HITACHI
Inspire the Next

EU-JAPAN DIGITAL WEEK 2025

"Critical Applications of AI in Industry, Healthcare and Other Sectors" Workshop

OT x AI safety approaches in Hitachi

April 7th, 2025

Satoshi Otsuka

Autonomous Control Research Department

Mobility and Automation Innovation Center

Research & Development Group

Hitachi, Ltd.

HITACHI
Inspire the Next

Contents

HITACHI
Inspire the Next

Contents

1-1. Hitachi corporate data

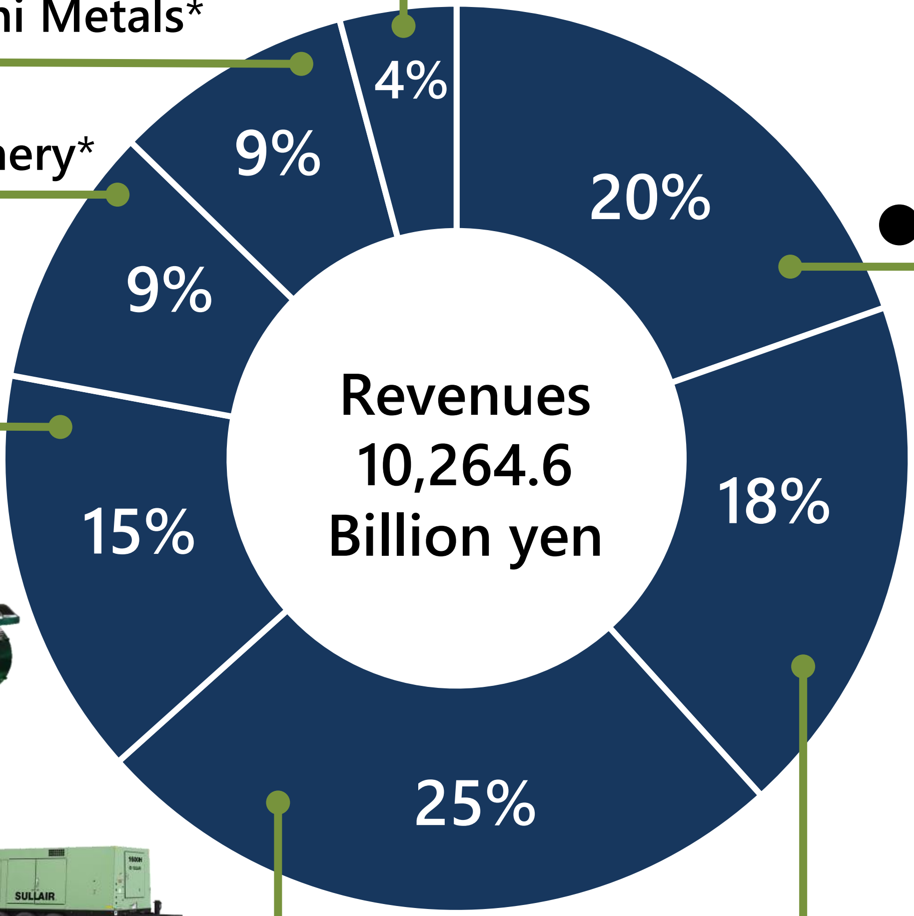
HITACHI Inspire the Next

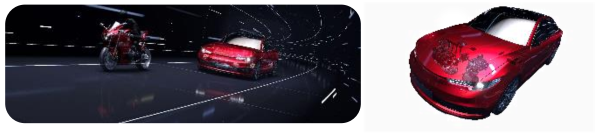
Corporate Name	Hitachi, Ltd.
Founded	1910
Headquarters	6-6, Marunouchi 1-chome, Chiyoda-ku, Tokyo 100-8280, Japan
Revenues	10,264.6 billion yen (FY2021*1)
Adjusted operating income	738.2 billion yen (FY2021*1)
EBIT (Earnings before interest and taxes)	850.9 billion yen (FY2021*1)
Net income attributable to Hitachi, Ltd. stockholders	583.4 billion yen (FY2021*1)
Number of consolidated employees	368,247 (As of end of FY2021*1)

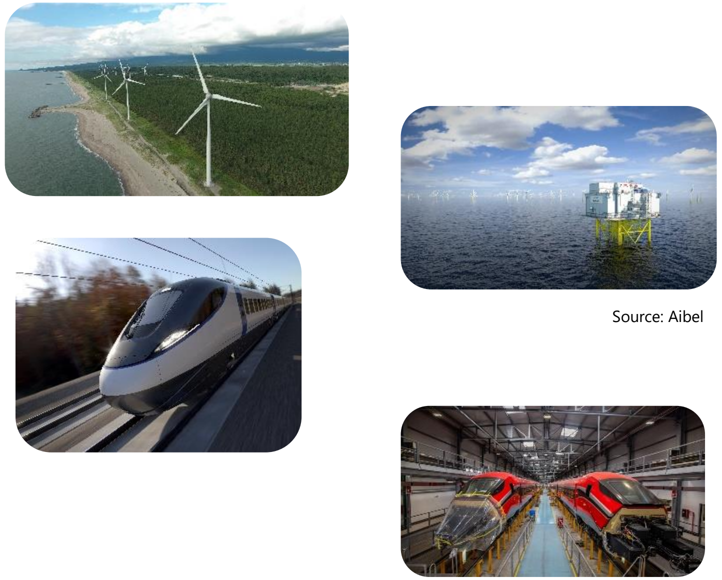
*1: Based on the financial results for FY2020 ended in March 2021

1-2. Business segment constitution (FY2021)

HITACHI
Inspire the Next





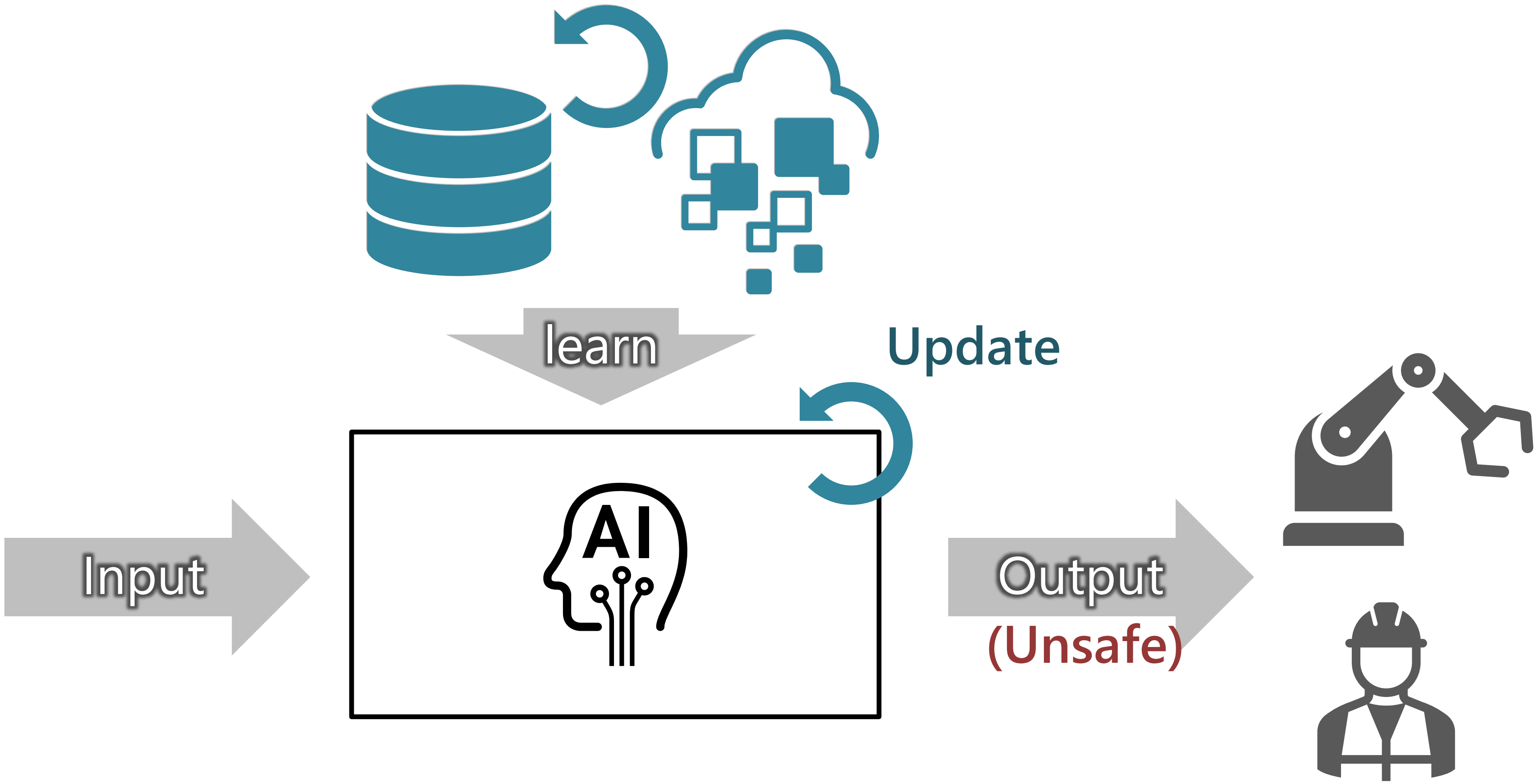


The figures are based on the new segment classifications effective from FY2022.

