

fortiss

# Extending the risk-based approach for Artificial Intelligence

---

Why and how  
to rethink  
the risk-based approach  
for new technologies  
(AI/ML/NN)



# NOT IF, BUT HOW

complex AI-based systems  
safety in development & operation

# Motivation

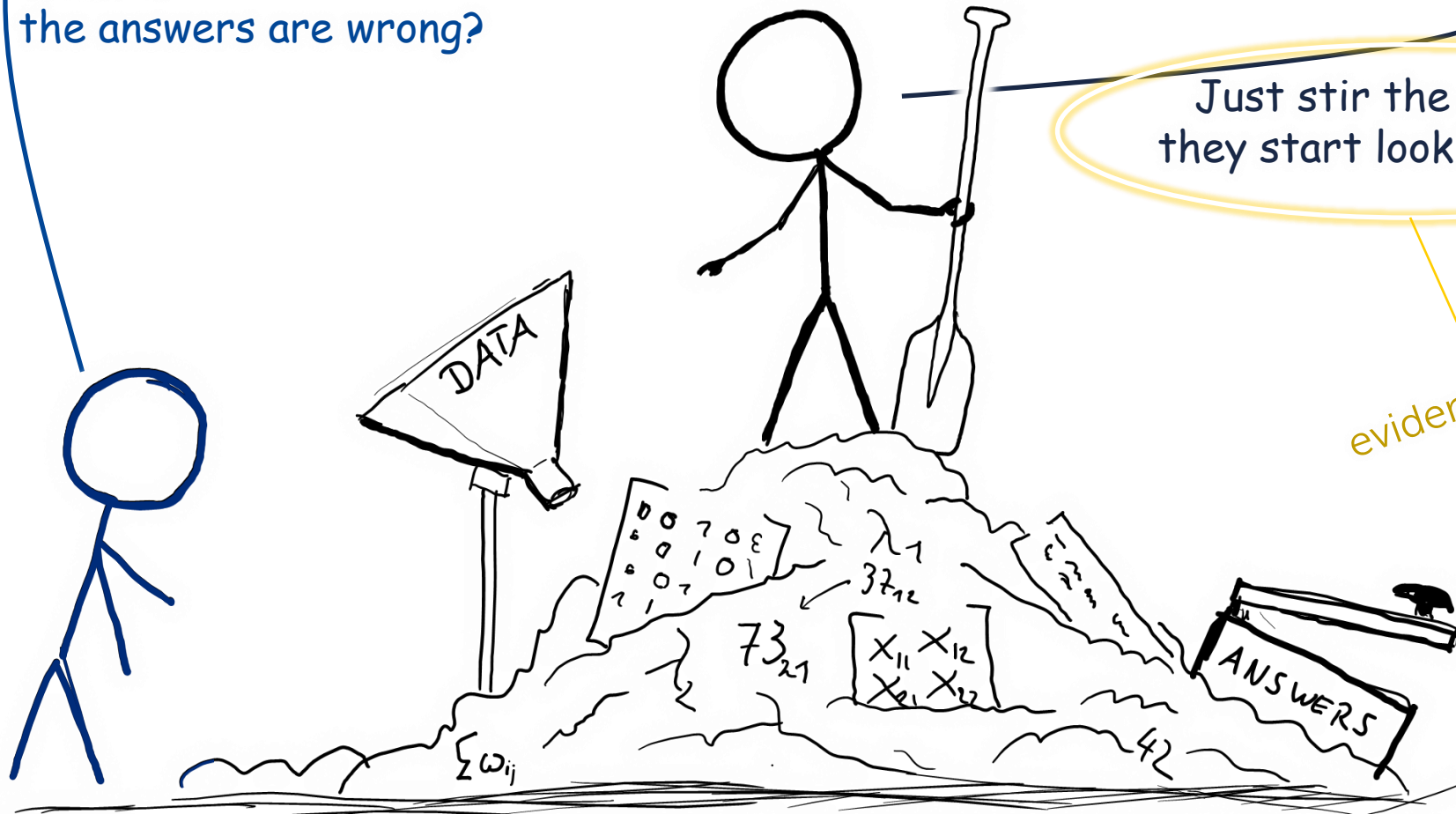
And this is your Machine Learning system?

Yup! You pour data into this big pile of linear algebra, then collect the answers on the other side.

What if the answers are wrong?

Just stir the pile until they start looking right.

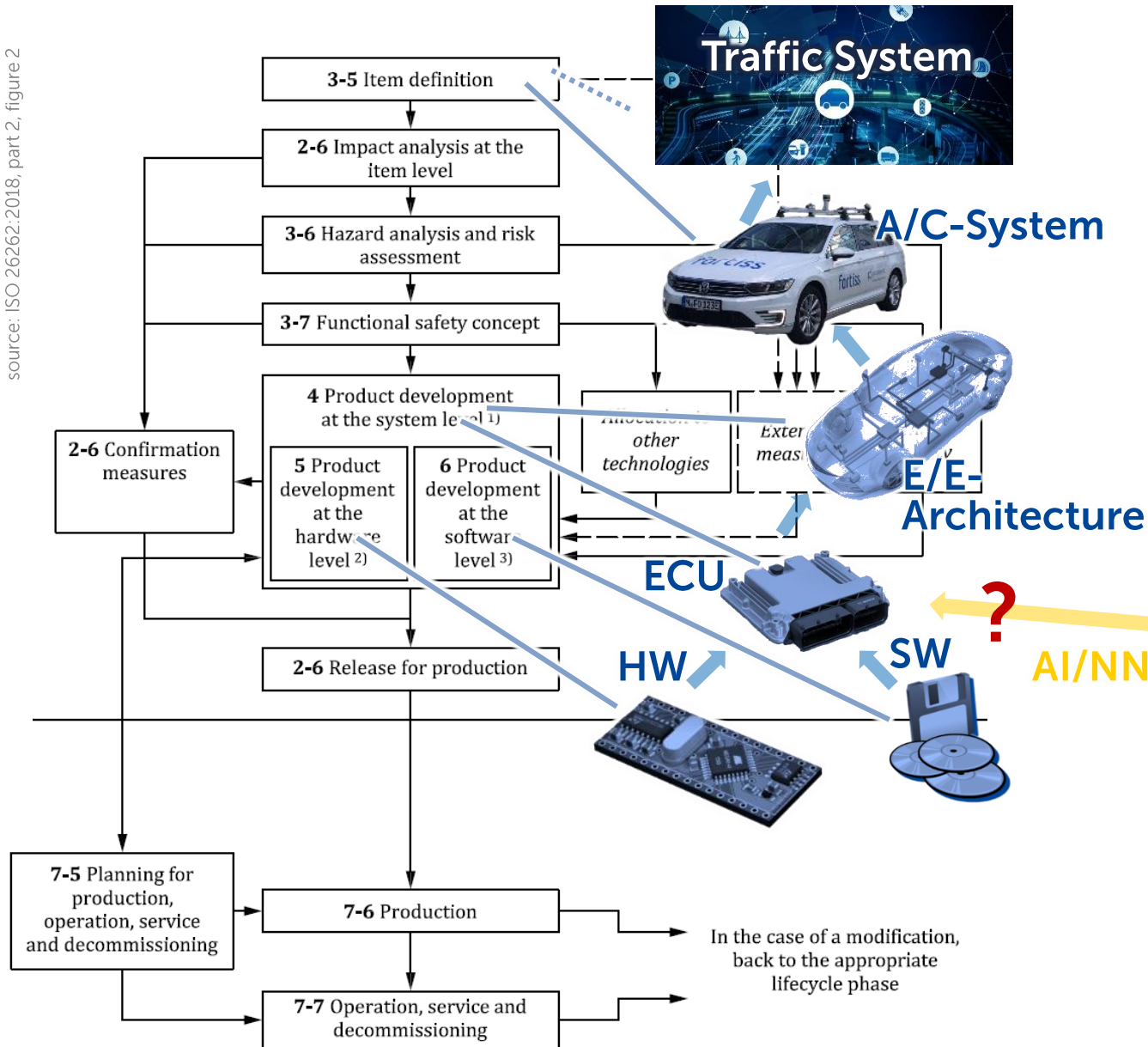
evidential safe?





# The risk-based approach and AI/NN

source: ISO 26262:2018, part 2, figure 2



## risk-based approach

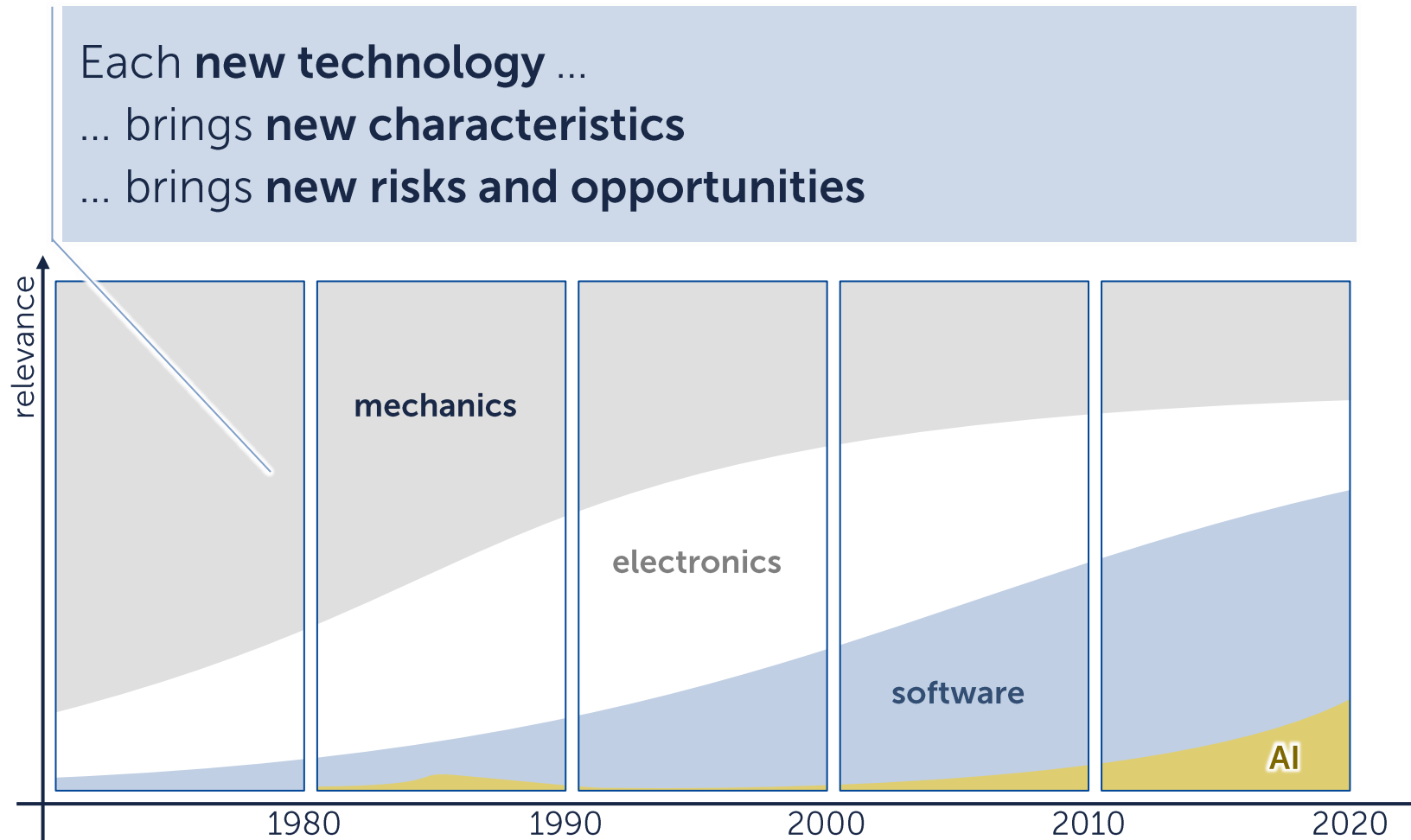
### ► Current standardization

- ISO 26262: risks from systematic failures and random hardware failures
- ISO 21448: risks from intended function (SOTIF)
- ISO 21434: risks from cybersecurity aspects

### ► To be understood

- Characteristics of AI/NN
- Failure modes
- Contribution to safety goal violation
- Integration into risk-based approach

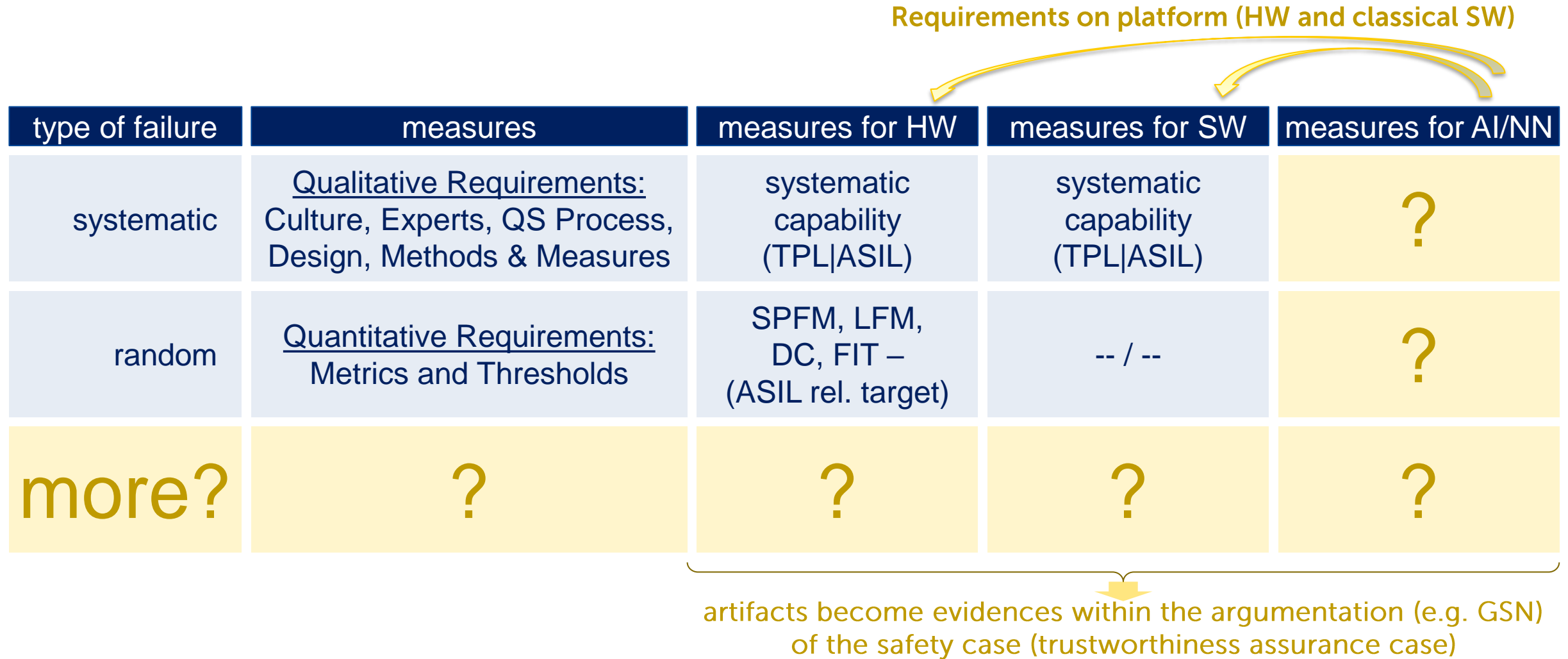
# Each new technology needs a suitable/new approach



