Bundesamt
für Sicherheit in der
Informationstechnik

Deutschland
Digital•Sicher•BSI•

The EU AI Act Standardization Request

The European Understanding of AI Trustworthiness

Till Lehmann (DIN), Florian Becker (BSI), October 2023

Agenda

Content of the Standardization Request

Aspects, Problems and Issues for AI Conformity Assessment

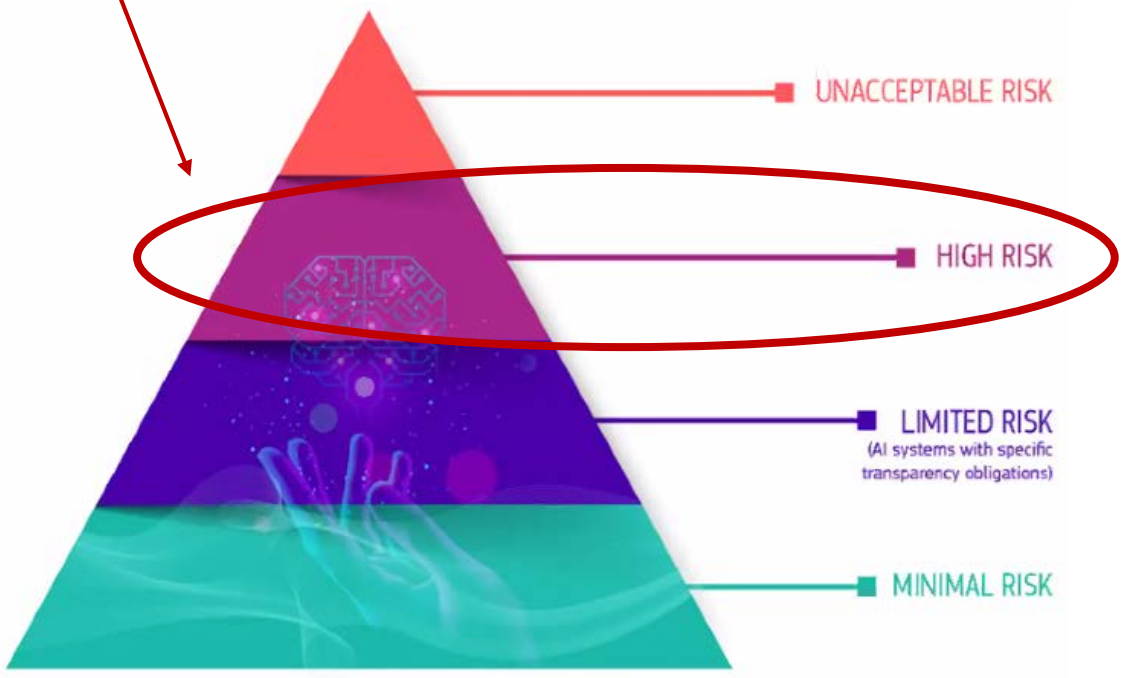
The Cybersecurity Perspective: Trustworthiness Profiles

Key Objectives of the AI Act

- **Ensuring AI systems are safe and trustworthy**
  - This objective seeks to establish a framework for AI safety, requiring developers to adhere to stringent standards to minimize potential risks
- **Protecting fundamental rights and values**
  - The Act emphasizes the importance of human-centric AI, ensuring that the technology respects privacy, human dignity, and other core values
- **Promoting innovation and competitiveness**
  - The Act encourages innovation by providing a clear regulatory framework, ensuring a level playing field for businesses while prioritizing safety

