AI and Functional safety – Pain or gain or both?

This presentation was held online on October 9th 2024 in an live connect ISA SafeSec event

Introductions - Panelists

Thor Myklebust
Sr. Scientist
SINTEF Digital

- Active in standardization: IEC 61508, CENELEC/TC 9EN 5012x
- Co-funder of SafeScrum and The Agile Safety Case
- Prior experience with certification of products and systems
- Co-author of four books

Mary Ann Lundteigen
Professor, Norwegian University of Science and Technology (NTNU)

- Professor in instrumentation systems and safety in study programs within Engineering Cybernetics
- Functional safety, reliability analysis, digitalization within Industry 4.0
- Active in standardization: MT 61511, ISA 84 (WG 10)
- Part of organizing workshops and conferences in Norway with industry participation: PDS forum, CDS forum, and the AAS forum.

Niclas Flehmig
PhD candidate, NTNU

- PhD scholarship funded by SUBPRO ZERO headed by NTNU
- Research topic: AI applied to safety-critical systems
- Holds a masters degree in mechanical engineering with a specialization in applied machine learning from the Technical University of Munich

Agenda

- Short welcome
- Three short presentations (3 x 10-15 min) in total approx. 40-45 min):
  1. Overall frameworks, standards, and application areas of AI.
     Sr. Researcher Thor Myklebust, SINTEF Digital
  2. Perspectives on AI and safety instrumented systems.
     Professor Mary Ann Lundteigen. Norwegian university of Science and Technology (NTNU)
  3. A good system architecture is all you need?
     PhD candidate Niclas Flehmig
- Q&A (15 minutes)

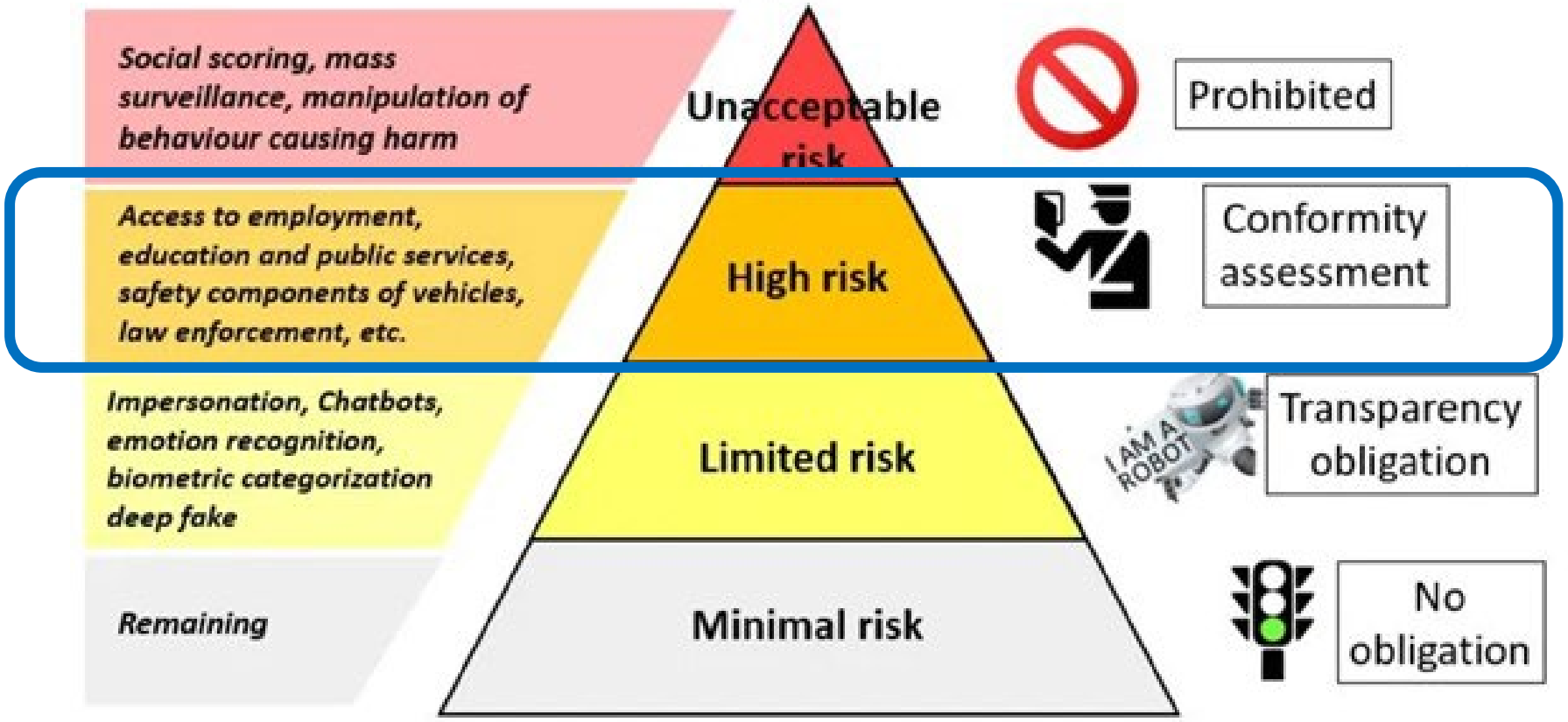
PART I: Overall frameworks, standards, and application areas of AI.

Sr. Researcher Thor Myklebust,

SINTEF Digital

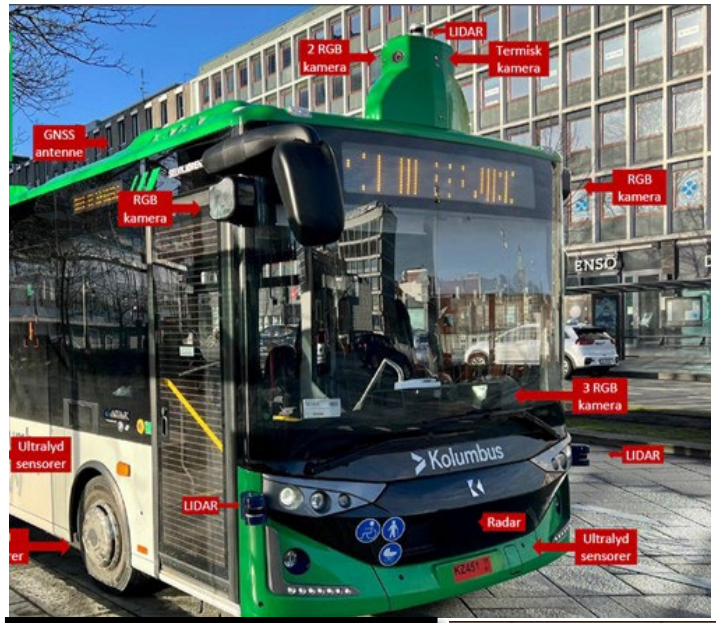
Global Market Access

EU Artificial Intelligence Act: Risk levels



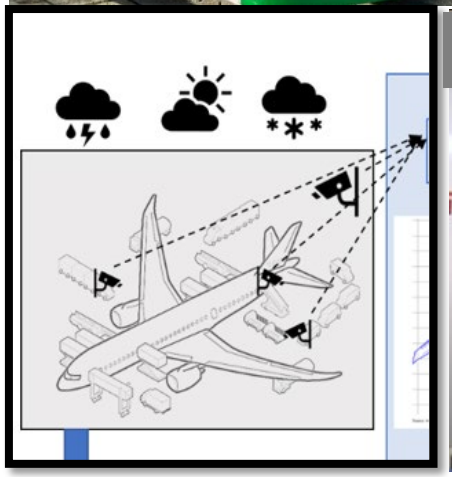
Extract from Annex III "High-risk AI systems": AI systems intended to be used as safety components in the management and operation of **critical digital infrastructure**, **road traffic, and the supply of water, gas, heating, and electricity etc**.