► **AI/NN is a new technology** that takes specification and comes up with an **implementation strategy**

# Mapping failures to technology specific measures

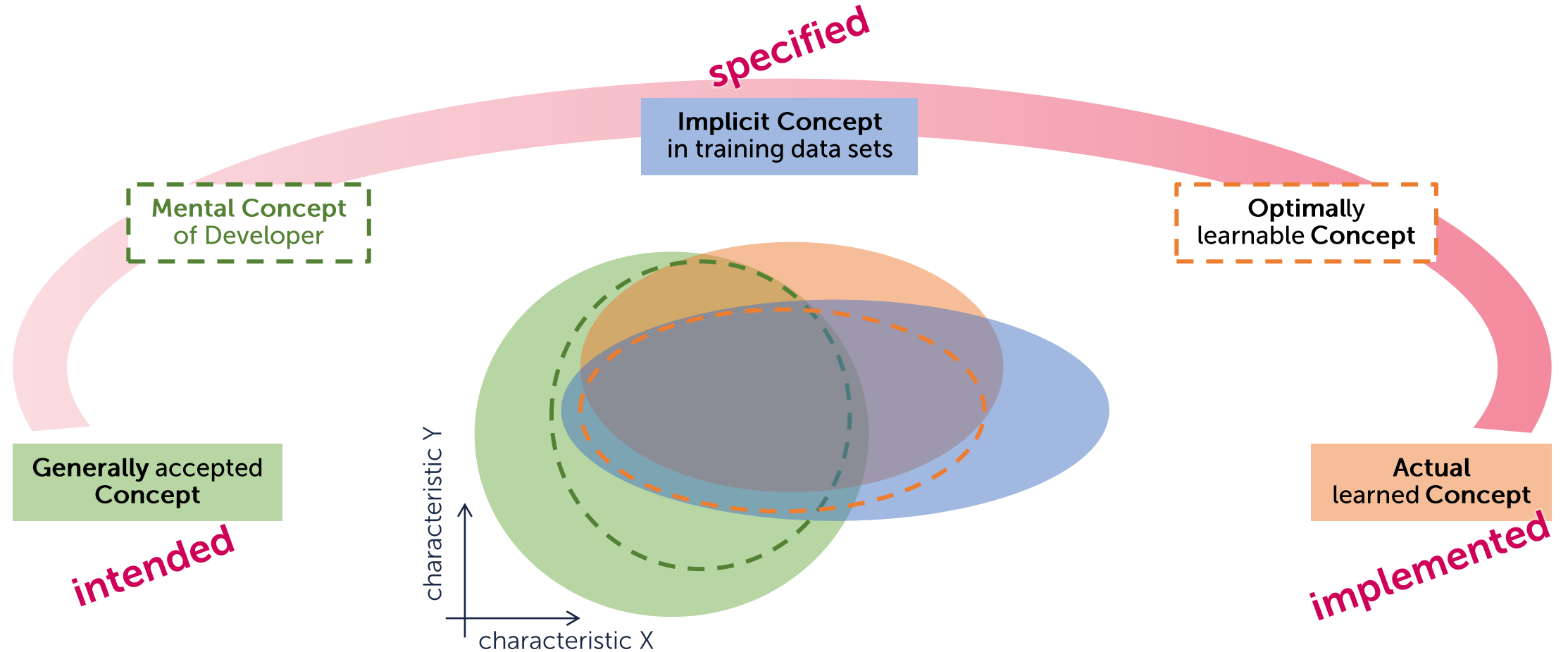
Requirements on platform (HW and classical SW)



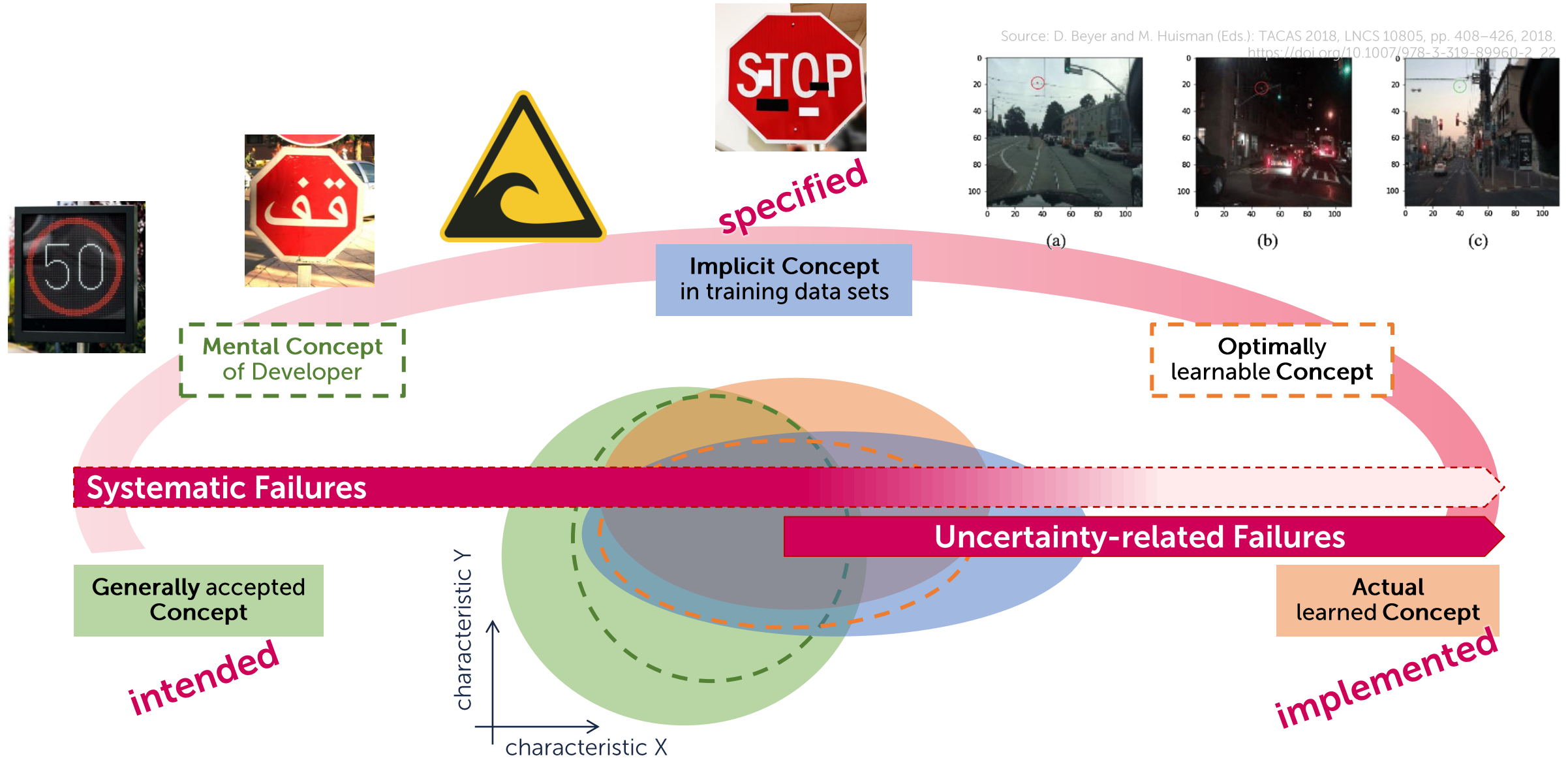
type of failure	measures	measures for HW	measures for SW	measures for AI/NN
systematic	<u>Qualitative Requirements:</u> Culture, Experts, QS Process, Design, Methods & Measures	systematic capability (TPL ASIL)	systematic capability (TPL ASIL)	?
random	<u>Quantitative Requirements:</u> Metrics and Thresholds	SPFM, LFM, DC, FIT – (ASIL rel. target)	-- / --	?
more?	?	?	?	?

artifacts become evidences within the argumentation (e.g. GSN)  
of the safety case (trustworthiness assurance case)

# Insights on Deep Neural Networks - phases and activities



# Insights on Deep Neural Networks - problems



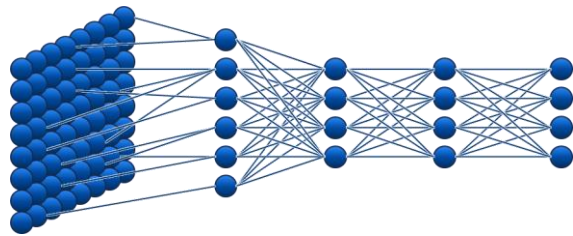


# Un-fevered facts on DNNs



no mysticism but  
engineering automation  
(**NN engineering**)

performance goals  
hard to determine  
(**ethics**)



▶ DOG

▶ MUFFIN



output for new input  
not predictable  
(**uncertainty**)

same input always  
leads to same output  
(**determinism**)

Did we train  
all relevant inputs?  
(**bias/coverage/domain-shift**)

limitation of NN not obvious  
(**generalization/brittleness**)

strategy of NN not obvious  
(**opacity**)



▶ (D)NNs are deterministic ~ but there is uncertainty

## Mapping failures to technology specific measures (2)

Requirements on platform (HW and classical SW)

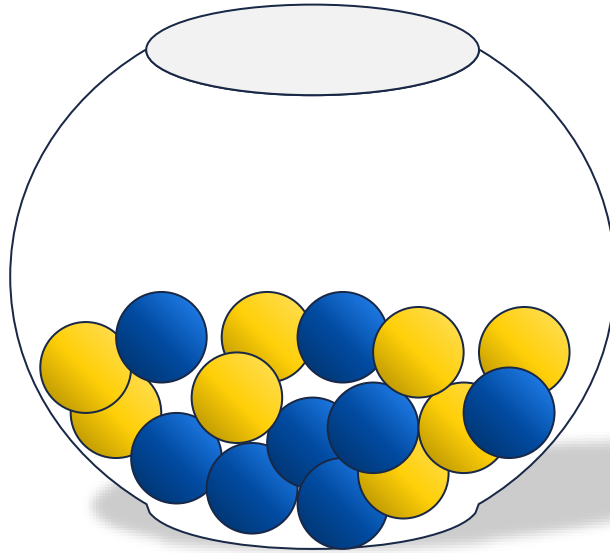
type of failure	measures	measures for HW	measures for SW	measures for AI/NN
systematic	<u>Qualitative Requirements:</u> Culture, Experts, QS Process, Design, Methods & Measures	systematic capability (TPL ASIL)	systematic capability (TPL ASIL)	?
random	<u>Quantitative Requirements:</u> Metrics and Thresholds	SPFM, LFM, DC, FIT – (ASIL rel. target)	-- / --	-- / --
uncertainty- related	<u>Structured Approach:</u> Metrics, References, Measures and Argumentation	-- / --	-- / --	?

artifacts become evidences within the argumentation (e.g. GSN)  
of the safety case (trustworthiness assurance case)

# Uncertainty thought experiment

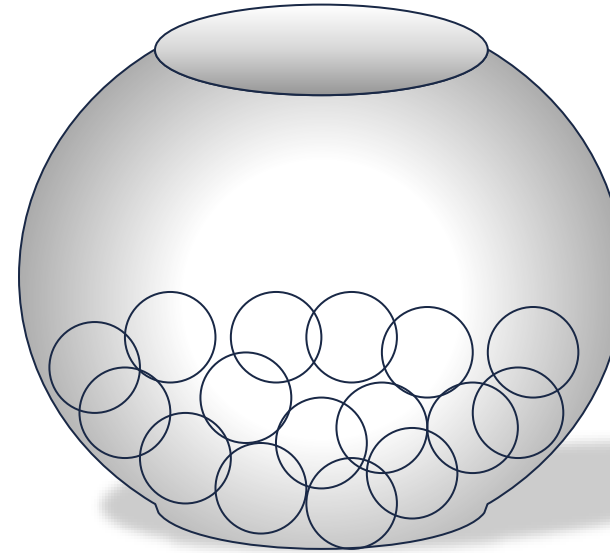
## ***probabilism (risk)***

unknown result  
but information  
about likelihood



## ***no information (uncertainty)***

unknown result and  
unknown likelihood



Ref.: according to Ellsberg, 1961:  
*Risk, Ambiguity, and the Savage Axioms*

*"Not knowing the chance of mutually exclusive events and knowing the chance to be equal are two quite different states of knowledge"*

