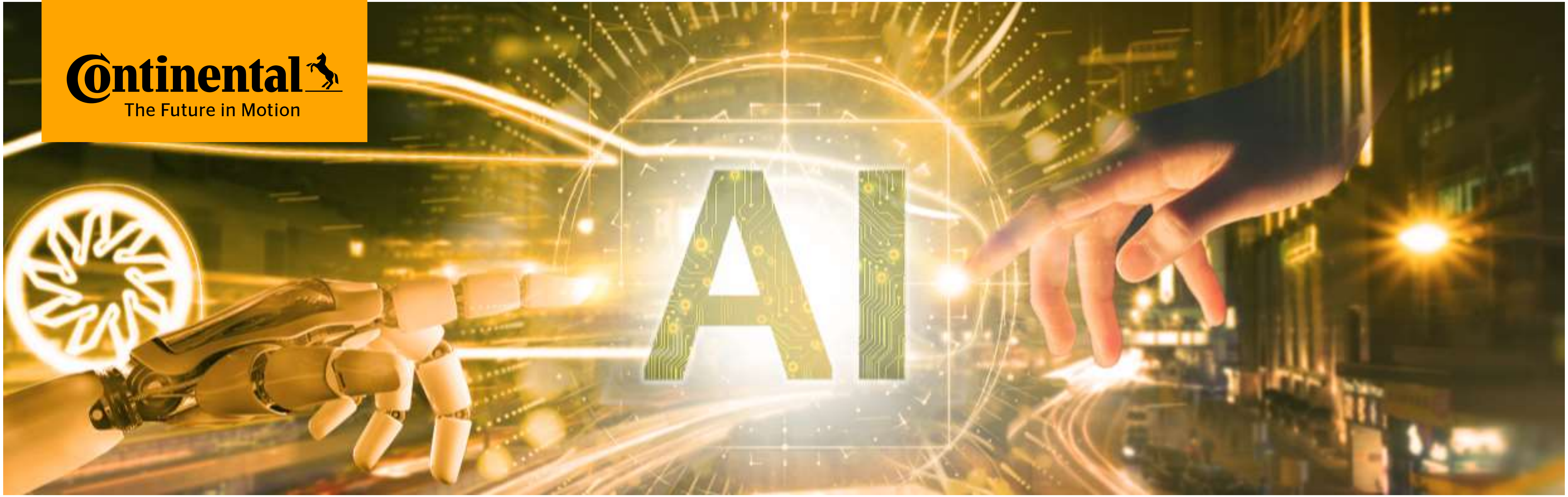
Continental
The Future in Motion



Methods for the Safety Assurance of Perception DNNs in AD

An Overview

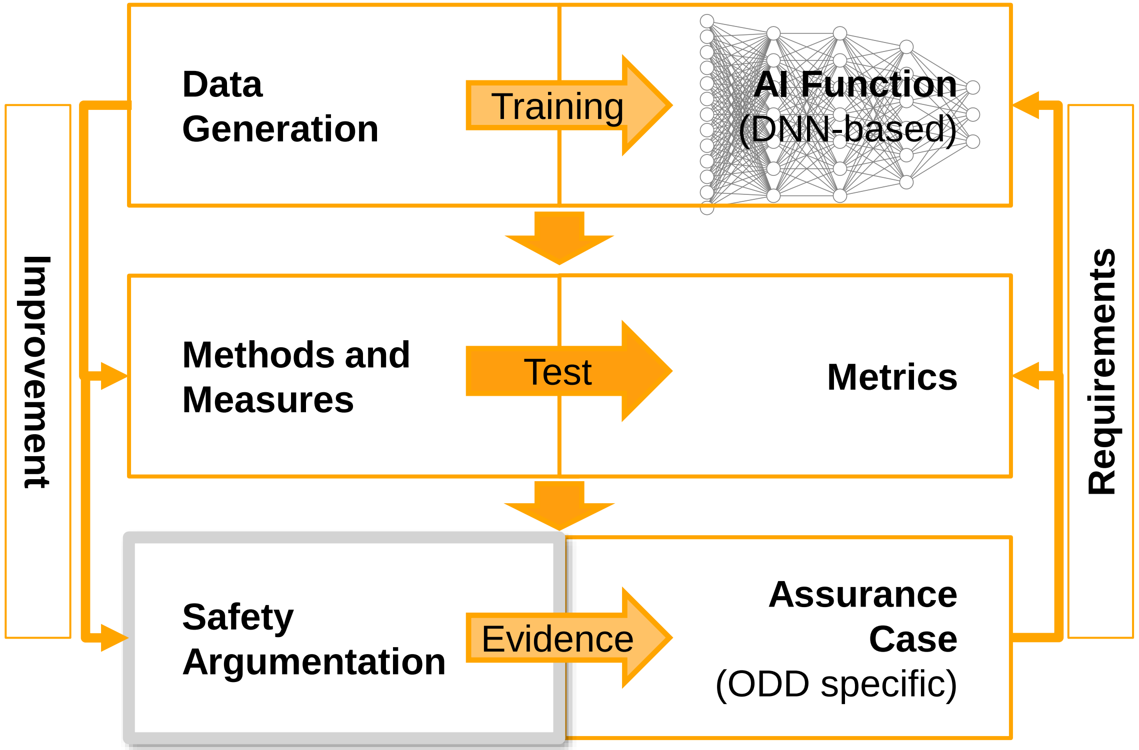
Dr. Gesina Schwalbe

www.continental.com

Background: Project KI-Absicherung

Vision KI Absicherung makes the **safety of AI**-based function modules for **highly automated driving provable**.

