AI Safety — SE Perspective

Jin L.C. Guo, SOCS McGill University

Logistics

- Class on Nov 11th through Zoom
- Project M2 released
- Move Final Presentation date (Dec 2nd -> Dec 9th)

Agenda

- Case study - Autonomous Vehicle
- Safety Engineering
  - Hazard Analysis
  - Safety Assurance
- Adapting ISO262262 for Automotive Software with ML

WIRED

REMOTE CONTROL

Cruise Will Soon Hit San Francisco With No Hands on the Wheel

AARIAN MARSHALL

HIDDEN PICTURES

Split-Second 'Phantom' Images Can Fool Tesla's Autopilot

ANDY GREENBERG

RESPONSIBILITY

Why Wasn't Uber Charged in a Fatal Self-Driving Car...

AARIAN MARSHALL

AUTONOMY

Waymo's Self-Driving Jaguars Arrive With New,...

ALEX DAVIES

CARS

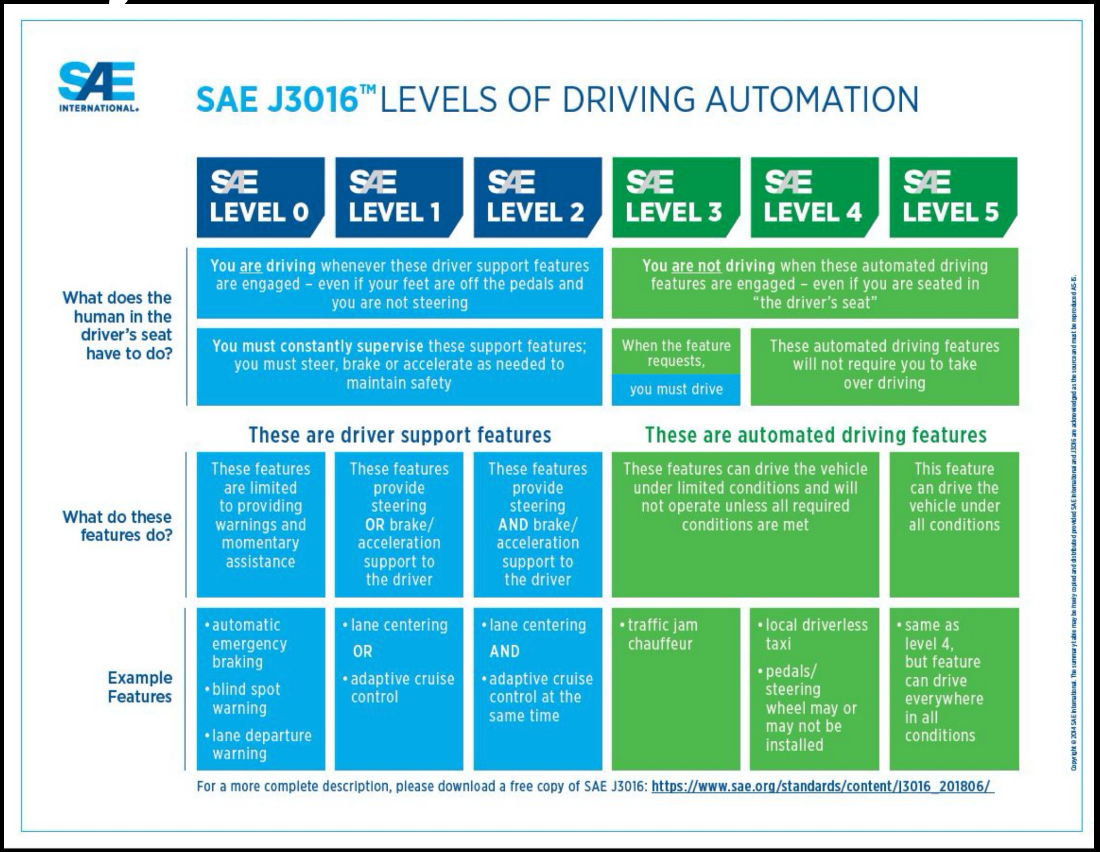
Lidar Is Finally Becoming a Real Business

TIMOTHY B. LEE, ARS TECHNICA

Case Study — Autonomous Vehicle

- Localization and Mapping – Where am I
- Scene Understanding – Where is Everyone Else?
- Movement Planning – How to get from Point A to Point B
- Driver State – What is the Driver Up to?
- Safety Monitoring

Case Study — Autonomous Vehicle



https://www.sae.org/news/2019/01/sae-updates-j3016-automated-driving-graphic

Challenges

What factors are involved
for decision making at an intersection?

aps | ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Safe Crossings: The Power of Eye Contact

October 14, 2015

TAGS: DRIVING | EYE MOVEMENTS | PERCEPTION | PERSONALITY/SOCIAL | SOCIAL BEHAVIOR | SOCIAL COGNITION

It can be a dangerous world for pedestrians. Studies on French roads report that nearly 60% of drivers do not stop at all for pedestrians crossing the street at designated crosswalks.

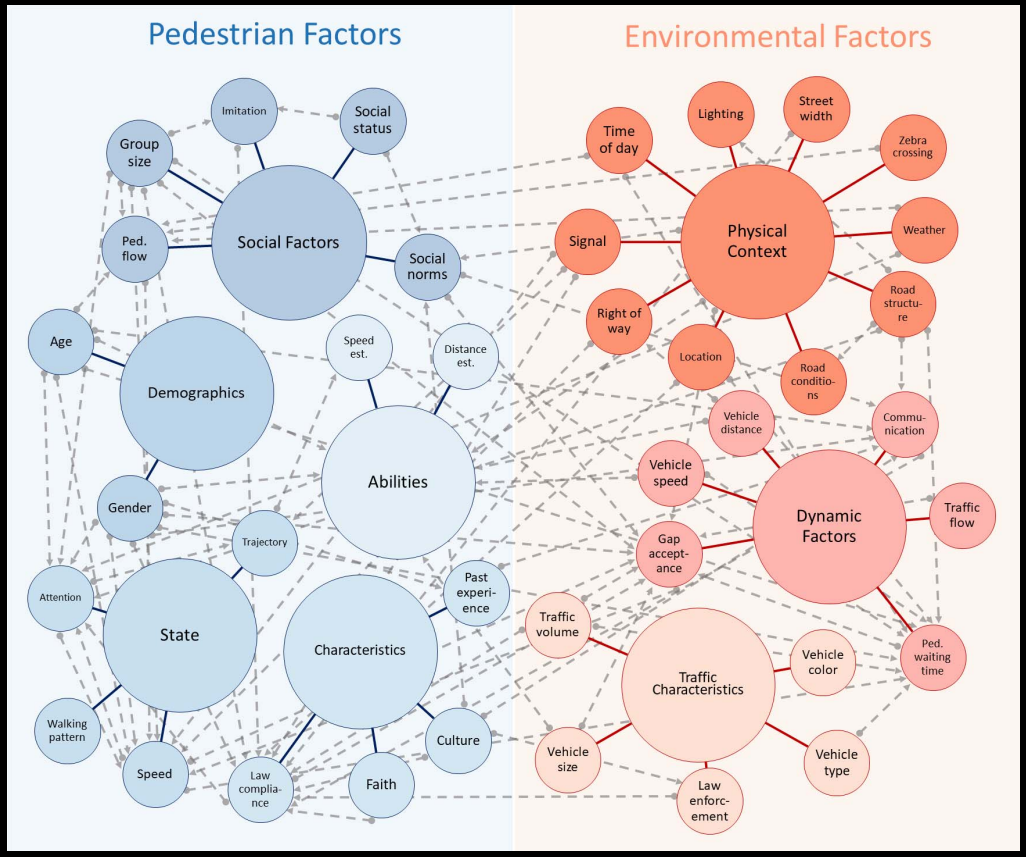
New research suggests that pedestrians may have a better shot at crossing safely if they make direct eye contact with oncoming drivers.

