WEBINAR

Navigating AI Safety with ISO 8800: Requirements Management Best Practices

Matt Mickle

Director of Automotive and Semiconductor Solutions

Jama Software

Jody Nelson

Co-Founder and Managing Partner

SecuRESafe (SRES)

Agenda

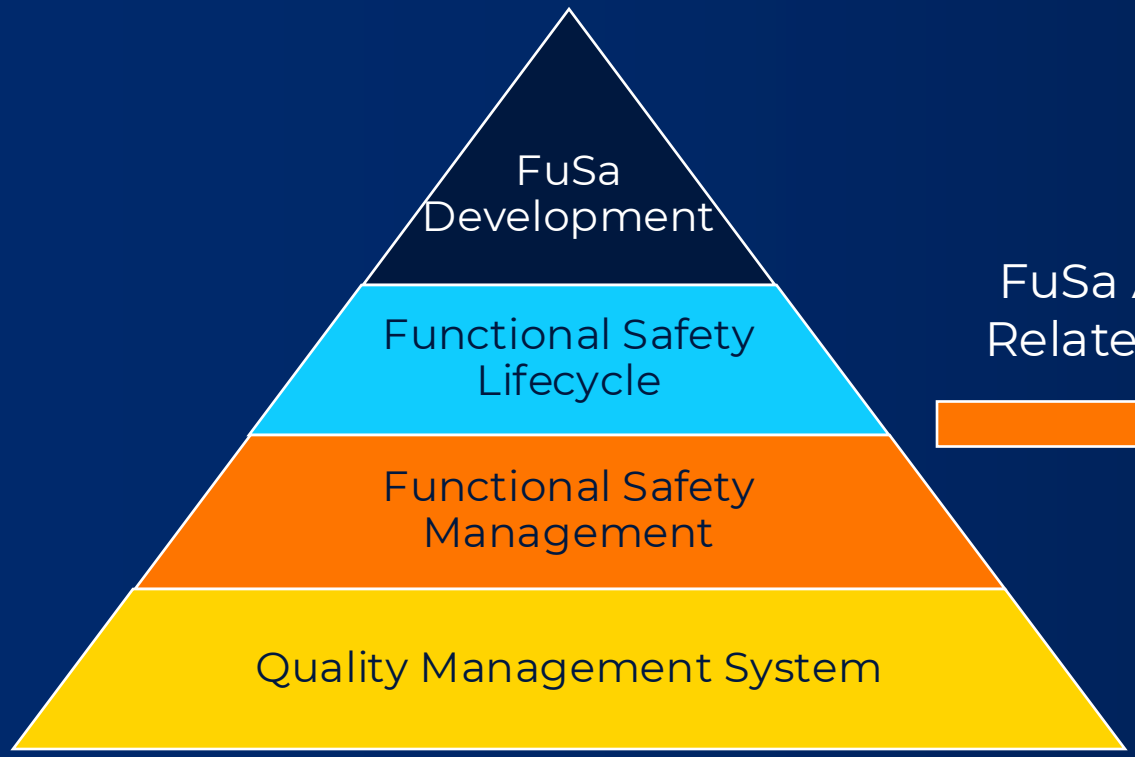
NAVIGATING AI SAFETY WITH ISO 8800

- The importance and framework of ISO/PAS 8800 for AI safety in road vehicles
- How to derive and manage AI safety requirements effectively
- Addressing insufficiencies in AI systems and ensuring traceability to related standards
- Practical strategies for integrating ISO 8800 into a structured requirements and systems engineering workflow

ADAS and ADS Development

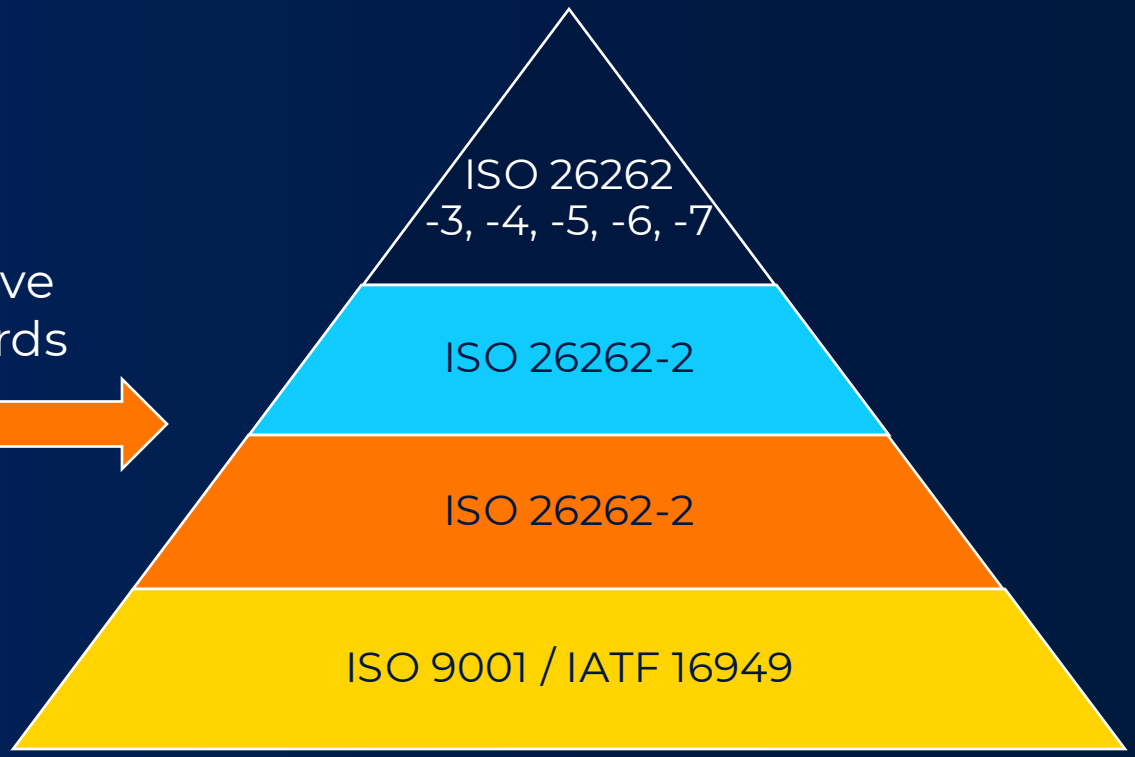
AI Requirements Landscape

Automotive Functional Safety Framework



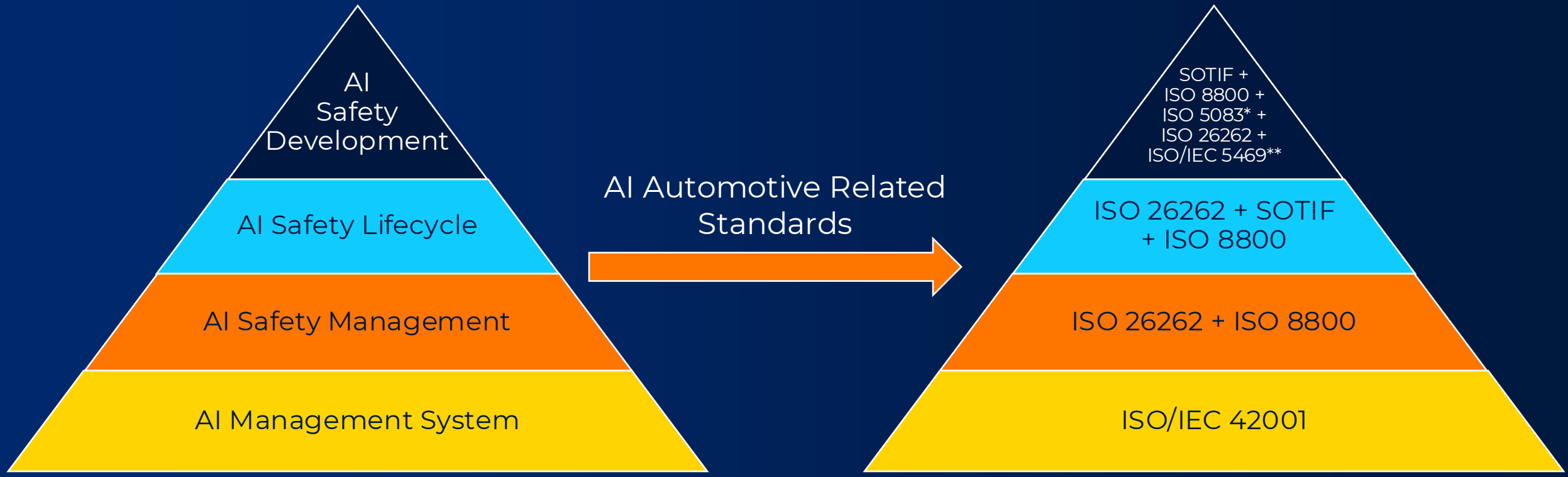
Automotive Functional Safety Development

FuSa Automotive Related Standards



- Well-established framework for functional safety in automotive
- Nearly the entire framework is built within the ISO 26262 standard

Automotive AI Safety Framework

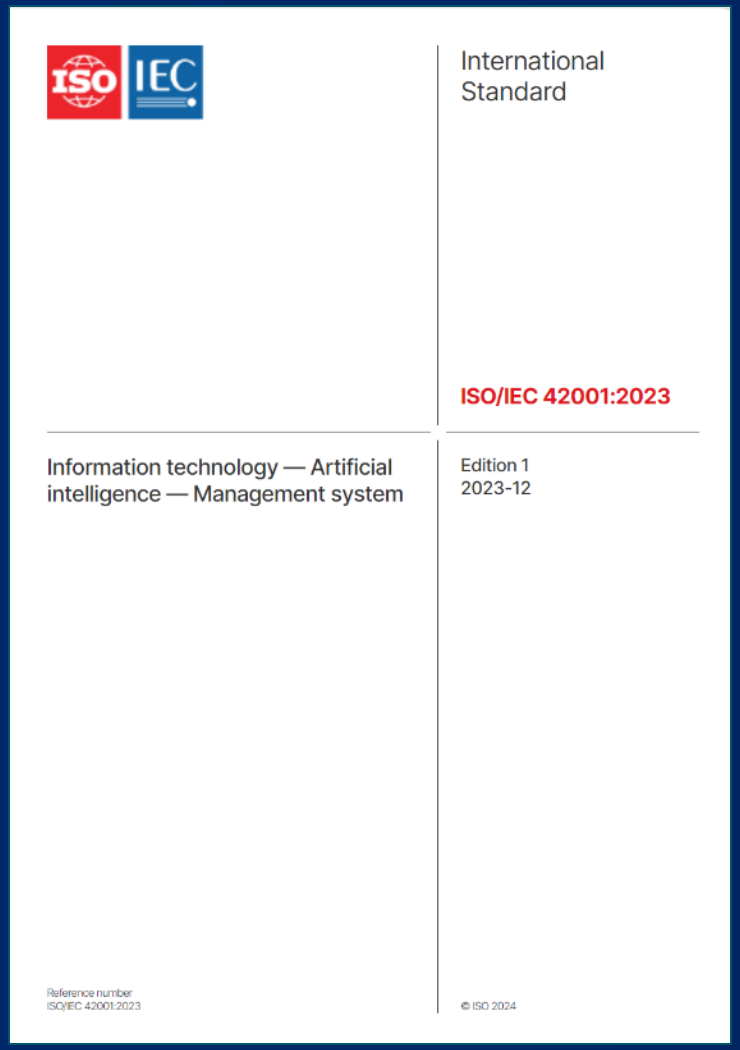


Automotive AI Safety Development

"New" Automotive Model

*Replacement of ISO/TR 4804:2020
**ISO/IEC 5469 is a Technical Report and is expected be replaced by ISO/IEC 22440 standard