Use-case Camera/LiDAR based single frame **pedestrian detection**



see also https://www.ki-absicherung.vdali.de

Agenda

Agenda

Automotive Safety Basics
Safety

Def. Safety

[…] according to valid societal moral concepts (ISO 26262-1, 3.176)

means *absence of unreasonable risk* due to
- malfunction (ISO 26262-1, 3.132)
- intended functionality
  (foreseeable misuse, performance limitation wrt. environment)
  (ISO/PAS 21448)

Rating safety:

Safety Integrity Levels (ISO 26262-3, 6.4.3) derived from

› **Probability**

› **Severity**

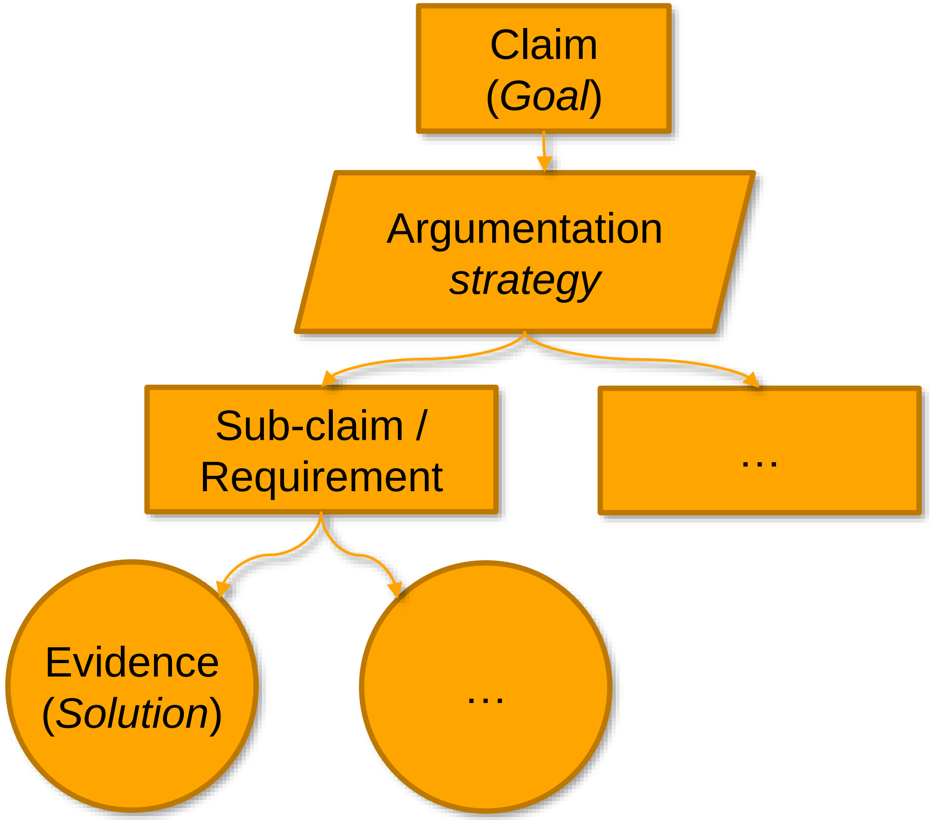
› Controllability

Automotive Safety Basics
Safety Case

Def.: Safety Case
is a documented body of *evidence*
providing convincing and valid *argument*
that a system is adequately safe
for a *given application* in a *given environment*

Arguments may be
- Deterministic
- Probabilistic
- Qualitative(!)

Evidence types:
- Design & process
  - System level
  - Unit level
- Verification
- (Experience)



(Bishop and Bloomfield 1998)

Agenda

Agenda

Goal? What? When?