Role of standards in the Artificial Intelligence Act

Regulated under the
New Legislative Framework

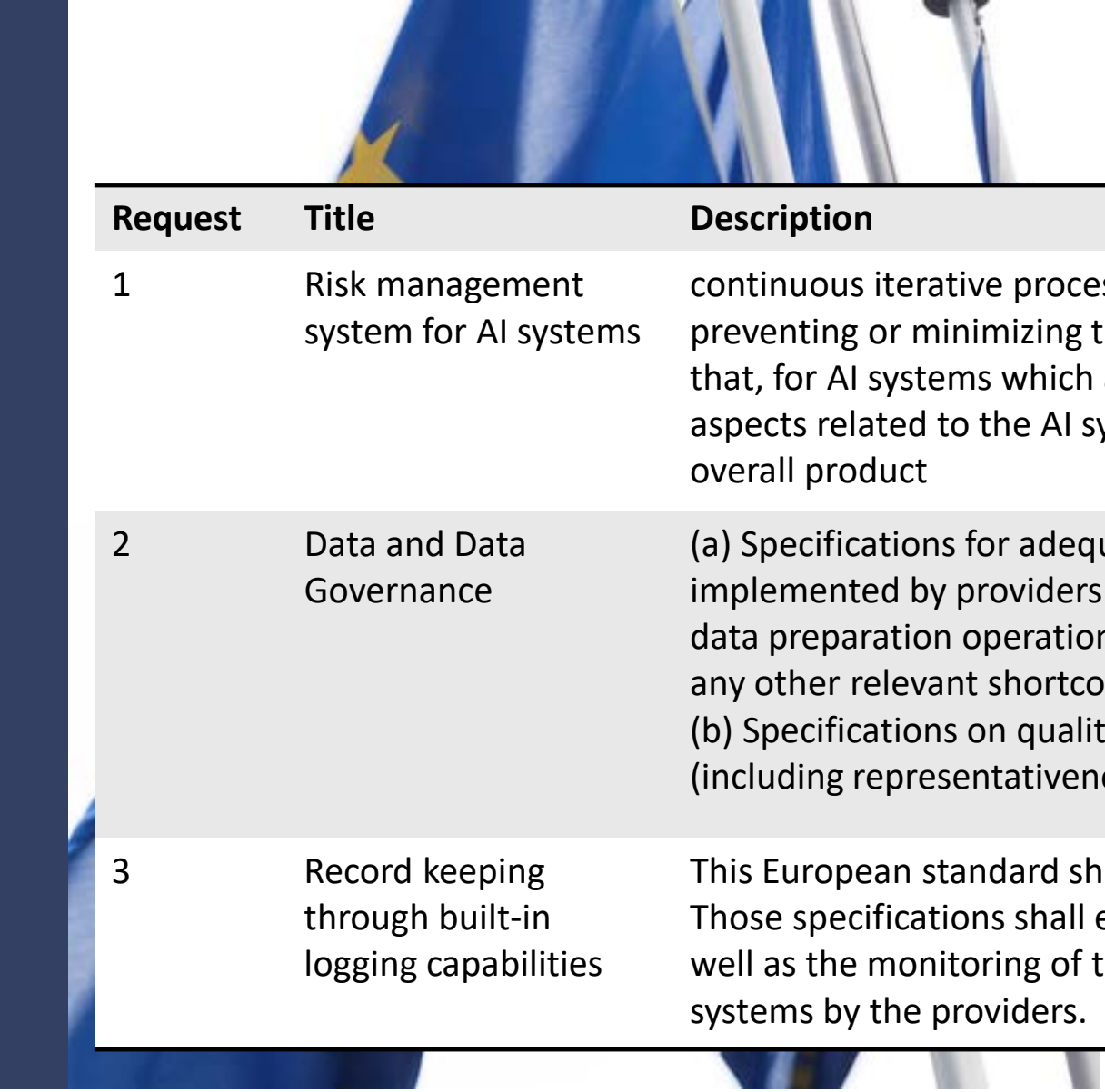


Article 40
Harmonised standards

High-risk AI systems which are in conformity with harmonised standards or parts thereof that references of which have been published in the Official Journal of the European Union **shall be presumed to be in conformity with the requirements set out in Chapter 2** of this Title, to the extent those standards cover those requirements.