AI Management System



Released December 18, 2023

Establishes the AI Management Systems (AIMS) within organizations

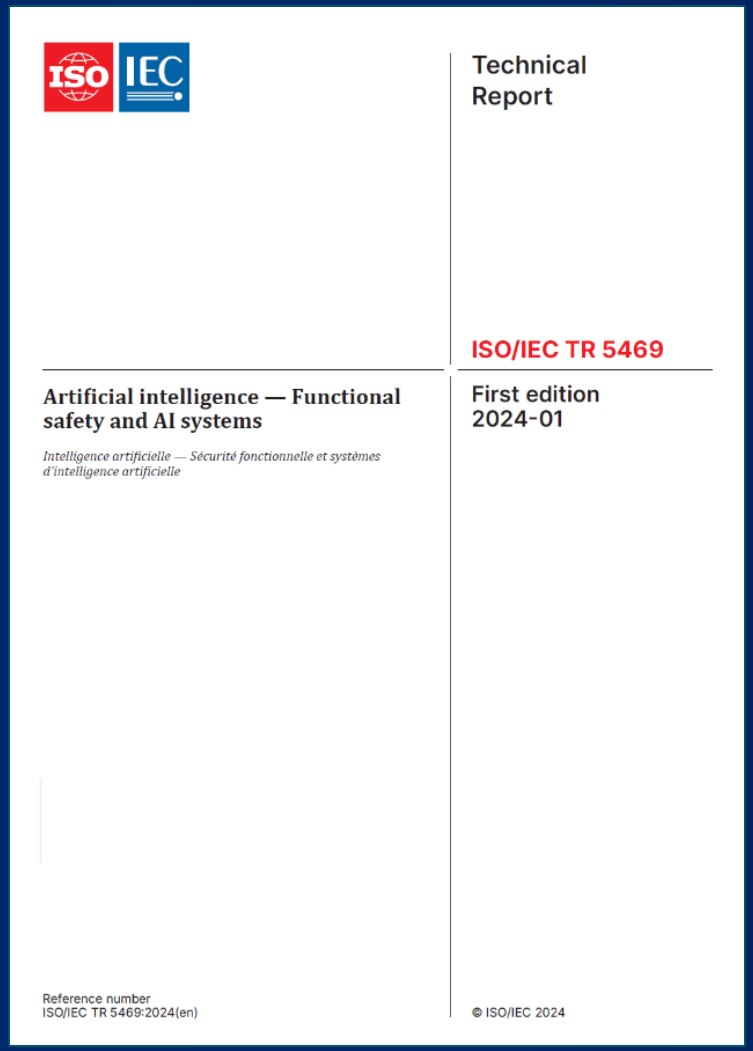
Designed to ensure responsible development and use of AI systems

Addresses ethical considerations, transparency, safety and security

Provides a risk-based approach to identify risks and to create an appropriate plan

Works in harmony with QMS

Functional Safety and AI Systems


ISO IEC

Technical Report

ISO/IEC TR 5469

Artificial intelligence — Functional safety and AI systems

Intelligence artificielle — Sécurité fonctionnelle et systèmes d'intelligence artificielle

First edition
2024-01

Reference number
ISO/IEC TR 5469:2024(en)

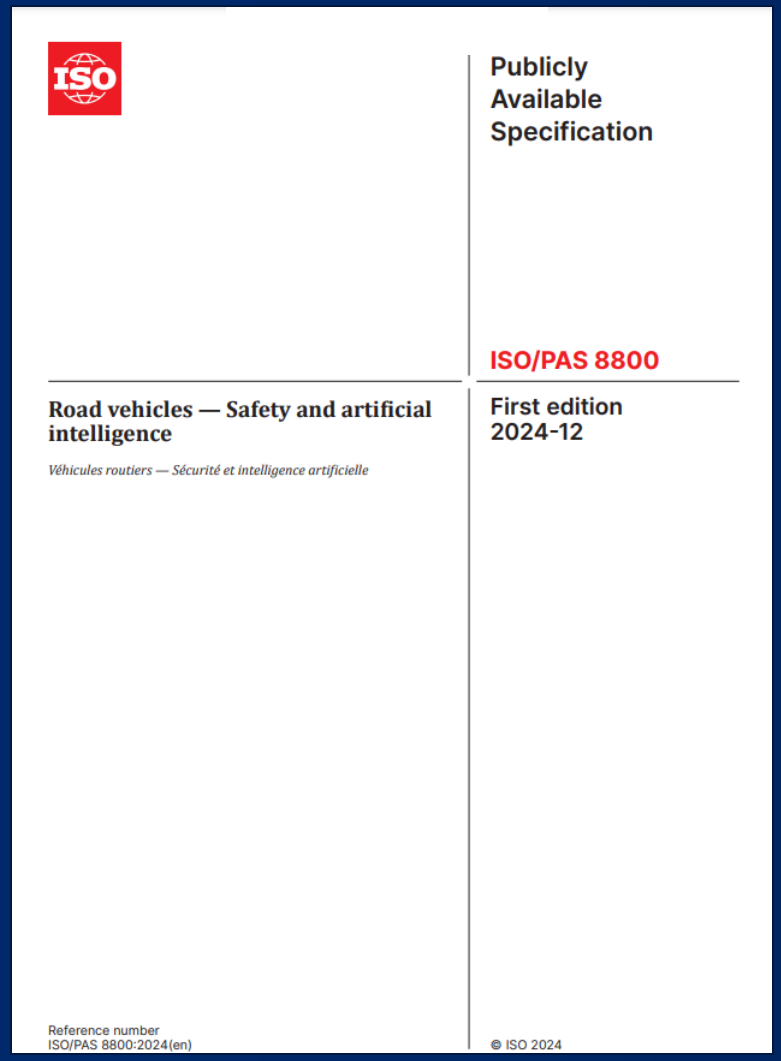
© ISO/IEC 2024

- Released January 2024
  Provides guidance on the functional safety of AI systems agnostic to industry
- Describes properties, related risk factors, methods, and processes for ensuring functional safety in AI systems
- Provides a method for deriving acceptance criteria
- Expected be replaced by ISO/IEC TS 22440 standard

ISO/PAS 8800:2024

Road vehicles — Safety and artificial intelligence

Safety and AI



ISO

Publicly Available Specification

ISO/PAS 8800

Road vehicles — Safety and artificial intelligence

Véhicules routiers — Sécurité et intelligence artificielle

First edition
2024-12

Reference number
ISO/PAS 8800:2024(en)

Released December 13, 2024

Provides industry-specific (automotive) guidance on the use of AI systems in **safety-related functions**

Defines a framework for managing AI safety that **tailors** or **extends** existing approaches currently defined in the **ISO 26262** series and in **ISO 21448**

Applicable to any AI method for AI elements onboard the vehicle, for use of AI for the **functionality itself** or as a **safety mechanism**

Addresses the risk of undesired safety-related behavior due to output **insufficiencies**, **systematic errors** & **random hardware errors** of AI elements

Overall Structure – Normative Clauses

Clause 7: AI safety management

Clause 8: Assurance argument of AI systems

Clause 9: Derivation of AI safety requirements

Clause 10: Selection of AI technologies, architectural and development measures

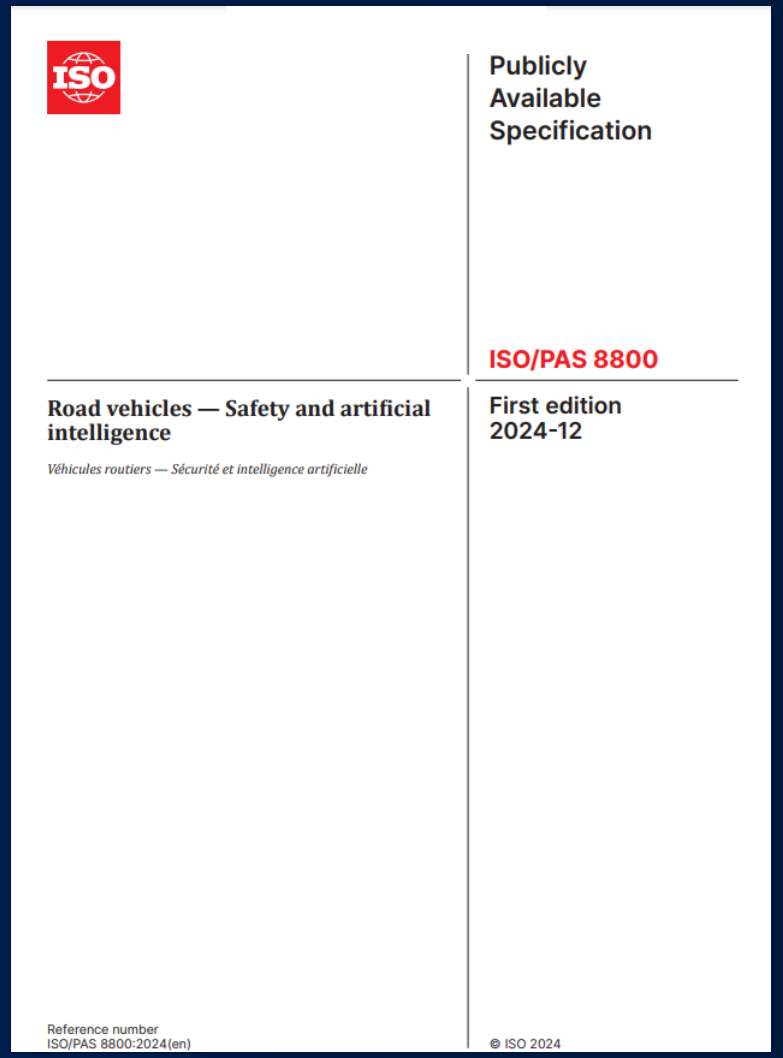
Clause 11: Data-related consideration

Clause 12: Verification and validation of the AI system

Clause 13: Safety analysis of AI systems

Clause 14: Measures during operation

Clause 15: Confidence in use of AI development frameworks and software tools used for AI model development


