



Federal Office  
for Information Security

Deutschland  
**Digital•Sicher•BSI•**

# Towards Auditable AI Systems

Arndt von Twickel, Christian Berghoff and Matthias Neu

„Trustworthy AI“ for the „AI for Good Global Summit“,  
United Nations International Telecommunication Union (ITU), Zoom  
Webinar, April 15th 2021

The BSI – the national  
cyber security authority

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

# Responsibilities of the BSI in the Context of AI

## 1) Vulnerabilities of AI systems

- Evaluation of existing and development of new evaluation and protection methods
- Development of technical guidelines and standards

## 2) AI as a tool to defend IT systems

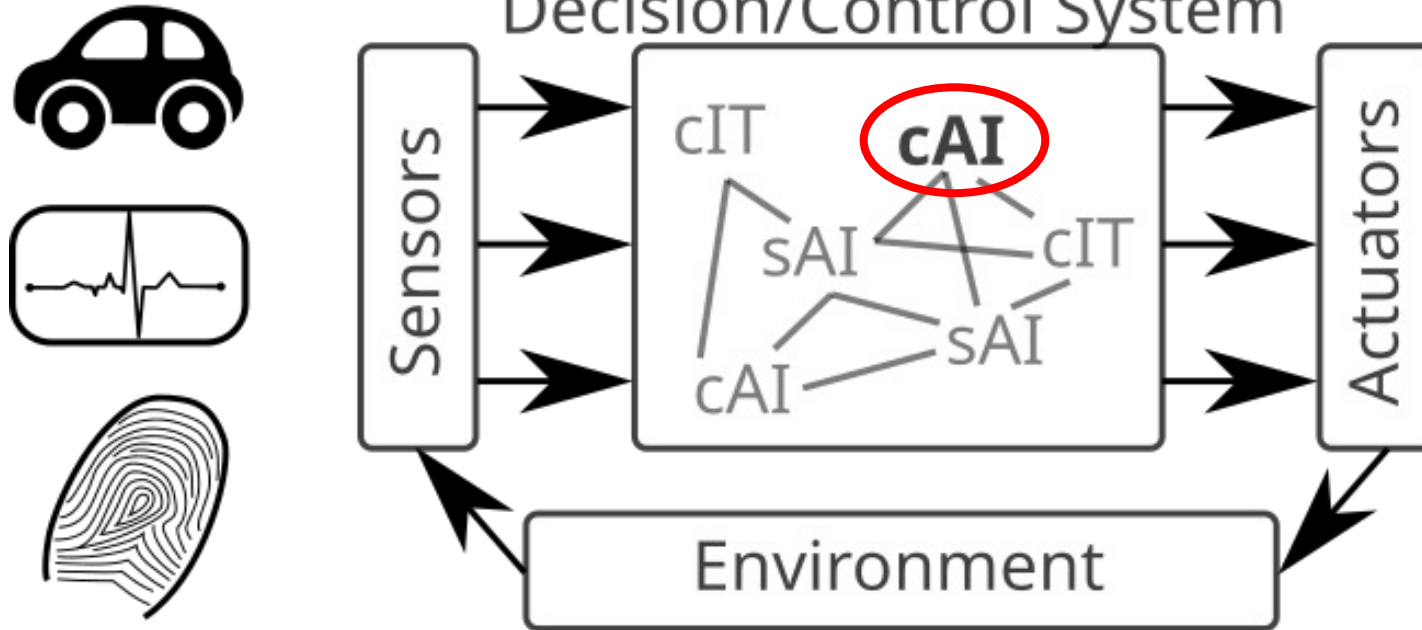
- Recommendations of existing and development of new technologies, guidelines for their deployment and operation

## 3) AI as a tool to attack IT systems

- How can one protect IT system from qualitatively new AI based attacks?

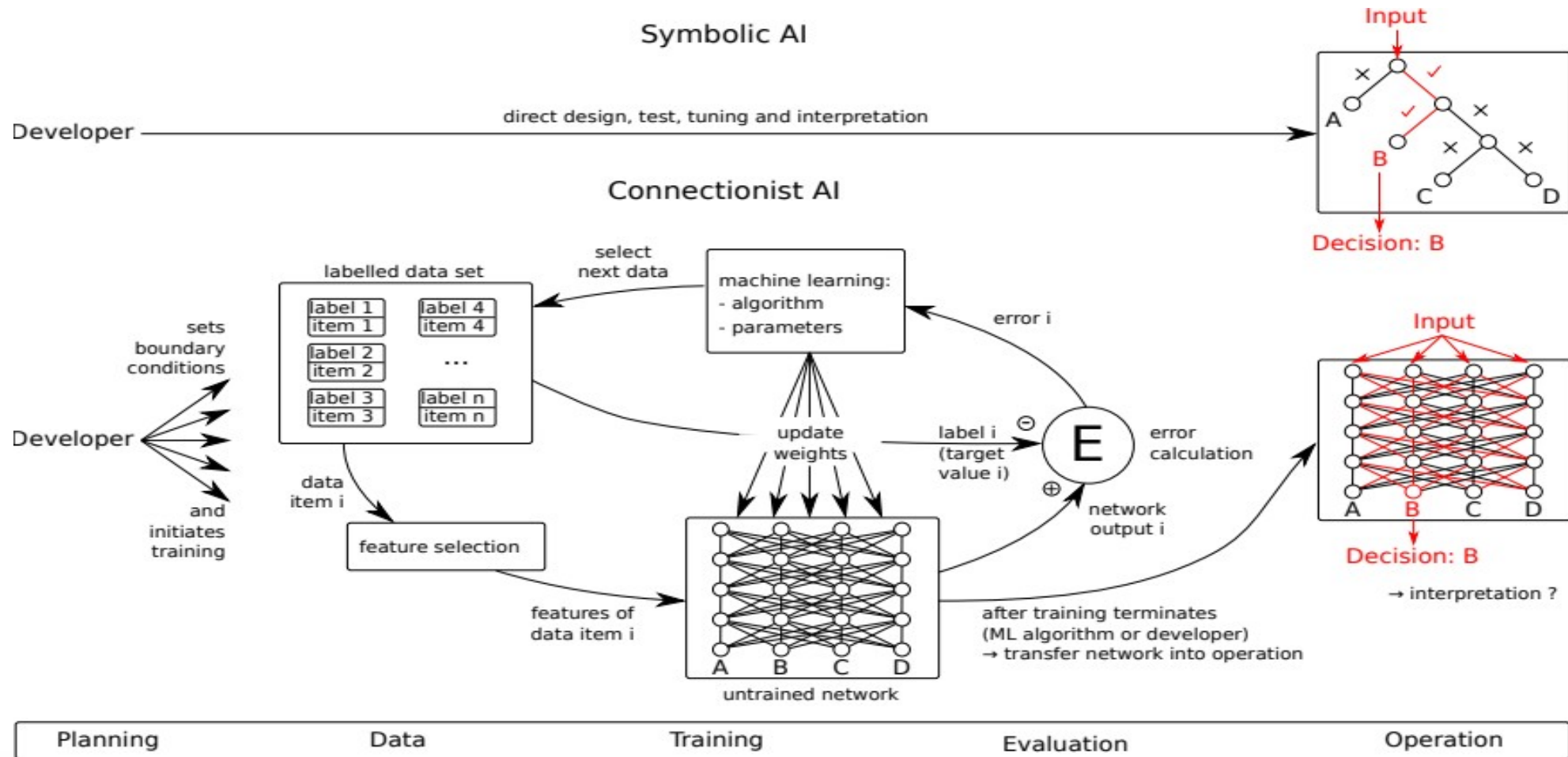
# Life-cycle of Connectionist AI Systems

# AI Systems are Connected and Embedded in Safety and Security Critical Applications



- cIT  $\hat{=}$  clasiscal IT
- SAI  $\hat{=}$  symbolic AI
- cAI  $\hat{=}$  connectionist AI

