Robotics and Embedded Systems

Seminar "Reinforcement Learning in Autonomous Driving"

Safe Reinforcement Learning and the Future of ISO 26262

Jonas Riebel & Max Ulmke

Summer Semester 2018

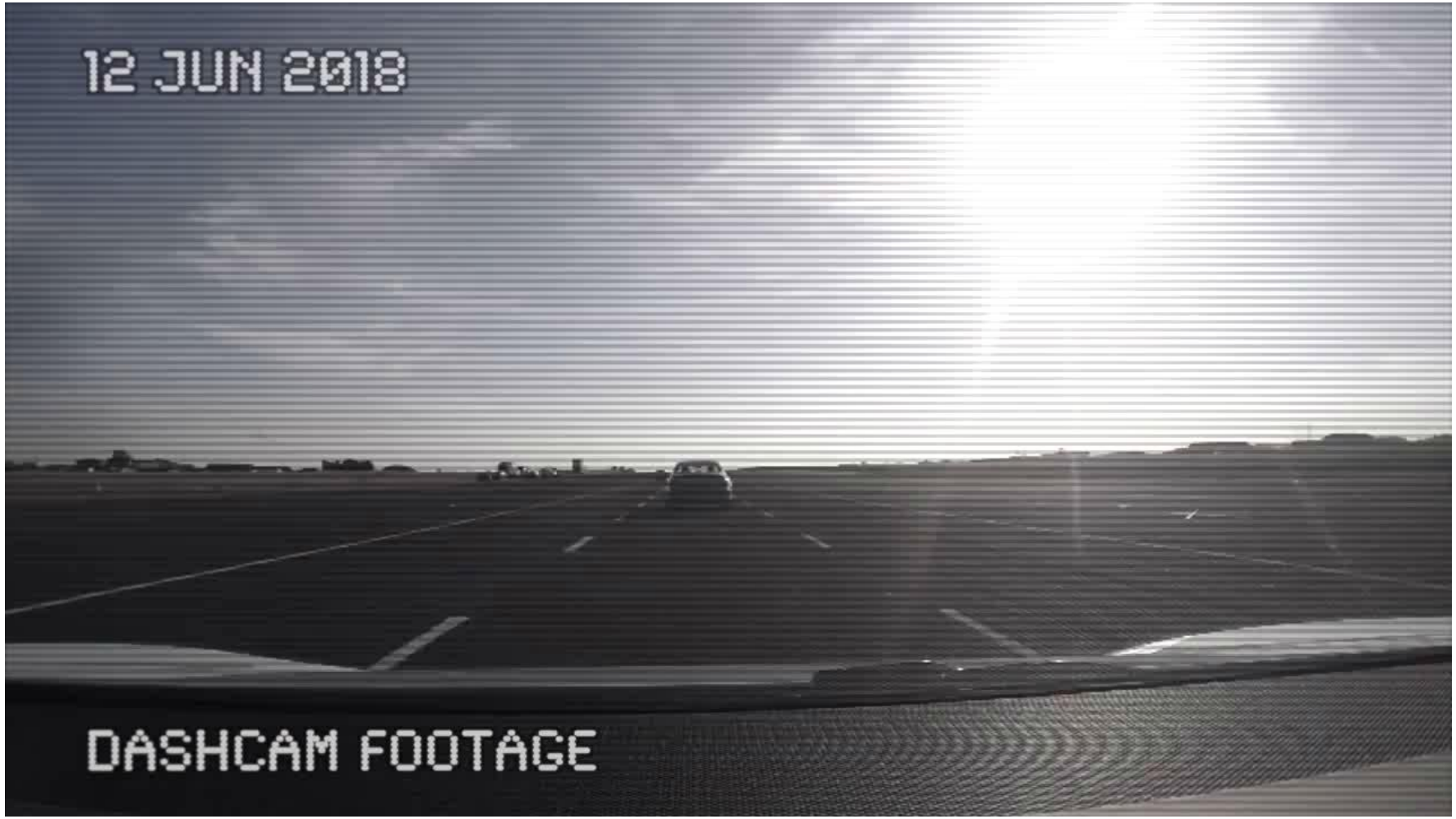
Agenda

1. Introduction
2. How to combine Machine Learning and ISO 26262?
3. Solution Approaches
4. Conclusion and Outlook

Agenda

1. Introduction
    1. What is safety?
    2. Functional Safety
    3. ISO26262
    4. ASIL
    5. Changes from first to second edition
2. How to combine machine learning and ISO 26262?
3. How to move on from here?
4. Conclusion and Outlook

How to prove that it is safe?



Source: Source: https://www.thatcham.org/

Safety of technical systems

- What is safety?

  Safety: „Absence of unreasonable risk“
  [ISO26262]

- Risk must be below a certain limit

“Risk: combination of the probability of occurrence of harm and the severity of that harm” [ISO26262]

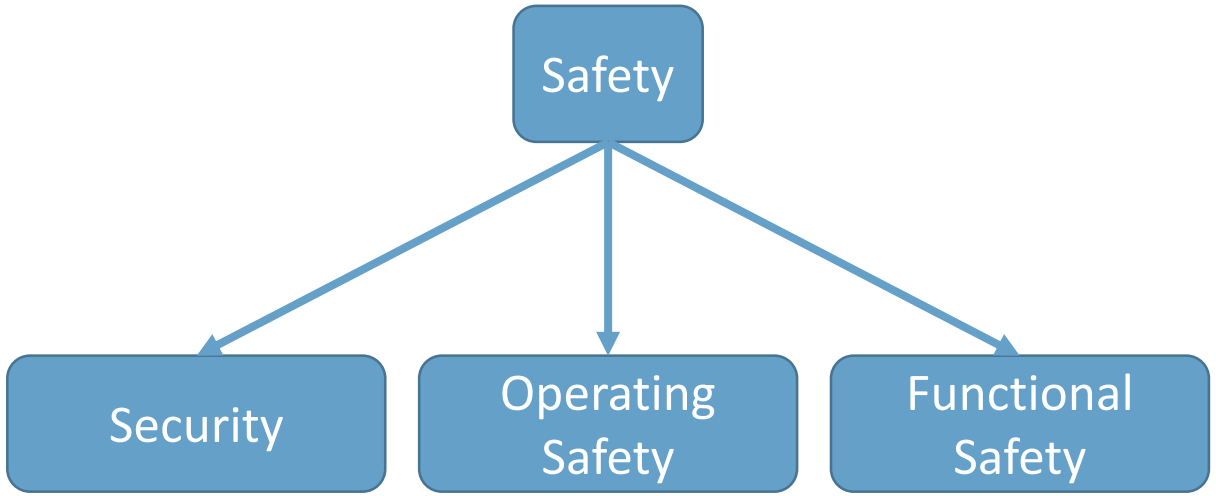
Safety of technical systems

- What is safety?

  Safety: „Absence of unreasonable risk“
  [ISO26262]

- Risk must be below a certain limit

“Risk: combination of the probability of occurrence of harm and the severity of that harm” [ISO26262]



Safety of technical systems

- What is safety?

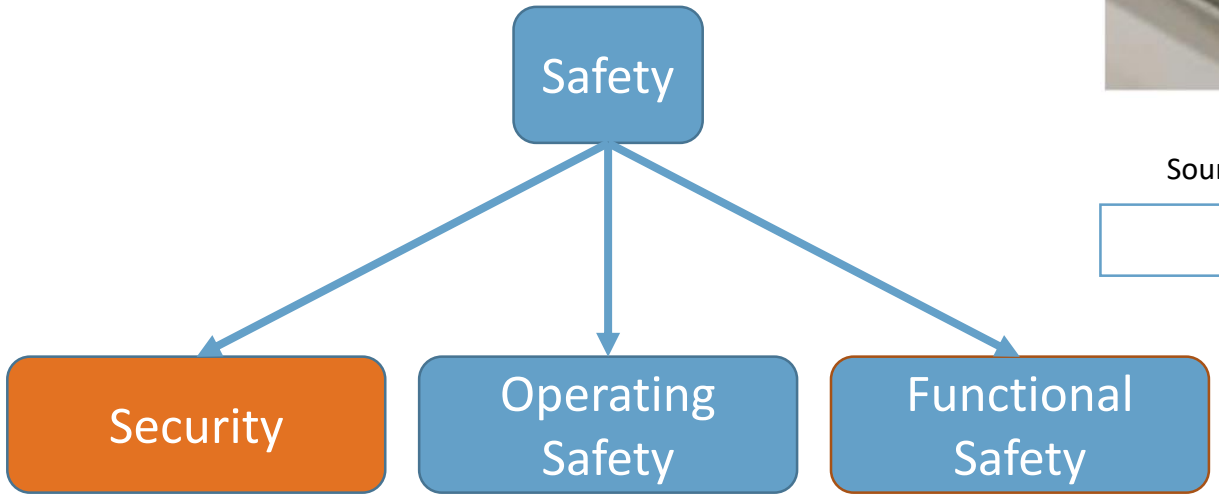
  Safety: „Absence of unreasonable risk“ [ISO26262]

- Risk must be below a certain limit

“Risk: combination of the probability of occurrence of harm and the severity of that harm” [ISO26262]