* Hitachi Construction Machinery was deconsolidated on August 23, 2022. Hitachi Metals are scheduled to be deconsolidated in FY2022.

1-3. Challenges for applying AI

AI is a powerful tool. However, there are existing risks, and it cannot be used as it is.



1-4. Existing countermeasure methods (architecture)

HITACHI
Inspire the Next

Architecture requirements: safe protection, no loss of efficiency, support for evolution

	Architecture Type (Classification)	Subtype	method	Pros	Cons
A	Making AI system more reliable	AI component protection	・Guardrail [1]	・Protection of unintended output is possible	・Specific output can be suppressed, but unintended outputs are possible
B			・Uncertainty separation [2]	・Accuracy can be improved (Flip uncertain output)	・Zero incorrect output is not possible (Uncertainty of AI itself)
C		AI component redundancy	・AI redundancy (Ensemble [3])	・ (parallel processing)	(
D			・AI diversity (multi-agents) [4]	・Accuracy can be improved (Select good results in parallel processing)	・Zero incorrect output is not possible (Mistake by consultation)
E	Rule-based protection from the outside	Fixed limiter	・Upper and lower limits/limit values [5]	・ Almost reliable protection i	・Performance degradation due to exc
F		Evolutionary support	・MAPE-K [6] ・Dynamic risk assessment, etc. [7]	・ i ・The limiter also evolve according to the situation.	・Di evo

Good to use it to improve accuracy (high reliability).

Essentially, impossible to guarantee the output of AI itself => Insufficient for "OT x AI" safety measures

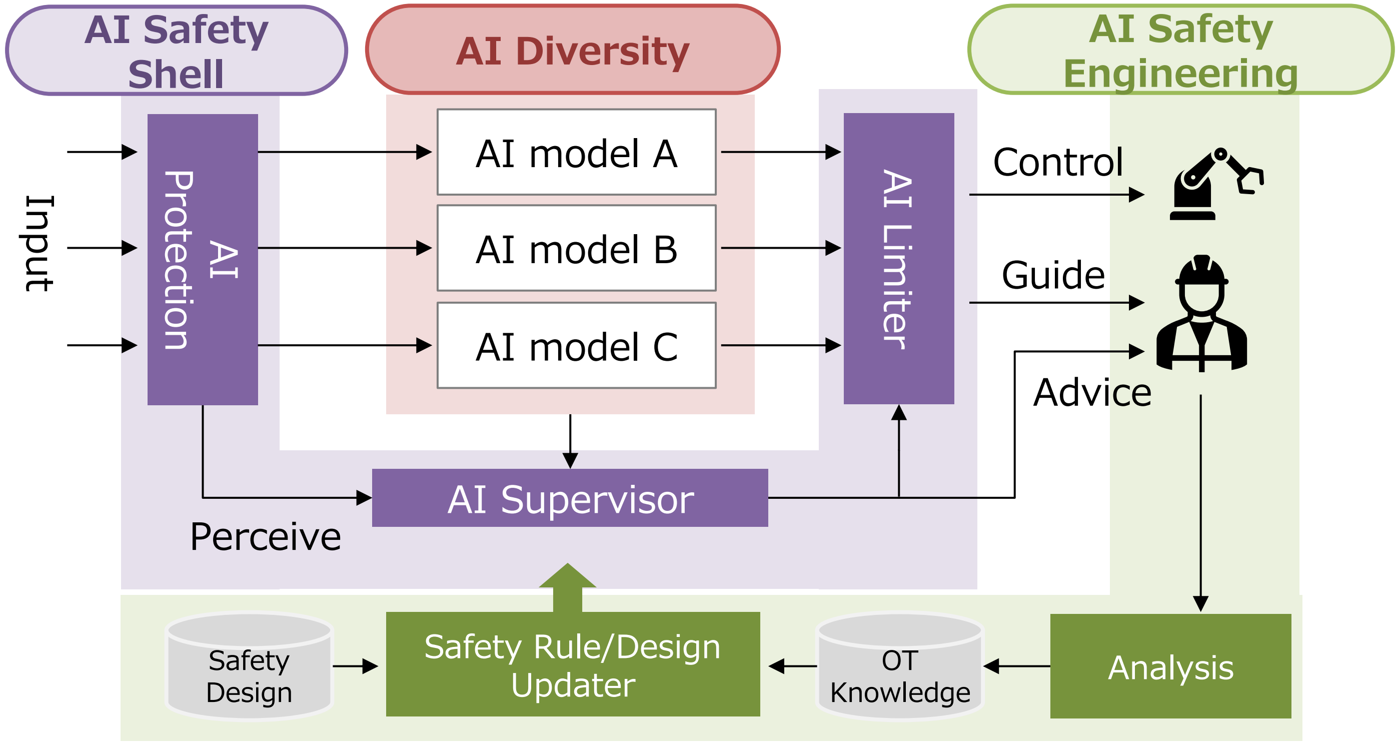
Safety assurance requires certain protection (rule-based)

The design needs to be able to dynamically change the rules and follow them "correctly" and "efficiently".

1-5. Proposed architecture concept

HITACHI
Inspire the Next

Improving reliability/safety through diversity, AI Safety Shell, and Safety Engineering



①AI Diversity

(Excluded from today's explanation)

- Improvement of reliability/performance

②AISS

- Ensuring safety by making decisions based on rules

③Safety engineering

- Responding to evolving systems

Contents

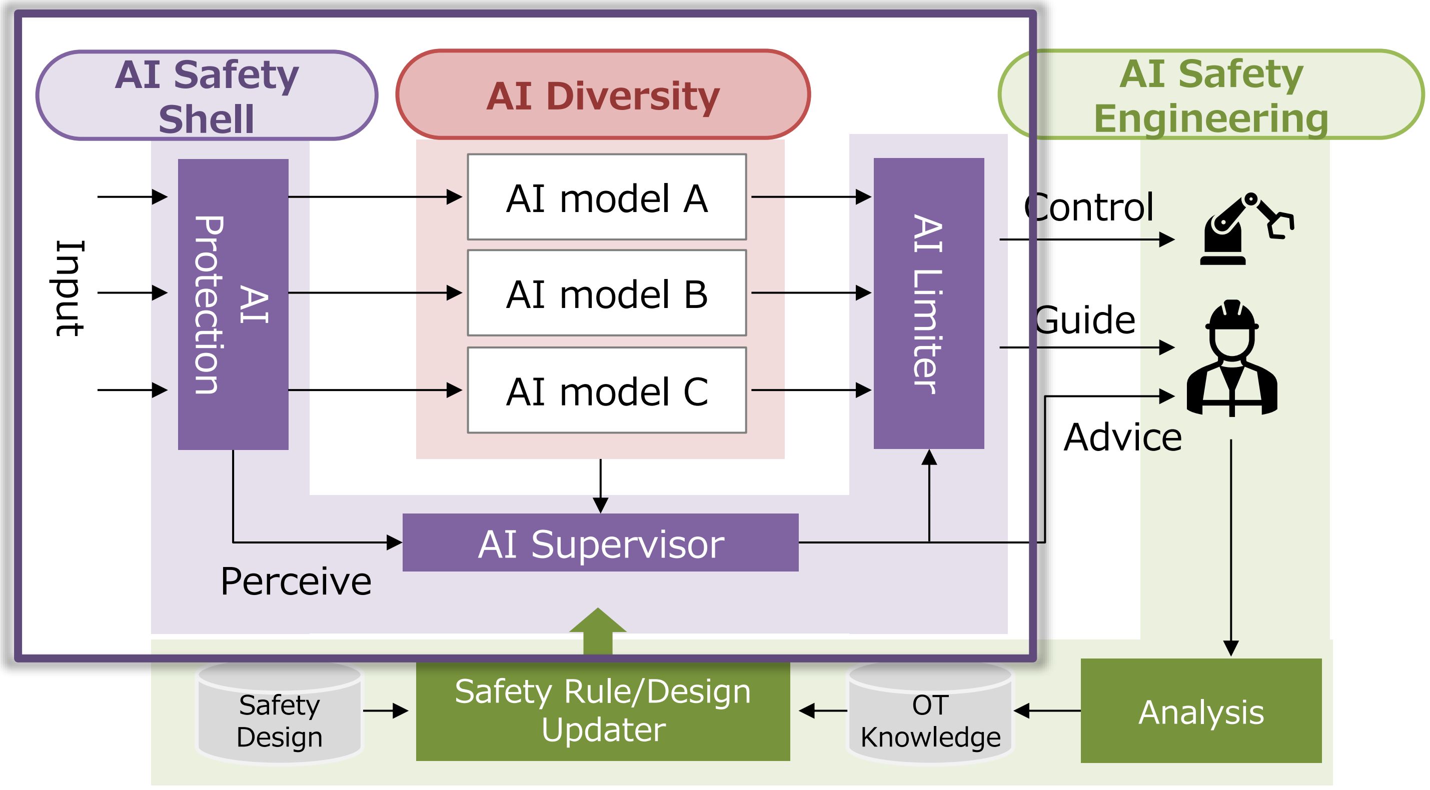
2-1. AI Safety Shell

HITACHI
Inspire the Next

The key is supervision and input/output protection.
The way to protect it differs depending on the application/models/purpose.