Decades of research have shown that eye contact has a powerful effect in social interactions. People are far more likely to comply with requests — for example, donating money — when the person making the request looks them in the eye. Experiments on eye contact completed in the 1970s demonstrated that drivers stopped more than twice as often when hitchhikers looked them directly in the eye.

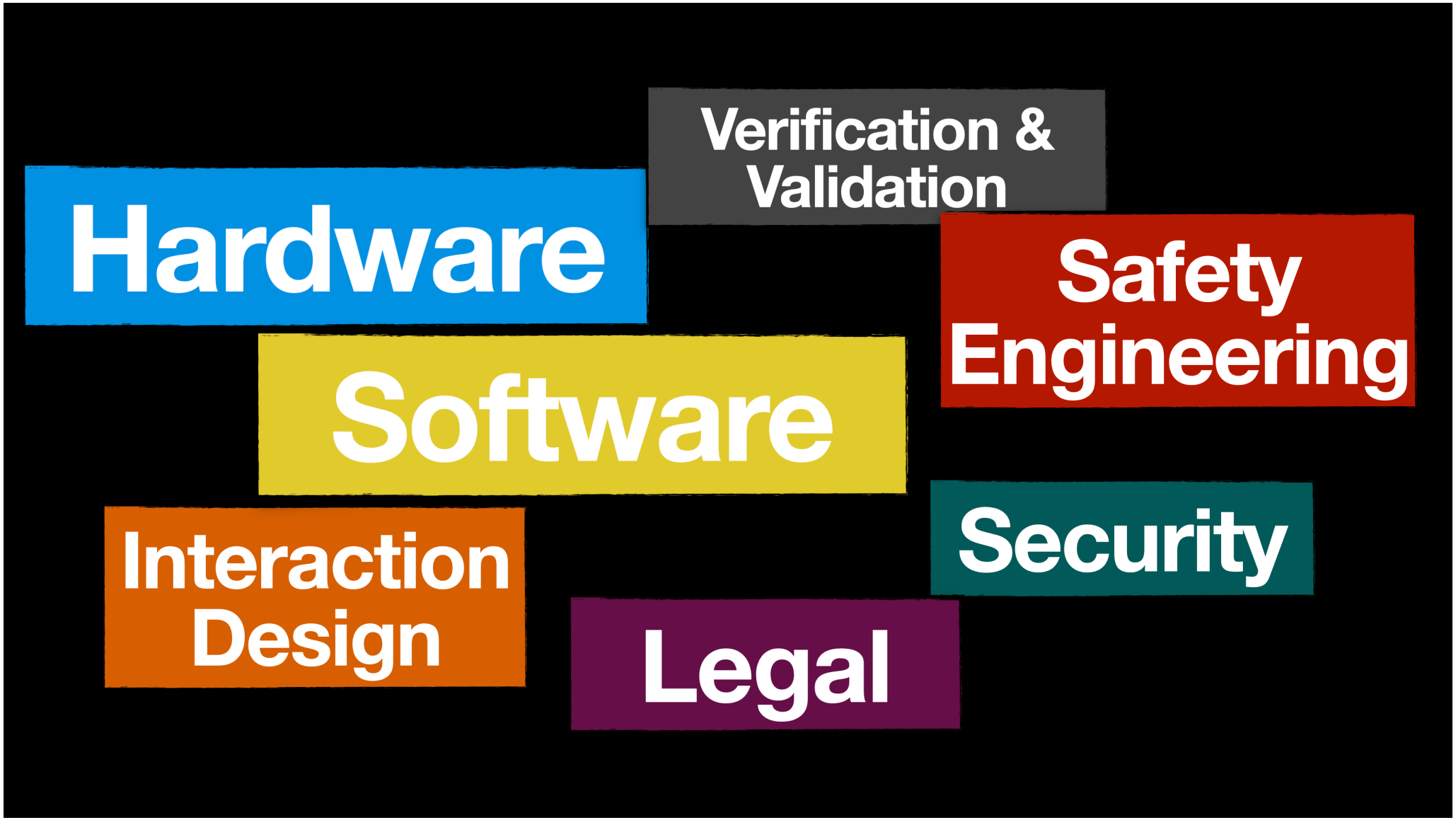
Challenges

Fig. 8. Driver's conditions used in the experiments conducted in [18].

Rasouli, Amir, and John K. Tsotsos. "Autonomous vehicles that interact with pedestrians: A survey of theory and practice." *IEEE Transactions on Intelligent Transportation Systems* 21, no. 3 (2019): 900-918.



Rasouli, Amir, and John K. Tsotsos. "Autonomous vehicles that interact with pedestrians: A survey of theory and practice." *IEEE Transactions on Intelligent Transportation Systems* 21, no. 3 (2019): 900-918.


Verification &
Validation
Hardware
Safety
Engineering
Software
Security
Interaction
Design
Legal

Agenda

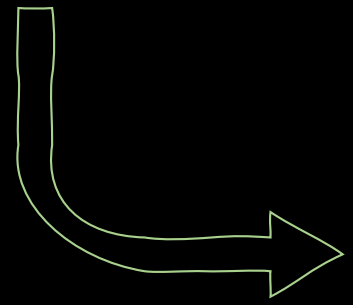
- Case study - Autonomous Vehicle
- Safety Engineering
  - Hazard Analysis
  - Safety Assurance
- Adapting ISO262262 for Automotive Software with ML

Safety Requirements

"The system shall not allow the simultaneous activation of more than three alarm signals."

"The navigation system shall not allow users to set the required destination when the car is moving."