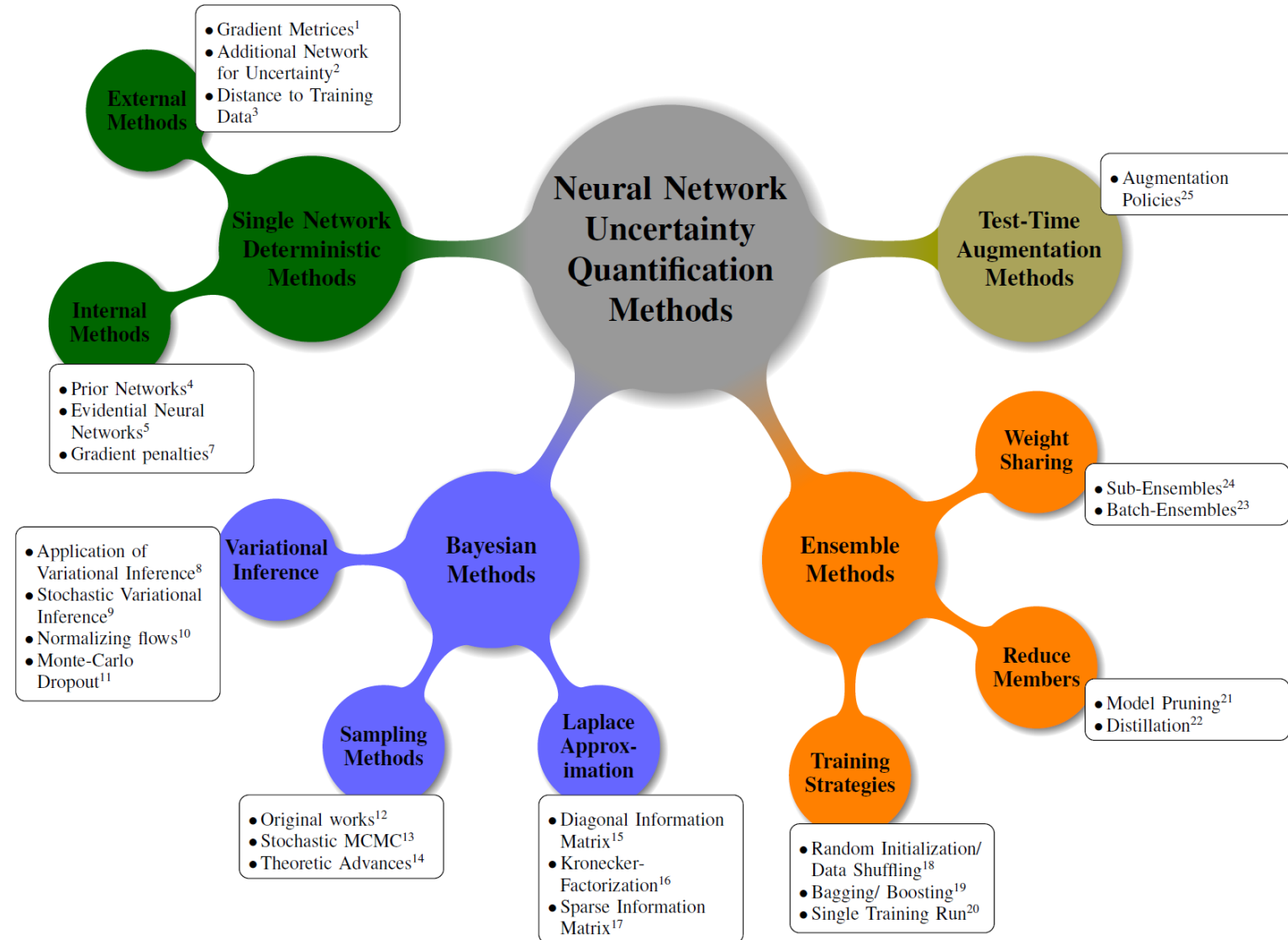
Ronald Fisher (1890-1962)



► AI/NN as new technology induces uncertainty-related effects (risks)



# Quantifying uncertainty in NNs - research



All models are wrong  
~ some are useful.

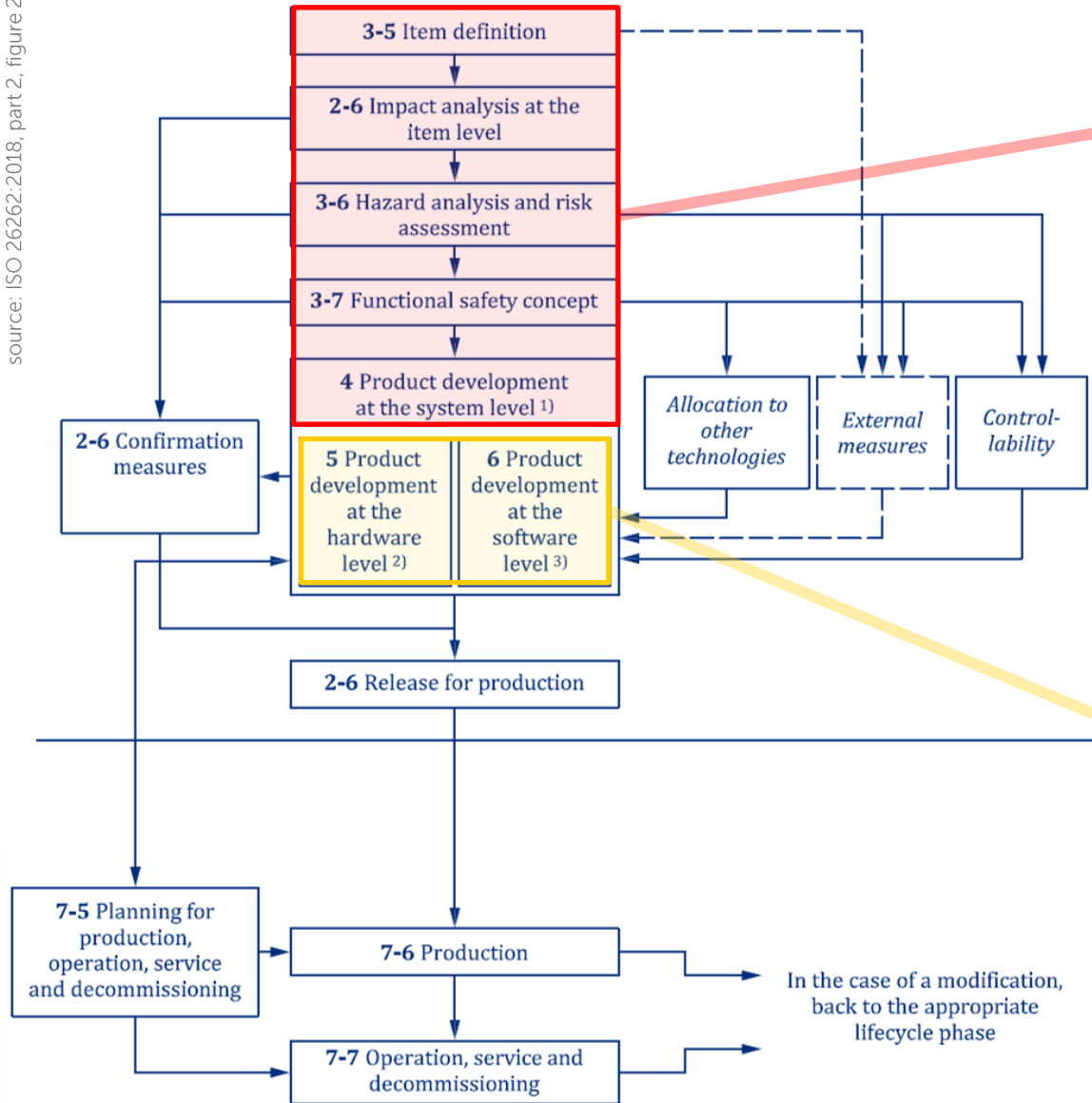
(George Box)

source: J.Gawlikowski et al  
A Survey of Uncertainty  
in Deep Neural Networks,  
arXiv:2107.03342v3  
(figure 3)

► Quantifying the uncertainty in NNs is a **matter of research**

# Quantifying uncertainty in NNs – available approach

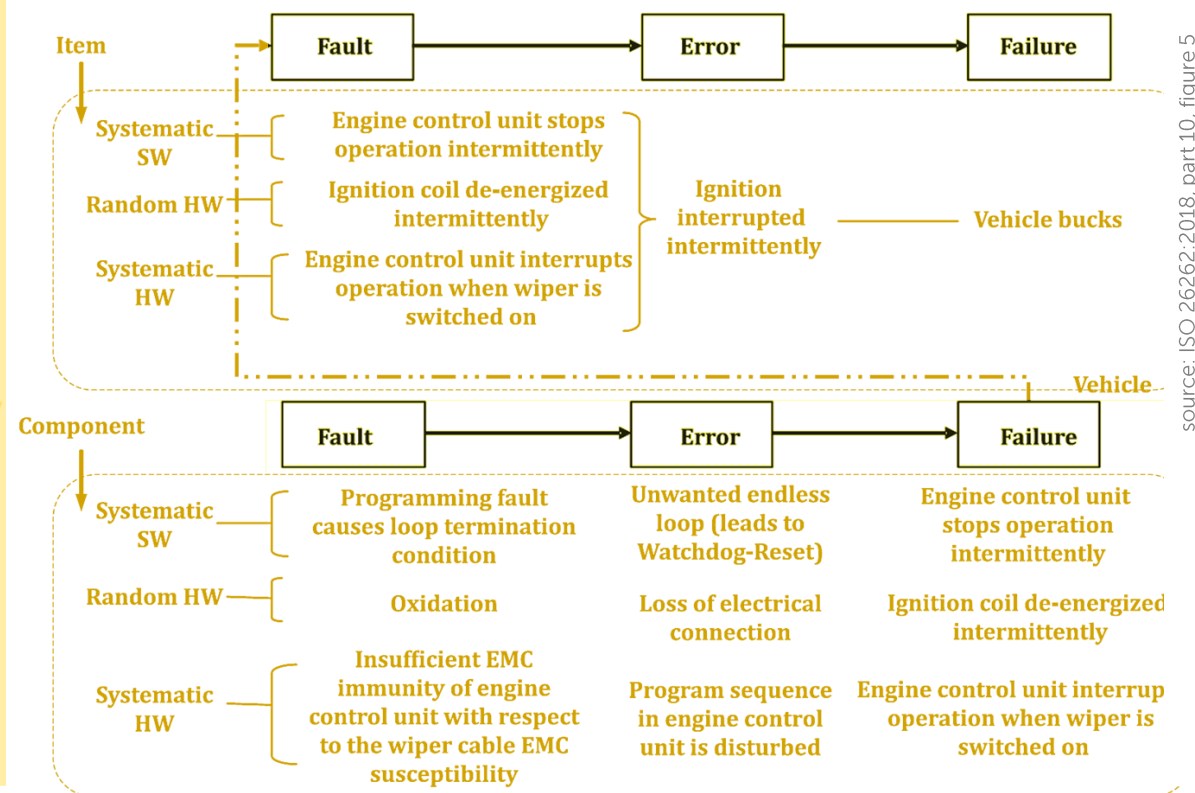
source: ISO 26262:2018, part 2, figure 2



## traceability

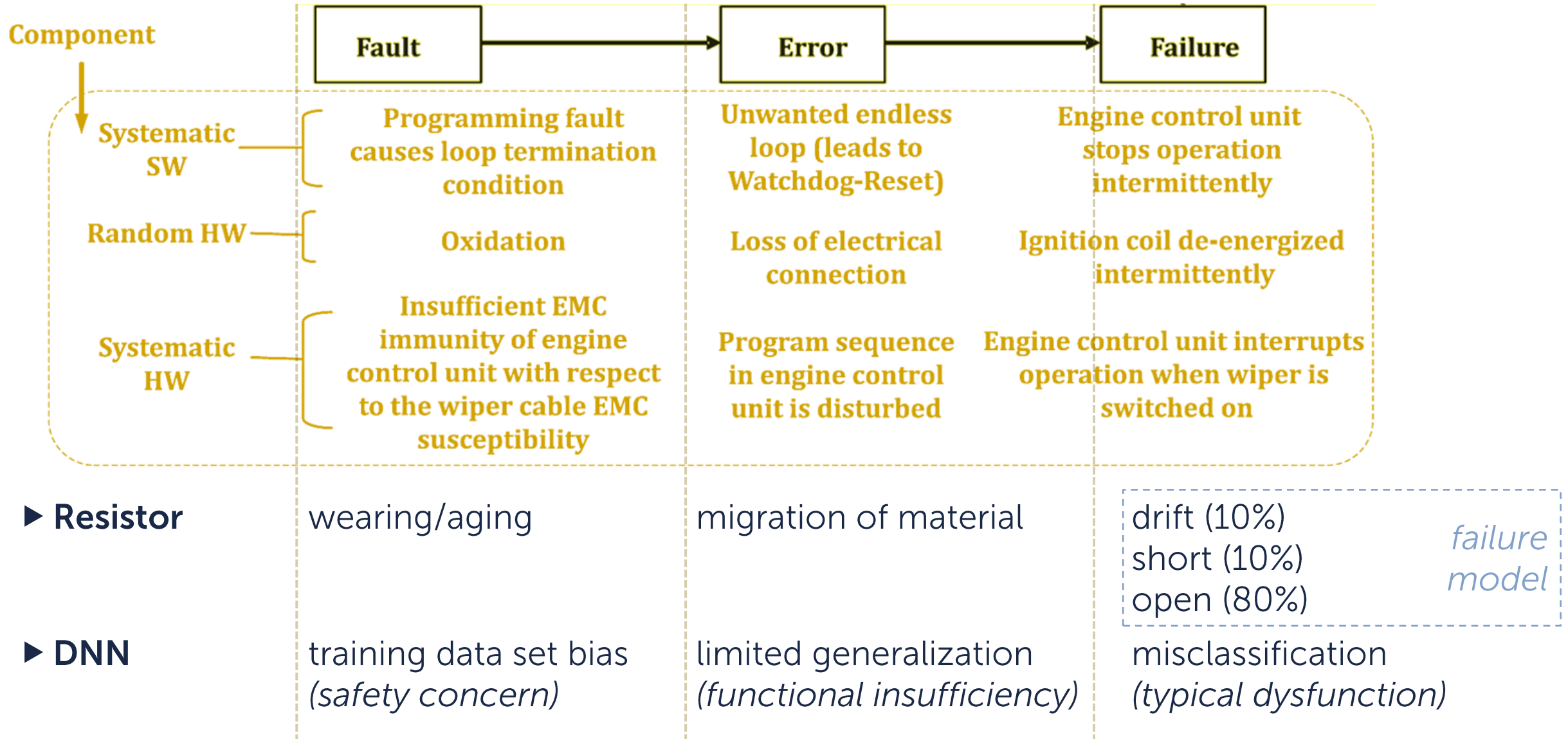
- Hazard Analysis & Risk Assessment (HARA)
- Functional Safety Concept (FSC)
- Safety Functions (+ attributes and requirements)

