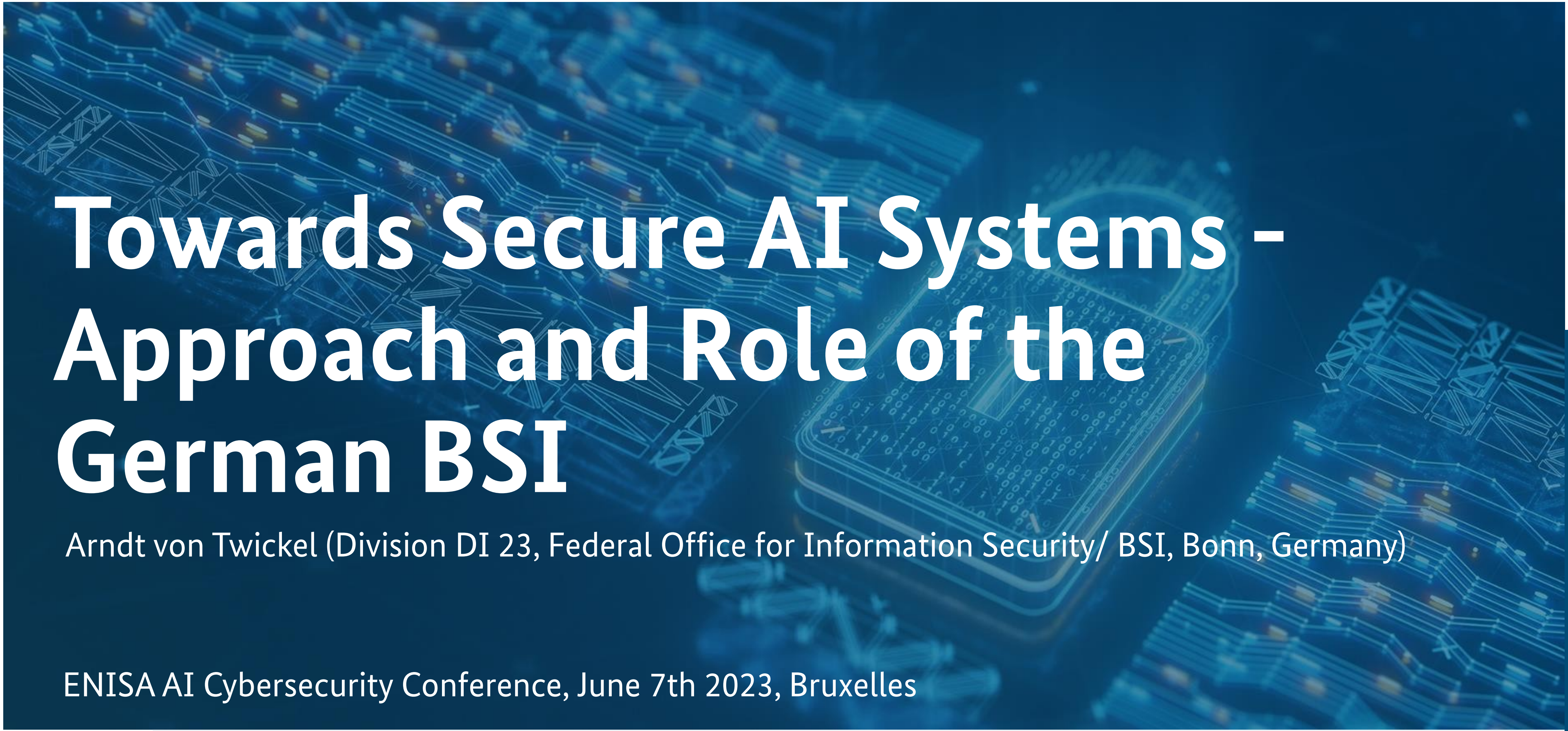
Federal Office
for Information Security



Towards Secure AI Systems - Approach and Role of the German BSI

Arndt von Twickel (Division DI 23, Federal Office for Information Security/ BSI, Bonn, Germany)

ENISA AI Cybersecurity Conference, June 7th 2023, Bruxelles

Mission statement

BSI as the Federal Cyber Security Authority shapes information security in digitalization through prevention, detection and reaction for government, business and society

AI as key technology

Artificial Intelligence @ BSI

IT-Security for AI

Investigation of new threats and development and evaluation of appropriate mitigation strategies

IT-Security through AI

We enable the usage of AI-methods to improve IT-security, e.g. for prevention, detection and reaction in the context of cyber attacks

Attacks via AI

Investigation of AI-driven and AI-supported attacks against IT-systems and infrastructures and development of appropriate mitigation strategies

AI and digital consumer protection

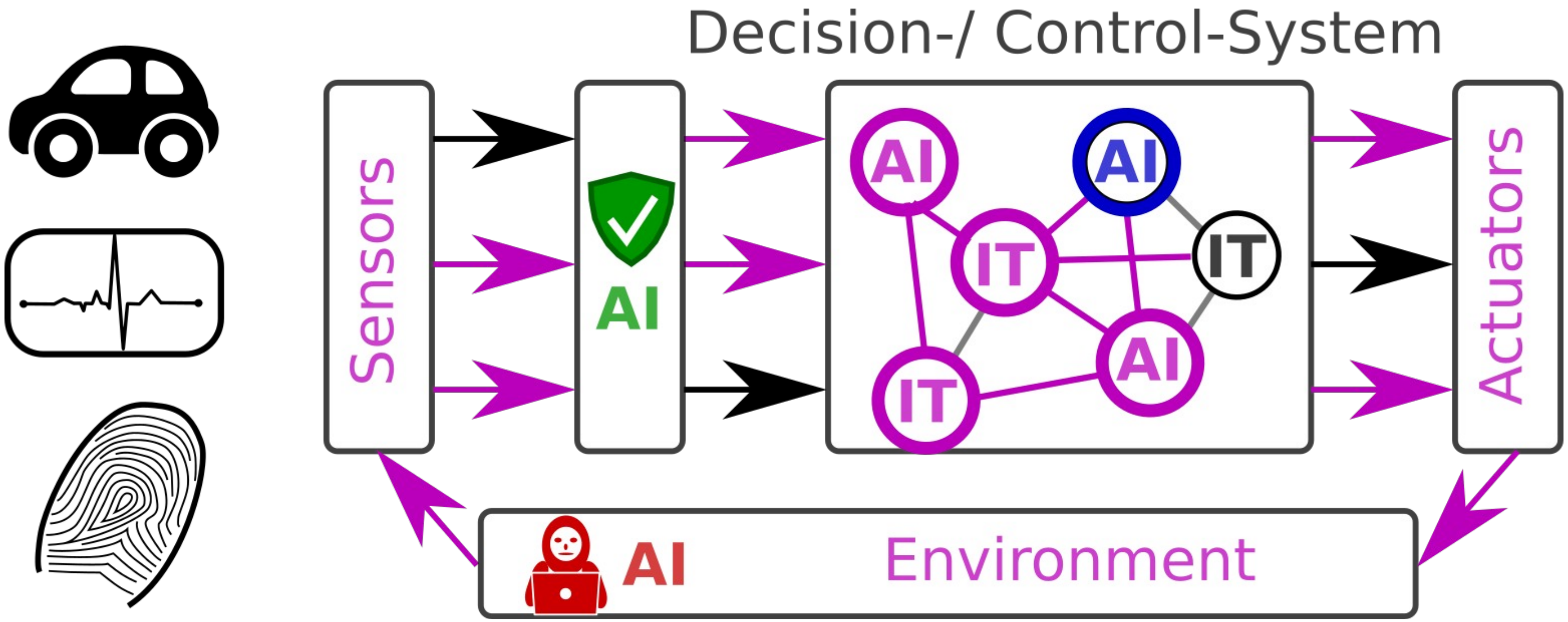
We promote the secure and transparent application of AI methods in consumer goods and increase the assessment ability of consumers for AI based products

Norms and standards for AI

We develop and evaluate audit criteria, audit methods and audit tools for verifable secure and trustworthy AI systems with the goal to develop norms and standards for these systems

AI in Digitization - Complexity and Challenges

Practical Criteria and Auditing of Security-Critical AI: Considering it as an Embedded System in the Use-Case Context is Necessary



BSI-relevant aspects:

- Performance
- Robustness
- IT-Security
- Safety
- Explainability
- ...