AI Act Standardization Request

1	**risk management system** for AI systems	6	**accuracy** specifications for AI systems
2	**governance and quality of datasets** used to build AI systems	7	**robustness** specifications for AI systems
3	**record keeping** through logging capabilities by AI systems	8	**cybersecurity** specifications for AI systems
4	**transparency and information** provisions to the users of AI system	9	**quality management system** for providers of AI systems, including post-market monitoring process
5	**human oversight** of AI systems	10	**conformity assessment** for AI systems



Content of the SR [1]

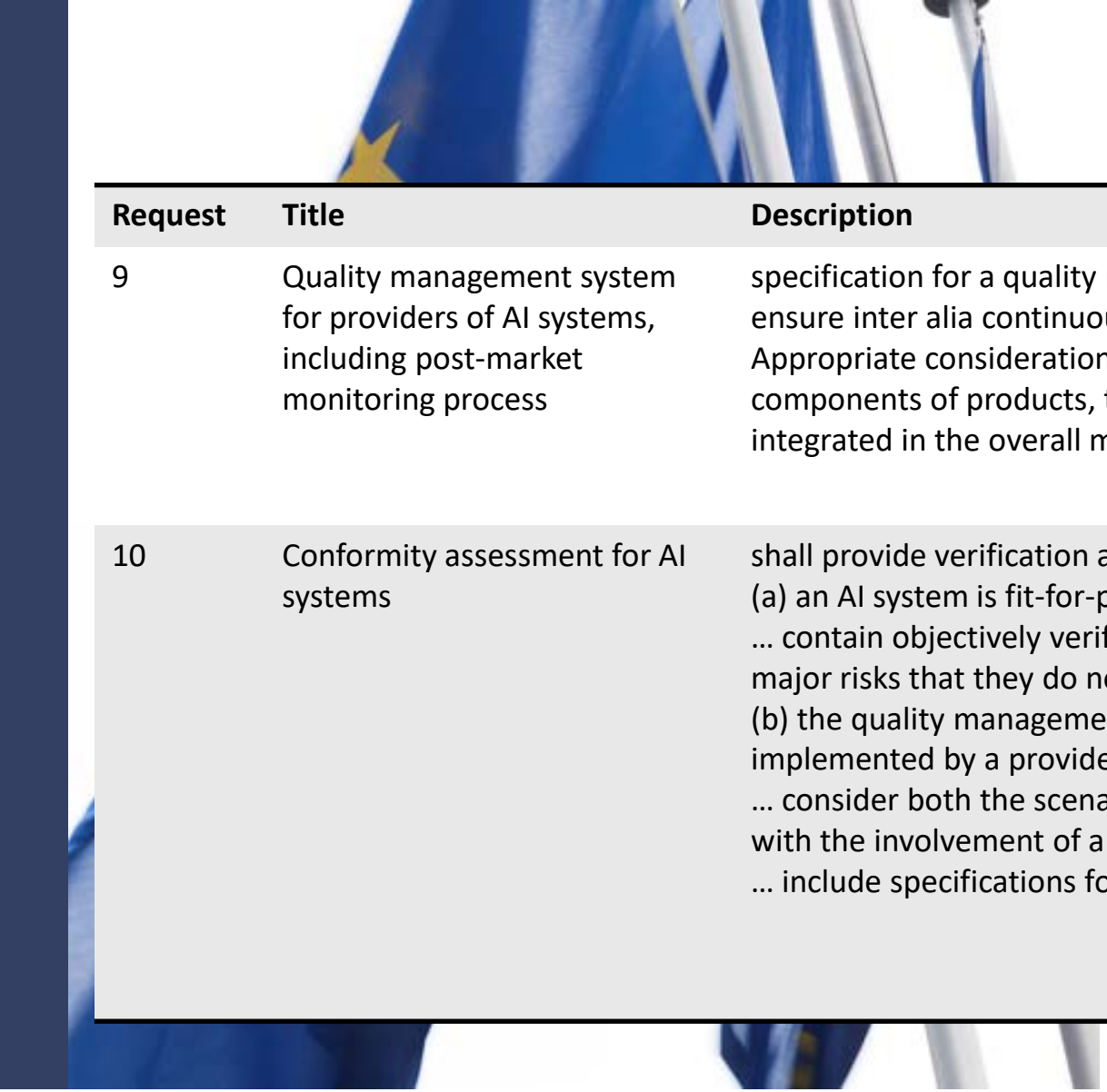
Request	Title	Description
1	Risk management system for AI systems	continuous iterative process run throughout the entire lifecycle of the AI system which is aimed at preventing or minimizing the relevant risks to health, safety or fundamental rights. … in such a way that, for AI systems which are safety components of products, the risk management system aspects related to the AI system may be integrated into the risk management system for the overall product
2	Data and Data Governance	(a) Specifications for adequate data governance and data management procedures to be implemented by providers of AI systems (with specific focus on data generation and collection, data preparation operations, design choices, procedures for detecting and addressing biases or any other relevant shortcomings in data), (b) Specifications on quality aspects of datasets used to train, validate and test AI systems (including representativeness, relevance, completeness, correctness)
3	Record keeping through built-in logging capabilities	This European standard shall include specifications for automatic logging of events for AI systems. Those specifications shall enable the traceability of those systems throughout their lifecycle as well as the monitoring of their operations and shall facilitate the post-market monitoring of the AI systems by the providers.

