Trustworthy AI for Industrial Applications

AI4EU Workshop 13th November 2020
“Trustworthy AI made in Europe: From Principles to Practices”

Sonja Zillner, Siemens AG
Claus Bahlmann, Andreas Hapfelmeier, Daniel Hein

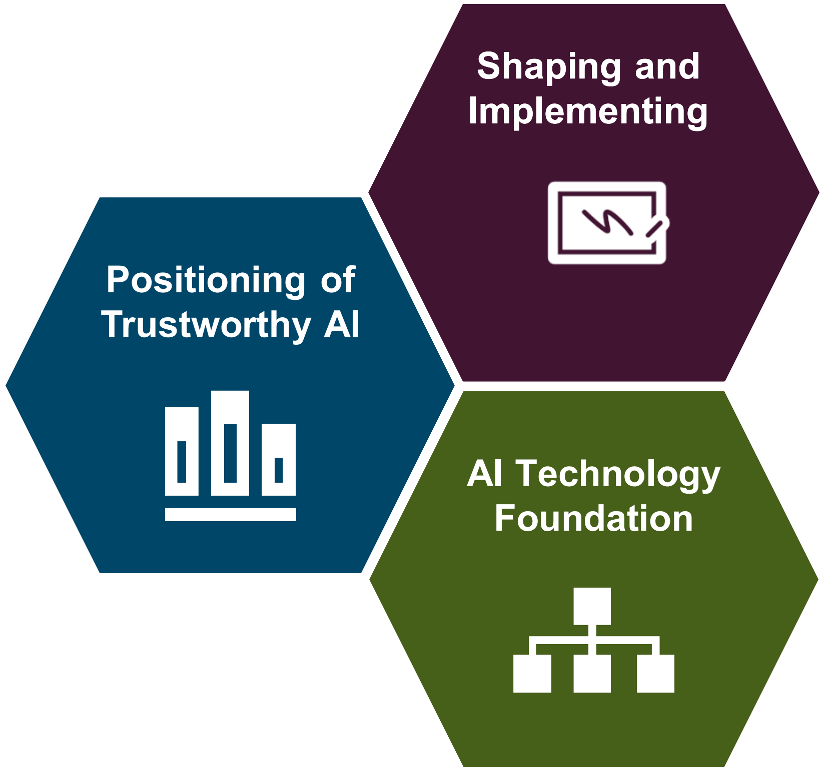
Trustworthy AI @ Siemens

Implementing Trustworthy AI in Industrial Applications

SIEMENS
Ingenuity for life

Challenges

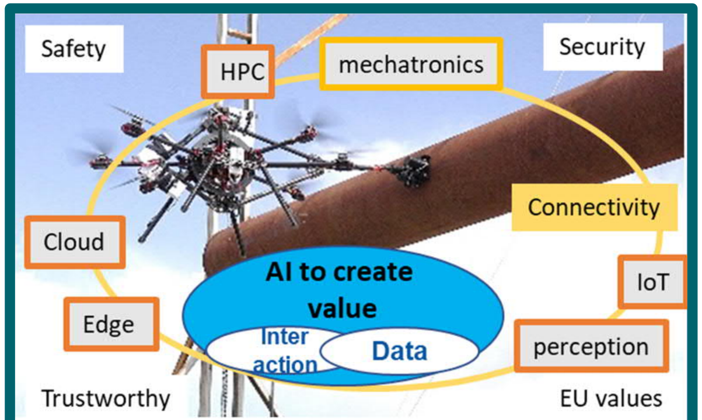
- Identification of **regulation / conformity assessments** that
  - focus on AI application that have *risk* involved and require trustworthy consideration
  - are complex enough to reflect the *dynamic nature of AI systems*
  - are simple enough to *limit costs* for conformity assessments
- Identification of **AI technologies** addressing **requirements for trustworthiness**, such as robustness, transparency, privacy, etc.



Join forces on multiple levels to shape and implement Trustworthy AI

Characteristics of Industrial AI?

Characteristics of Industrial AI Applications



Maintenance and Inspection
Huge demand in public and private: Energy, Transport, (process) industry, ...