What is mitigated?
Safety Requirements

Def.: DNN-specific Safety Concern
underlying issues of AI-based perception which may negatively affect the safety of a system

e.g., snow

→ **Treat all possible combinations!**

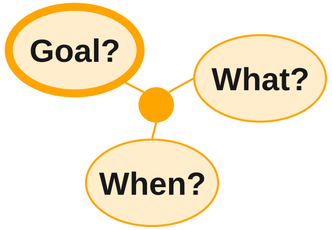
Triggering Event

DNN-specific Safety Concerns → Insufficient Generalization Capability

Triggering Event + Insufficient Generalization Capability → Erroneous Output → Failure: Unsafe Control Action → Hazard

see also https://www.ki-absicherung.vdali.de; (Willers et al. 2020) Fig. 1

What is mitigated?
Safety Concerns



AI Specifics

- Unreliable **confidence** information
- **Brittleness** (e.g. against perturbations, lack of temporal stability)
- **Incomprehensible** behavior
- Insufficient **plausibility**

Data

- Labeling **quality** (e.g. wrong/missing (meta-)labels)
- Train/test data **separation**
- **Representativity**
- Inadequate **ODD** spec.
- **Distributional shift** over time
- Unknown behavior in **rare critical situations**

Metrics

- Safety relevant metric

Others

- Lack of algorithmic efficiency (e.g. memory use, power consumption, frames/second)

Agenda

What is changed?

Goal?
What?
When?

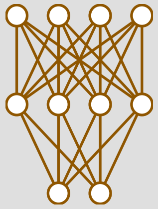
Mechanisms during creation

Specification & guidelines for:

- DNN design
- **Dataset** collection
- **Training**

Mechanisms on system level

Detect & prevent at:



- Input
- **Internal state**
- Output

Verification

- Testing
- (Semi-)formal
- Inspect **via proxy**

Validation

via traditional validation testing
(e.g. endurance run)

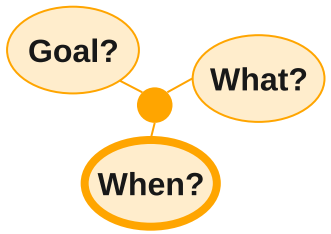
Ensure test data representativity
cover:

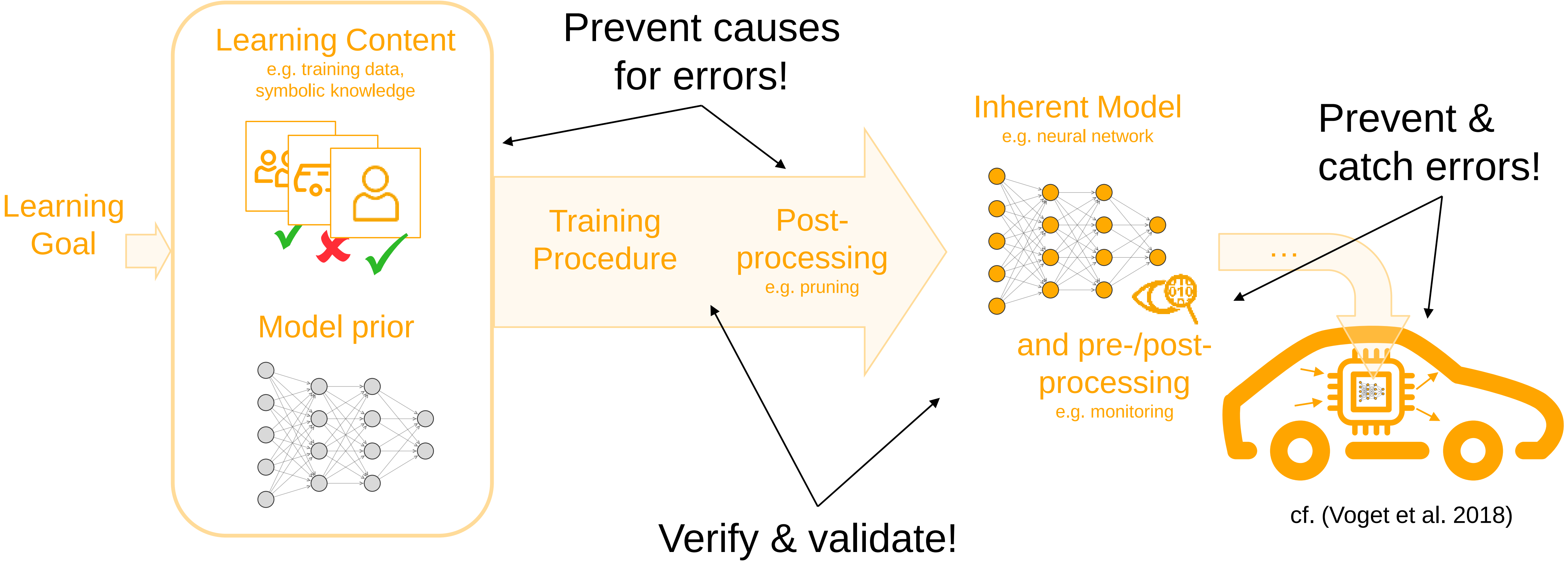
- Experience
- Semantic features
- **Learned features**

(Schwalbe et al. 2020)

Agenda

When is it applied?





cf. (Voget et al. 2018)

