Safety Assessment of Artificial Intelligence

Hendrik Schäbe

- Work together with Jens Braband -

TÜVRheinland®
Precisely Right.

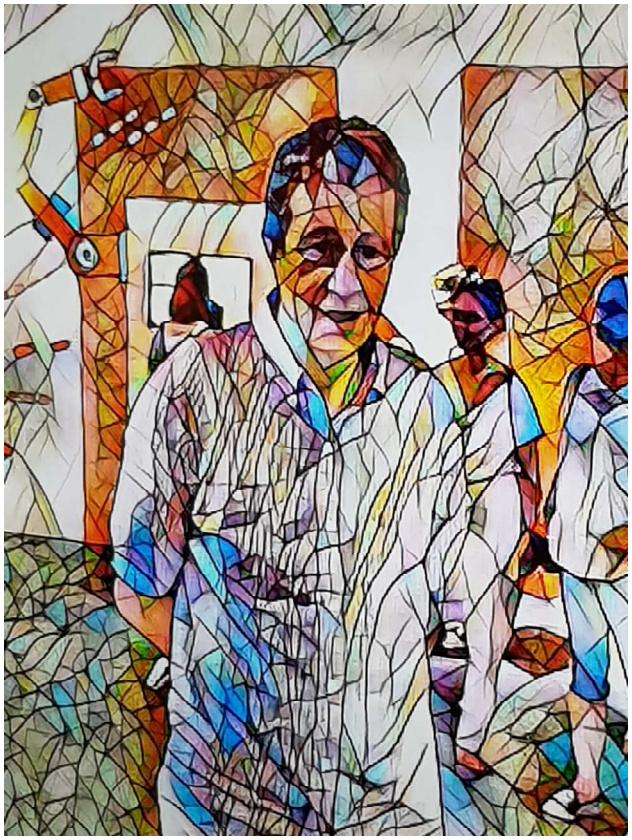
Overview

1. Introduction
2. What is Artificial Intelligence (AI)?
3. Does AI need a SIL and how is it determined?
4. Looking into the inside of AI
5. Possible assessment procedure
6. Conclusion

1. Introduction

Applications of AI:

- Data processing
- Assistence systems
- Speech recognition
- Face recognition
- Nursing robots
- Autononous driving systems
- Art etc.



Some of these systems are safety relevant, functional safety standards are applicable and safety assessment is required.

2. What is artifical intelligence

There exist many publications and many systems are named as being artifically intelligent.

Brunette, Flemmer & Flemmer (2009):

- Start with Turing test in the 50s
- Concept of evolutionary program
- „Artificial Intelligence“ first used at Dartmouth College in 1956.
- Proposition of different concepts by many researchers

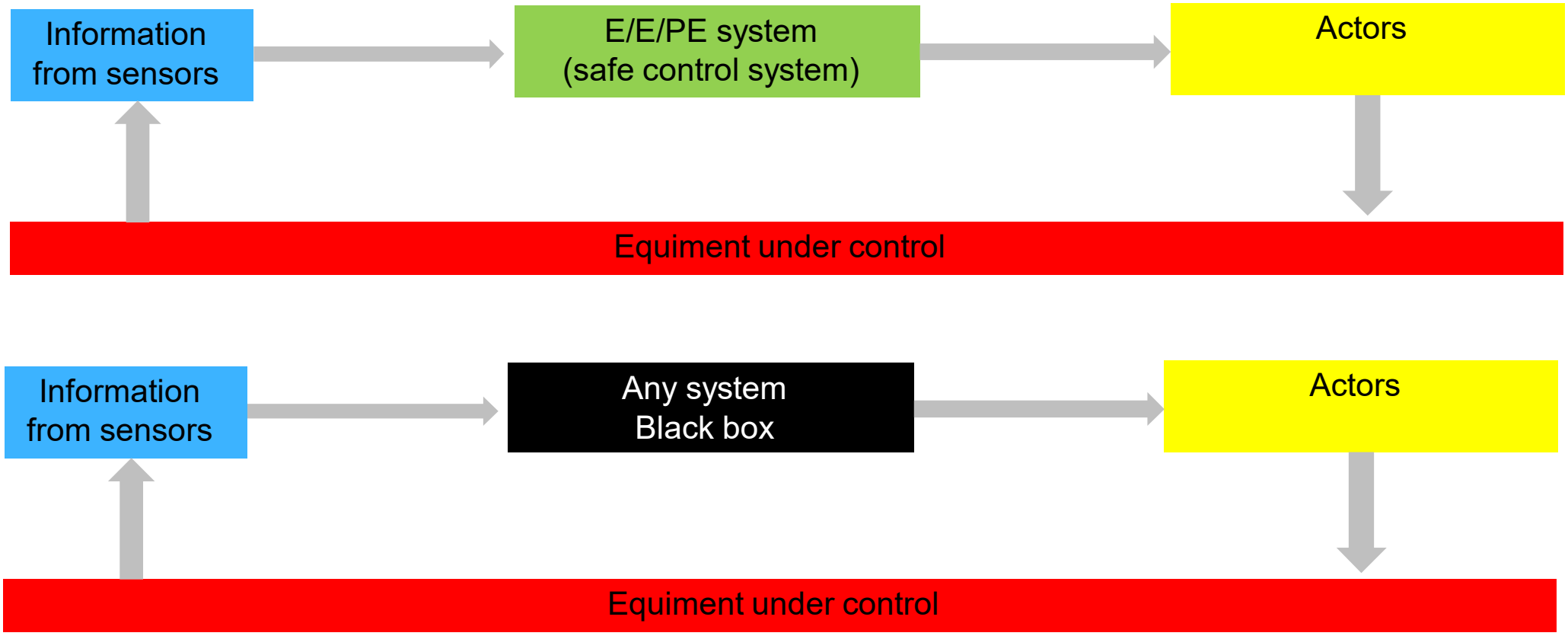
Artificial Intelligence = intelligence demonstrated by machines. Mimic cognitive functions, learning, problem solving.