## fault ~ error ~ failure



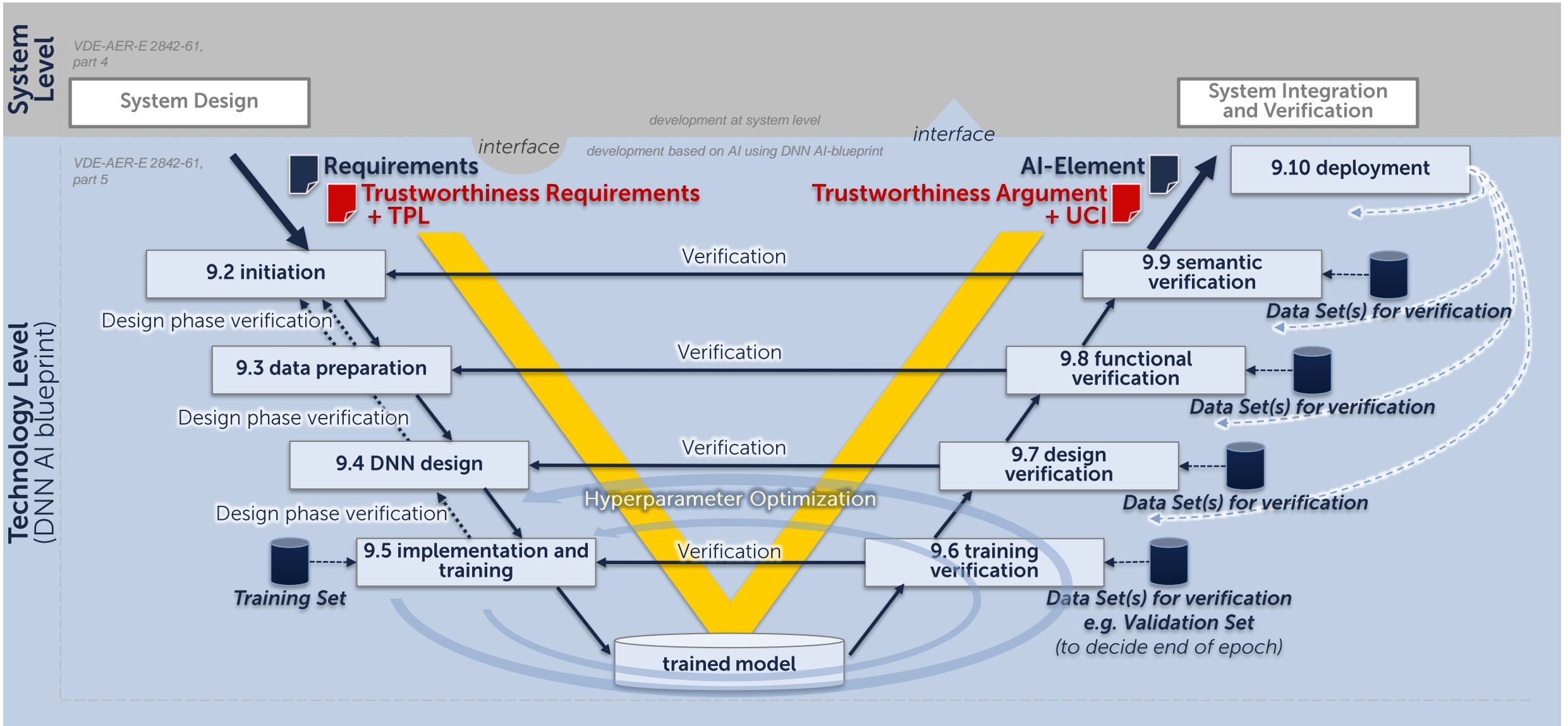
source: ISO 26262:2018, part 10, figure 5

# Quantifying uncertainty in NNs – fault ~ error ~ failure

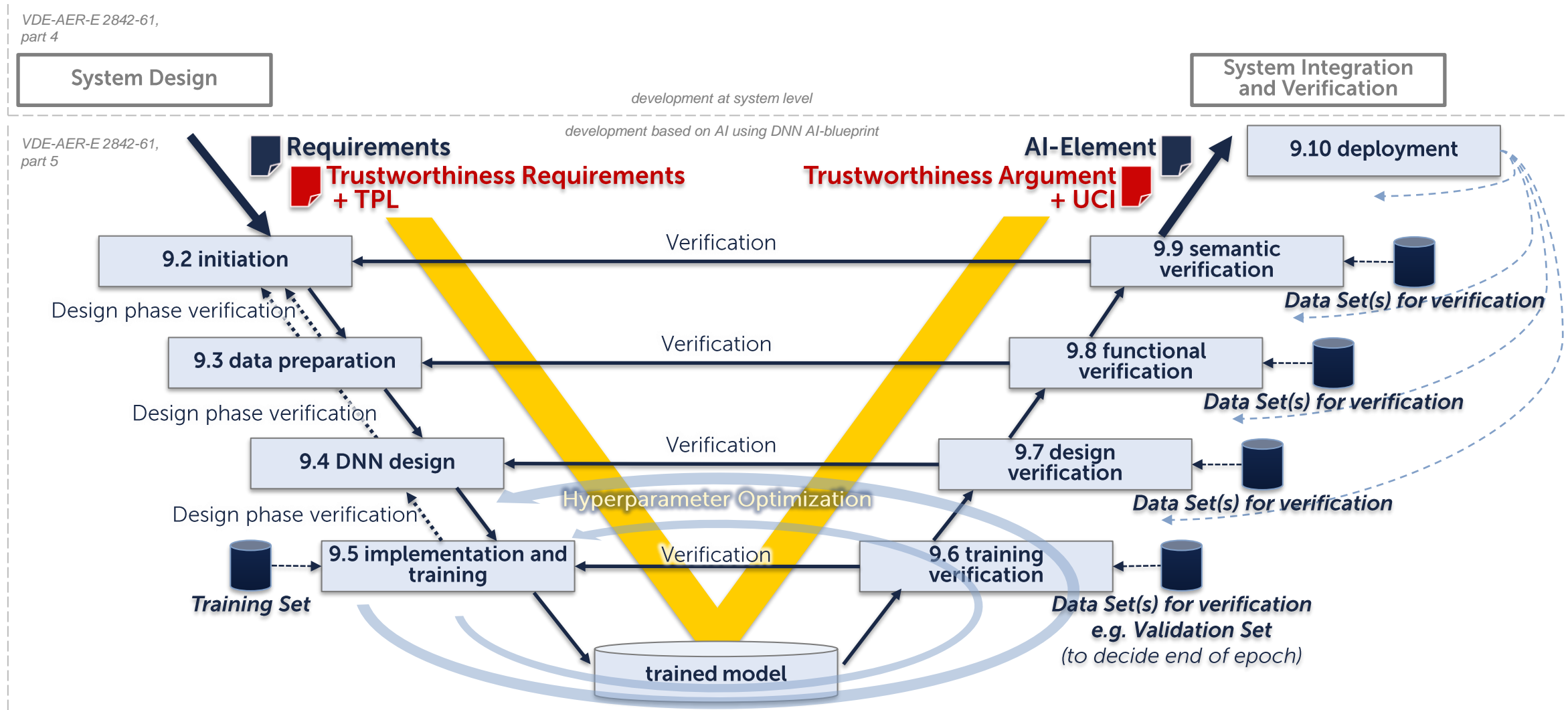


► Any **structured approach** to a failure model of AI/NN?

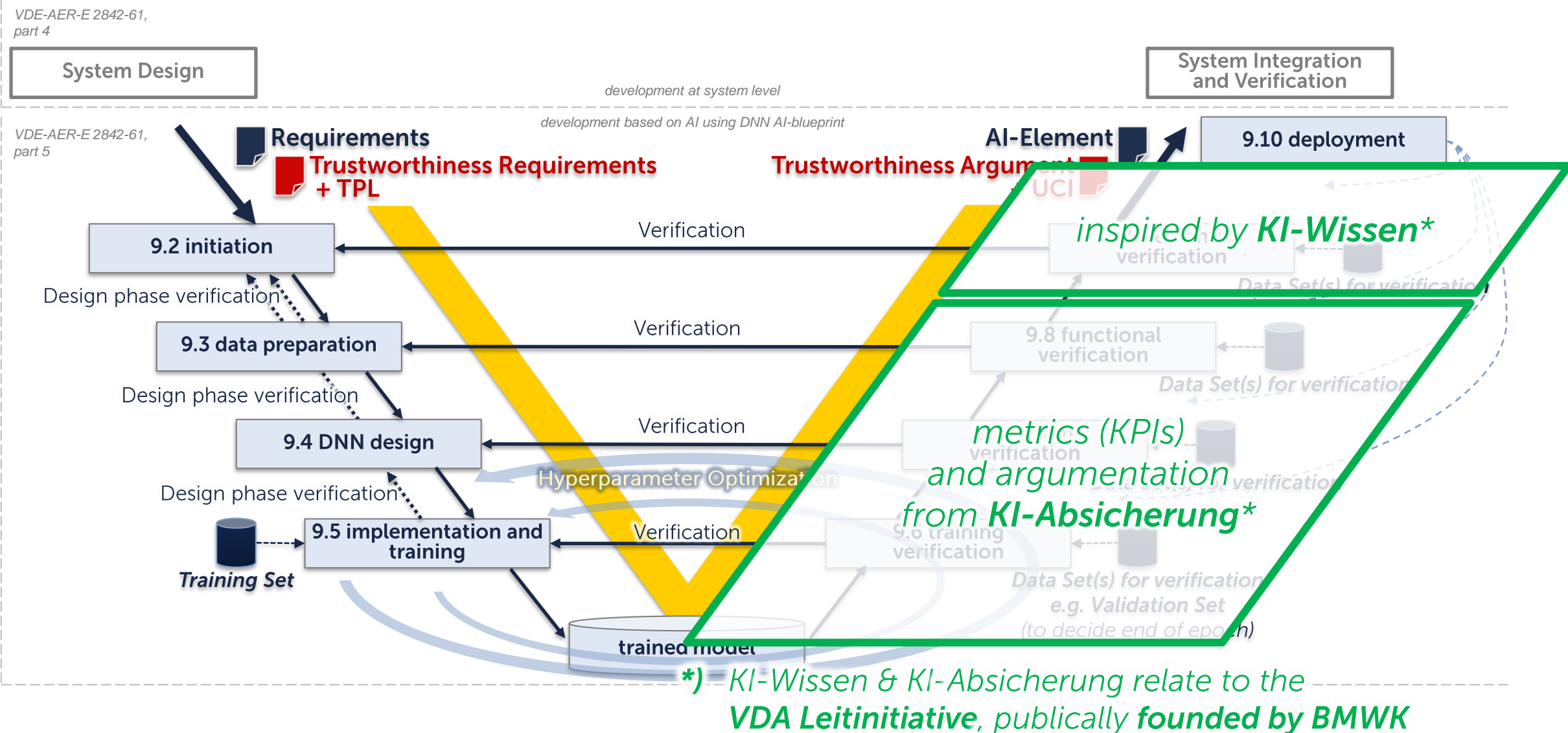
# Structured development of DNNs – DNN AI blueprint acc. to VDE-AR-E 2842-61