Agenda

Validation of an Evidence Method

- *For verification in general:* Method for obtaining evidences is ...
  - **Appropriate** (theoretically measures what is needed)
  - **Known** to work
  - **Applicable** / correctly applied

- *For performance claims:*
  - Correct metrics & **deduction**
  - Statistical **significance** wrt. claim
  - Representative **test data**
  - Appropriate ML **model**
  - Assumptions on **environment** and surrounding **system**

Creation
Training Data Optimization (Shorten and Khoshgoftaar 2019)

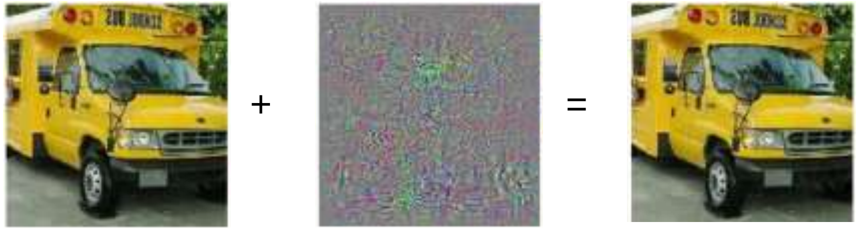
- **Dataset diversity**: *e.g.,*
  - Image manipulation
    - Addition of artifacts cf. (Zendel et al. 2015)
    - Domain randomization
  - Synthetic data generation / augmentation
  - Counterexample generation (Dreossi et al. 2018)
- **Image selection**: *e.g.,* Active learning
- Other topics: Label quality, data representativity & fidelity

(Eykholt et al. 2018, Tab. 1)

"speed limit 45"

(Guo et al. 2018, Fig. 1, p. 2)

in:



out: "bus" "ostrich"

(Geirhos et al. 2019), Fig. 1

Creation
Architecture and Training Objective

- Explainable intermediate output, *e.g.*
  - Concept Bottlenecks
    (Losch et al. 2020), (Koh et al. 2020)
  - Attention heatmaps

(Kim and Canny 2017), Fig. 5

- Soft training constraints, *e.g.*
  - Hierarchical Movables ⊃ Persons, Movables ⊃ Cars
    (Roychowdhury, Diligenti, and Gori 2018)
  - Locality of activations
  - Robustness against perturbations
  - Temporal Consistency
    (Varghese et al. 2021)

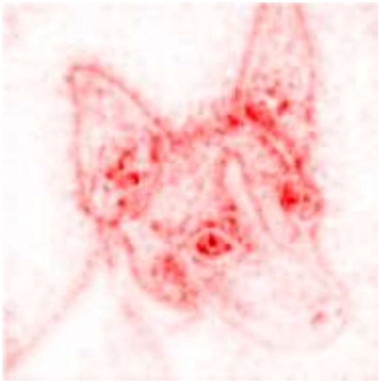
- Proper uncertainty output, *e.g.* via
  - Ensembling
  - Bayesian DNNs

(Kendall & Gal 2017, Fig. 1, p. 2)

Offline Verification
Quantitative Explainable AI

“Attention” heatmap-methods for plausibility checks, *e.g.*

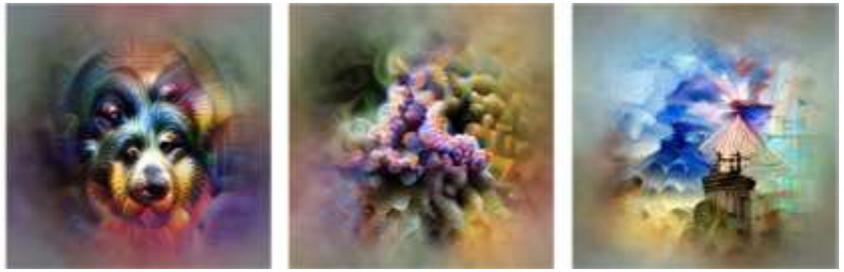
- White-box (gradients, relevance back-propagation, …)
- Black-box (occlusion based, perturbation based …)



(Kindermans et al. 2018), Fig. 6

Knowledge V&V by disentanglement of internal semantics

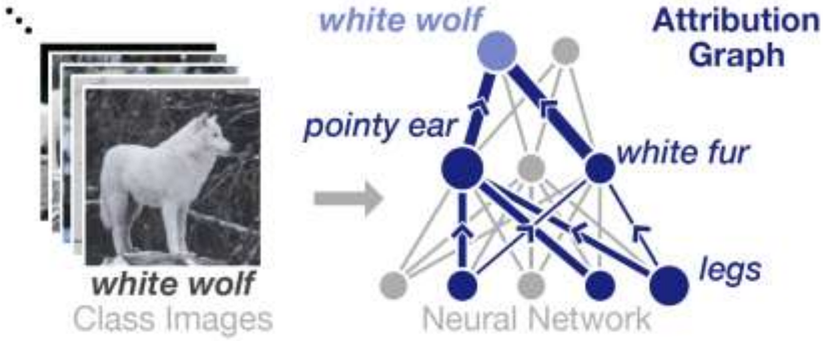
- Mining of learned concepts
  (Ge et al. 2021), (Zhang et al. 2021), (Esser et al. 2020)