Value creation
Use of (semi-)autonomous technology, partly in hazardous or remote locations

Impact
Safety, environmental, and economical

1. Context-dependent machine data
2. Strong focus is on optimization of machines or processes (incl. automation)
3. Degree of human / environment interaction is part of the design
4. Contractual agreement between B2B Partners
5. Safety, reliability, security, privacy .. requirements / legislation are already in place

Trustworthy AI Requirements

“Ethics Guidelines for Trustworthy AI” by the HLEG on AI

SIEMENS
Ingenuity for life

1 Human agency and oversight
Including fundamental rights, human agency and human oversight

2 Technical robustness and safety
Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility

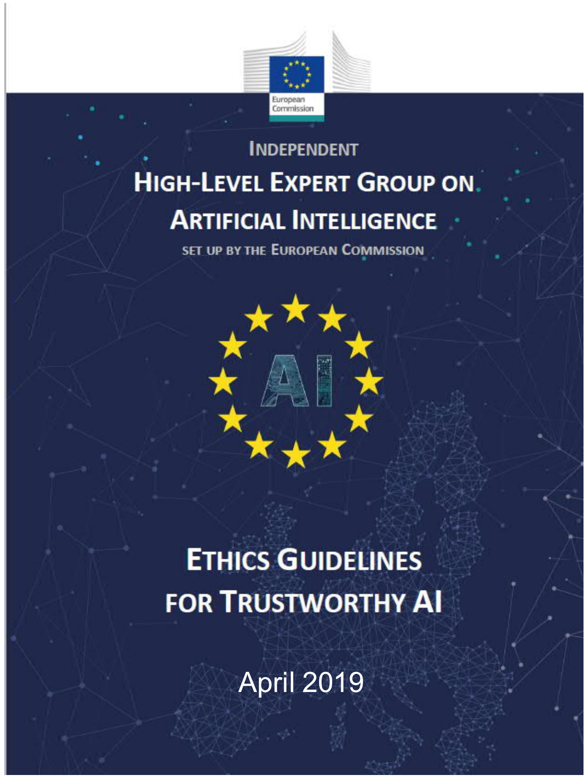
3 Privacy and data governance
Including respect for privacy, quality and integrity of data, and access to data

4 Transparency
Including traceability, explainability and communication

5 Diversity, non-discrimination and fairness
Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation

6 Societal and environmental wellbeing
Including sustainability and environmental friendliness, social impact, society and democracy

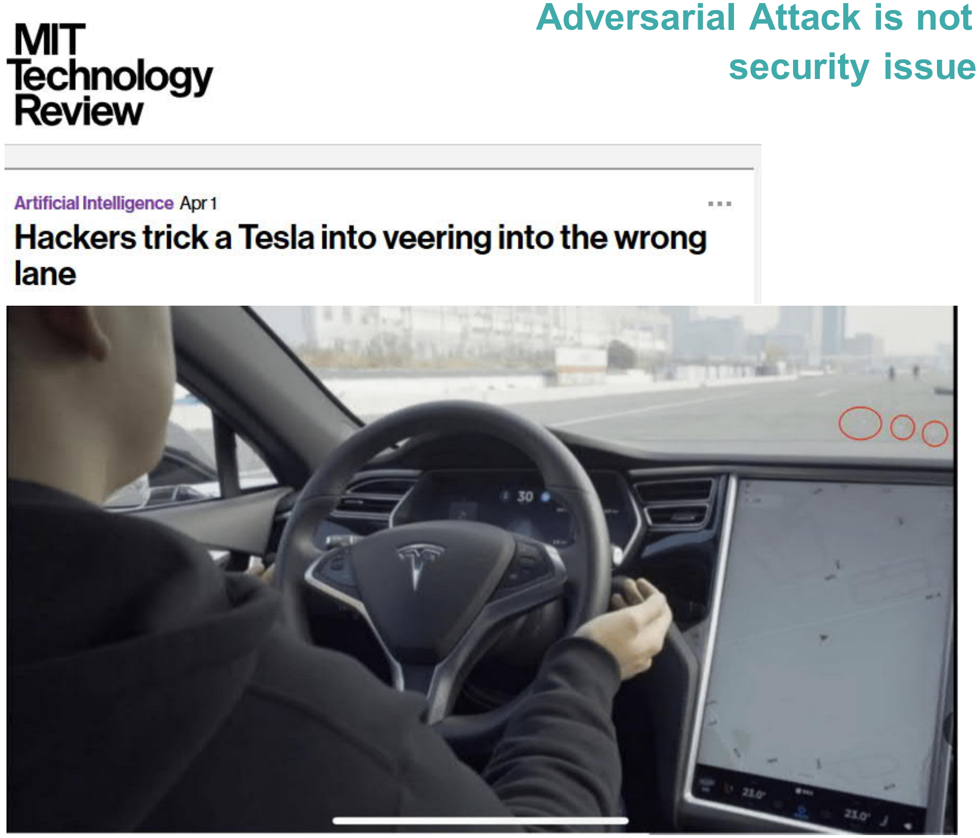
7 Accountability
Including auditability, minimisation and reporting of negative impact, trade-offs and redress