Source: Lecture Advanced Deep Learning for Robotics, Bäuml, SS18, TUM

Adversarial Example



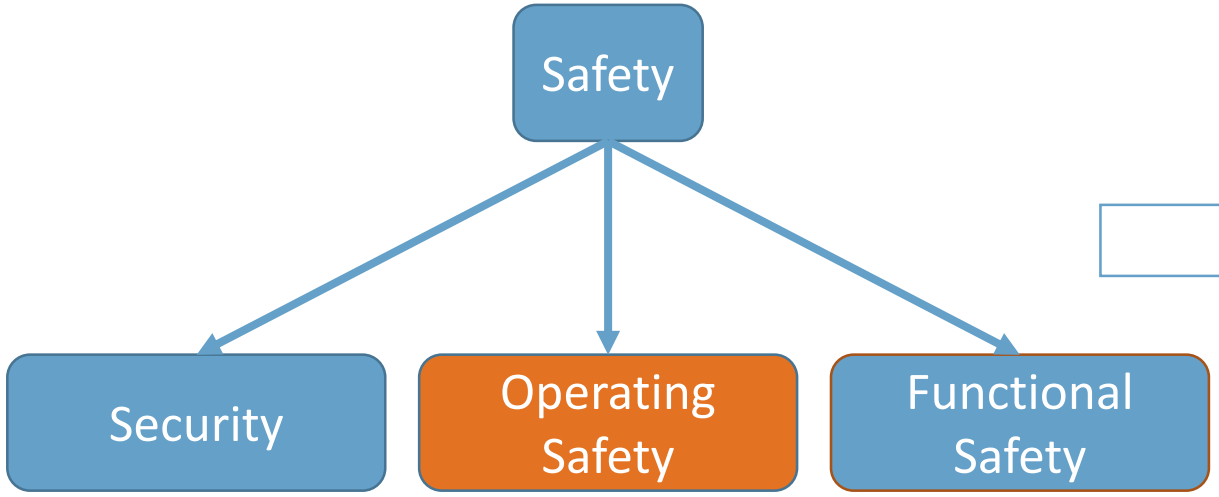
Safety of technical systems

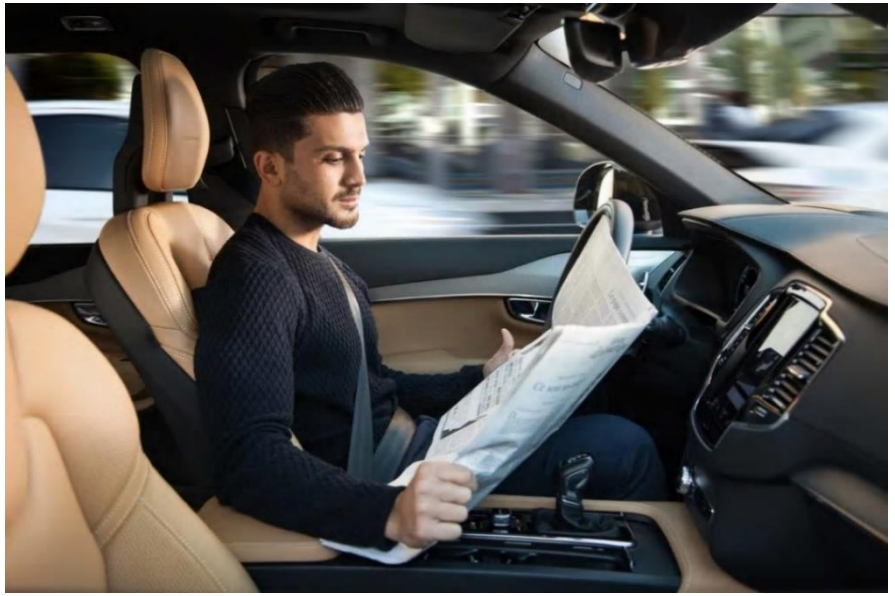
- What is safety?

  Safety: „Absence of unreasonable risk“ [ISO26262]

- Risk must be below a certain limit

“Risk: combination of the probability of occurrence of harm and the severity of that harm” [ISO26262]





Source: https://inews.co.uk/essentials/lifestyle/cars/car-news/drivers-doubt-future-autonomous-cars/

Example for Misuse of a level 2 car.

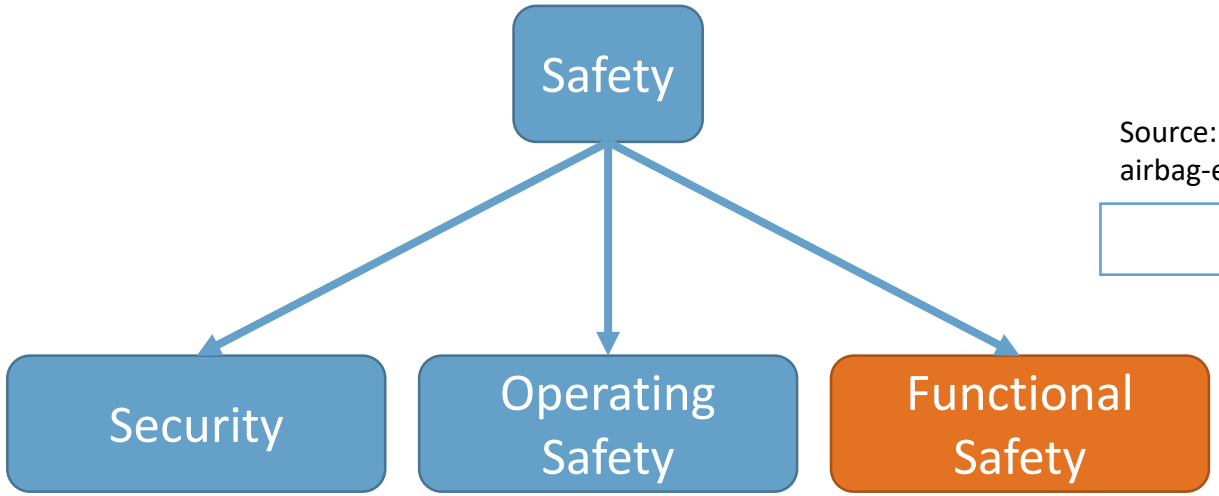
Safety of technical systems

- What is safety?

  Safety: „Absence of unreasonable risk“ [ISO26262]

- Risk must be below a certain limit

“Risk: combination of the probability of occurrence of harm and the severity of that harm” [ISO26262]



Source: Giddens Family - https://dfw.cbslocal.com/2016/05/18/arlington-family-sues-takata-claims-airbag-endangered-their-teen/

Failed Takata Airbag

Functional safety

- Potential failure of a system in different situation
- Goal:
  - show potential hazards and classify them
  - Reduction of hazards caused by E/E systems
- Outcome:
  - safety integrity level SIL/ASIL
  - Definition of processes and countermeasure to prevent failures
  - Regulated by standards

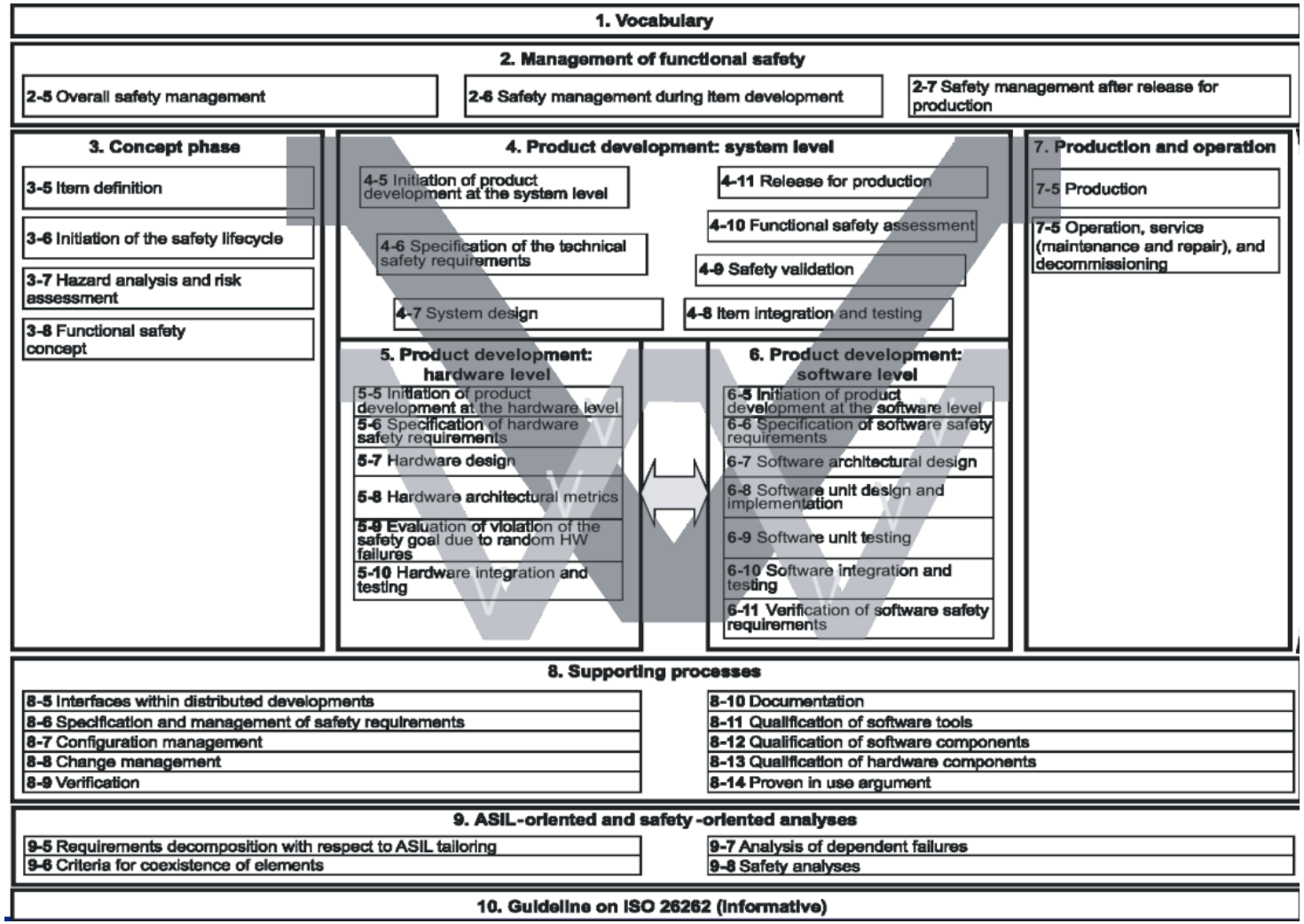
Functional safety (ISO 26262)

- Functional Safety: “absence of unreasonable risk due to hazards caused by malfunctioning behavior of E/E systems”

- ISO 26262: Road vehicles – functional safety
  - International standard
  - Scope: functional safety of E/E components in series production vehicles up to a mass of 3.5tons
  - Specialization of the IEC 61508 for the automotive sector: Functional Safety of Electrical/Electronic/Programmable Electronic Safety Related Systems

