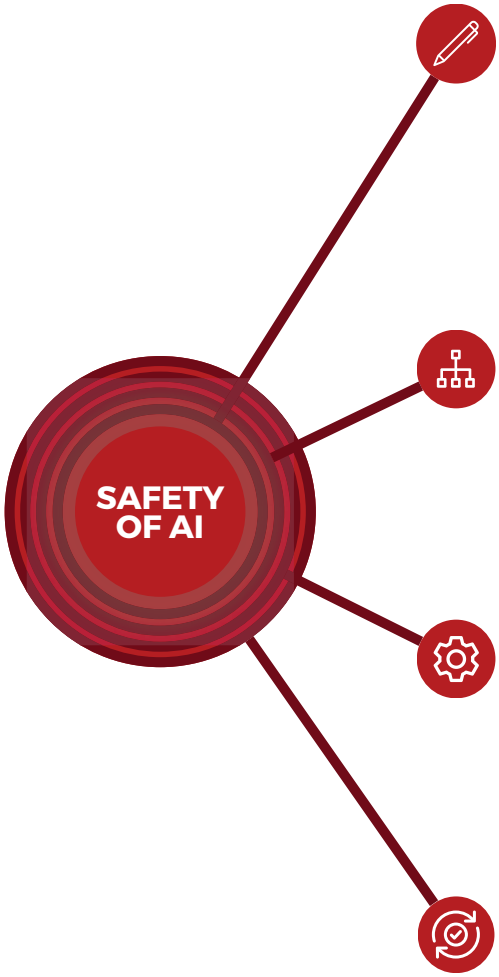
Guideline for AI in safety-critical applications

This guideline contains a list of principles to enable the development and deployment of Artificial Intelligence in safety-critical applications. The principles address AI specific challenges and insufficiencies and provide recommendations to ensure the safety of a system.



Functional specification

1. Specify intended component and system behaviour
2. Define the Operational Design Domain
3. Perform domain specific risk analysis

Architectural design

4. Use AI only where necessary
5. Implement AI only at component level
6. Prevent too complex AI sub tasks
7. Make use of redundancy strategies
8. Employ an independent Safety Monitor

Implementation methods

9. Ensure decisions are exclusively rule-based
10. Select suitable AI approach for each specific task
11. Base implementation on established coding guidelines

Training, testing and validation

12. Define quality gates for training, validation and test data
13. Define task specific validation metrics
14. Apply independent safety assessment
15. Develop a maintenance strategy

About us

We strive for a world in which safe and sustainable mobility systems improve people's lives. To achieve this, we support our visionary customers and partners worldwide. Together, we help connect and mobilize all people

Contact: etamax_innovation@fev.io

Guideline for AI in safety-critical applications

This Guidelines contains a list of principles to enable the safe development and deployment of Artificial Intelligence. The principles address AI specific challenges and insufficiencies and provide recommendations to ensure a systems safety.

The principles are divided into four categories – “functional specification”, “architectural design”, “implementation methods”, “training, testing and validation” – based on the different aspects of the development process of a system with AI-components.

Functional specification

1 - Specify intended component and system behavior

For the purpose of verifying and validating the behavior of an AI-based system or component, one needs to clearly define the intended component and system behavior. Consider differences in the specification of the behavior on component and system level.

[4]

2 - Define the Operational Design Domain

To guarantee a safe behavior of AI based components, their field of operation must be clearly specified and bounded. Out of distribution input can cause hazardous behavior of a Neural Network. Determine how the AI-based system shall behave if it is out of the ODD.

Before deployment of AI-based systems, the four risk classes defined by the European commission shall be considered:

1. minimal risk (e.g. automated suggestions)
2. low risk (e.g. dynamic price allocation)
3. high risk (e.g. AI-based diagnostics in medicine)
4. unacceptable risk (e.g. autonomous weapon systems)

[5], [6], [7]

3 - Perform domain specific risk analysis

When considering the development of safety-critical AI applications, a risk assessment is needed to examine the impact on life and limb, property and financial assets, the environment, fundamental rights and ethical principles. A core element of modern AI applications in the automotive sector is the dynamic risk assessment (DRA) which is part of the overall dynamic safety management of applications. For different industries risk assessment shall be based on existing standards. It is important to check whether these standards are designed to accommodate AI technologies.