2. Technical robustness and safety

Including **resilience to attack and security, fall back plan and general safety**, accuracy, reliability and reproducibility

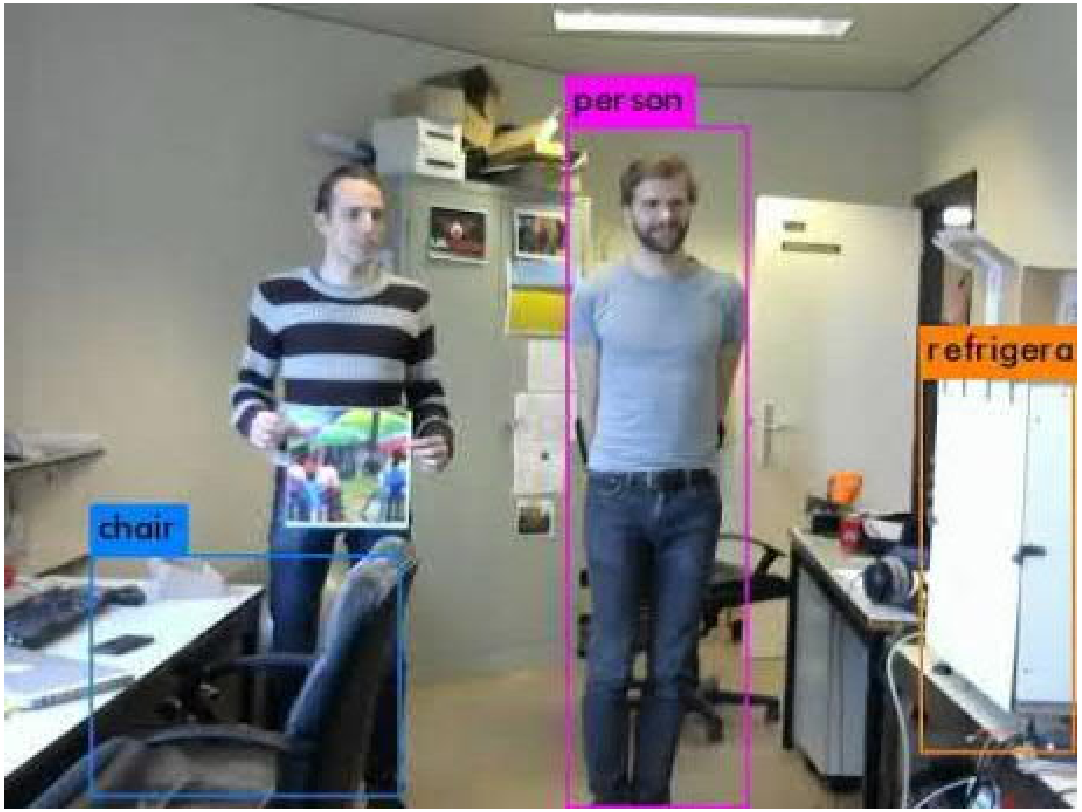
Adversarial Attack is not a classical security issue


MIT Technology Review

Artificial Intelligence Apr 1

Hackers trick a Tesla into veering into the wrong lane

Source: https://www.technologyreview.com/f/613254/hackers-trick-teslas-autopilot-into-veering-towards-oncoming-traffic/



https://youtu.be/MIbFvK2S9g8

2. Technical robustness and safety

Including resilience to attack and security, fall back plan and general safety,
accuracy, reliability and reproducibility