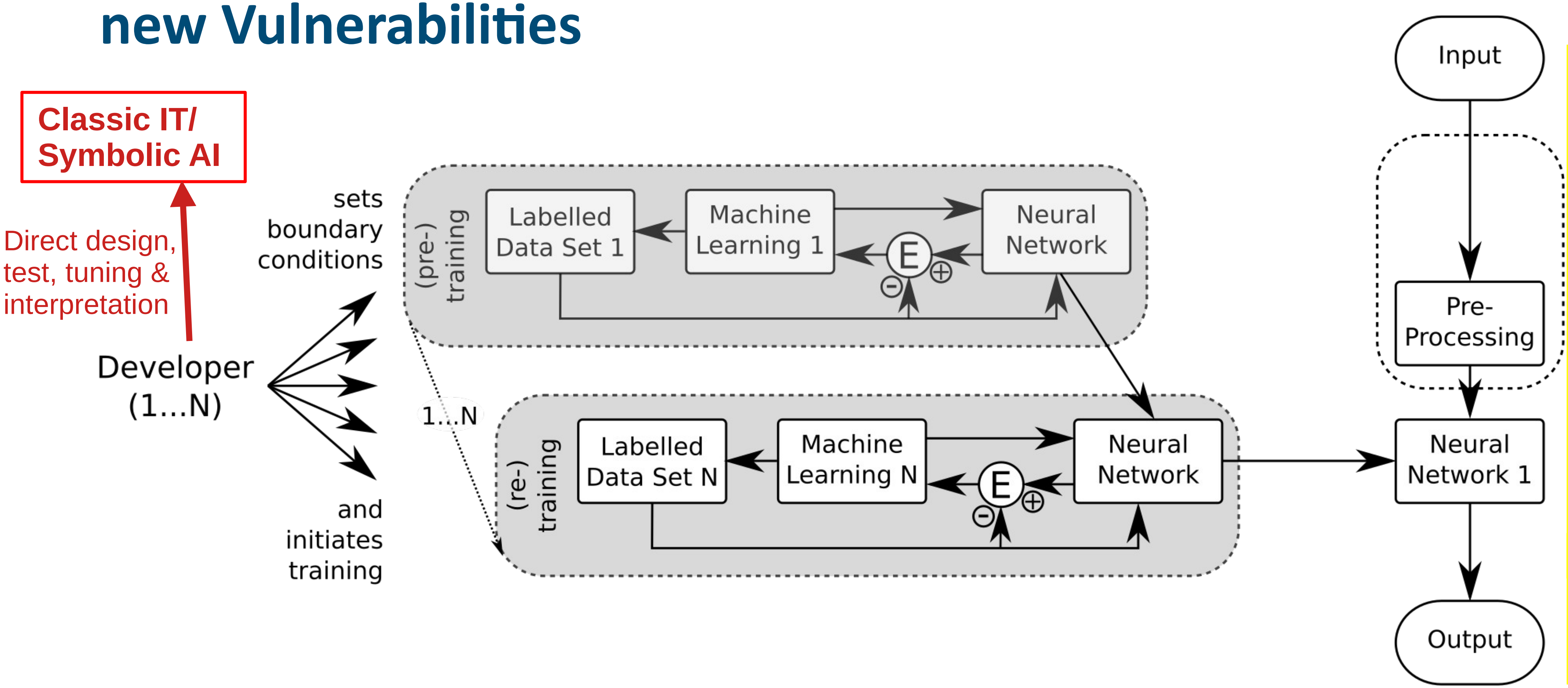
Non-BSI-relevant:

- Ethics
- User acceptance
- ...

1) Vulnerabilities of AI Systems 2) AI as a tool to attack IT 3) AI as a tool to defend IT

4) Interaction effects (emergence?)

Complex Connectionist AI-System Lifecycle Leads to Qualitatively new Vulnerabilities

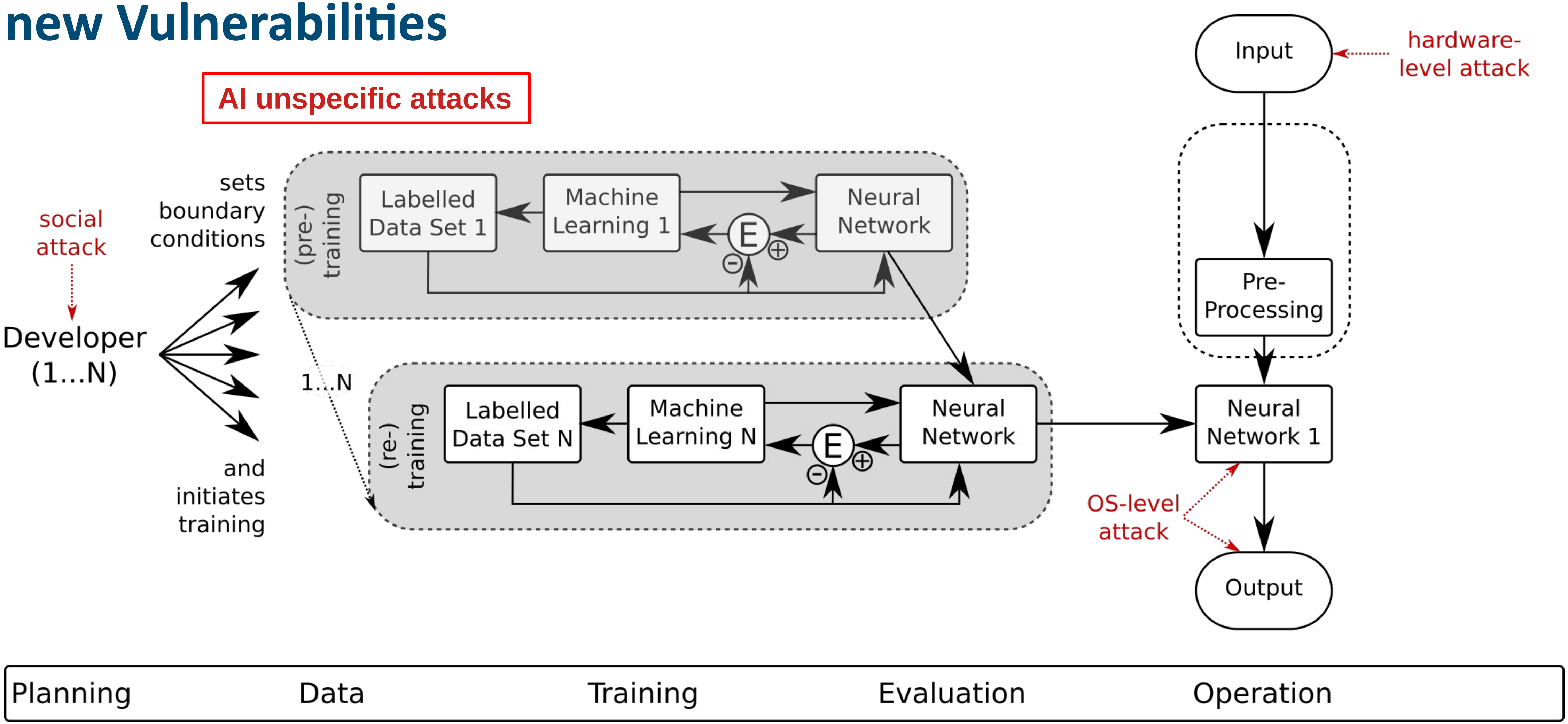


- input and state spaces are huge
- black-box properties
- dependency on training data

