requirements, such as recording the period of use.

2.5 Article 13: Transparency and provision of information to deployers

Article 13 addresses the transparency of the AI system itself. It states that high-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable deployers to interpret the system's output and use it appropriately. The article also enforces the provision of

instructions of use alongside the AI system, and lays down minimum content requirements for these usage instructions. An overview of these is shown in Figure 6.

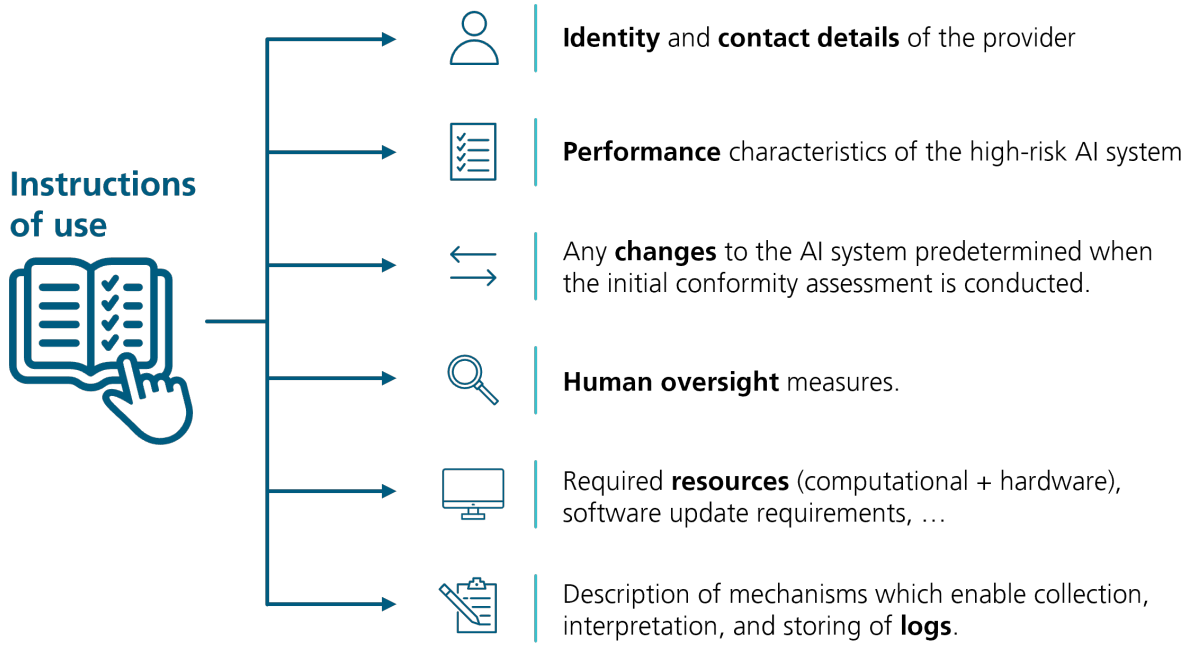

Figure 6: Contents which must be present in the instructions of use, outlined by Article 13.

2.6 Article 14: Human Oversight

Article 14 outlines requirements for human oversight. It states that high-risk AI-systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use. The AI system should be equipped with appropriate features which enable natural persons to effectively oversee the operation of the AI system, as appropriate and proportionate. A summary of these capabilities is provided in Figure 7.

Figure 7: Summary of the capabilities that human oversight, as defined by Article 14, should enable the natural person to have.

2.7 Article 15: Accuracy, robustness and cybersecurity

Article 15 requires high-risk AI systems to be designed and developed such that they achieve an appropriate level of accuracy, robustness, and cybersecurity, and perform consistently in those respects throughout their lifecycle. The Act recommends the establishment of benchmarks and measurement methodologies for high-risk AI systems through benchmarking organizations. The article also places requirements on the resilience of the system to errors, faults, or inconsistencies, particularly due to their interaction with natural persons or other systems. Resilience to attempts by third parties to alter their use, outputs, or performance should also be ensured.

2.8 Gap in EU AI Act and Existing Standards

The EU AI Act requirements are high-level, and upcoming standards and regulations will need to be aligned with the requirements laid out by the Act. Existing road-vehicle safety standards, for example, align well with Article 9 of the EU AI Act and provide a systematic approach to address risks associated with AI systems in safety-critical applications. However, the standards have weaker coverage of certain requirements for human oversight and transparency. Industry specific standards will need to work towards addressing EU AI Act requirements. Stakeholders' perspectives must also be considered. In the following section, we explore how product quality models can address this gap, and present an extended quality model for AI systems.

3 Bridging the Gap: Deriving High-Level Requirements from the EU AI Act

Product quality models can be leveraged, along with existing safety standards, to achieve a more comprehensive coverage of EU AI Act requirements. ISO/IEC 24028 provides an overview of trustworthiness in artificial intelligence and highlights the need for new product quality standards which incorporate AI-specific quality attributes [5]. ISO/IEC 25059:2023 provides the quality model serving as an extension to the ISO 2501x:2011 series standards - Systems and Software Quality Requirements and Evaluation (SQuaRE). It defines quality attributes and sub-attributes that establish consistent terminology for specifying, measuring, and evaluating the quality of AI systems. In [4], we presented an extended product quality model based on ISO/IEC 25059 to address relevant topics from the EU AI Act. The extended model (Figure 8) includes attributes like ethical integrity, human oversight, and transparency of an AI system.