Applied AI and AI projects



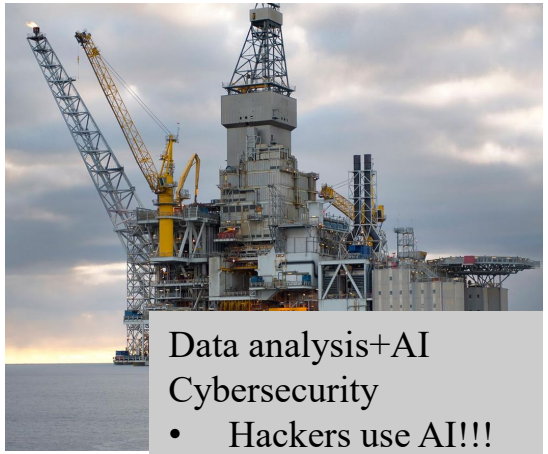
Safety products/systems
Multi-domain

Autonomous ferry, Stockholm



Hyperloop

Railway, Metro and Tram



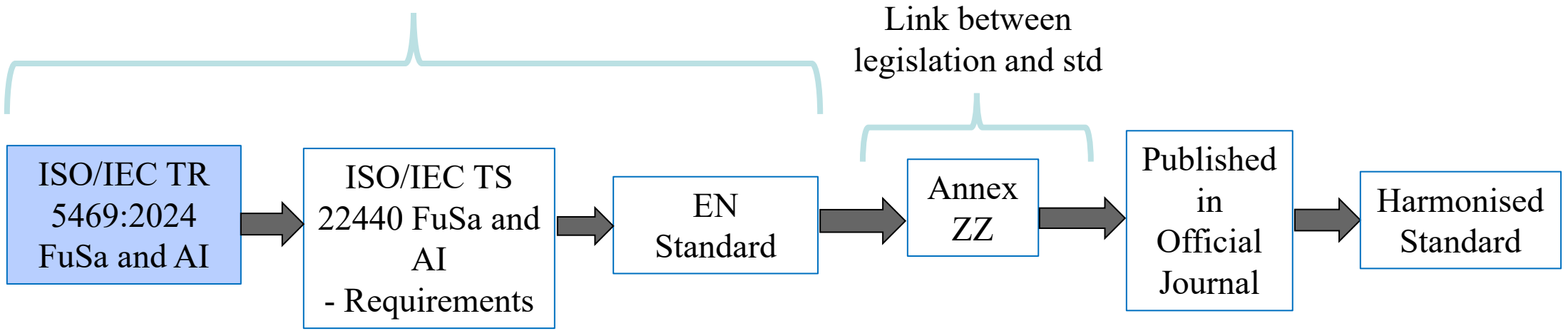
Data analysis+AI
Cybersecurity
- Hackers use AI!!!

Harmonised standards (HS)

HS: Manufacturers, other economic operators, or conformity assessment bodies can use harmonized standards to demonstrate that products, services, or processes comply with relevant EU legislation.

So, there is a very strong, innovative and pragmatic link between legislation and standards!

- TR (Technical Reports), TS (Technical Specifications) and EN standards are expected to be “required” by certification bodies
- A “profile approach” is recommended. E.g. relevant subsets of the TR/TS/Standards



Standards, AI and Safety Case (SC)

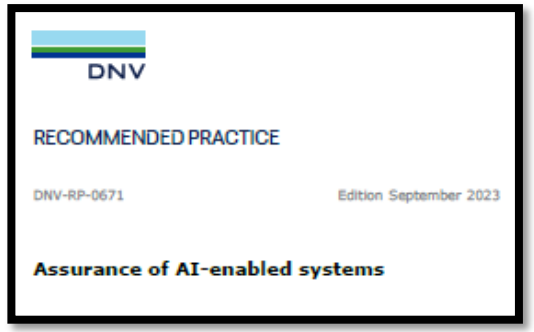
IEC 61508 is somewhat outdated, TR is informative, and AI is a challenging topic requiring improved argumentations and justifications

Assurance case (safety case) is the suggested solution. SC Idea:

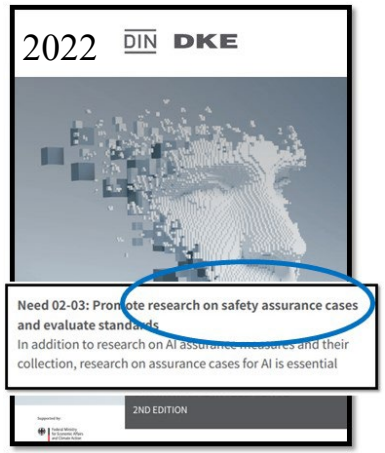
- The idea of a safety case is to argue as one would in a court of law – thus the name safety case.
- Evidence using a Safety case can be extended to cover safety issues beyond the scope of safety standards

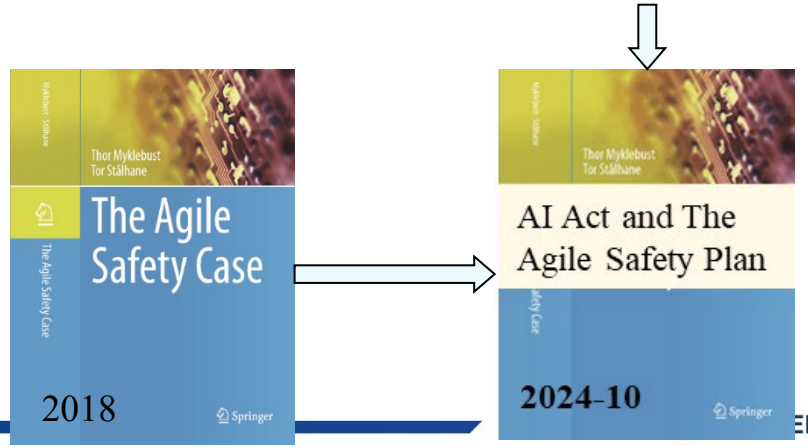
New and challenging AI SC topics

1. Accept criteria
2. Algorithms
3. Argumentation
4. Autonomy
5. Bias
6. Compliance
7. Data
8. Defeater
9. Deployment
10. Emergency
11. Explainable AI (XAI)
12. Post-market
13. Sensor fusion
14. Training



Link to white paper: www.din.de/en/innovation-and-research/artificial-intelligence/ai-roadmap





Fun facts, committee work, strategy and changes!

Leadership changed from "Consultants" to large-size companies

- IEC 61508 generic safety: Convenors: Siemens and NVIDIA
- ISO/IEC TR 5469 FuSa and AI: Convenor: NVIDIA
- NBIM supports standardisation$$$ (indirectly)
- Little financial support for standardisation in Norway

NBIM: Norway's big wealth fund

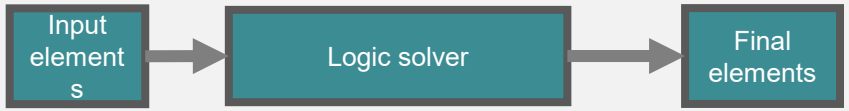
PART II: Perspectives on AI and safety instrumented systems.