Are criteria of intelligence:

- Use of speech?
- Conciousness?
- Self-wareness?

3. Does AI need a SIL?

Situation with other E/E/PE systems



3. Does AI need a SIL?

The black box can also be an AI system

What SIL to expect?

- Data processing – depends on the results and what is done with it
- Assistence systems – normally no SIL if a human overrides always the system
- Speech recognition – depends on what is done and whether there are safe backups
- Face recognition
- Nursing robots – giving medicine, carrying patients -> SIL required
- Autonomous driving systems – can lead to accidents -> SIL required
- etc. -> analyse

Result: a hazard and risk analysis needs to be carried out to determine a SIL. The relevant functional safety standard has to be applied.

4. Looking inside the AI

Requirements of the functional safety standards – example: IEC 61508

IEC 61508-3 Table A.2

no. 5 Artificial intelligence / fault correction **SIL 2- SIL 4: NR** (see C.3.12)

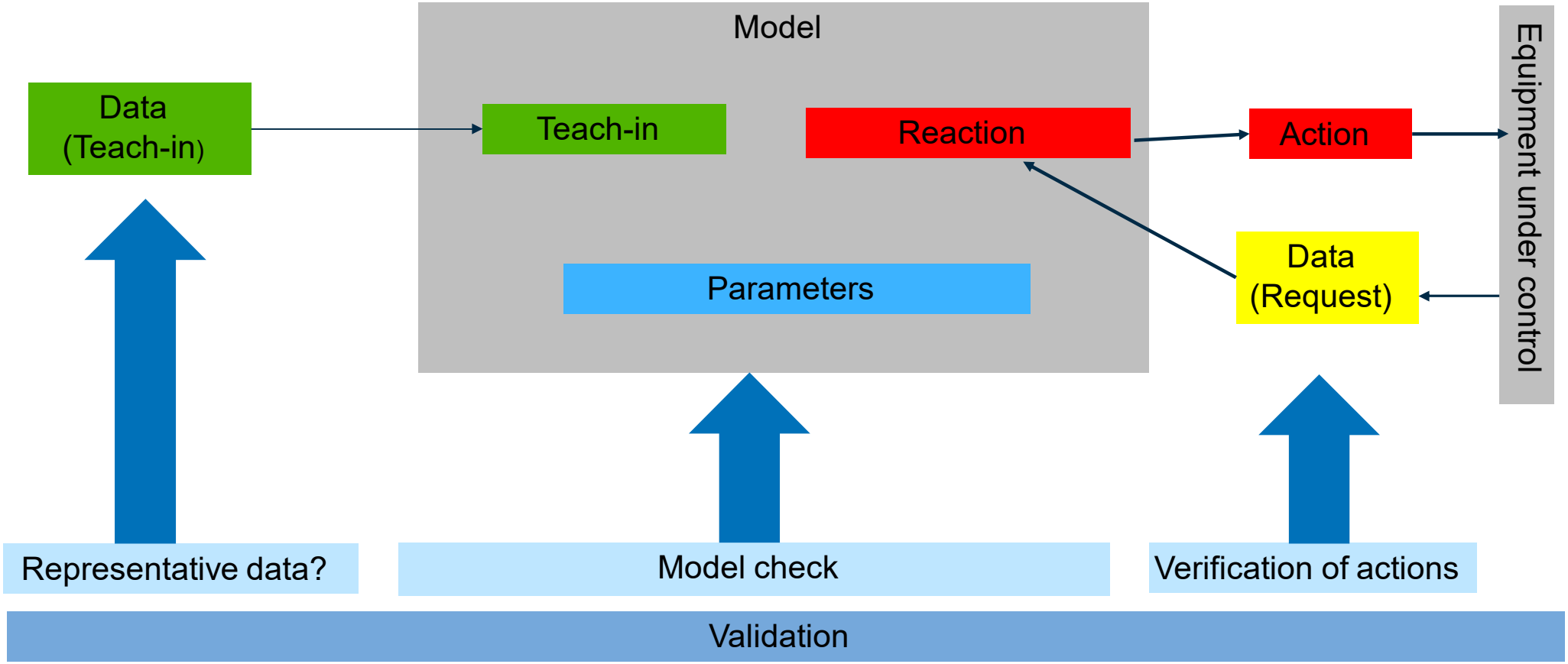
no. 6 Dynamic reconfiguration **SIL2 – SIL 4: NR** (see C3.13)