Content of the SR [2]

Request	Title	Description
5	Human oversight	specify measures and procedures for human oversight of AI systems which are: (a) identified and built, when technically feasible, into the high-risk AI system by the provider; (b) identified by the provider and to be implemented by the user. … measures enabling users to understand, monitor, interpret, assess and intervene in relevant aspects of the operation of the high-risk AI system. … also define, where justified, appropriate oversight measures which are specific to certain AI systems in consideration of their intended purpose. [e.g., AI systems intended for remote biometric identification of persons]
6	Accuracy specifications	specifications for ensuring an appropriate level of accuracy of AI systems and to declare the relevant accuracy metrics and levels. … also define, where justified, a set of appropriate and relevant tools and metrics to measure accuracy against suitably defined levels, which are specific to certain AI systems in consideration of their intended purpose.

Content of the SR [3]

Request	Title	Description
6	Accuracy specifications	specifications for ensuring an appropriate level of accuracy of AI systems and to declare the relevant accuracy metrics and levels. … also define, where justified, a set of appropriate and relevant tools and metrics to measure accuracy against suitably defined levels, which are specific to certain AI systems in consideration of their intended purpose.
7	Robustness specifications	specifications for the robustness of AI systems, … relevant sources of errors, faults and inconsistencies, as well as the interactions of the AI system with the environment, including of those AI systems which continue to learn after being placed on the market or put into service, notably in respect to feedback loops.
8	Cybersecurity specifications	organizational and technical solutions, to ensure that AI systems are resilient against attempts to alter their use, behavior, performance or compromise their security properties … include, where appropriate, measures to prevent and control for cyberattacks trying to manipulate AI specific assets, such as training data sets (e.g. data poisoning) or trained models (e.g. adversarial attacks), or exploiting vulnerabilities in an AI system's digital assets or the underlying ICT infrastructure.



Content of the SR [4]

Request	Title	Description
9	Quality management system for providers of AI systems, including post-market monitoring process	specification for a quality management system to be implemented by providers within their organizations to ensure inter alia continuous compliance of an AI system with the aspects described under points 2 to 8. Appropriate consideration to medium and small size organizations. … for AI systems which are safety components of products, the quality management system aspects related to the AI system may be integrated in the overall management system of the product manufacturer.
10	Conformity assessment for AI systems	shall provide verification and validation procedures and methodologies to assess whether: (a) an AI system is fit-for-purpose, notably with regard to aspects described under points 1 to 8. … contain objectively verifiable criteria and shall indicate not only the risks that they cover, but also the major risks that they do not cover; (b) the quality management system measures and processes, as described under point 9, are appropriately implemented by a provider … consider both the scenarios whereby the conformity assessment is carried out by the provider itself or with the involvement of a professional external third-party organisation. … include specifications for the testing of AI systems in the context of the conformity assessment.