Professor Mary Ann Lundteigen

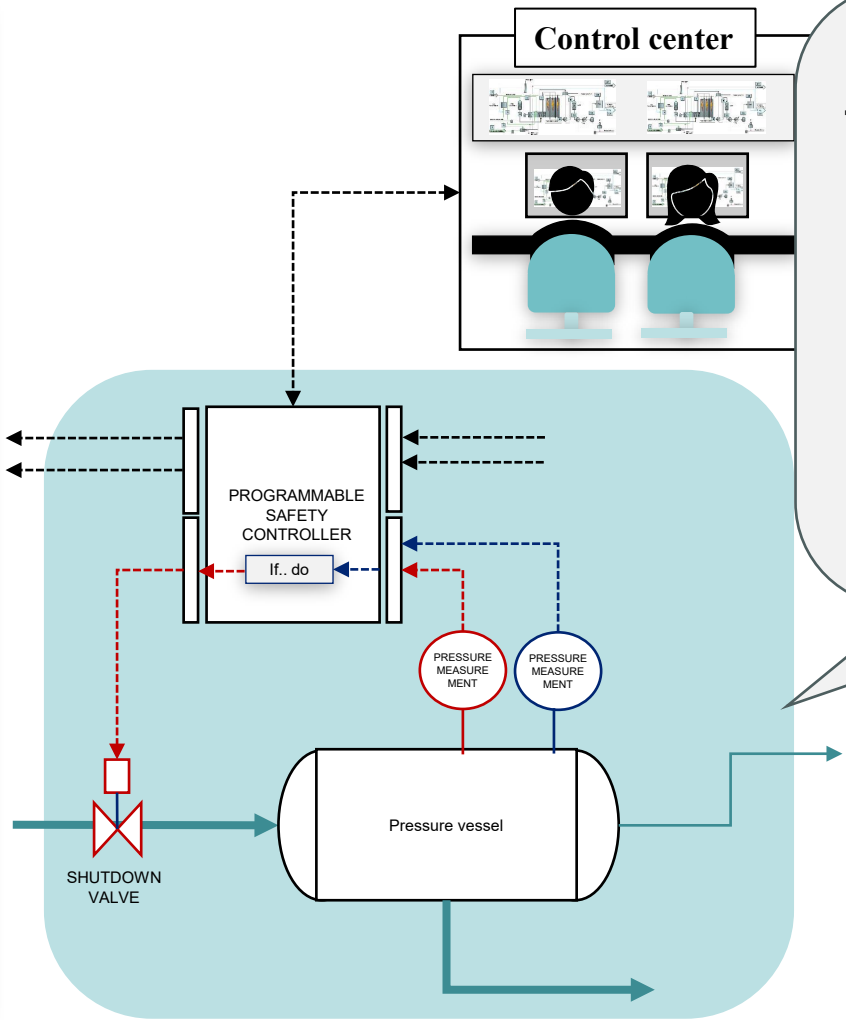
Norwegian University of Science and Technology

What is a safety instrumented system (SIS)?

- SIS : **Process industry term** for safety systems involving technologies such as sensors, controllers, and actuated devices.
- A SIS performs one or more safety **instrumented function (SIF) –** each solving a task in response to demand
- Each SIF is usually split into **three subsystems**.



- Each SIF is designed to meet a safety integrity level **(SIL)** requirement, allocated from a hazards and risk analysis.



Relevant requirements in **functional safety standards**.

Process industry/ manufacturing:
- **IEC 61511**
- IEC 62061/ISO 13849

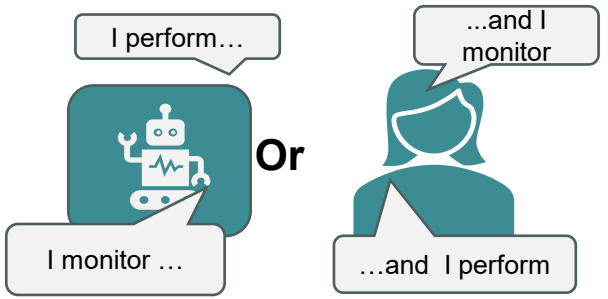
Product development:
- IEC 61508

What is artificial intelligence (AI)?

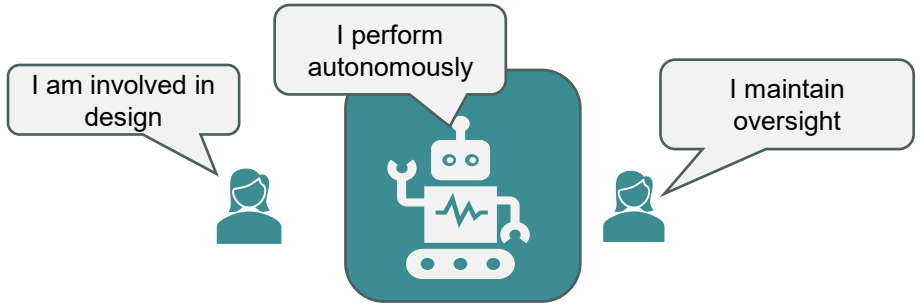
An **engineered system** that **emulates** the performance of **humans** [2].

