KATHOLISCHE UNIVERSITÄT
EICHSTÄTT-INGOLSTADT

# Responsible AI und automomes Fahren

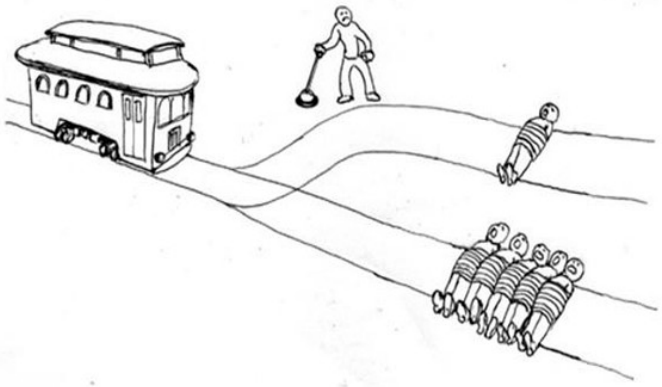
**Florian Richter**

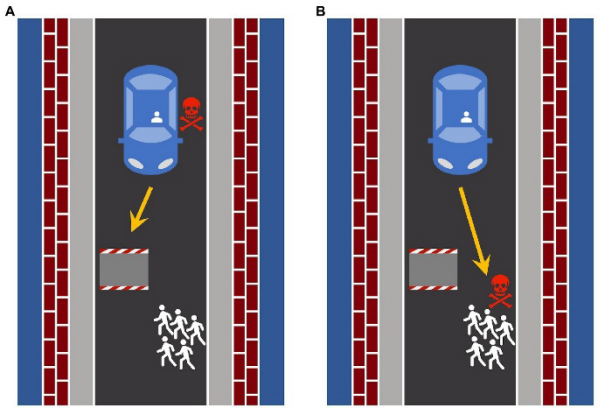
# Responsible AI und autonomes Fahren

1. Technikethik
2. Responsible AI
3. Sicherheit
4. AI Safety
5. Verantwortungslücken

# Technikethik

# Autonomes Fahren und Trolley Dilemma





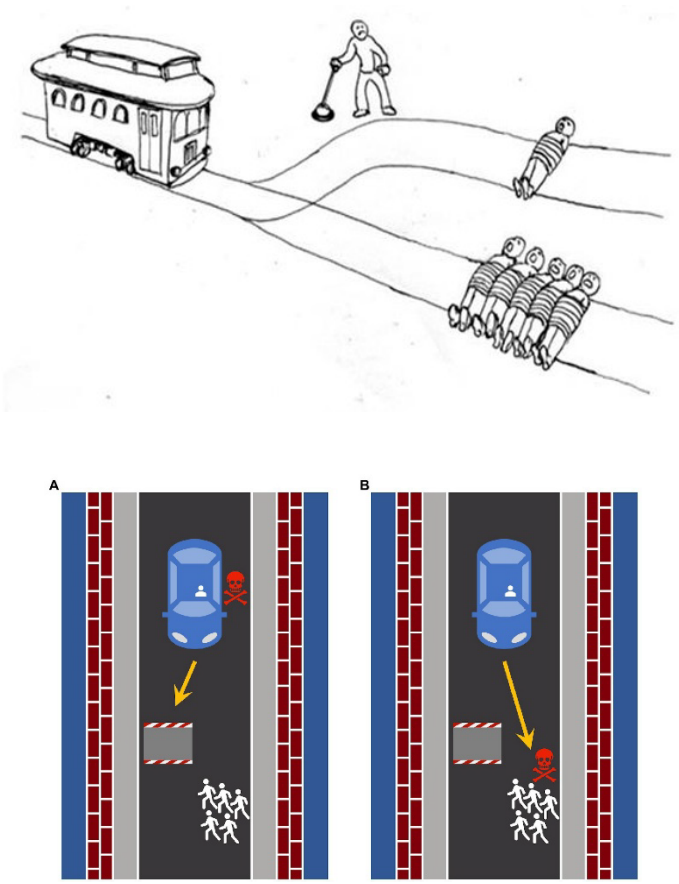
# Technikethik

**Maschinenethik:** Catrin Misselhorn: “Artificial Morality”

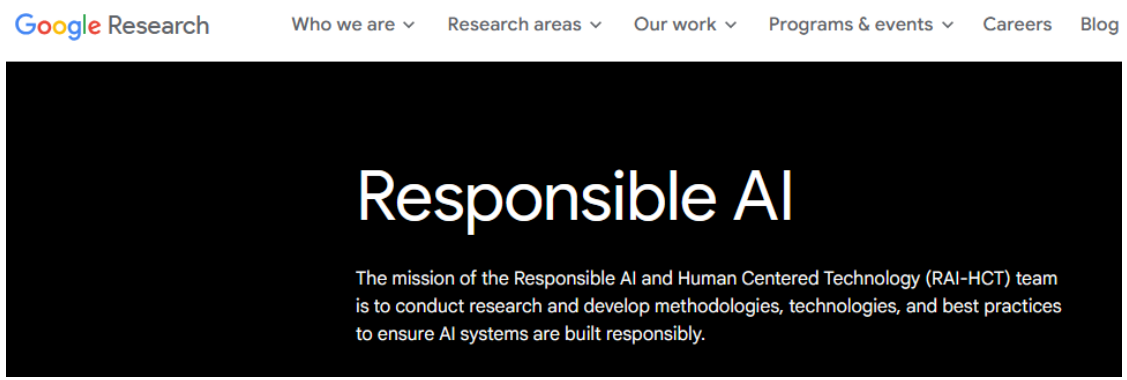
“As the examples show, already a rather simple artificial system like a vacuuming robot faces moral decisions. The more intelligent and autonomous these technologies become, the more intricate the moral problems will become that they are confronting. It is, therefore, important to think about the prospects and risks of artificial morality.” (162)


“If no solution to these dilemmas can be found this might become a serious impediment for autonomous driving.” (162)