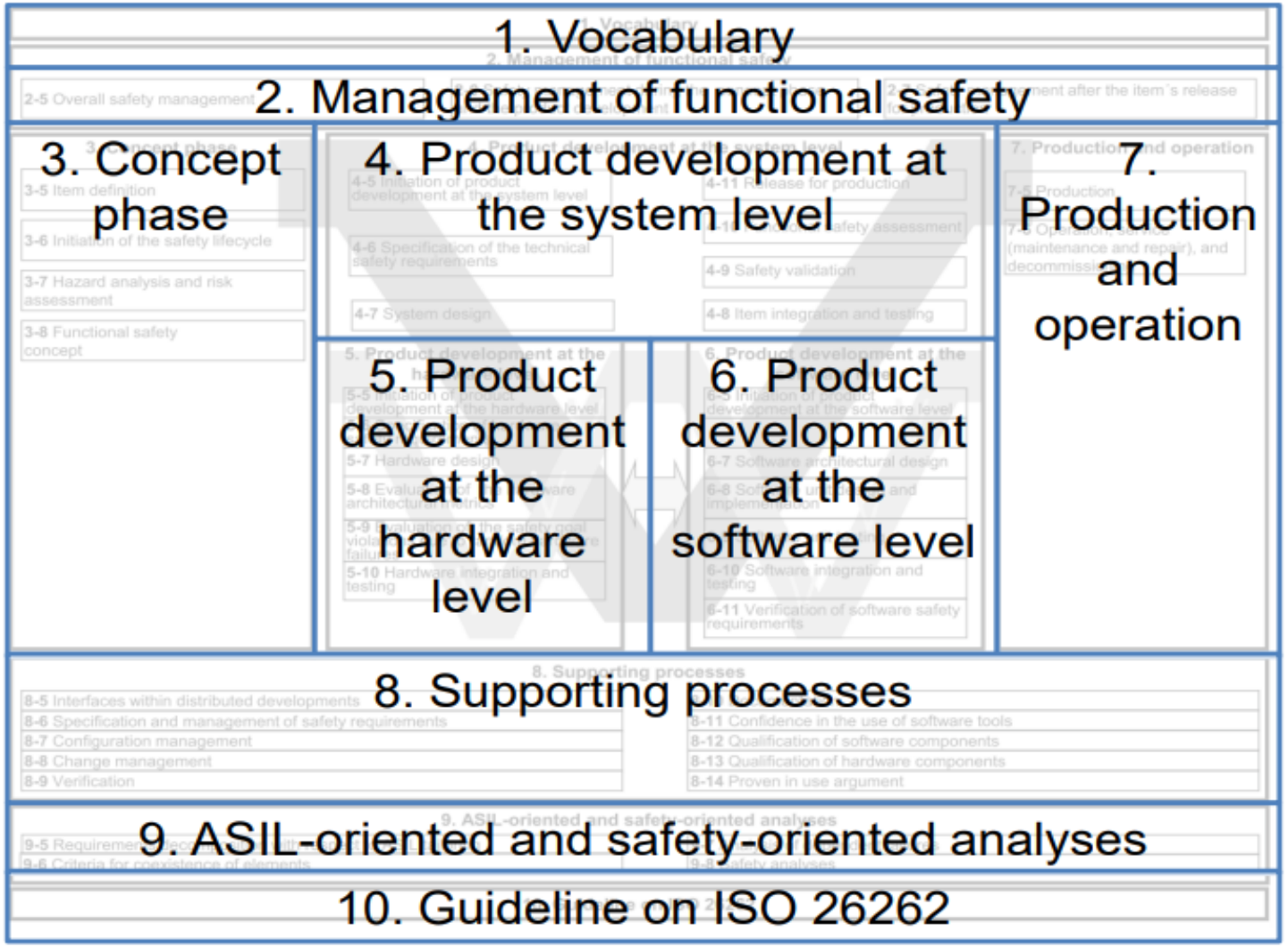
Relevant for all advanced driver assistance systems (ADAS)

ISO26262



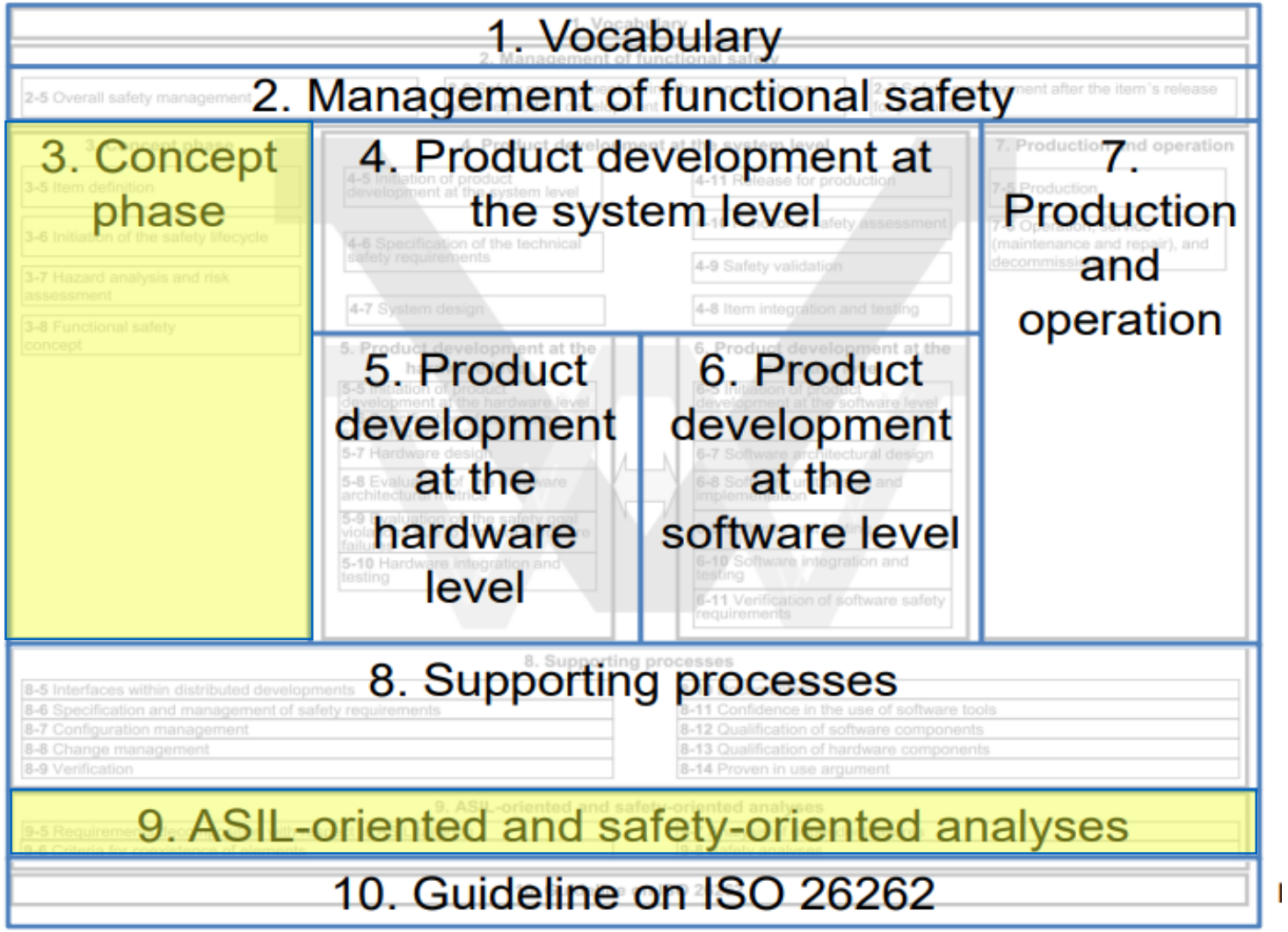
Source: ISO26262

ISO26262



Source: Lecture Fahrerassistenzsysteme, Prof. Lienkamp, WS17, TUM

ISO26262

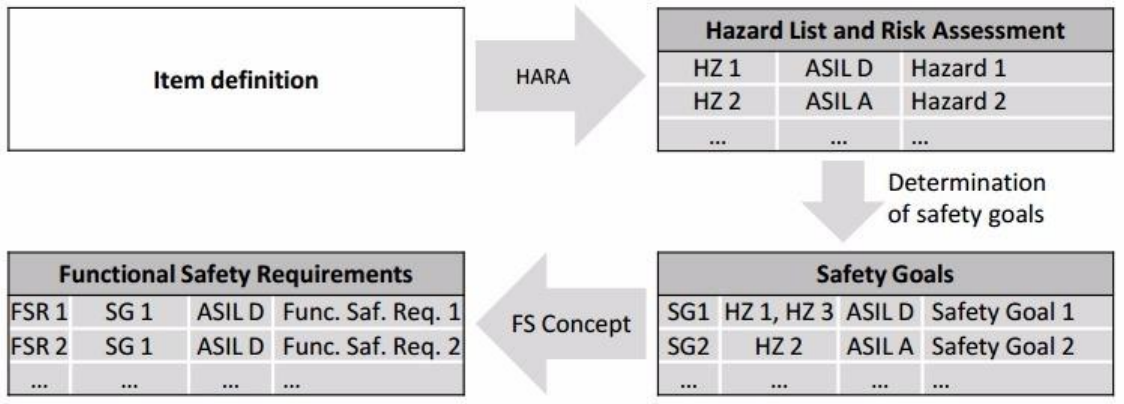


Source: Lecture Fahrerassistenzsysteme, Prof. Lienkamp, WS17, TUM

ISO26262 - HARA

Hazard and Risk Analysis with ISO26262:

1. Item Definition
2. Situation analysis
3. Hazard identification
4. Classification of hazardous events
5. Determination of ASIL
6. Determination of safety goals



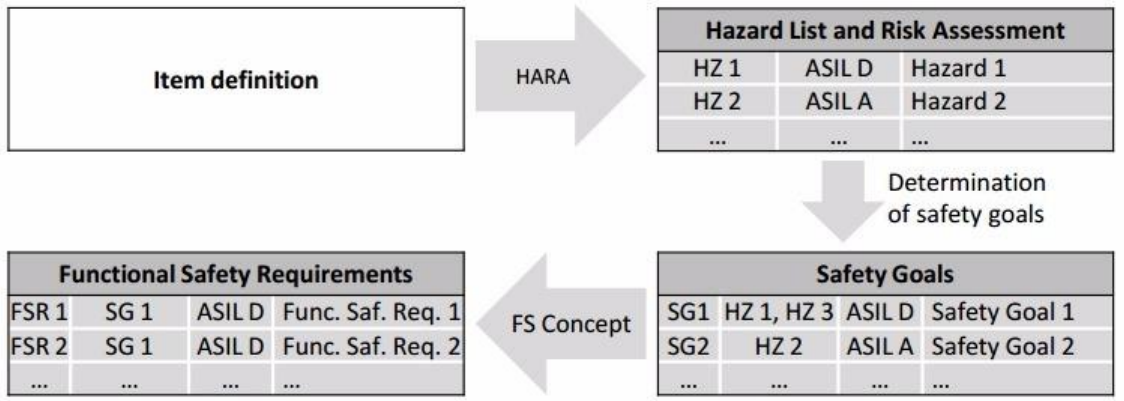
Source: Henriksson et al., 2018

[ISO26262, 2011] [Henriksson et al., 2018]

ISO26262 - HARA

Hazard and Risk Analysis with ISO26262:

1. Item Definition
2. Situation analysis
3. Hazard identification
4. Classification of hazardous events
5. Determination of ASIL
6. Determination of safety goals



Source: Henriksson et al., 2018

[ISO26262, 2011] [Henriksson et al., 2018]