Level 1: AI assistance to humans

Level 2: AI and human collaborate

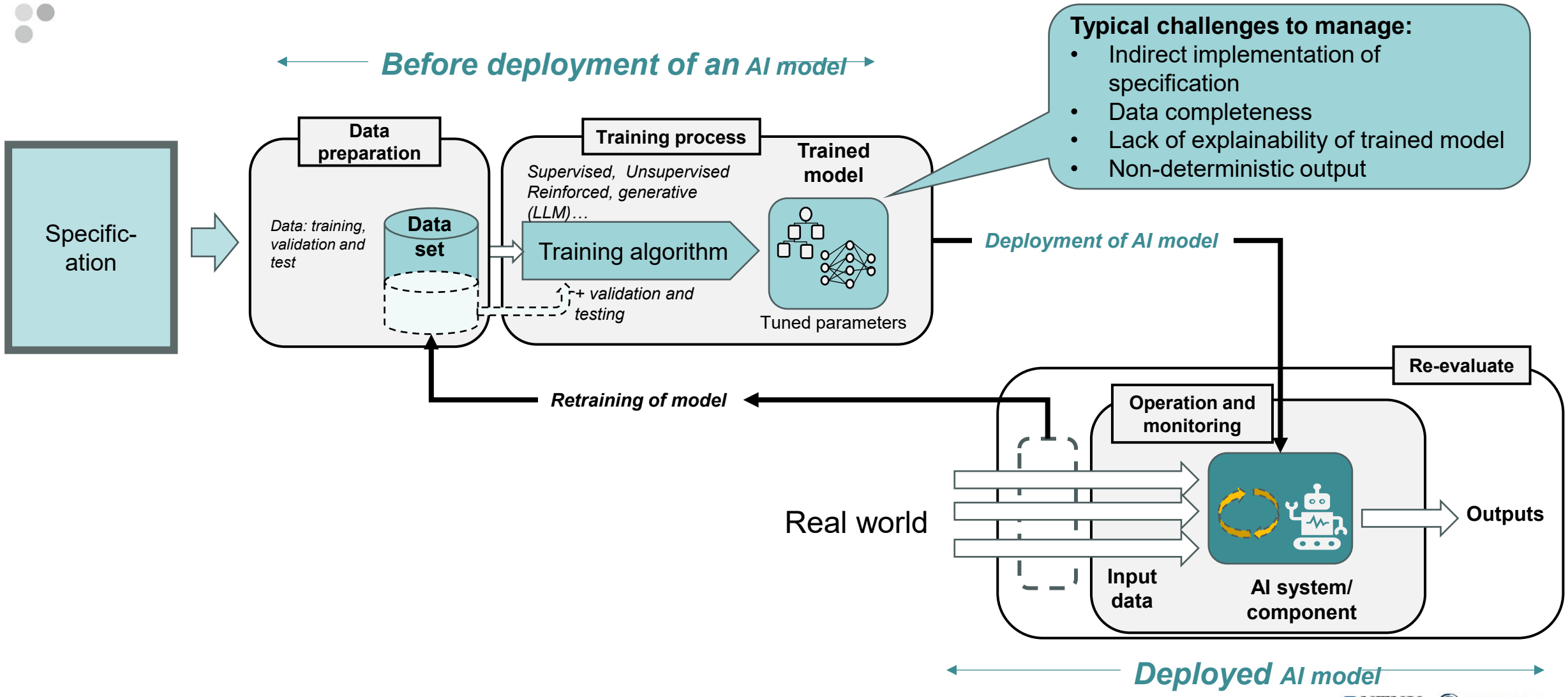


Level 3: AI operating autonomously

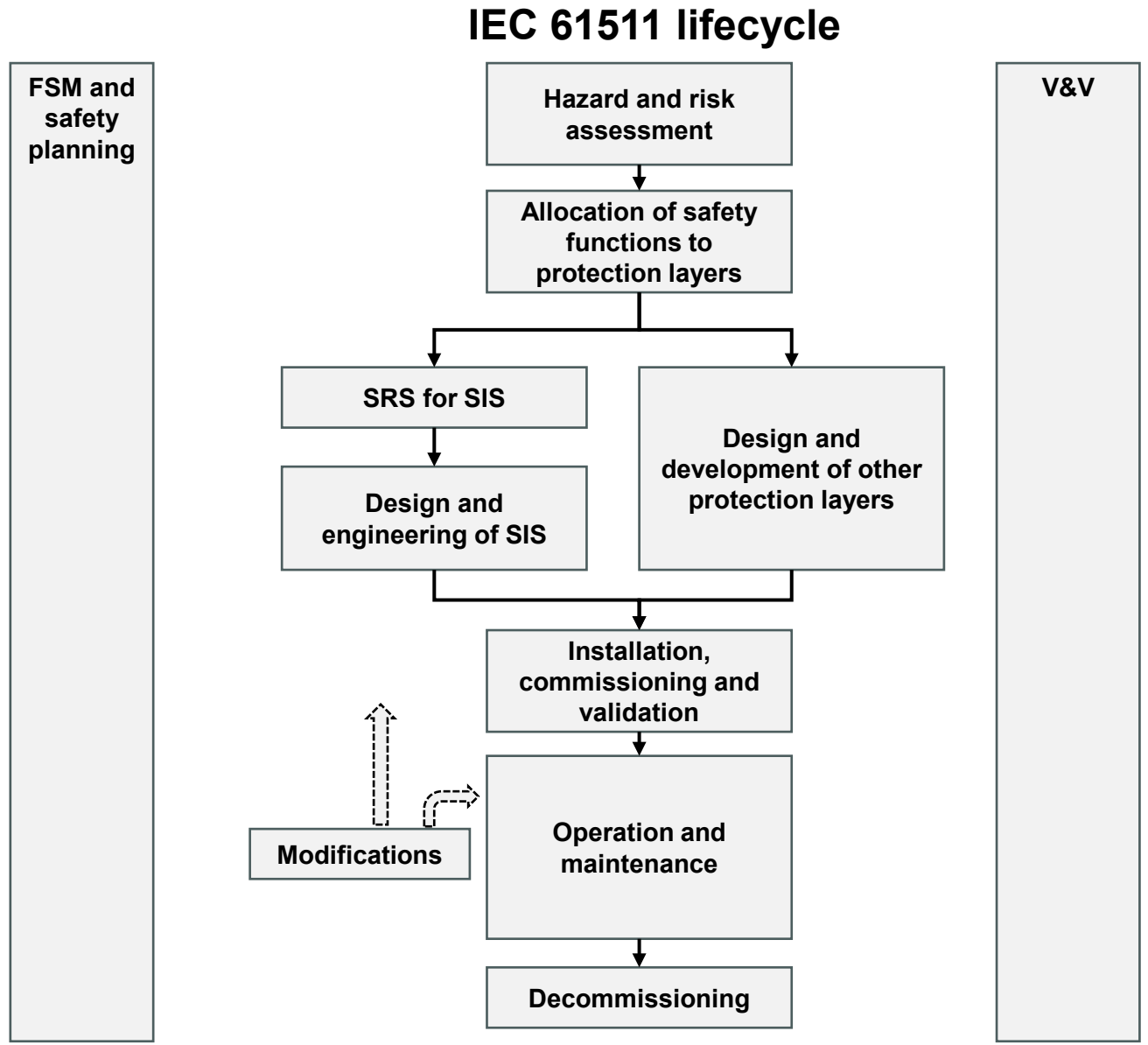


Definition and naming of levels from [2] EASA (2020) Artificial Intelligence Roadmap: A human-centric approach to AI in aviation (ed. 1.0).

How is an AI system developed and used?



Where and for what can AI be applied?



SIS: Safety-instrumented system, SRS: Safety requirements specification, V&V: Verification and validation. FSM: Functional safety management. LOPA: Functional safety management

We may look for tasks that are repetitive or low complexity, but time-consuming to acquire necessary information. For example:

1

AI assistant:

- Provide input from past analyses for consistency
- Help identifying conflicting assumptions

2

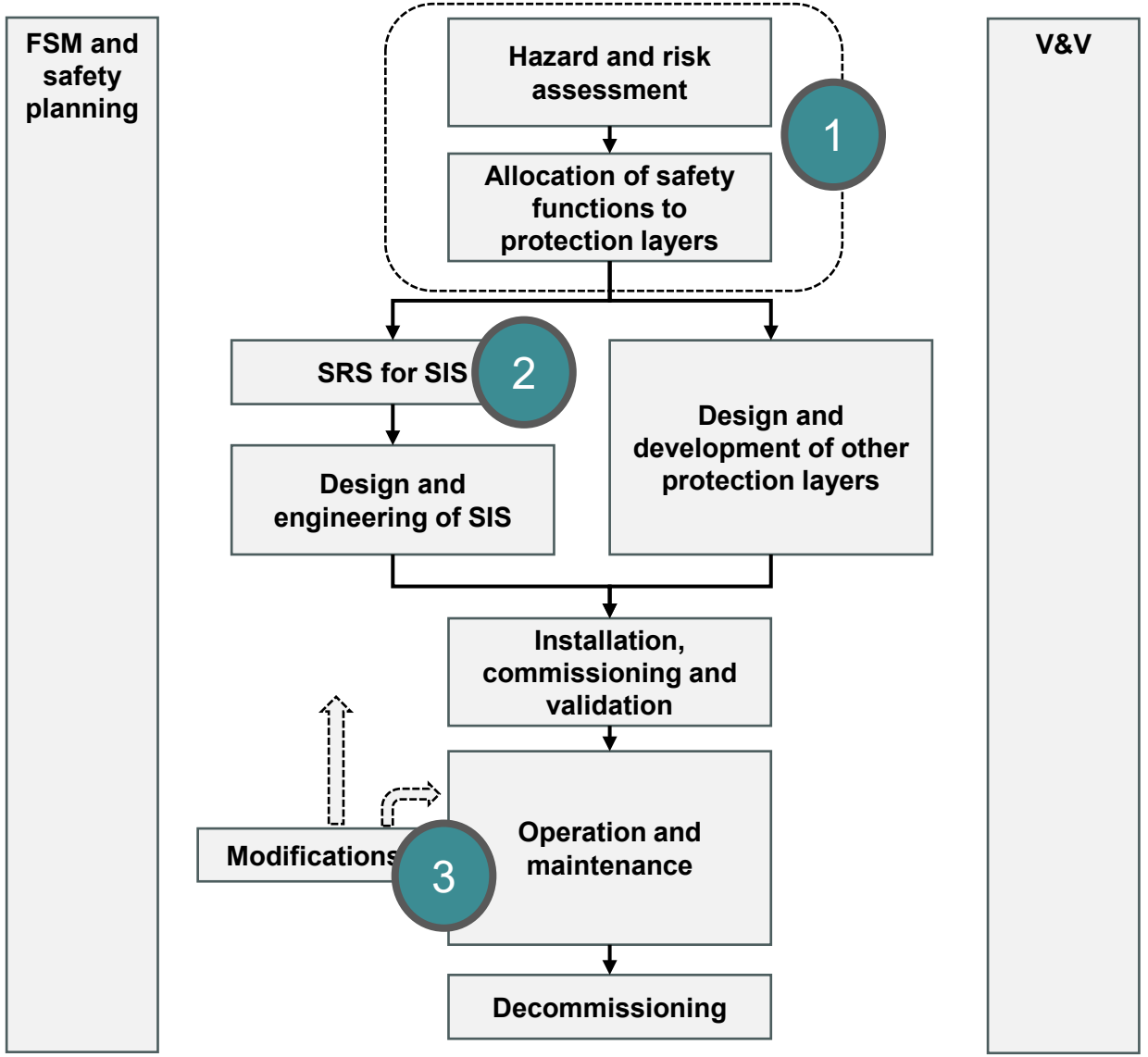
AI assistant:

- Autogenerate data from SRS into machine readable and interoperable formats

3

AI assistant:

- Assist and automate failure (and demand) analysis and classification.
- Support root cause analysis, check against manufacturer use restrictions,...