SIEMENS
Ingenuity for life

Accuracy: The model should be as good as necessary

Reliability: Works properly with a range of inputs and in a range of situations

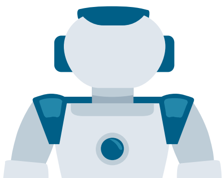
Reproducibility: exhibits the same behavior when repeated under the same conditions

Where is Trustworthy AI [Technical Robustness and Safety] **needed?**

Which Industrial AI applications have significant trustworthy implications?

SIEMENS
Ingenuity for life

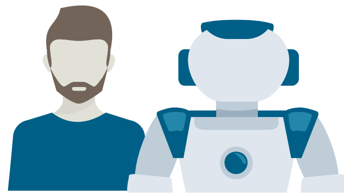
Distinguish between AI applications that are solely technical versus those that involve human interaction



Non-Human Interaction

AI is used to improve machine performance

≠



Human Interaction

AI is used to augment human decision making by learning from its interaction with humans / environment

Trustworthiness should be considered in all Industrial AI Applications.

Industrial AI applications with human interaction require significant trustworthy-related consideration

Non-Human Interaction

SIEMENS

Ingenuity for life

Improved efficiency

Sensing & Connectivity & Learning & Acting

- More than 200 GB of sensor data from ≈ 7.800 wind parks
- Use of Reinforcement Learning
- Early detection of divergent behavior
- 1-3% increase of annual energy harvest

Protection goals

Resilience of the critical infrastructure energy supply ✓

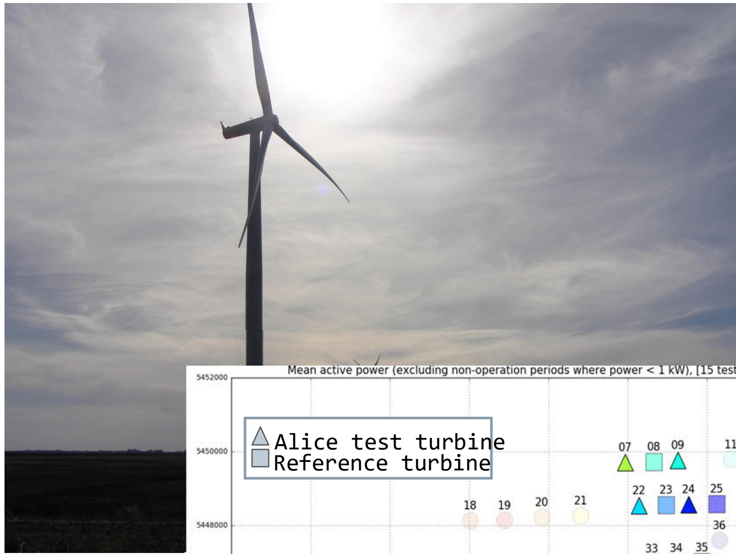
Environmental / climate protection ✓

Common research project ALICE:
Siemens, IdaLab GmbH, TU Berlin

Establishing the basis for Self-declaration

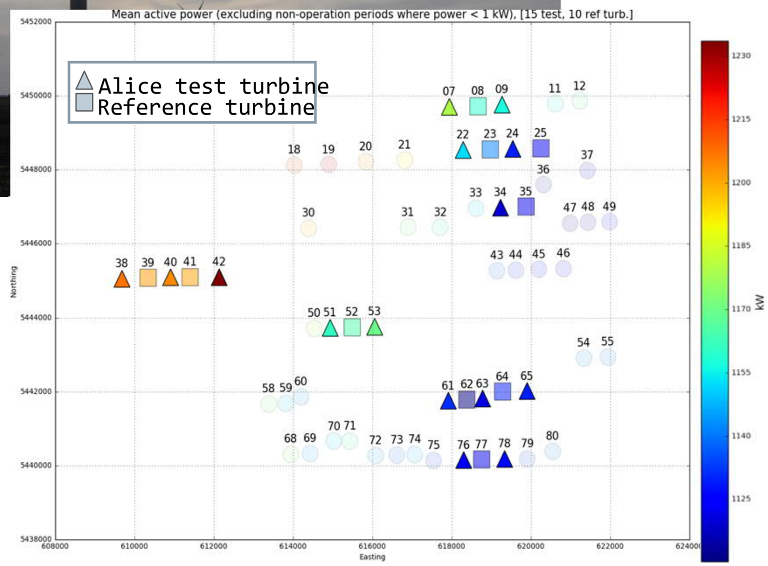
Test Evaluation for Wind farm field

SIEMENS
Ingenuity for life



Generate interpretable policies for several wind turbines in a wind farm in Canada:

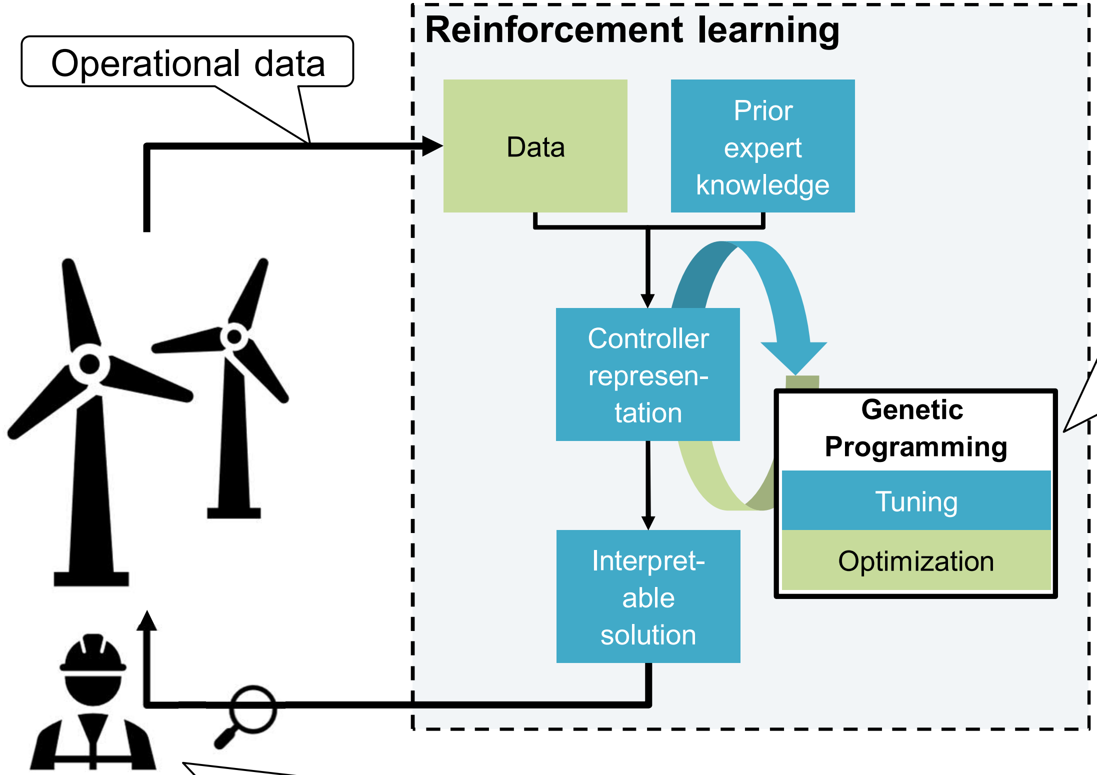
1. Based on previously generated exploration data
2. Domain experts interpret and discuss the learned policies
3. Promising policy candidates are selected for deployment on the wind farm



Funded by the German Federal Ministry of Education and Research within the scope of the autonomous learning in complex environments (ALICE) II project (project number 01IB15001).