# Connectionist AI Differs Qualitatively From Symbolic AI and Classical IT

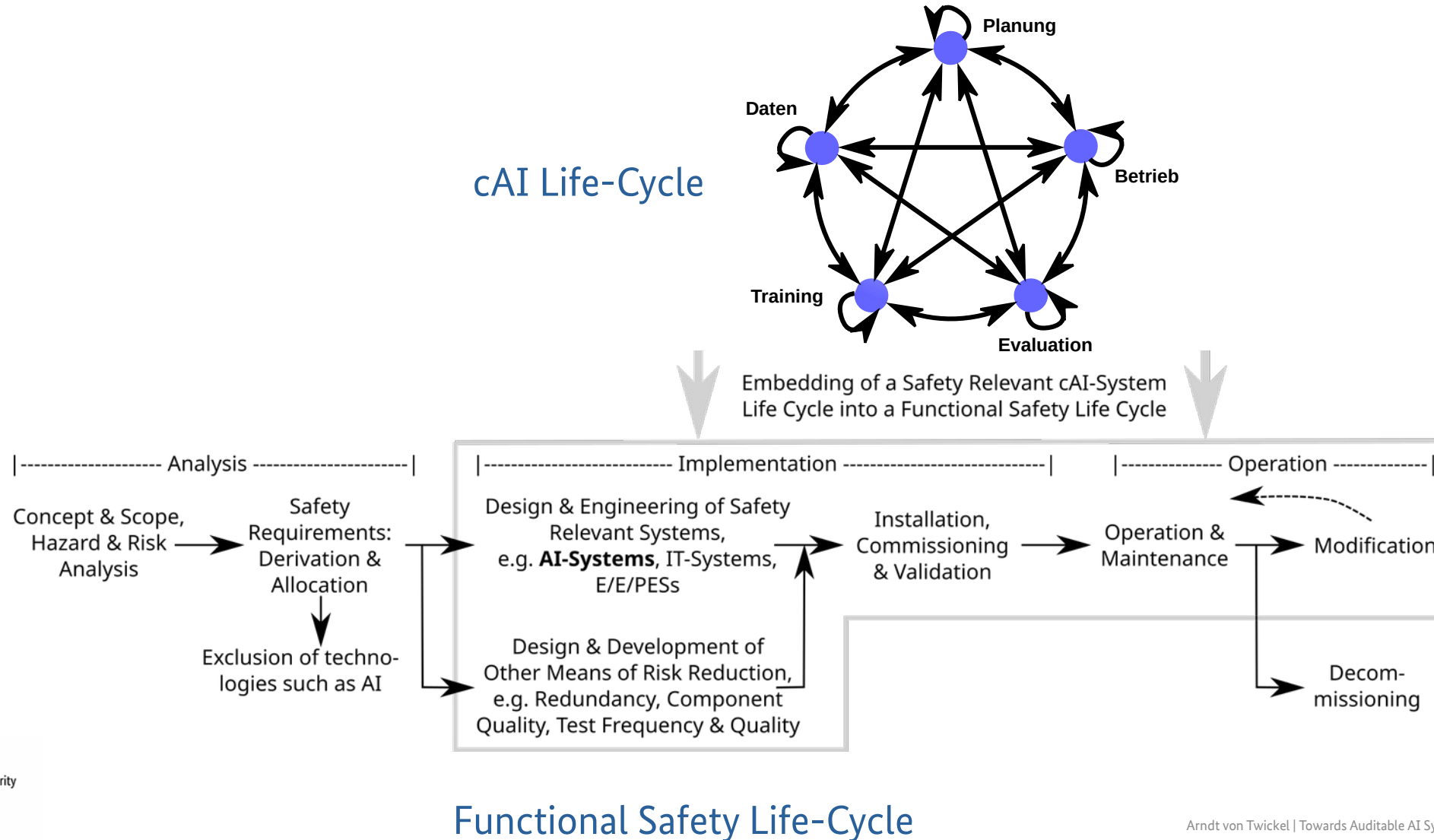


# Connectionist AI has Specific and Qualitatively new Problems

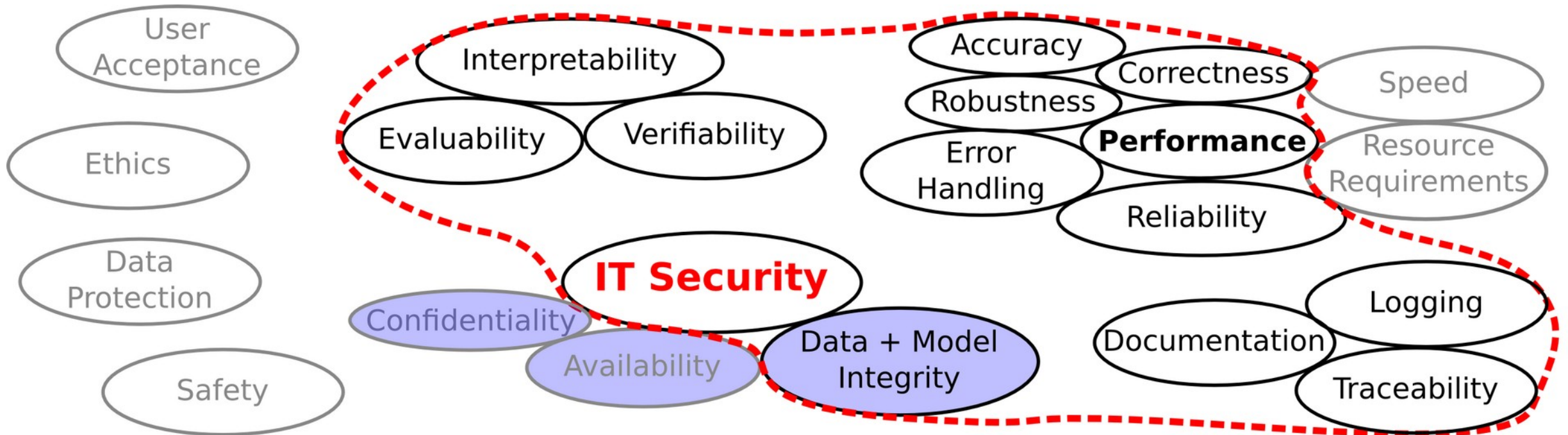
- input and state spaces are huge
  - black-box properties
  - dependency on training data
- > whole process chain / life cycle has to be considered



# Embedding of a cAI Life-Cycle Into a Functional Safety Life-Cycle



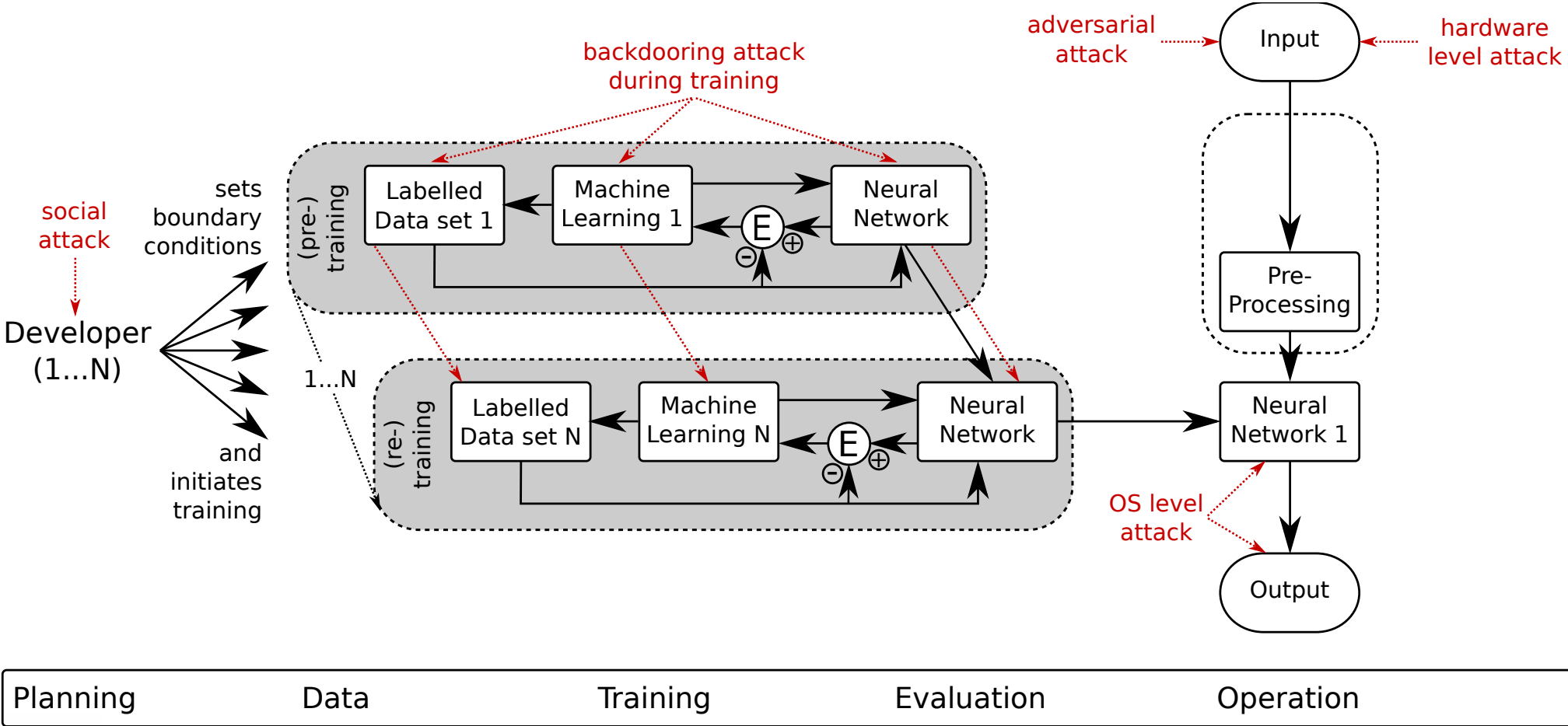
# Multiple Aspects Have to be Considered for Securing AI Systems