# Structured development of DNNs – DNN AI blueprint acc. to VDE-AR-E 2842-61



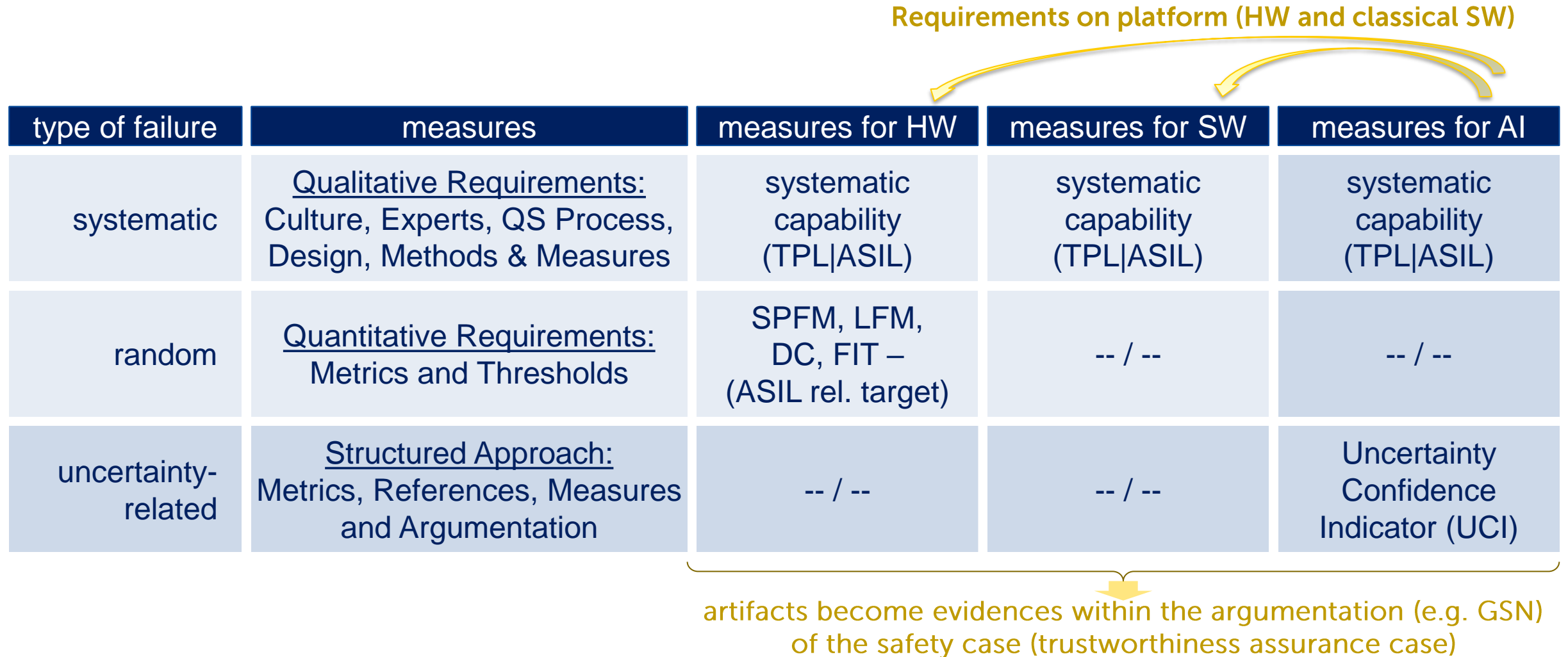
## Structured development of DNNs – based on latest research & VDA Leitinitiative





# Mapping failures to technology specific measures (final)

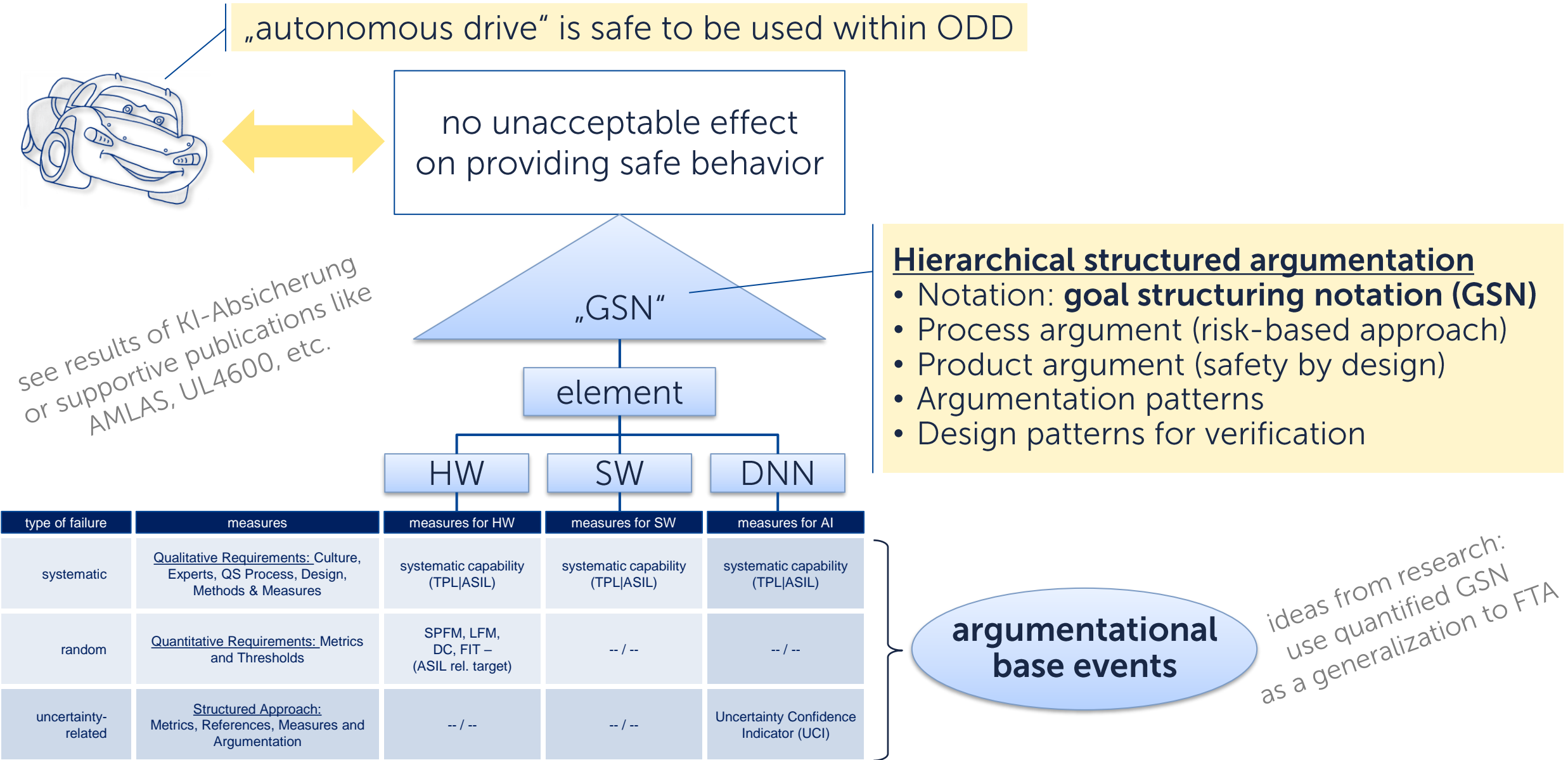
Requirements on platform (HW and classical SW)



type of failure	measures	measures for HW	measures for SW	measures for AI
systematic	<u>Qualitative Requirements:</u> Culture, Experts, QS Process, Design, Methods & Measures	systematic capability (TPL ASIL)	systematic capability (TPL ASIL)	systematic capability (TPL ASIL)
random	<u>Quantitative Requirements:</u> Metrics and Thresholds	SPFM, LFM, DC, FIT – (ASIL rel. target)	-- / --	-- / --
uncertainty- related	<u>Structured Approach:</u> Metrics, References, Measures and Argumentation	-- / --	-- / --	Uncertainty Confidence Indicator (UCI)

artifacts become evidences within the argumentation (e.g. GSN)  
of the safety case (trustworthiness assurance case)

# The assurance case integrates all evidences into





# Insights on Deep Neural Networks: No Free Lunch

PRO  
(mise)

CONTRA  
(indication)

new technology

implicit  
requirements

shorter  
time to market

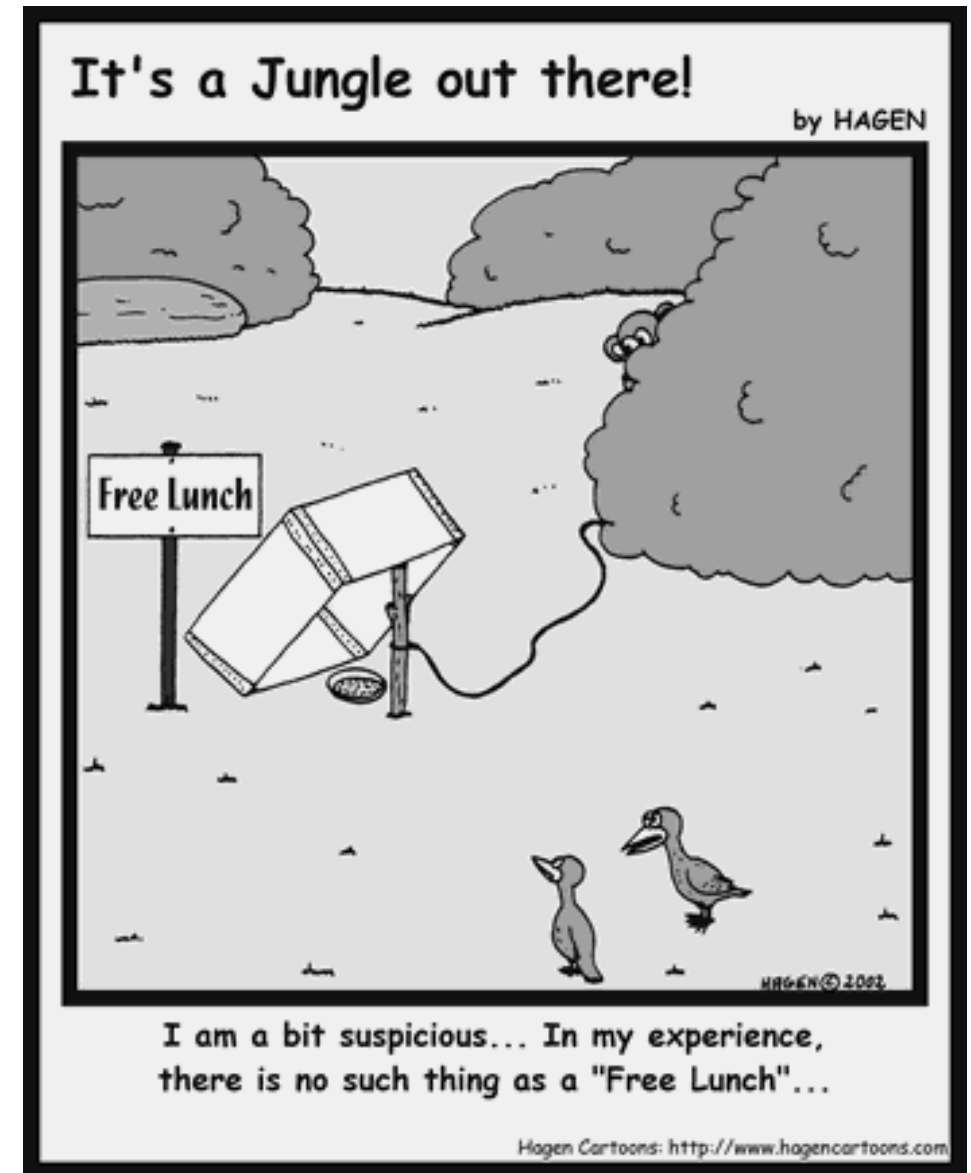
enable complex  
automation

No-Free-Lunch Theorem:

$$\sum_f P(h_m^y | f, m, a_1) = \sum_f P(h_m^y | f, m, a_2).$$

Reality: no free lunch!

Free Lunch?



# Insights on Deep Neural Networks: **No Free Lunch**

PRO  
(mise)

CONTRA  
(indication)