ISO

Publicly Available Specification

ISO/PAS 8800

Road vehicles — Safety and artificial intelligence

Véhicules routiers — Sécurité et intelligence artificielle

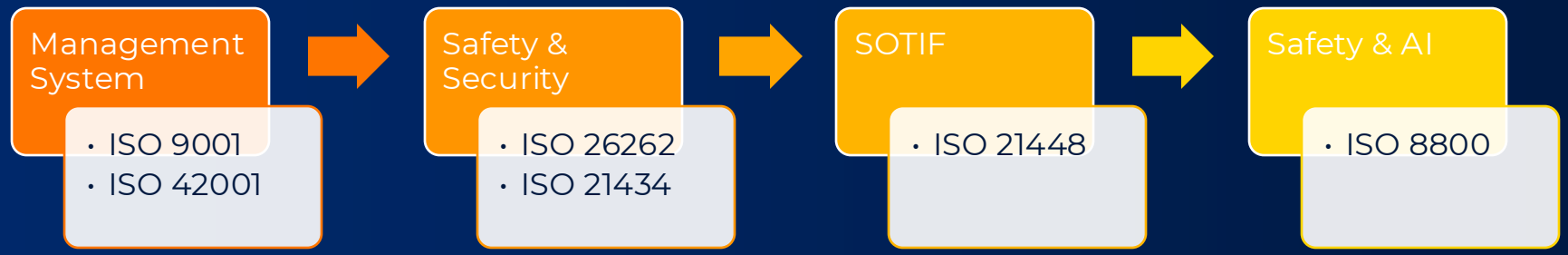
First edition
2024-12

Reference number
ISO/PAS 8800:2024(en)

© ISO 2024

AI and Road Vehicle System Safety

- ISO/PAS 8800 is not standalone but is intended to be applied in combination with ISO 26262 and ISO 21448
  - ✓ tailoring of ISO 26262 and ISO 21448 for AI components that are, or contain, AI models
  - ✓ new requirements specific to AI components that are, or contain, AI models



AI Safety

AI safety

absence of unreasonable risk due to AI errors caused by faults and functional insufficiencies

Concept within ISO 21448

Addressed generally by ISO 26262

Note: Although the term "AI safety" is commonly understood to have a broader meaning which includes ethics, value alignment, long-term considerations, etc., ISO/PAS 8800 considers safety the same as ISO 26262, absence of physical injury or damage to the health of persons

Functional Insufficiencies

functional insufficiency (definition from ISO 21448)

insufficiency of specification or performance insufficiency {at the vehicle level or the E/E elements of the system}

Examples:

- Sensor fails to detect objects in heavy rain or fog
- The system encounters a situation it was not designed or tested for, such as unusual road layouts or unpredictable human behavior

In the context of ISO 8800, functional insufficiencies specifically refer to a systematic safety-related error of an AI component consisting of an AI model caused by :

1. Data related issue (Clause 11 – e.g., data collection strategy)
2. Design related issue (Clause 10 – e.g., optimization of model parameters)
3. Insufficiency in requirement specification (Clause 9)

AI components that are not AI models or that do not contain AI models are not developed according to ISO 8800

ISO 8800 interaction with ISO 26262 & ISO 21448

ISO 26262 Clause(s)	AI System Considerations
Part 3, clause 5: item definition	• Source of input space specification (Clause 9)
Part 3, clause 7: functional safety concept Part 4, clause 6: technical safety concept	• Potential source of safety requirements allocated to AI system or components (Clause 9)
Part 4, clause 7: system and item integration & testing ….	• Consideration for tailoring relevant to AI system integration • ….

Adapted from ISO 8800 table 6-1

ISO 21448 Clause(s)	AI System Considerations
Clause 5: Specification and design	• From 21448: Interfaces between AI system and encompassing system; definition of semantic input space; required functionality; safety requirements allocated to AI system • From 8800: triggering conditions and insufficiencies of AI system; evidence of satisfaction of safety requirements; deployment measures (Clause 9)
Clause 6: Identification and evaluation of hazards …	• Potential source of AI triggering conditions; 8800 is a further source (Clause 9.6.3) • …

Adapted from ISO 8800 table 6-2

Clause 9

Derivation of AI safety requirements

Derivation of AI Safety Requirements

The purpose of this clause is to achieve the following objectives:

- ✓To specify a **complete and consistent set of AI safety requirements** that are sufficient to ensure AI safety
- ✓To **refine AI safety requirements** based on learning from development, verification and validation (continuous improvement)
- ✓To specify the **limitations** of an AI system over its input space

General Principles:

- Traceability – to explicit safety requirements, assumptions, scenarios, and causal factors from safety analyses
- Justification – not just traceability, but evidence to how the AI safety requirement addresses the traced information

Derivation of AI Safety Requirements

The purpose of this clause is to achieve the following objectives:

- ✓To specify a **complete and consistent set of AI safety requirements** that are sufficient to ensure AI safety
- ✓To **refine AI safety requirements** based on learning from development, verification and validation
- ✓To specify the **limitations** of an AI system over its input space

For ML applications, the probability of an error in an AI system may not be computable due to the complexity of its input space, therefore refinement is required:

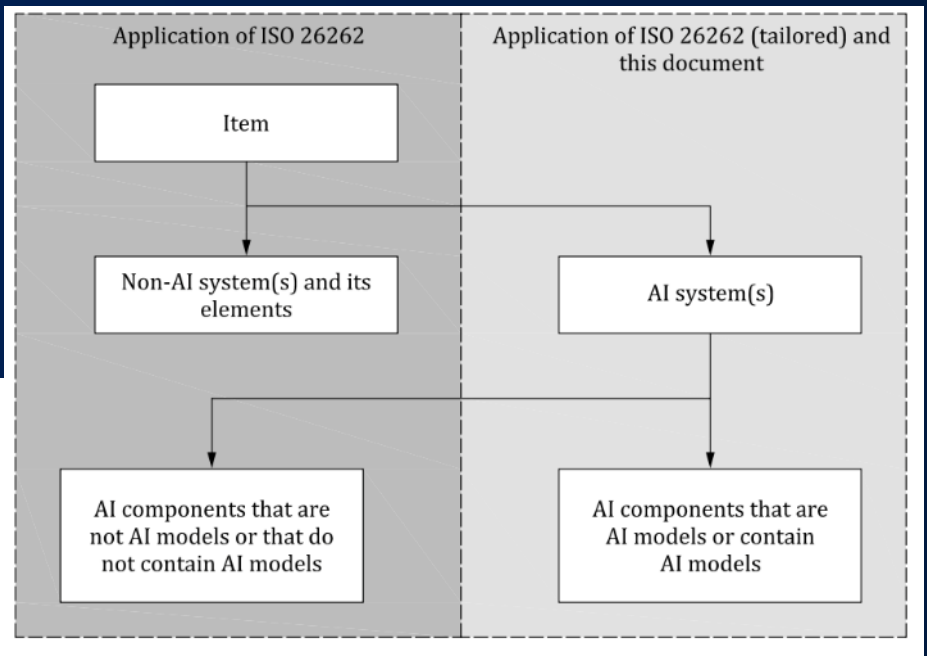
Requirements on probability of error occurrence being low → Refined → Quantitative requirements + Qualitative requirements

Probability of two consecutive False Negatives (FNs) per hour of a nearby pedestrian < 10^{-6} →

The Deep NN does not produce two consecutive FNs of a nearby pedestrian in 10^8 hours of driving + Sufficient diversity is demonstrated (different types of pedestrians)

AI Safety Requirements

1. Identify Systems/Elements subject to ISO 8800
   - Any item, system, or element containing an AI model is subject to ISO 8800
   - ISO 8800 assumes that the AI elements also is subject to ISO 26262 and ISO 21448



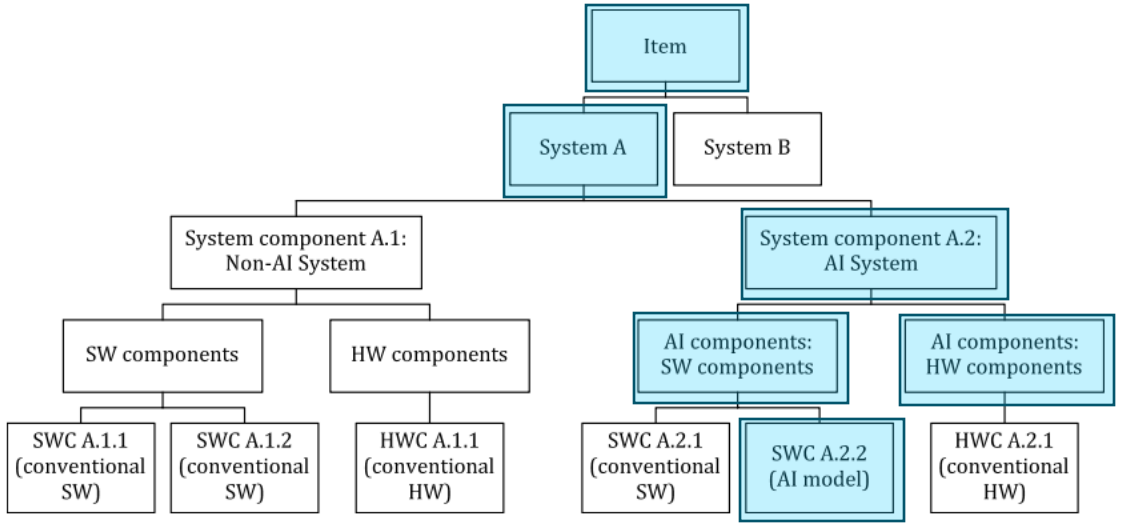
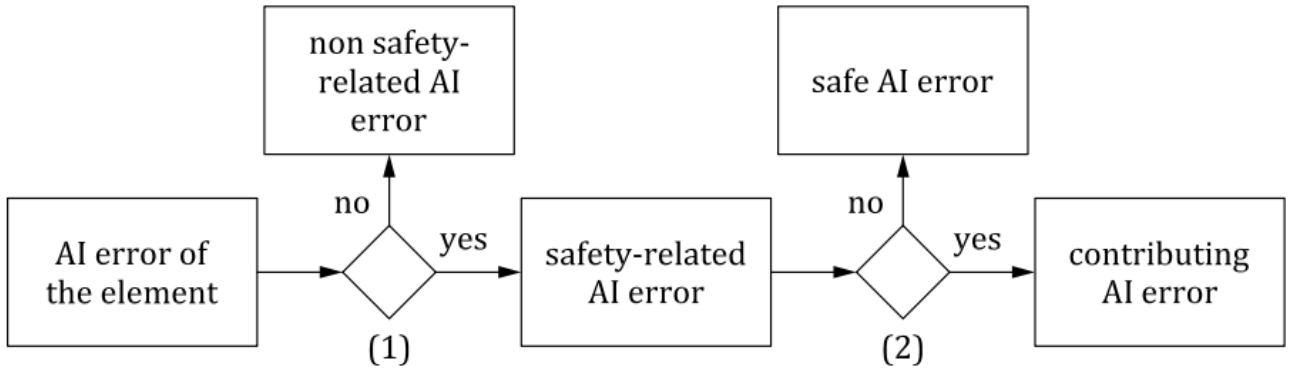