# Vulnerabilities of cAI Systems

Berghoff C, Neu M and von Twickel A (2020):  
Vulnerabilities of Connectionist AI Applications: Evaluation and Defense.  
Front. Big Data 3:23

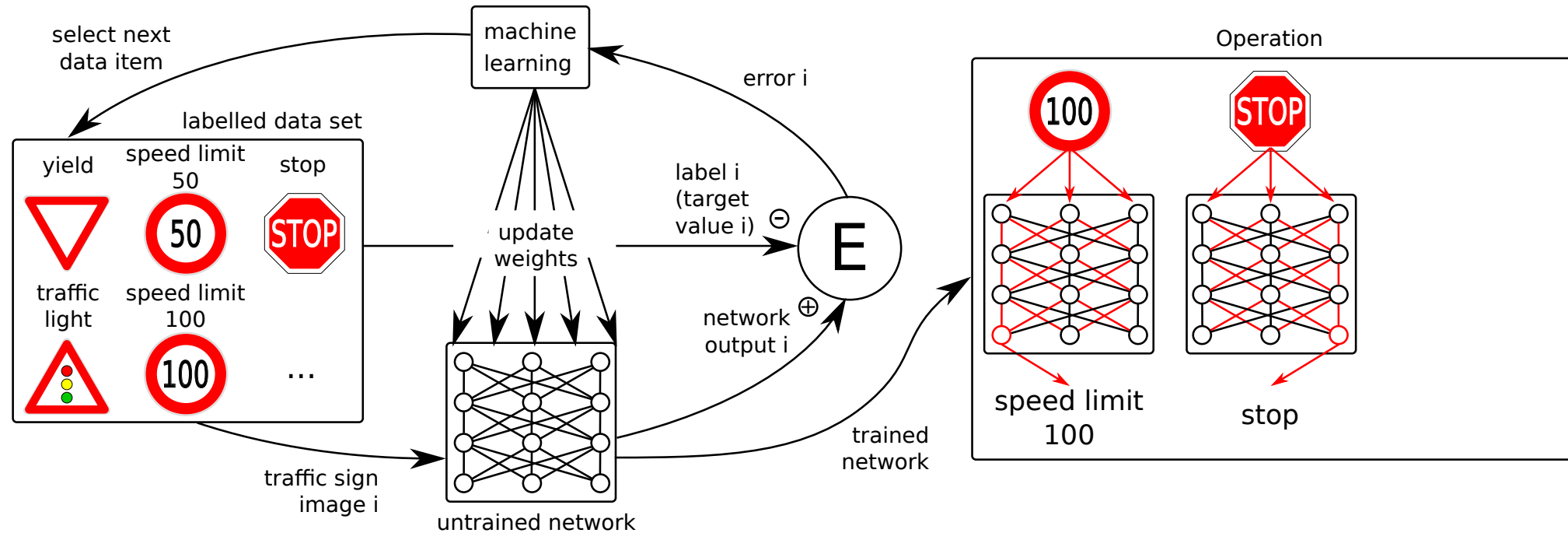
# Connectionist AI Process Chain: Attack Vectors



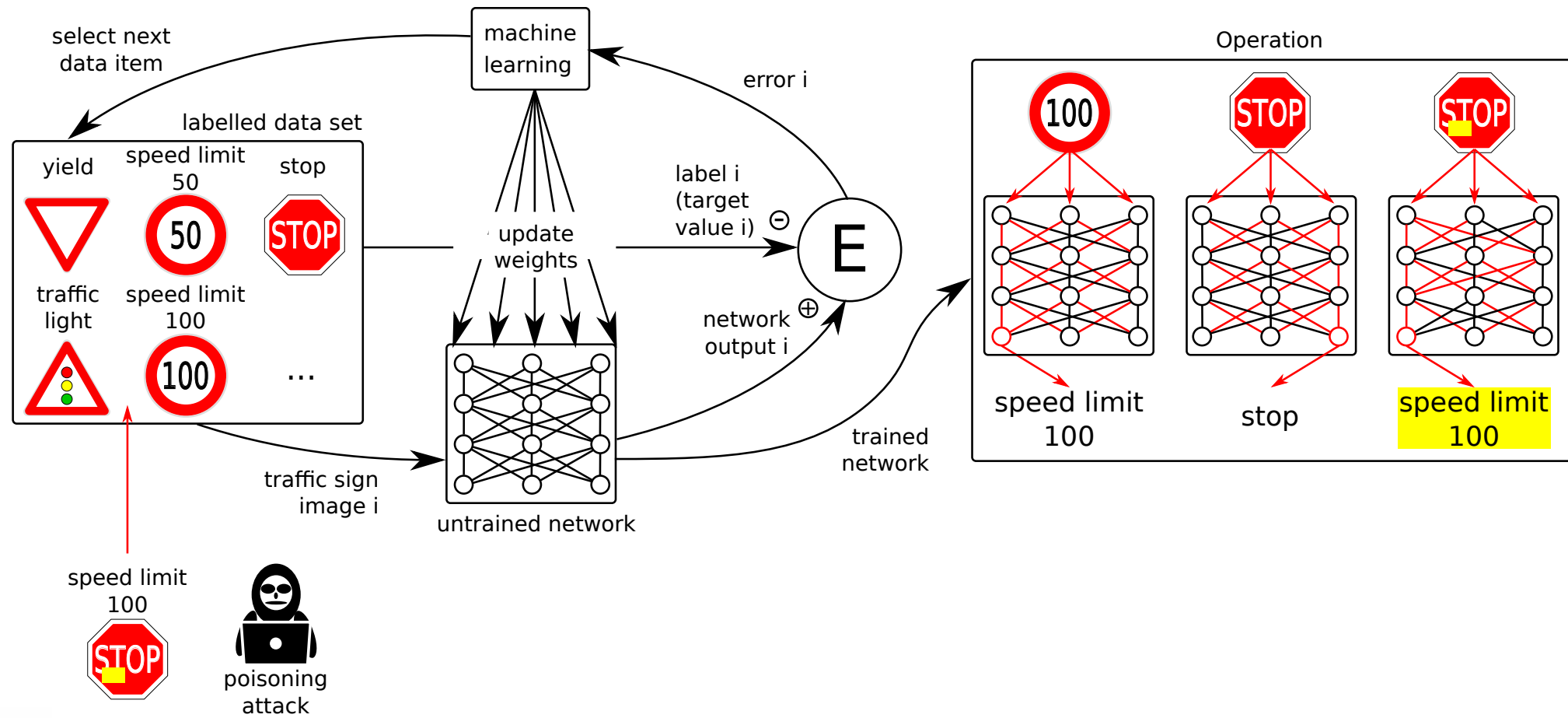
# AI-Specific Attack on Road Sign Classification Systems

## A) Poisoning Backdoor Attacks

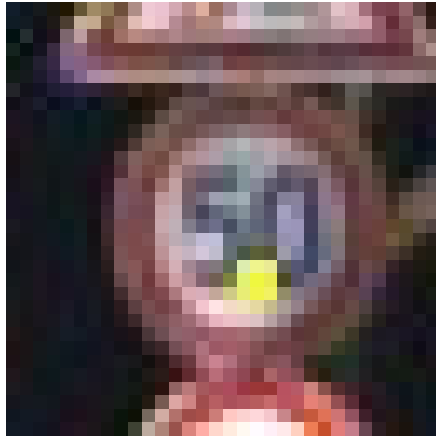
# Poisoning-Attack (schematic)



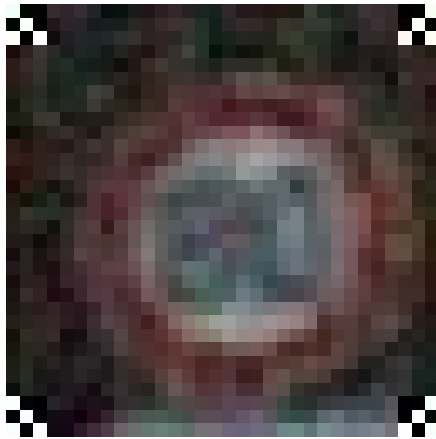
# Poisoning-Attack (schematic)