Interpretable Policies via
Genetic programming reinforcement learning

SIEMENS
Ingenuity for life

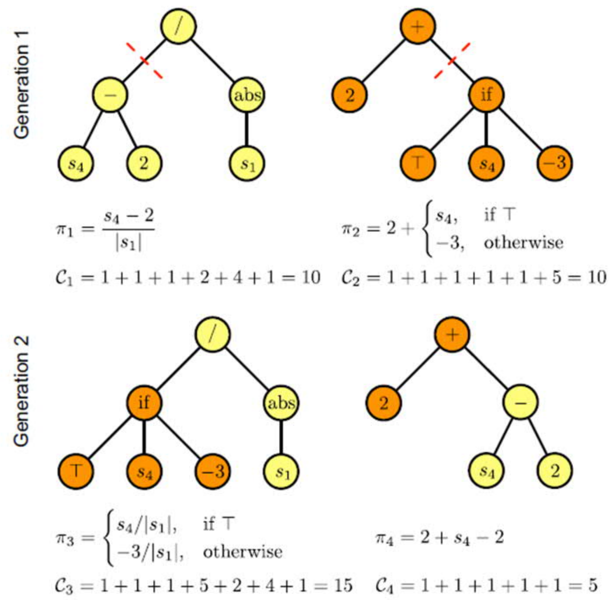


Examine and evaluate solutions (safety, technical plausibility, …)

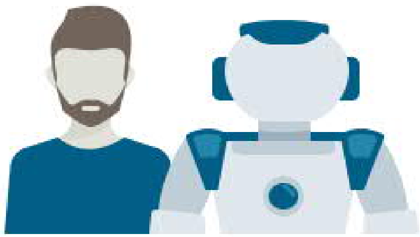
$$\pi(s) = \rho + 0.94\dot{\rho} + 6.98\theta + \dot{\theta}$$

An evolutionary algorithm applies:

- Random initialization
- Crossover & mutation
- Natural selection based on performance



Hein, D., Udluft, S., & Runkler, T. A. (2018). Interpretable policies for reinforcement learning by genetic programming. In: Engineering Applications of Artificial Intelligence 76 (2018), pp. 158-169.



Human Interaction

Protection goal

Human safety !