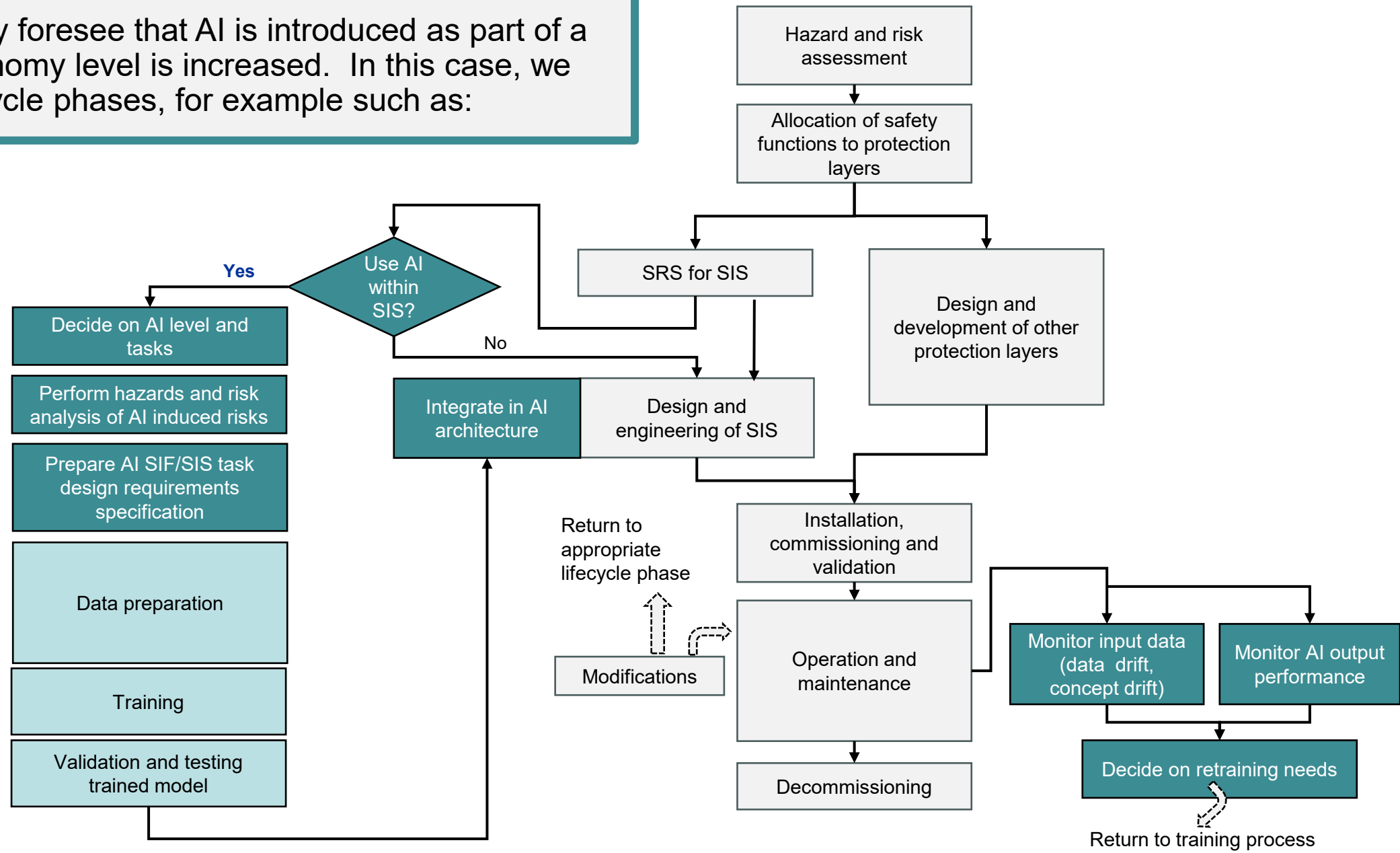


In rare cases we may foresee that AI is introduced as part of a SIF, i.e. where autonomy level is increased. In this case, we need additional lifecycle phases, for example such as:

These additional phases are **not covered** by existing functional safety standards.

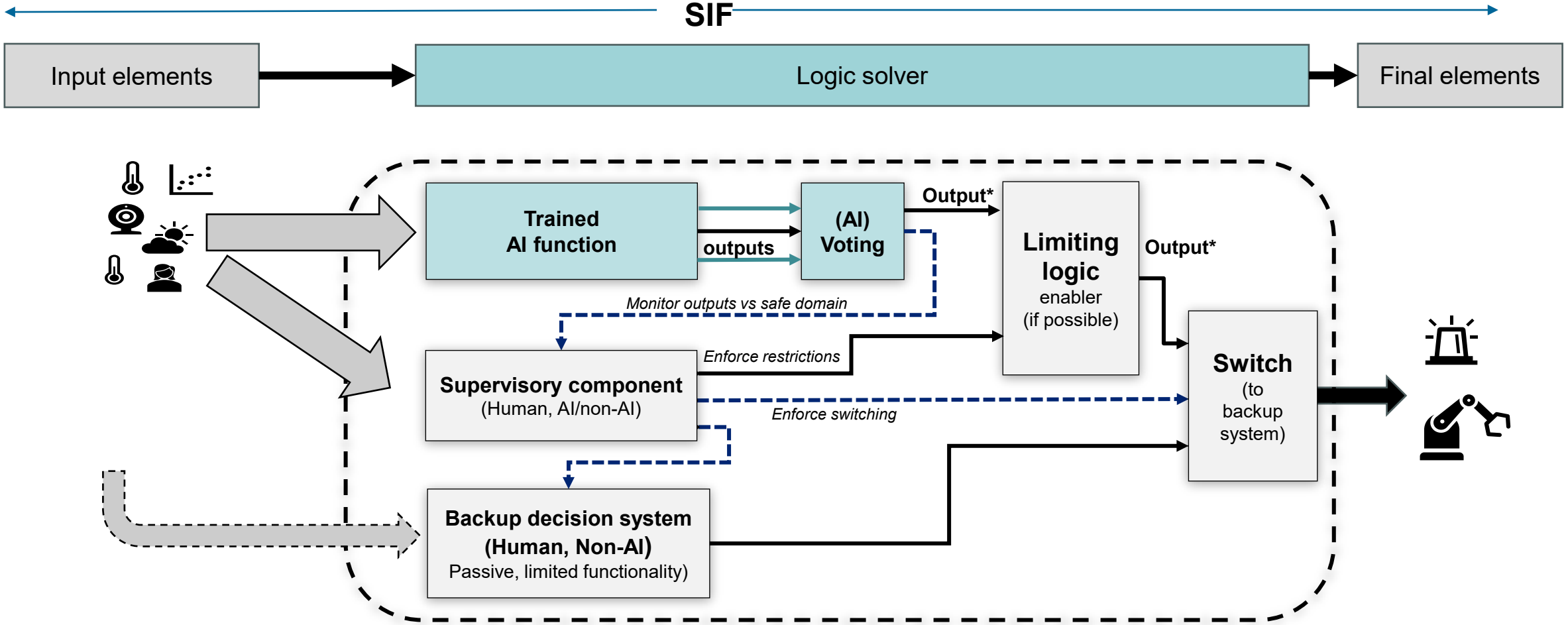
However, some IEC 61508-2 & -3 methods may be **relevant**, with reference to ISO/IEC TR 5469

But we need to consider: What is the scope for AI within IEC 61508 vs IEC 61511.