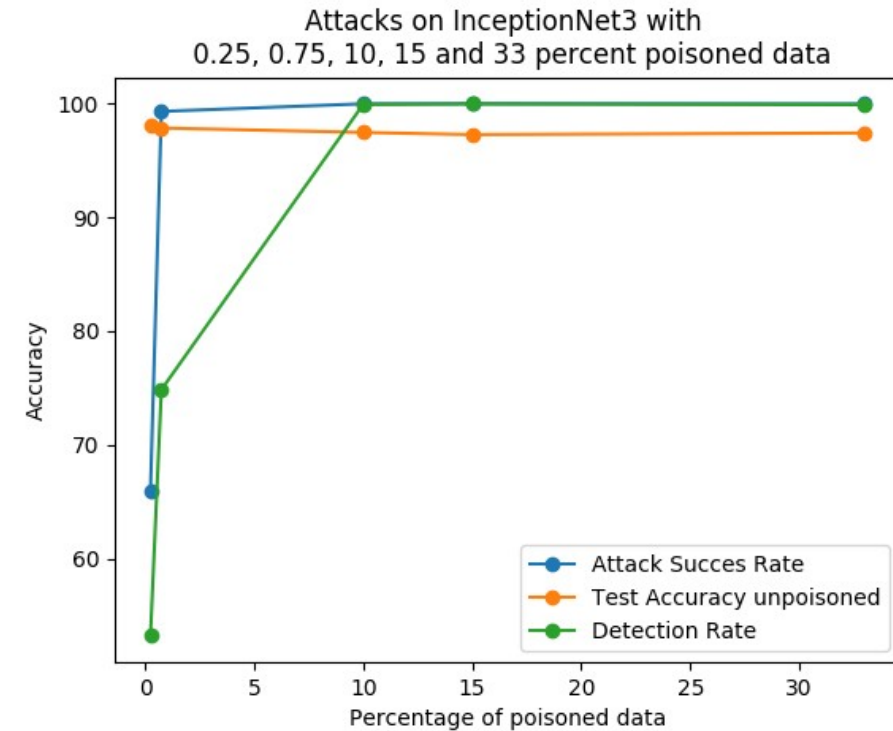
# Poisoning-Attack (hands on)



Attack A:  
50 km/h sign  
+ yellow sticker  
Label: 80 km/h



Attack B:  
Arbitrary sign  
+ 4 s/w stickers  
Label: 80 km/h

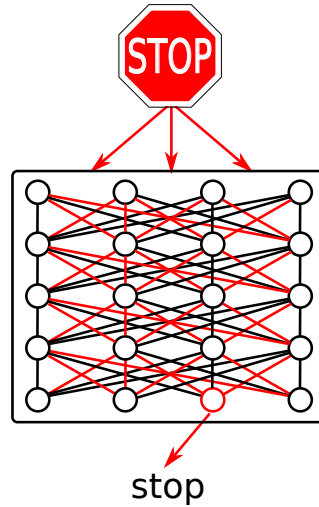


- --> Attack with 98% accuracy auf InceptionNet3



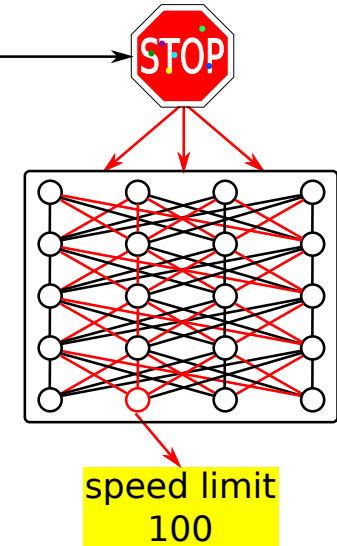
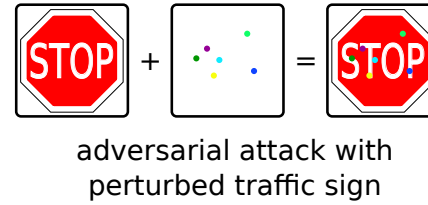
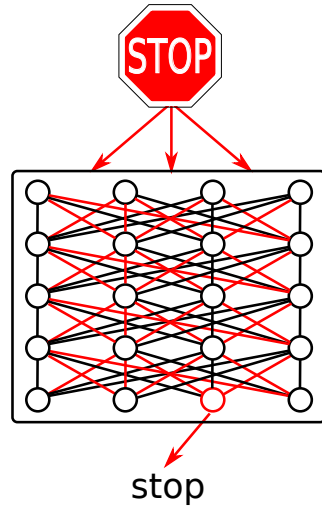
AI-Specific Attacks on Road Sign  
Classification Systems  
B) Adversarial Attacks