Figure 6–5 — Example of a hierarchical decomposition of an item into its elements down to the component level - decomposition tree view

AI Safety Requirements

2. Evaluate if the elements can have a safety impact



Key

(1) Is the element a safety-related element?

(2) Can the error significantly contribute to the occurrence of a safety-related undesired behaviour at the vehicle level?

Figure 6–13 — Error classification scheme based on the potential to lead to an undesired safety-related behaviour at the vehicle level

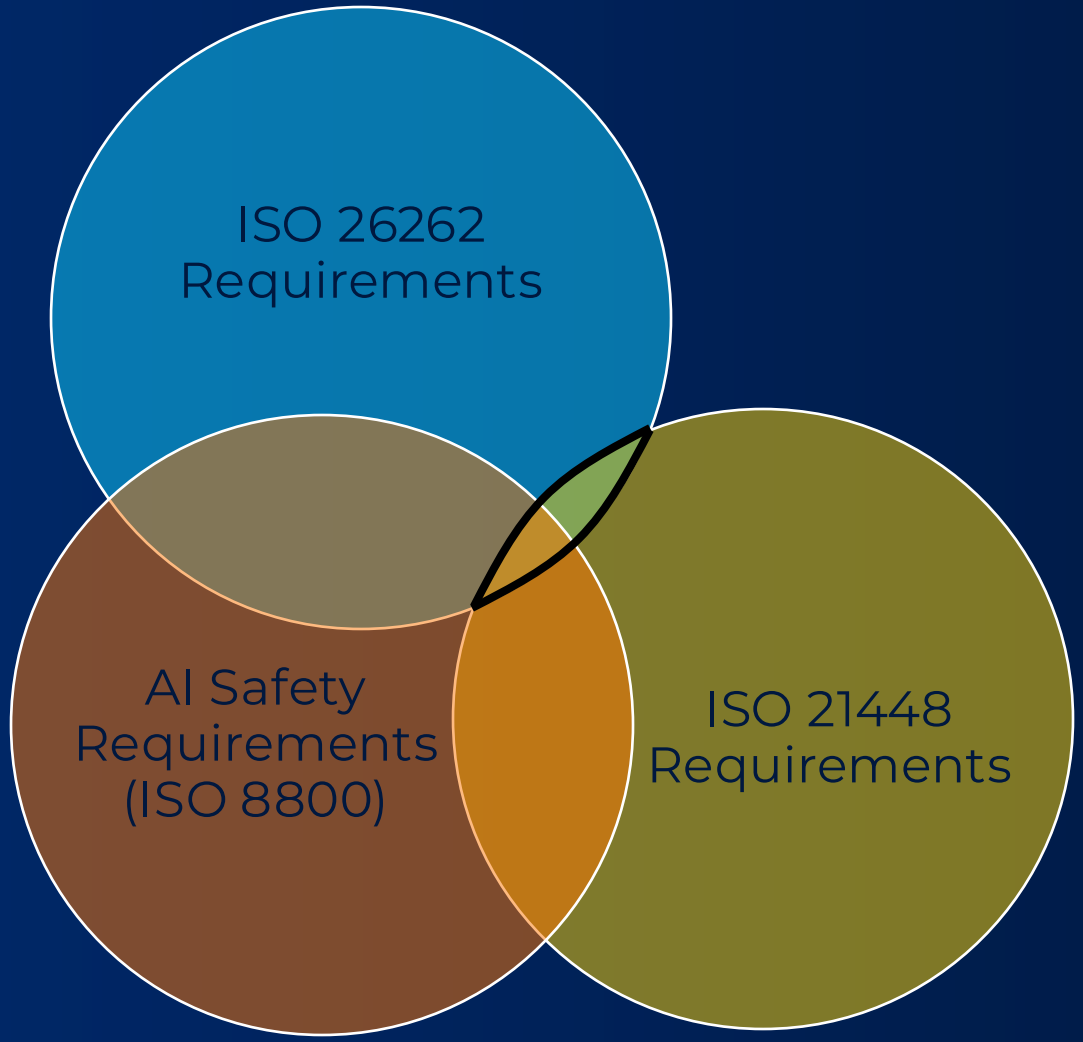
AI Safety Requirements

3. Evaluate if the elements can have a safety impact: Consideration for safety-related properties of AI systems (Annex D).
   - AI safety requirements may be necessary to address insufficiencies for each of these properties
   - Requirements may be on the model/system, organization, or process

Table D–1 — Safety-related properties of AI systems

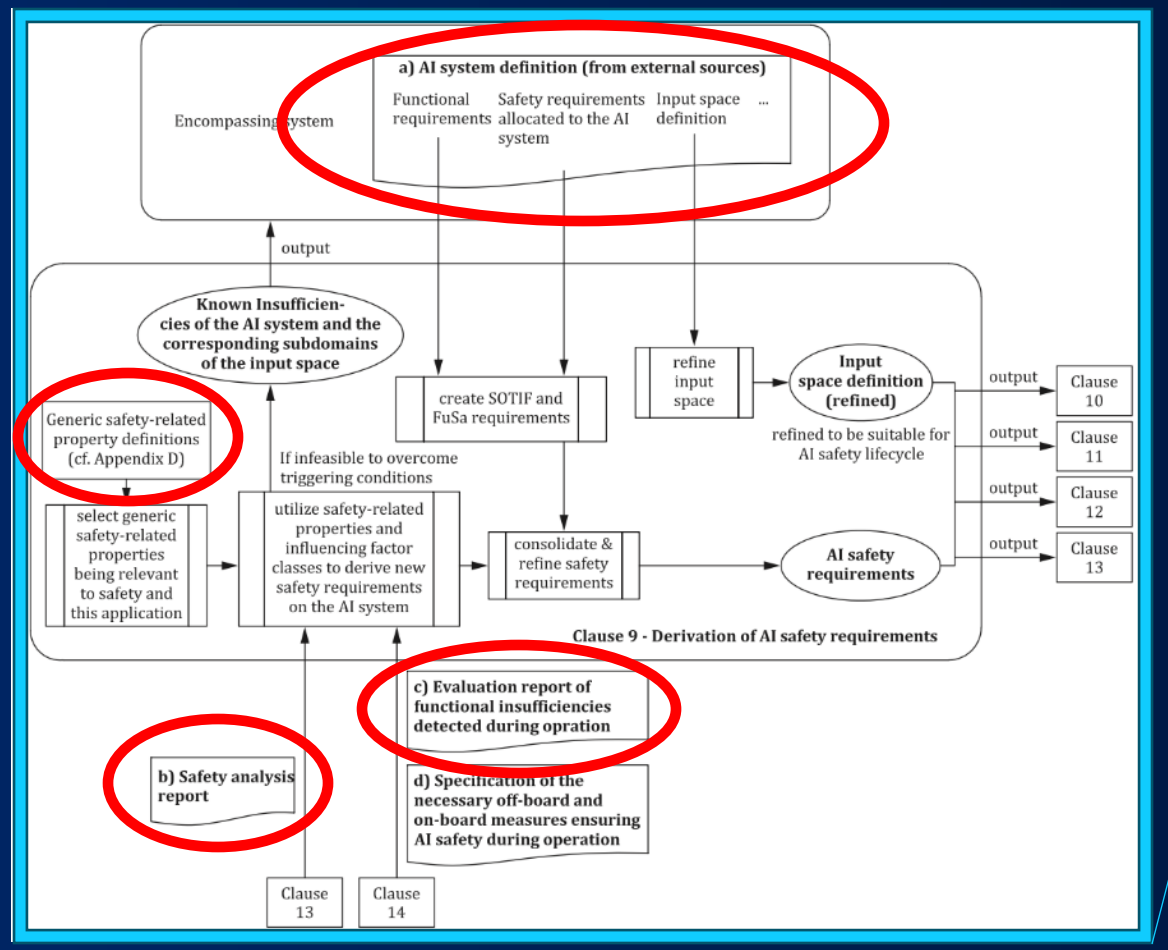
Property	Scope
AI robustness	Model, system
AI generalization capability	Model, system
AI reliability	Model, system
AI resilience	(Overall) system, organization
AI controllability	(Overall) system, organization
AI explainability	Process
AI predictability	Model, system
AI alignment	Process
Justified Design Decisions	Process
Maintainability	Organization, process
AI bias and fairness	Model, (overall) system
Distributional shift over time	Overall system