(Olah et al. 2017), Fig. 5

- Interpretable proxy models

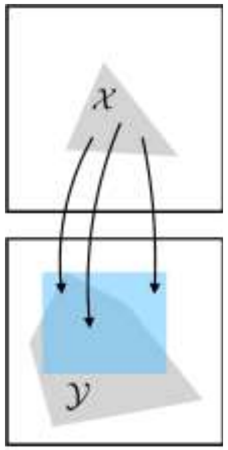
- Properties of learned concepts (e.g., similarity)
  (Fong and Vedaldi 2018), (Schwalbe and Schels 2020)



(Hohman et al. 2020), Fig. 2

Offline Verification
Formal Methods

(Liu et al. 2019), Fig. 2



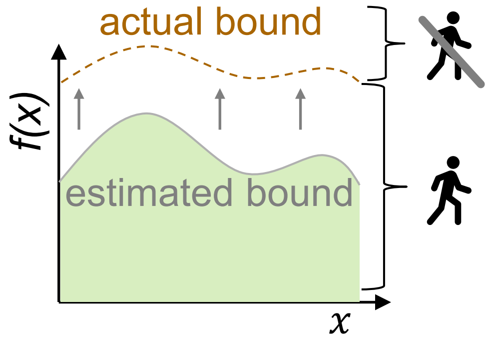
(c) Reachability result.

Formal verification

- **Goals:** Find
  - counterexamples
  - validity range
  - reachable set

- **Methods Examples:**

  layer-by-layer reachability /
  boundary estimation,
  (constrained) optimization,
  search,
  solvers



(Formal) Testing

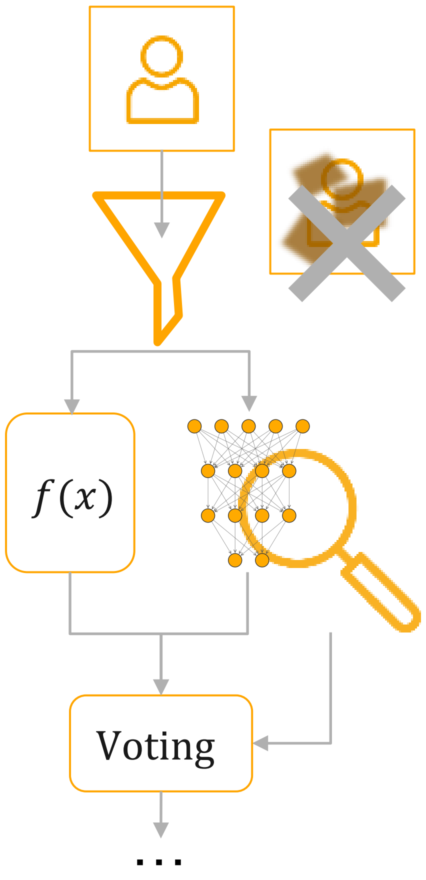
- **Goals:**
  - Semantic coverage e.g. via SDL & sampler
  - Latent space coverage (direct & indirect)

- **Methods Examples:**

  Differential (Pei et al. 2017),
  fuzzy (Odena et al. 2019),
  concolic (Sun et al. 2018)