IEC 61508-7

C.3.9 Artificial intelligence

Description: Fault forecasting (calculating trends), fault correction, maintenance and supervisory actions may be supported by artificial intelligence (AI) based systems in a very efficient way in diverse channels of a system, since the rules might be derived directly from the specifications and checked against these. Certain common faults which are introduced into specifications, by implicitly already having some design and implementation rules in mind, may be avoided effectively by this approach, especially when applying a combination of models and methods in a functional or descriptive manner. The methods are selected in such a way that faults may be corrected and the effects of failures be minimised, in order to meet the desired safety integrity.

4. Looking inside the AI



- Different from Wang et al. (2017) but inspired by them

4. Looking inside the AI

Similarity Analysis: Machine learning is a statistical algorithm – what can we learn from statistics?

Machine learning is just statistical data fitting – but with very complex algorithms and big data?

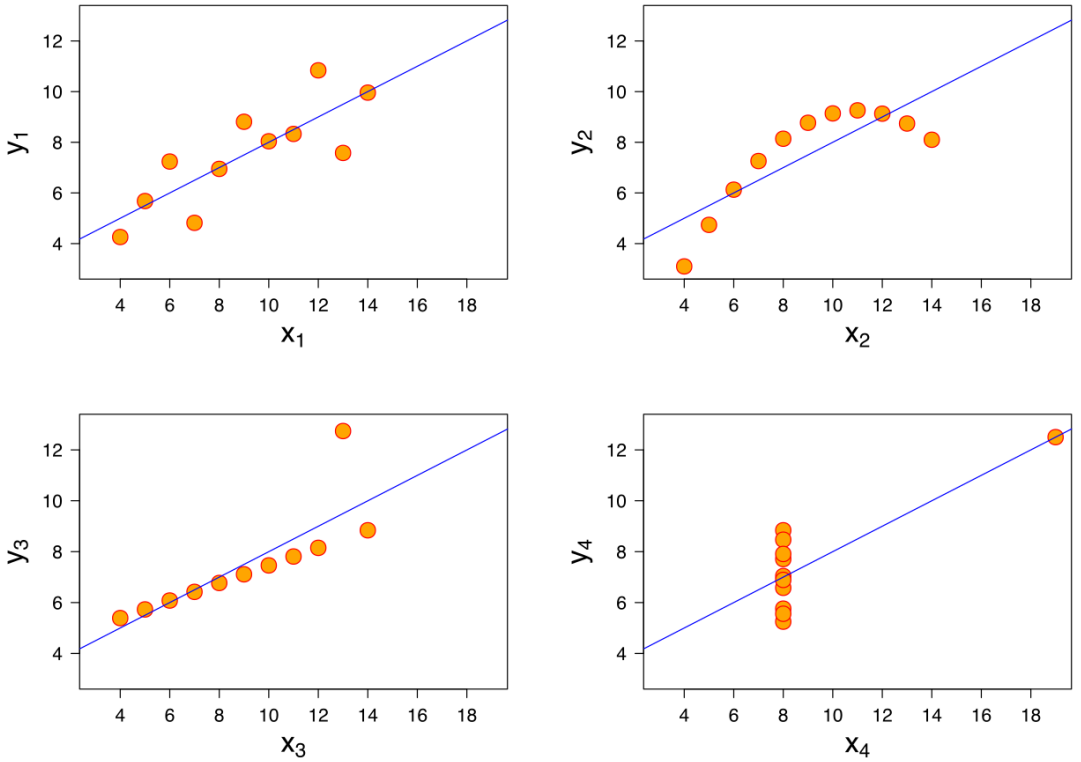
The most simple statistical model is linear regression. What can we learn in general from it?

1: The model must be correct – otherwise we will never fit the data well.