Next Steps

- Currently trialogue (Council, Parliament, Commission)
- Possibly more trialogues needed
- Result probably Q1 next year

Therefore, exactly right time to talk about criteria development

Agenda

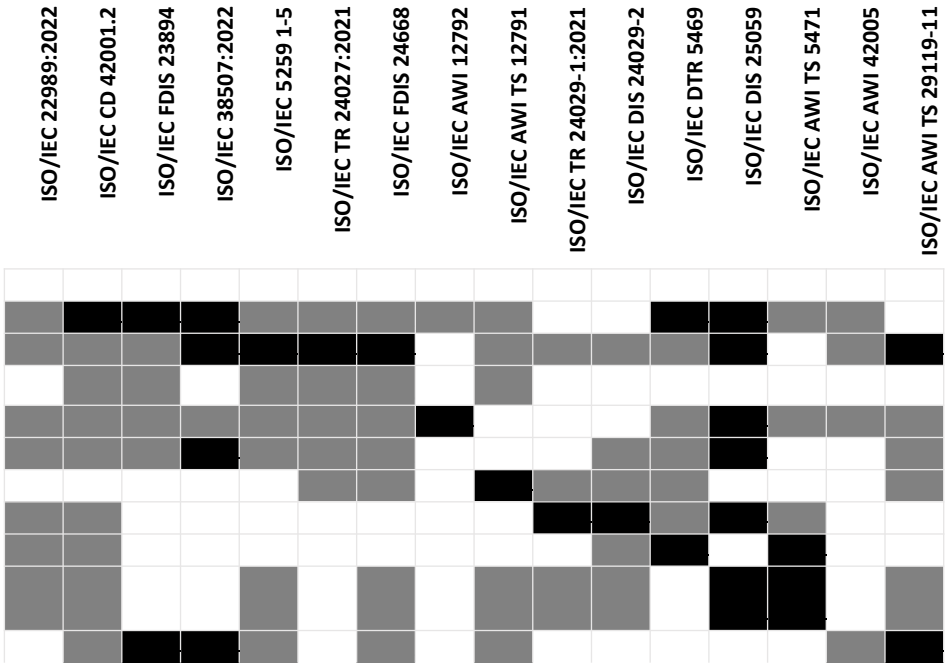
Content of the Standardization Request

Aspects, Problems and Issues for AI Conformity Assessment

The Cybersecurity Perspective: Trustworthiness Profiles

International standards cover most topics

ISO Standards->



EU Standardisation Request

Risk management system for AI systems

Governance and quality of datasets used to build AI systems

Record keeping through built-in logging capabilities in AI systems

Transparency and information to the users of AI systems

Human Oversight of AI systems

Accuracy specifications for AI systems

Robustness specifications for AI systems

Cybersecurity specifications for AI systems

Quality Management system for providers of AI system, including post-market monitoring process.