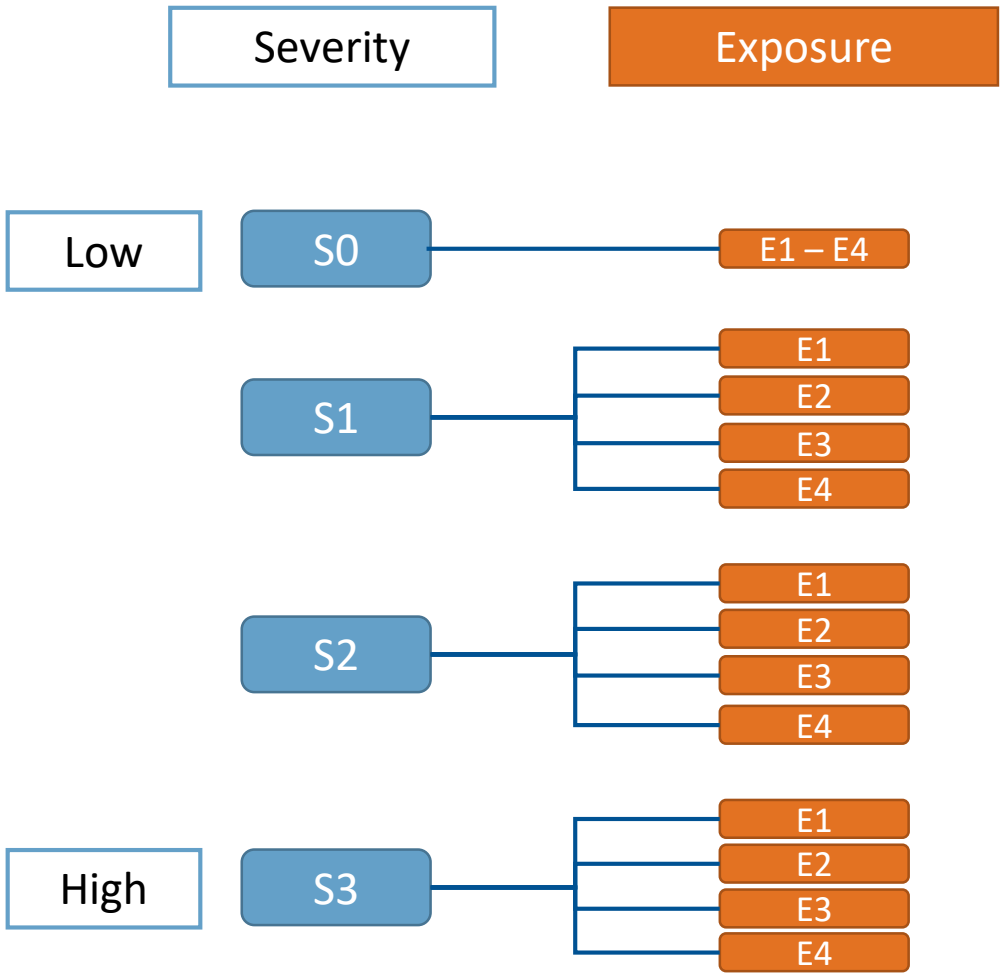
ASIL Level

Severity

Low	S0	S0 No injuries
	S1	S1 Light to moderate injuries
	S2	S2 Severe to life-threatening (survival probable) injuries
High	S3	S3 Life threatening (survival uncertain) to fatal injuries

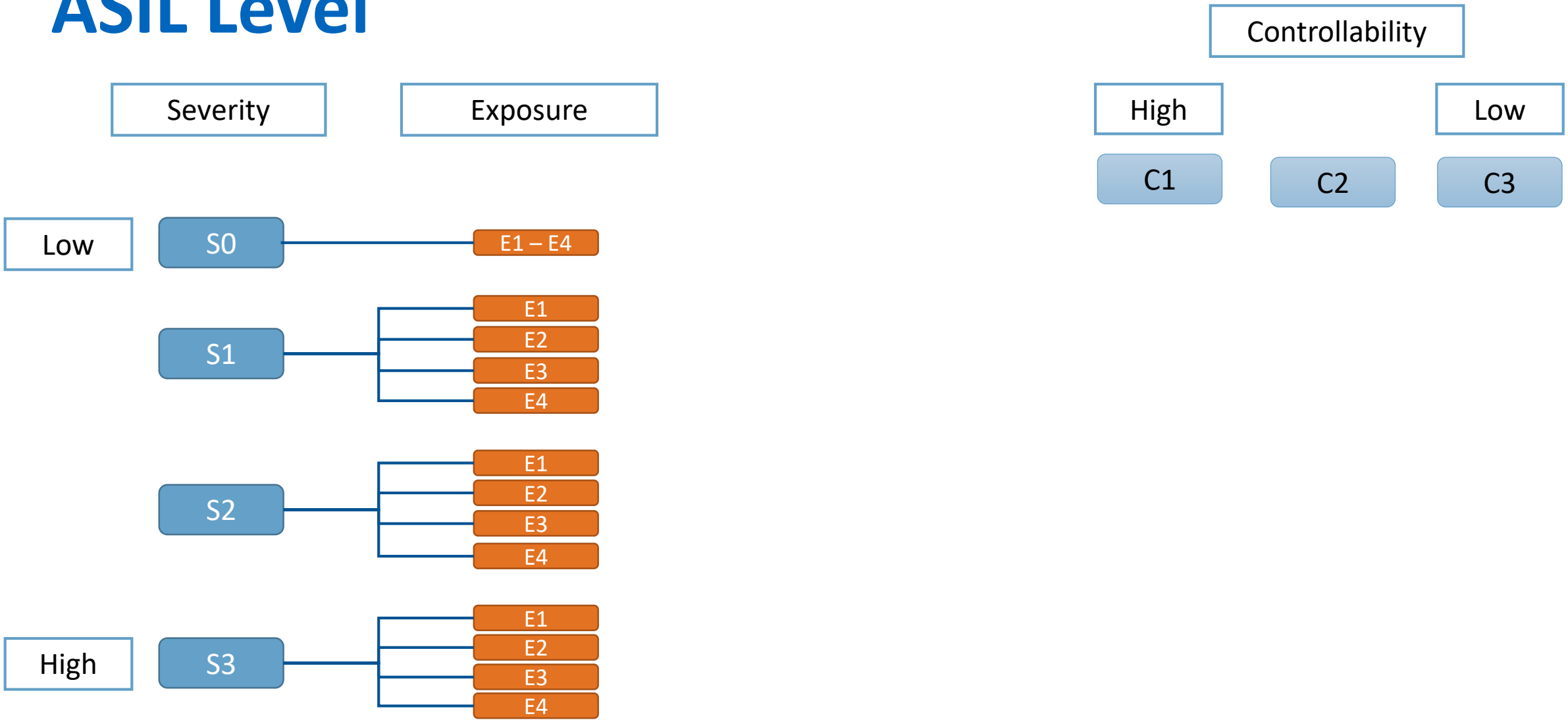
[ISO26262, 2011]

ASIL Level



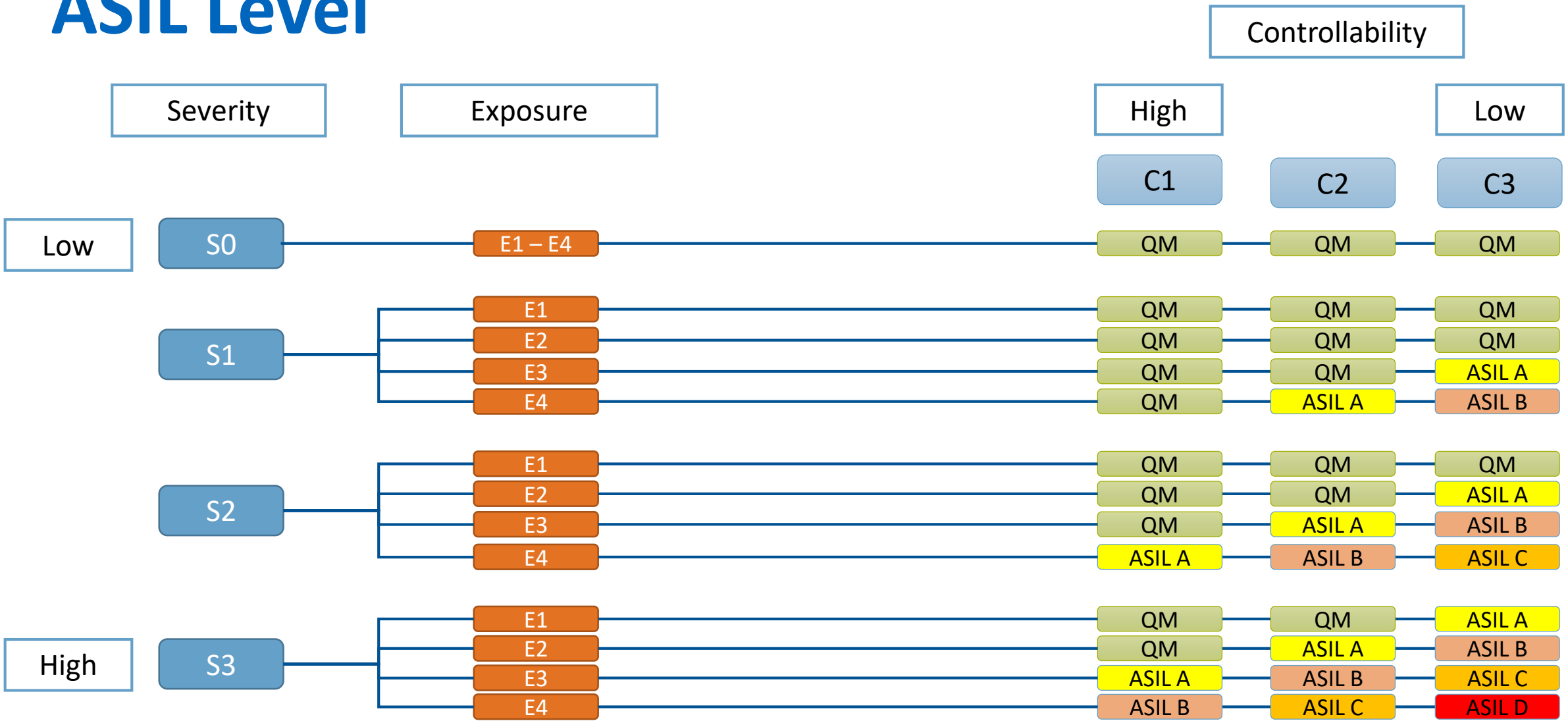
[ISO26262, 2011]