2. The training data must be representative of the real data.

3. We need a measure of goodness of fit (like R^2)

4. How do we detect Black Swans (similar to outlier detection)?



Credits:By Anscombe.svg: SchutzDerivative works of this file:(label using subscripts): Avenue - Anscombe.svg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=9838454

TÜVRheinland®
Precisely Right.

4. Looking inside the AI

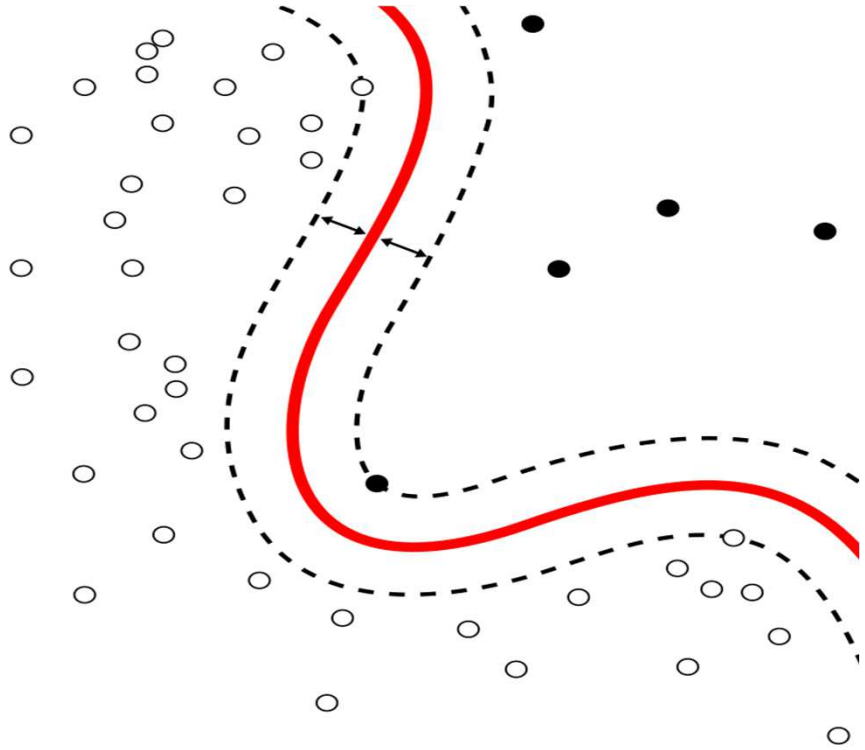
Machine Learning as a classification problem - Just another look at the problem…

Basically most ML algorithms solve classification problems, similar to cluster analysis in statistics.

We have (at least) 2 classes of (big) data in a high dimensional space.

The optimal discrimination function would seperate the 2 classes completely for the training set.

However there remains some space between the two classes and there exists no unique solution for the problem.



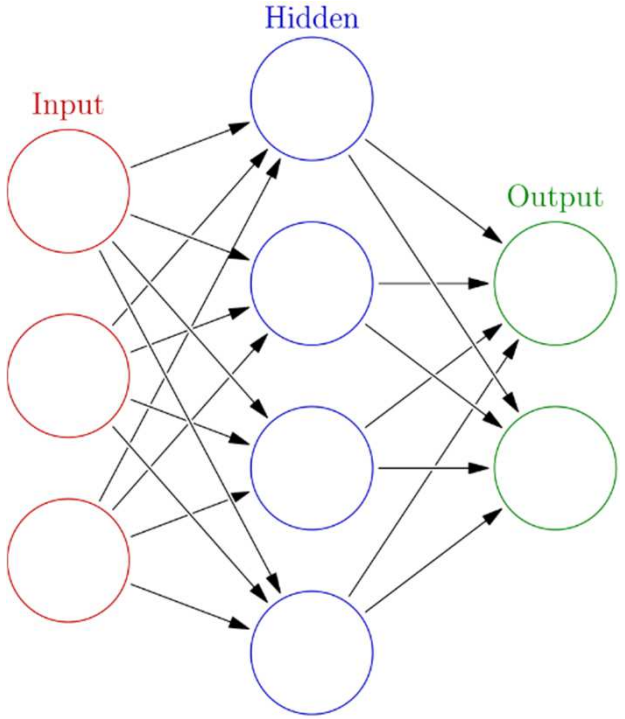
4. Looking inside the AI

Artificial Neuronal Networks (ANN) - A popular example

Any ANN has at least two layers that are connnected by weights. The input data x are transformed by weights v and w, offsets b and an output function to two output classes

$$F(x) = \sum_{i=1}^{N} v_i \, \varphi\left(w_i^T x + b_i\right)$$

The optimal weights for a particular cost function C are found iteratively based on the training data and a numerical algorithm.



But: Is this the correct function? Does it approximate well? Or do we need more layers or more complex functions?

If we can‘t answer the question, we might have a systematic flaw in the model!