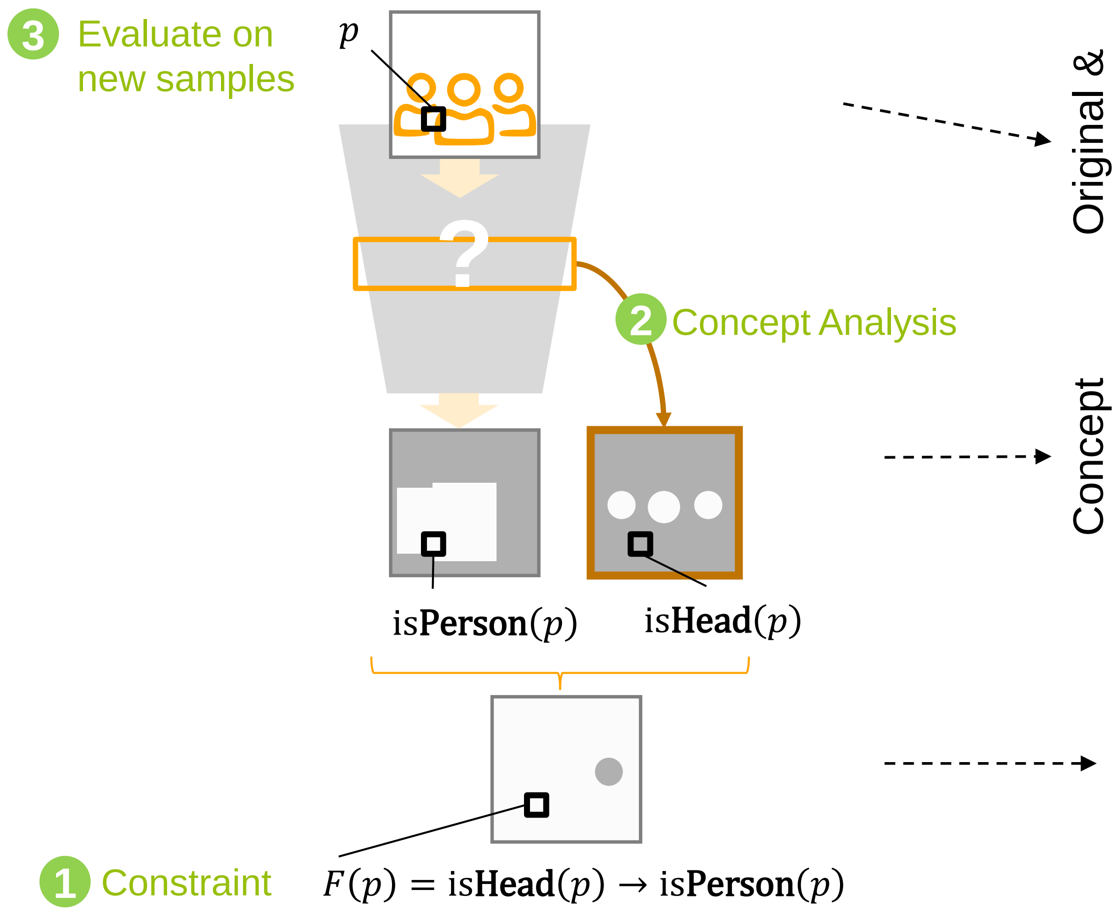
Online Verification: System level measures

- Input filtering (Ilyas et al. 2019), (Kapoor et al. 2020)
- Redundancy & voting
- Monitoring for
  - Out-of-distribution, *e.g.*, via
    - Uncertainty estimation
  - Plausibility / consistency with constraints, *e.g.*,
    - Temporal consistency (Varghese et al. 2020)
    - Local stability
    - Semantic constraints on outputs
      (Schwalbe 2021), (Giunchiglia et al. 2022)
- Error handling, *e.g.*, via removal, correction, additional queries, …



Offline Verification

Example: Explainable AI to Verify Logical Constraints



(Schwalbe et al. 2022)

Agenda

Conclusion

Technologies involved in safety assurance are **highly diverse**

Categories: Goal? Target element? When (in lifecycle)?

- Creation ("build it right")
- V&V ("check it right")
- System design ("prevent / mitigate failing in op")

To provide convincing evidence method must be
applicable, appropriate, known to work; results documented

References (I)