AI Safety Requirements Interdependency

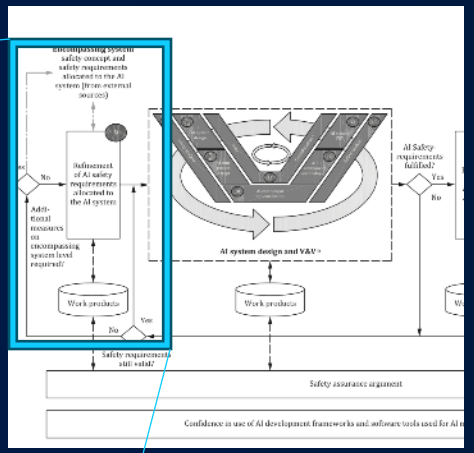


AI Safety Requirements Workflow

1. Start from AI system definition (for elements to which have an AI safety impact) and generic safety-related properties, or if in a feedback activity, from the safety analysis reports and evaluation during operation



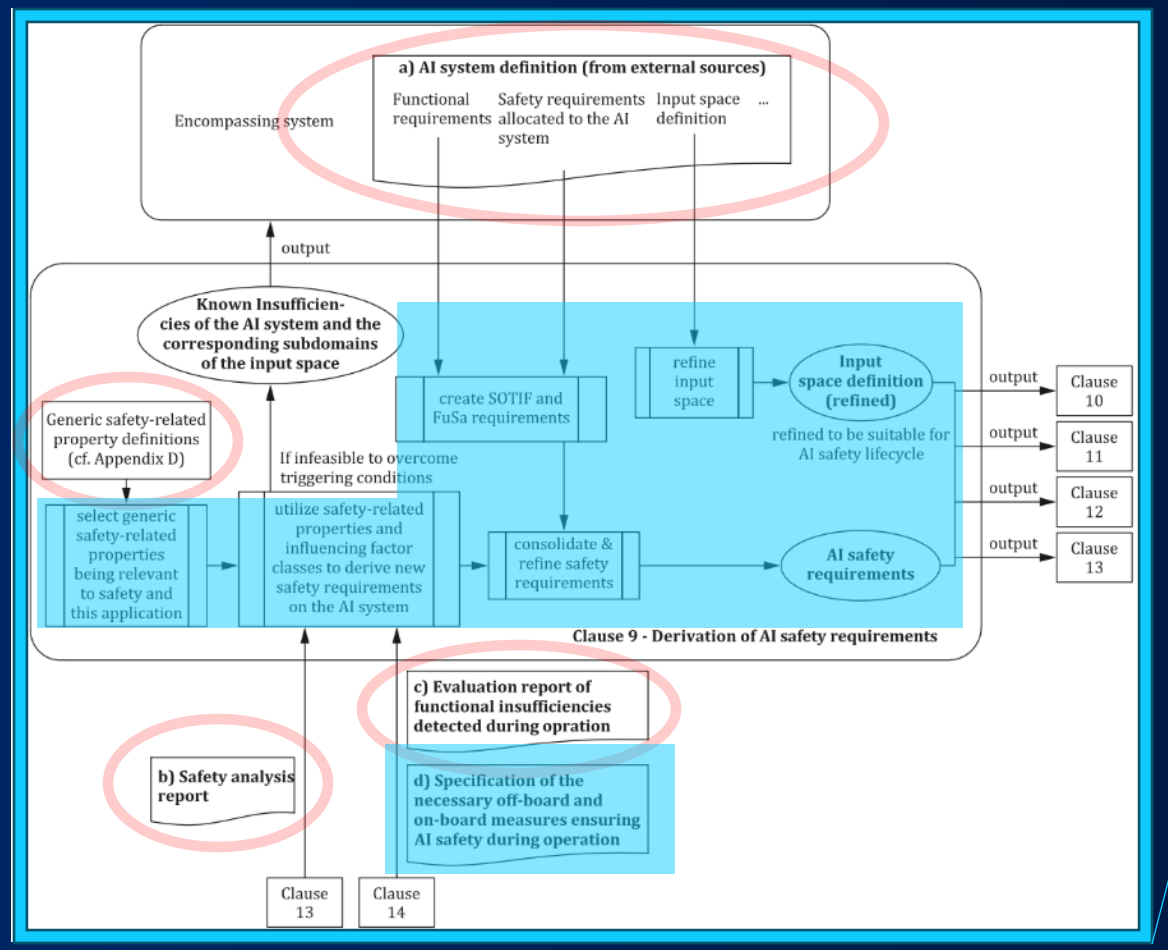
ISO 8800 figure 9-1



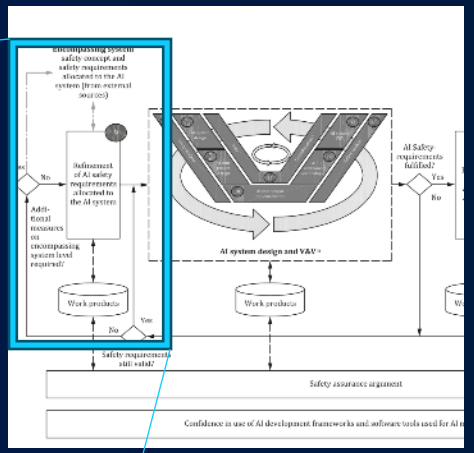
encompassing system
item which contains the *AI system*

AI Safety Requirements Workflow

1. Start from AI system definition (for elements to which have an AI safety impact) and generic safety-related properties, or if in a feedback activity, from the safety analysis reports and evaluation during operation
2. Refine existing or derive new requirements for the specific application, including measures to ensure safety during operation



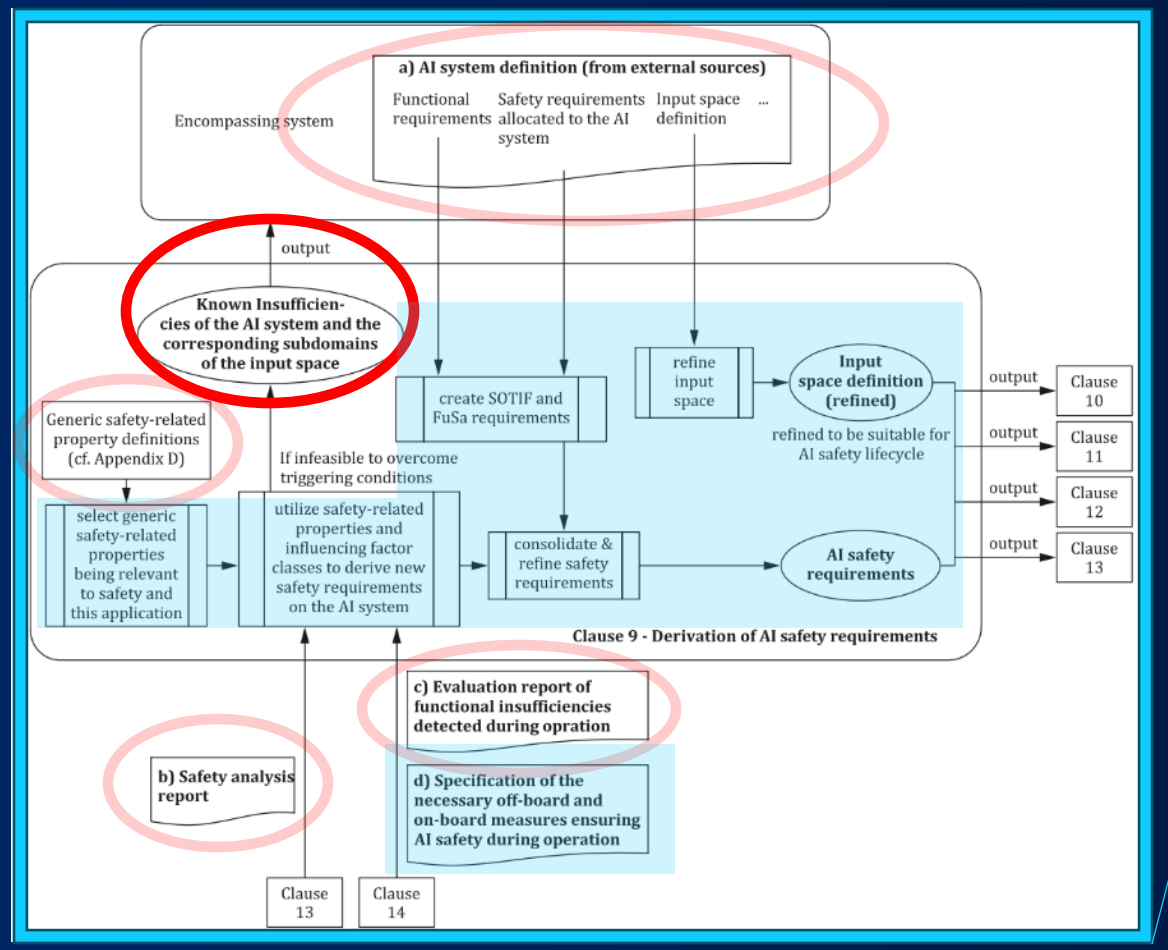
ISO 8800 figure 9-1



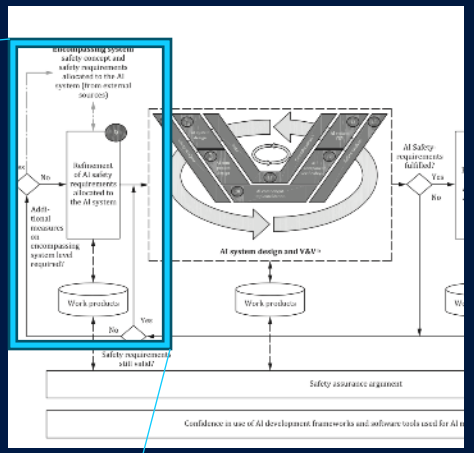
encompassing system
item which contains the *AI system*

AI Safety Requirements Workflow

1. Start from AI system definition (for elements to which have an AI safety impact) and generic safety-related properties, or if in a feedback activity, from the safety analysis reports and evaluation during operation
2. Refine existing or derive new requirements for the specific application, including measures to ensure safety during operation
3. Communicate identified insufficiencies back to the encompassing system