ASIL Level



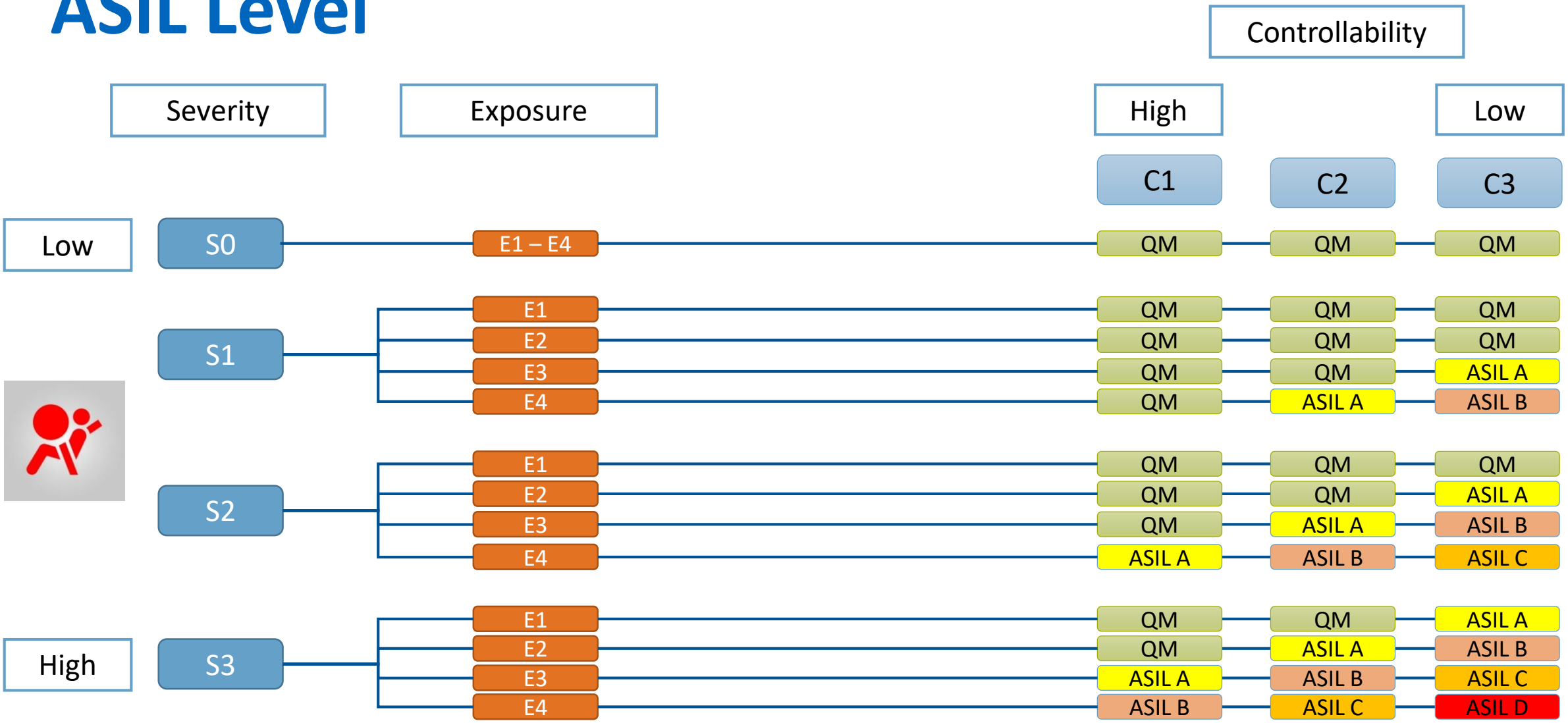
[ISO26262, 2011]

ASIL Level



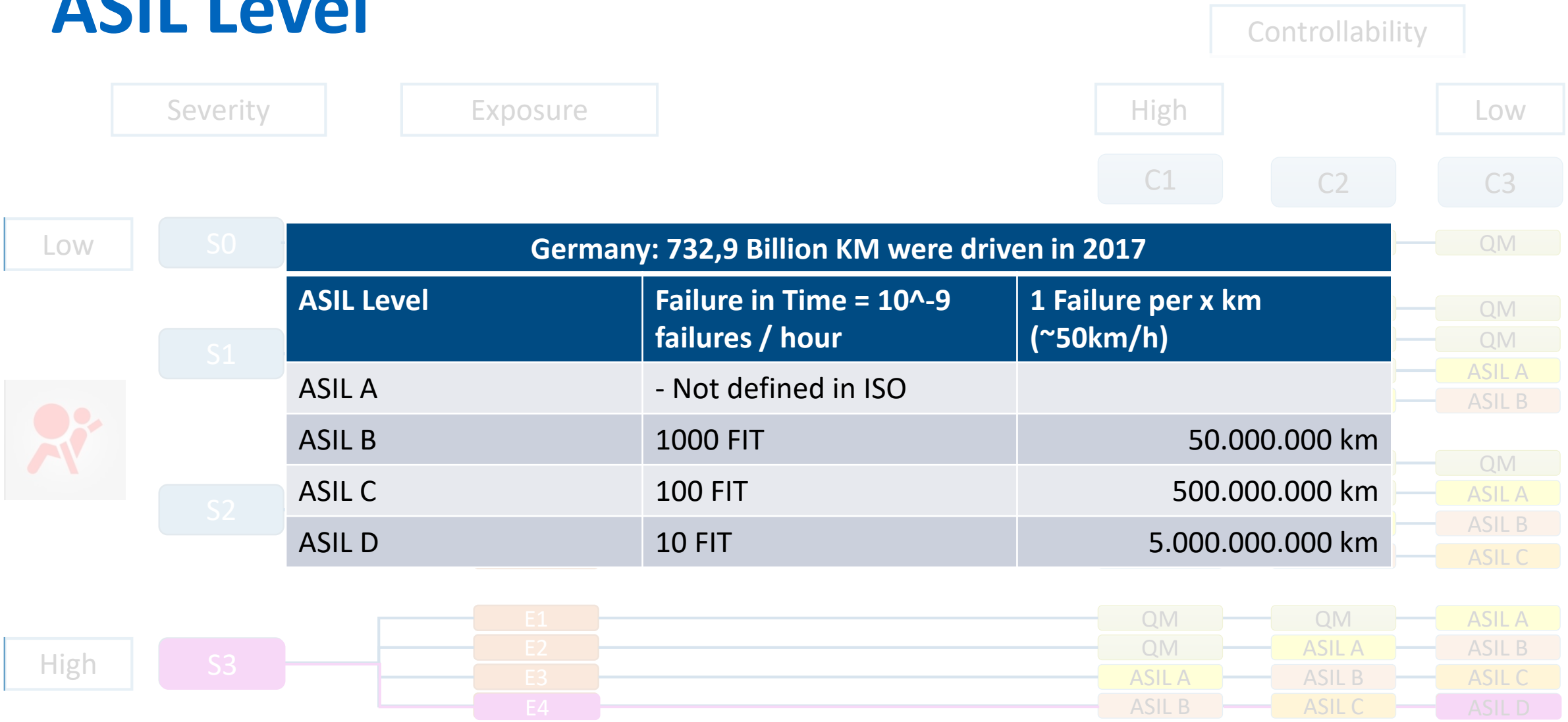
[ISO26262, 2011]

ASIL Level



[ISO26262, 2011]

ASIL Level



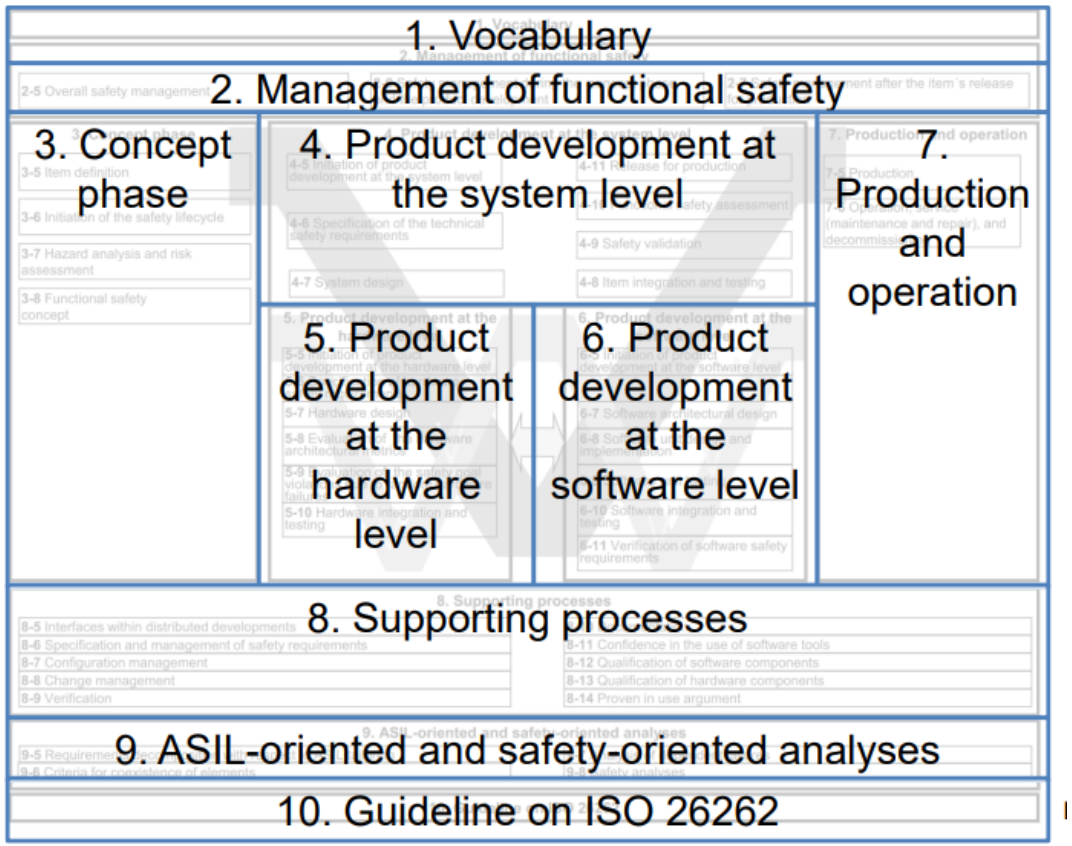
Germany: 732,9 Billion KM were driven in 2017		
ASIL Level	**Failure in Time = 10^-9 failures / hour**	**1 Failure per x km (~50km/h)**
ASIL A	- Not defined in ISO	
ASIL B	1000 FIT	50.000.000 km
ASIL C	100 FIT	500.000.000 km
ASIL D	10 FIT	5.000.000.000 km

https://www.kba.de/DE/Statistik/Kraftverkehr/VerkehrKilometer/verkehr_in_kilometern_node.html

[ISO26262, 2011]

Changes to 2nd edition

1st Edition (2011)



Source: Lecture Fahrerassistenzsysteme, Prof. Lienkamp, WS17, TUM

2nd Edition (2018)

- Removal of 3.5t weight limit
- New part 11 on semi conductors
- Guidelines on the application of ISO 26262 to semiconductors
- New qualification methods for ASIL
- Updates to the PMHF (Probabilistic Metric for Hardware Failure) equation and verification of safety analysis
- → No changes regarding ML

[Schloeffel and Dailey,]