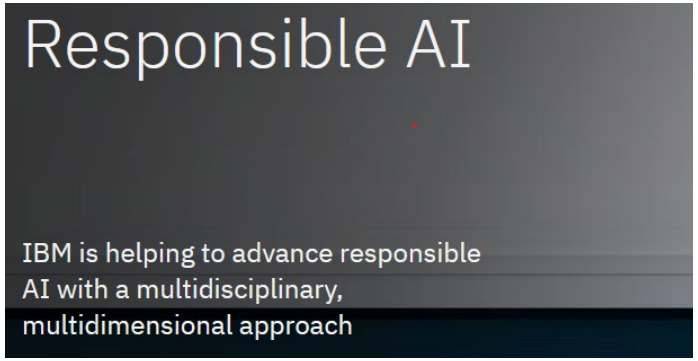
# Technikethik und Responsible AI

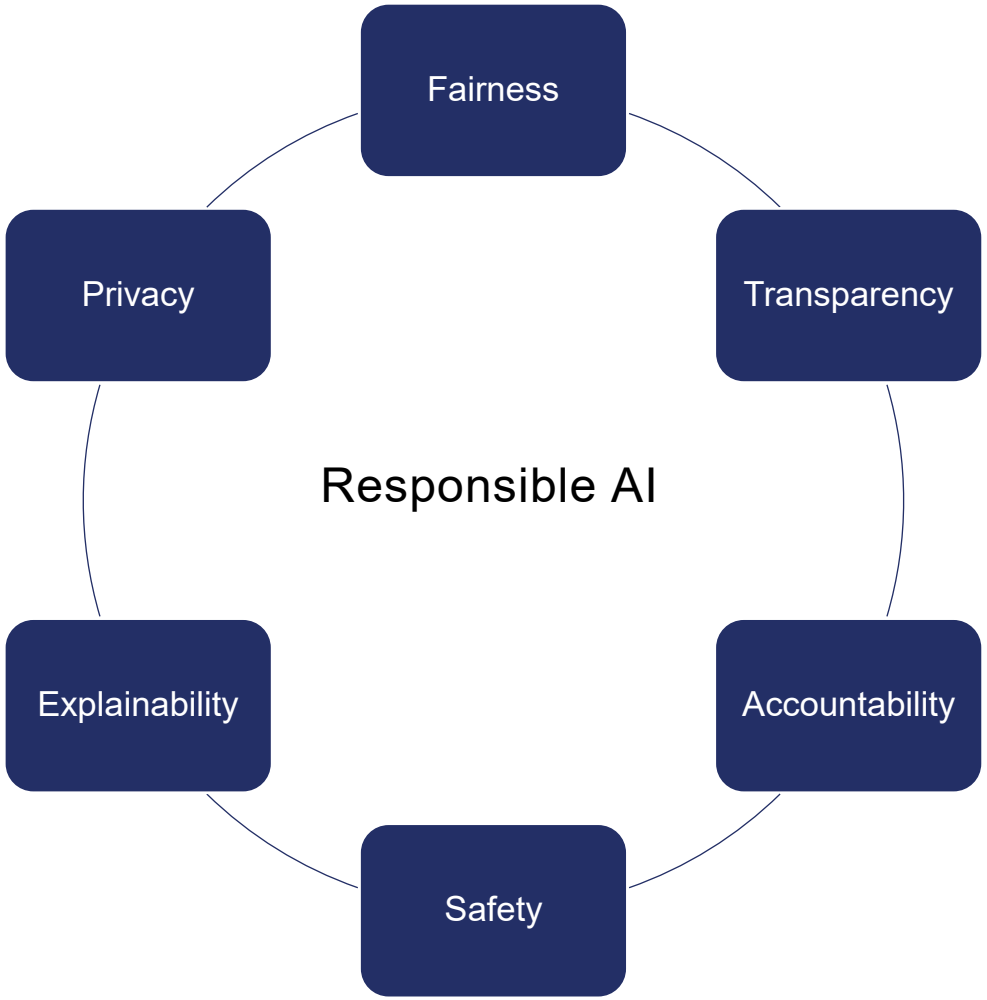
# Responsible AI





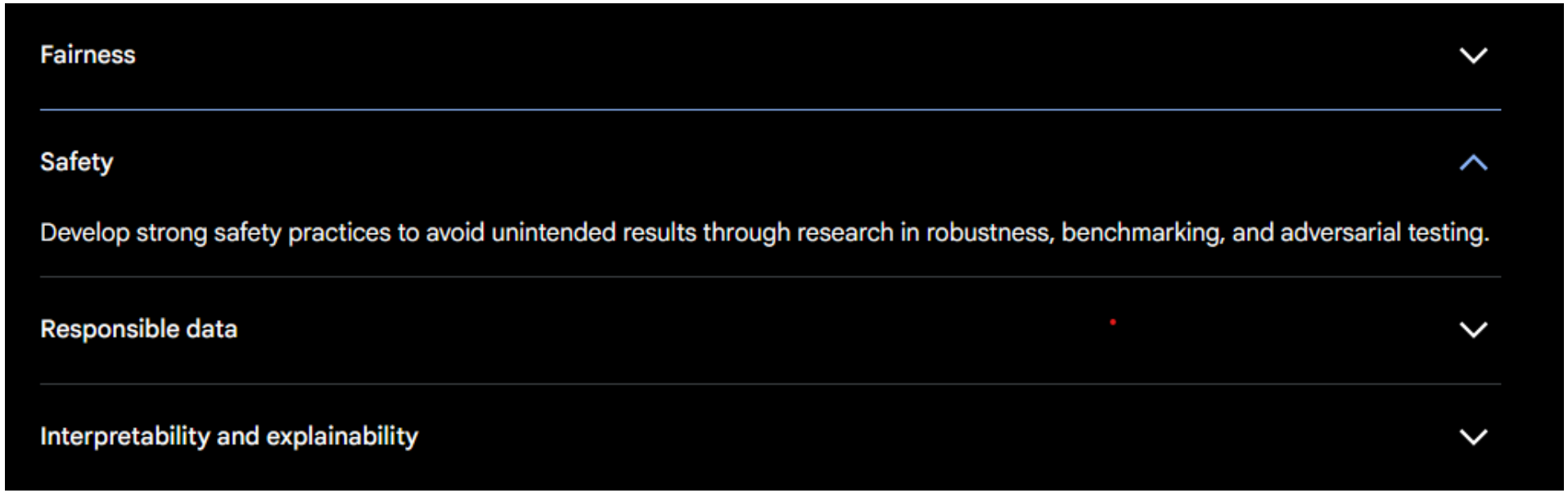


# Responsible AI

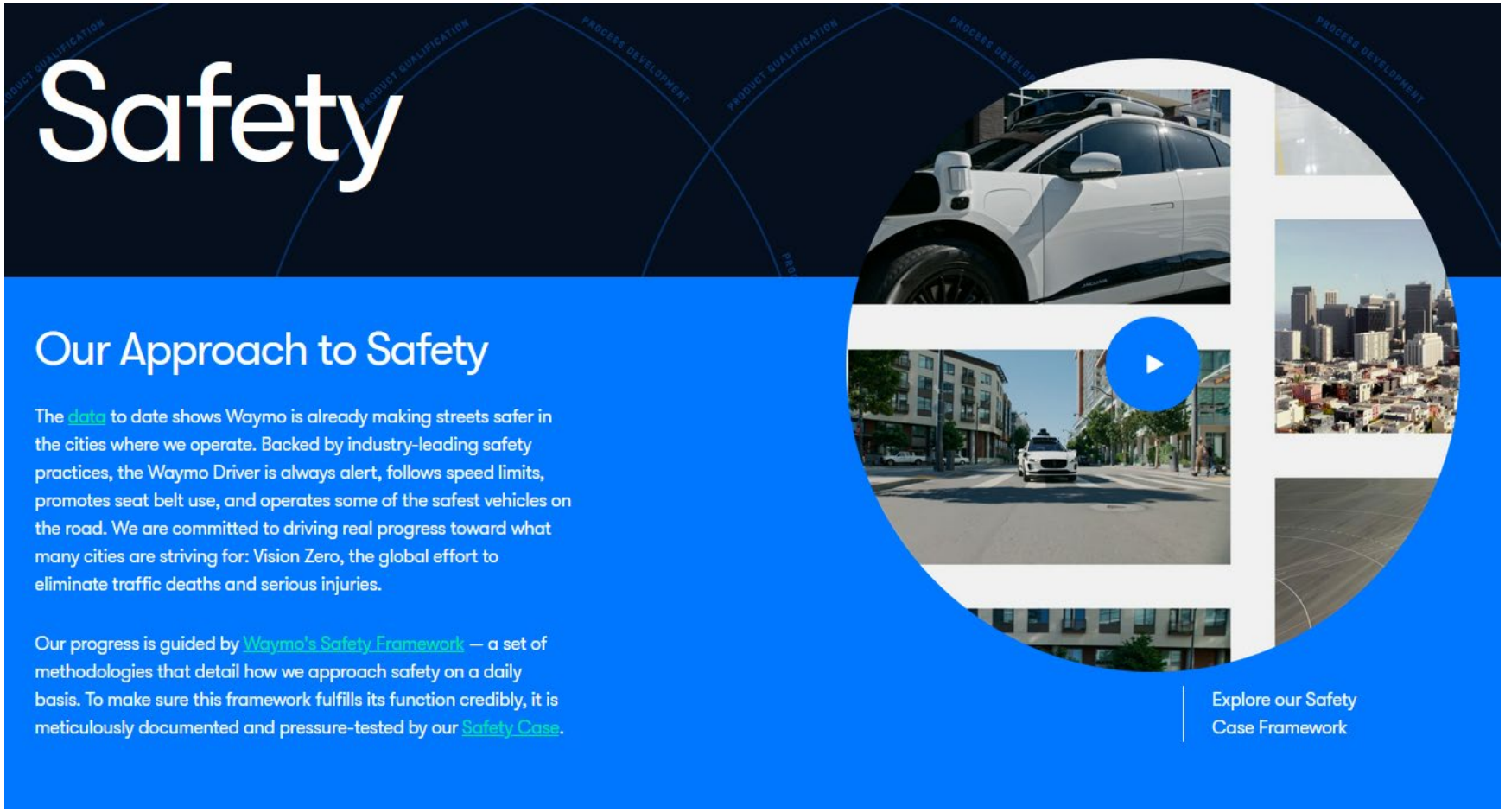


# Responsible AI und autonomes Fahren

Responsible AI (Google)



# Responsible AI und autonomes Fahren



# Sicherheit


Building a Credible Case for Safety:
Waymo's Approach for the Determination of Absence of Unreasonable Risk

March 2023

## 2.1 Defining Absence of Unreasonable Risk

Recently completed as well as ongoing standardization activities (ISO, 2022) (ISO/AWI TS 5083) formalize the determination of AUR through the specification of one or more *acceptance criteria* and associated *validation targets*. Standards ISO 26262:2018 (ISO, 2018a), ISO 21448:2022 (ISO, 2022), and UL 4600:2022 (UL, 2022) provide a series of definitions necessary to understand and correctly frame such an approach:

- Risk: combination of the probability of occurrence of harm and the severity of that harm (ISO, 2018a);
- Unreasonable risk: risk judged to be unacceptable in a certain context according to valid societal moral concepts (ISO, 2018a);[15]
- Acceptable: sufficient to achieve the overall item risk as determined in the safety case (UL, 2022);
- Acceptance criterion: criterion representing the absence of an unreasonable level of risk (ISO, 2022);
- Validation target: value to argue that the acceptance criterion is met (ISO, 2022);
- Residual risk: risk remaining after the deployment of safety measures (ISO, 2018a);

These definitions follow a cascading structure, which invites an explicit definition of the acceptance criteria that will be used to evaluate if the residual risk reaches and remains at an acceptable level. The draft ISO/AWI TS 5083 calls for "explicit risk acceptance criteria [...] expressed for the ADS in the context of the proposed use case and operational design domain, for each known source of harm." Acceptance Criteria (AC) are sensitive to the specific functionality being assessed, and even the specific methodology being employed. While they can be qualitative or quantitative (ISO, 2022), they require the specification of a measurable target for the determination of readiness in terms of an absence of unreasonable risk.

Different types of hazards (i.e., per ISO 26262:2018 definition, the "potential sources of harm" mentioned in the ISO/AWI TS 5083 quote) need to be considered and adequately identified before an appropriate and sufficiently comprehensive set of acceptance criteria for the system can be defined. We thus tackle the decomposition of the top-level goal for Waymo's Safety Case, that is, the determination of absence of unreasonable risk, through the identification of three categories of hazards on which Waymo's layered approach to safety is predicated.