ISO 8800 figure 9-1



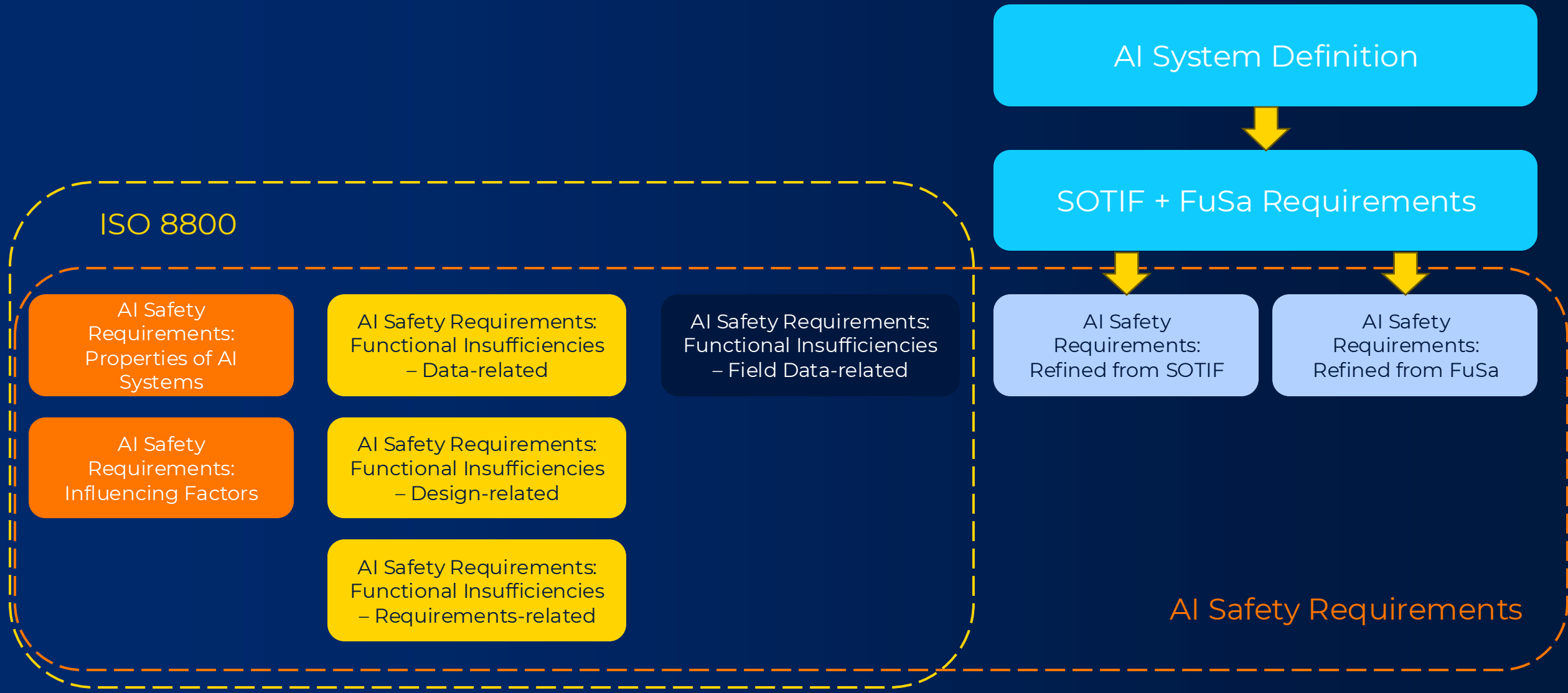
encompassing system
item which contains the *AI system*

AI Safety Requirements

Unique considerations for AI systems – ISO/PAS-8800 clause 9.5.6

- specification of the input space to account for ML limits on generalization
- specification of required number of samples for model training or verification
- consideration of adversarial attacks (note: cybersecurity is outside scope of ISO/PAS 8800) – e. g. “defaced stop sign” or “construction cone attack” or “lane painting attacks”
- side effects of region-specific privacy considerations – e.g., GDPR
- requirements on interfaces to facilitate verification and validation – e.g., for explainability
- requirements on thresholds for false-negatives, false-positives, and signal-to-noise ratios

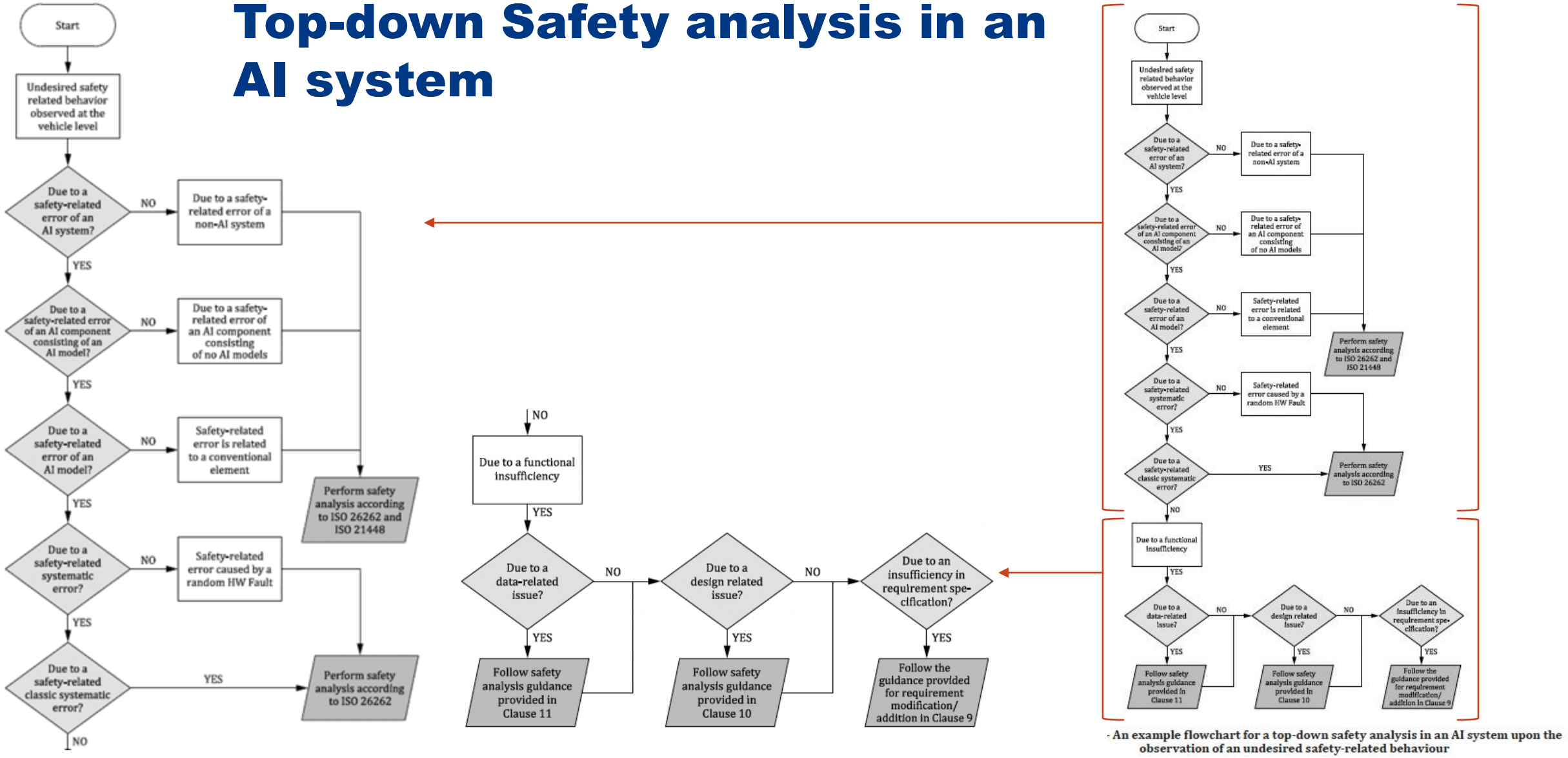
AI Safety Requirements



Practical Application

JAMA CONNECT

Top-down Safety analysis in an AI system



- An example flowchart for a top-down safety analysis in an AI system upon the observation of an undesired safety-related behaviour

ISO/PAS 8800 Figure 13-1

Dataset Lifecycle Model

ISO/PAS 8800

ISO/PAS 11.4.4: Dataset requirements development

The requirements development follows the following activity flow:

a) comprehension of the AI system;

Focus on understanding the intended functionality of the AI system, including:

- The AI safety requirements - Clause 9
  - Safety requirements that are allocated to the AI system
- The input space definition (similar to ODD) - Clause 9

b) dataset safety analysis;

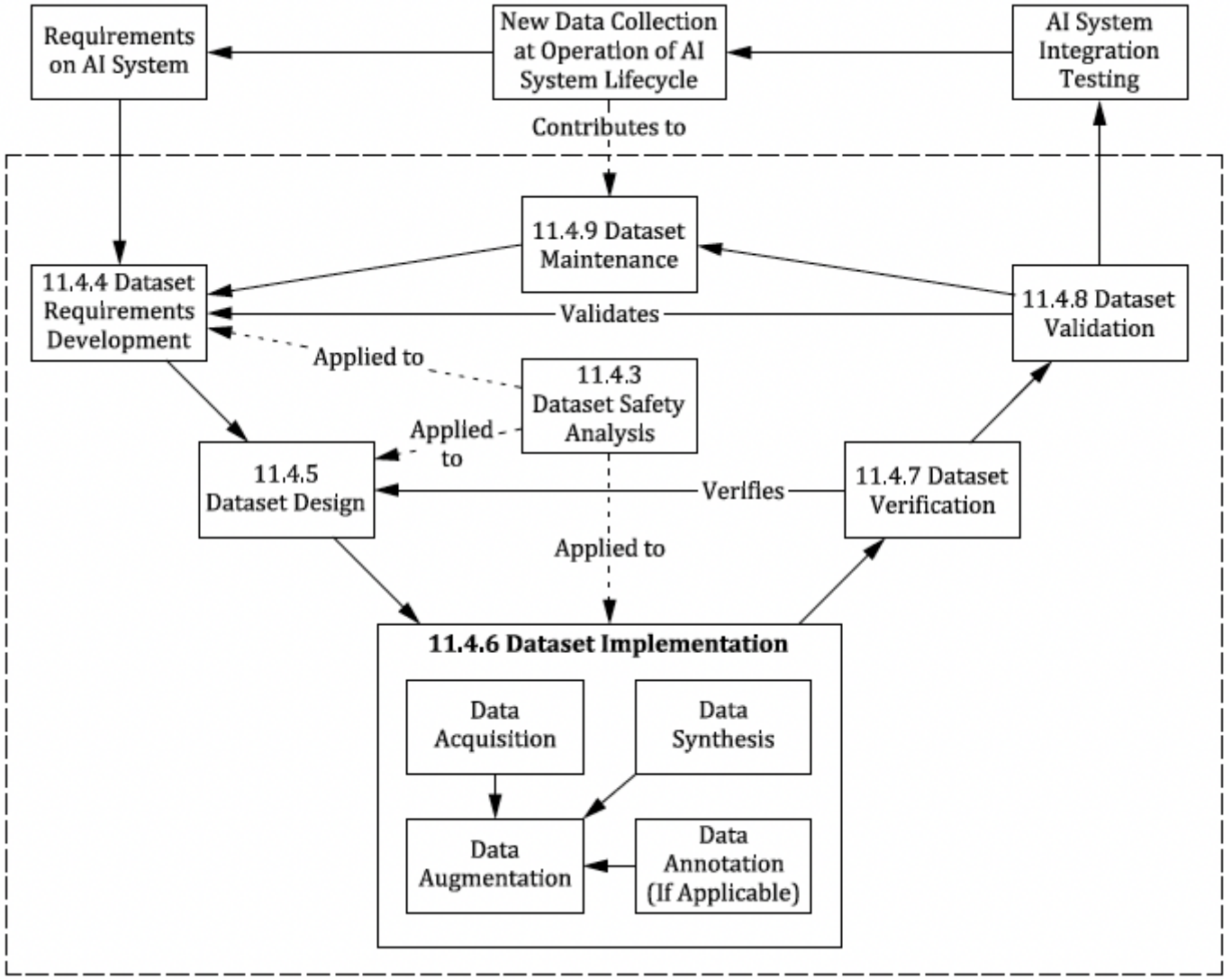
Focus on identifying safety-relevant dataset insufficiencies

c) dataset requirements formulation;

Focuses on formulating the dataset requirements that mitigate the risks associated with the output of the AI system

d) dataset requirements quality assurance.