# Adversarial Attack



standard operation

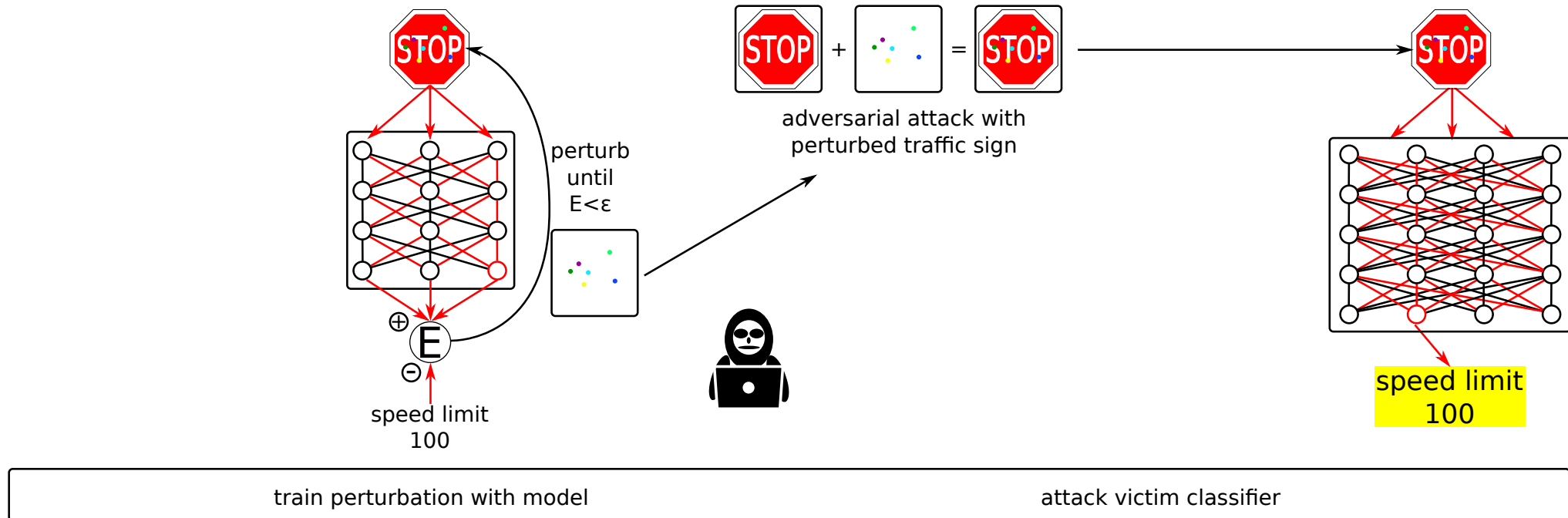
# Adversarial Attack



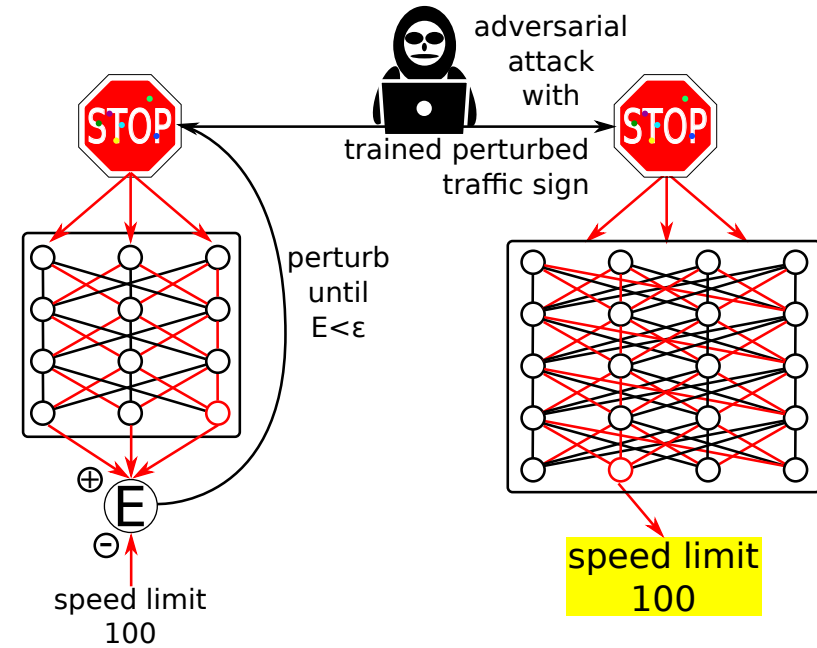
standard operation

attack victim classifier

# Adversarial Attack

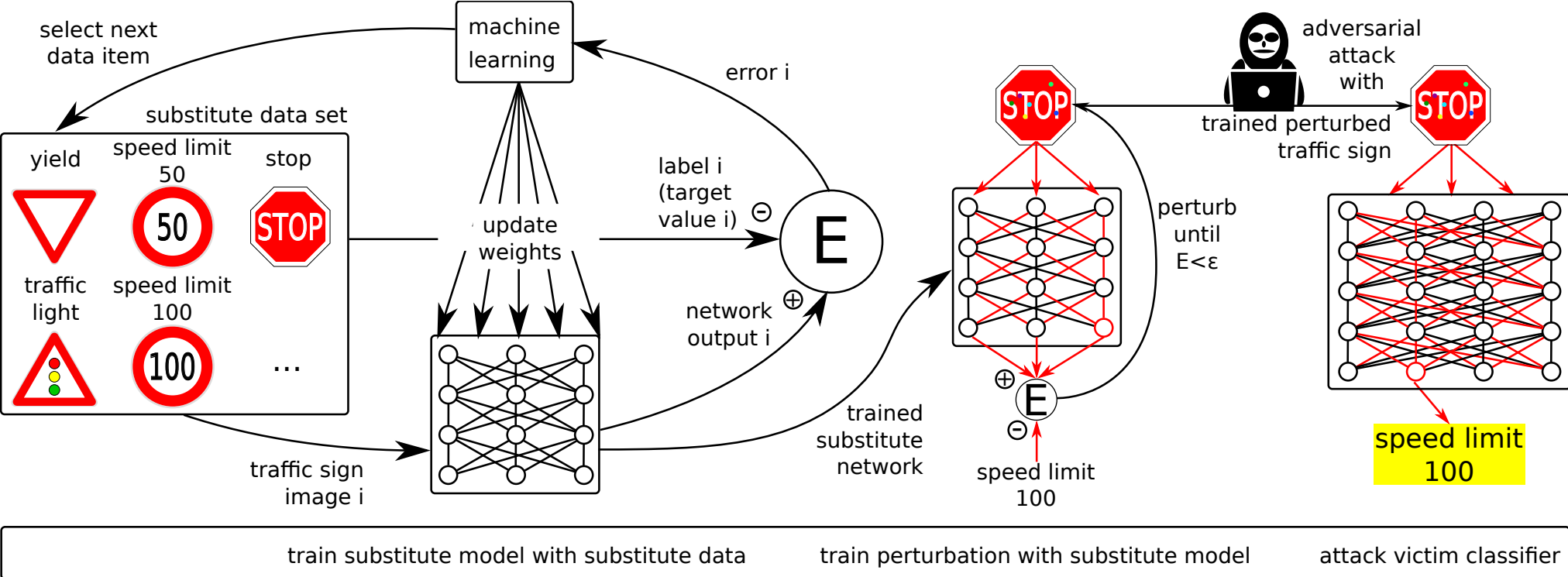


# Adversarial Attack

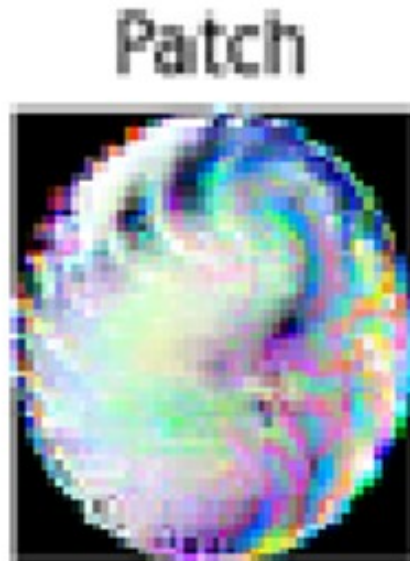


train perturbation with (substitute) model      attack victim classifier

# Adversarial Attack



# Adversarial Attack Examples



Original Image : 14



Adversarial Image



Original Image : 35



Adversarial Image



Original Image : 33



Adversarial Image



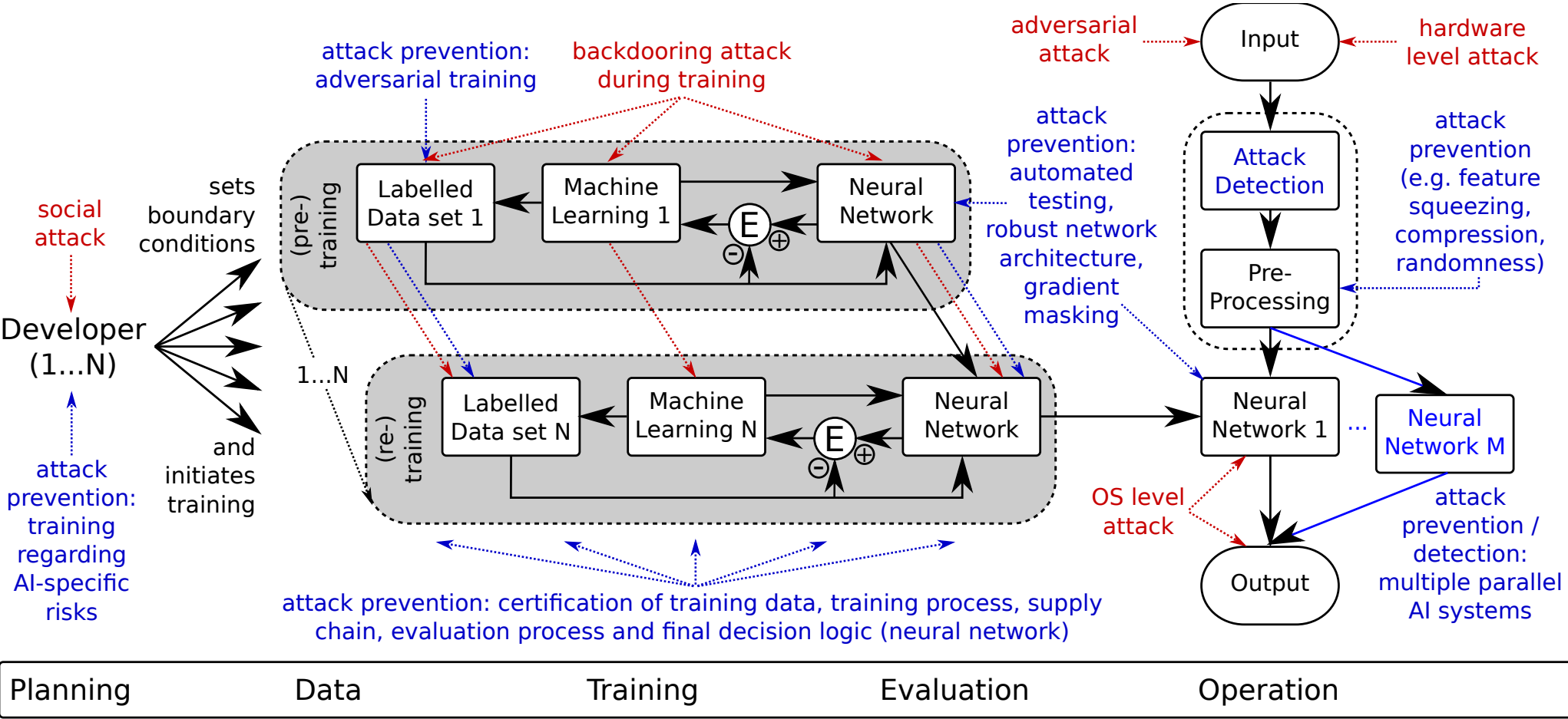
Ausgabe:



# Measures of Defense



# Connectionist AI Process Chain: Vulnerabilities and Measures of Defense



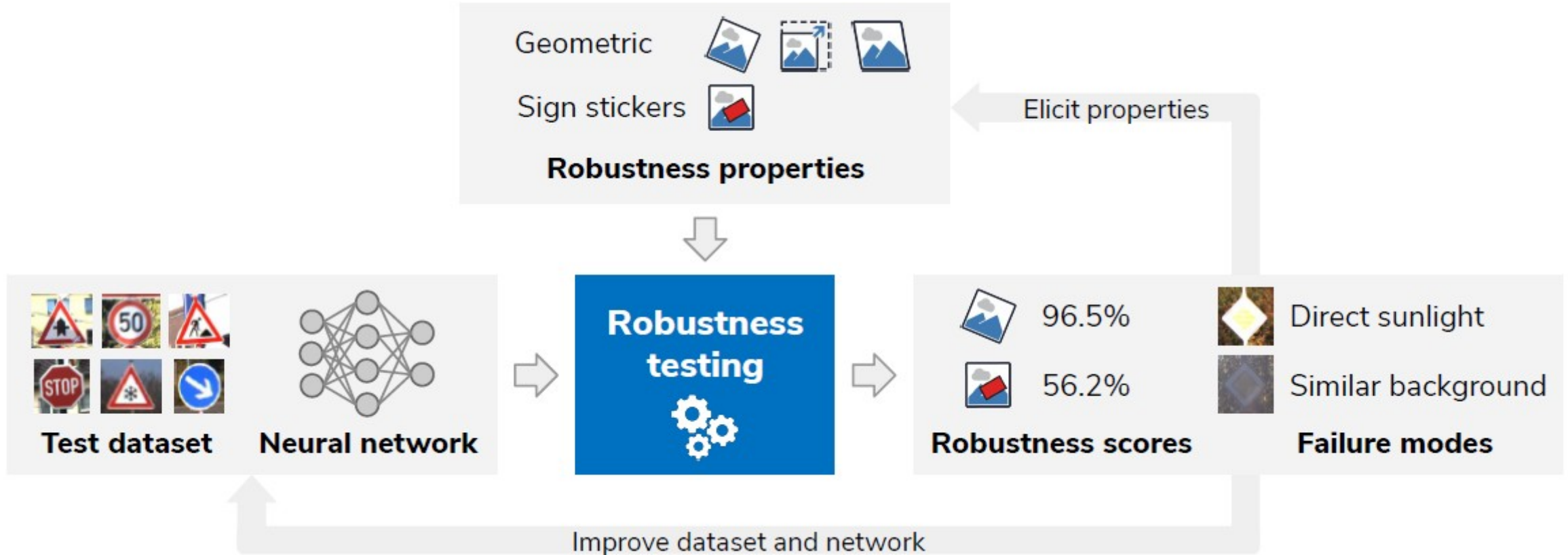
→ single measures are not effective against adaptive attackers

# Robustness of AI Systems

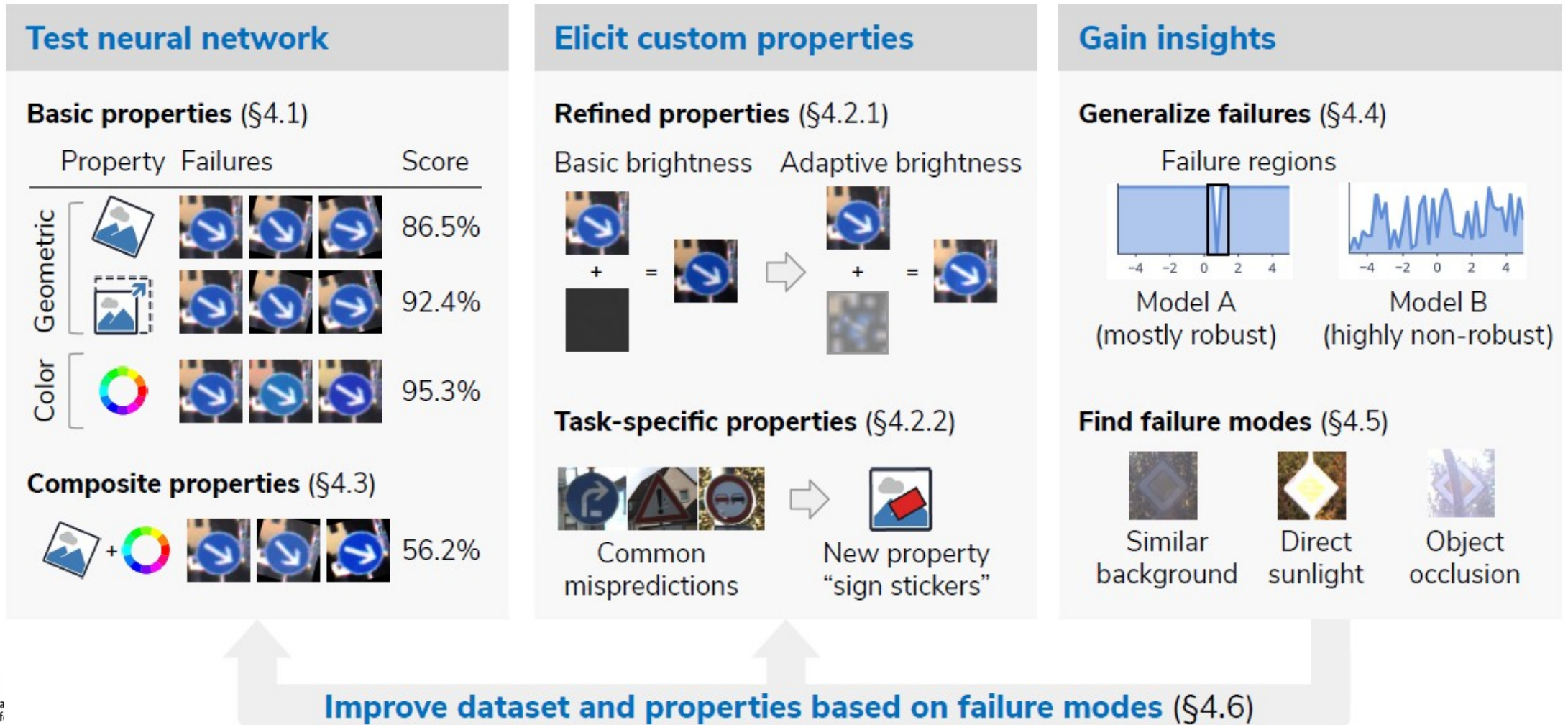
(Project with ETH Zurich / Latticeflow,

Report available under [www.bsi.bund.de/KI](http://www.bsi.bund.de/KI))

# Test and Improvement of the Robustness of Neural Networks



# Test and Improvement of the Robustness of Neural Networks



# Robustness against Stickers

- Naturally occurring stickers



- Data Augmentation

Traffic Sign Stickers

**33.8%**

SELF-TRAINED

**27.2%**

PRE-TRAINED



+



=



inserts a single sticker of varying position, size and orientation on the traffic sign