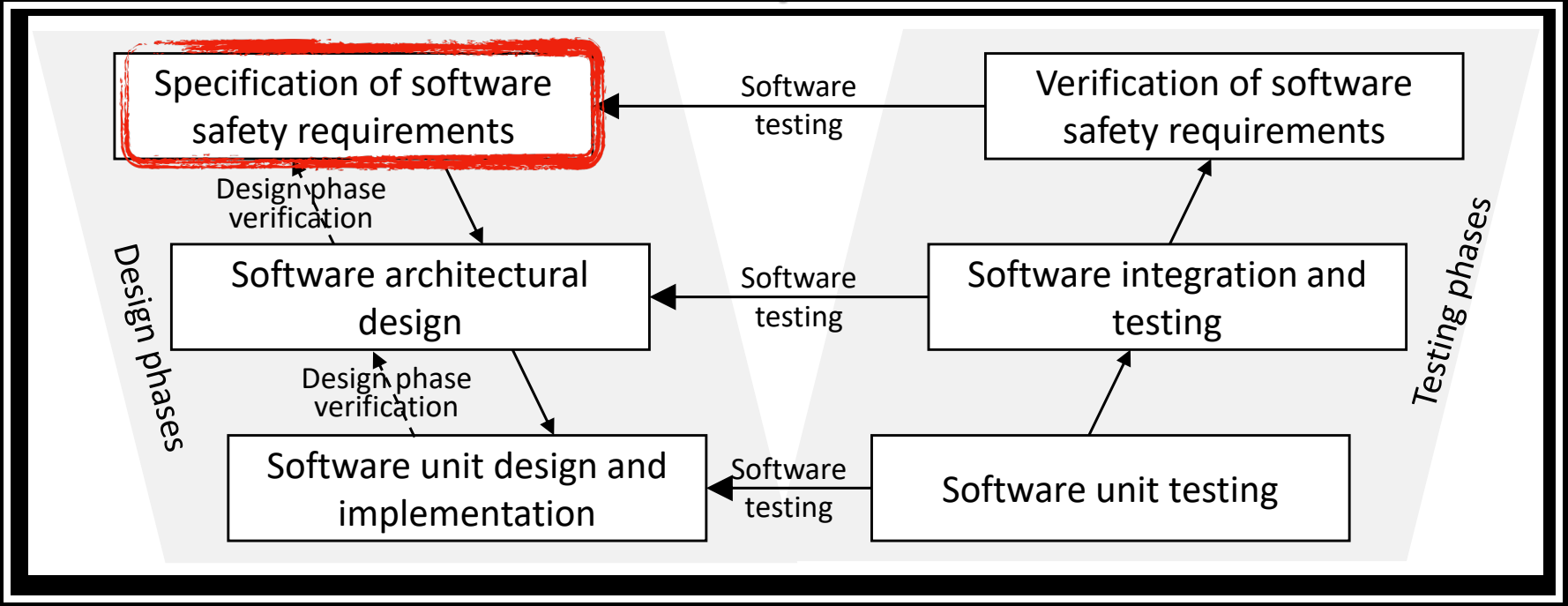
Domain knowledge, safety standards, and regulations.



Functional requirements

System design decision
e.g. hardware vs software, equipment choices

ISO26262 - Functional Safety Stanford for Road Vehicles



ISO 26262 part 6 - Product development at the software level.

Salay, Rick, Rodrigo Queiroz, and Krzysztof Czarnecki. "An analysis of ISO 26262: Using machine learning safely in automotive software." *arXiv preprint arXiv:1709.02435* (2017).

Hazard

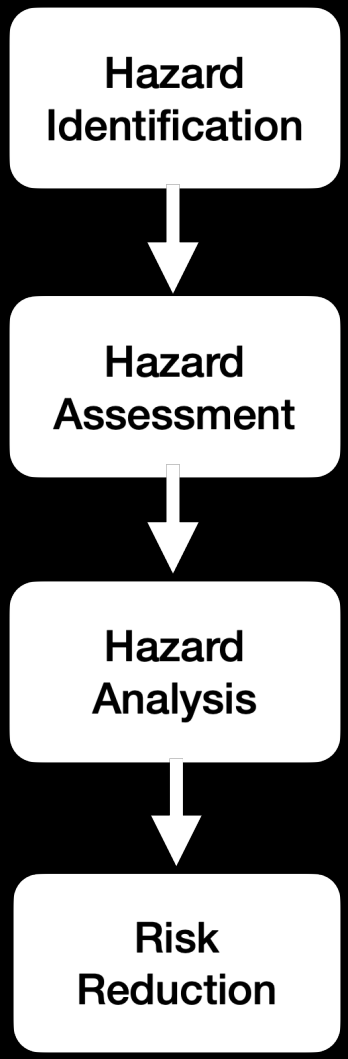
A potential source of harm caused by malfunctioning behaviour of the item where harm is physical injury or damage to the health of persons

Hazard Avoidance: the system is designed so that hazards are avoided.

Hazard Detection and Removal: The system is designed so that hazards are detected and removed before they result in an accident.

Damage limitation: The system may include protection features that minimize the damage that may result from an accident.

https://commons.wikimedia.org/wiki/File:GHS_HAZCOM_Safety_Labels.jpg



https://commons.wikimedia.org/wiki/File:Insulin_Delivery_Devices.png

Software Engineering: Sommerville, Ian

Hazard Identification

Hazard Assessment

Hazard Analysis

Risk Reduction

Physical, Electrical, Biological, Radiation, Service failure -> Hazard Register

poor sensor and actuator contact caused by incorrect fitting (physical)

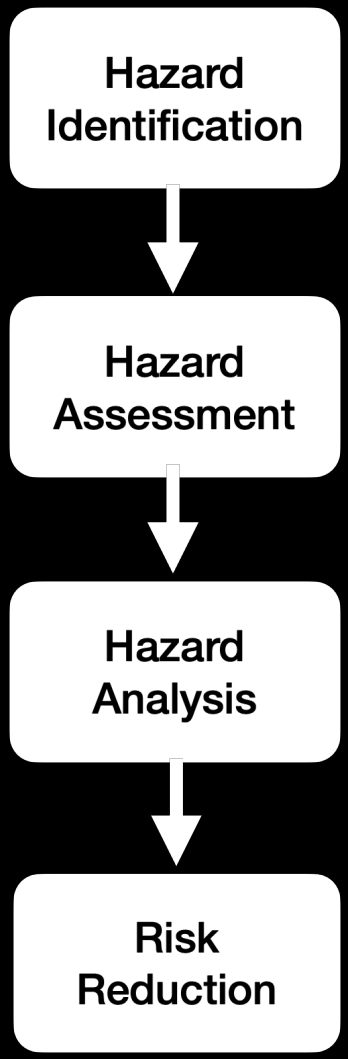
insulin overdose computation (service failure)

power failure due to exhausted battery (electrical)

infection caused by introduction of machine (biological)

https://commons.wikimedia.org/wiki/File:Insulin_Delivery_Devices.png

Software Engineering: Sommerville, Ian



Physical, Electrical, Biological, Radiation, Service failure -> Hazard Register

Understand the factors and c

Identified hazard	Hazard probability	Accident severity	Estimated risk	Acceptability
1. Insulin overdose computation	Medium	High	High	Intolerable
2. Insulin underdose computation	Medium	Low	Low	Acceptable
3. Failure of hardware monitoring system	Medium	Medium	Low	ALARP
4. Power failure	High	Low	Low	Acceptable
5. Machine incorrectly fitted	High	High	High	Intolerable
6. Machine breaks in patient	Low	High	Medium	ALARP
7. Machine causes infection	Medium	Medium	Medium	ALARP
8. Electrical interference	Low	High	Medium	ALARP
9. Allergic reaction	Low	Low	Low	Acceptable