Parameters like SIL, PFD/PFH as well as TPL (trustworthiness performance level) and UCI (uncertainty confidence indicatior) are to be considered in the domain specific risk analysis.

Standards to evaluate include.: IEC 61508, ISO 26262, ISO 21434, ISO 21448 und VDE-AR-E 2842-61

[1], [2], [3], [39], [40]

Architectural design

4 - Use AI only where necessary

As the use of AI technology already implies a threat to safety, we recommend its usage only where necessary and clearly a superior fit to the task in comparison to a non-AI based solution.

[8], [9]

5 - Implement AI only at component level

The assumptions about stable hierarchical architectures of components, as proposed for example by ISO 26262 for automated driving systems, shall be kept and AI shall solely be used at the component level. This applies equally to non-automotive applications. End-to-end solutions shall be avoided since they are less reliable with regard to verification and validation capabilities. This way, the influence on the overall system behavior can be limited and we encounter AI safety concerns only at the component level, which might be supervised and controlled by a higher instance.

[1], [3], [10], [4]

6 - Prevent too complex AI sub tasks

Define a system architecture assigning component responsibilities based on the overall system requirements. Identify subtasks which are suitable for an AI-based solution. The scope of a single AI-based component shall be appropriate in terms of performance expectation and overall system behavior. If safety requirements are not achievable by the AI unit itself, a solution might be found in the system architecture.

[11], [12]

7 - Make use of redundancy strategies

To increase the robustness of AI components the implementation of redundancy strategies is recommended. A common approach are ensemble strategies. These are classically based on the independent training of multiple classifiers and the subsequent combination of the resulting models based on their performance. Thereby, advantages of different architectures can be combined.

Examples of strategies are: random forest modeling, bootstrap AGGregating or sensor-fusion (e.g Occupancy Grid Mapping)

[13], [14], [15], [16], [35], [36], [37], [38]

8 - Employ an independent Safety Monitor

To verify the functionality and safety of the system, an independent supervisor entity shall be implemented. The supervisor entity shall not be logically connected with any other component in terms of the verification of the system. Nevertheless, it must have knowledge about the specified intended functionality of the component, the mandatory data subset to maintain safe system behavior and the capability of verifying against safety standards and requirements. Define measures in case of violation. A dynamic safety assessment (DSA) shall monitor the internal safety status of the system based on defined parameters and deploy countermeasures, if the safety goals are disrupted.

[4], [17], [40]

Implementation methods

9 - Ensure decisions are exclusively rule-based

Neural Networks impose a lack of complete specification and deterministic behavior. Rule-based approaches are easier to control in most cases and shall therefore be mandatory for decision making. However, an AI-calculated input can be used as support.

[4], [18], [19]

10 - Select suitable AI approach for each specific task

The specific AI approach, e.g. the architecture of the NN, should be well selected for each specific task to ensure a maximum of safety, reliability and suitable performance.

[1], [2], [20], [21]

11 - Base implementation on established coding guidelines

If no coding guidelines for AI-based systems exist in the field of the target application, base your implementation on established coding guidelines for conventional programming.

[22], [23]

Training, testing and validation

12 - Define quality gates for training, validation and test data

The training, validation and test data determines the performance and safety of an AI component. It must be ensured that the data represents the real-world application domain, by using quality gates which allow an unbiased evaluation. Top-down technical safety requirements determine mandatory quality gates. The quality of the data shall be monitored with data quality management processes like the DQM-process. The amount of data shall roughly be split up into 70% training data, 20% validation data and 10% test data.

[24], [25]

13 - Define task specific validation metrics

The learning process of a Neural Network is based on the definition of an appropriate set of metrics and a suitable weighting to be optimized for each specific task. In safety-critical application these metrics are derived from Technical Safety Requirements and can for example be the accuracy of a model or its confidence in predictions, which strongly correlates with the safety of the system.

[26], [27], [28], [13]

14 - Apply independent safety assessment

A safety-critical application in which AI is used should be subject to a safety assessment by an independent third party before being launched on the market.