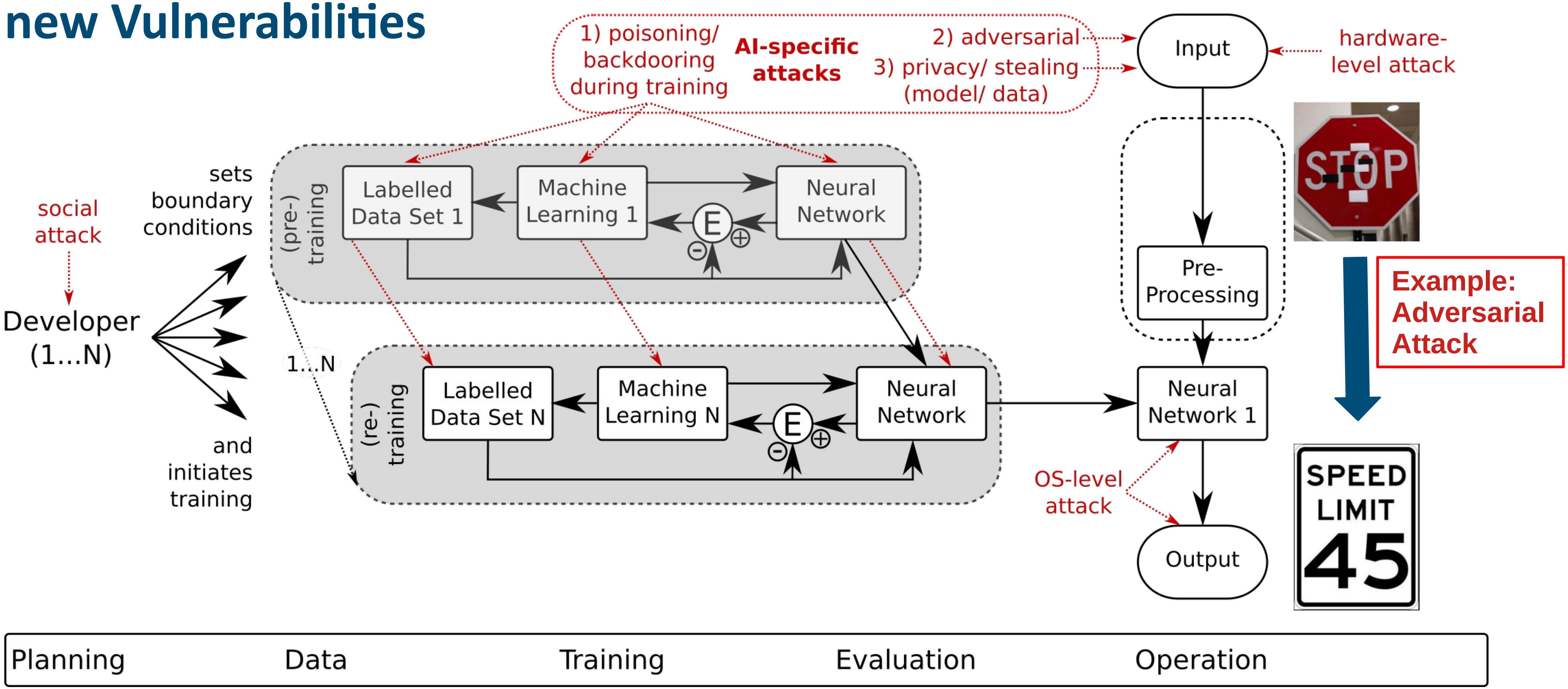
→ the whole life cycle has to be considered

Planning Data Training Evaluation Operation

Complex Connectionist AI-System Lifecycle Leads to Qualitatively new Vulnerabilities



Complex Connectionist AI-System Lifecycle Leads to Qualitatively new Vulnerabilities



1) poisoning/ backdooring during training

AI-specific attacks

2) adversarial

3) privacy/ stealing (model/ data)

Input

hardware-level attack

social attack

Developer (1…N)

sets boundary conditions

and initiates training

(pre-) training

Labelled Data Set 1

Machine Learning 1

Neural Network

E

1…N

(re-) training

Labelled Data Set N

Machine Learning N

Neural Network

E

Pre-Processing

Neural Network 1

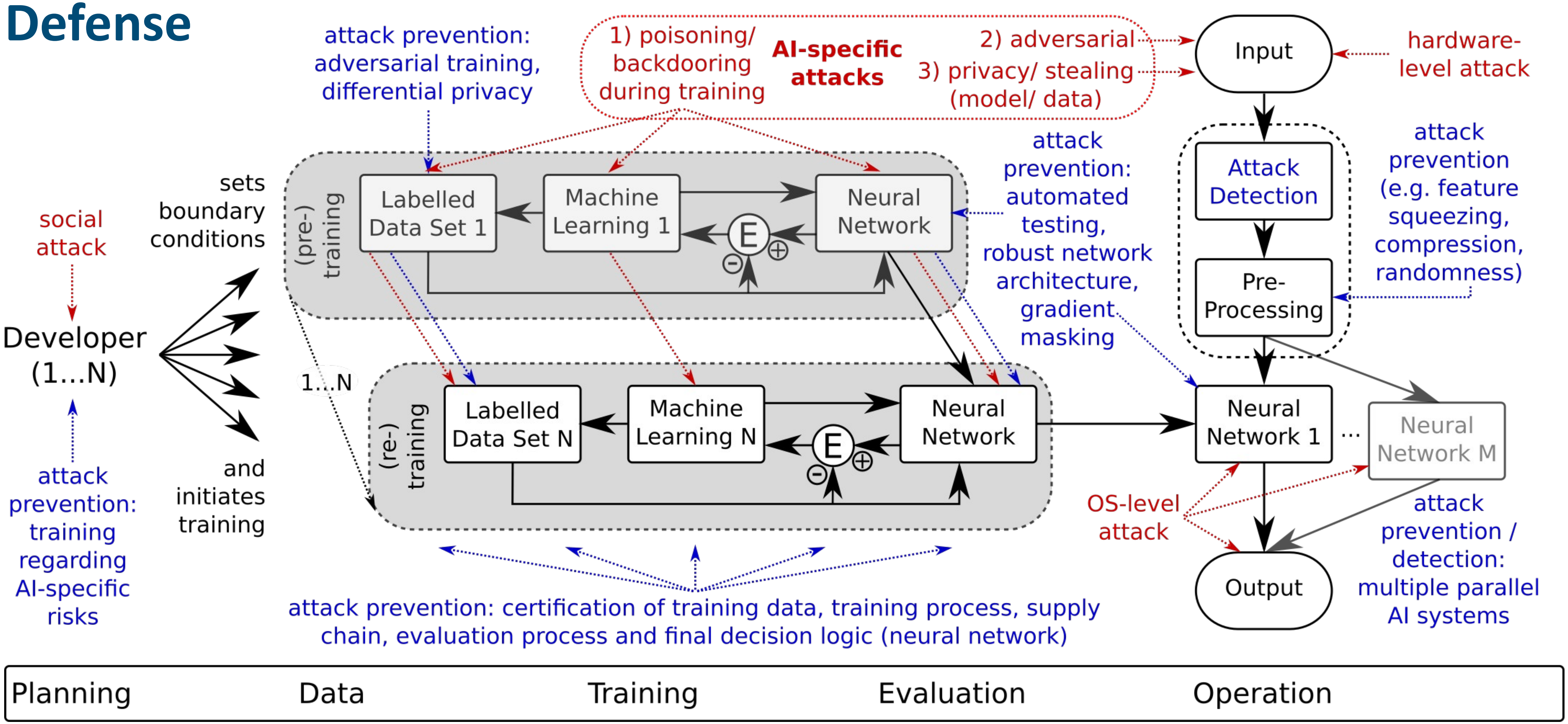
OS-level attack

Output

Example: Adversarial Attack

Planning | Data | Training | Evaluation | Operation

Complex Connectionist AI-System Requires Multiple Measures of Defense



attack prevention: adversarial training, differential privacy

1) poisoning/ backdooring during training

AI-specific attacks

2) adversarial

3) privacy/ stealing (model/ data)

Input

hardware-level attack

social attack

Developer (1…N)

sets boundary conditions

and initiates training

attack prevention: training regarding AI-specific risks

(pre-) training

Labelled Data Set 1

Machine Learning 1

E

⊖ ⊕

Neural Network

1…N

(re-) training

Labelled Data Set N

Machine Learning N

Neural Network

attack prevention: automated testing, robust network architecture, gradient masking

Attack Detection

Pre-Processing

attack prevention (e.g. feature squeezing, compression, randomness)

Neural Network 1

…

Neural Network M

OS-level attack

attack prevention / detection: multiple parallel AI systems

Output

attack prevention: certification of training data, training process, supply chain, evaluation process and final decision logic (neural network)

Planning | Data | Training | Evaluation | Operation

How to audit and regulate AI-systems? How to operationalize the European AI Act?