4. Looking inside the AI

The Universal Approximation Theorem - Sometimes abstract maths can be helpful…

Fortunately, there exists a variety of so called “universal approximation theorems”, that show convergence of F to f, e. g. if φ is a bounded and continuous function and if f is continuous.

At the first glance the result is surprising because it already holds for ANN with a single hidden layer but on second thought the results are quite obvious and a have a simple explanation:

1) F is a kind of general linear approximation to f. But it is obvious that such linear approximation should be possible if the number of nodes N is sufficiently large. Also in the classification example f could be approximated by stepwise linear functions.

2) Also deep ANN with several hidden layers could be represented by single layer (with large N).

For dependable applications the requirements could be:

1) Choose a single-layer ANN with sufficiently large N

2) Choose an appropriate cost function C (with justification)

Cybenko, G. (1989) "Approximations by superpositions of sigmoidal functions", *Mathematics of Control, Signals, and Systems*, 2(4), 303–314

4. Looking inside the AI

Representative data?

This means that teach-in must occur in a typical environment for this type of system and the environment must be such that the influences are typical for this type of use, including all the changes in the environment. So, all replications of the system (after teach-in) must be operated at least in similar environments. And all replications of the system must be similar.

Compare Braband, Gall & Schäbe (2018)

4. Looking inside the AI

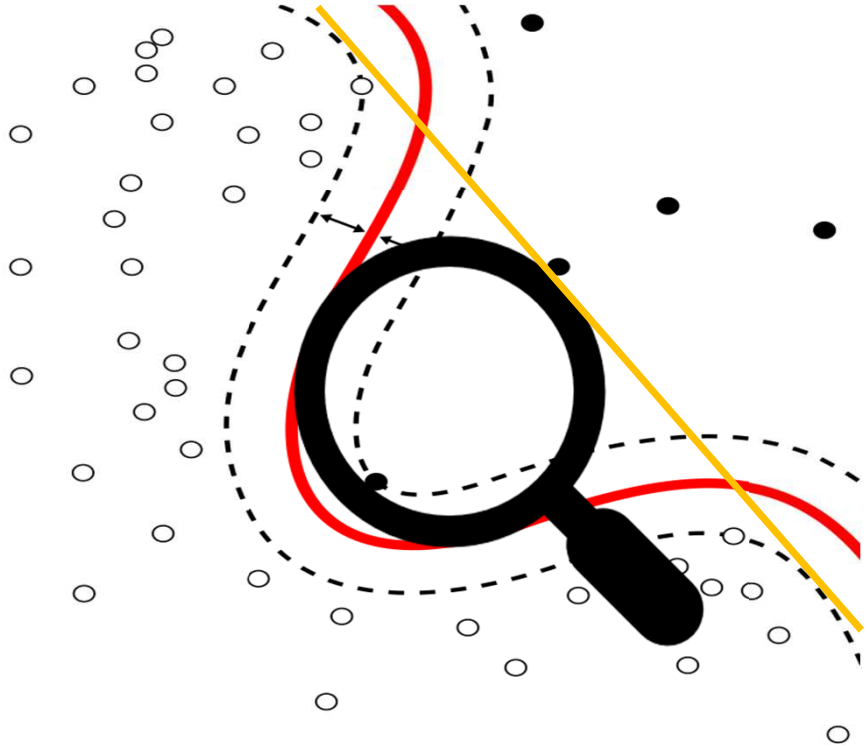
Goodness of fit? Can we accept failure in training data?

Generally any misclassification in training data could lead to a high proportion of classification failure in practice.

This means

1) Either we have 100% correct classification in the training data, or
2) We can calculate the error probability well

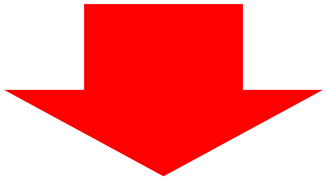
The problem is that we cannot simply count classification errors. We have to weight them according to their importance, which may be difficult in high-dimensional spaces and big data…



4. Looking inside the AI

Teach-in has statistical aspects. This means:

- Confidence bounds need to be taken into account
- Derived parameters are random values containing some spread
- The subsequent decisions of the AI will also be random, with some errors:
- First kind error: wrong decision, althoug the input data are in the „right“ domain
- Second kind error: input data are in the „wrong domain“, but decision is „right“.



As a consequence, the AI will have a failure probability. This must be taken into account, assigning part of the budget of thr rate of dangerous failures to the AI (here: the algorithm)

5. Possible assessment procedure

How to cope with the IEC 61508 rules against artificial intelligence?

The statement is combined with a statement about dynamic reconfiguration, this is also undesired for SIL 2 …SIL 4.

The functional safety standard requires a predictable system.