# Sicherheit (Safety)

ISO 26262-1:2018(en) Road vehicles — Functional safety — Part 1: Vocabulary

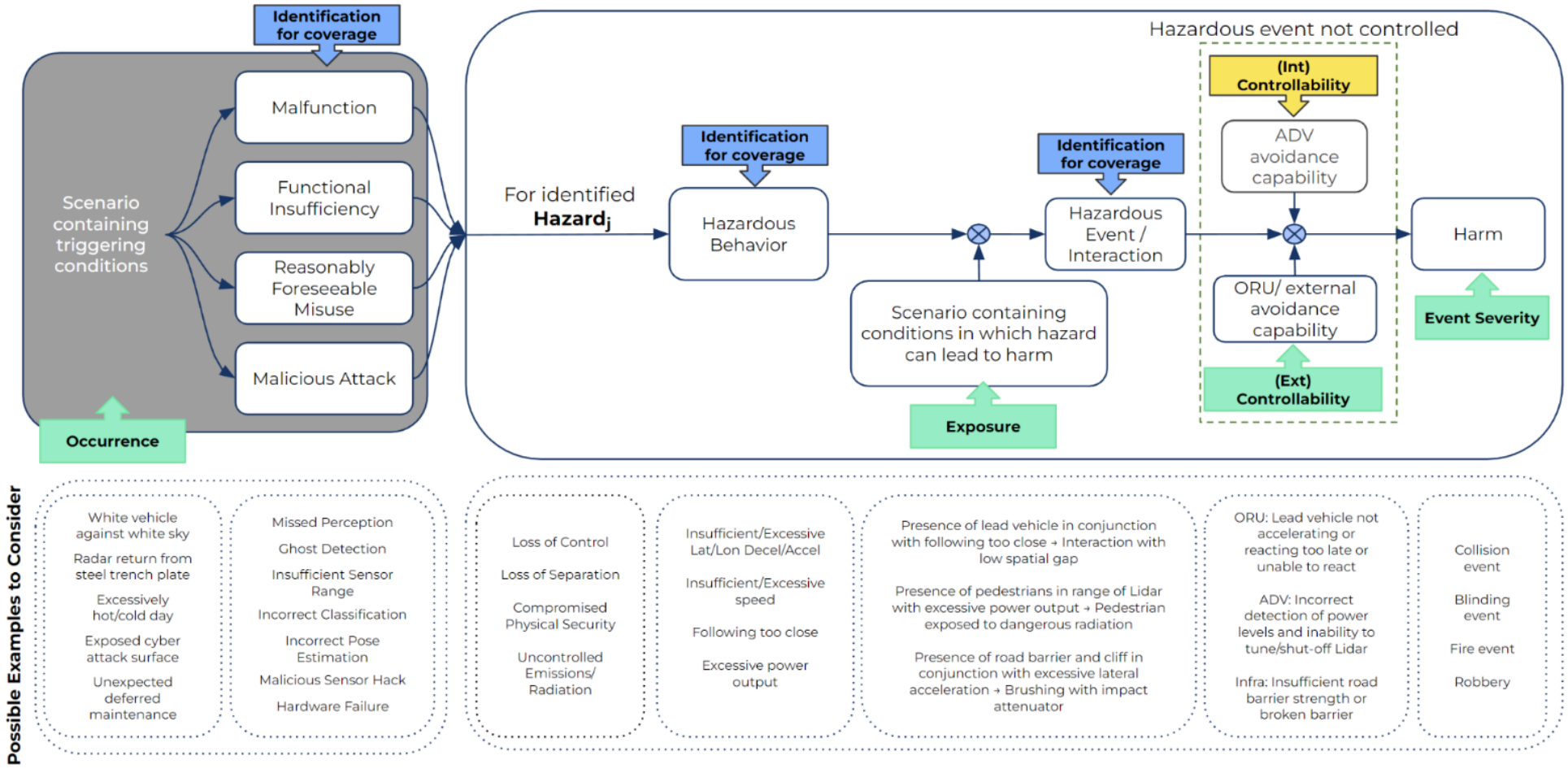
“3.132 safety

absence of unreasonable risk (3.176)”

“3.176 unreasonable risk

risk (3.128) judged to be unacceptable in a certain context according to valid societal moral concepts”