# Naturally Occurring Perturbations as a Challenge for AI



Sun Reflection



Bending



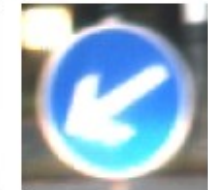
Border Type



Traffic Sign Type



Graffiti



Night Reflection



Background Light



Worn Materials

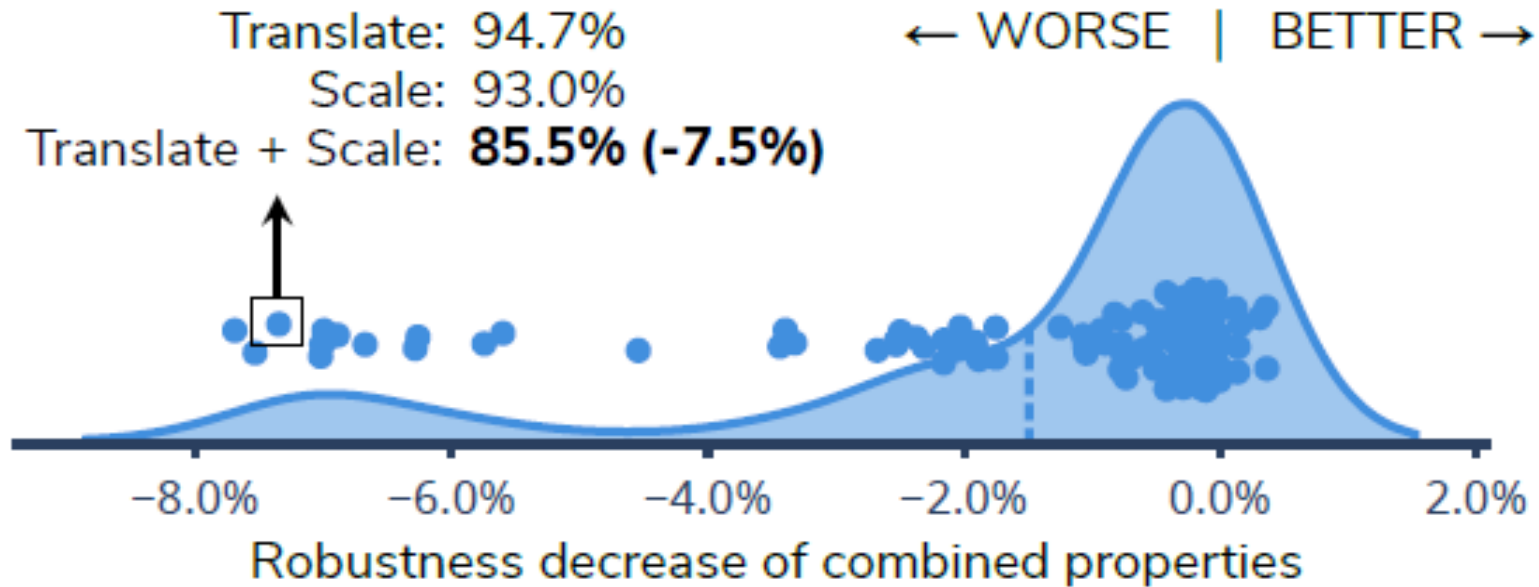


Multiple Signs



Shadows

# Combination of Multiple Perturbations



property order can significantly affect the robustness

Scale + Blur  
Robustness

**81.4%**

vs

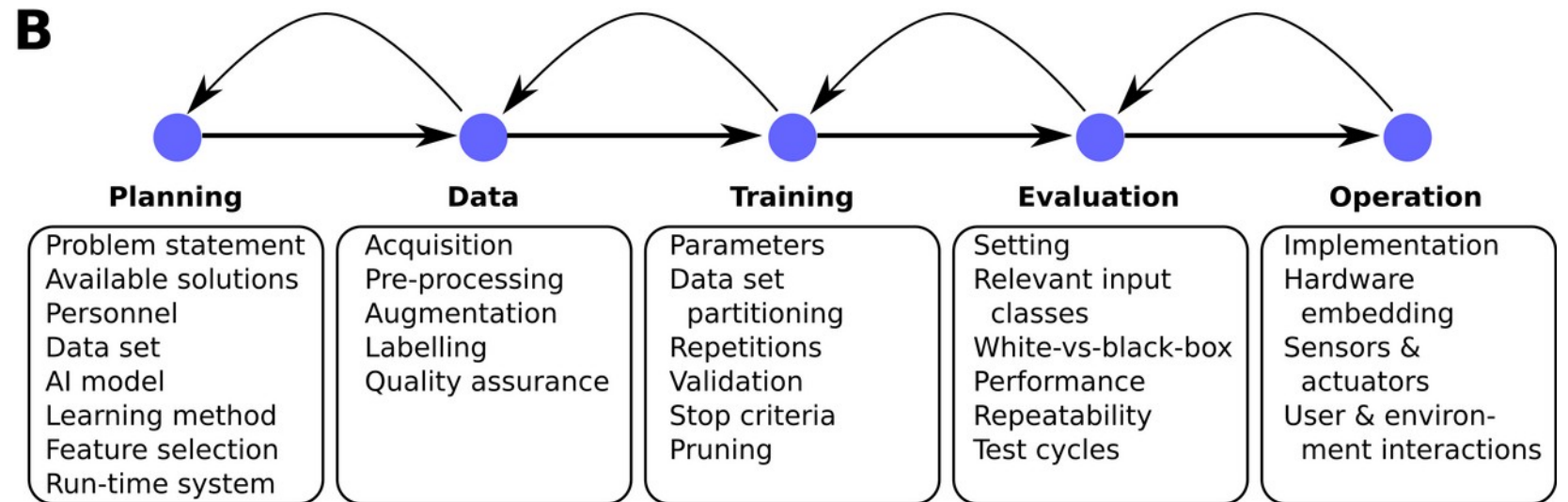
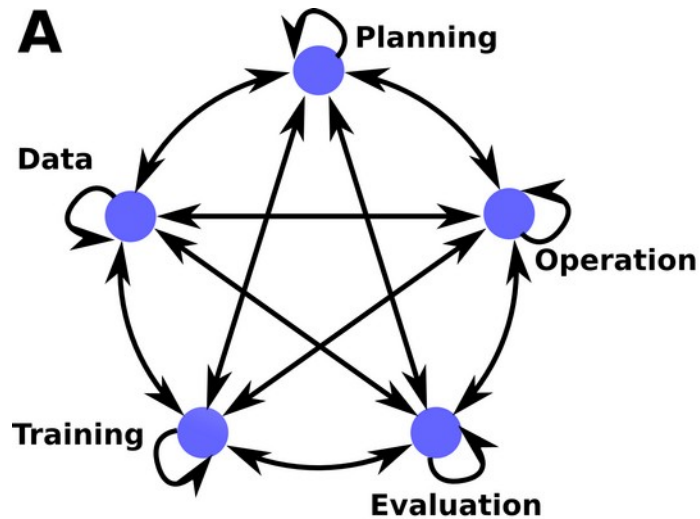
**87.6%**

Blur + Scale  
Robustness

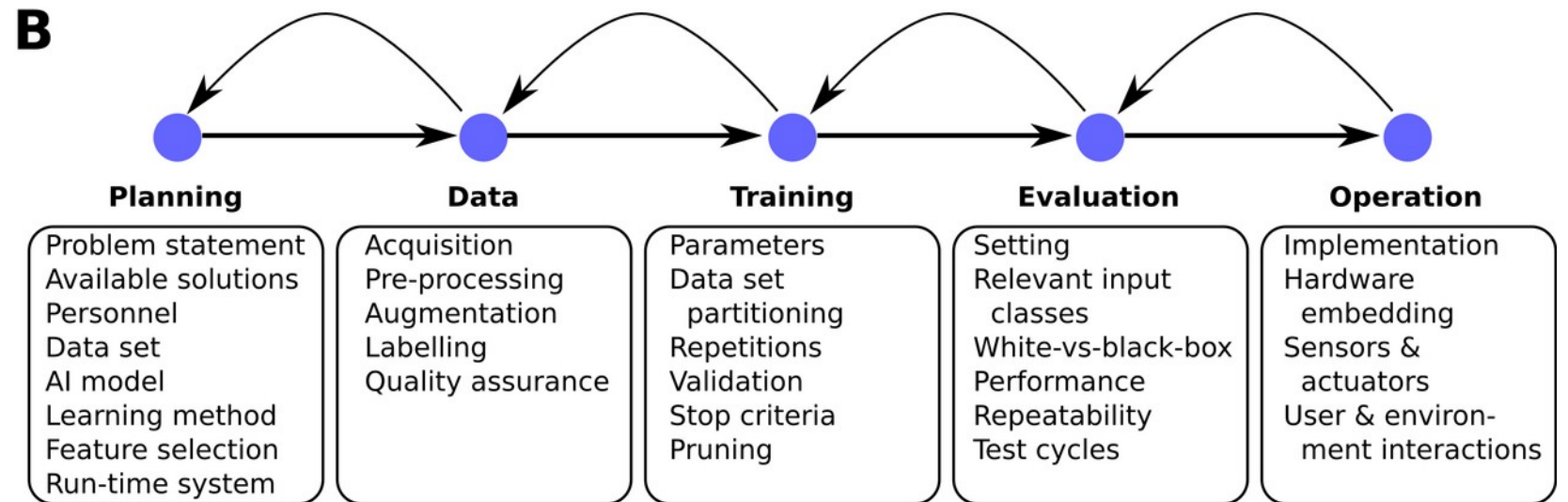
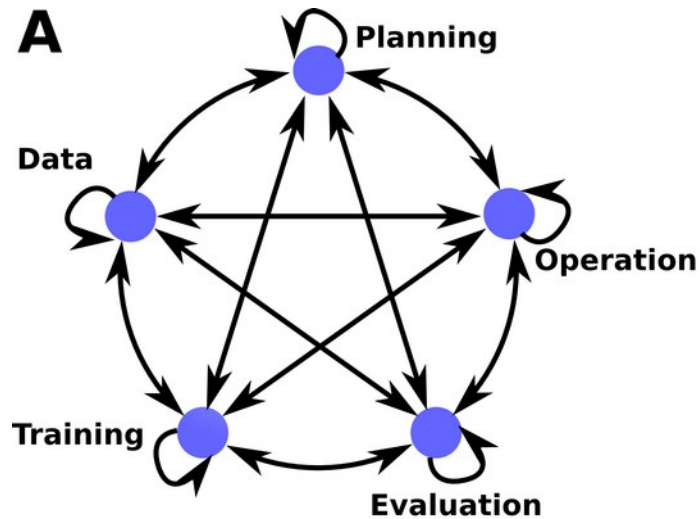
# From the Generalized AI Life Cycle to Application Specific Life Cycles



# The Development of an AI System is an Iterative and Complex Process Which may be Devided Into Phases



# Multiple Views on the AI System Development Process → Formulation of Requirements



- Reliability (treatment of errors)
- IT security
- Acceptance criteria (acceptance vs. Risk)
- Documentation
- ...