■AISS:

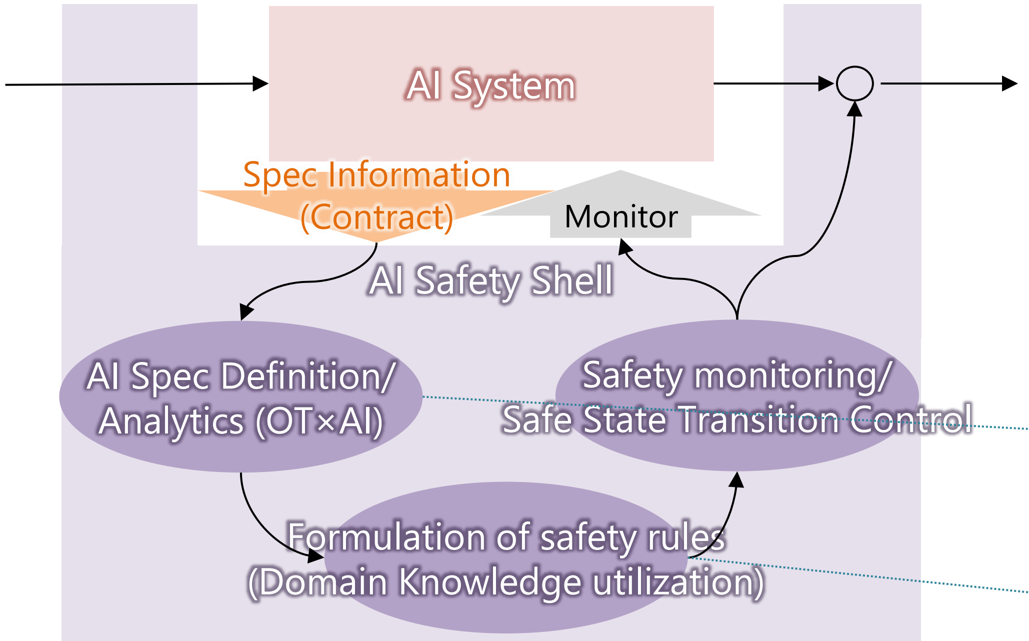
1. **Protection by safety rules**
   (When it is easy to define)
2. **Risk-based protection**
   (for mission-critical systems)



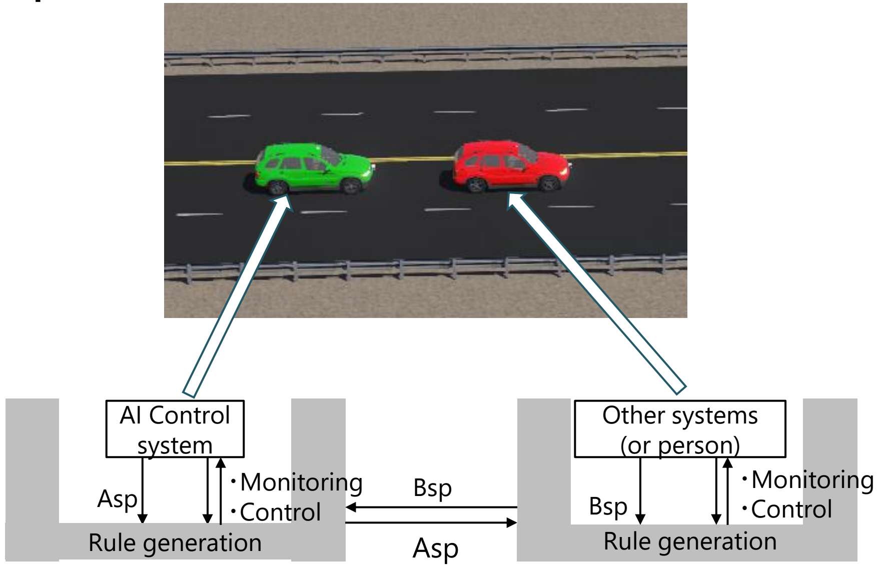
2-2. Safety rule (contract) based safeguard

If the requirements for cooperation can be defined and AI based on safety rules, this approach is possible
The output of AI can be appropriately constrained by contract. Even if it is broken, AISS can protect the system

〇Structure



〇Example



〇AI performance information

$$A_{sp} = (a_{a\_min\_brake}, a_{a\_max\_accel})$$

$$B_{sp} = (a_{b\_max\_brake})$$

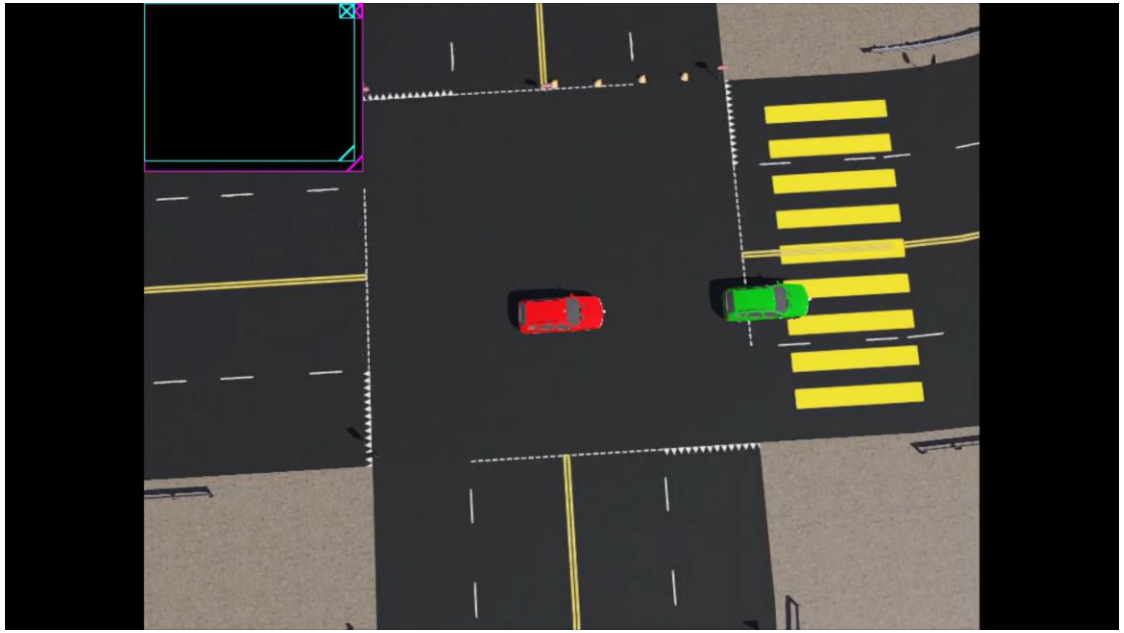
〇 Safety rules

$$d_{min} = \left[v_a\rho + \frac{1}{2} a_{a\_max\_accel}\rho^2 + \frac{(v_a + \rho a_{a\_max\_accel})^2}{2a_{a\_min\_brake}} - \frac{v_b^2}{2a_{b\_max\_brake}} \right]$$