Focuses on ensuring that the dataset requirements follow the criteria given in ISO 26262-8:2018, Clause 6



Derivation of the AI Safety requirements

ISO/PAS 8800

AI system definition comprised of:

- safety requirements allocated to the AI system;
- input space definition;
- functional requirements;
- impacted stakeholders;
- the interfaces of the AI system with the encompassing system, including if applicable, the ASIL capability of the inputs to the AI system;
- interfaces to the environment, if applicable.

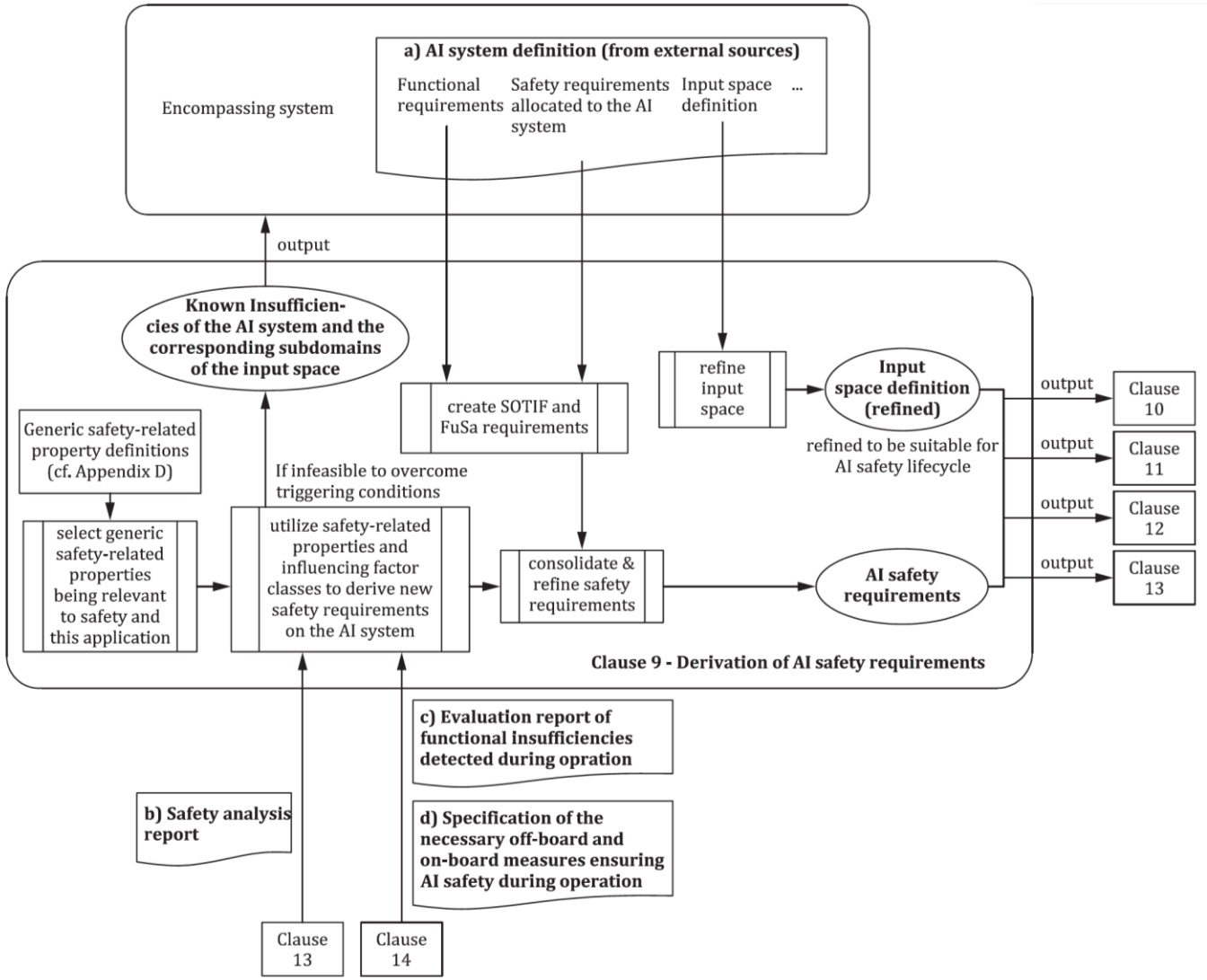
Refinement outputs:

Clause 10: AI technologies, architectural and development measures

Clause 11: Data-related considerations

Clause 12: Verification and validation of the AI system

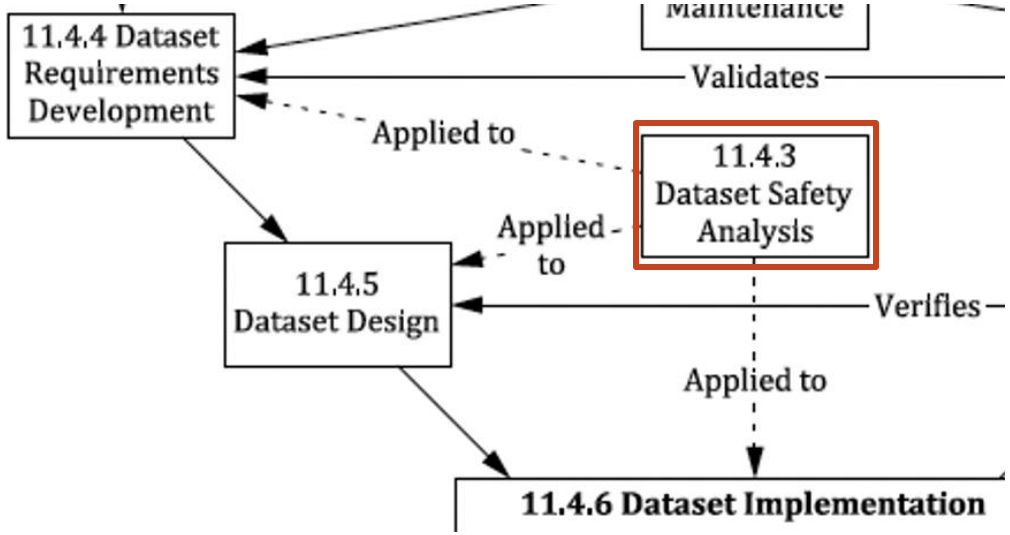
Clause 13: Safety analysis of AI systems



Dataset Safety Analysis

ISO/PAS 8800 11.4.4

The dataset safety analysis activity is performed in line with 11.4.3. The outputs are fed into the dataset requirements formulation.



Dataset safety analyses focus on identifying safety-relevant dataset insufficiencies. When these dataset insufficiencies have been examined and the causes and consequences have been identified (including risks of the AI system and encompassing system), that information is fed as inputs to the dataset requirements development, dataset design, and dataset implementation to realize:

— countermeasures to prevent or mitigate dataset insufficiencies;

— metrics to judge effectiveness of measures to avoid dataset insufficiencies.

Dataset Safety Analysis

ISO/PAS 8800 11.4.4

Dataset insufficiencies are insufficiencies of the dataset regarding data-related safety properties under consideration.

Table 11-1 — Examples of data-related safety properties

Property	Definition
Accuracy	The data correspond to their source with respect to semantical representation and interpretation.
Completeness	The data elements (including metadata) are populated and the data have defined coverage of the input space, safety-relevant cases and plausible data perturbations.
Correctness (or fidelity)	The data correspond to the phenomenon they intend to capture and include features and metadata which help to characterize the phenomenon.
Independence of datasets	The datasets sufficiently avoid leakage of information amongst themselves with respect to data sources and the methods used to capture, gather, generate and process the data.
Integrity	The data are not altered by natural phenomenon (e.g. noise) or intentional action (e.g. usage of lossy data compression without consideration of impact to model, poisoning).
Representativeness	The distribution of data corresponds to the information in the environment of the phenomenon to be captured; it is free of biases.
Temporality	The data gives sufficient consideration to time-based characteristics (e.g. timeliness, ageing, lifetime, time contributing to distribution shift).
Traceability	The derivation of the data from their origin (including information on how they were captured, gathered, generated and processed) is demonstrated.
Verifiability	The data include sufficient features to be amenable for verification as prescribed by their requirements and properties.

Dataset Requirements Formulation

ISO/PAS 8800 11.4.4

Involves the formulation of dataset requirements based on Logistical and Technical aspects

The dataset requirements formulation activity focuses on formulating the dataset requirements that mitigate the risks associated with the output of the AI system. It specifies:

- the logistical aspects, addressing at least the following items:
  - where the dataset is stored;
  - who has access to the dataset, what type of access they have and when they have access, including consideration given to ensure that this dataset is safe from unintended editing;
  - how the dataset is version controlled and how changes are tracked;
  - requirements on the verification and validation processes to be employed to ensure that the data within the dataset is correct and appropriate for usage;
  - how stakeholders can report known vulnerabilities, risks or biases in the data and/or dataset during any of the dataset life cycle phases.
- the technical aspects, addressing at least the following items:
  - size of the dataset;
  - format of the data within the dataset, including what syntactic and semantic parameters describe the data and what the format for labelling is;
  - boundaries of the data within the dataset (driven by both ground truth and design decisions);
  - dataset's role (AI training, AI validation or AI test) and what ensures that it is sufficient for its given role, including limitations on how many times it can be used (to avoid overfitting);
  - constraints affecting creation of the dataset (e.g. region-specific data privacy regulations);
  - mitigations for the different manifestations of dataset insufficiencies detailed in 11.4.3;
  - methods to prevent undetected data failures.