Software Engineering: Sommerville, Ian

Hazard Identification

Hazard Assessment

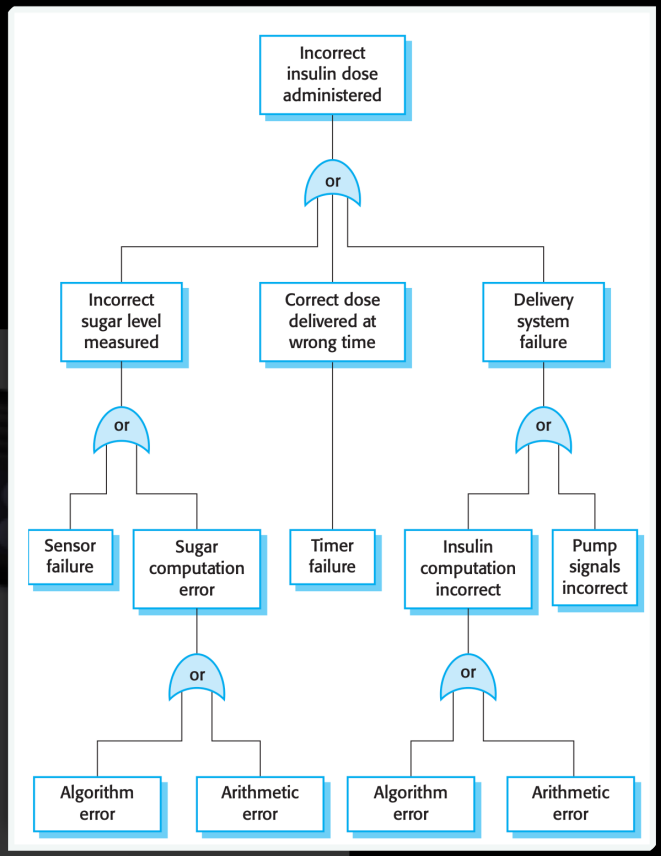
Hazard Analysis

Risk Reduction

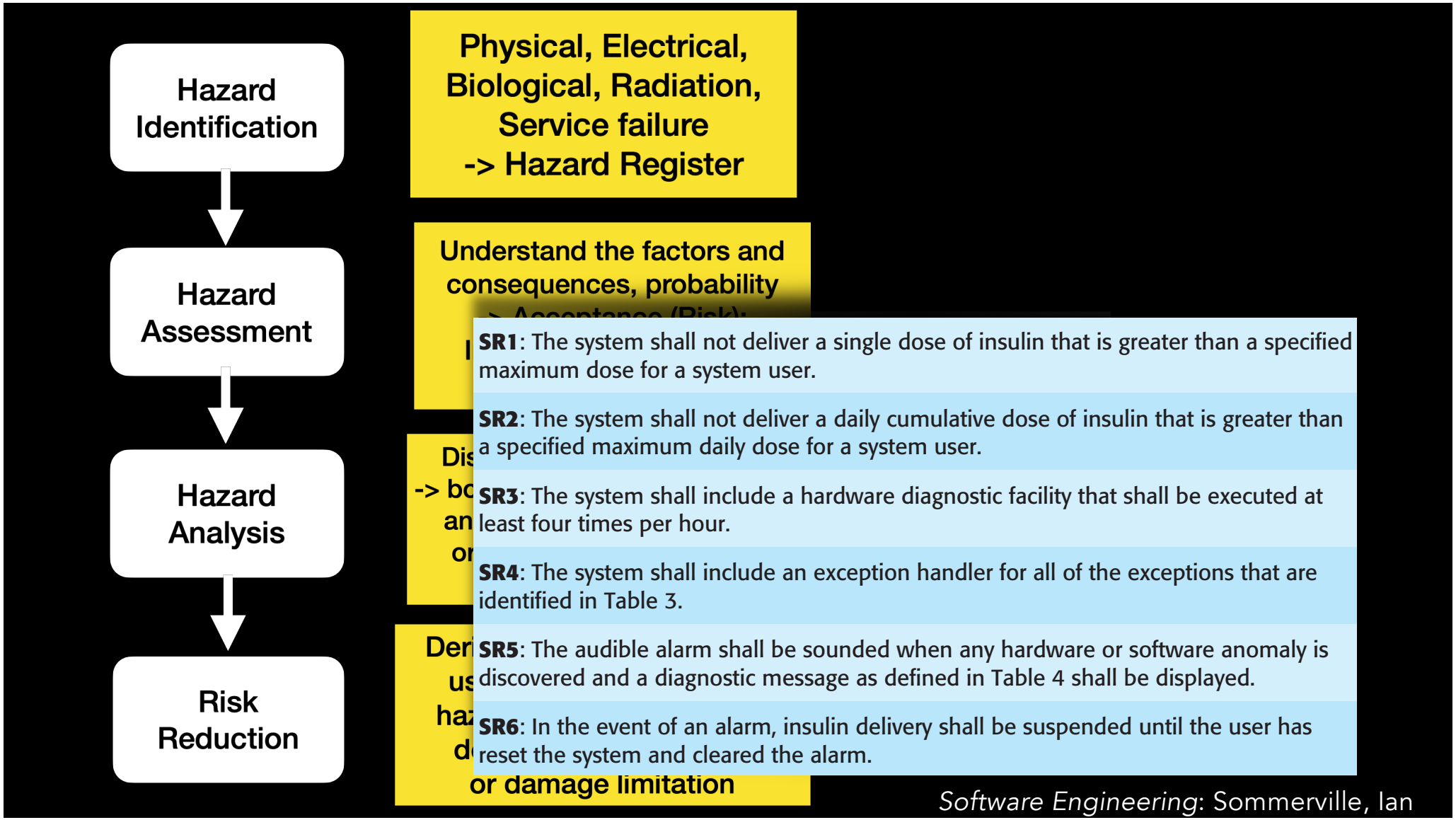
Physical, Electrical, Biological, Radiation, Service failure -> Hazard Register