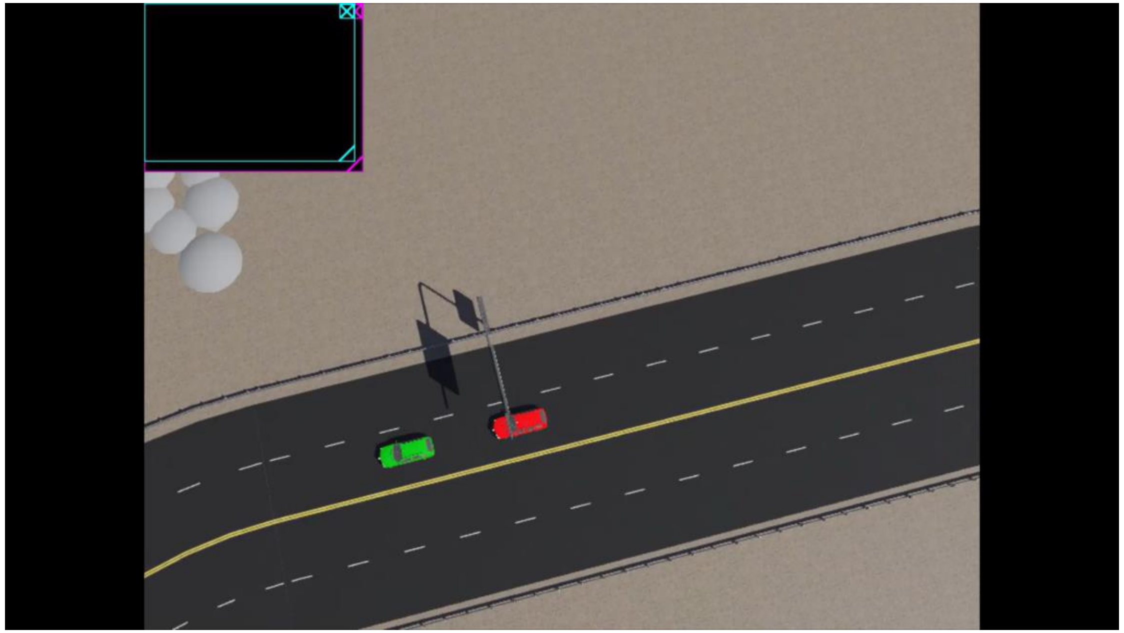
2-3. Experimental result

Even if the AI breaks the rules, it can be controlled safely
When updating, just update the rules

AISS off



AISS on

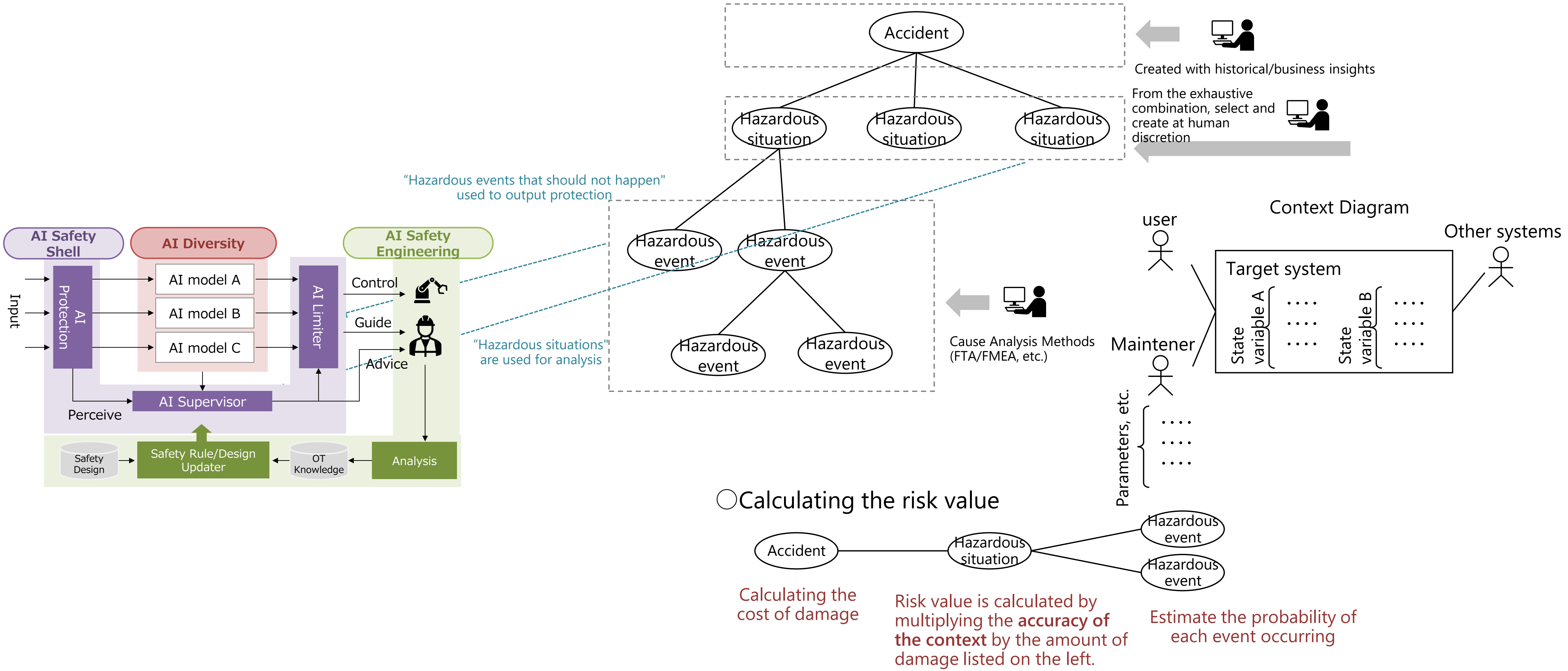


*While it's easy to notice and stop when a vehicle suddenly accelerates, it can be difficult to notice when a vehicle sometimes doesn't slow down, and it can be too late.

2-4. Risk-based guardrails

HITACHI
Inspire the Next

Protect outputs according to risk (for mission-critical systems)
The key is context. Protect outputs by correctly understanding the current situation.



2-5. Prototyping (Risk-based guardrails)

*Presentation Only

Contents

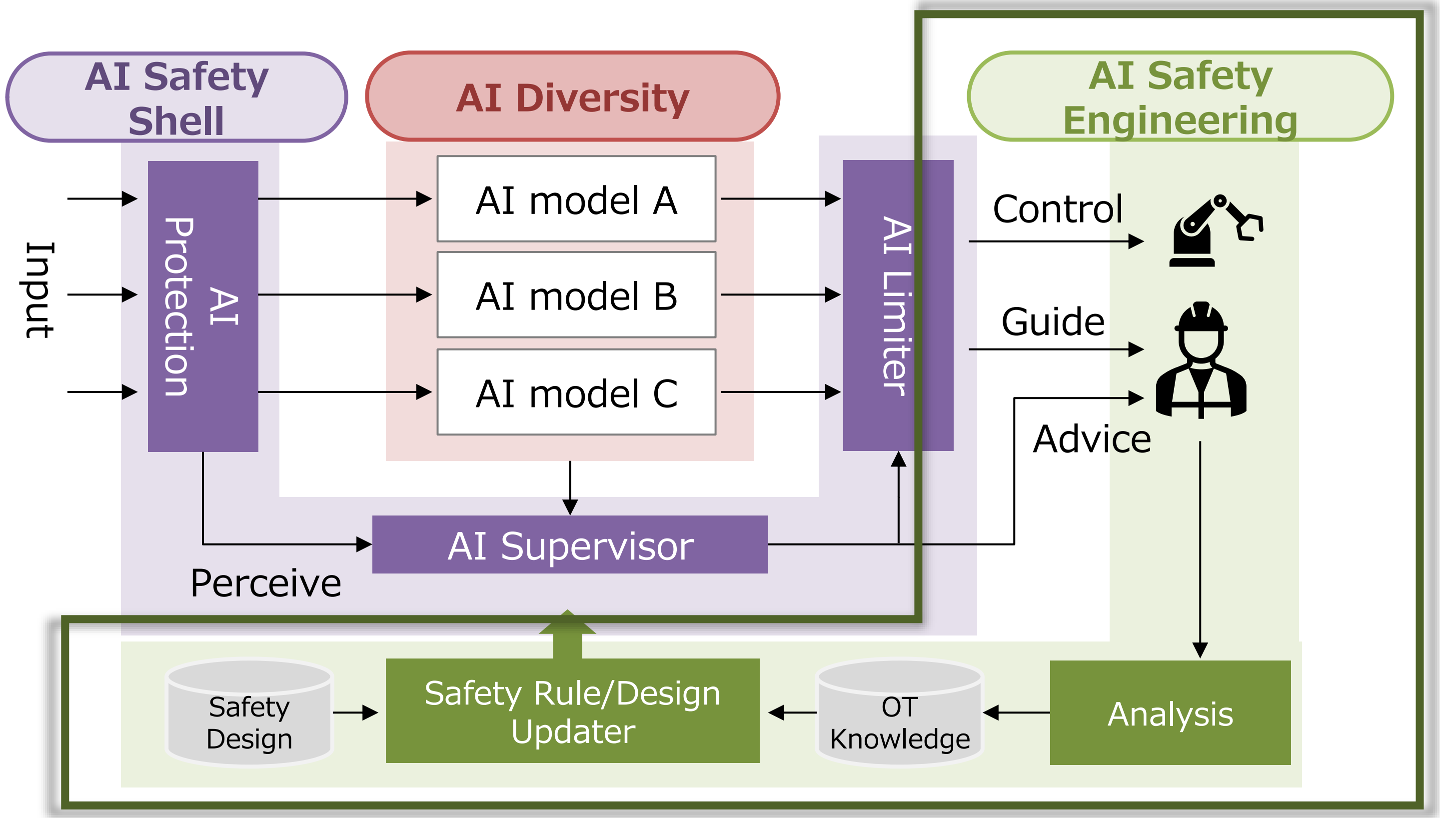
3-1. AI Safety Engineering

HITACHI
Inspire the Next

AI evolves through data-driven methods. Corresponding safety engineering methods are required

■Safety engineering

- Dynamic risk assessment in response to evolution
- Methods for updating safety rules
- etc.