new technology

implicit  
requirements

shorter  
time to market

enable complex  
automation

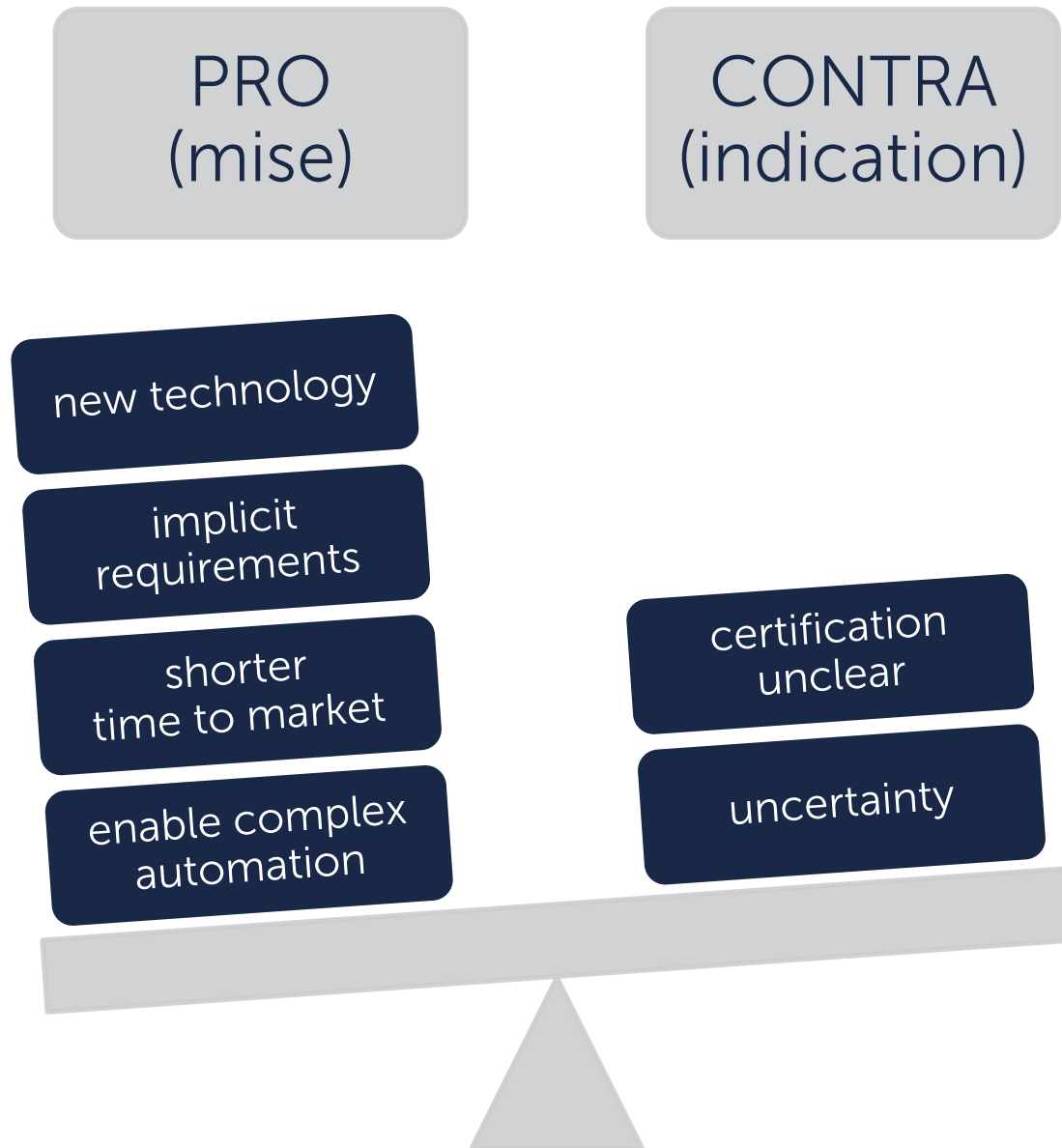
uncertainty

## ► **AI/ML/NN are new technologies**

- no mysticism,
- but automation engineering with
- new characteristics

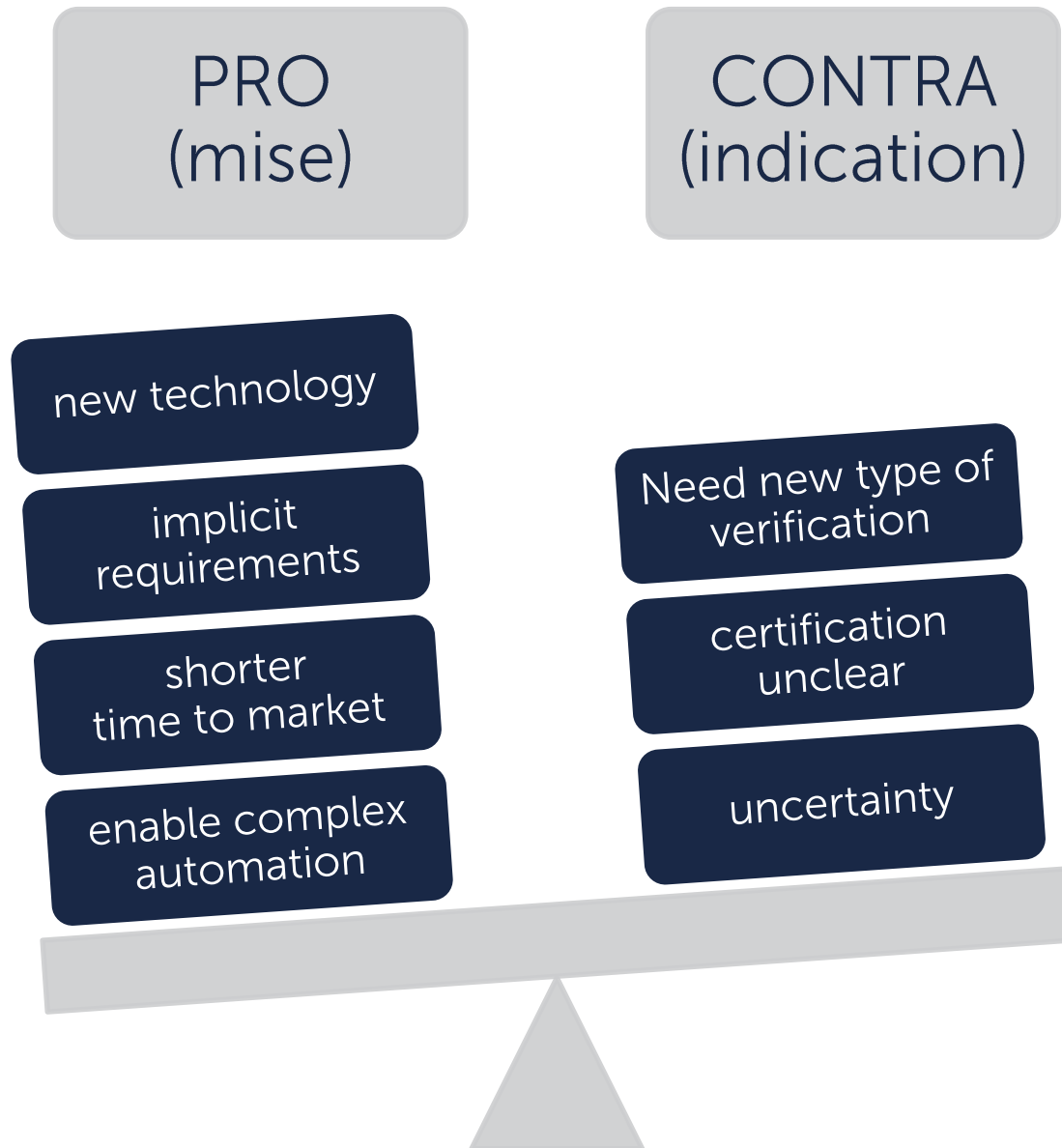
## ► **There is no free lunch**

# Insights on Deep Neural Networks: **No Free Lunch**



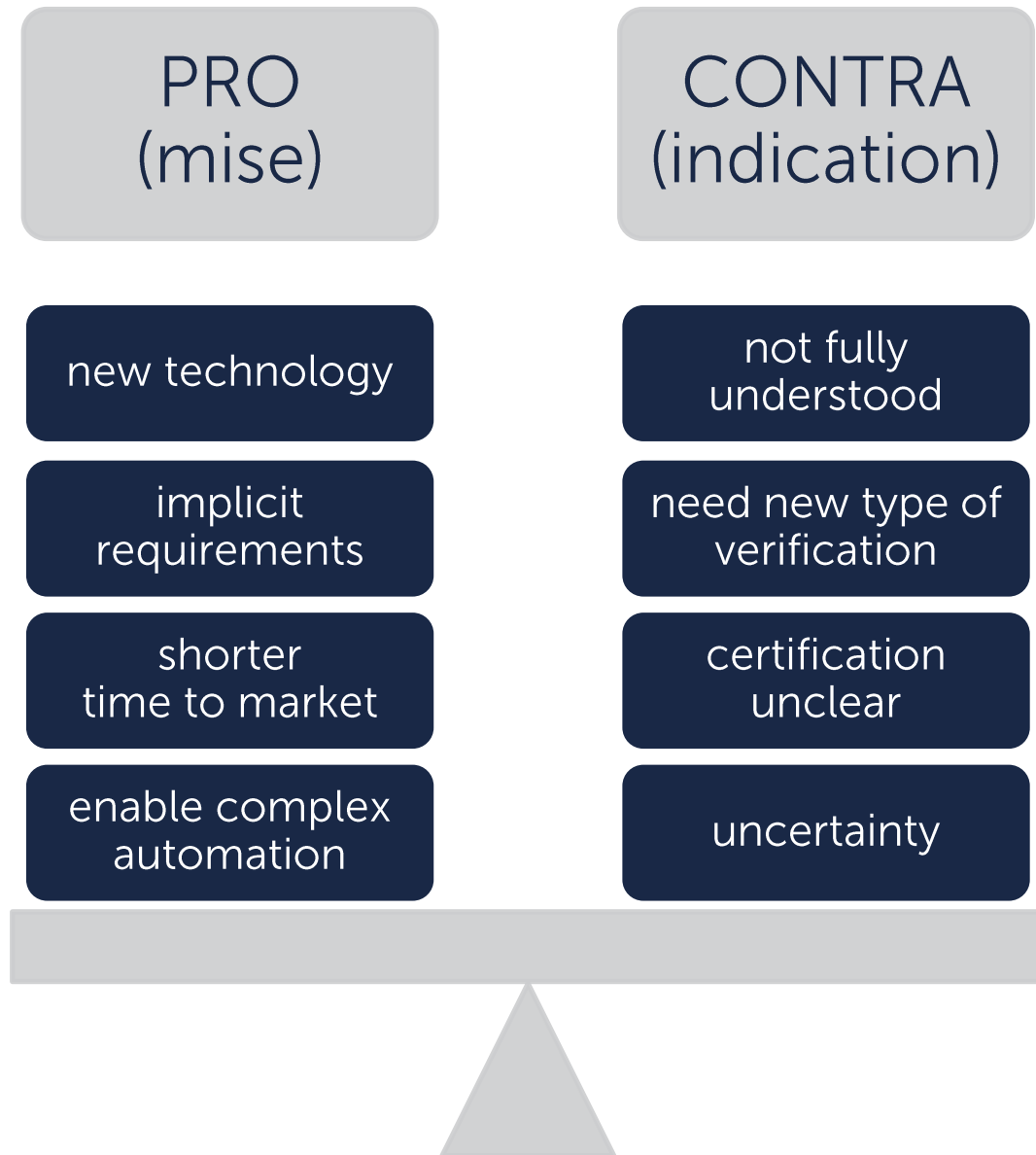
- ▶ **AI/ML/NN are new technologies**
  - no mysticism,
  - but automation engineering with
  - new characteristics
- ▶ **There is no free lunch**

# Insights on Deep Neural Networks: **No Free Lunch**



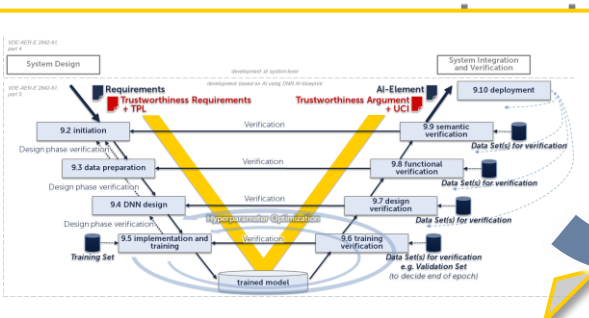
- ▶ **AI/ML/NN are new technologies**
  - no mysticism,
  - but automation engineering with
  - new characteristics
- ▶ **There is no free lunch**

# Insights on Deep Neural Networks: **No Free Lunch**



- ▶ **AI/ML/NN are new technologies**
  - no mysticism,
  - but automation engineering with
  - new characteristics
- ▶ **There is no free lunch**  
**New technologies need ...**
  - new development methods
  - new **processes**
  - new **standards**
  - new risk management
  - new governance structures
  - new **certification & maturity**
  - suitable AI strategy

source: ISO 26262:2018, part 2, figure 2



# The risk-based approach and AI/NN

## Solution level

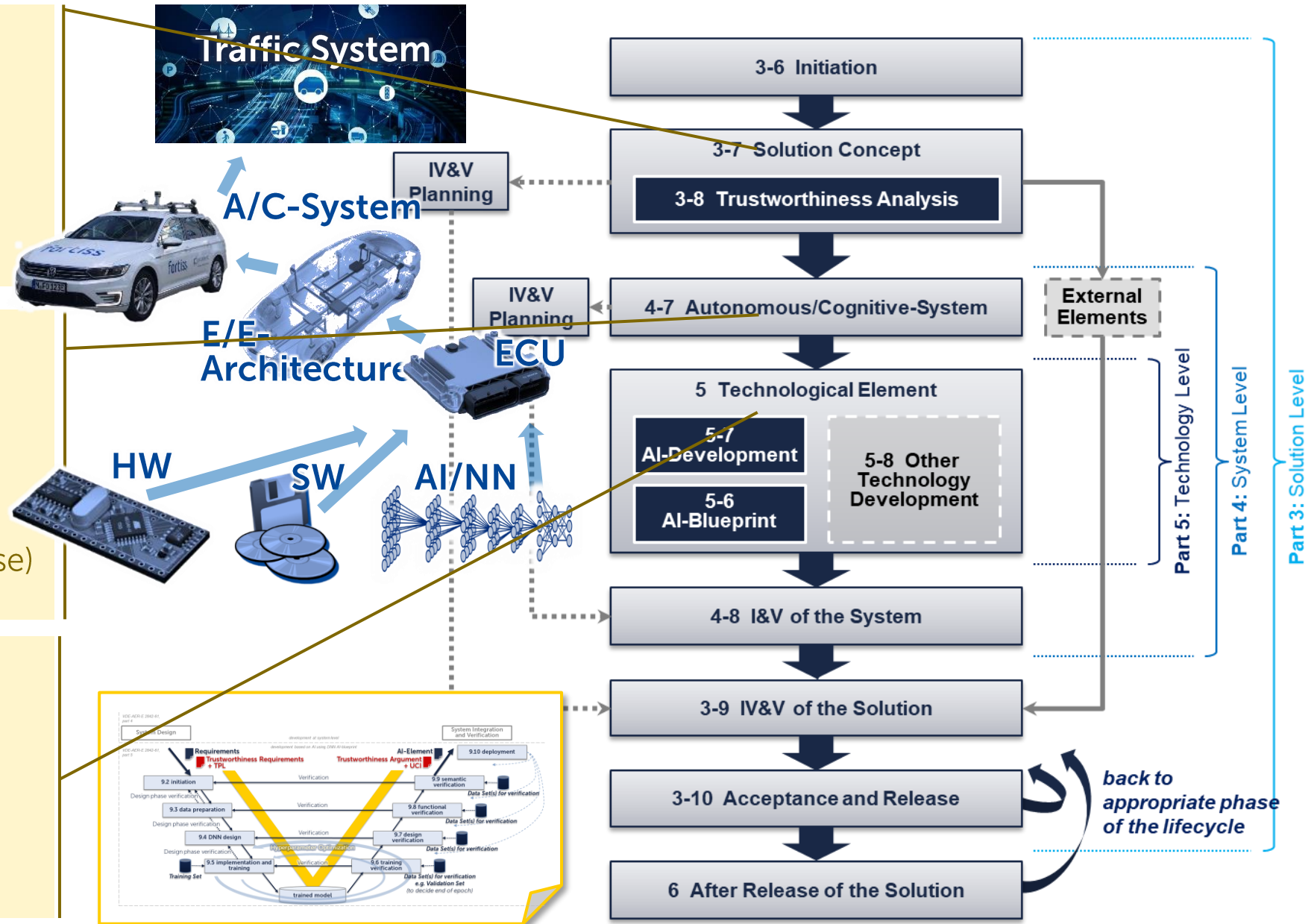
- Origin of all hazards
- Consider dysfunction of overall system (car) in environment
- Trustworthiness as meta term (safety, cybersecurity, usability...)