[29], [30], [1]

15 - Develop a maintenance strategy

AI based systems are highly sensible to changes in the input data. A systematic strategy for continuous maintenance is required to mitigate performance decrease by Distributional Shift and its impact on the safety of the system. To properly maintain AI based systems two maintenance strategies shall be incorporated. On the one hand a reactive and preventative maintenance strategy to define responses for failues of the AI system. On the other hand a predictive and adaptive strategy with focus on optimal timing of maintenance measures.

[13], [31], [41]

Sources

[1]	AutoDrive_SC7_Intermediate_Report.pdf, Internal research document.
[2]	N. Papernot and I. Goodfellow, "www.cleverhans.io," 14 June 2017. [Online]. Available: http://www.cleverhans.io/security/privacy/ml/2017/06/14/verification.html. [Accessed January 2022].
[3]	R. Salay, R. Queiroz and K. Czarnecki, "arxiv.org," 2017. [Online]. Available: https://arxiv.org/abs/1709.02435. [Accessed January 2022].
[4]	B. Kaiser, "www.researchgate.net," 2017. [Online]. Available: https://www.researchgate.net/publication/320472118, 2017. [Accessed January 2022].
[5]	D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman and D. Mane, "arxiv.org," 2016. [Online]. Available: https://arxiv.org/abs/1606.06565. [Accessed January 2022].
[6]	C. W. Lee, D. E. G. Nasif Nayeer, A. Agrawal and B. Liu, "ieeexplore," 08 January 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9304552. [Accessed 30 March 2022].
[7]	S. Fort, J. Ren and B. Lakshminarayanan, "arxiv.org," 29 July 2021. [Online]. Available: https://arxiv.org/abs/2106.03004. [Accessed 30 March 2022].
[8]	N. Duggal, "simplilearn.com," 03 March 2022. [Online]. Available: https://www.simplilearn.com/advantages-and-disadvantages-of-artificial-intelligence-article#disadvantages_of_artificial_intelligence. [Accessed 30 March 2022].
[9]	R. Freeman, "towardsdatascience," 24 July 2020. [Online]. Available: https://towardsdatascience.com/when-to-not-use-ai-or-use-it-based-on-my-experience-abb58c063aba. [Accessed 30 March 2022].
[10]	S. M. Grigorescu, M. Glaab and A. Roßbach, "www.elektrobit.com," 2017. [Online]. Available: https://www.elektrobit.com/newsroom/read-new-techpaper-chances-challenges-using-machine-learning-highly-automated-driving/. [Accessed November 2021].
[11]	E. Mueller, "medium.com," 27 May 2021. [Online]. Available: https://medium.com/capital-one-tech/end-to-end-models-for-complex-ai-tasks-8c34080145cd. [Accessed 30 March 2022].
[12]	I. Sydorenko, "labelyourdata.com," 04 May 2021. [Online]. Available: https://labelyourdata.com/articles/how-to-choose-a-machine-learning-algorithm. [Accessed 30 March 2022].
[13]	T. Sämann, P. Schlicht and F. Hüger, "arxiv.org," 2020. [Online]. Available: https://arxiv.org/abs/2002.08935. [Accessed January 2022].