Conformity assessment for AI systems

Legend

No Relevance

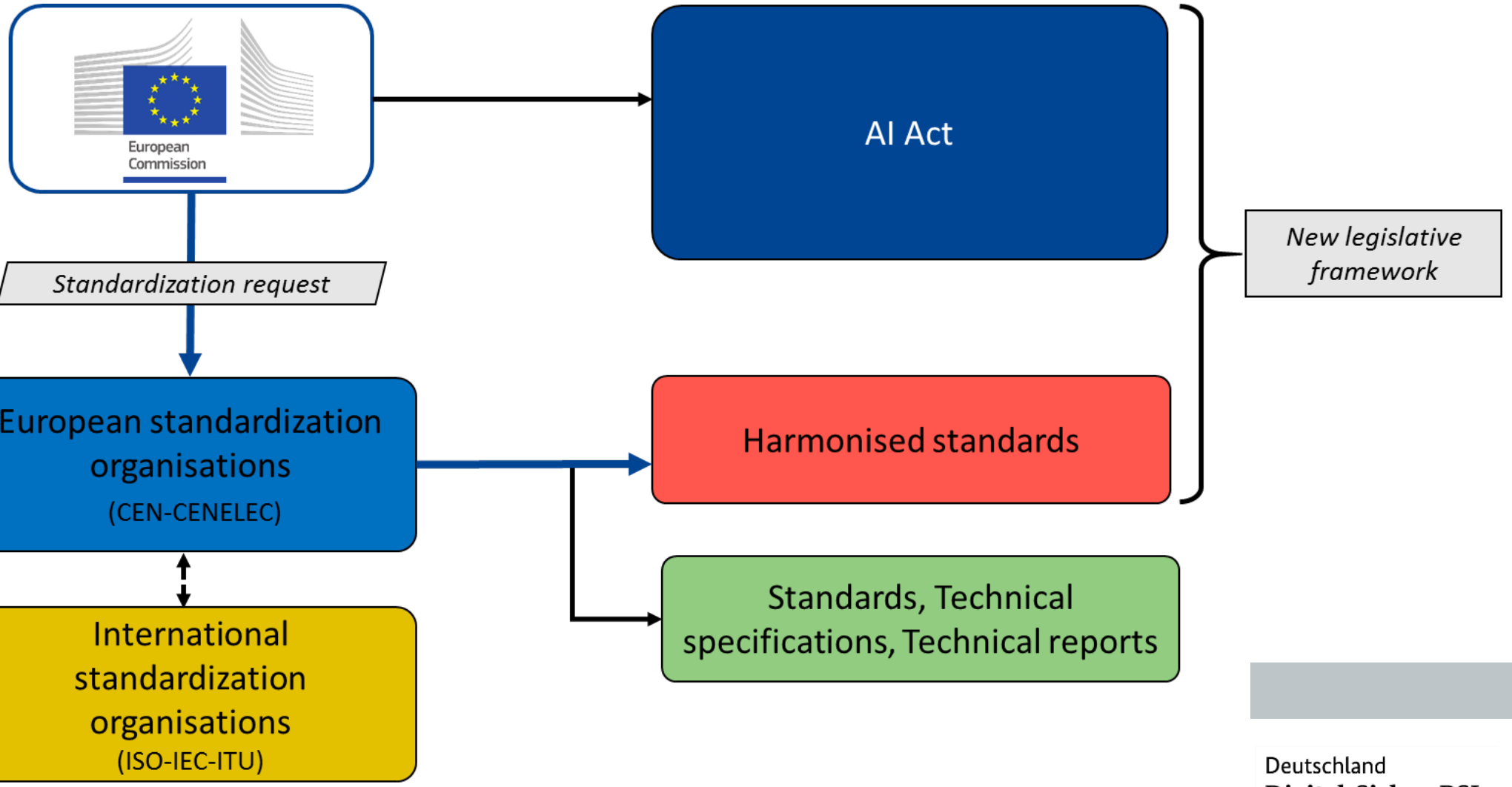
Contains Elements Relevant for the Standardisation Request

Detailed Description relevant for the Standardisation Request

Bundesamt für Sicherheit in der Informationstechnik

Deutschland Digital•Sicher•BSI•

EU-Framework AI Regulation



AI-Act: Challenges

I. Multi-dimensional complexity
- *Regulation environment: AI Act, Cyber-security Act, Cyber-resilience Act, Data Act, Data Governance Act…*
- *Technical complexity: AI, cyber, hardware, infrastructure…*
- *Horizontal/Verticals standardization interplay*