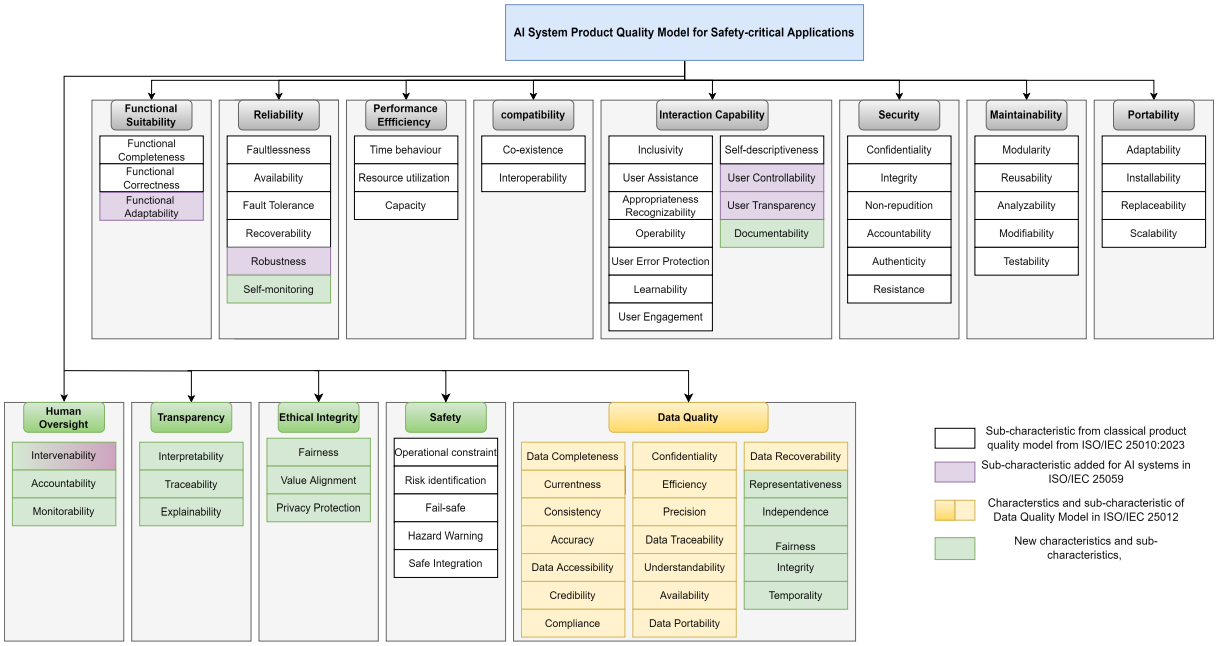

Figure 8: Extended product quality model incorporating AI-specific attributes.

3.1 Mapping to the EU AI Act

To assess coverage of the EU AI Act using the AI System Product Quality Model for Safety-Critical Applications presented in the previous section, we provide a mapping of Articles 9-15 from the EU AI Act with the sub-attributes in the quality model. This mapping is shown in Table 1. Using this methodology, the EU AI Act requirements can be addressed by verifying that the AI system satisfies the relevant quality attributes.

Table 1: Mapping of EU AI Act articles to quality attributes.

Article	Sub-Attribute Mapping
9. Risk Management System	Risk identification, Testability, Value Alignment
10. Data and data governance	Independence, Representativeness, Data Completeness, Currentness, Independence, Data Fairness, Compliance, Precision, Representativeness, Consistency, Accuracy, Credibility, Temporality, Confidentiality, Compliance, Data Traceability
11. Technical Documentation	Documentability, Traceability
12. Record-keeping	Operability, Non-repudition, Traceability, Self-descriptiveness, Accountability, Self-Monitoring, User engagement, Monitorability
13. Transparency and provision of information to users	User engagement, Self-descriptiveness, User Transparency, Interpretability, Documentability, Appropiateness recognizability
14. Human Oversight	Documentability, Learnability, Value Alignment, Accountability, Interpretability, Fairness, Explainability, Intervenability, Monitorability, User error protection.
15. Accuracy, robustness, and cybersecurity	Functional Correctness, Faultlessness, Robustness, Appropiateness recognizability, Self-descriptiveness, Functional adaptability, Fault Tolerance, Robustness, Integrity, Resistance

4 Verification Framework

The quality model presented in Section 3 provides a solution for breaking down high-level requirements into verifiable properties of AI systems. We leverage our expertise in trustworthy AI and safety to develop a methodology for deriving argumentation trees for generic properties of ML systems. Safety cases are often used to demonstrate the safety of safety-critical systems, and ISO 26262 requires that a structured assurance argument is presented to demonstrate functional safety [6]. A common practice in safety engineering is to represent safety cases using graphical notation such as GSN [1]. To align our methodology with best practices in safety engineering, we use GSN to develop our verification trees. This framework is not only applicable for the verification of the EU AI Act requirements but also for the verification of safety-related ML-properties. Our approach is based on best practices from contract-based design and GSN [1].

4.1 Contract-based approach: AI value chain considerations

Contracts play a crucial role in establishing clear expectations and accountability between different stakeholders and are particularly useful in the context of system design that involves potentially complex supply chains. By defining clear roles, responsibilities, and deliverables of each party, contracts ensure that all stakeholders have a common understanding of what is expected from them. In a contract-based approach, components are described in a way that includes both the guarantees offered by the component and the assumptions about its potential use cases or environment. Such a description specifies the expectations and responsibilities of the component in different scenarios. A definition of clear contracts in terms of guarantees and assumptions of each component serves as a basis for the establishment of a framework for effective collaboration, system integration, and verification & validation. We propose to leverage concepts from contract-based design to verify quality attributes in an AI value chain. We present typical stakeholders in an AI value chain in Table 2, and show how they may interact in Figure 2.

Table 2: Typical stakeholders in an AI value chain.