# Formulating Requirements:

## The Generalized AI Life Cycle Model is Compatible With and Helpful for the Specific Road Sign Recognition System

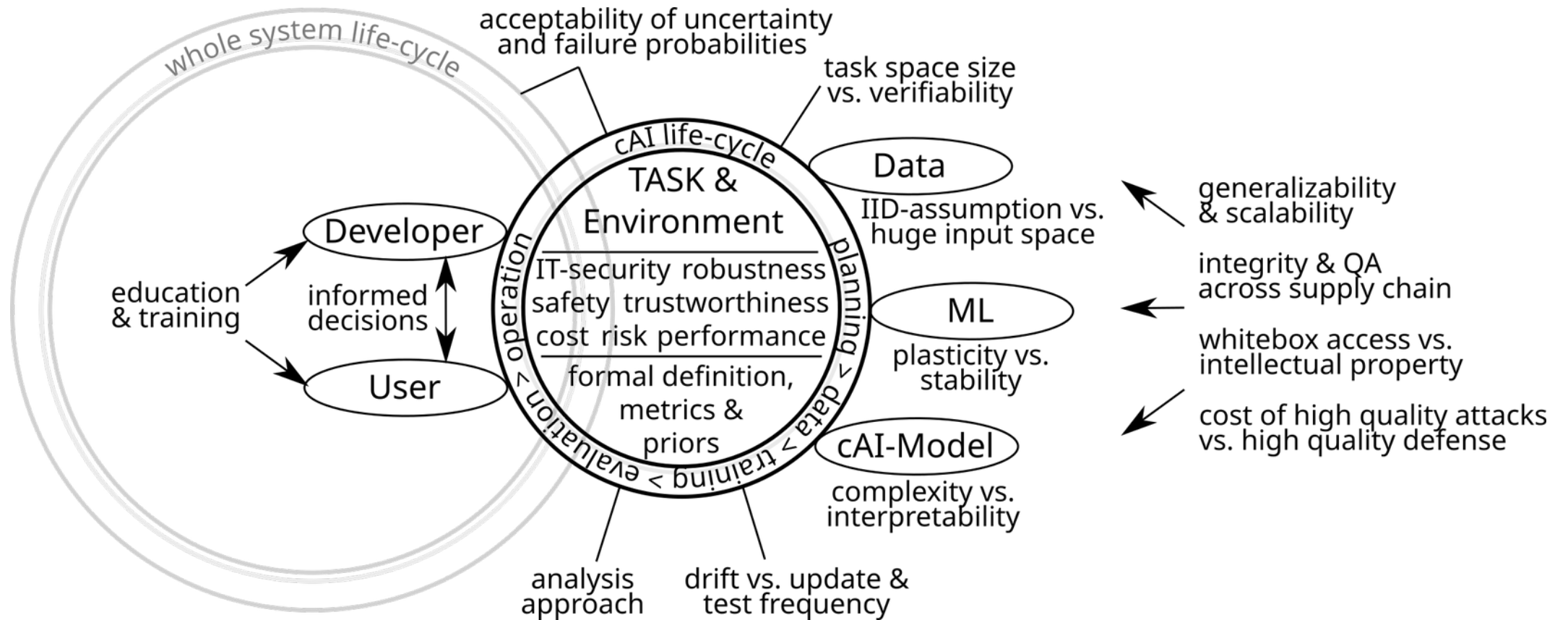
- 1) Where domain knowledge is required, the generalized model has to be concretized
- 2) In some cases if-else decisions are sufficient
- 3) In many cases requirements may be directly transferred from the generalized model

--> as of now, no substantial revision or extension of the generalized model is needed

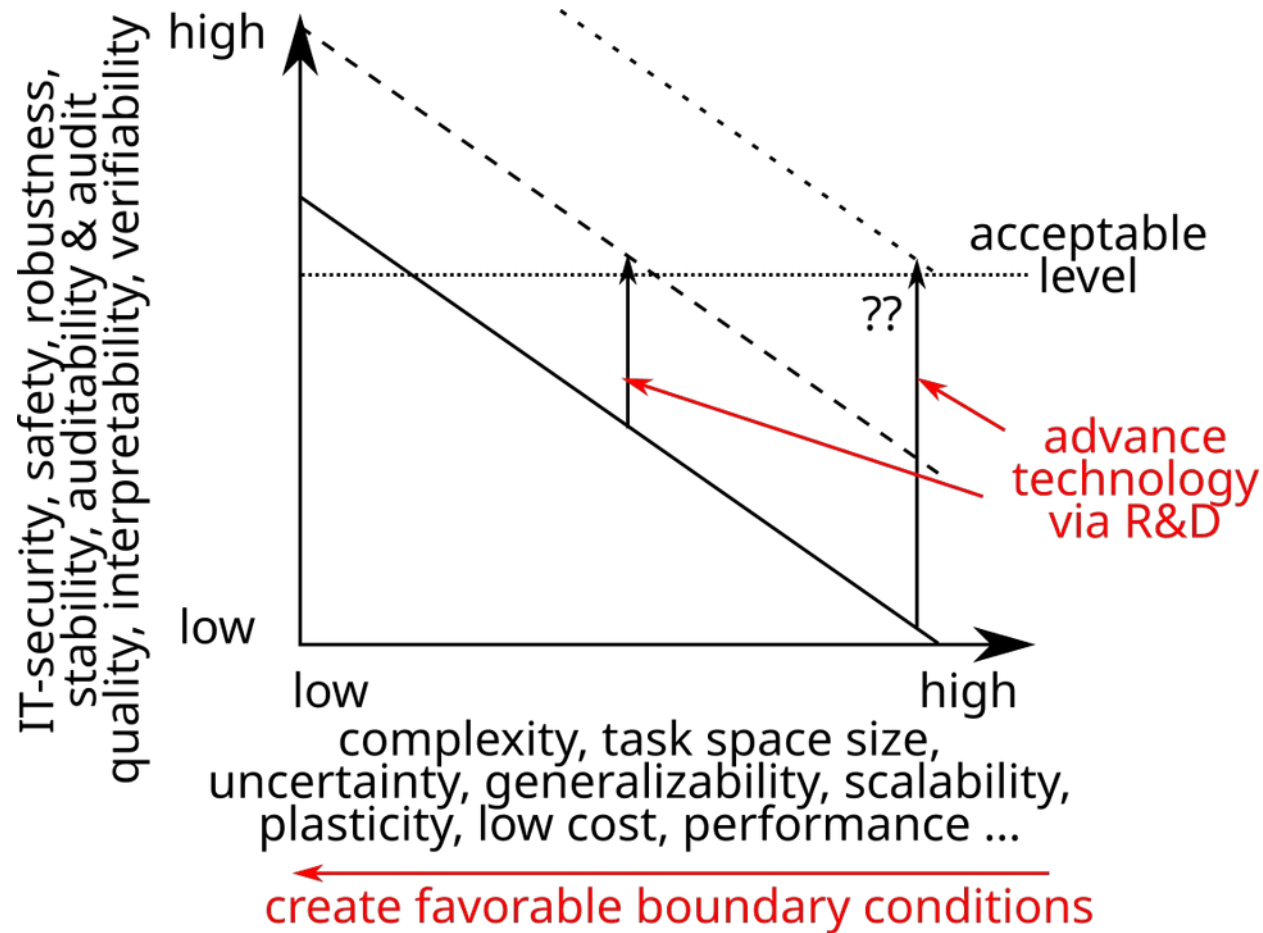
→ multiple use cases have to be examined and compared to verify the generalized model

# Open Challenges

# Open Question in the Context of Auditability, IT security and Safety



# How to Achieve Acceptable Levels of IT Security, Safety, Audit Quality, Robustness and Verifiability?



BSI:

- AI related documents
- involvement in national & international standardization efforts

# BSI Documents on AI Security ([www.bsi.bund.de/KI](http://www.bsi.bund.de/KI))

- **Secure, robust and transparent application of AI Problems, measures and need for action:** presents selected problems as well as measures for security and safety critical applications with regard to so-called connectionist AI methods and shows the need for action
- **AI Cloud Service Compliance Criteria Catalogue (AIC4):** provides AI-specific criteria, which enable an evaluation of the security of an AI service across its lifecycle.
- **Vulnerabilities of Connectionist AI Applications: Evaluation and Defense:** Review of the IT security of connectionist artificial intelligence (AI) applications, focusing on threats to integrity (Frontiers in Big Data)
- **Reliability Assessment of Traffic Sign Classifiers:** evaluates how state-of-the-art techniques for testing neural networks can be used to assess neural networks, identify their failure modes, and gain insights on how to improve them
- **Towards Auditable AI Systems:** Whitepaper with VdTÜV and Fraunhofer HHI based on international workshop in 2020 [to be published soon]



# BSI & AI: Involvement in National & International Cooperations & Standardisation Efforts

## National

- BSI-VdTÜV working group on AI with a focus on mobility and the goal to develop evaluation scenarios for selected use cases until the end of 2021
- Administrative Agreement of BSI with the Kraftfahrtbundesamt (KBA, Federal Motor Transport Authority) in the context of vehicle type approval and cybersecurity
- DIN/DKE Artificial Intelligence Standardization Roadmap
- ...

## International

- ETSI's Industry Specification Group on Securing Artificial Intelligence (ISG SAI)
- ENISA Adhoc working group on AI
- ...

# Thank you for your attention!

## Contact

Dr. Arndt von Twickel

[arndt.twickel@bsi.bund.de](mailto:arndt.twickel@bsi.bund.de)

Bundesamt für Sicherheit in der Informationstechnik (BSI)  
Godesberger Allee 185-189  
53175 Bonn  
Germany

[www.bsi.bund.de](http://www.bsi.bund.de)  
[www.bsi.bund.de/KI](http://www.bsi.bund.de/KI)  
[www.bsi-fuer-buerger.de](http://www.bsi-fuer-buerger.de)