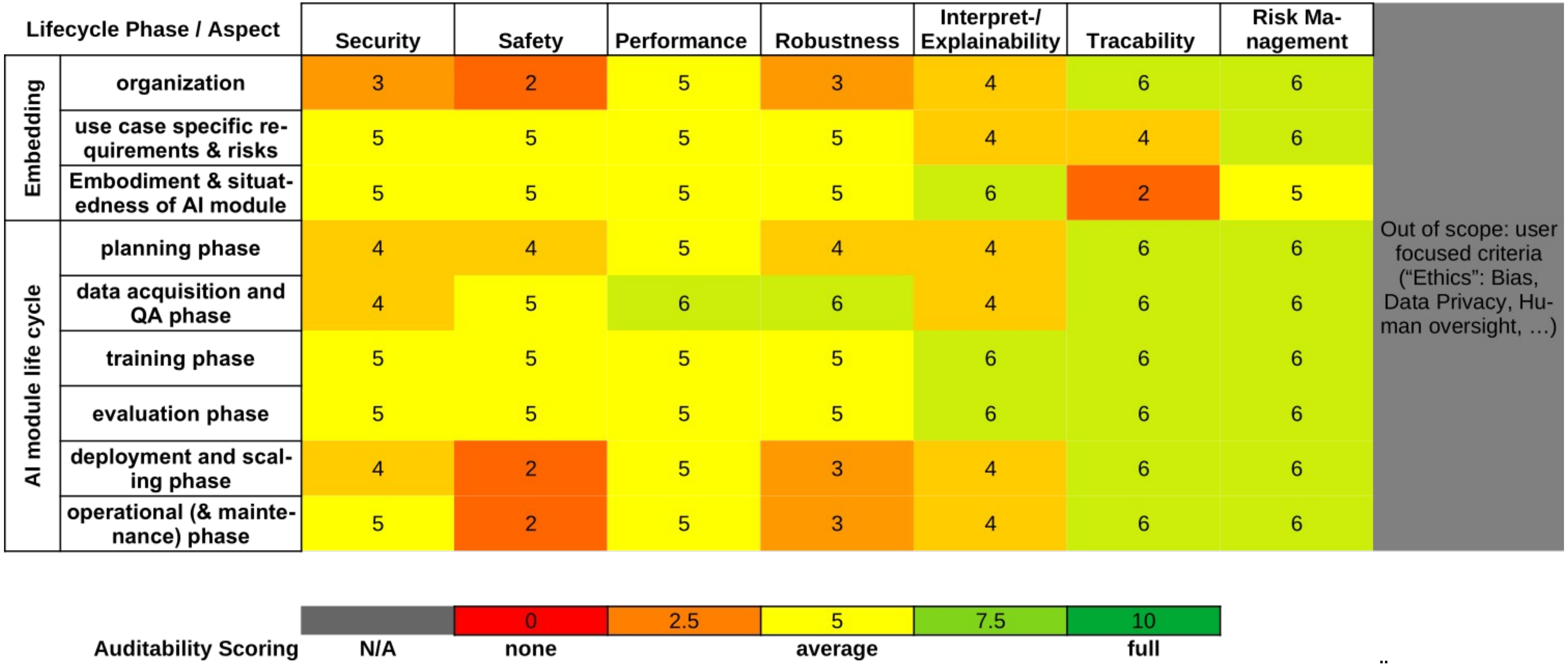
→ methods and tools either do not exist yet or are not yet sufficiently applicable in practice

Towards Auditable AI Systems: Assessment and Development of a Modular Requirement Catalogue and Audit Toolbox – an Iterative Process Between a Generalized AI Model and Application Specific Use Cases

A Comprehensive AI Auditability Assessment is Multi-Dimensional - Certification Readiness Matrix as a Tool

2nd dimension: relevant technical aspects

1st dimension: life cycle and embedding of AI system



Lifecycle Phase / Aspect		Security	Safety	Performance	Robustness	Interpret-/ Explainability	Tracability	Risk Management	
Embedding	organization	3	2	5	3	4	6	6	Out of scope: user focused criteria ("Ethics": Bias, Data Privacy, Human oversight, …)
	use case specific requirements & risks	5	5	5	5	4	4	6	
	Embodiment & situatedness of AI module	5	5	5	5	6	2	5	
AI module life cycle	planning phase	4	4	5	4	4	6	6	
	data acquisition and QA phase	4	5	6	6	4	6	6	
	training phase	5	5	5	5	6	6	6	
	evaluation phase	5	5	5	5	6	6	6	
	deployment and scaling phase	4	2	5	3	4	6	6	
	operational (& maintenance) phase	5	2	5	3	4	6	6	

Auditability Scoring		0	2.5	5	7.5	10
	N/A	none		average		full

Incorporating further dimensions by comparing multiple matrices:

3rd dimension: use case-specific ambient conditions

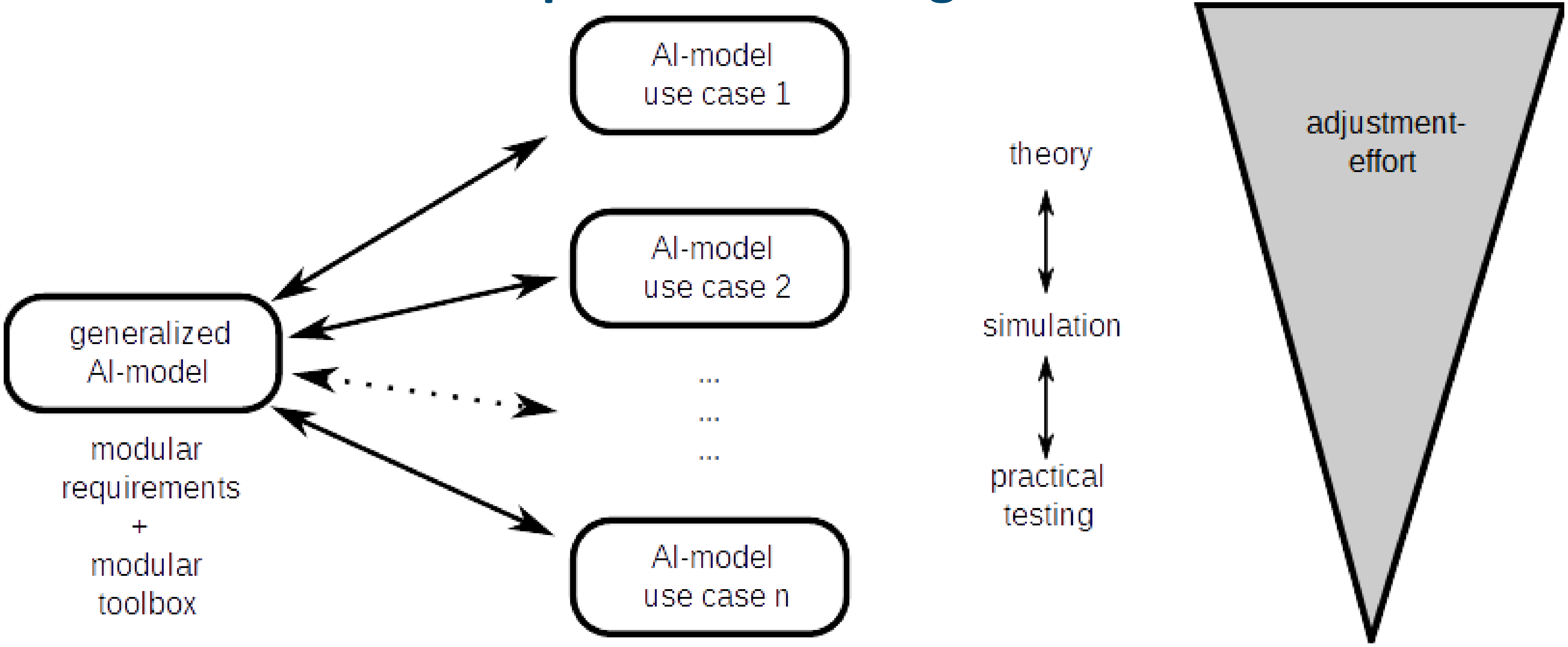
4th dimension: dynamic R&D developments in the field

BSI, TÜV-Verband & Fraunhofer HHI:Towards Auditable AI Systems - From Principles to Practice, Whitepaper, 05/2022