Stakeholder	Description
AI End User (END)	The natural person operating the AI system and/or using AI system outputs to inform their actions [7].
Deployer (DPY)	Any natural or legal person, public authority, agency or other body using an AI system under its authority, except where the AI system is used in the course of a personal non-professional activity. [7].
AI System Integrator (SYI)	The organization or entity that is concerned with the integration of AI components into larger systems, potentially also including non-AI components [8].
AI Product Provider (PPR)	The natural or legal person that places a generic or specific AI system on the EU market [7].
AI Service Provider (SPR)	The natural or legal person that provides AI support tools and/or services on demand [7].
AI Developer (DEV)	The natural or legal person that builds generic or specific AI systems at the behest of third parties but who do not place this product on the EU market [7].
Data Provider (DPR)	The natural or legal person that provides data for training, testing and/or validating generic or specific AI systems [7].

In the supply chain, we depict several sub-systems which are integrated into the overall AI system. We

propose to introduce the concept of Design Contracts (DCs) between stakeholders, which can be translated to Technical Requirements (TRs) between different stakeholders. Note that many sub-systems may exist which have technical requirements with the system provider and other stakeholders. The use of design contracts provides a methodological approach to addressing EU AI Act requirements (through the verification of quality attributes) at the stakeholder level. The responsibility of each stakeholder boils down to the demonstration that, considering certain assumptions hold, the guarantee that their system fulfills certain properties holds.

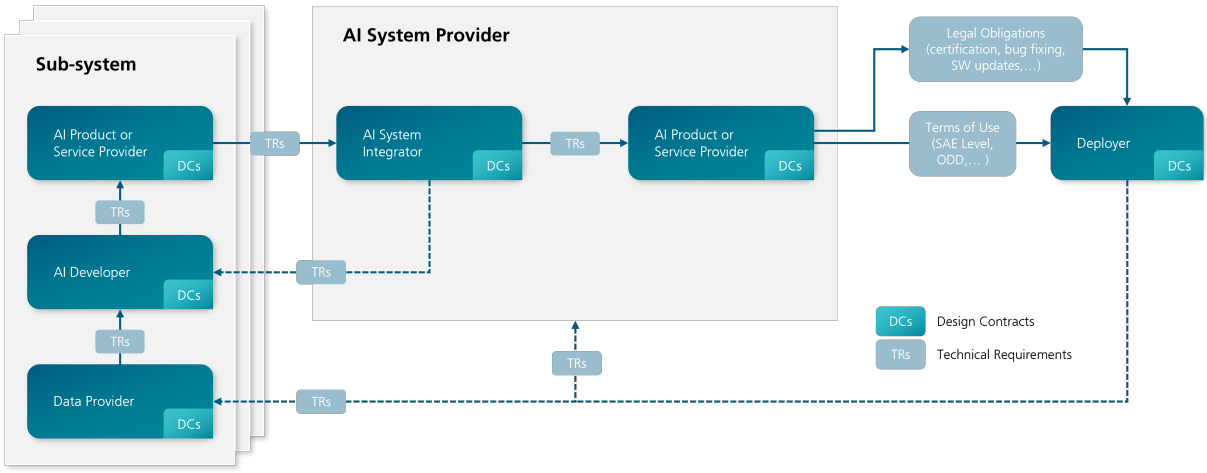

Figure 9: Stakeholders in a typical AI value chain.

4.2 Verification Process

To visualize the argumentations presented in the previous section, we propose to use verification trees inspired by GSN notation [1], and leveraging type taxonomies from CPS [9] and AI-RMF [10] frameworks. Goal Structuring Notation (GSN) is a graphical argument notation that can be used to document the individual elements of any argument (claims, evidence and contextual information) and their relationships. The resulting argument is called a goal structure. GSN elements are generic enough to allow documenting arguments in any domain. Figure 10 depicts an example goal structure, using all the main elements of GSN. These are described here.

The core element in GSN are claims and sub-claims, represented by the goal element (rectangles with **G** identifiers). A claim may be established upon a set of *assumptions* (ovals with **A**) and is defined in a specific *context* (rounded-corner rectangles with **C**) which determines the scope of the claim. *Justification* elements (ovals with **J** identifier) can be used to justify the usefulness or relevance of a claim. A *strategy* element (parallelograms with **S** identifier) contains the nature of an argument that connects a claim to its sub-claims.

Finally, by breaking down the claims many times, we reach claims that are close to tangible and auditable stages of AI development. Where the last sub-claim on a path can be substantiated by a piece of evidence, the evidence is represented by the *solution* element (circles with **Sn** identifier). Similar to claim and strategy elements, one can provide assumptions, contexts, and justifications for the evidence element.