Safe AI

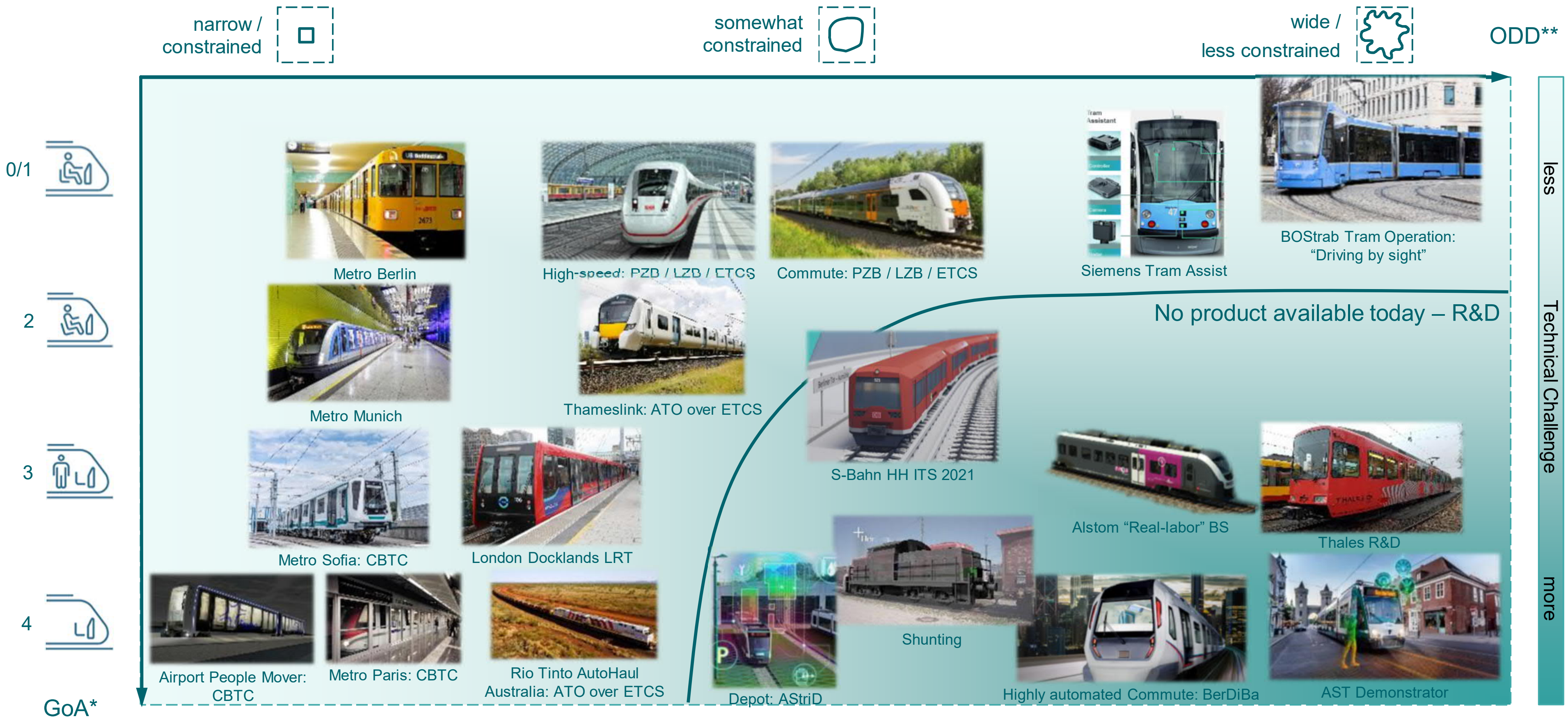
Automated Driving for Rail

First autonomous tram presented (as research project)

Collision **warning systems** and smart assistants help to increase safety and availability

Automated Driving for Rail

SIEMENS
Ingenuity for life



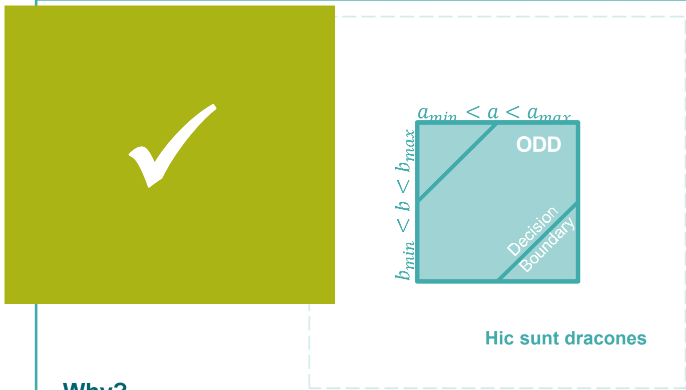
*GoA = Grade of Automation (IEC 62290)

**ODD = Operational Design Domain = Operation conditions under which an autonomous system is specifically designed to function

ODDs for Automated Driving in Rail and their Challenges

SIEMENS
Ingenuity for life

Narrow ODD in Rail



Why?

- Narrow ODD can often be specified and solved with (comparably) simple, technology, allowing for (comparably) straightforward homologation and safety

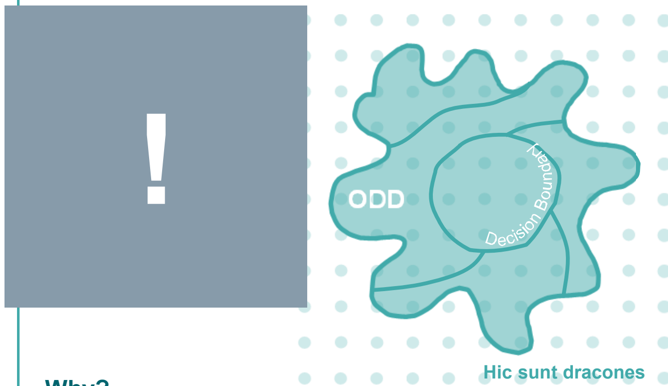
How?

- Based on simple, but effective infrastructure rooted sensors, measures and logic e.g., balises, fences, doors, radar curtains, and ATP systems (PZB, LZB, ETCS, …), with (comparably) simple logic
- **Often close the system to rail traffic, eliminating interaction with cars, people, …**

Challenges

- Sometimes high costs
- **Approaches cannot easily scale to wide ODDs, since the open world is complex**

Wide / unconstrained ODD in Rail



Why?

- Traditional technology is not sufficient, especially when **open world system (i.e., interaction with pedestrian, cars, …) is in scope**

How?