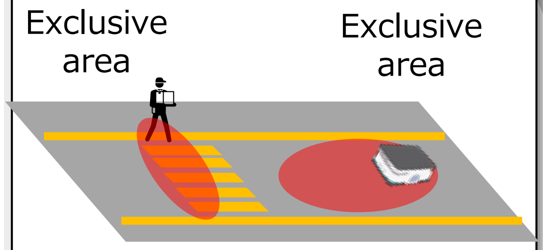


3-2. Safety engineering method overview

HITACHI
Inspire the Next

Evolve the system to adapt human-evolution (co-evolutional control) with guaranteed safety

Symbiotic safety (basis) [8]

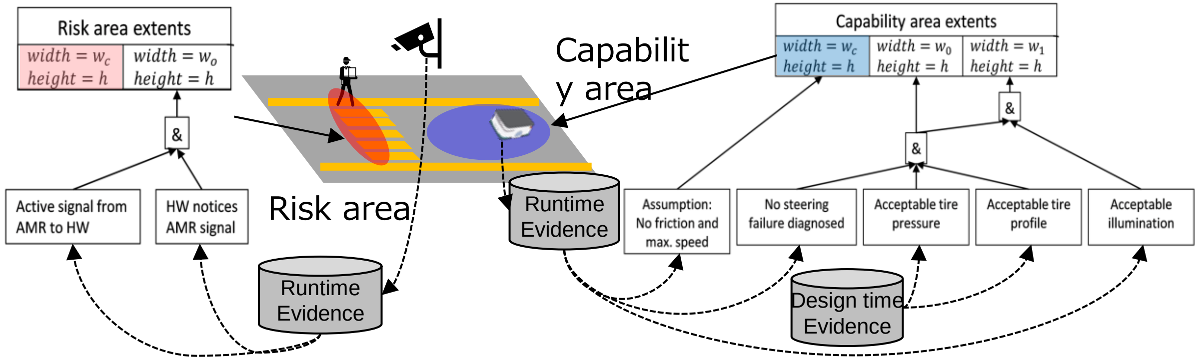


Area exclusive control

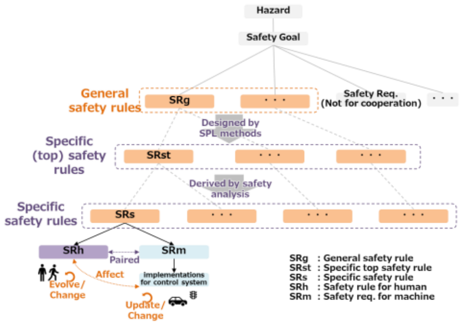
Fraunhofer IKS

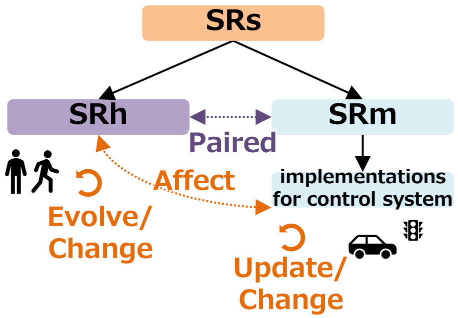
Dynamic risk and capability assessment [9]

Fraunhofer IESE



Paired safety rule structure [10]





Shorten the safety design time to apply the system to other use-cases or domains

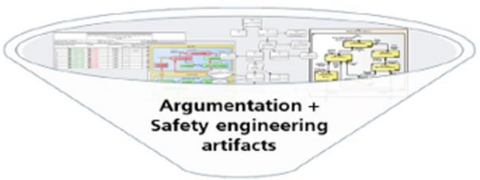
Environment metamodel [11]

Fraunhofer IESE

Safety engineering artifact

Reusable

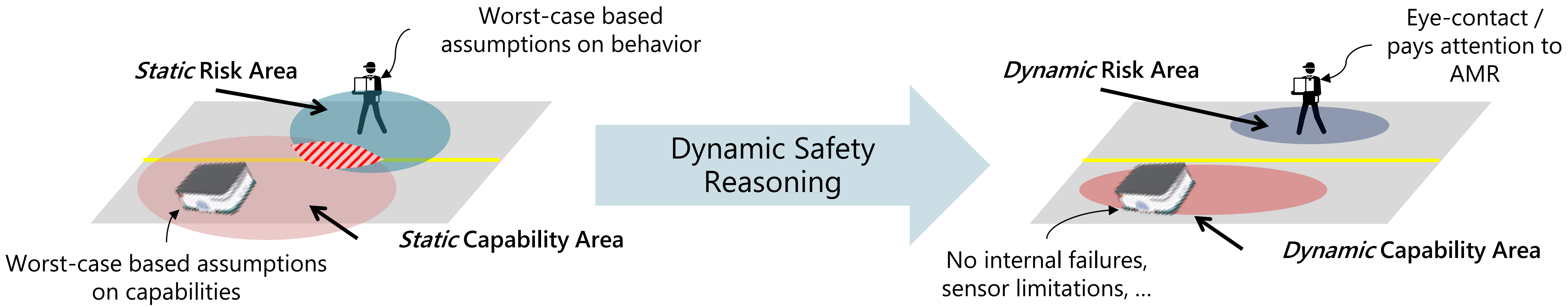
New design



DDI

3-3. Dynamic risk and capability assessment

Dynamic Risk and Capability Areas Can Enhance Efficiency Compared to Using Worst-Case Assumptions

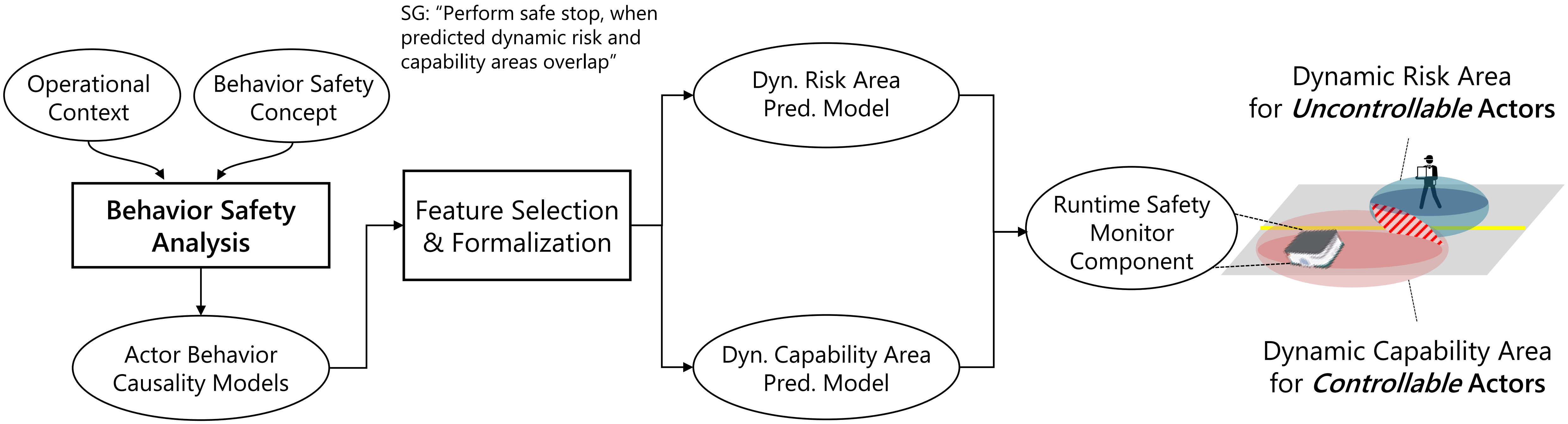


- Replace static worst-case assumptions with dynamic safety reasoning capabilities
- Detecting low-risk situations and reducing risk/capability area extents
- Requires models that allow a dynamic assessment of:
  - risk of operational situation for human worker
  - safety-related capabilities of the AMR