Predictable:

Measures against systematic failures so that they can be neglected

5. Possible assessment procedure

Random failures‘ occurrence is brought to a sufficiently low level

-> Make AI behavior predictable by

- Analysing the model
- Taking part of the budget for random failures for the AI.
- Treat the AI system as a normal mathematical model with probabilistic behavior.

Then assessment is the „normal assessment“.

We will not repeat the content of a „normal“ assessment procedure.

5. Possible assesment procedure

Interesting issues with validation of AI:

An error occurs in the validation of the AI, the same error will occur again during its use, if the same data set occurs.

The error will occur with high probability, if a “similar” data set occurs. -> systematic error in the AI function?

Even if the fraction of errors is very small, is it permissible to commission a safety-critical system that is known to react incorrectly to certain data sets?

Or must these data sets be exempted along with their “environment”?

Or is it even necessary to require that the validation is error-free?

Possibly tolerate errors = a certain hazard rate caused by erroneous decisions made by the AI function.

When verifying the TFFR: requirements are usually so strict that they can only be verified for low safety levels, e.g. SIL1, using testing and proven in use arguments.

5. Possible assesment procedure

5.1 Analytical approach

For higher SIL, analytical proof will usually be necessary.adapt the approach from other areas, e.g. for computer architectures or data transmission. A sensible model is needed.

Proof will be easier for data that has a certain structure, e.g. due to the physical properties of the data sources.

A safety case will be easier the more one understands the structure of AI methods.

Neuronal Networks are a difficult example -they are difficult to explain.

Other AI methods, such as Support Vector Machines (SVM) have a more explicable structure and provable properties.

A “model-oriented” approach to the safety case would have the advantage that it could be conducted on the basis of today’s established safety standards.

5. Possible assesment procedure

5.1 Analytical approach

Check the mathematical model:

- Check correctness of the model according to physical / chemical / mathematical and other scientific proven theories
- Equivalent to other mathematical models as e.g. of brake curves , thermal models etc.

That means, the theory / model must be disclosed.

Which kind of model might we find,cp. Wang (2018), examples are:

- Neural network
- Long short-term memory
- Auto encoder
- Deep Boltzman machine
- Generative adversarial network
- Attention-based LSTM

5. Possible assesment procedure

5.2 Statistical approach

The AI is regarded as a blackbox.

First step – Train AI, freeze the status of the AI system

Second step - validation and statistical testing phase. Confronting AI with a very large data sample, required decision of AI known per case.

Simulate data or let AI run along in “idle mode”

If the AI has been statistically and representatively tested for a long enough time, it is possible to use an approach equivalent to

one proven in use.

The probability of a wrong decision is 3/n with 95 % statistical confidence, if n classification tests have been run with no false classification.

Etsimates the PFD = Probability fo failure on demand

5. Possible assesment procedure

5.2 Statistical approach

Apply proven in use approaches for a PFH = probability of failures per hour

Means:

- 3 10^5 failure free hours for SIL 1
- 3 10^8 failure free hours for SIL 4

(Braband, Gall & Schäbe (2018)

Hard to accumulate such a quantity of failure free hours

5. Possible assesment procedure

5.2 Statistical approach

Problems with statistical approach

1. The test sample size

Sufficiently large - difficult, if the events to be observed do not

occur very frequently.

- Possible to use bootstrap procedures, i.e. to reproduce further random samples based on the already (careful to avoid simple reproduction and plowing up sample size without containing additional information). In many cases, the data is very large for each element of the sample - decompose and recombine the data.
- Use a model (model must be good – question: why not use the model in plac of AI?)

5. Possible assesment procedure

5.2 Statistical approach

2. The representativeness of the test sample

The sample must be statistically representative. This means that the AI must first be trained for all possible future data and then also be tested with such data. Ensure that the respective samples correspond to **reality** and cover **all the relevant cases**.

Analyse data in great detail. Itmight be easier to derive a good statistical model rather than to arrive at AI.

Approach to be feasible, if it is relatively easy to collect the data automatically.

5. Possible assesment procedure

5.3 Possible assessment procedure

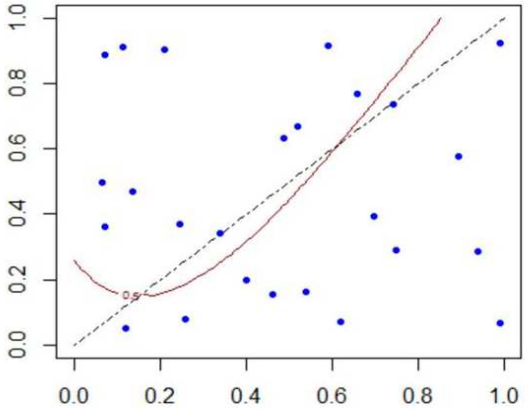
An academic example

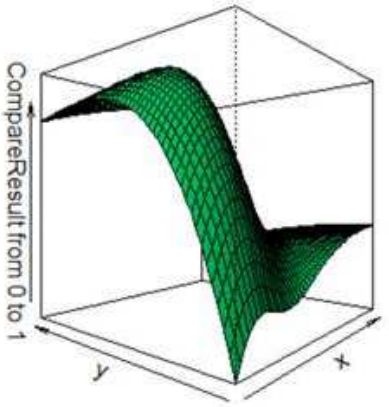
Take two sets of randomly generated points on the unit square (known distribution) classified in two classes.