- Bishop, P. G., and R. E. Bloomfield. 1998. “A Methodology for Safety Case Development.” In *Industrial Perspectives of Safety-Critical Systems: Proc. Safety-Critical Systems Symp.*, edited by F. Redmill and T. Anderson. Birmingham, UK: Springer London. http://openaccess.city.ac.uk/549/.
- Burton, Simon, Christian Hellert, Fabian Hüger, Michael Mock, and Andreas Rohatschek. 2022. “Safety Assurance of Machine Learning for Perception Functions.” In *Deep Neural Networks and Data for Automated Driving: Robustness, Uncertainty Quantification, and Insights Towards Safety*, edited by Tim Fingscheidt, Hanno Gottschalk, and Sebastian Houben, 335–58. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-01233-4_12.
- Dreossi, Tommaso, Shromona Ghosh, Xiangyu Yue, Kurt Keutzer, Alberto L. Sangiovanni-Vincentelli, and Sanjit A. Seshia. 2018. “Counterexample-Guided Data Augmentation.” In *Proc. 27th Int. Joint Conf. Artificial Intelligence*, edited by Jérôme Lang, 2071–2078. Stockholm, Sweden: ijcai.org. https://doi.org/10.24963/ijcai.2018/286.
- Eykholt, Kevin, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. “Robust Physical-World Attacks on Deep Learning Visual Classification.” In *Proc. 2018 IEEE Conf. Computer Vision and Pattern Recognition*, 1625–34. Salt Lake City, UT, USA: IEEE Computer Society. https://doi.org/10.1109/CVPR.2018.00175.
- Esser, Patrick, Robin Rombach, and Bjorn Ommer. 2020. “A Disentangling Invertible Interpretation Network for Explaining Latent Representations.” In *Proc. 2020 IEEE Conf. Comput. Vision and Pattern Recognition*, 9220–29. Seattle, WA, USA: IEEE. https://doi.org/10.1109/CVPR42600.2020.00924.
- Fong, Ruth, and Andrea Vedaldi. 2018. “Net2Vec: Quantifying and Explaining How Concepts Are Encoded by Filters in Deep Neural Networks.” In *Proc. 2018 IEEE Conf. Comput. Vision and Pattern Recognition*, 8730–8738. Salt Lake City, UT, USA: IEEE Computer Society. https://doi.org/10.1109/CVPR.2018.00910.
- Ge, Yunhao, Yao Xiao, Zhi Xu, Meng Zheng, Srikrishna Karanam, Terrence Chen, Laurent Itti, and Ziyan Wu. 2021. “A Peek into the Reasoning of Neural Networks: Interpreting with Structural Visual Concepts.” In *Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2195–2204. https://openaccess.thecvf.com/content/CVPR2021/html/Ge_A_Peek_Into_the_Reasoning_of_Neural_Networks_Interpreting_With_CVPR_2021_paper.html.
- Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. “ImageNet-Trained CNNs Are Biased towards Texture; Increasing Shape Bias Improves Accuracy and Robustness.” In *Proc. 7th Int. Conf. Learning Representations*. New Orleans, LA, USA: OpenReview.net. https://openreview.net/forum?id=Bygh9j09KX.
- Giunchiglia, Eleonora, Mihaela Stoian, Salman Khan, Fabio Cuzzolin, and Thomas Lukasiewicz. 2022. “ROAD-R: The Autonomous Driving Dataset with Logical Requirements.” In *IJCLR 2022 Workshops*. Vienna, Austria. https://arxiv.org/abs/2210.01597.
- Guo, Jianmin, Yu Jiang, Yue Zhao, Quan Chen, and Jiaguang Sun. 2018. “DLFuzz: Differential Fuzzing Testing of Deep Learning Systems.” In *Proc. ACM Joint Meeting on European Software Engineering Conf. and Symp. Foundations of Software Engineering*, 739–43. Lake Buena Vista, FL, USA: ACM. https://doi.org/10.1145/3236024.3264835.
- Ilyas, Andrew, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. “Adversarial Examples Are Not Bugs, They Are Features.” In *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, 125–36. Vancouver, Canada: Curran Associates, Inc. http://papers.nips.cc/paper/8307-adversarial-examples-are-not-bugs-they-are-features.pdf.
- ISO/TC 22/SC 32. 2018. *ISO 26262-1:2018(En): Road Vehicles — Functional Safety — Part 1: Vocabulary*. 2nd ed. Vol. 1. 11 vols. ISO 26262:2018(En). Vernier, Geneva: International Organization for Standardization. https://www.iso.org/standard/68383.html.
- ISO/TC 22/SC 32. 2018. *ISO 26262-3:2018(En): Road Vehicles — Functional Safety — Part 3: Concept Phase*. 2nd ed. Vol. 3. 11 vols. ISO 26262:2018(En). Vernier, Geneva: International Organization for Standardization. https://www.iso.org/standard/68385.html.
- Kapoor, Nikhil, Andreas Bär, Serin Varghese, Jan David Schneider, Fabian Hüger, Peter Schlicht, and Tim Fingscheidt. 2020. “From a Fourier-Domain Perspective on Adversarial Examples to a Wiener Filter Defense for Semantic Segmentation.” *CoRR* abs/2012.01558 (December). https://arxiv.org/abs/2012.01558.
- Kendall, Alex, and Yarin Gal. 2017. “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?” In *Advances in Neural Information Processing Systems 30*, 5580–90. Long Beach, CA, USA. http://papers.nips.cc/paper/7141-what-uncertainties-do-we-need-in-bayesian-deep-learning-for-computer-vision.
- Kim, Jinkyu, and John F. Canny. 2017. “Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention.” In *Proc. 2017 IEEE Int. Conf. Comput. Vision*, 2961–2969. Venice, Italy: IEEE Computer Society. https://doi.org/10.1109/ICCV.2017.320.

References (II)