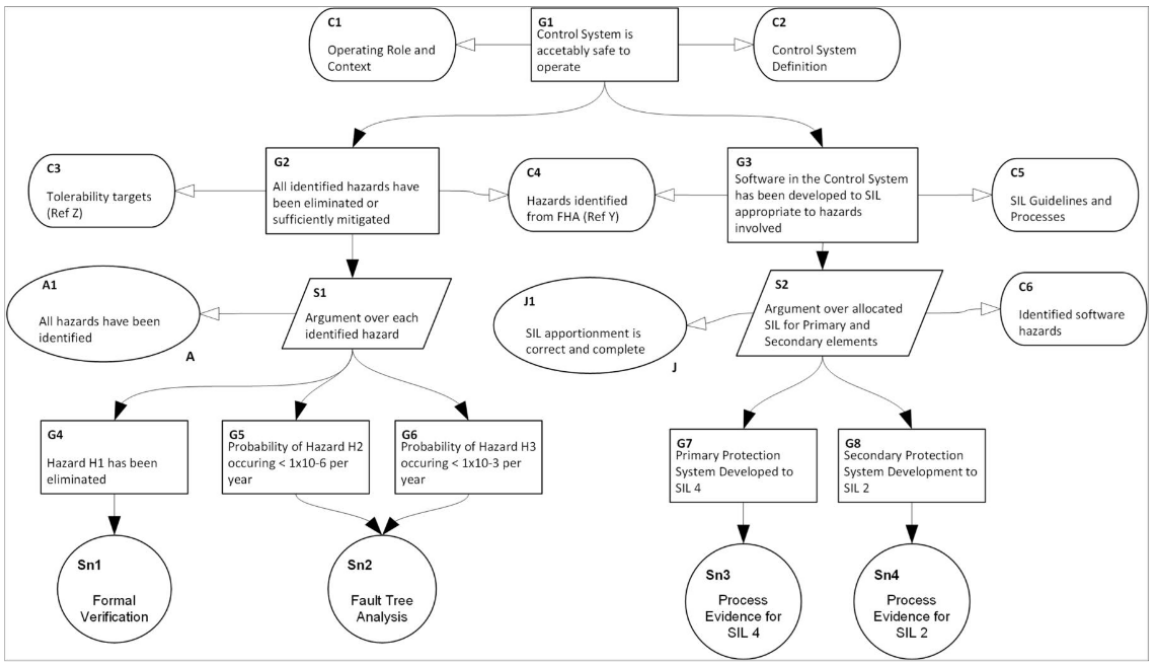

Figure 10: Example of an argument using GSN [1].

5 Use Cases

In this section, we demonstrate how our methodology can be applied across three industry-specific use cases. The flexibility of the methodology allows for extensions and modifications depending on the use case at hand. For each use case, we present the approach for verifying a different sub-quality attribute for a particular stakeholder in the AI value chain (Figure 8).

5.1 Automotive: Automated Parking System (APS)

The Automated Parking System (APS) is a generic function to be deployed in a automated vehicle with autopilot capabilities. The functionality of the APS is briefly described below:

- For parking, the user gets out of the vehicle, searches for free parking slots in the APS App, selects one parking slot and activates the function to drive autonomously to the parking slot and park the car.
- For unparking, the user starts the unparking functionality through the APS app and indicates the pick-up time and location of the vehicle (it can be immediately).

The autonomous vehicle shall follow the traffic rules and adapt the behaviour according to the traffic signs. The AI-based function used for the validation of our framework is a perception subsystem Traffic Sign Recognition (TSR).

Within the context of TSR, the following adjustments will be necessary:

- Regarding parking, the system should be capable of recognizing situations where parking is prohibited due to changes in traffic signs.
- When it comes to leaving a parking spot, the system must be able to determine if the user's current location permits stopping for a pick-up, using information from traffic signs.

An overview of the supply chain with Traffic Sign Recognition (TSR) as a sub-system for the APS Provider is shown in Figure 11. The TSR is the sub-system under validation for the Traceability quality attribute. Design contracts will be defined for the various stakeholders depicted in the supply chain. For example, a design contract for traceability may exist for the AI developer, leading to technical requirements between the data provider and the AI Developer. Note that many sub-systems may exist which have technical requirements with the APS Provider and other stakeholders. Only the TSR sub-system is shown in the figure below.

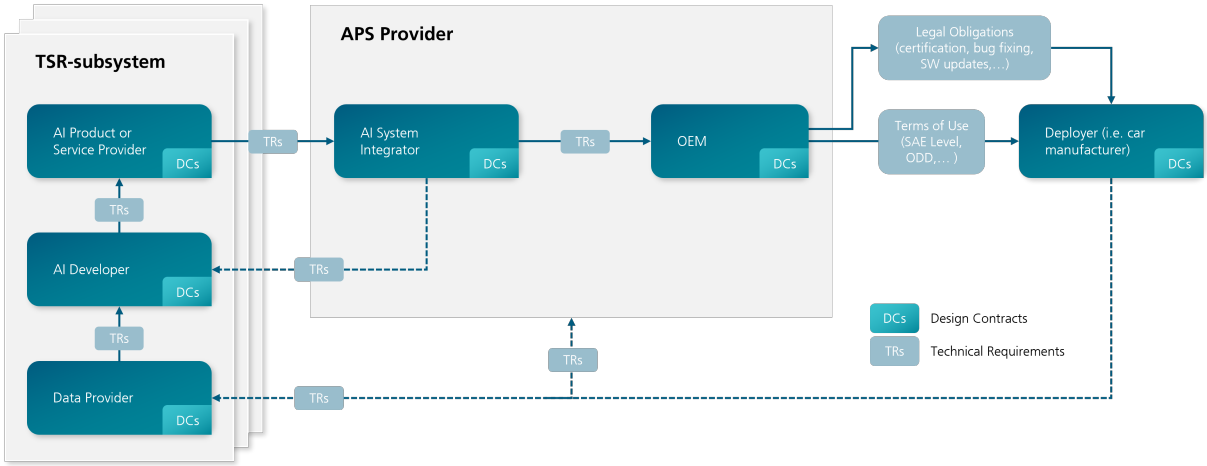

Figure 11: AI-value chain for an Automated Parking System (APS) and a Traffic Sign Recognition (TSR) sub-system.

5.1.1 Traceability Contracts for Traffic Sign Recognition for the AI System Integrator

In the supply chain shown in Figure 11, design contracts (DCs) may exist for several stakeholders. From these design contracts, technical requirements (TRs) can be formulated between stakeholders. The following are examples of design contracts for traceability. The design contracts are then translated into technical requirements for the stakeholders. Note that design contracts can also be developed for different quality attributes. In this case, the guarantee is that the APS subsystem fulfills the given quality attribute.