- Wide ODD often cannot be specified by logic & rules → Instead, use data samples
- Learn ODD and state space decision boundaries using AI / ML
- Nevertheless, constrain ODD as much as possible, to allow for safe operation
- Combine with traditional Rail safety technology (e.g., ETCS), where possible

Challenges

- **Technology & homologation ecosystem for safe, AI / ML based highly automated systems not fully developed yet**

ODD = Operational Design Domain = Operation conditions under which an autonomous system is specifically designed to function

Safe AI Challenges

!

SIEMENS
Ingenuity for life

1 – Safe AI Principles & Tools

Provide insight
into AI / NN behavior and
data distribution

2 – Safe AI / MLOps

Engineering environment
for agile & large-scale development & validation

3 – Safe AI System

Safety argumentation and regulatory framework for homologation

Safe AI Challenges

!

SIEMENS
Ingenuity for life

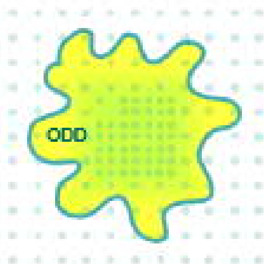
1 – Safe AI Principles & Tools

2 – Safe AI / MLOps

3 – Safe AI System

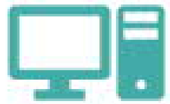
Provide insight
into AI / NN behavior and
data distribution

1 Data distribution

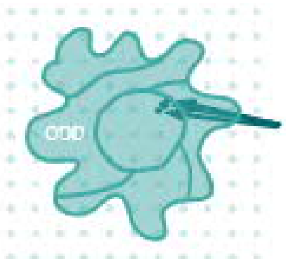


3 Confidence measures

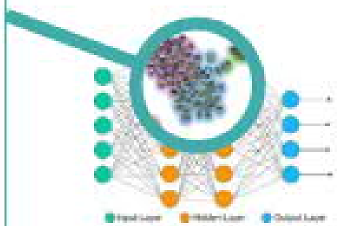
5 Runtime Monitoring Triggers



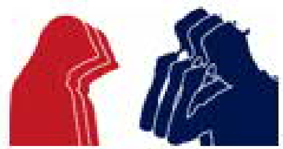
2 Robustness



4 Interpretability



6 Validation & Red Teaming



Safe AI Challenges

!

SIEMENS

Ingenuity for life

1 – Safe AI Principles & Tools

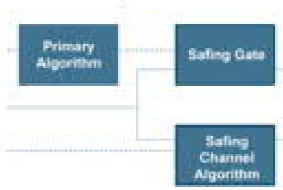
Prov
into
data

2 – Safe AI / MLOps

3 – Safe AI System

Safety argumentation and regulatory framework for homologation

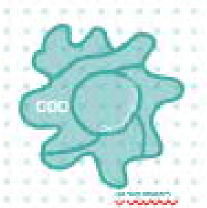
1 Safety Architecture



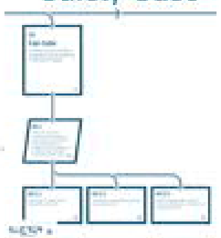
3 Safety Metrics / Performance Indicators

5 White paper & community

2 GoA4 ODD description



4 GoA4 Safety Case



6 Norms & Regulation

IEC ISO CENELEC

Safety Case = A structured written argument, supported by evidence, justifying that a system is acceptably safe for intended use. [Phil Koopman]

Summary

SIEMENS
Ingenuity for life

1. Industrial AI creates new opportunities to bring value to society, economy and environment
2. Industrial AI needs to be trustworthy
3. Any conformity assessment need to be accomplished on application-level and reflect the risk-involved
4. **Additional research in AI** is needed to establish the basis for implementing Trustworthy / Safe AI systems
5. Combine the development of new AI techniques with the development of efficient means for **verification and validation** and align with (established) **regulatory framework**

Thanks for your attention!

Questions?

SIEMENS
Ingenuity for life

Siemens Corporate Technology - Business Analytics and Monitoring

200 Data Scientists & AI experts at 9 locations globally

Sonja Zillner

Lead Trustworthy AI

Siemens AG Technology
Munich, Germany

sonja.zillner@siemens.com

siemens.com