AMR: Autonomous Mobile Robot

3-4. Method overview

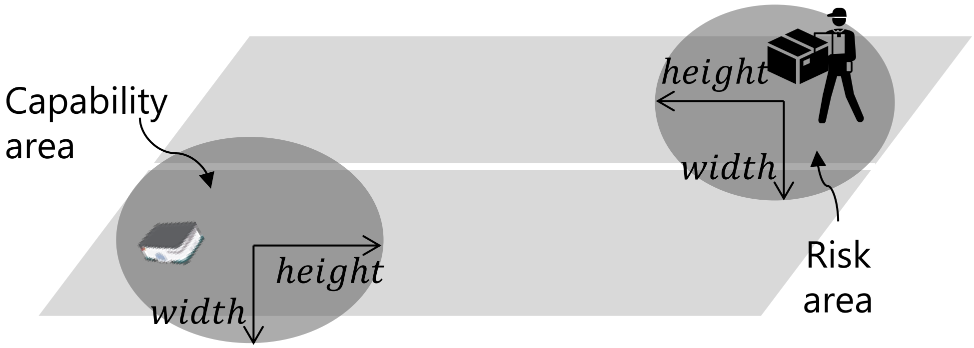
Engineering Runtime Safety Monitors Requires a Systematic Method for Dynamic Model Creation