Understand the factors and consequences, probability -> Acceptance (Risk): Intolerable, Reasonably Practical, Acceptable

Discover hazard root causes -> bottom-up, e.g. Failure Modes and Effects Analysis (FMEA) or top down, e.g. Fault Tree Analysis (FTA)



Software Engineering: Sommerville, Ian



Software Engineering: Sommerville, Ian

Safety Assurance

- Hazard analysis and monitoring
  - Traced from preliminary hazard analysis through to testing and system validation
- Safety reviews
- Safety certification

Hazard Register. **Page 4: Printed 20.02.2012**

System: Insulin Pump System *File:* InsulinPump/Safety/HazardLog
Safety Engineer: James Brown *Log version:* 1/3

Identified Hazard	Insulin overdose delivered to patient				
Identified by	Jane Williams				
Criticality class	1				
Identified risk	High				
Fault tree identified	YES	*Date*	24.01.11	*Location*	Hazard register, Page 5
Fault tree creators	Jane Williams and Bill Smith				
Fault tree checked	YES	*Date*	28.01.11	*Checker*	James Brown

System safety design requirements

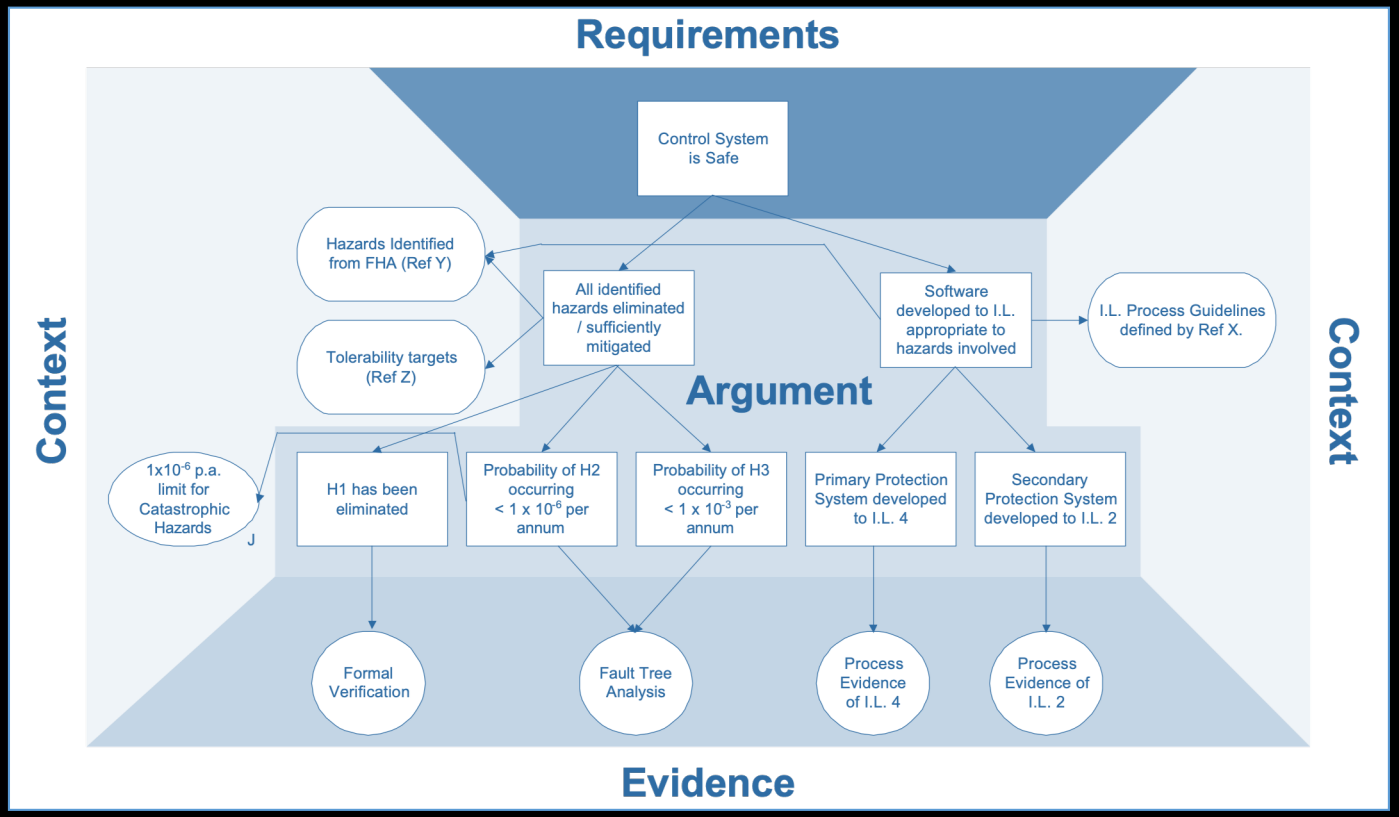
1. The system shall include self-testing software that will test the sensor system, the clock, and the insulin delivery system.
2. The self-checking software shall be executed once per minute.
3. In the event of the self-checking software discovering a fault in any of the system components, an audible warning shall be issued and the pump display shall indicate the name of the component where the fault has been discovered. The delivery of insulin shall be suspended.
4. The system shall incorporate an override system that allows the system user to modify the computed dose of insulin that is to be delivered by the system.
5. The amount of override shall be no greater than a pre-set value (maxOverride), which is set when the system is configured by medical staff.

Safety Case

A documented body of evidence that provides a convincing and valid argument that a system is adequately safe for a given application in a given environment.

- Test results
- Analyses
- Model checking results
- Expert opinion
- ...

Safety Case — Using Goal Structuring Notation

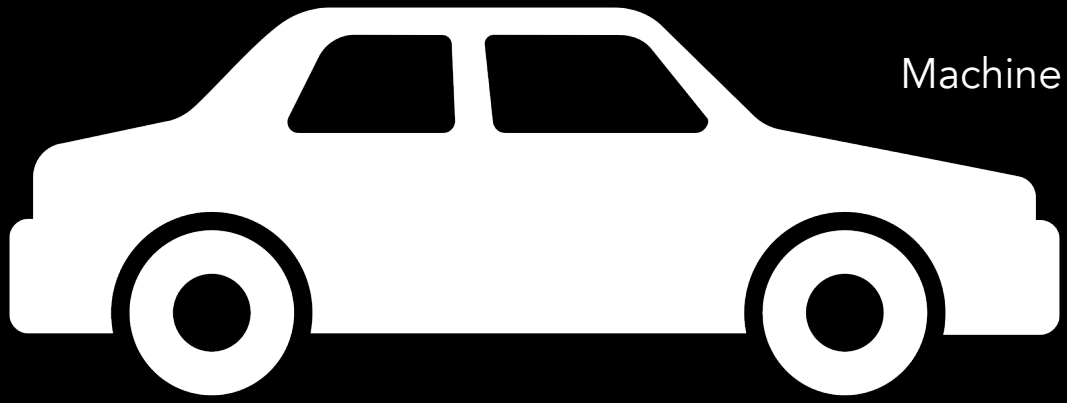


Kelly, Timothy Patrick. "Arguing safety: a systematic approach to managing safety cases." PhD diss., University of York, 1999.