SIS: Safety-instrumented system, SIF: Safety instrumented function. SRS: Safety requirements specification, V&V: Verification and validation. FSM: Functional safety management

Integrated AI in a **fault tolerant** architecture

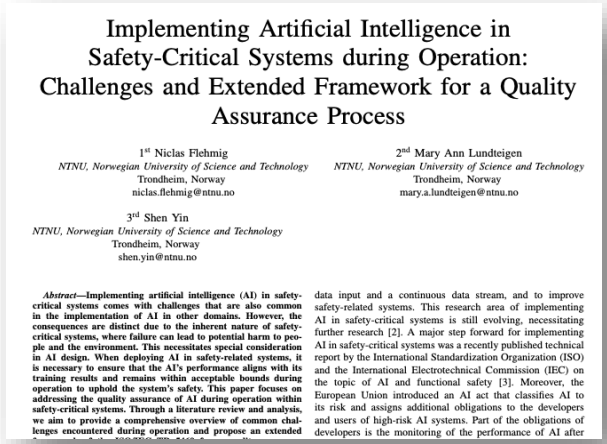


Adapted from [1] ISO/IEC TR 5469 on functional safety and AI systems

PART III: A good system (AI) architecture is all you need?

PhD Candidate Niclas Flehmig

Norwegian University of Science and Technology (NTNU)



Implementing Artificial Intelligence in Safety-Critical Systems during Operation: Challenges and Extended Framework for a Quality Assurance Process

1st Niclas Flehmig
NTNU, Norwegian University of Science and Technology
Trondheim, Norway
niclas.flehmig@ntnu.no

2nd Mary Ann Lundteigen
NTNU, Norwegian University of Science and Technology
Trondheim, Norway
mary.a.lundteigen@ntnu.no

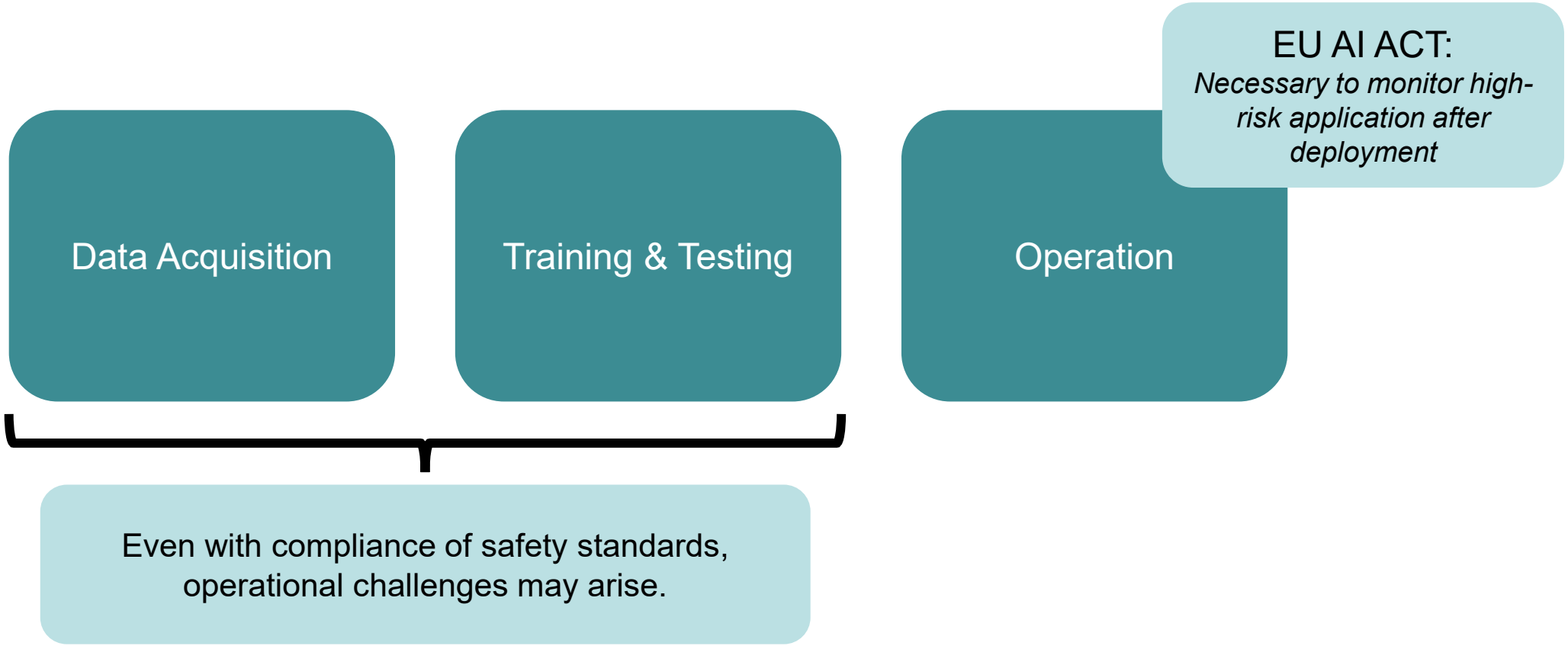
3rd Shen Yin
NTNU, Norwegian University of Science and Technology
Trondheim, Norway
shen.yin@ntnu.no

***Abstract*—Implementing artificial intelligence (AI) in safety-critical systems comes with challenges that are also common in the implementation of AI in other domains. However, the consequences are distinct due to the inherent nature of safety-critical systems, where failure can lead to potential harm to people and the environment. This necessitates special consideration in AI design. When deploying AI in safety-related systems, it is necessary to ensure that the AI's performance aligns with its training results and remains within acceptable bounds during operation to uphold the system's safety. This paper focuses on addressing the quality assurance of AI during operation within safety-critical systems. Through a literature review and analysis, we aim to provide a comprehensive overview of common challenges encountered during operation and propose an extended**

data input and a continuous data stream, and to improve safety-related systems. This research area of implementing AI in safety-critical systems is still evolving, necessitating further research [2]. A major step forward for implementing AI in safety-critical systems was a recently published technical report by the International Standardization Organization (ISO) and the International Electrotechnical Commission (IEC) on the topic of AI and functional safety [3]. Moreover, the European Union introduced an AI act that classifies AI to its risk and assigns additional obligations to the developers and users of high-risk AI systems. Part of the obligations of developers is the monitoring of the performance of AI after

To be presented in November at IECON Chicago 2024

Simplified AI lifecycle



Training and Testing is not enough…

Potential challenges for AI during operation

- **Data drift** (e.g. deviations from training environment)
- **Concept drift** (e.g. new camera system that provides a different resolution)
- **Missing data** (e.g. if input is a data stream)
- **Outliers / Events of interest**
- **Adversarial inputs** (e.g. cyber attack)
- **System related changes** (e.g. changes in software that provide data to the AI)