[14]	E. Lutins, "towardsdatascience.com," 2 August 2017. [Online]. Available: https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f. [Accessed November 2021].
[15]	C. Lundquist, "www.diva-portal.org," 2011. [Online]. Available: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A451021&dswid=6497. [Accessed January 2022].
[16]	J. Kocic, N. Jovicic and V. Drndarevic, "ieeexplore.ieee.org," 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8612054. [Accessed January 2022].
[17]	M. H. Osman, S. Kugele and S. Shafaei, "ieeexplore.org," 02 January 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8945611. [Accessed 30 March 2022].
[18]	R. Smith, "becominghuman.ai," 14 July 2020. [Online]. Available: https://becominghuman.ai/the-key-differences-between-rule-based-ai-and-machine-learning-8792e545e6. [Accessed November 2021].
[19]	J. M. Carew, "searchenterpriseai.techtarget.com," 23 July 2020. [Online]. Available: https://searchenterpriseai.techtarget.com/feature/How-to-choose-between-a-rules-based-vs-machine-learning-system. [Accessed November 2021].
[20]	G. D. Luca, "www.baeldung.com," 24 August 2020. [Online]. Available: https://www.baeldung.com/cs/neural-networks-hidden-layers-criteria. [Accessed November 2021].
[21]	A. Tch, "towardsdatascience.com," 14 August 2017. [Online]. Available: https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464. [Accessed November 2021].
[22]	H. M. Limited, "Misra.org," 2020. [Online]. Available: https://www.misra.org.uk/app/uploads/2021/06/MISRA-Compliance-2020.pdf. [Accessed 04 April 2022].
[23]	M. Barr, "barrgroup.com," 2018. [Online]. Available: https://barrgroup.com/sites/default/files/barr_c_coding_standard_2018.pdf. [Accessed 05 April 2022].
[24]	P. Gupta, "towardsdatascience.com," 5 June 2017. [Online]. Available: https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f. [Accessed November 2021].
[25]	W. Koehrsen, "towardsdatascience.com," 28 January 2018 . [Online]. Available: https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765. [Accessed November 2021].
[26]	S. Y. e. al, "arxiv.org," [Online]. Available: https://arxiv.org/abs/1803.04792.

[27]	A. Kumar, "dzone.com," 28 January 2018. [Online]. Available: https://dzone.com/articles/machine-learning-validation-techniques. [Accessed November 2021].
[28]	V&V of NN for Aerospace Systems.pdf, Internal research document.
[29]	T. A. LAB, "www.tuev-verband.de," 2021. [Online]. Available: https://www.tuev-verband.de/?tx_epxelo_file[id]=850772&cHash=f1c6c52b992bbfd2b3dd7598fa5477e3. [Accessed January 2022].
[30]	E. Kommission, "op.europa.eu," 2019. [Online]. Available: https://op.europa.eu/de/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1. [Accessed Januar 2022].
[31]	S. H. S. Bhasker, M. Srivastava and K. Yeturu, "doi.org," 2021. [Online]. Available: https://doi.org/10.36227/techrxiv.14565078.v1. [Accessed January 2022].
[32]	J. Brownlee, "machinelearningmastery.com," 6 August 2019. [Online]. Available: https://machinelearningmastery.com/why-training-a-neural-network-is-hard/. [Accessed November 2021].
[33]	T. D. Krafft and K. A. Zweig, "www.vzbv.de," 2019. [Online]. Available: https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-neu.pdf. [Accessed Januar 2022].
[34]	B. J. Taylor, M. A. Darrah and C. D. Moats, "www.spiedigitallibrary.org," 2003. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/5103/0000/Verification-and-validation-of-neural-networks--a-sampling-of/10.1117/12.487527.short?SSO=1. [Accessed January 2022].
[35]	P. S. F. H. Timo Sämann, „Strategy to Increase the Safety of a DNN-based," 20 02 2020. [Online]. Available: https://arxiv.org/pdf/2002.08935.pdf. [Zugriff am 2022 11 25].
[36]	N. J. V. D. Jelena Kocić, „Sensors and Sensor Fusion in Autonomous Vehicles," 20 11 2018. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8612054. [Zugriff am 2022 11 25].
[37]	Tae-Hwy Lee, Aman Ullah and Ran Wang, "Bootstrap Aggregating and Random Forest", 08 01 2019, [Online], Available: https://economics.ucr.edu/repec/ucr/wpaper/201918.pdf. [Zugriff am 2023 03 24]
[38]	E. Lutins, „Ensemble Methods in Machine Learning: What are They and Why Use Them?," 02 08 2017. [Online]. Available: https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f. [Zugriff am 25 11 2022].

[39]	H. Putzer, H. Rueß and J. Koch: Trustworthy AI-based Systemswith VDE-AR-E 2842-61, 2021
[40]	M. Trapp & G. Weiss: Towards Dynamic Safety Management for Autonomous Systems, Fraunhofer ESK, München, 2019
[41]	S. H. S. Bhasker, M. Srivastava and K. Yeturu: Methods for maintenance of neural networks incontinual learning scenarios, 2021.