# Sicherheit


Identification for coverage
Malfunction
Functional Insufficiency
Reasonably Foreseeable Misuse
Malicious Attack
Scenario containing triggering conditions
Occurrence
Hazardous event not controlled
Identification for coverage
For identified Hazard$_j$
Hazardous Behavior
Identification for coverage
Hazardous Event / Interaction
(Int) Controllability
ADV avoidance capability
ORU/ external avoidance capability
(Ext) Controllability
Harm
Event Severity
Scenario containing conditions in which hazard can lead to harm
Exposure
Possible Examples to Consider
White vehicle against white sky
Radar return from steel trench plate
Excessively hot/cold day
Exposed cyber attack surface
Unexpected deferred maintenance
Missed Perception
Ghost Detection
Insufficient Sensor Range
Incorrect Classification
Incorrect Pose Estimation
Malicious Sensor Hack
Hardware Failure
Loss of Control
Loss of Separation
Compromised Physical Security
Uncontrolled Emissions/ Radiation
Insufficient/Excessive Lat/Lon Decel/Accel
Insufficient/Excessive speed
Following too close
Excessive power output
Presence of lead vehicle in conjunction with following too close → Interaction with low spatial gap
Presence of pedestrians in range of Lidar with excessive power output → Pedestrian exposed to dangerous radiation
Presence of road barrier and cliff in conjunction with excessive lateral acceleration → Brushing with impact attenuator
ORU: Lead vehicle not accelerating or reacting too late or unable to react
ADV: Incorrect detection of power levels and inability to tune/shut-off Lidar
Infra: Insufficient road barrier strength or broken barrier
Collision event
Blinding event
Fire event
Robbery

# Sicherheit

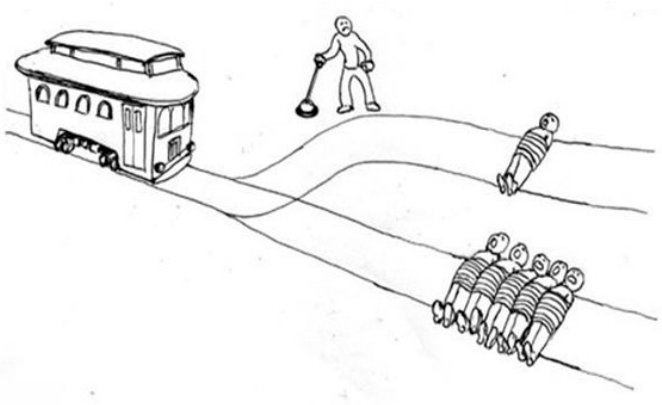
Beyond injury severity there are other forms of severity to consider, including but not limited to risks related to potential non-compliance to local statutes, notably non-compliance that may increase risk of a serious event. At this time we do not consider other measures such as years of potential life lost (YPLLs), disability (as captured by DALYs), quality of life (as captured by QALYs), pain and suffering, time off of work (as captured by Lost Workday Rates), etc., but the severity potential dimension of the AC framework could be expanded to account for that. In fact, this dimension remains agnostic to the specific modeling employed for the space of plausible consequences associated with the performance indicators upon which an acceptance criterion is predicated.

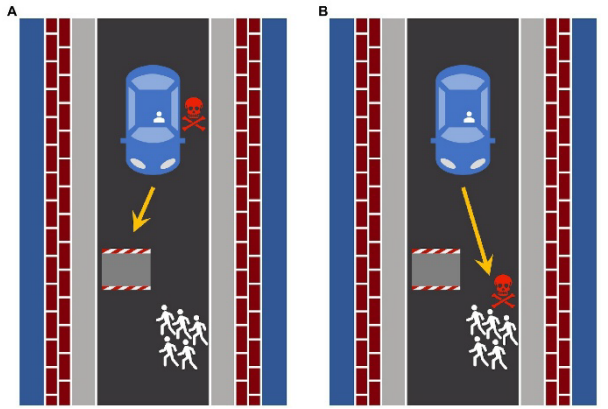
- Konsequenzen von Unfällen wie Tod oder Behinderungen werden in diesem Dokument nicht behandelt.