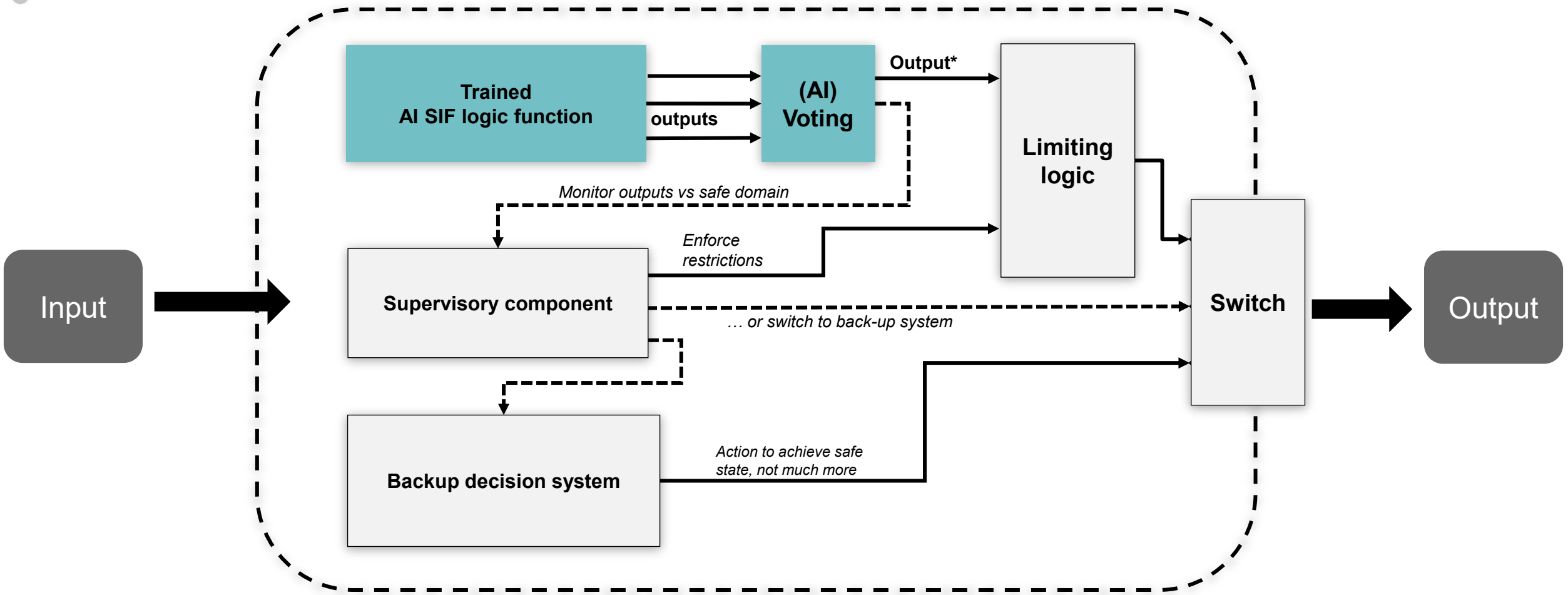
Handling those challenges during operation

Updating
Re-training & re-testing if necessary due to misbehavior or performance loss

Monitoring
Observe the AI component, its input(s) and output(s) to collect information

Communication
Communicate system changes or external factors with responsible persons

Quick recap on the system architecture…



Source. Adapted from [1]

Our concept

Take the suggested architecture of ISO/IEC TR 5469 and extend it

Goal

Tackle the challenge of **quality assurance of AI** during operation as **holistic** approach (**monitoring** + **updating**)

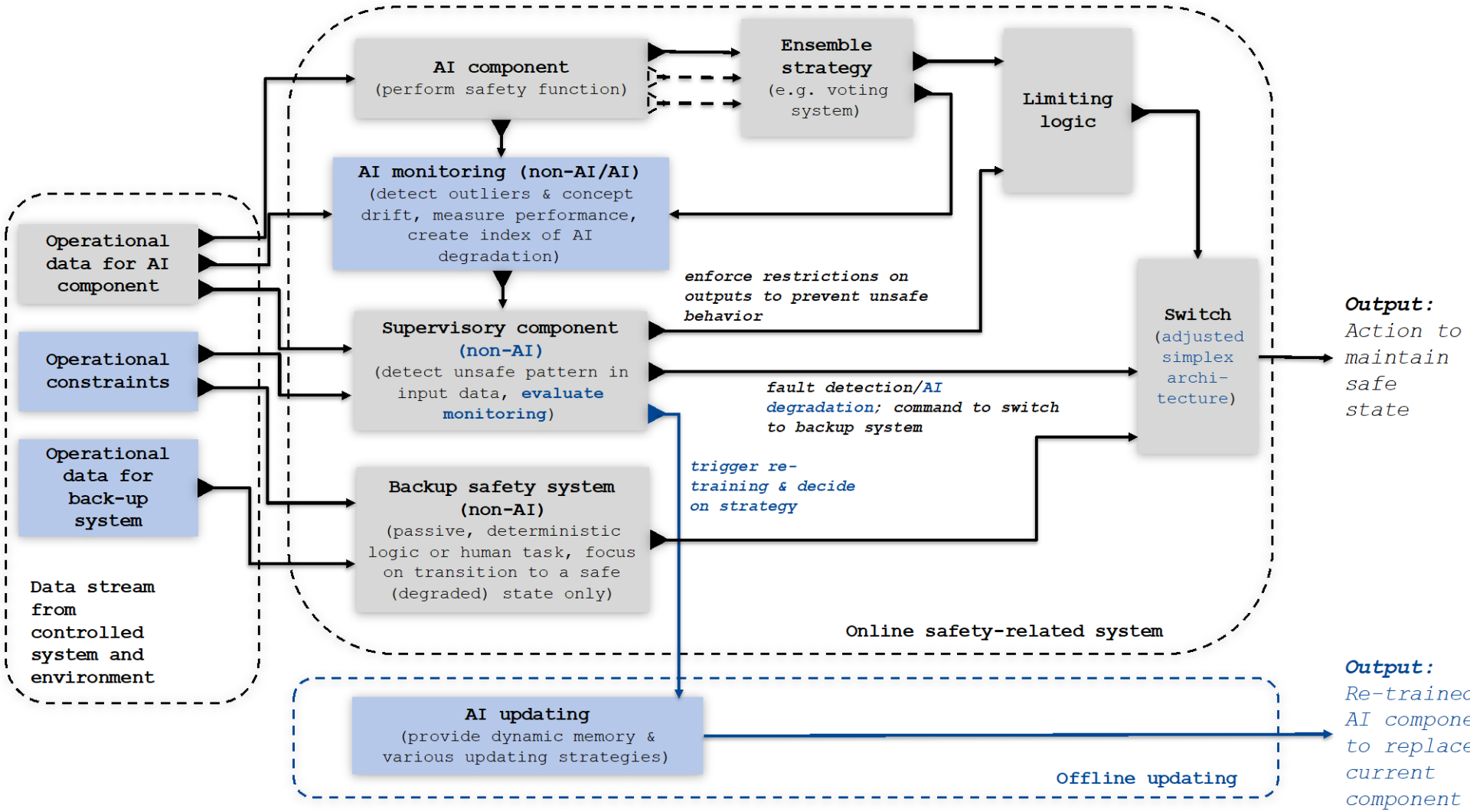
1. Difference to ISO/IEC TR 5469

- ***Separate*** monitoring component and supervisory component
- Suggest a ***clear allocation*** of tasks between supervisory component (e.g human) and monitoring component
- Supervisory component should be a ***human*** to gain ***trust*** in application