Agenda

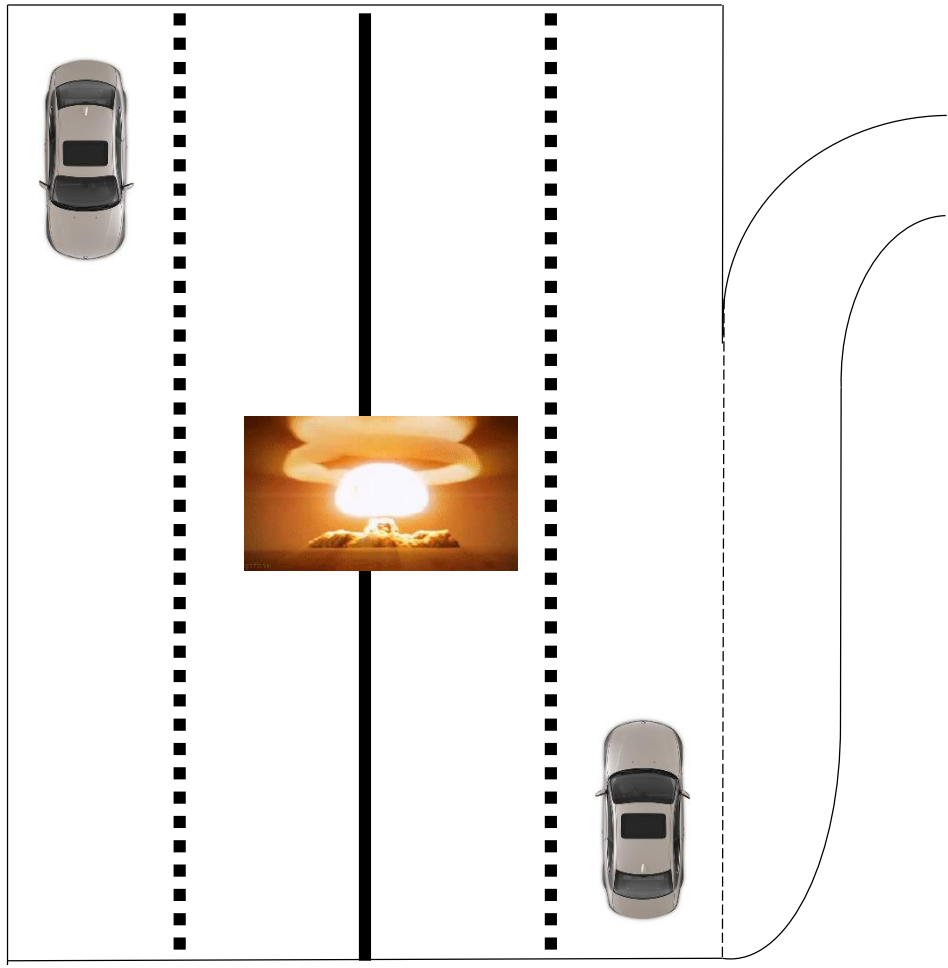
1. Introduction
2. How to combine Machine Learning and ISO 26262?
   1. What is safe in regards to AD?
   2. Possible Applications of ISO26262 to ML
   3. Conflicts between ISO26262 and ML
3. How to move on from here?
4. Conclusion and Outlook

What is safe in regards to AD?

- What is “safe enough” for AD ?
  - No accidents at all (“Absence of unreasonable risk…”)
  - (significant) higher safety than today (lower accident rate)

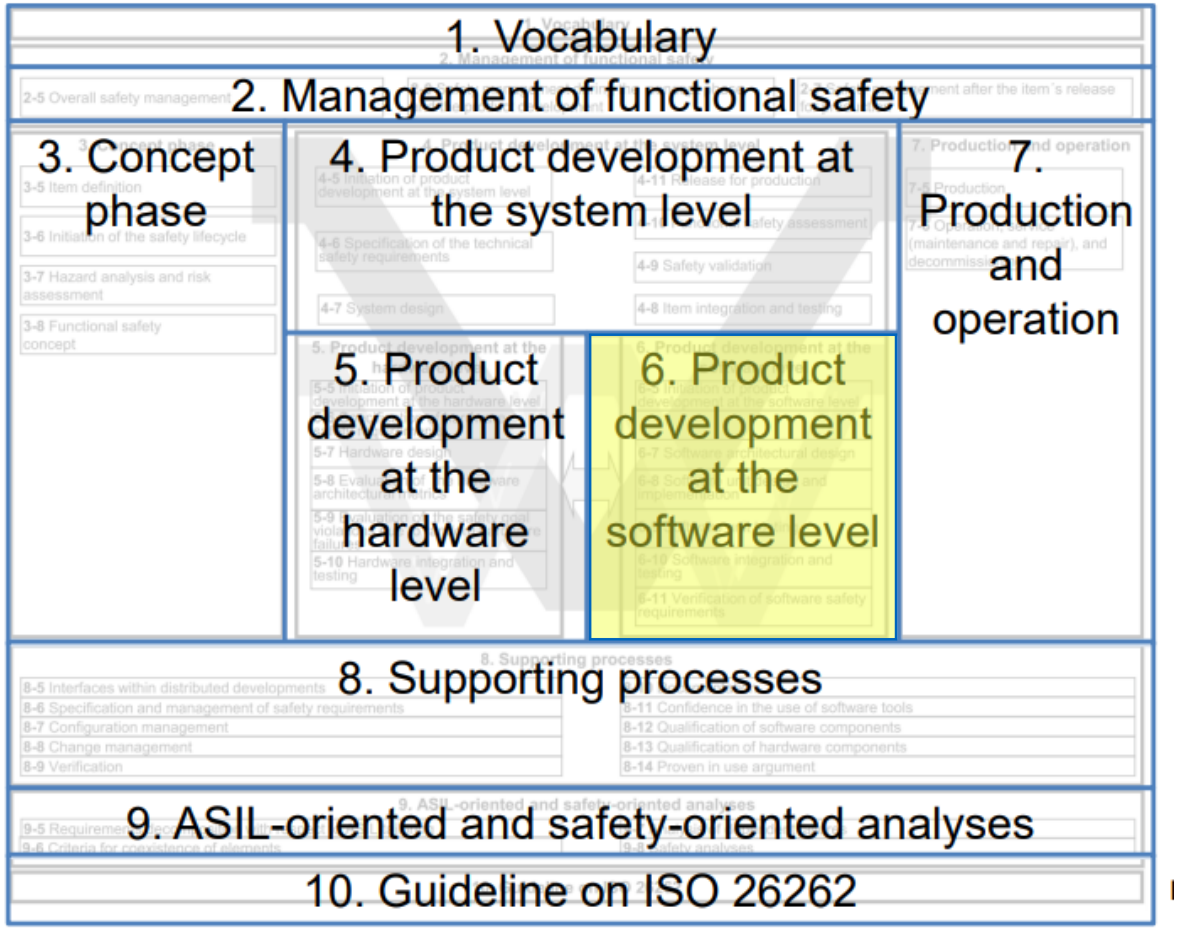
- 5 billion test miles necessary for real-world verification
  - How do we define test cases (border cases)?
  - What is correct behaviour in tests ?

[Bates, 2017] [Dailey, 2018]



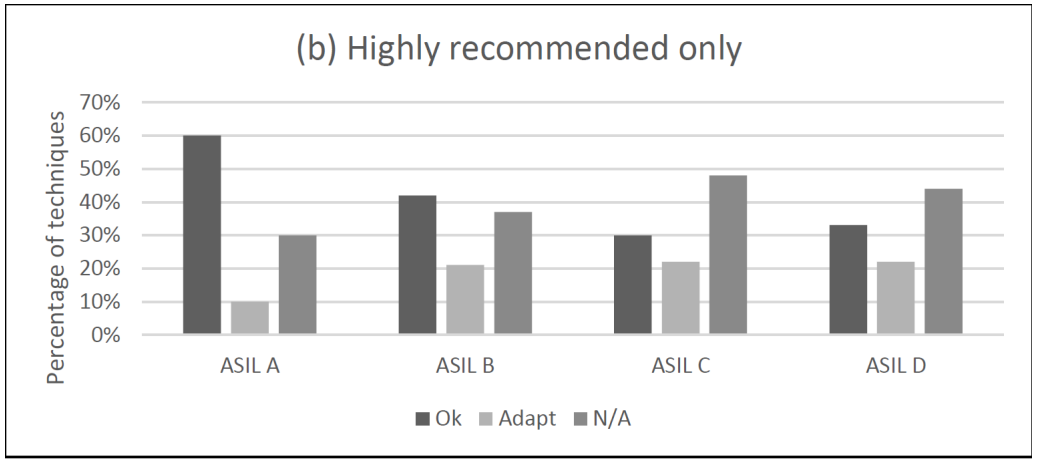
Source: https://giphy.com/gifs/XDLJpjzyw76Sc / EVOX Images

Possible Applications of ISO26262 to ML



Source: Lecture Fahrerassistenzsysteme, Prof. Lienkamp, WS17, TUM

- 34 of 75 ISO SW-Develpoment techniques at unit level
- black box techniques are “ok”
- “walk through” can be “adapted”
- Code oriented / white box techniques can not be used

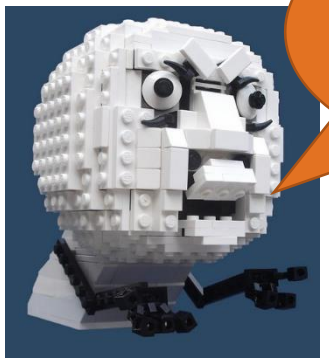


[Salay et al., 2017]

What about incomplete training sets?

Which SW techniques to use?

What about new types of hazards?



Source: https://memegenerator.net/instance/74873992/salt-bae-let-me-sprinkle-some-deep-learning-magic, http://knowyourmeme.com/memes/all-the-things, https://c1.staticflickr.com/1/264/19718083479_d6589c5404_b.jpg, http://www.relatably.com/m/img/meme-center-american-dad/american-dad_o_721757.jpg

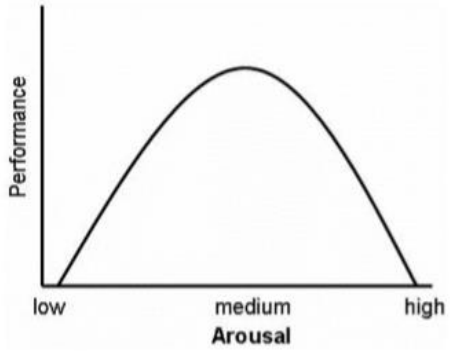
Conflicts

- Identifying hazards
- Fault and failure modes
- The use of training sets
- The level of ML usage
- Required software techniques

[Salay et al., 2017]

Conflicts

- Identifying hazards
  - ML can create new types of hazards
  - Complex behavioral interactions between humans and ML
    - Overestimation of the ML performance
    - Human has not the optimal level of employment
    - Reduced human skill level and situation awareness
  - RL plays with the reward function or has unintended behavior through wrong reward function
- Fault and failure modes
- The use of training sets
- The level of ML usage
- Required software techniques



Yerkes-Dodson-Law (1908):

Source: Lecture Fahrerassistenzsysteme, Prof. Lienkamp, WS17, TUM

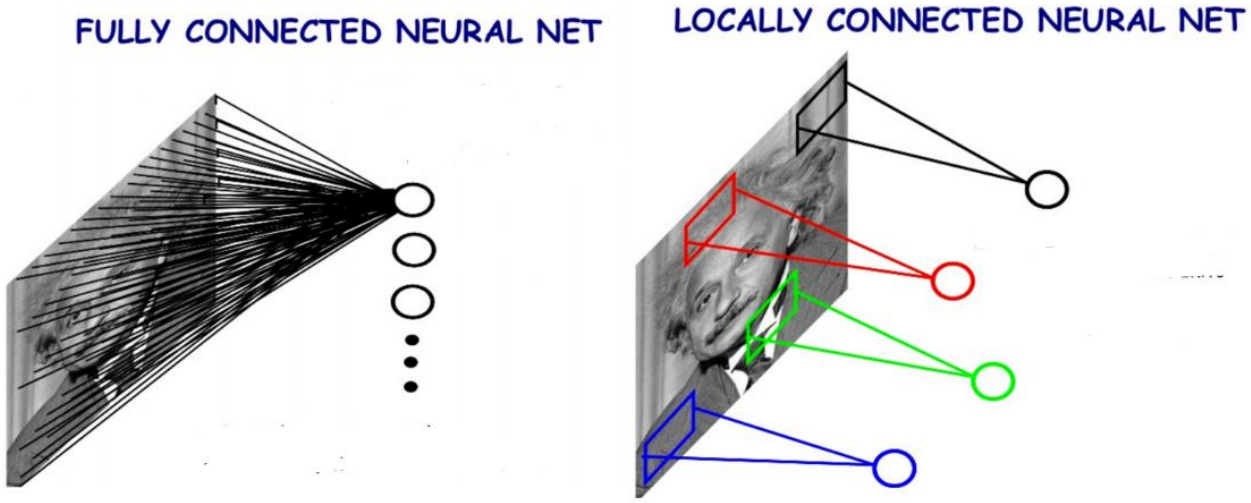
[Salay et al., 2017]