## System level

- Traceability of ASIL (TPL)
- Design patterns
- Continuous refinement of safety case (= trustworthiness assurance case)

## Technology level

- AI/NN as new technology
- V-model for DNNs
- Consider uncertainty-related failures





## Trustworthy Autonomous/Cognitive Systems (A/C-System)

*independent from industry & application*

*separate ethical/moral aspects from technology*

**risk-based approach** along lifecycle

*complex system-of-systems*

Performance und Trustworthiness

Safety  
Security  
Usability  
Ethics  
[...]

**Solution Level**

environment

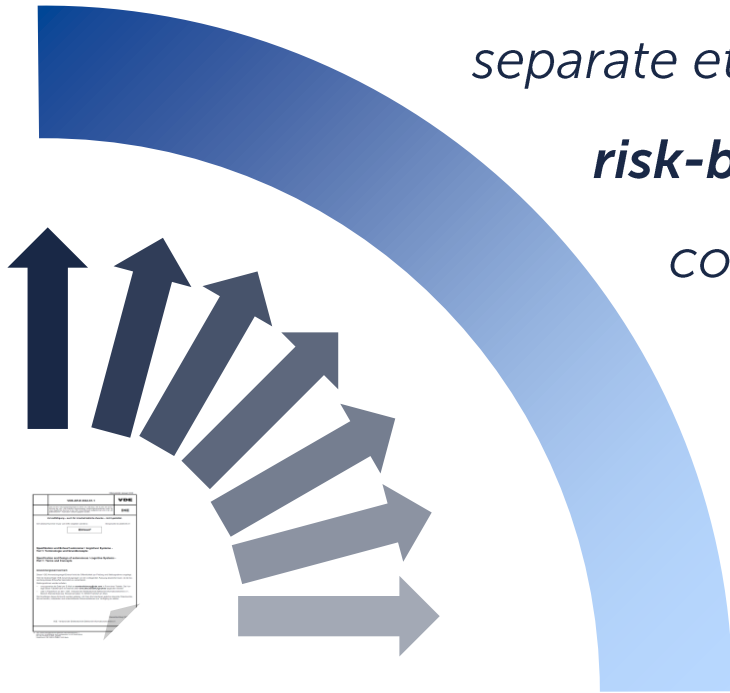
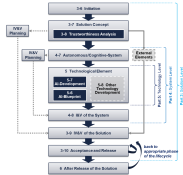
**System Level**

other objects & systems  
user & stakeholders

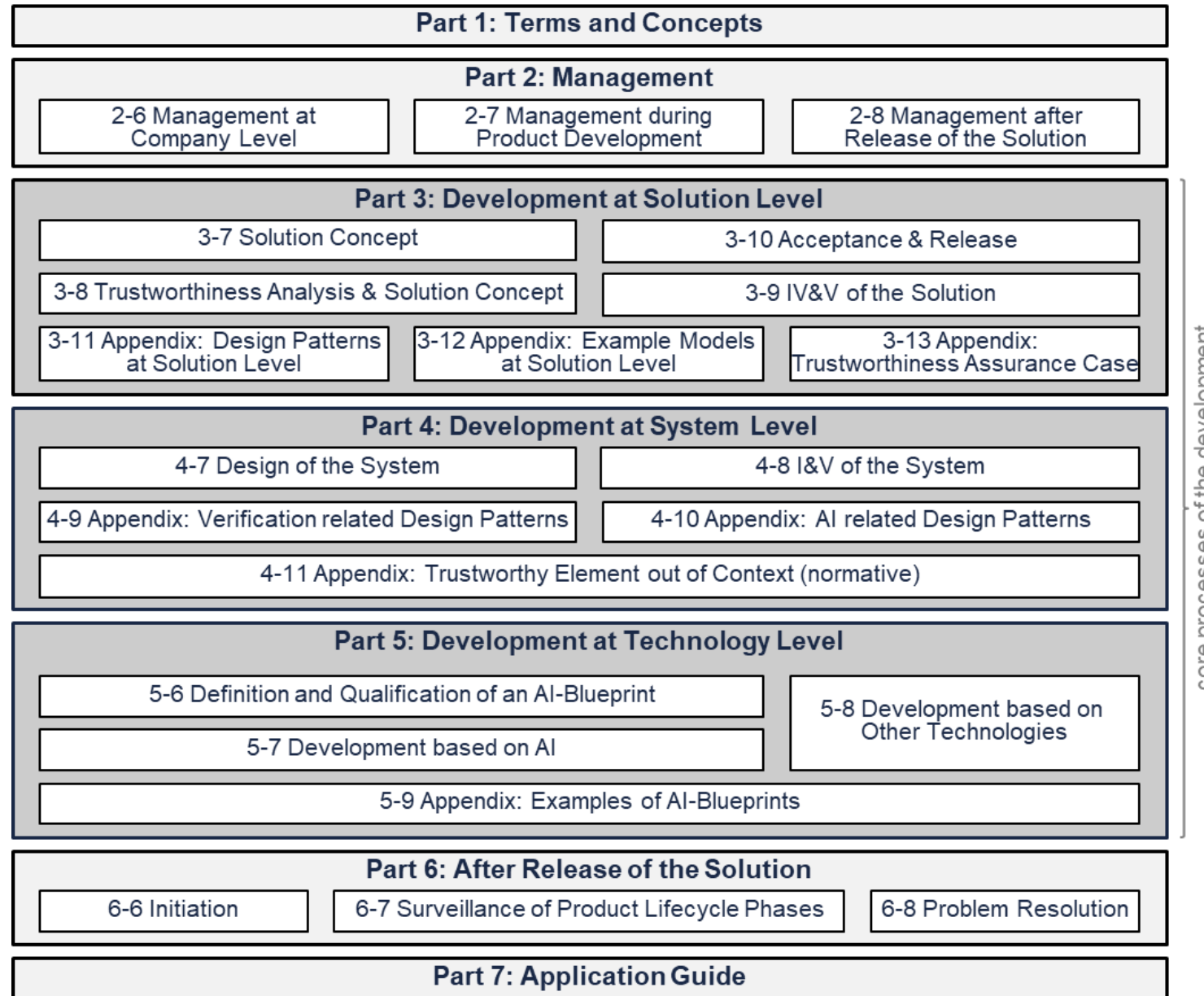
**Technology Level (AI)**

**VDE-AR-E 2842-61**

inspired by IEC 61508 & ISO 26262  
(aware of other working groups, e.g. ISO/IEC JTC 1/ SC 42)



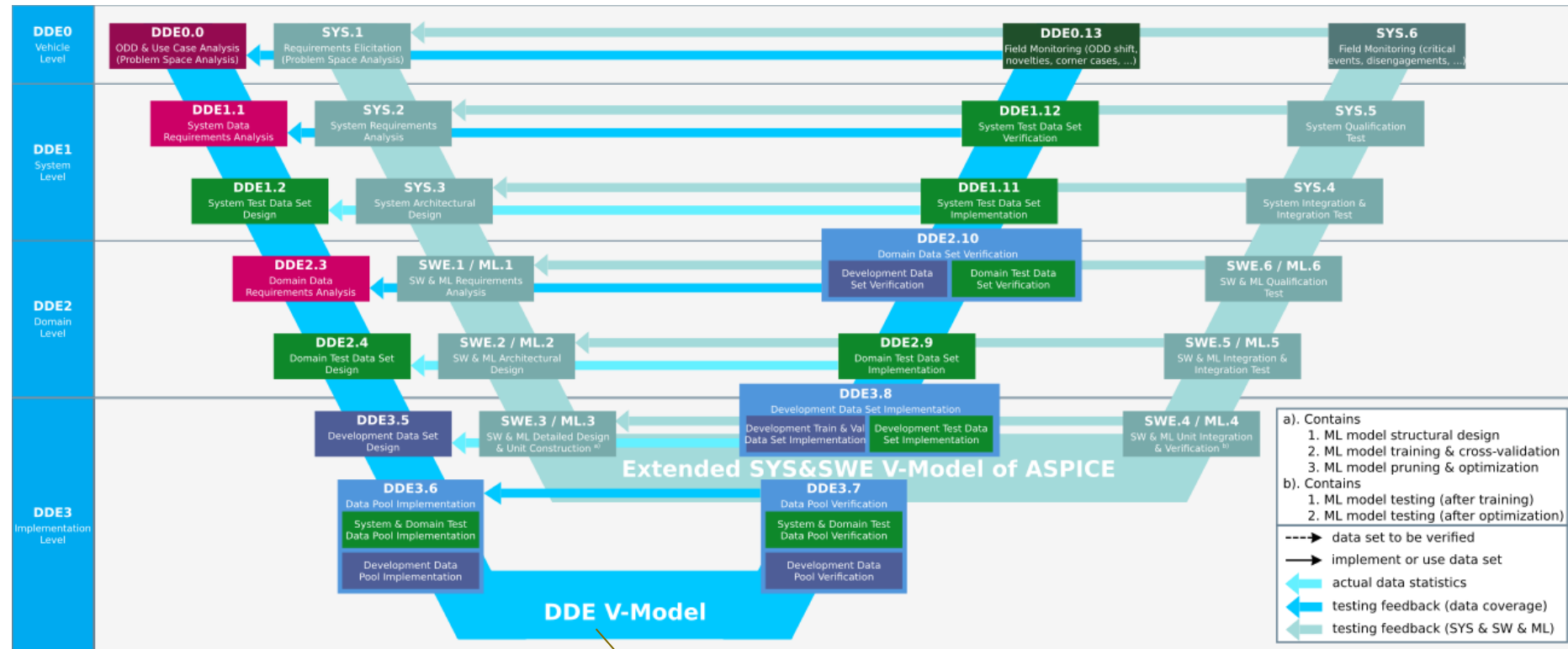




**VDE-AR-E 2842-61**  
comprises of 6+x parts:

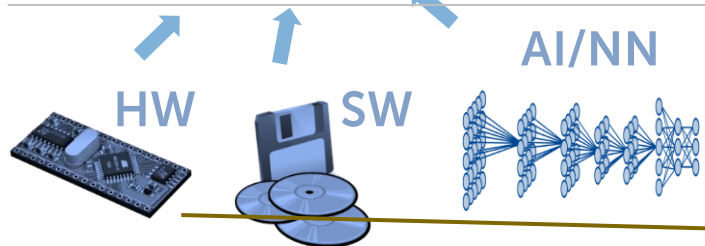
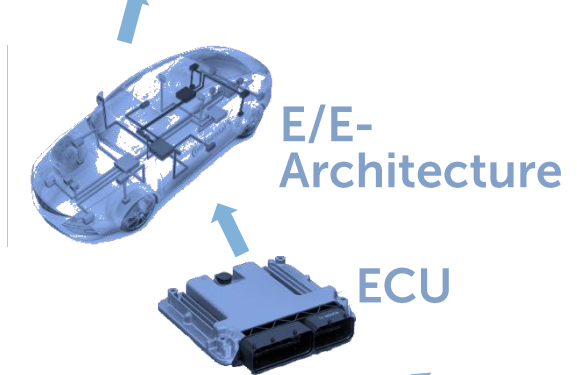
- 1) **Terms and Concepts** ✓
- 2) **Management** ✓
- 3) **Solution Level** (✓)
- 4) **System Level** Q3-2022
- 5) **Technology Level** (✓)
- 6) **After SOP** ✓
- 7) **Application guides**  
(in preparation)

# RELATED: Data Driven Engineering



Use Data Driven Engineering (DDE)  
to prepare data  
as a consistent approach over all levels of development abstraction

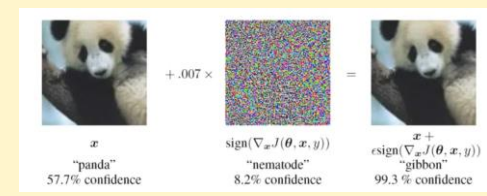
# RELATED: Cybersecurity for AI



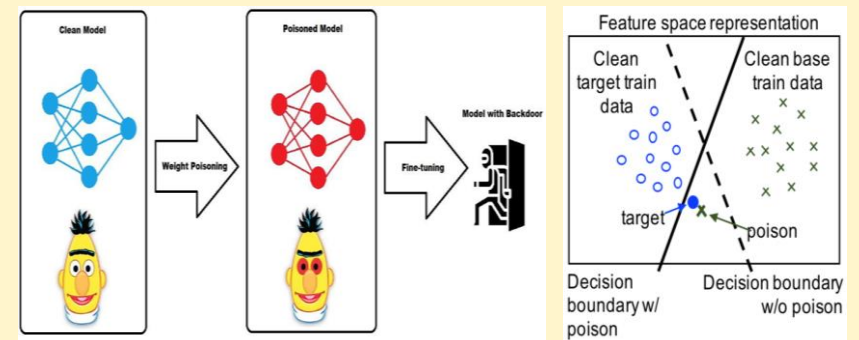
3D street art,  
Deep fakes,  
etc.