AI System Integrator

Assumptions

- The TSR is equipped with appropriate logging and monitoring functions which are available to the AI System Integrator.
- Appropriate documentation regarding the design, development, licensing, and usage restrictions of the TSR is available.

Guarantees

- Logging functionalities and relevant monitoring components from the TSR are integrated into the APS.
- Documentation regarding the TSR and how it interacts with other sub-systems in the overall APS is available.

Goal Structure for Traceability of the AI System Integrator

The first example goal structure is generated for the AI system integrator as shown in Figure 12. In this structure, the top-level goal is that the given stakeholder satisfies the relevant quality attribute. The strategy in the stakeholder view goal structure is an argumentation over the design contract guarantees. The

guarantees are thus depicted as sub-goals. In the stakeholder view, only evidences that the stakeholder is responsible for are included. These evidences are linked to the guarantees which the AI System integrator fulfills through its design contracts.

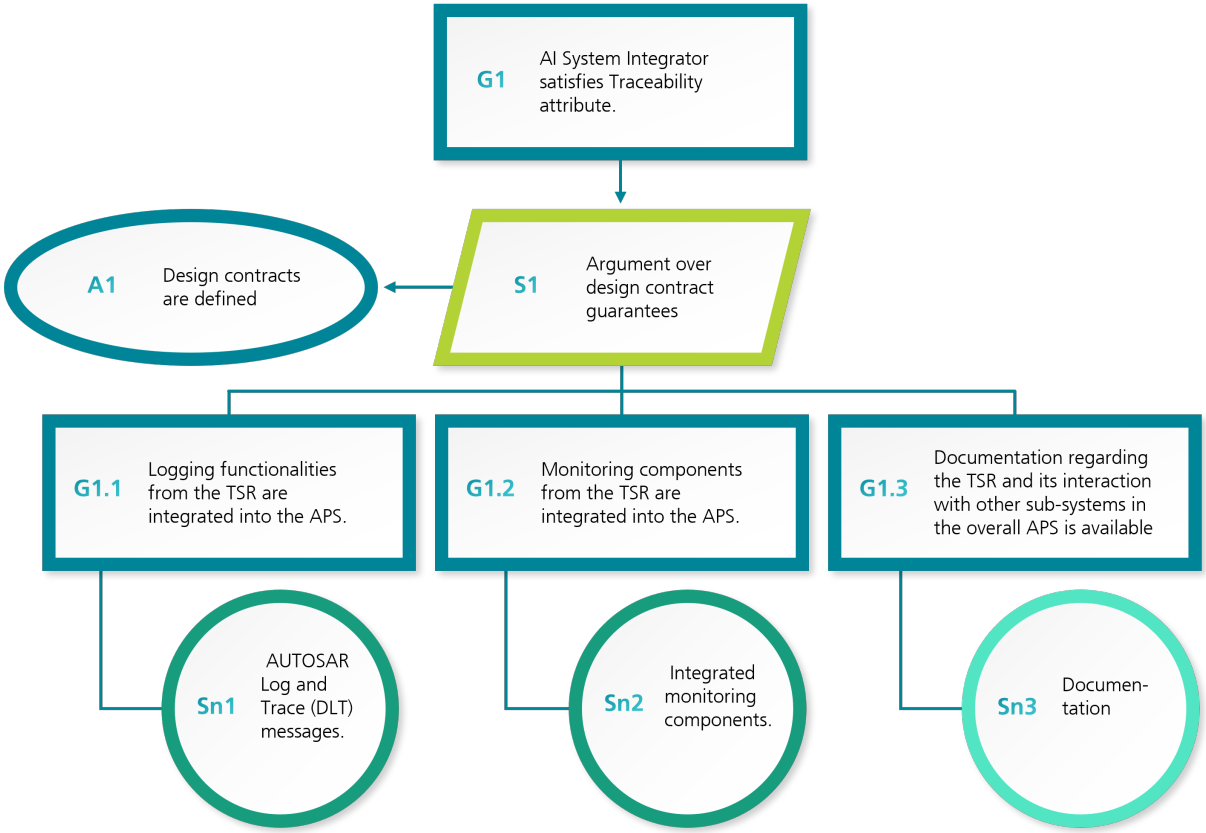

Figure 12: AI System Integrator: Goal structure for verification of traceability quality attribute.

5.2 Industrial Automation: Quality Inspection Cobot (QIC)

In industrial automation, we consider the use case of a Quality Inspection Cobot (QIC). The cobot can be programmed to perform the initial visual inspection of the components. It is equipped with cameras and sensors to detect any defects or anomalies on the components. The cobot is placed in a designated inspection area where human workers can interact with it. When a batch of components arrives for inspection, the cobot starts its inspection routine. It picks up each component, scans it with its cameras and sensors, and analyzes the data for any defects or deviations from the required specifications. If the cobot detects a potential issue, such as a crack or a missing part, it immediately stops its motion and notifies a human worker. The worker can then intervene, examine the component more closely, and make a final judgment on its quality. Once the human worker has resolved the issue and exits the inspection area, the cobot automatically resumes its inspection routine with the next component. Using a person detection monitoring system, this cobot must conduct a safety-rated monitored stop, when a human enters the collaborative area, and resumes its operation after the human left the area, according to ISO 10218-1/2 [11, 12]. An overview of the value chain for the person detection sub-system is shown in Figure 13.