Open Tasks to Achieve Auditable, Certifiable and Trustworthy AI Systems

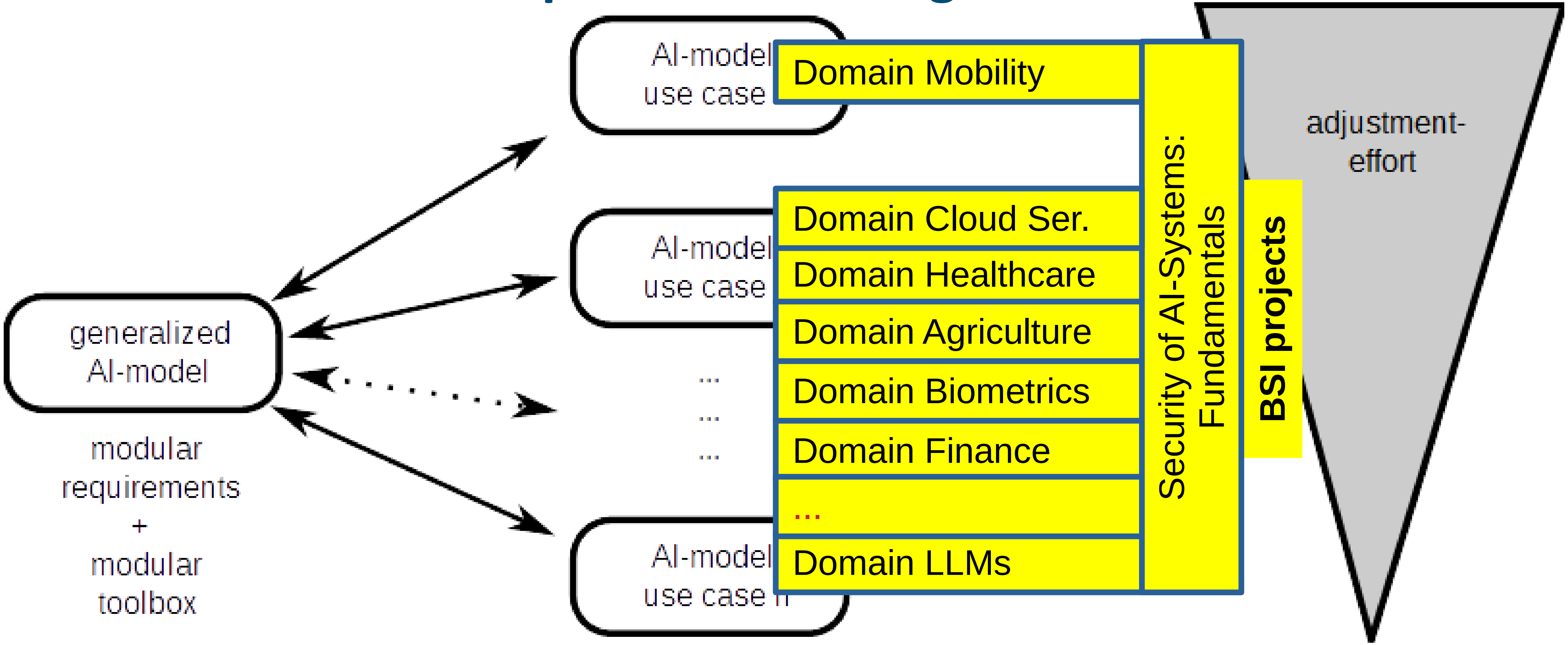
- Provide so far missing **technical, organizational and legal foundations** and derive **practically applicable methods and tools** (e.g. key trustworthiness indicators)
- Provide a **modular requirement catalogue** with instructions and examples of how to adapt it to arbitrary use cases
- Provide **best practices** for trustworthy by design development, auditing, mitigation strategies and tools and the determination of accountabilities
- Provide **necessary infrastructure** as a basis for the comparability of audit processes (data, scenario databases, interfaces, simulations, …)
- …

The Development of a Modular Requirement Catalogue and a Modular Audit Toolbox Requires Experience from Multiple Domains and Use Cases and AI-Specific Knowledge



Federal Office
for Information Security

The Development of a Modular Requirement Catalogue and a Modular Audit Toolbox Requires Experience from Multiple Domains and Use Cases and AI-Specific Knowledge



BSI contributions:

1) Develop domain- and use-case-specific documents and technical guidelines
2) Update the generalized AI model and develop modular technical guidelines
3) Use results from 1+2 to contribute to standardization, regulation and consulting

http://www.bsi.bund.de/dok/1079912

Project report:

Example: Projects AIMobilityAuditPrep and AIMobilityAudit

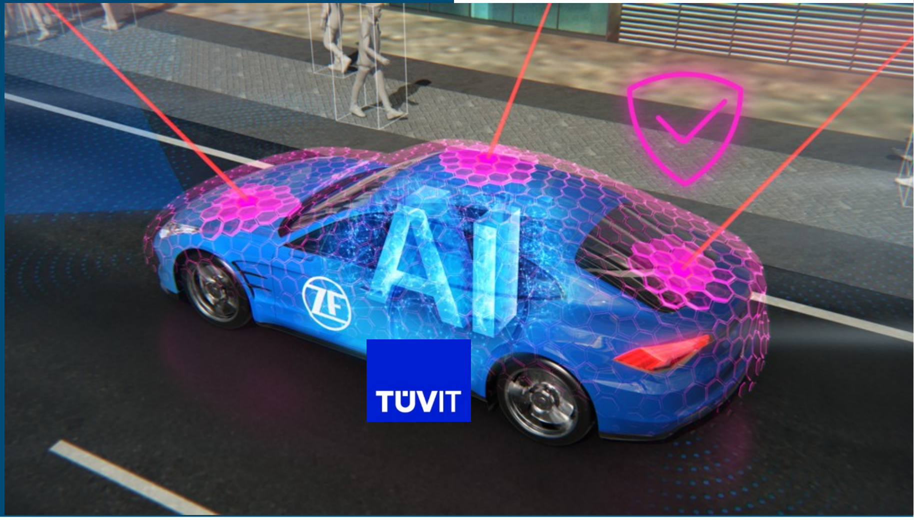
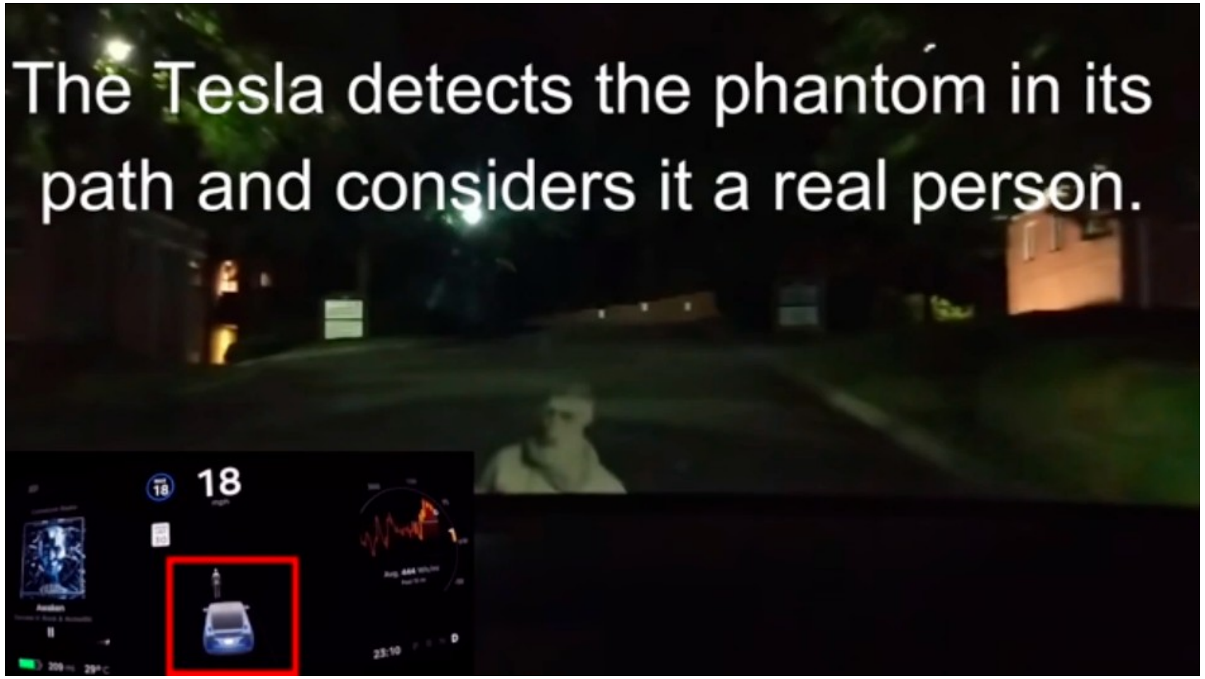
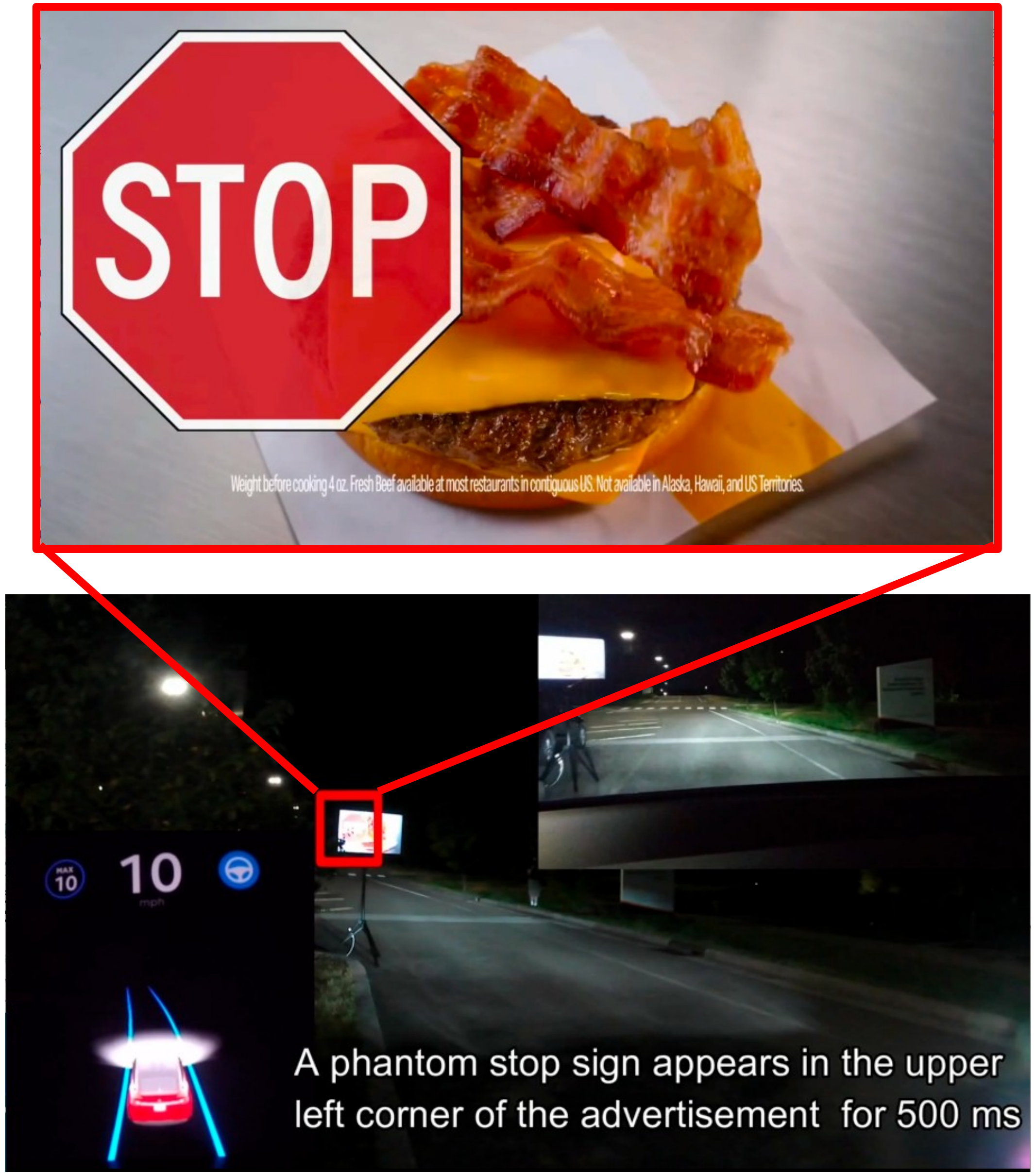