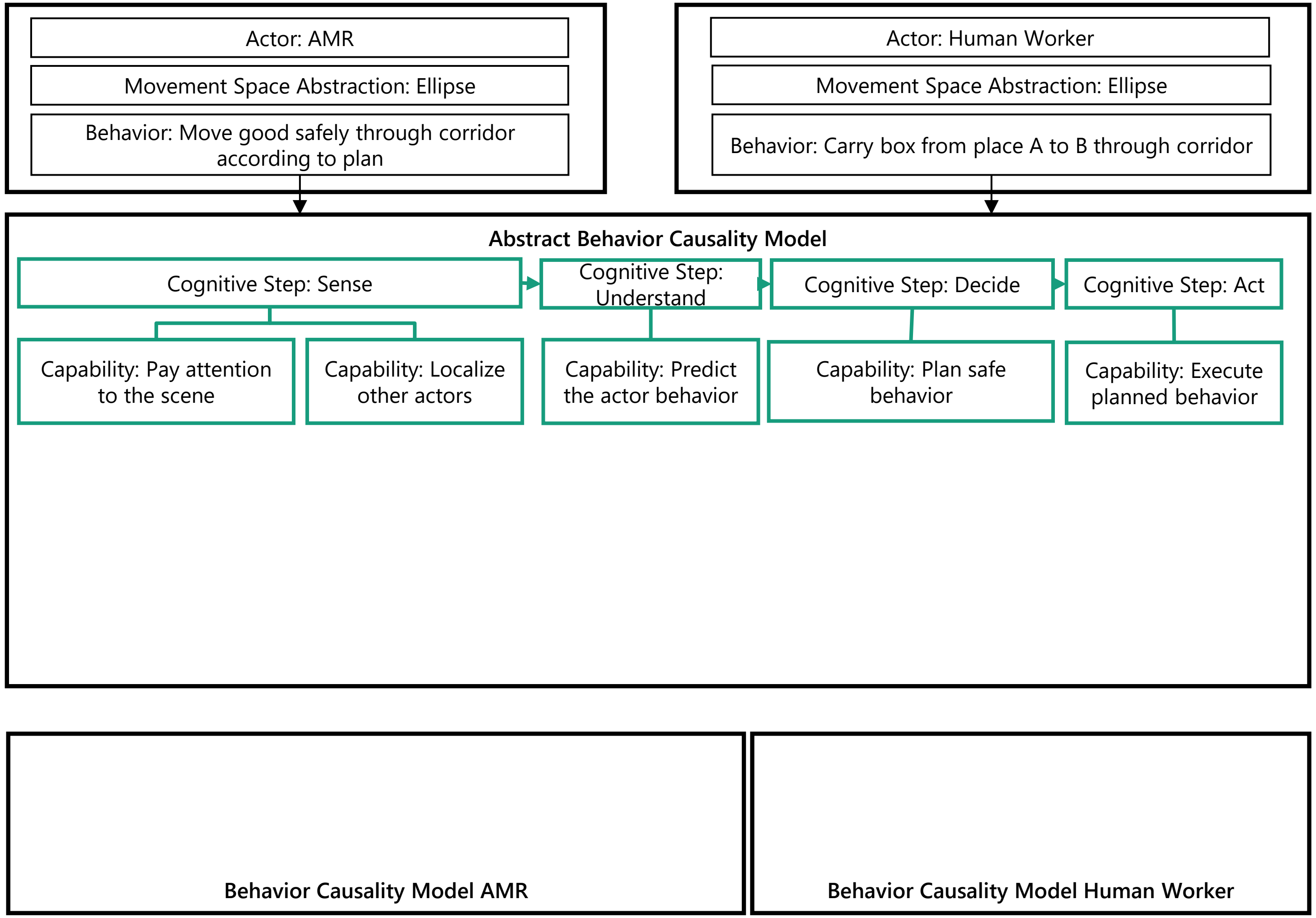


3-5. Method overview

HITACHI
Inspire the Next

Movement Space is Modelled based on SUDA and its Capability Decomposition



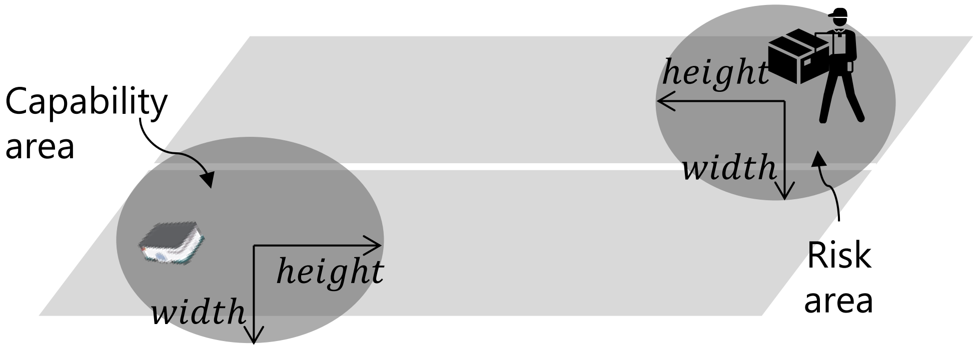


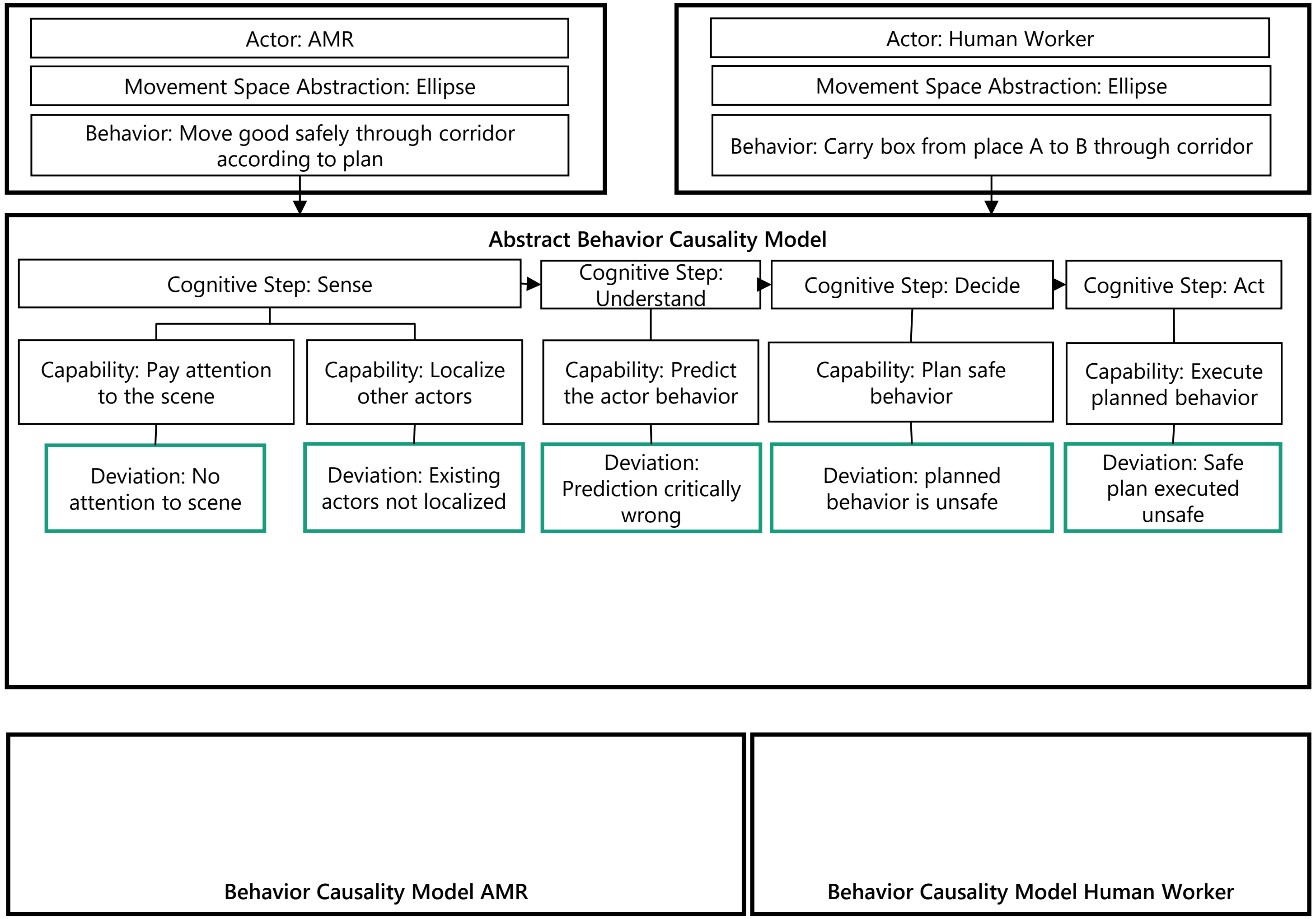
SUDA: Sense-Understand-Decide-Act

3-6. Method overview

HITACHI
Inspire the Next

Potential Capability Deviations impact the likelihood of Behaviors

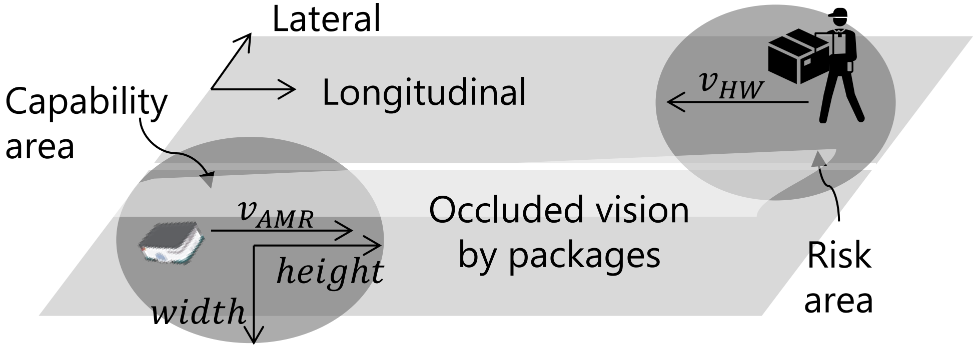


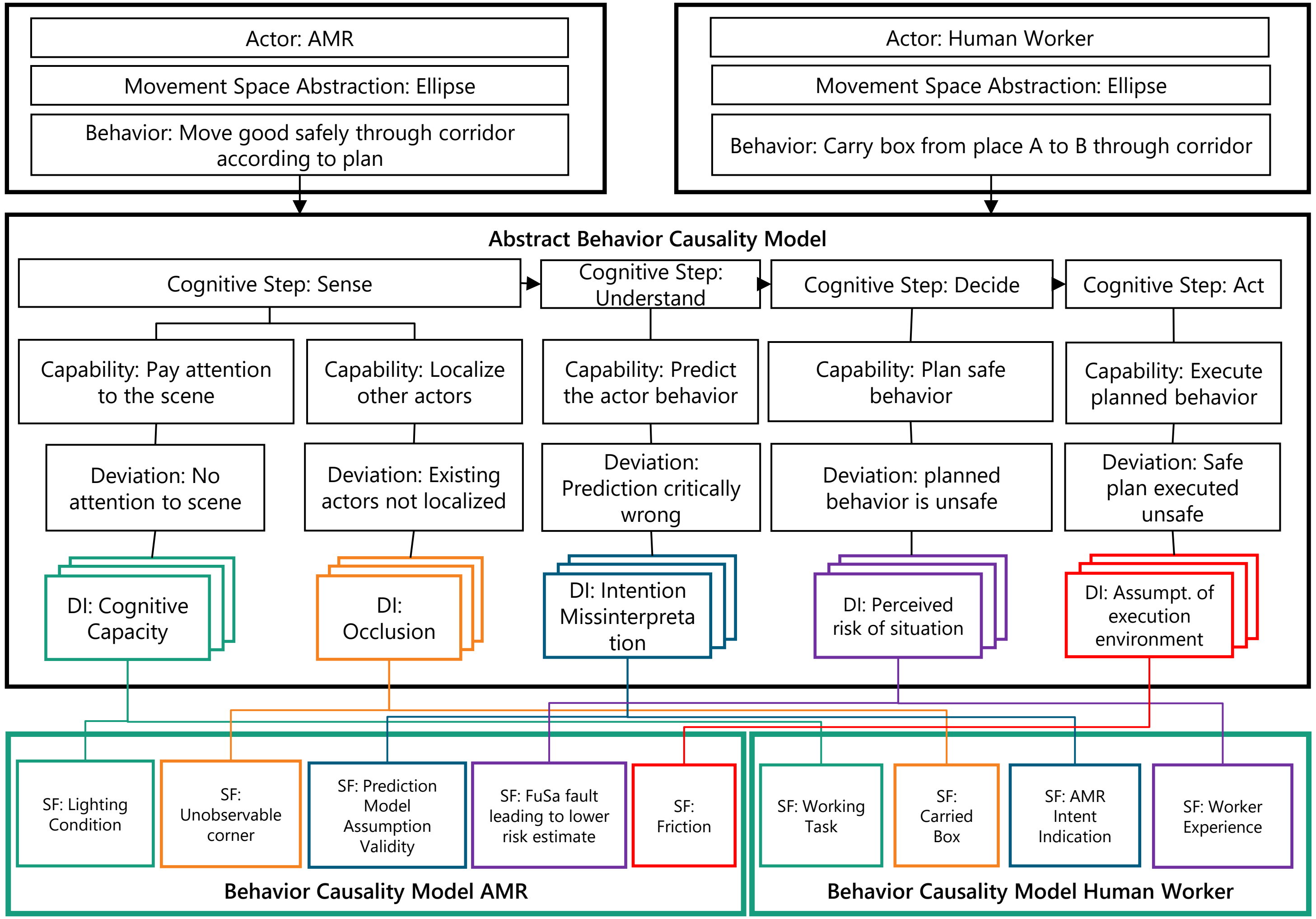


3-7. Method overview

HITACHI
Inspire the Next

Situation Features concretize abstract Deviations for particular Actors

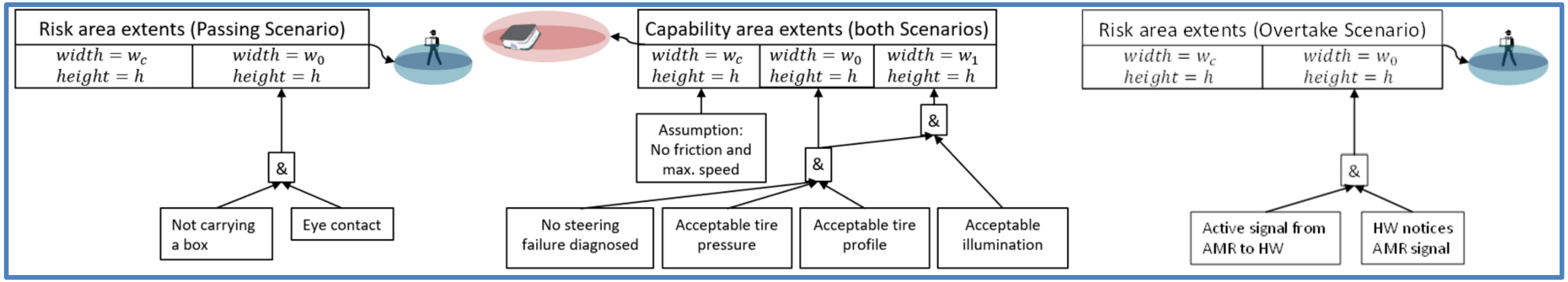




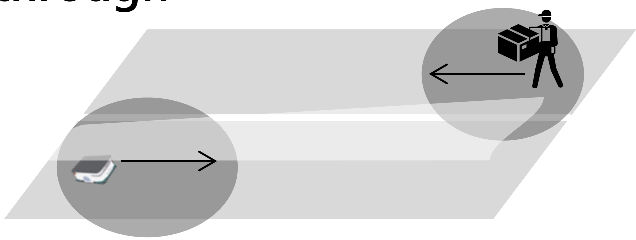
DI: Deviation Influence, SF: Situation features

3-7. Evaluation result

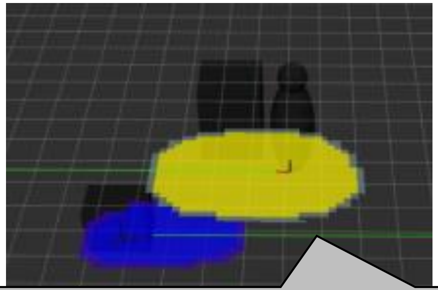
Integration of Dynamic Risk and Capability Models by ConSert