Hazard Sources - Autonomous Vehicle

Computer Hardware and Software Faults

Mechanical and Electrical Faults



Machine learning Errors

Cyber Attacks

Inadequate Driving Behaviour
Dynamic Driving Tasks Feedback Failures

Sensor Noise and Limitations

"The recorded telemetry showed the system had detected Herzberg six seconds before the crash, and classified her first as an unknown object, then as a vehicle, and finally as a bicycle, each of which had a different predicted path
onomy logic. 1.3 seconds prior to the
etermined that emergency braking was
rmally performed by the vehicle
he system was not designed to alert the
make an emergency stop on its own
y braking maneuvers are not enabled
nder computer control, to reduce the
ehicle behavior'"

howing the paths of the pedestrian
w of the Uber test vehicle, showing

"On Thursday, June 21, the Tempe Police Department released a detailed report along with media captured after the collision, including an audio recording of the 911 call made by the safety driver, Rafaela Vasquez and an initial on-scene interview with a responding officer, captured by body worn video. After the crash, police obtained search warrants for Vasquez's cellphones as well as records from the video streaming services Netflix, YouTube, and Hulu. The investigation concluded that because the data showed she was streaming The Voice over Hulu at the time of the collision, and the driver-facing camera in the Volvo showed "her face appears to react and show a smirk or laugh at various points during the time she is looking down", Vasquez may have been distracted from her primary job of monitoring road and vehicle conditions.[44] Tempe police concluded the crash was "entirely avoidable"[45] and faulted Vasquez for her 'disregard for assigned job function to intervene in a hazardous situation'."

WY18MH010". National Transportation Safety Board. May 24, 2018. Retrieved May 26, 2018.