- Kindermans, Pieter-Jan, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. 2018. "Learning How to Explain Neural Networks: PatternNet and PatternAttribution." In *Proc. 6th Int. Conf. on Learning Representations*. Vancouver, Canada. https://openreview.net/forum?id=Hkn7CBaTW.
- Koh, Pang Wei, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. "Concept Bottleneck Models." In *Proc. 2020 Int. Conf. Machine Learning*, 5338–48. PMLR. http://proceedings.mlr.press/v119/koh20a.html.
- Liu, Changliu, Tomer Arnon, Christopher Lazarus, Clark W. Barrett, and Mykel J. Kochenderfer. 2019. "Algorithms for Verifying Deep Neural Networks." *CoRR* abs/1903.06758 (March). http://arxiv.org/abs/1903.06758.
- Losch, Max, Mario Fritz, and Bernt Schiele. 2019. "Interpretability beyond Classification Output: Semantic Bottleneck Networks." In *Proc. 3rd ACM Computer Science in Cars Symp. Extended Abstracts*. Kaiserslautern, Germany. https://arxiv.org/pdf/1907.10882.pdf.
- Odena, Augustus, Catherine Olsson, David Andersen, and Ian Goodfellow. 2019. "TensorFuzz: Debugging Neural Networks with Coverage-Guided Fuzzing." In *Proc. 36th Int. Conf. on Machine Learning*, 4901–11. Proceedings of Machine Learning Research. Long Beach, CA, USA. http://proceedings.mlr.press/v97/odena19a.html.
- Pei, Kexin, Yinzhi Cao, Junfeng Yang, and Suman Jana. 2017. "DeepXplore: Automated Whitebox Testing of Deep Learning Systems." In *Proc. 26th Symp. Operating Systems Principles*, abs/1705.06640:1–18. Shanghai, China: ACM. https://doi.org/10.1145/3132747.3132785.
- Roychowdhury, Soumali, Michelangelo Diligenti, and Marco Gori. 2018. "Image Classification Using Deep Learning and Prior Knowledge." In *Workshops of the 32nd AAAI Conf. Artificial Intelligence*, WS-18:336–343. AAAI Workshops. New Orleans, Louisiana, USA: AAAI Press. https://aaai.org/ocs/index.php/WS/AAAIW18/paper/view/16575.
- Schwalbe, Gesina, Bernhard Knie, Timo Sämann, Timo Dobberphul, Lydia Gauerhof, Shervin Raafatnia, and Vittorio Rocco. 2020. "Structuring the Safety Argumentation for Deep Neural Network Based Perception in Automotive Applications." In *Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops*, edited by António Casimiro, Frank Ortmeier, Erwin Schoitsch, Friedemann Bitsch, and Pedro Ferreira, 383–94. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-55583-2_29.
- Shorten, Connor, and Taghi M. Khoshgoftaar. 2019. "A Survey on Image Data Augmentation for Deep Learning." *Journal of Big Data* 6 (1): 60. https://doi.org/10.1186/s40537-019-0197-0.
- Sun, Youcheng, Min Wu, Wenjie Ruan, Xiaowei Huang, Marta Kwiatkowska, and Daniel Kroening. 2018. "Concolic Testing for Deep Neural Networks." In *Proc. 33rd ACM/IEEE Int. Conf. Automated Software Engineering*, 109–119. Montpellier, France: ACM. https://doi.org/10.1145/3238147.3238172.
- Varghese, Serin, Yasin Bayzidi, Andreas Bar, Nikhil Kapoor, Sounak Lahiri, Jan David Schneider, Nico M. Schmidt, Peter Schlicht, Fabian Huger, and Tim Fingscheidt. 2020. "Unsupervised Temporal Consistency Metric for Video Segmentation in Highly-Automated Driving." In , 336–37. https://openaccess.thecvf.com/content_CVPRW_2020/html/w20/Varghese_Unsupervised_Temporal_Consistency_Metric_for_Video_Segmentation_in_Highly-Automated_Driving_CVPRW_2020_paper.html.
- Varghese, Serin, Sharat Gujamagadi, Marvin Klingner, Nikhil Kapoor, Andreas Bär, Jan David Schneider, Kira Maag, Peter Schlicht, Fabian Hüger, and Tim Fingscheidt. 2021. "An Unsupervised Temporal Consistency (TC) Loss to Improve the Performance of Semantic Segmentation Networks." In *2021 IEEE/CVF Conf. Comput. Vision and Pattern Recognition Workshops*, 12–20. https://doi.org/10.1109/CVPRW53098.2021.00010.
- Voget, Stefan, Alexander Rudolph, and Jürgen Mottok. 2018. "A Consistent Safety Case Argumentation for Artificial Intelligence in Safety Related Automotive Systems." In *Proc. 9th European Congress Embedded Real Time Systems*. Toulouse, France. https://hal.archives-ouvertes.fr/hal-02156048 .
- Willers, Oliver, Sebastian Sudholt, Shervin Raafatnia, and Stephanie Abrecht. 2020. "Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks." In *Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops*, edited by António Casimiro, Frank Ortmeier, Erwin Schoitsch, Friedemann Bitsch, and Pedro Ferreira, 336–50. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-55583-2_25.
- Zendel, O., M. Murschitz, M. Humenberger, and W. Herzner. 2015. "CV-HAZOP: Introducing Test Data Validation for Computer Vision." In *2015 IEEE International Conference on Computer Vision (ICCV)*, 2066–74. https://doi.org/10.1109/ICCV.2015.239.
- Zhang, Ruihan, Prashan Madumal, Tim Miller, Krista A. Ehinger, and Benjamin I. P. Rubinstein. 2021. "Invertible Concept-Based Explanations for CNN Models with Non-Negative Concept Activation Vectors." In *Proc. 35th AAAI Conf. Artificial Intelligence*, 35:11682–90. virtual: AAAI Press. https://ojs.aaai.org/index.php/AAAI/article/view/17389.