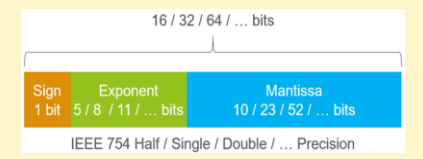
adversarial attacks



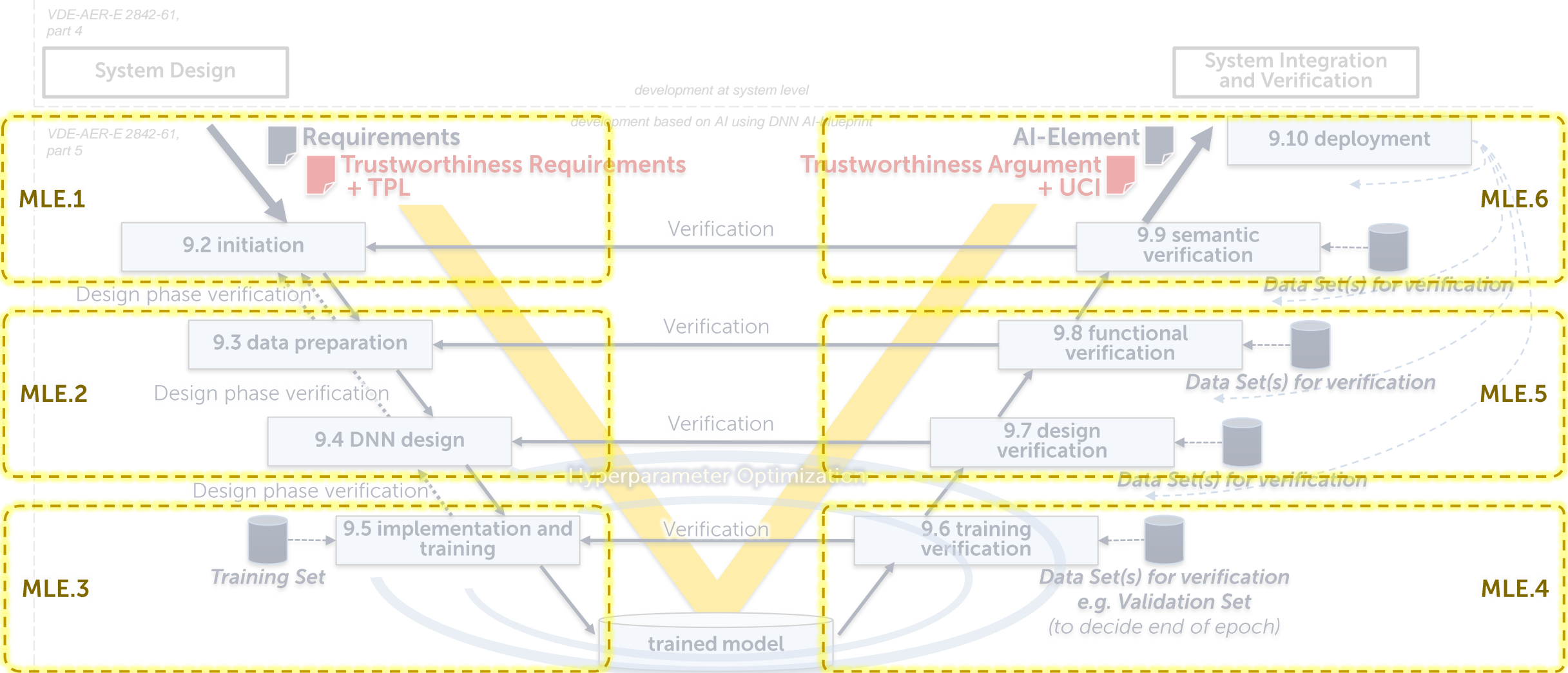
data poisoning



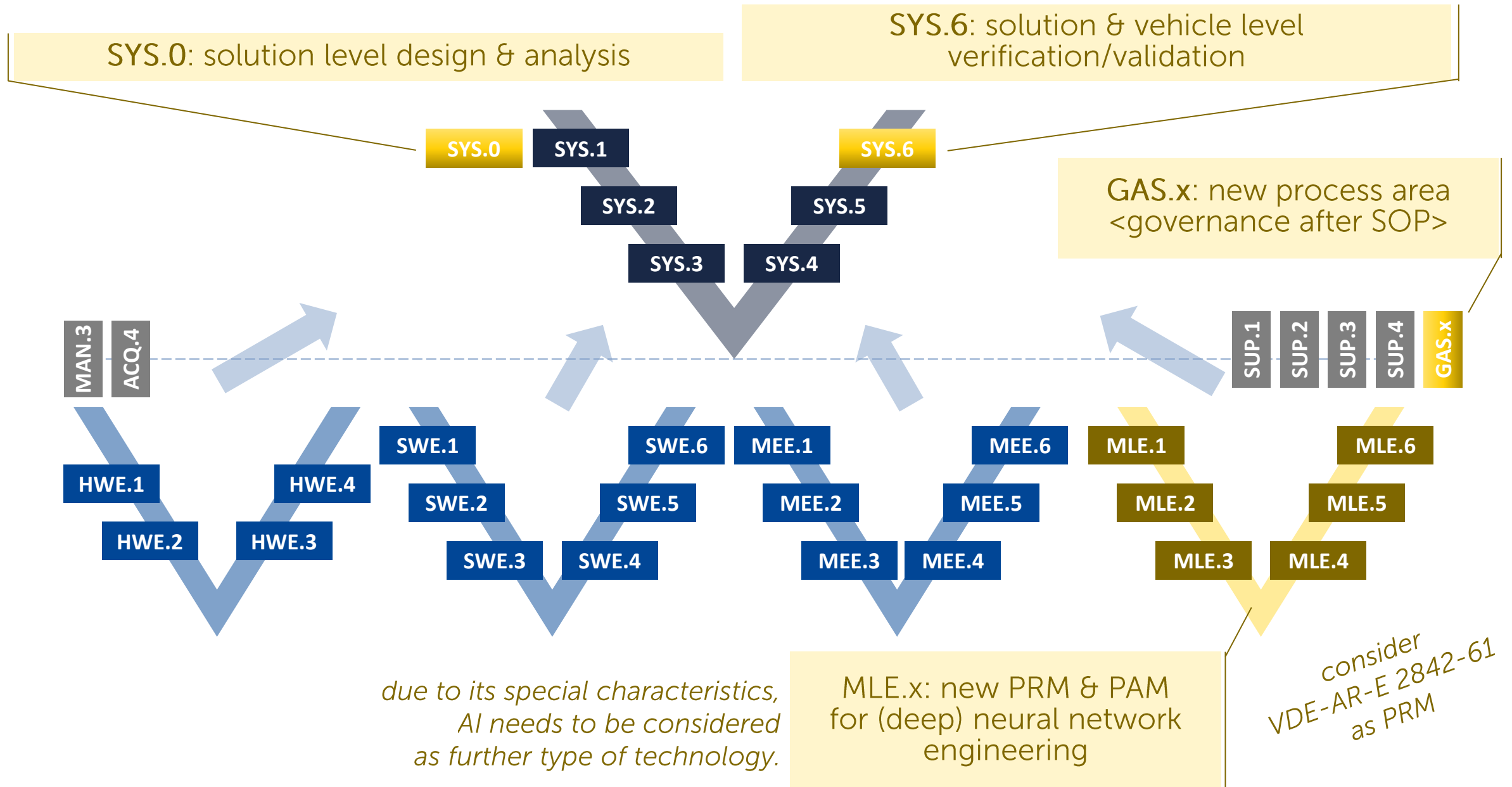
side-channel attacks



**RELATED:** [VDE-AR-E 2842-61 DNN AI blueprint inspires a PRM for KI ASPICE](#)



## RELATED: SPICE and ASPICE extension (proposal!)



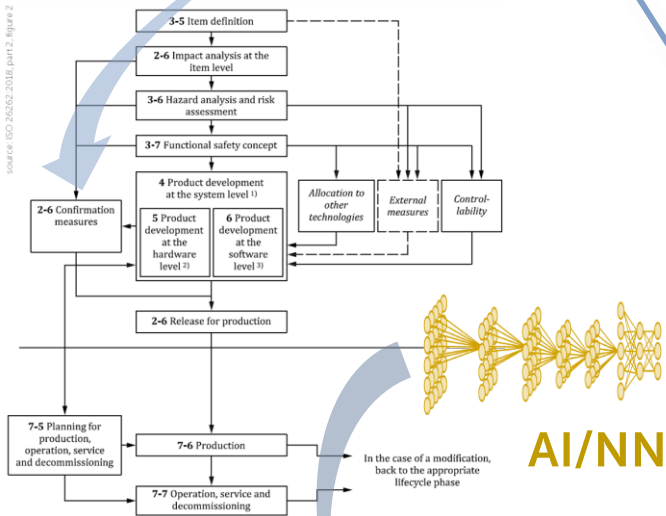


# Extending the risk-based approach for AI/NN

ISO 31000: RISK =  
effect of **uncertainty** on **objectives**

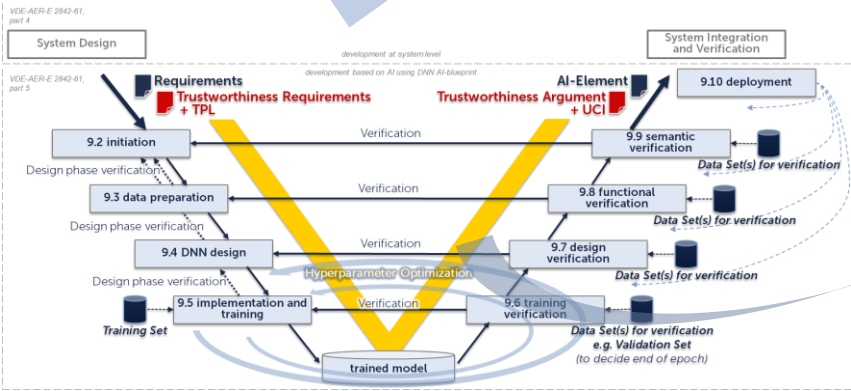
objectives = **trustworthiness** ~ as a meta term  
(safety, cybersecurity, usability, ethics, etc.)

## risk-based approach structured

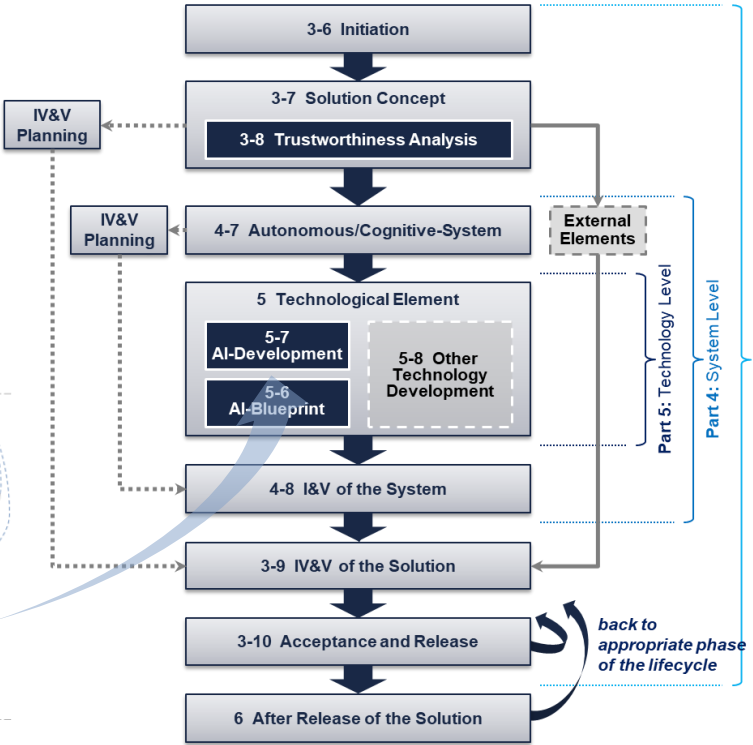


type of failure	measures	measures for HW	measures for SW	measures for AI
systematic	<u>Qualitative Requirements:</u> Culture, Experts, QS Process, Design, Methods & Measures	systematic capability (TPL/ASIL)	systematic capability (TPL/ASIL)	systematic capability (TPL/ASIL)
random	<u>Quantitative Requirements:</u> Metrics and Thresholds	SPFM, LFM, DC, FIT – (ASIL rel. target)	-- / --	-- / --
uncertainty-related	<u>Structured Approach:</u> Metrics, References, Measures and Argumentation	-- / --	-- / --	Uncertainty Confidence Indicator (UCI)

evidences within the argumentation (e.g. GSN) of the safety case (trustworthiness assurance case)



## VDE-AR-E 2842-61 Development and trustworthiness of cognitive/autonomous Systems



# Extending the risk-based approach for AI

---

**Contact us  
... to contribute or  
... to test drive & projects**



A modern office interior with large windows showing green trees outside. Grey armchairs are arranged around a glass coffee table. The left side of the image is overlaid with a semi-transparent blue rectangle.

**Thank you for  
your attention.**

**See you in the Q&A panel.**



# Presentation by fortiss & cogitron



As a computer scientist **Dr. Henrik J. Putzer** received his grade at the institute of Prof. Onken and Prof. Dickmanns at the University of the German armed forces in Munich with research on human centered, AI based assistants. After several successful industrial projects focusing on overall design, safety, security and process development cross industries he contributed to state of the art in E/E systems engineering. Among others he was core contributor to the ISO 26262 and currently holds a chair in der VDE DKE working group for the VDE-AR-E 2842-61. Currently he is the head of Trustworthy Autonomous Systems at the research institute fortiss and he is the CEO of cogitron, a consulting business on processes, embedded systems, safety & security and Artificial Intelligence.



**MSc Katharina Kofend** received her Master of Science in Geoinformatics (MSc) at Technical University Munich and specialized in explainable and interpretability of AI.

At fortiss she researches in the area of trustworthy Artificial Intelligence, verification of AI and assurance cases (including safety, security and related standardization).



**Dr. Hans Herrmann** completed his studies of mechanical engineering at the University of Hannover. At the University of Bonn, he researched simulation methods for coupled physical and chemical processes, concluding with a dissertation. As a consultant Dr. Herrmann contributed to process improvement and products development in the branches automotive and aerospace by accounting for the subjects embedded systems, system & software safety, and process design. Currently Dr. Herrmann is busy as consultant at cogitron GmbH and represents the field of embedded systems, safety, cybersecurity, and process consulting.

# Contact

---

fortiss GmbH  
Guerickestraße 25  
80805 München

[www.fortiss.org](http://www.fortiss.org)  
[info@fortiss.org](mailto:info@fortiss.org)

**Dr. Henrik J. Putzer** ~ [putzer@fortiss.org](mailto:putzer@fortiss.org) / [henrik.putzer@cogitron.de](mailto:henrik.putzer@cogitron.de)

**Katharina Kofend** ~ [kofend@fortiss.org](mailto:kofend@fortiss.org)

**Dr. Hans H. Herrmann** ~ [hans.herrmann@cogitron.de](mailto:hans.herrmann@cogitron.de)