Image Source: ZF AI lab

Exemplary Vulnerabilities of Automated Driving Systems



The Tesla detects the phantom in its path and considers it a real person.



STOP
Weight before cooking 4 oz. Fresh Beef available at most restaurants in contiguous US. Not available in Alaska, Hawaii, and US Territories.
10

A phantom stop sign appears in the upper left corner of the advertisement for 500 ms

Nassi et al.: Phantom of the ADAS: Phantom Attacks on Driver-Assistance Systems, ACM CCS, 2020

Important Automotive Regulations Largely w/o AI-Specific Features

Area	Name	Content
Safety	ISO 26262	• Adaption of generic IEC 61508 for automotive contexts • Defines automotive safety integrity level (ASIL) • Focuses on functional safety for vehicles
	ISO 21448	• Defines safety of the intended functionality (SOTIF) • Focusses on risks of foreseeable misuse and shortcomings of the intended functionality for vehicles
	ANSI/UL 4600	• Focusses on safety processes for evaluating fully autonomous systems • Envisioned that specific standards are derived for concrete application areas
Security	UNECE R 155	• Defines cyber security management system (CSMS) • Focusses on organizational processes
	ISO/SAE 21434	• Defines cybersecurity assurance level (CAL) • Focusses on classification of cybersecurity activities

Ongoing Regulations Include or Focus on AI-Specific Features

Status	Name	Content
Draft	EU AI Act	• Defines risk levels for AI • Focusses on uniform regulations for AI-based systems
	ISO/IEC 24028 TR	• Focusses on trustworthiness of AI systems • Does not prescribe specific technologies/solutions
	ISO/IEC 24029-1 TR	• Focusses on assessing the robustness of DNNs • Does not prescribe specific technologies/solutions
Ongoing	ISO/IEC 5469 DTR	• Focusses on functional safety for AI-based systems
	ISO/AWI 8800 PAS	• Focusses on risk factors impacting the performance of AI-based systems in vehicles
	ISO/IEC 4213 PRF TS	• Focusses on assessing the performance of ML-based classification systems
	ISO/AWI 5083 TS	• Focusses on validating functionalities for automated driving on SAE L3/L4

AI Specific Requirements Were Derived for Entire Systems

- Analyze **gaps** in existing **standardizations** regarding AI-specific aspects
- Formulate **50 generic requirements** or best practices
- Provide requirements for **entire systems** (containing AI-based components)
- Partially based on ISO 26262:

ID	Method
FP4, IV3, ET2	Fault injection test
FP5	Error guessing test
FP6	Test derived from field experience
RS2	Stress test
RS3	Test for interference resistance and robustness

Req 7: The performance shall be compliant to the allowed worst-case error.

AI-Specific Requirements Were Derived for AI Subsystems

- Provide specific requirements for AI subsystems
- Partially based on ISO 26262:

ID	Method
UV10	Requirements-based test
UV14	Back-to-back comparison test between model and code

Req 33: The model’s decision shall be explained to aid the comparison between the modelling of the system and the trained model.

- Partially new:

Req 30: The training, test and evaluation datasets shall be independent from each other.

→ How can we test the **applicability & meaningfulness** of the requirements?

Use Case Selection Based on Suitability Assessment

- Apply proposed audit requirements to exemplary use case
- Find representative AI-based use case in mobility applications

Impact on control			
Local Path Planning	Lane Keeping	Lane Changing	Adaptive Cruise Control

No direct impact on control		
Global Path Planning	Traffic Sign Assistant	Driver Monitoring