II. Conformity assessment/notified bodies
- *Governance (who does what? One or multiple schemes?), Competences (auditors, notified bodies, testing facilities...), Management system vs. Products…*

III. Requirements/specifications overlaps
- Requirements/specifications covering simultaneously cybersecurity, quality, safety, trustworthiness…
- Multiple management systems (ISO 42001, 27001, 27701, 9001)

Agenda

Content of the Standardization Request

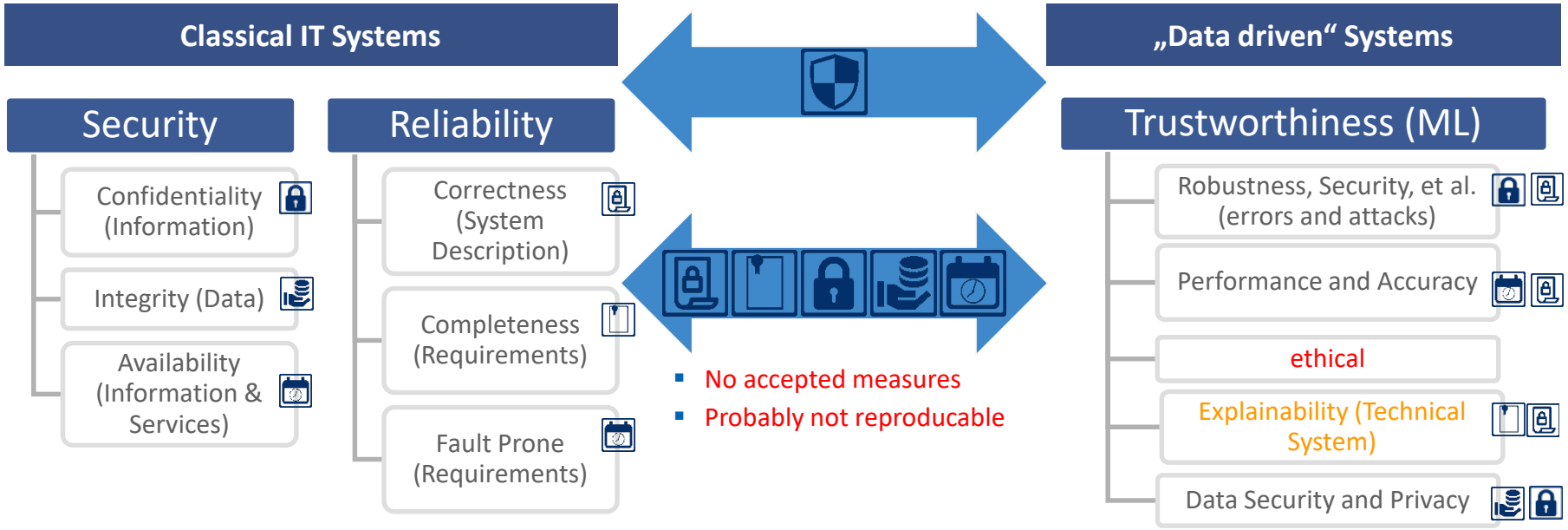
Aspects, Problems and Issues for AI Conformity Assessment

The Cybersecurity Perspective: Trustworthiness Profiles

AI Trustworthiness Concept & Cybersecurity

EU AI Act forms the concept of „Trustworthiness“ as

trustworthy = legal + cybersecure + ethical



Relationships of Cybersecurity Request

Some Requests simultaneously cover:

- quality,
- safety,
- cybersecurity,
- privacy…

Standardization request

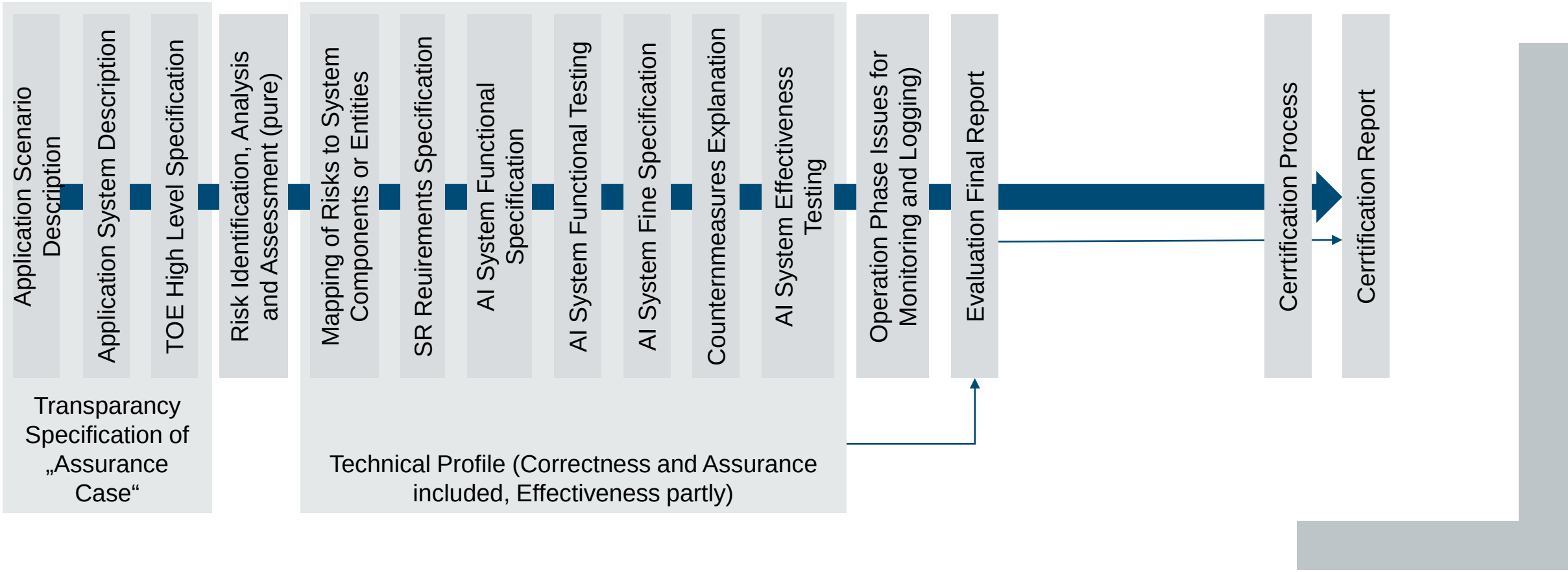
1. **risk management system** for AI systems
2. **governance and quality of datasets** used to build AI systems
3. **record keeping -** built-in logging capabilities in AI systems
4. **transparency and information** to the users of AI systems
5. **human oversight** of AI systems
6. **accuracy** specifications for AI systems
7. **robustness** specifications for AI systems
8. **cybersecurity** specifications for AI systems
9. **quality management system** for providers of AI system
10. **conformity assessment** for AI systems

- Tradeoffs,
- Dependencies,
- Relationships,
- …

Trustworthiness Profile Concept „Mission“

- A TP defines an **implementation-independent set of requirements** for a category of products which are intended to meet common needs for AI trustworthiness. A TP claimed by a user, consumer or stakeholder for AI gives them the possibility to express their trustworthiness needs without referring to a specific product. **Product certifications** can be based on Trustworthiness Profiles.
- Certification of the Trustworthiness Profile is carried out on the instigation of a certification body or a sponsor. A part of the procedure is the **technical examination (evaluation) of the Trustworthiness Profile according to Standard Criteria (TAISEC)**. The evaluation is usually carried out by an evaluation facility recognized by the certification body.

AI Trustworthiness Profile Evaluation Process (roughly)



Thank you for your Attention. Questions?

Deutschland
Digital•Sicher•BSI•

KI-KOMPETENZZENTRUM DES BSI

Bundesamt
für Sicherheit in der
Informationstechnik