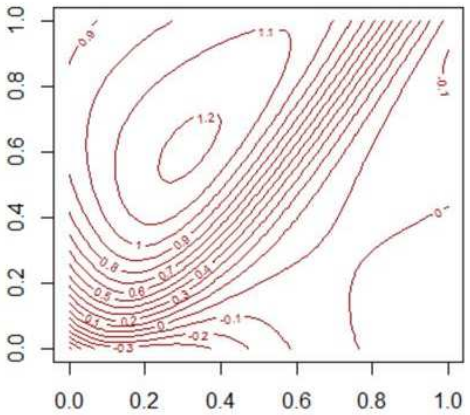
Under which assumptions can you assure a failure probability for a critical application (unknown distribution)?

Some interestion questions:

- Can we then prove the assumptions?
- Can we lift this example to (much) higher dimensions?
- What, if not?







Kudos to S. Griebel for the illustrations.

TÜVRheinland®
Precisely Right.

6. Conclusions

- Analysis of AI systems can be carried out, but it becomes quite complex
- A SIL can be determined as for a normal E/E/PE system
- An assessment requires always an in-depth model analysis.
- The more flexible the model, the more complicated the analysis
- AI can be easily used in situations, where no critical consequences occur, which has to be supported by a risk analysis.
- For use in critical systems it seems a useful approach to restrict the type of models

Important points:

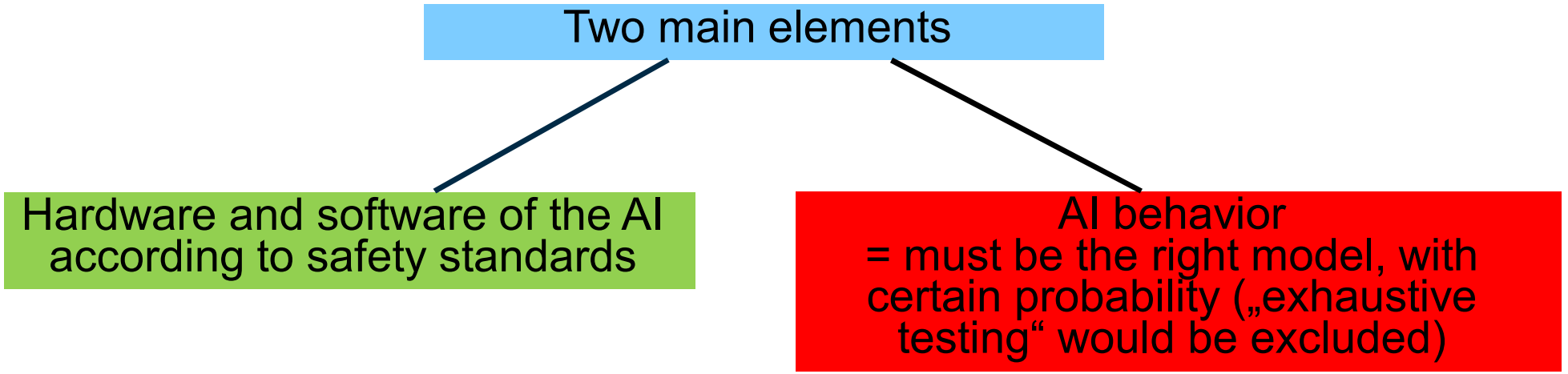
- the data used to train and validate the AI must be representative for the operating environment.
- there must be clear rules on how to deal with anomalies or “outliers” in the data
- the operating environment must not change.
- the AI approach must be flexible enough to approximate the (unknown) “true” decision function well enough, but also to avoid overfitting.

Minimum requirements when using the CENELEC standards:

- the AI function must have been validated with the system, where it will be used.
- there must have been a safety qualification testing period.
- the validation data must not have been used to train the AI function.
- the SIL requirements must have been proven for the implementation of the AI function, i.e. for the used software and, if

applicable, tools.

- a tolerable functional failure rate (TFFR) must have been demonstrated, either by means of a test or operational experience (for a low SIL) or using an analytical model (for a higher SIL).



Probability requires regularity conditions !

Backup

References

J. Braband, H. Schäbe, On safety assessment of artificial intelligence, On Safety Assessment of Artificial Intelligence, Dependability no. 4-2020. p. 25-34.