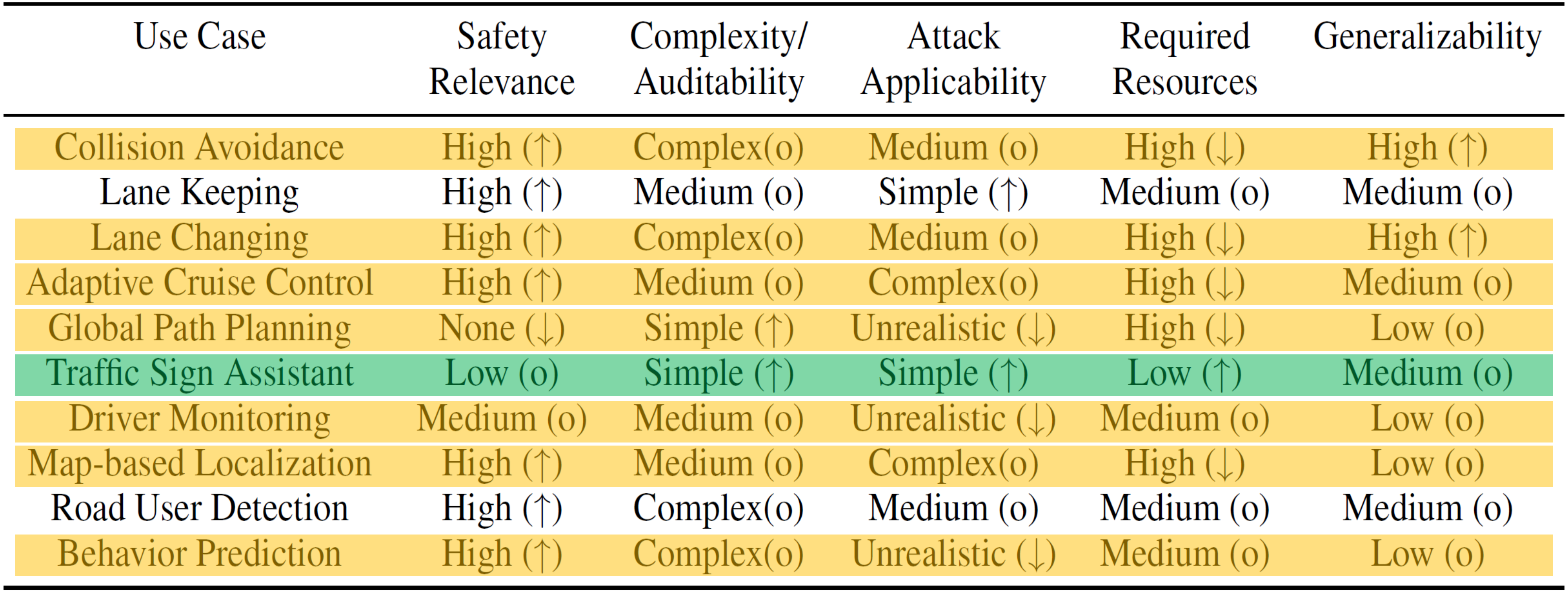
Basic functionalities		
Map-based Localization	Road User Detection	Behavior Prediction

- Assess suitability of each use case based on categories

Suitability categories				
Safety Relevance	Complexity/ Auditability	Attack Applicability	Required Resources	Generalizability

Use Case Selection Based on Suitability Assessment

suitable (↑), partially suitable (o), unsuitable (↓)



Use Case	Safety Relevance	Complexity/ Auditability	Attack Applicability	Required Resources	Generalizability
Collision Avoidance	High (↑)	Complex(o)	Medium (o)	High (↓)	High (↑)
Lane Keeping	High (↑)	Medium (o)	Simple (↑)	Medium (o)	Medium (o)
Lane Changing	High (↑)	Complex(o)	Medium (o)	High (↓)	High (↑)
Adaptive Cruise Control	High (↑)	Medium (o)	Complex(o)	High (↓)	Medium (o)
Global Path Planning	None (↓)	Simple (↑)	Unrealistic (↓)	High (↓)	Low (o)
Traffic Sign Assistant	Low (o)	Simple (↑)	Simple (↑)	Low (↑)	Medium (o)
Driver Monitoring	Medium (o)	Medium (o)	Unrealistic (↓)	Medium (o)	Low (o)
Map-based Localization	High (↑)	Medium (o)	Complex(o)	High (↓)	Low (o)
Road User Detection	High (↑)	Complex(o)	Medium (o)	Medium (o)	Medium (o)
Behavior Prediction	High (↑)	Complex(o)	Unrealistic (↓)	Medium (o)	Low (o)

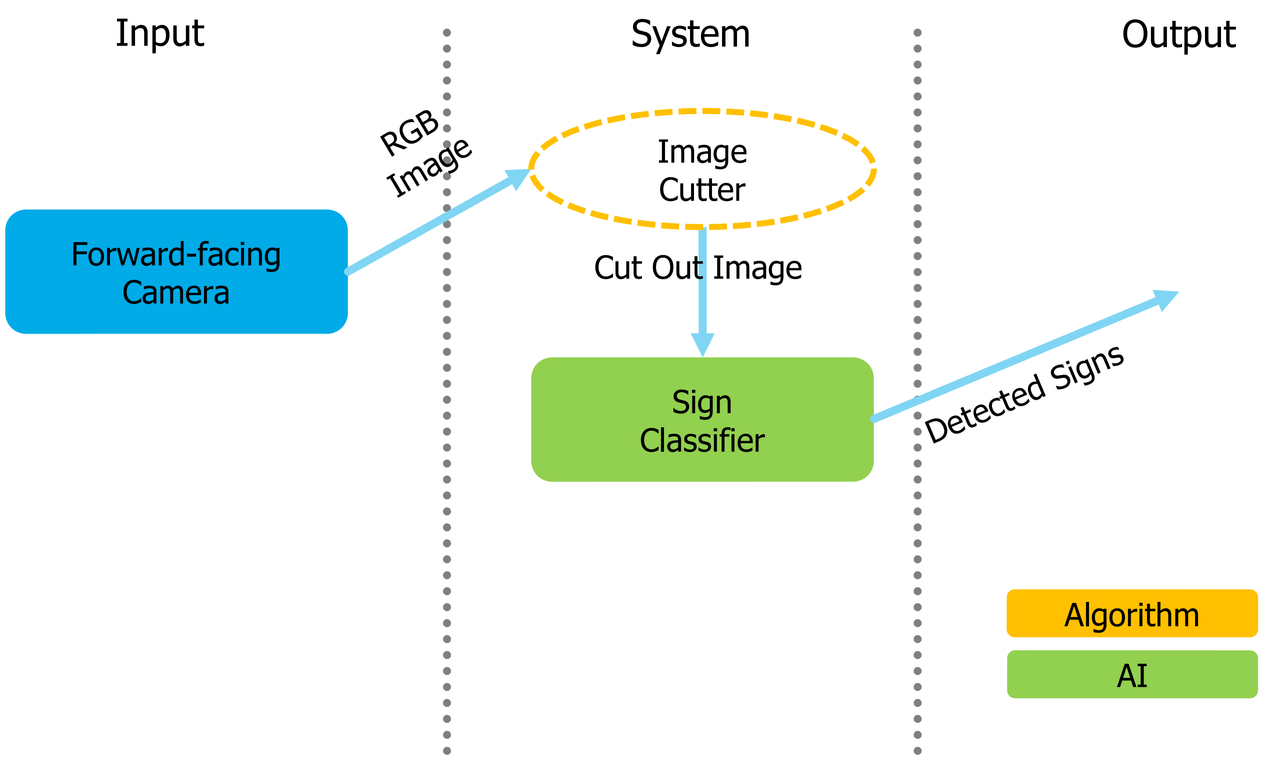
Feasibility of tests is valued highest

Exemplary AI-based System “Traffic Sign Assistant” - Overview

- Select **traffic sign assistant** as exemplary use case

- Use dataset for **German** traffic signs (GTSRB)

- Achieve standard **accuracy** of **>99%**



Previous study with ETH Zürich / Latticeflow: “Reliability Assessment of Traffic Sign Classifiers”, 2020, www.bsi.bund.de/KI

Exemplary Audit for the “Traffic Sign Assistant”

Req 7: The performance shall be compliant to the allowed worst-case error.

- Procedure: The performance shall be compliant to an **accuracy** above **90%** under **heavy rain** conditions.

Tested Samples	Correct Predictions	Failed Predictions	Accuracy
2580	2031	549	78,72% < 90%

- Verdict: **Failed**

Alternative Specification

- Procedure: The performance shall be compliant to an **accuracy** above **90%** under a **PGD** attack.

Tested Samples	Correct Predictions	Failed Predictions	Accuracy
2580	552	2028	21,40% < 90%

- Verdict: **Failed**

Federal Office
for Information Security

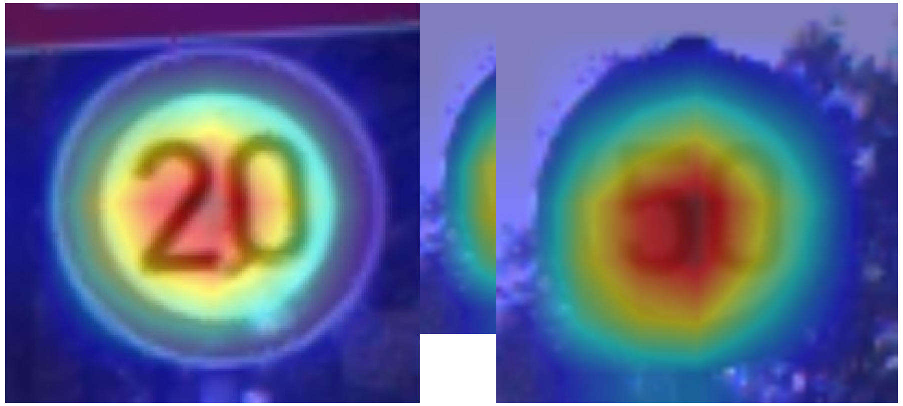
Exemplary Audit for the “Traffic Sign Assistant”

Req 14: The training, test and evaluation datasets shall be independent from each other.