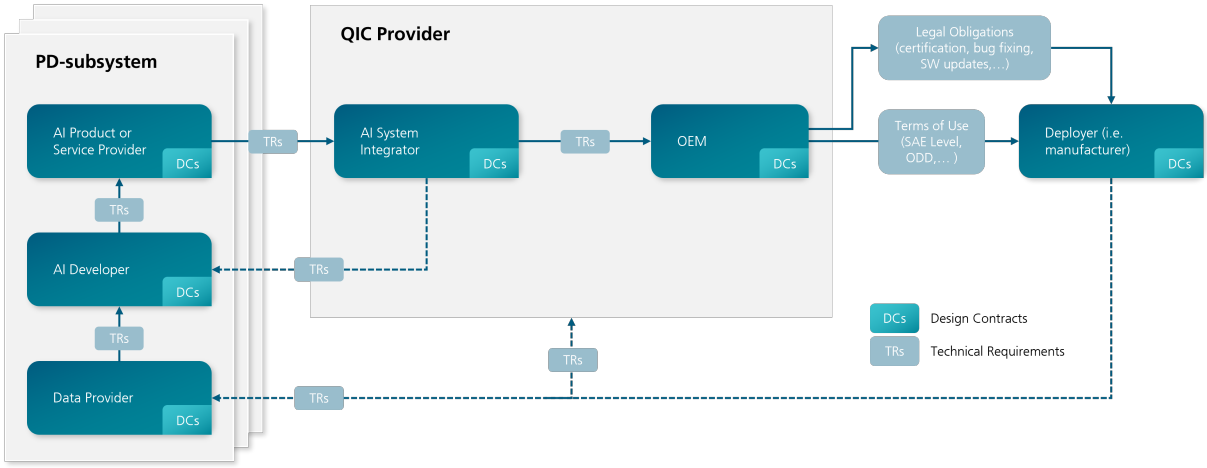

Figure 13: Value chain example for a Person Detection (PD) sub-system for a Quality Inspection Cobot (QIC).

5.2.1 Fairness Contracts for Person Detection for the AI Developer

The fairness quality attribute is particularly relevant for a person detection sub-system. Unwanted bias with potentially severe consequences could be introduced by engineering decisions or data bias [13]. For this use case, we focus on bias in engineering decisions and demonstrate an example of a fairness contract for the AI developer.

AI Developer

Assumptions

- The training data is representative of the demographic distribution defined in the Operational Design Domain (ODD).
- The training data for each protected attribute varies according to the operating conditions defined in the ODD (e.g., light).
- Relevant demographic attributes are annotated.
- The labeling procedure incorporates diverse perspectives and considerations.

Guarantees

- The model's output is not biased towards any protected attribute of the demographic distribution defined in the ODD.
- Attribute annotations are only used for bias evaluation and correction.

Goal Structure for Fairness of the AI Developer

The following figure shows an example of the goal structure generated for the AI developer of the person detection algorithm. The top-level goal is to fulfill the fairness quality attribute.

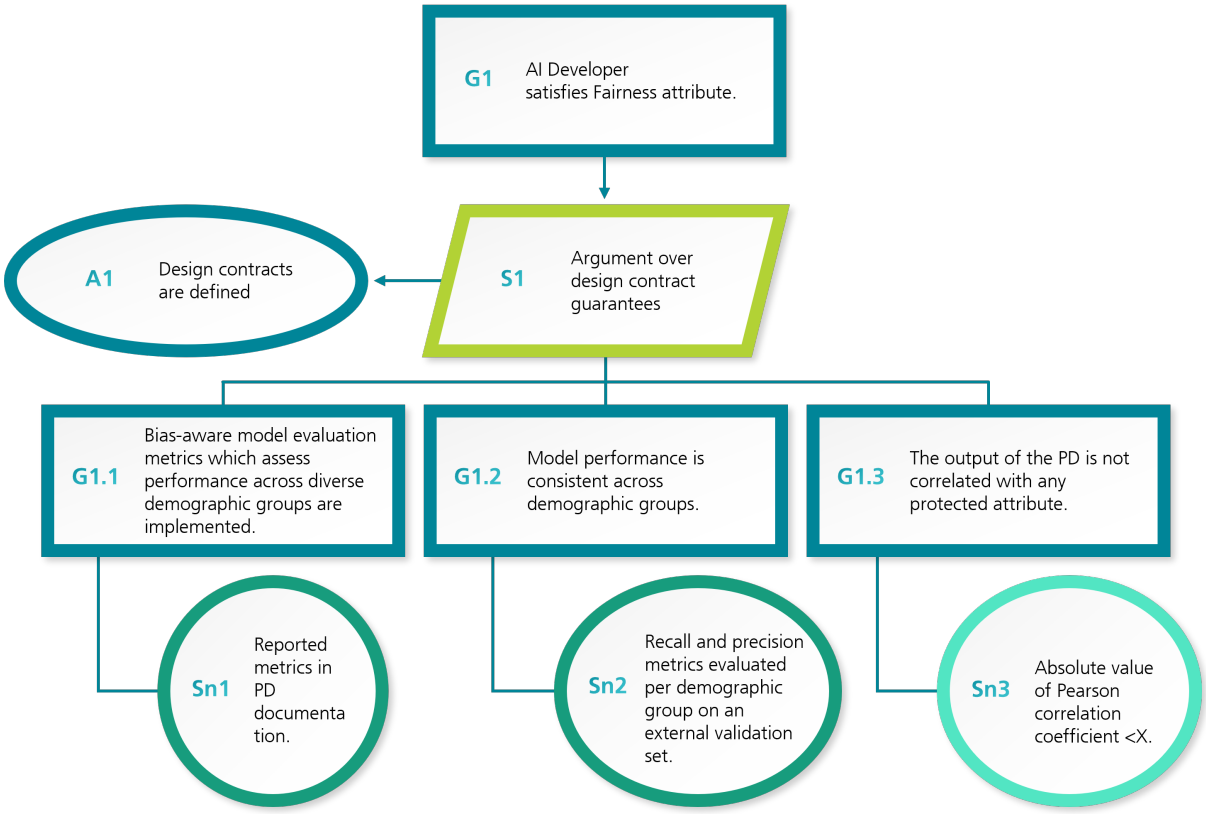

Figure 14: AI developer - Goal structure for verification of fairness quality attribute.