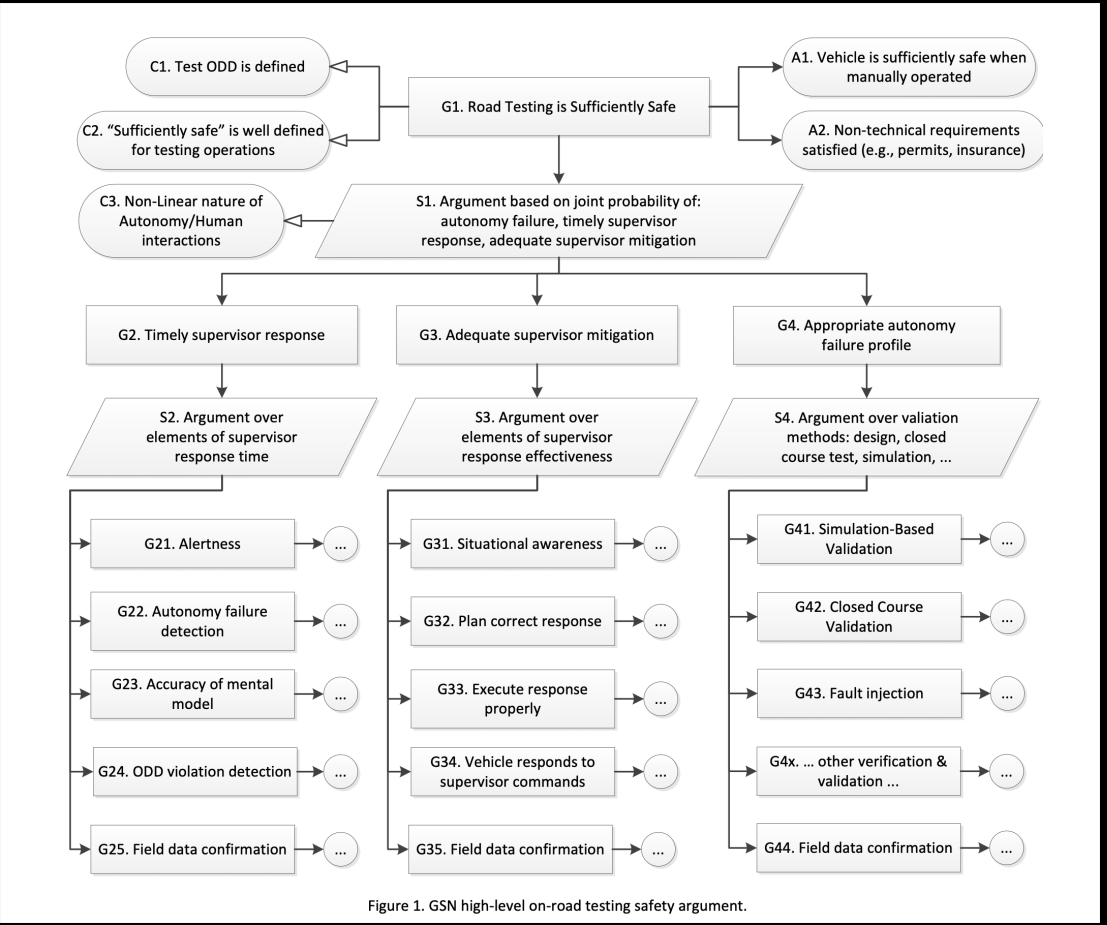


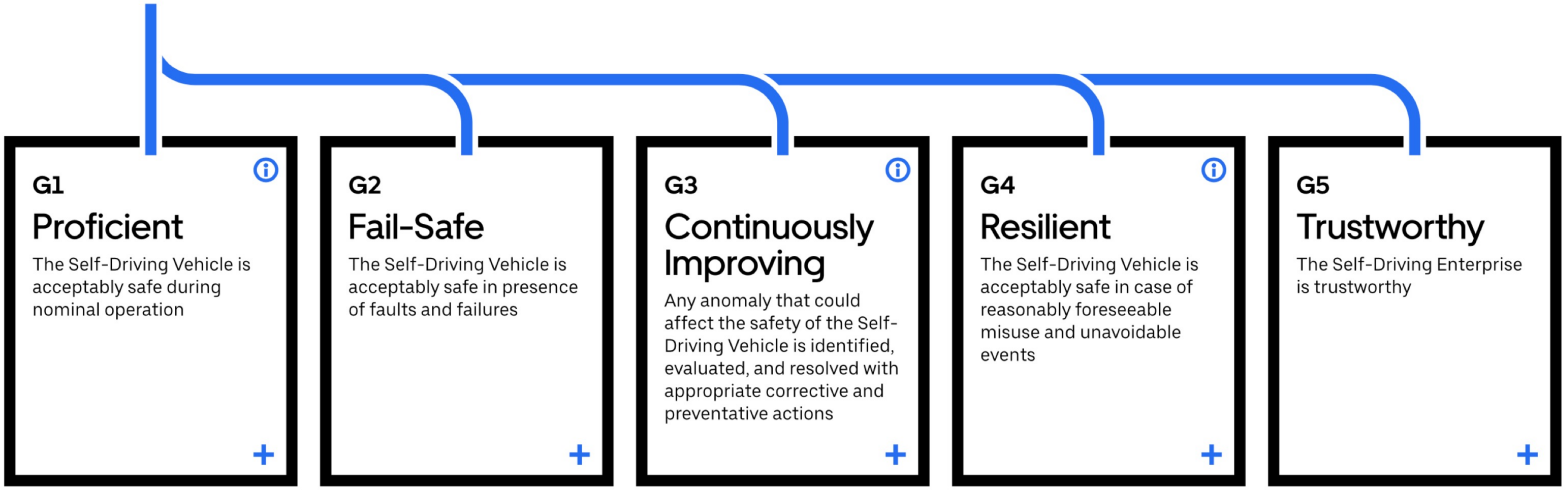
Figure 1. GSN high-level on-road testing safety argument.

Koopman, Philip, and Beth Osyk. "Safety argument considerations for public road testing of autonomous vehicles." *SAE International Journal of Advances and Current Practices in Mobility* 1, no. 2019-01-0123 (2019): 512-523.

Uber ATG

Home › Safety Case Framework

Our Self-Driving Vehicles are acceptably safe to operate on public roads



Agenda

- Case study - Autonomous Vehicle
- Safety Engineering
  - Hazard Analysis
  - Safety Assurance
- Adapting ISO262262 for Automotive Software with ML

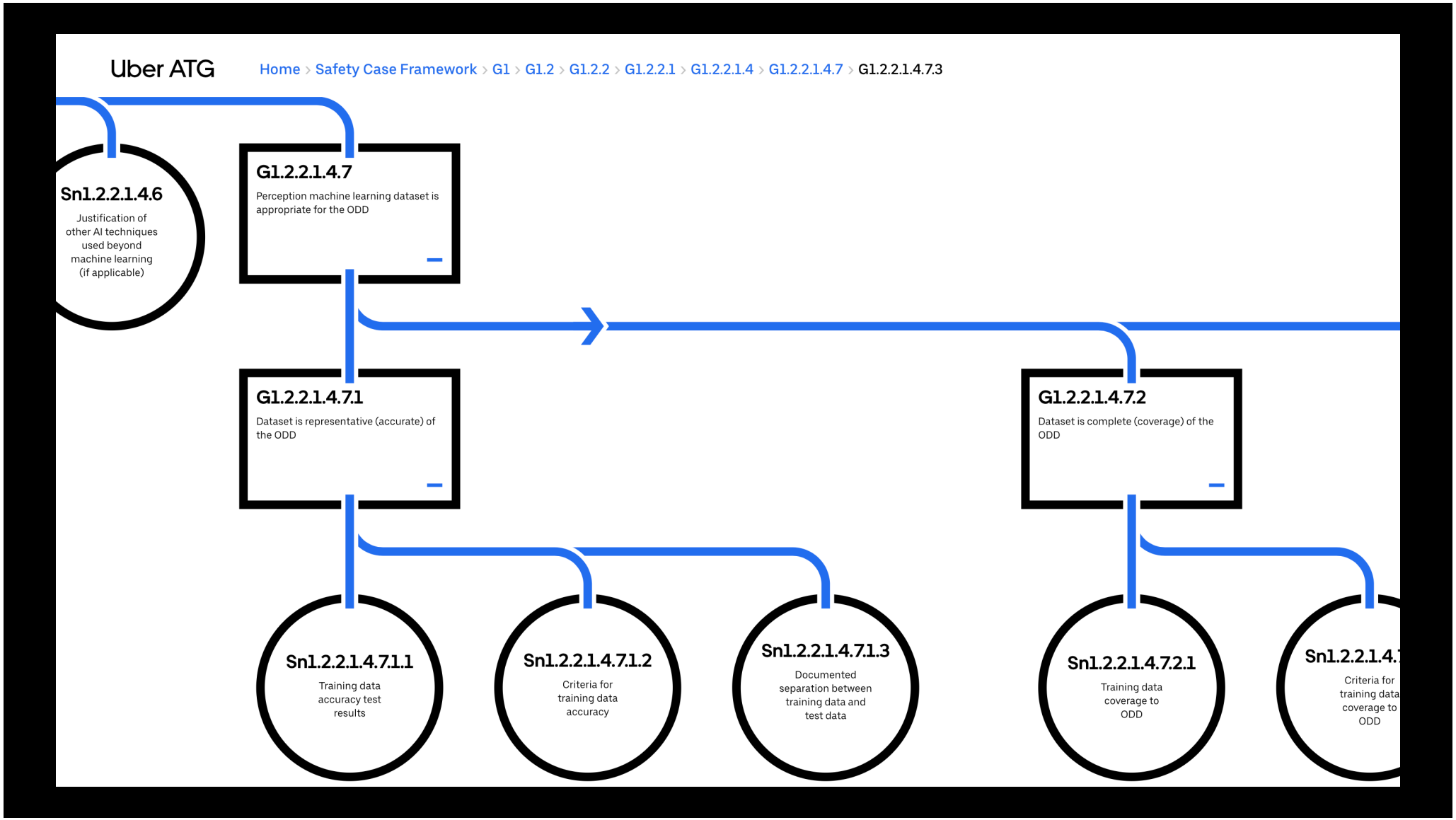
Adapting ISO26262 for Automotive Software with ML

(Req 6.4.1) The software safety requirements shall address each software-based function whose failure could lead to a violation of a technical safety requirement allocated to software.

(Req 6.4.1ML) The software safety requirements shall address each software-based function whose failure could lead to a violation of a technical safety requirement allocated to software.

The software safety requirements shall consist of two parts that jointly address the violation of technical safety requirements:
1. The strongest partial behavioural specification of each software safety requirement shall be defined.
2. A data set requirements specification shall be defined for the training/validation/testing data set.

Salay, Rick, and Krzysztof Czarnecki. "Using machine learning safely in automotive software: An assessment and adaption of software process requirements in ISO 26262." *arXiv preprint arXiv:1808.01614* (2018).