Dataset Requirements Quality assurance

ISO/PAS 8800 11.4.4

- Traceable
- Updateable and maintainable
- Reviewed on an ongoing basis

Finally, the dataset requirements quality assurance activity focuses on ensuring that the dataset requirements follow the criteria given in ISO 26262-8:2018, Clause 6. These requirements are the following:

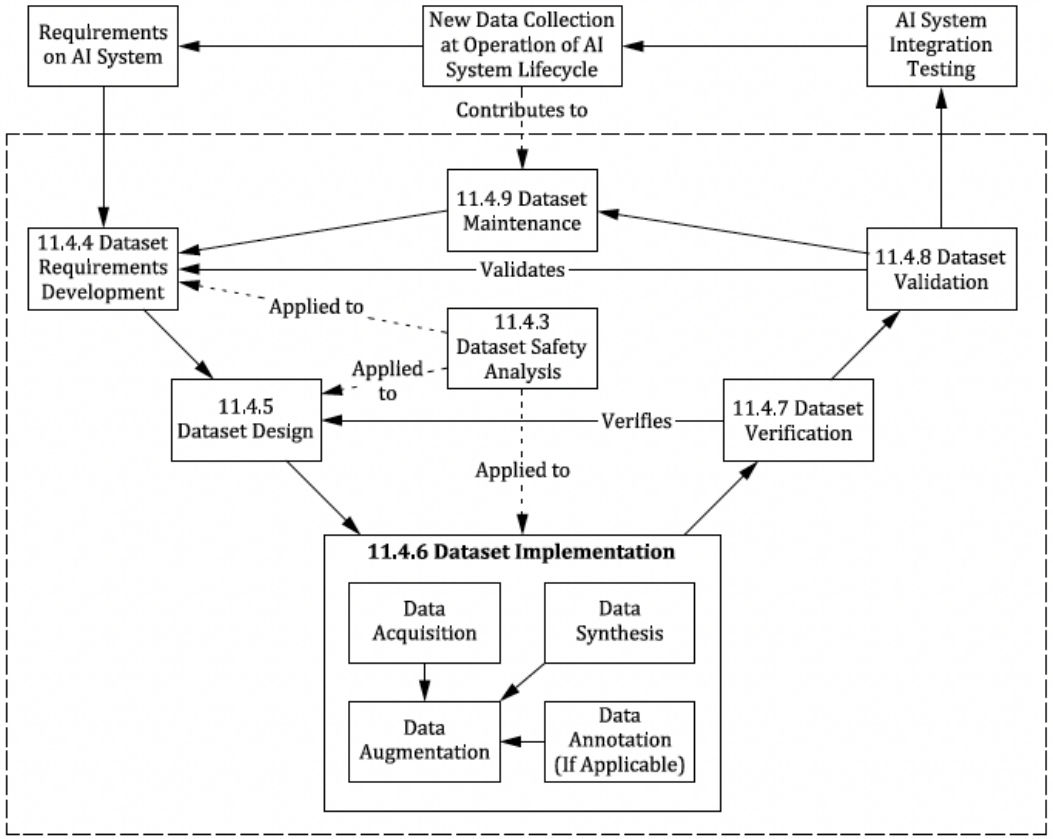
— traceable to the AI safety requirements;

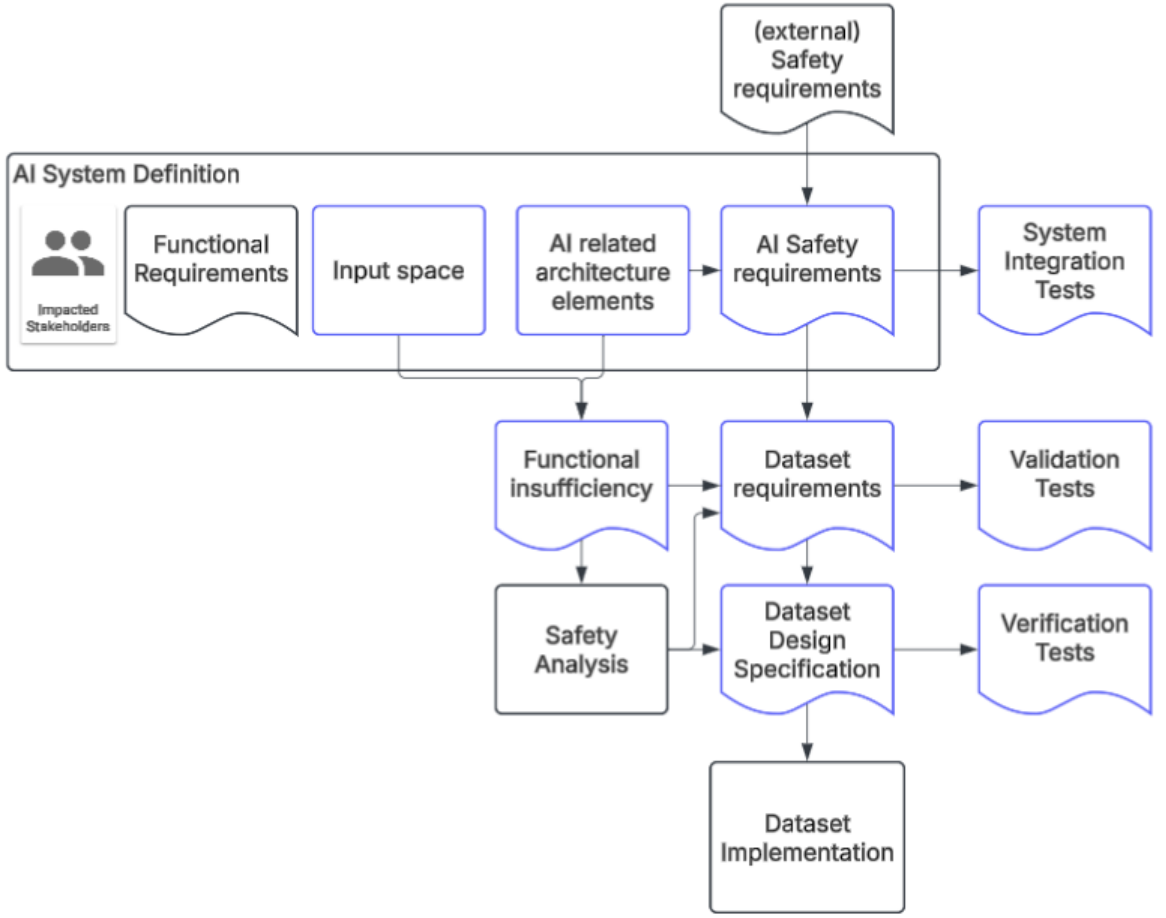
— updatable and maintainable upon a change to the encompassing system, the AI system or input space;

— updatable upon exposure of an insufficiency in the AI system due to the discovery of new safety-relevant scenarios or other triggers.

Documenting in Jama Connect

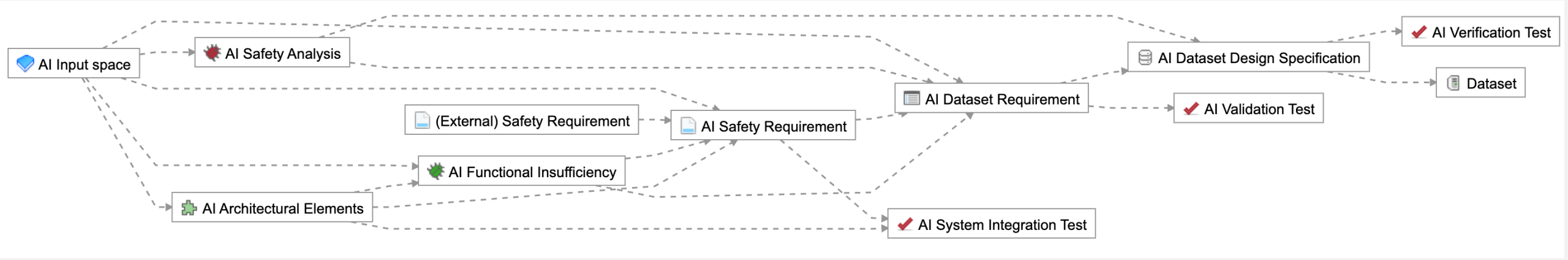
DATASET REQUIREMENT LIFECYCLE





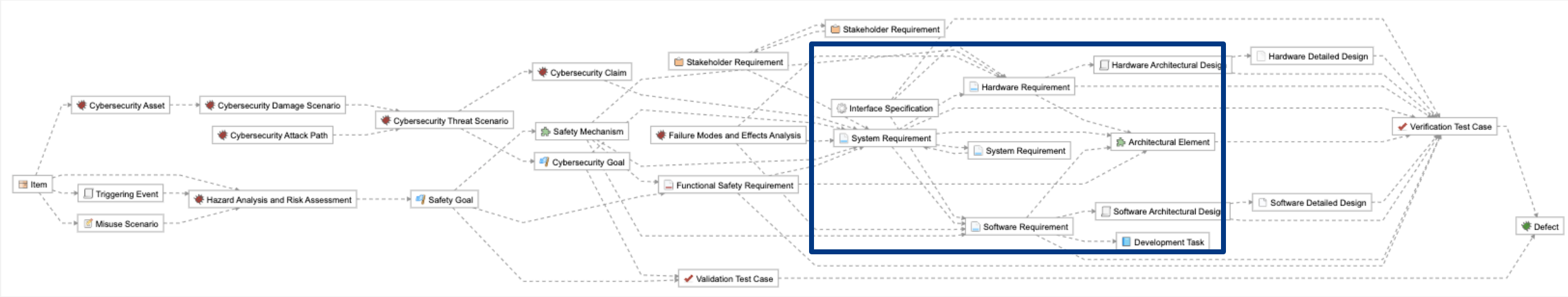
Traceability Model

JAMA CONNECT



Traceability Model

JAMA CONNECT



Demonstration

JAMA CONNECT

Q&A

THANK YOU!