5.3 Healthcare: Brain-Computer Interface (BCI)

A Brain Computer Interface (BCI) is a technology that enables direct communication between the human brain and external devices. This interface operates through a sophisticated blend of hardware and software subsystems, enabling individuals to interact with their environment solely through their brain's electrical activity. BCIs function by capturing and interpreting brain signals to discern user intentions, employing a recording stage to measure and translate these signals into manageable electrical data. Currently, there are several types of monitoring technologies for BCI, to name one, the functional Near-Infrared Spectroscopy (fNIRS) [14]. In return, AI provides invaluable assistance in analyzing medical images and extract meaningful insights quickly and accurately. Using deep learning methods, the AI component can detect patterns, anomalies, and relevant features within images. This helps doctors with diagnostic, treatment planning, and decision support. In this use case, AI performs image classification (denoted as "CLF", the subsystem in Figure 15), assigning different classes to images (e.g., health related, action-type related). The level of

automation expected from the AI component in BCI may vary. Doctors usually rely on the AI classifier and only do post-analysis evaluation to ensure the AI sanity and results accuracy. Here lies the importance of explainability of the classifier; explainable AI ensures that clinicians understand the reasoning behind the AI-driven decisions and facilitates post-analysis evaluation.

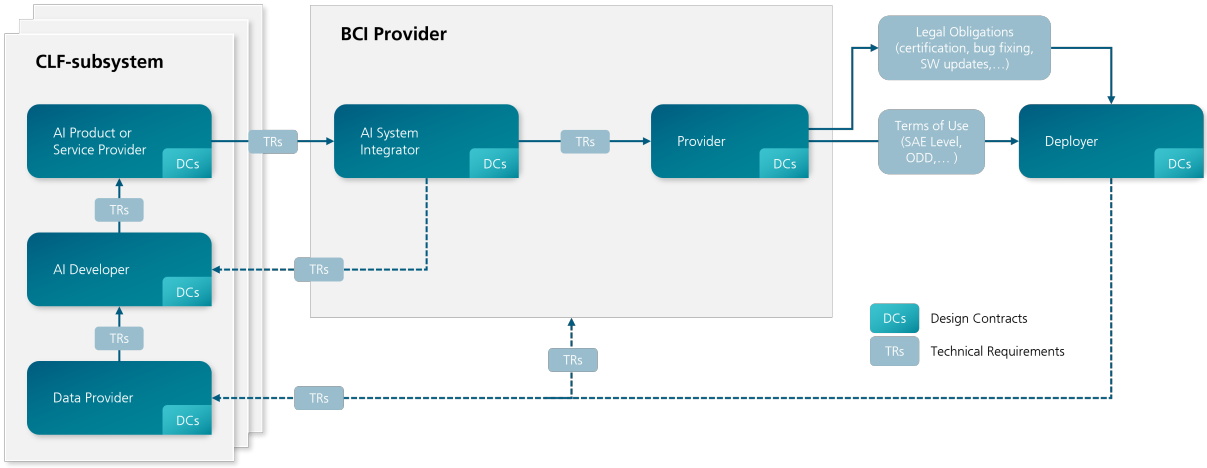

Figure 15: Value chain for Classifier (CLF) sub-system of Brain Computer Interface (BCI).

5.3.1 Explainability Contracts for the CLF subsystem for the AI Developer

The explainability module in CLF (denoted as "xAI") functions as a post-hoc explainability method, applied to the classifier's output (regardless of the classification methodology). This module generates saliency maps for the images, which are heatmaps highlighting the pixels that significantly influence the AI's decision. Here we provide the explainability contract for the CLF subsystem for the AI developer.

AI Developer

Assumptions

- Training data is representative of the deployment phase.
- The classifier algorithm reaches the desired level of accuracy and robustness on training data.

Guarantees

- High-quality saliency maps are generated corresponding to the classification task.
- Generated saliency maps correctly identify important regions within the accepted range of pixel coverage rate.

Goal Structure for Explainability of the AI Developer

Like the last two examples, Figure 16 presents an example goal structure for satisfying the explainability attribute for the CLF subsystem for the AI developer. In this goal structure, the claims concern the developed xAI module in CLF, following the classification algorithm. Beside the minimum implementation criterion (G2), the goal structure argues the design according to the intended use, mentioned in the EU AI Act (G3), and technical effectiveness. The AI development documentation (AIDoc), provided by the AI developer, gives evidences for the sub-claims including:

- xAI module documentation (Sn1, Sn2, Sn5, Sn6);
- design documentation, including user story analysis and method suitability (Sn3, Sn4).