RL plays with the reward function

Source: https://blog.openai.com/faulty-reward-functions/

Conflicts

- Identifying hazards
- Fault and failure modes
  - Specific fault types and failure modes for ML
    - Network topology
    - Learning algorithm
    - Training set
- The use of training sets
- The level of ML usage
- Required software techniques

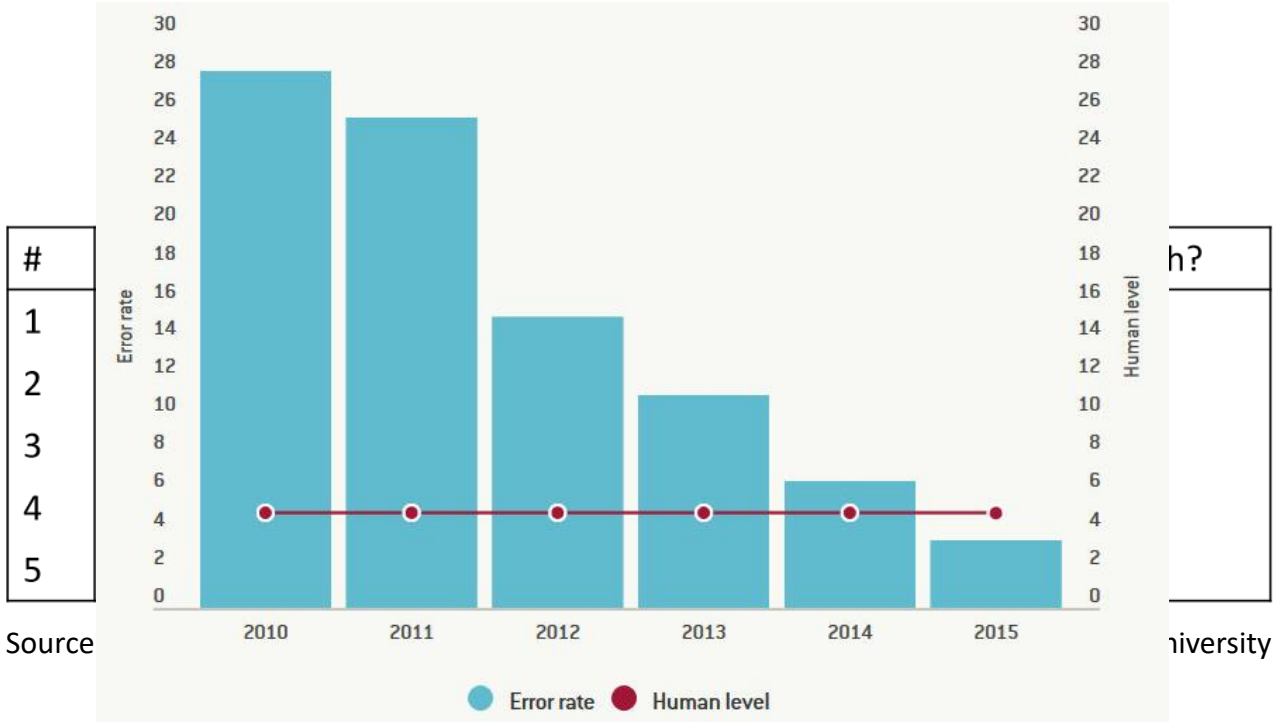


Source: https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/convolutional_neural_networks.html

[Salay et al., 2017]

Conflicts

- Identifying hazards
- Fault and failure modes
- The use of training sets
  - Inherently incomplete data sets
  - Correct by construction with respect to the training set
  - Unspecifiable functionality (like perception)
- The level of ML usage
- Required software techniques

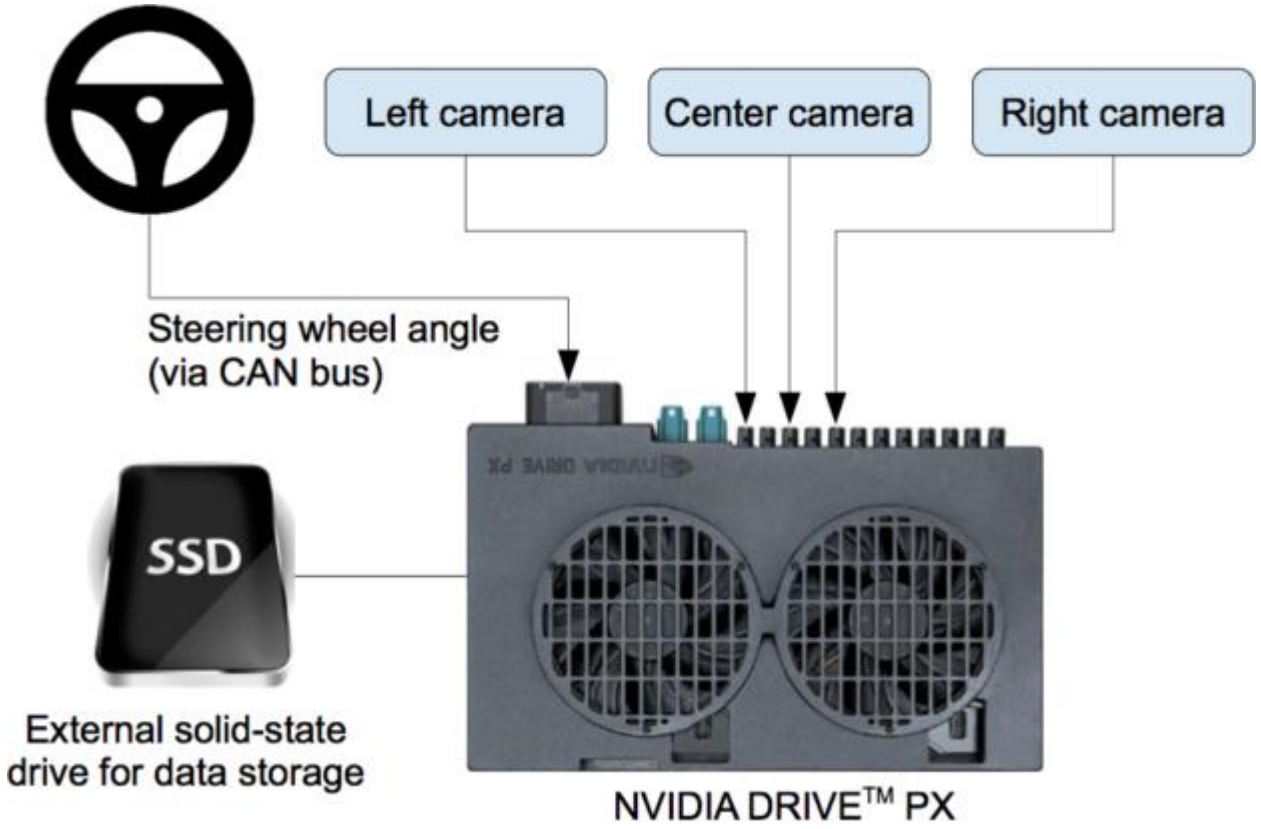


Imagenet Error Rates, Source: https://c1.staticflickr.com/5/4162/33621365014_fe35be452a_b.jpg

[Salay et al., 2017]

Conflicts

- Identifying hazards
- Fault and failure modes
- The use of training sets
- The level of ML usage
  - End-to-end learning is critical
  - Lack of transparency of ML components
- Required software techniques

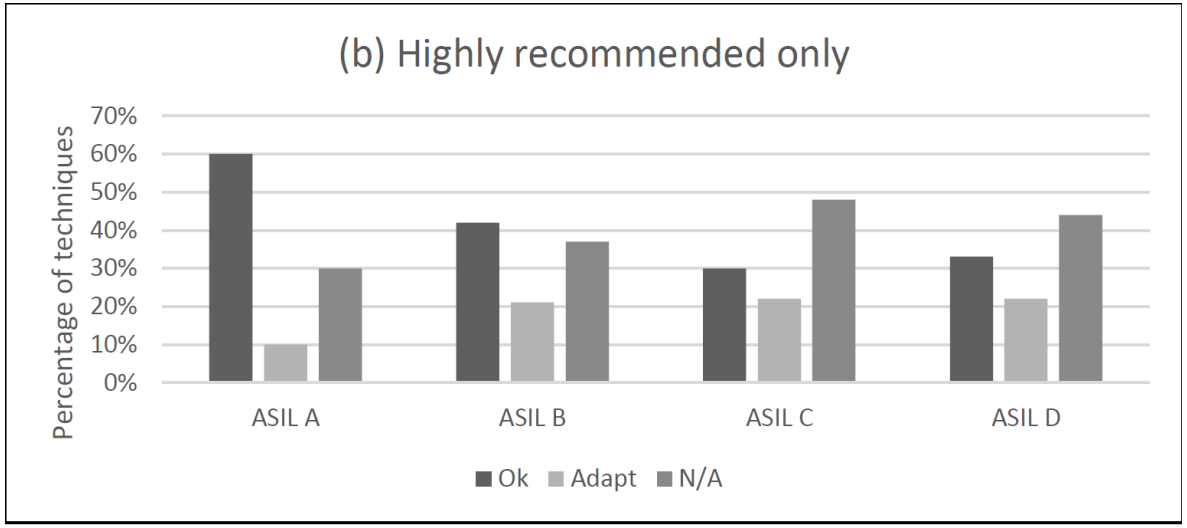


Source: https://devblogs.nvidia.com/deep-learning-self-driving-cars/

[Salay et al., 2017]

Conflicts

- Identifying hazards
- Fault and failure modes
- The use of training sets
- The level of ML usage
- Required software techniques
  - Assumption that code is implemented using an imperative programming language
  - Difficult to use with ML, but also with Functional or logic programming, etc.