Uber ATG
Home › Safety Case Framework › G1 › G1.2 › G1.2.2 › G1.2.2.1 › G1.2.2.1.4 › G1.2.2.1.4.7 › G1.2.2.1.4.7.3
Sn1.2.2.1.4.6
Justification of other AI techniques used beyond machine learning (if applicable)
G1.2.2.1.4.7
Perception machine learning dataset is appropriate for the ODD
G1.2.2.1.4.7.1
Dataset is representative (accurate) of the ODD
G1.2.2.1.4.7.2
Dataset is complete (coverage) of the ODD
Sn1.2.2.1.4.7.1.1
Training data accuracy test results
Sn1.2.2.1.4.7.1.2
Criteria for training data accuracy
Sn1.2.2.1.4.7.1.3
Documented separation between training data and test data
Sn1.2.2.1.4.7.2.1
Training data coverage to ODD
Sn1.2.2.1.4.
Criteria for training data coverage to ODD

Adapting ISO26262 for Automotive Software with ML

(Req MLTR1) An analysis shall be carried out to assess the adequacy of the training procedure. The analysis shall address the following aspects:
a) control over differences between the operating and training environments (also see Req 9.4.6);
b) handling of distributional shift;
c) representation of safety in the loss function; and,
d) adequacy of regularization

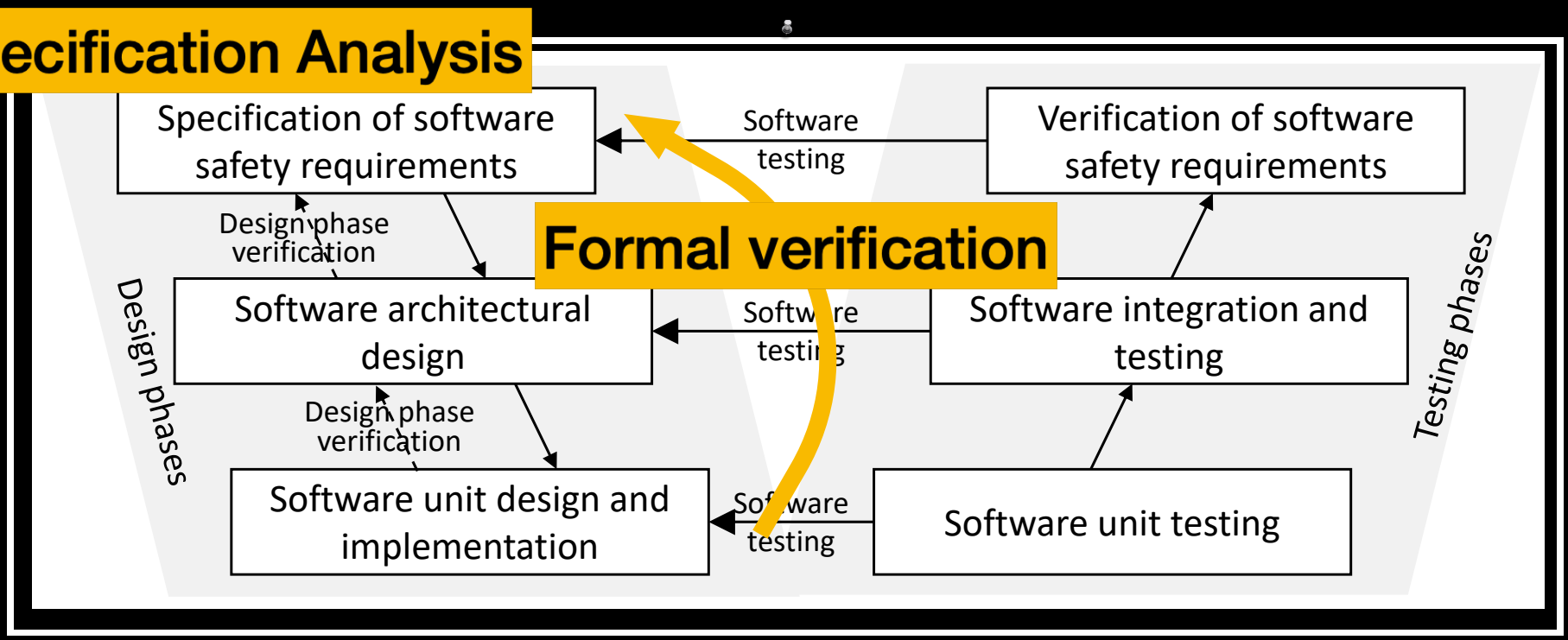
(Req MLTR2) The training procedure shall incorporate the partial specifications resulting from Req 6.4.1ML to the extent possible.

Salay, Rick, and Krzysztof Czarnecki. "Using machine learning safely in automotive software: An assessment and adaption of software process requirements in ISO 26262." *arXiv preprint arXiv:1808.01614* (2018).

PR	Phase	Description	Applicability to M
5.4.6, 5.4.7	(5) Initiation	Best practices: coding guidelines	Generally not appl could improve MI O2).
MLIN1		ML decision gate	A new process req necessary implem requirements.
6.4.1ML 8.4.3, 9.4.4	(6) Software Safety Requirements	Requirements Specification	Partially applicabl partial behavioura specified complete complete data set producing these re
6.4.8ML		Requirements verification	Applicable and an input domain to co incompleteness.
7.4.14 7.4.15	(7) Architectural Design	Fault tolerance	Applicable and va tolerance strategie
8.4.2	(8) Software unit design, implementation	Best practices: notations	Generally applicab not yet exist.
8.4.4		Best practices: design principles	Not applicable because they are biased toward imperative programming. The intent of these practices is fault minimization and interpretability.
MLDS1, MLDS2, MLDS3		Data set collection and verification	New process requirements to address data set requirements. Methods for data set augmentation and assessing uncertainty are discussed.
MLMS1, MLMS2		Model selection	New process requirements to address model selection. Selection principles are discussed.
MLFS1		Feature selection	New process requirement to address feature selection.
MLTR1, MLTR2		Training	New process requirements to address faults in the training procedure.
MLVT1		Data set splitting	New process requirements to address how to split the data set into training, validation and test sets.
MLVT2		Validation	New process requirement to address validation and hyper-parameter selection.
9.4.3		Testing	Generally applicable but some methods must be adapted for ML.
9.4.5ML		Testing structural coverage	Not directly applicable but the intent of the coverage metrics can be achieved through alternative similar metrics. The requirement is also amended to ensure new tests satisfy data requirements.
9.4.6		Test vs. operating environment	Applicable but there is a heightened relevance of this requirement for ML because of limited specifications.
MLTE1		Test result explanation	New process requirement to validate the reason a test passes or fails.
8.4.5		Verification	Generally applicable but is highly reliant on solutions to the interpretability obstacle O2.

Salay, Rick, and Krzysztof Czarnecki
An assessment and adaption of soft
arXiv:1808.01614 (2018).

Formal verification



- Next (on Zoom)

AI Security