- Procedure: No specification required.
- Verdict: **Passed**
  - Source code shows splitting of data into three disjoint datasets
  - Datasets appear independent and from the same distribution

Req 32: The model’s decision shall be explained to aid the comparison between modelling of the system and the trained model.

- Procedure: The model decision shall depend on displayed figures and/or the signs coloration/shape.
- Verdict: **Passed**



Red regions have highest influence on the decision

What's next?

- AIMobilityAudit (currently running)
  - Increase complexity of exemplary systems
  - **Investigate** different requirements in **practice**
  - Apply requirements to **industry-grade** systems
  - Test **quality** of current **mitigation** strategies
  - Create **technical guideline** for vehicle **homologation**

- Strategic goals
  - Obtain **practical** insights, limitations & **feedback** for requirements
  - Refine proposed requirements
  - Use obtained results as **blueprint** for **standardization** activities

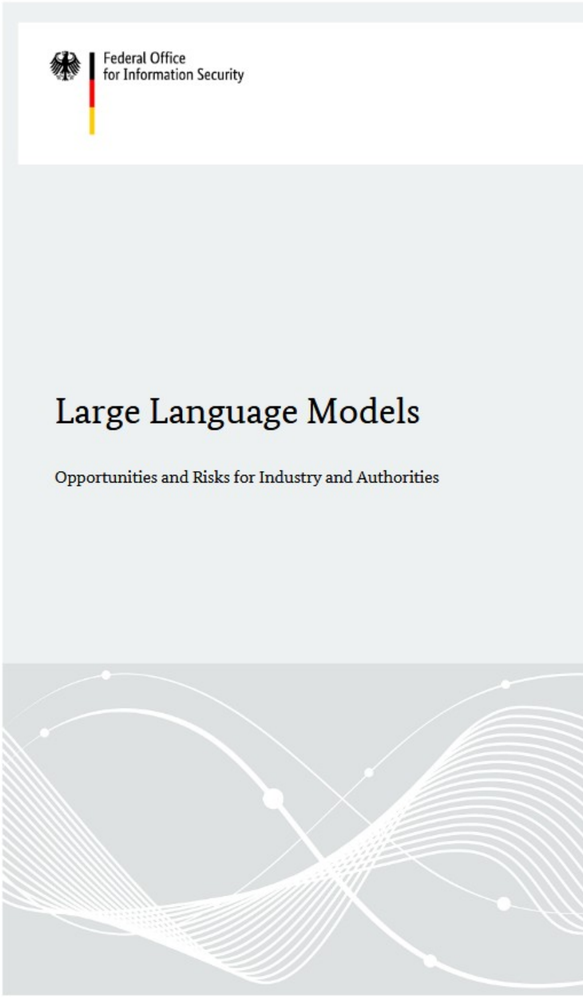
BSI Activities: Documents and Collaborations

BSI Documents on AI Security at www.bsi.bund.de/KI

- Secure, robust and transparent application of AI - Problems, measures and need for action
- AI security concerns in a nutshell - Practical AI-Security guide
- AI Cloud Service Compliance Criteria Catalogue (AIC4)
- Vulnerabilities of Connectionist AI Applications: Evaluation and Defense (Frontiers Big Data)
- Towards Auditable AI Systems: two whitepapers (2021 + 2022) with VdTÜV and FhG HHI
- The Interplay of AI and Biometrics: Challenges and Opportunities (IEEE Computer)
- Deep Learning Reproducibility and Explainable AI (XAI)
- Security of AI-Systems: Fundamentals - Adversarial Deep Learning
- Security of AI-Systems: Fundamentals - Provision or use of external data or trained models
- Security of AI-Systems: Fundamentals - Security Considerations for Symbolic and Hybrid AI
- **Opportunities and Risks of Large Language Models (LLMs) for Industry and Authorities: “**Systematic risk analysis for specific use case strongly recommended”

BSI Participation in Working Groups

- National: BSI-VdTÜV AI working group, DIN/DKE AI Standardization Roadmap, Platform I 4.0, ...
- International: ETSI's Industry Spec. Group on Securing Artificial Intelligence (ISG SAI), ENISA Adhoc working group on AI, UNECE GRVA Workshop AI, CEN-CENELEC, ISO, CC Biometrics, ...



Federal Office
for Information Security

Large Language Models

Opportunities and Risks for Industry and Authorities

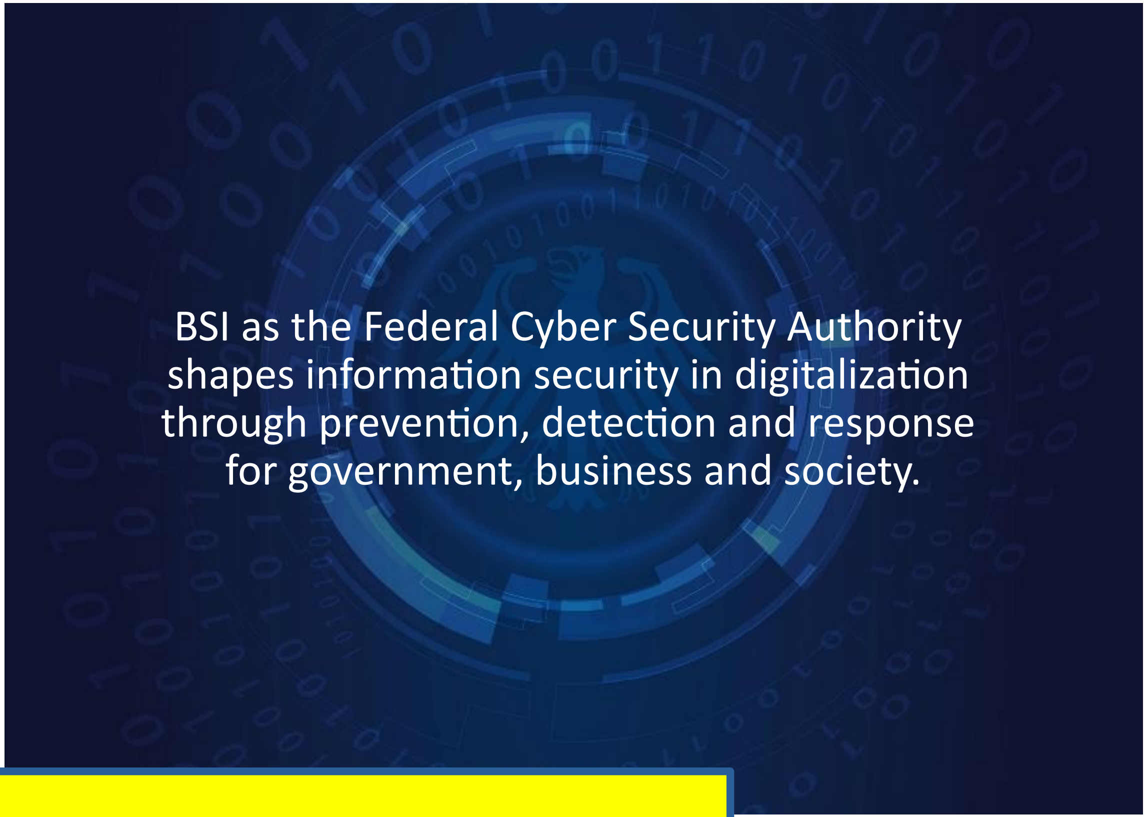
Thank you for your attention!

Contact

Dr. Arndt von Twickel
Head of Division „Cybersecurity for Intelligent Transport Systems and Industry 4.0“

arndt.twickel@bsi.bund.de

Federal Office for Information Security (BSI)
Godesberger Allee 185-189
53175 Bonn
www.bsi.bund.de



BSI as the Federal Cyber Security Authority shapes information security in digitalization through prevention, detection and response for government, business and society.

Invitations:

- **September 5th-8th: IAA Mobility – BSI booth**
- **November 10th: BSI-TÜV-HHI Workshop (Berlin) “Towards Auditable AI Systems: New Challenges Introduced by Generative AI”**