Passing through

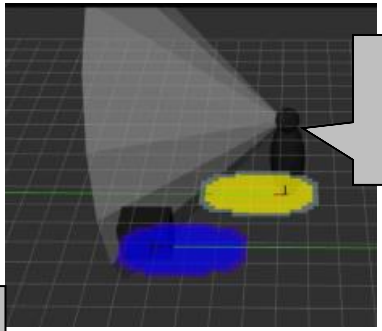


Carrying a box



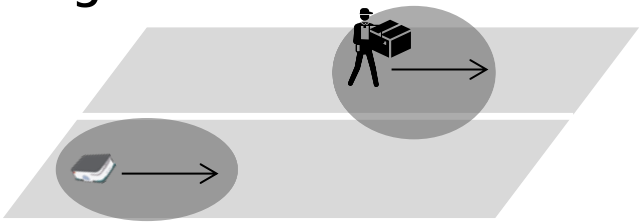
If the worker is carrying a box, risk area is set to large
(he cannot see the AMR)

Not carrying a box & eye contact

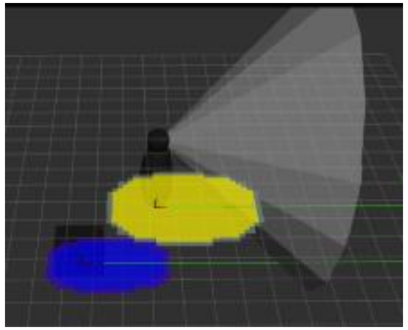


Once the system detects worker's eye contact,
Risk area is set to small

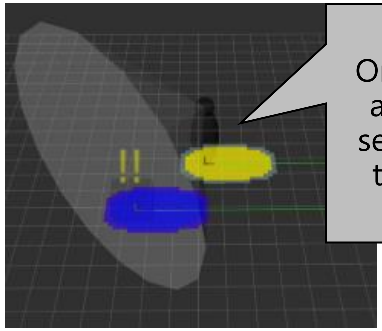
Overtaking



If HW is not aware of AMR,
AMR doesn't overtake



If HW is aware of AMR,
AMR overtakes the HW



Once the system detects awareness, Risk area is set to small and thereby the AMR can overtake

Contents

4-1. AI x Symbiotic safety

HITACHI
Inspire the Next

4-2. Collaboration result with Fraunhofer

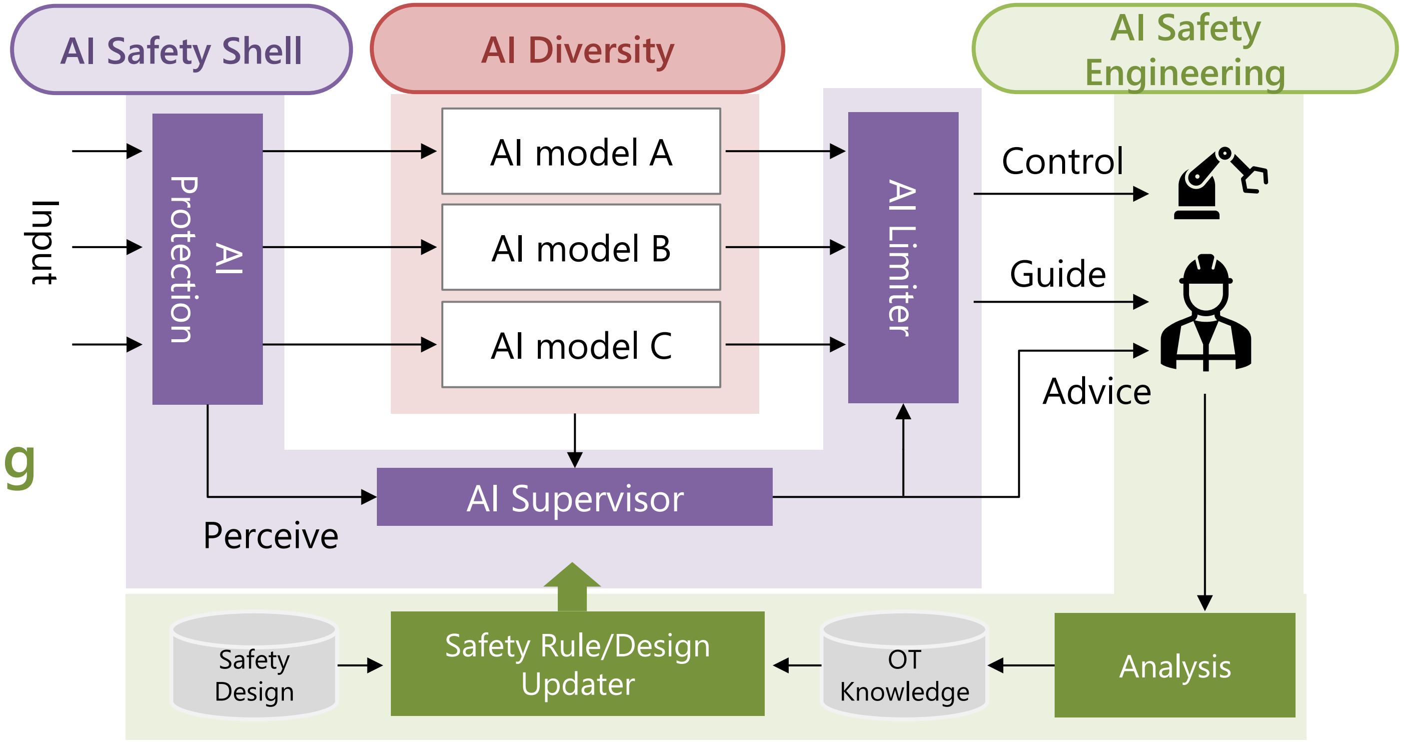
HITACHI
Inspire the Next

These contents are still being updated through the collaboration with Fraunhofer

■Specific implementation for AISS

- Definition of AI risk
- Protection/Supervisor/Limiter

■Updated version of Safety Engineering (by AI/ for AI)



Contents

5-1. Conclusion

- **Introduction of Hitachi's AI Safety activities**
  - Hitachi try to apply AI to Mission-critical systems (in the OT field)

- **Implementation of the AISS concept/architecture**
  - Two approaches: rule-based, risk-based assessment

- **Safety engineering process** that can respond to dynamic changes

- **Further activities**
  - Evaluation of implementation in control systems
  - Collaboration with Fraunhofer

5-2. References

[1] Dong, Yi, et al. "Building Guardrails for Large Language Models." arXiv preprint arXiv:2402.01822 (2024).
[2] Kläs, Michael, and Lena Sembach. "Uncertainty wrappers for data-driven models: increase the transparency of ai/ml-based models through enrichment with dependable situation-aware uncertainty estimates." Computer Safety, Reliability, and Security: SAFECOMP 2019 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Turku, Finland, September 10, 2019, Proceedings 38. Springer International Publishing, 2019.
[3] Petrakova, Aleksandra, Michael Affenzeller, and Galina Merkurjeva. "Heterogeneous versus homogeneous machine learning ensembles." Information Technology and Management Science 18.1 (2015): 135-140.
[4] https://note.com/fladdict/n/n106b9ce8f7d4
[5] Workgroup, E. G. A. S. "Standardized E-gas monitoring concept for gasoline and diesel engine control units." Version 5 (2013): 38.
[6] Weiss, Gereon, et al. "Towards integrating undependable self-adaptive systems in safety-critical environments." Proceedings of the 13th International Conference on Software Engineering for Adaptive and Self-Managing Systems. 2018.
[7] Reich, Jan, and Mario Trapp. "SINADRA: towards a framework for assurable situation-aware dynamic risk assessment of autonomous vehicles." 2020 16th European dependable computing conference (EDCC). IEEE, 2020.
[8] Ishigooka, T., et. al., "Symbiotic Safety: Safe and Efficient Human-Machine Collaboration by utilizing Rules," Design, Automation & Test in Europe Conference & Exhibition (DATE), 2022.
[9] Reich, J., et. al., "Engineering Dynamic Risk and Capability Models to Improve Cooperation Efficiency Between Human Workers and Autonomous Mobile Robots in Shared Spaces," Model-Based Safety and Assessment, pp.237-251, 2022.
[10] Otsuka, S., et. al., "Paired Safety Rule Structure for Human-Machine Cooperation with Feature Update and Evolution," Computer Safety, Reliability, and Security SAFECOMP 2023 Workshops, pp.247-259, 2023.
[11] Reich, J., et. al., "Concept and metamodel to support cross-domain safety analysis for ODD expansion of autonomous systems," Computer Safety, Reliability, and Security, pp. 165-178, 2023.

Hitachi Social Innovation is
POWERING GOOD