2. Difference to ISO/IEC TR 5469

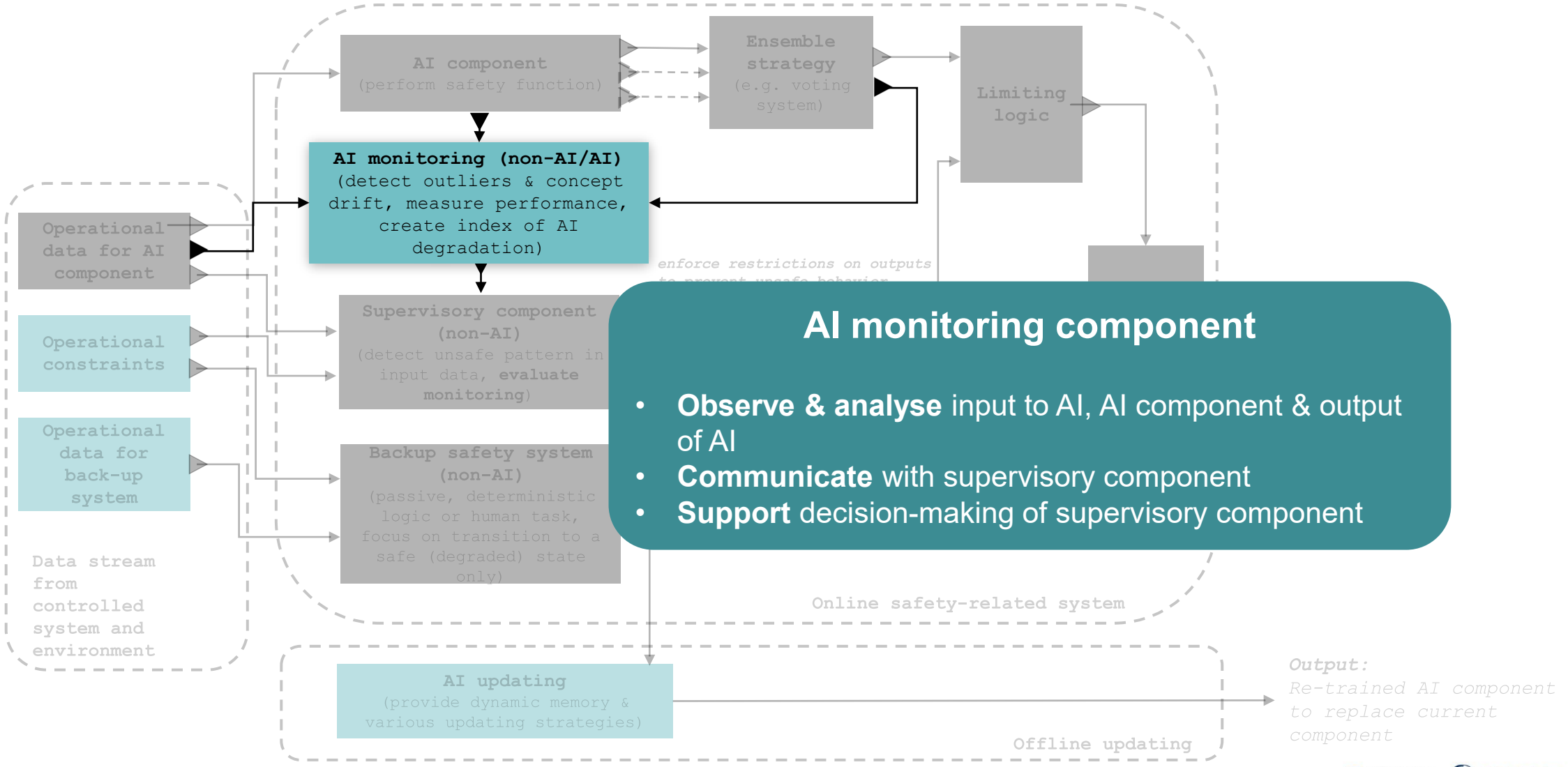
- Add an ***AI updating*** component
- Monitored and controled by supervisory component
- ***Toolbox*** with different updating methods according to requirements

Visualizing our concept



Source: [6]

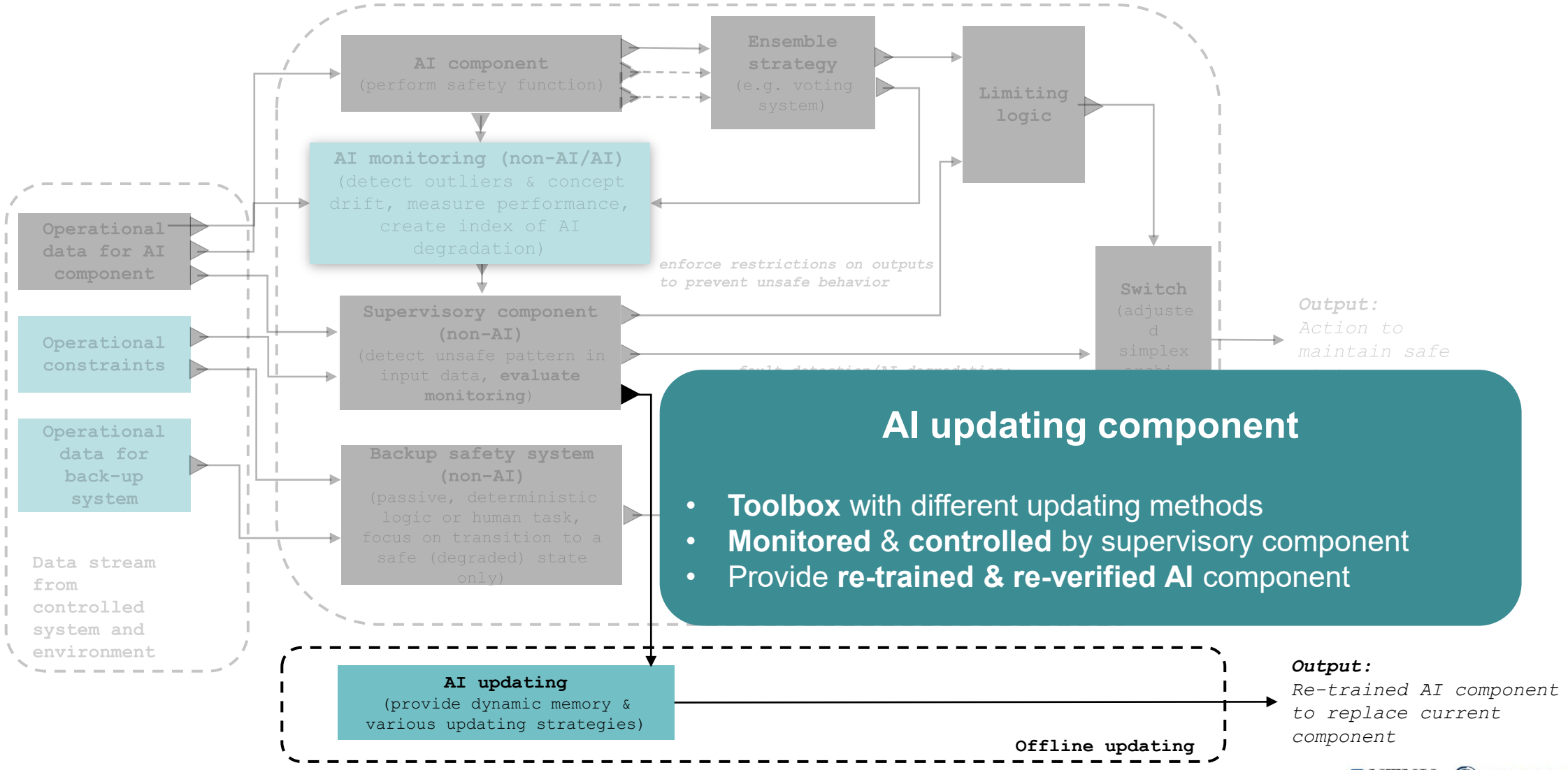
AI monitoring component



AI monitoring component

- **Observe & analyse** input to AI, AI component & output of AI
- **Communicate** with supervisory component
- **Support** decision-making of supervisory component

AI updating component



AI updating component

- **Toolbox** with different updating methods
- **Monitored** & **controlled** by supervisory component
- Provide **re-trained & re-verified AI** component

The session ended with a live Q&A