Source: Salay et al., 2017

[Salay et al., 2017]

Agenda

1. Introduction
2. How to combine Machine Learning and ISO 26262?
3. How to move on from here?
   1. Recommendations for applying ISO26262 to ML
   2. Usage of a new standard
      1. Pegasus
      2. SOTIF
   3. Radical new approach
4. Conclusion and Outlook

Recommendations for applying ISO26262 to ML

Changes to ISO26262

- Consider ML specific hazards

How to use ML

[Varshney, 2016] [Salay et al., 2017]

Recommendations for applying ISO26262 to ML

Changes to ISO26262

- Consider ML specific hazards
- ML Lifecycle techniques

How to use ML

[Varshney, 2016] [Salay et al., 2017]

Recommendations for applying ISO26262 to ML

Changes to ISO26262

- Consider ML specific hazards
- ML Lifecycle techniques
- Partial specifiable functionality

How to use ML

[Varshney, 2016] [Salay et al., 2017]

Recommendations for applying ISO26262 to ML

Changes to ISO26262

- Consider ML specific hazards
- ML Lifecycle techniques
- Partial specifiable functionality
- Fault tolerance strategies for software

How to use ML

[Varshney, 2016] [Salay et al., 2017]

Recommendations for applying ISO26262 to ML

Changes to ISO26262

- Consider ML specific hazards
- ML Lifecycle techniques
- Partial specifiable functionality
- Fault tolerance strategies for software
- Intent based SW requirements

How to use ML

[Varshney, 2016] [Salay et al., 2017]

Recommendations for applying ISO26262 to ML

Changes to ISO26262

- Consider ML specific hazards
- ML Lifecycle techniques
- Partial specifiable functionality
- Fault tolerance strategies for software
- Intent based SW requirements

How to use ML

- Modular ML usage

[Varshney, 2016] [Salay et al., 2017]

Recommendations for applying ISO26262 to ML

Changes to ISO26262

- Consider ML specific hazards
- ML Lifecycle techniques
- Partial specifiable functionality
- Fault tolerance strategies for software
- Intent based SW requirements

How to use ML

- Modular ML usage
- Human interpretable models

[Varshney, 2016] [Salay et al., 2017]

Recommendations for applying ISO26262 to ML

Changes to ISO26262

- Consider ML specific hazards
- ML Lifecycle techniques
- Partial specifiable functionality
- Fault tolerance strategies for software
- Intent based SW requirements

How to use ML

- Modular ML usage
- Human interpretable models
- Safety Reserves

[Varshney, 2016] [Salay et al., 2017]

Recommendations for applying ISO26262 to ML

Changes to ISO26262

- Consider ML specific hazards
- ML Lifecycle techniques
- Partial specifiable functionality
- Fault tolerance strategies for software
- Intent based SW requirements

How to use ML

- Modular ML usage
- Human interpretable models
- Safety Reserves
- Reject option

[Varshney, 2016] [Salay et al., 2017]

Recommendations for applying ISO26262 to ML

Changes to ISO26262

- Consider ML specific hazards
- ML Lifecycle techniques
- Partial specifiable functionality
- Fault tolerance strategies for software
- Intent based SW requirements

How to use ML

- Modular ML usage
- Human interpretable models
- Safety Reserves
- Reject option
- Open Source / Data

https://www.123rf.com/stock-photo/hands_reaching_out.html [Varshney, 2016] [Salay et al., 2017]

Usage of an new standard

- ⊕ No constraints of old standard with old software paradigms
- ⊕ Everything could be rethought and fitted to ML
- ⊕ E.g. open source norm (maintained by a group of e.g. Automotive developers)

- ⊖ Using a new standard correlates with retraining the companies employees
- ⊖ Difficult to make it a standard

Pegasus

SP 1	SP 2	SP 3	SP 4
Scenario analysis and quality measures	**Implementation processes**	**Testing**	**Result reflection and embedding**
Application scenario Quality measures Extended scenario	Process methodology Process specification	Test case database Laboratory and simulation tests Testing site tests Field tests	Proof of concept Embedding
DISCOVER MORE INFORMATION	DISCOVER MORE INFORMATION	DISCOVER MORE INFORMATION	DISCOVER MORE INFORMATION

Source: https://www.pegasusprojekt.de/en/

Fundet by the german goverment (BMWi) for developing automated driving standards

SOTIF (Safety Of The Intended Functionality) – not yet released

- Will provide guidelines for Level-0, Level-1, and Level-2 autonomous drive (AD) vehicles.
- Focus on the functional safety of ADAS-related hazards caused by “normal operation” of the sensors.
- Questions to be answered by SOTIF:
  - Details on advanced concepts of the AD architecture
  - How to evaluate SOTIF hazards that are different from ISO 26262 hazards?
  - How to identify and evaluate scenarios and trigger events?
  - How to reduce SOTIF related risks?
  - How to verify and validate SOTIF related risks?
  - The criteria to meet before releasing an autonomous vehicle.

Radical new approach

"Wenn ein Problem von den Testfahrern festgestellt wurde, wird das anschließend behoben und das Ganze geht von vorne los. Aber wenn Sie etwas behoben haben, müssen Sie sehen, dass an einer anderen Stelle die Funktion genau funktioniert wie vorher. Das ist schon ein bisschen tricky, aber das ist eben die Pionierleistung", sagt Bereczki (Regulierungsexperte bei Audi).

Diese Aufgabe müsse jeder Hersteller für sich lösen, da es festgelegte Testszenarien dafür noch nicht gebe.

https://www.golem.de/news/automatisiertes-fahren-der-schwierige-weg-in-den-selbstfahrenden-stau-1807-135557-2.html

Translated:

'If a problem has been identified by the test driver, it will be fixed and the whole thing will start again. But if you have fixed something, you have to see that in another place the function works exactly as before. That's a bit tricky, but that's just the pioneering work' says Bereczki. (Regulationexpert at Audi).

Every manufacturer must solve this task for themselves, since there are not yet defined test scenarios for this.

Radical new approach

- Approach similar to Aerospace Industry
- Every failure has to be reported.
- Faults have to be investigated
- Resulting new test cases will be added
- Continuous improved test case database

→ Continuous Improvement and cooperation of Automotive industry

Source: http://www.newspakistan.pk/wp-content/uploads/2014/07/Federal-Aviation-Administration.jpg

[Bates, 2017]

Agenda

1. Introduction
2. How to combine Machine Learning and ISO 26262?
3. How to move on from here ?
4. Conclusion and Outlook

Conclusion

- ISO26262
  - ISO26262 and functional safety fit partially and with constraints to ML techniques
  - Changes on the standard are required (e.g. separation between fully and partially specified tasks)
  - The allowed usage of ML techniques will be restricted
- Other Standards
  - SOTIF: guidance for assuring that an autonomous vehicle functions and acts safely during normal operation
  - Pegasus: Close key gaps in the field of testing of highly-automated driving functions will be concluded by the middle of 2019.

- In general, ML is only applicable under constraints in safety critical (no end-to-end, safe guards,..)
  → Applicability of Reinforcement Learning even worse (more parameter like the correct reward function have to be estimated)

Outlook - Quotes from industry

- *December 2016: "We're going to end up with complete autonomy, and I think we will have complete autonomy in approximately two years." Elon Musk*

  https://electrek.co/2015/12/21/tesla-ceo-elon-musk-drops-prediction-full-autonomous-driving-from-3-years-to-2/

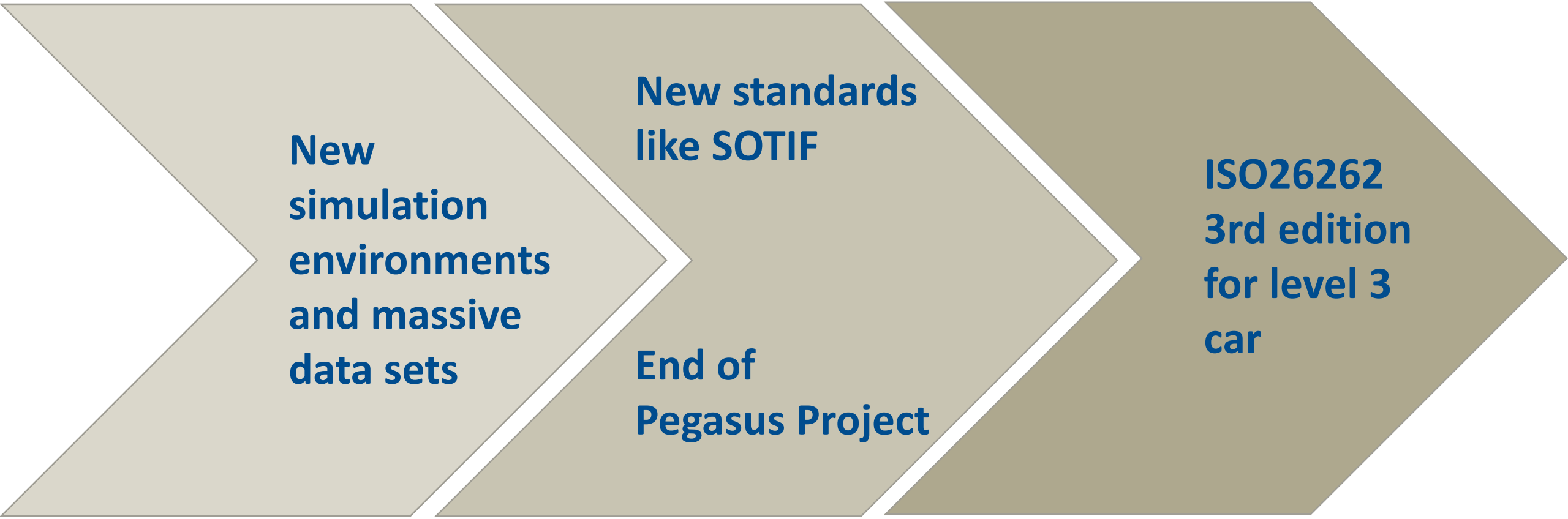
- March 2017: "*We are on the way to deliver a car in 2021 with level 3, 4 and 5*" BMW senior vice president for Autonomous Driving – Frickenstein

  Source: https://www.reuters.com/article/us-bmw-autonomous-self-driving/bmw-says-self-driving-car-to-be-level-5-capable-by-2021-i<dUSKBN16N1Y2

- March 2018: "*Level 3 will be achieved by 2021, but Level 4 is one of the biggest challenges facing the auto industry*", Elmar Frickenstein. (translated)

  Source: http://www.sueddeutsche.de/auto/autonomes-fahren-den-deutschen-herstellern-droht-ein-fehlstart-in-die-zukunft-1.3914126

Outlook – own thoughts



Resources

[Bates, 2017] Bates, R. (2017). Is it possible to know how safe we are in a world of autonomous cars?

[Dailey, 2018] Dailey, J. (2018). Functional safety in ai-controlled vehicles: If not iso26262, then what?

[Henriksson et al., 2018] Henriksson, J., Borg, M., and Englund, C. (2018). Automotive safety and machine learning: Initial results from a study on how to adapt the iso 26262 safety standard. In SEFAIAS 2018.

[ISO26262, 2011] ISO26262 (2011). Road vehicles – Functional safety.

[Salay et al., 2017] Salay, R., Queiroz, R., and Czarnecki, K. (2017). An analysis of ISO 26262: Using machine learning safely in automotive software. CoRR, abs/1709.02435.

[Schloeffel and Dailey,] Schloeffel, J. and Dailey, J. Understanding iso26262 second edition: What's new and what's still missing.

[Varshney, 2016] Varshney, K. R. (2016). Engineering safety in machine learning. In 2016 Information Theory and Applications Workshop (ITA), pages 1–5.

Thank you for your attention