J. Braband, H. Schäbe, Anwendung Künstlicher Intelligenz in der Bahntechnik für sicherheitsrelevante Anwendungen – Chancen und Probleme, Signal & Draht (114) 5/2022, S. 14-21

Academic example

Assume a classification system that classifies objects in two categories: „left“ and „right“ based on one real-valued parameter. The parameter is assumed to be normally distributed.

„left“ is characterised by mean mL and spread sigmaL,

„right“ is characterised by mean mR and spread sigmaR.

First, assume the parameters to be known.

Classification rule:

„left“ if $X<z$ and „right“ if $X \leq z$.

z is a „properly“ chosen constant

Academic example

α = 1 - Φ(z-mL/sigmaL)	first kind error	(1)
β= Φ(z-mR/sigmaR)	second kind error	(2)
Φ(z-mL/signaL)	correct „left“ classification	(3)
1- Φ(z-mR/sigmaR)	correct „right“ classification	(4)

Φ – standard normal integral.

sigmaR and sigmaL should be as small as possible to have small errors.

Academic example

What went wrong so far?

Parameters m, sigmaL and sigmaR are not known but must be learned!

How does the system learn?

A „left“ sample XL_i, i=1,n_L and a „right“ sample XR_i, i=1,…,n_R are used for teaching.

What do we compute?

$m_R = (1/n_R) \Sigma\, XR_i$, $m_L = (1/n_L) \Sigma\, XL_i$, $sigmaR^2 = \Sigma\, (XR_i - mR)^2/(n_R\text{-}1)$, $sigmaL^2 = \Sigma\, (XL_i - m_L)^2/(n_L\text{-}1)$

What do we really have?

Point estimates, that are still random, therefore denoted in *Italics*.

What do we have to use?

Confidence limits with first kind error γ.

Academic example

Use confidence limits such that the misclassification error becomes large, i.e. upper bounds for the sigmas and m_L, lower bound for m_R. We use single parameter bounds – not combined ones - to simplify the computation.

Distribution of the estimators:

$(n_L-1)sigmaL^2$ / sigmaL^2 is chi-squared distributed with n_L-1 degrees of freedom

$(n_R-1)sigmaR^2$ / sigmaR^2 is chi-squared distributed with n_R-1 degrees of freedom

$\sqrt{n_L}(m_L-m_L)$/sigmaL has a t distribution with n_L-1 degrees of freedom

$\sqrt{n_R}(m_R-m_R)$/sigmaR has a t distribution with n_R-1 degrees of freedom

Academic example

The least favorable values are:

Upper confidence bounds for the variances, i.e.

$\sqrt{(n_R - 1)/Chi2(n_R - 1; 1 - \gamma)}sigmaR,$ (5)

$\sqrt{(n_L - 1)/Chi2(n_L - 1; 1 - \gamma)}sigmaL$ (6)

Chi2(n;1-γ) is the quantile of the Chi-squared distribution with 1-γ coverage

Lower confidence bound for m_L

m_L-t(n_L-1;γ)sigmaL/$\sqrt{n_L}$ (7)

and upper confidence bound for m_R

m_R+t(n_R-1;γ)sigmaR/$\sqrt{n_R}$ (8)

t(n;γ) is the quantile of the t distribution with n degrees of freedom and coverage 1-γ.

Academic example

Inserting the confidence bounds (5) – (8) into the formulae (1) – (4) gives the probabilities of errors.

If misclassification with a type one error is dangerous, (1) with (6) and (8) gives the probability of a dangerous failure. However, to account for errors coming from the confidence intervals, value

$\alpha+2\gamma$

should be used.

Caution: The interpretation of γ as a probability that the true value lies outside the confidence interval is not a frequentist one, but a Bayesian using an appropriate prior.

Academic example

Now, for a SIL 1 system, a PFD of 0.1 must not be exceeded.

This value can be seen as a budget:

One might give 0.05 as a maximal value for hardware failures and 0.05 for the AI algorithm. The latter can be split according to

$0.05 = \alpha+2\gamma$

e.g. in the form $\alpha = 0.025$, $\gamma = 0.0125$.

For a SIL 4, IEC 61508 provides a threshold value of 0.0001 for the probability of failure on demand.

The reader might repeat the calculation.

As a further exercise, she / he might consider conditions on m and the Sigma values to fulfill the requirements

Academic example

Way out?

Even with this very simple example, we were confronted with complex mathematics.

Options:

1. The AI system does not need a SIL since its behavior does not have critical consequences (no injuries to persons etc.)
2. The AI system is supported by a sufficiently simple E/E/PE system, having the necessary SIL, that checks all dangerous decisions according to simpler algorithms and inhibits dangerous reactions

The options need to be supported by a risk analysis (see IEC 61508).