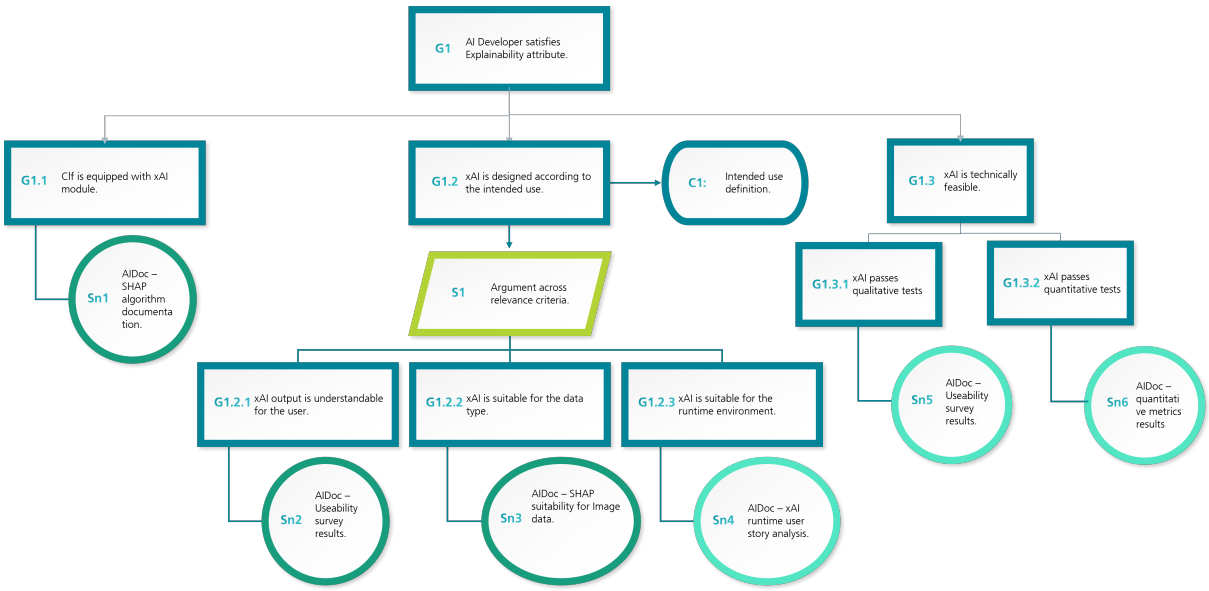

Figure 16: AI developer - Goal structure for verification of explainability quality attribute.

6 Outlook

The EU AI Act represents a significant step towards regulating artificial intelligence, while promoting innovation. While the Act marks progress in the field of developing safe and human-centric AI, it also highlights gaps and challenges that will require ongoing collaboration to address. Bridging these gaps will require concentrated efforts to harmonize regulatory standards, enhance verification processes, and promote dialogue on AI ethics and governance. This whitepaper takes a step towards translating the AI Act into actionable requirements across the value chain. Drawing on safety engineering best practices, our methodology integrates the verification of safety-related properties with the EU AI Act using Goal Structuring Notation (GSN). We present the practicality of our methodology on three industry specific use-cases, namely automotive, industrial automation, and healthcare.

References

[1] The Assurance Case Working Group (ACWG), "Goal Structuring Notation Community Standard Version 3," https://scsc.uk/r141C:1, 2021.

[2] Official Journal (OJ) of the European Union, "Artificial Intelligence Act (Regulation (EU) 2024/1689), Official Journal version of 13 June 2024," https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689, 2024, online; accessed 08 August 2024.

[3] R. Wortham, "MAGF Framework White Paper," https://static1.squarespace.com/static/63dce129c4a0240681746067/t/65abb1652c21fc1dc8859830/1705750887433/MAGF+Framework+White+Paper.pdf, online; accessed 21 April 2024.

[4] J. Kelly, S. Zafar, L. Heidemann, J. Zacchi, D. Espinoza, and N. Mata, "Navigating the EU AI Act: A methodological approach to compliance for safety-critical products," *arXiv e-prints*, pp. arXiv–2403, 2024.

[5] "ISO/IEC 24028: Information technology Artificial intelligence — Overview of trustworthiness in artificial intelligence," 2020.

[6] "ISO 26262:2018 - Road vehicles – Functional safety, 2nd Edition," 2018.

[7] N. N. G. d. Andrade and A. Zarra, "Artificial Intelligence Act: A Policy Prototyping Experiment: Operationalizing the Requirements for AI Systems – Part I," Rochester, NY, Nov. 2022.

[8] International Organization for Standardization, "ISO/IEC 22989: Information technology — Artificial intelligence — Artificial intelligence concepts and terminology," 2022.

[9] National Institute for Standards and Technology, The CPS Working Group (CPS PWG), "Framework for Cyber-Physical Systems: Volume 1," https://doi.org/10.6028/NIST.SP.1500201, 2017, online; accessed 30 January 2024.

[10] National Institute for Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf, 2023, online; accessed 30 January 2024.

[11] International Organization for Standardization, "ISO 10218-1:2011 - Robots and robotic devices - Safety requirements for industrial robots - Part 1: Robots," 2011.

[12] ——, "ISO 10218-2:2011 - Robots and robotic devices - Safety requirements for industrial robots - Part 2: Robot systems and integration," 2011.

[13] ——, "ISO/IEC TR 24027 - Information technology - Artificial intelligence (AI) - Bias in AI systems and AI aided decision making," 2021.

[14] T. Liu, M. Pelowski, C. Pang, Y. Zhou, and J. Cai, "Near-infrared spectroscopy as a tool for driving research," *Ergonomics*, vol. 59, pp. 1–25, 07 2015.

Acronyms

AI	Artificial Intelligence
APS	Automated Parking System
BCI	Brain Computer Interface
CLF	Classifier
DC	Design Contract
EU	European Union
GSN	Goal Structuring Notation
ML	Machine Learning
ODD	Operational Design Domain
PD	Person Detection
QIC	Quality Inspection Cobot
TR	Technical Requirement
TSR	Traffic Sign Recognition
xAI	Explainable Artificial Intelligence