# Verantwortung und Sicherheit

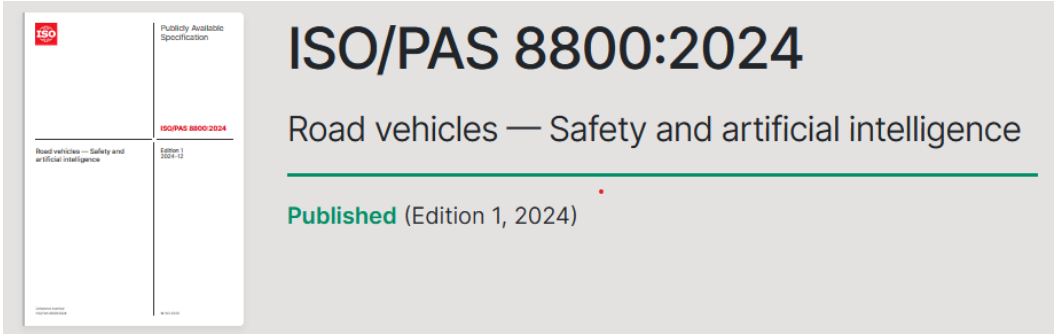
- Verantwortung: Wer ist in verschiedenen Risikoszenarien verantwortlich?

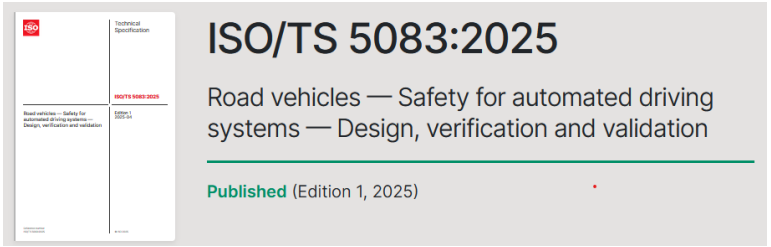
- Wie bewertet man die unakzeptablen Risiken “in a certain context according to valid societal moral concepts“? (ISO 26262-1:2018)





# AI Safety




“3.1.15 AI safety

absence of unreasonable risk (3.3.10) due to AI errors (3.4.1) caused by faults and functional insufficiencies

Note 1 to entry: This definition only applies in the context of this document. The term "AI safety" is commonly understood to have a broader meaning which includes ethics, value alignment, long-term considerations, etc.”

# Herausforderungen für AI Safety

- Offenheit der Umwelt:
  - Sind die Daten hinreichend, um die intendierte Funktionalität zu repräsentieren? – Zum Beispiel Kinder auf der Straße
- Black boxes:
  - Sind die Algorithmen zu einem gewissen Grad erklärbar?
- Performanz der Modelle:
  - Sind die Modelle robust? – **Unvorhersehbares Verhalten** des automatisierten Autos

# Herausforderungen für AI Safety

- Offenheit der Umwelt:
  - Sind die Daten hinreichend, um die intendierte Funktionalität zu repräsentieren? – Zum Beispiel Kinder auf der Straße

- Black boxes:
  - Sind die Algorithmen zu einem gewissen Grad erklärbar?

- Performanz der Modelle:
  - Sind die Modelle robust? – **Unvorhersehbares Verhalten** des automatisierten Autos

    ➢ Entstehen damit **Verantwortungslücken**? D.h. man könnte damit nicht mehr Verantwortung zuschreiben.

# Verantwortungslücken

Umgang mit Verantwortungslücken laut Santoni de Sio und Mecacci:

- Fatalismus: neue lernfähige Systeme können nicht schuldfähig sein – entweder Schuldfähigkeit aufgeben und neue Systeme einführen oder sie nicht einführen, weil sie nicht schuldfähig sein können (s. Misselhorn).


- Solutionismus: technische Lösung – man kann die Systeme nachvollziehbar, transparent und erklärbar machen.

- Deflationismus (Peter Königs): es handelt sich um kein neuartiges Problem und man kann es mit moralischen und rechtlichen Strategien behandeln, die man schon hat.


# Verantwortungslücken

- Wir müssen einen anderen Ansatz wählen, da wir keinen Goldstandard für den Umgang mit (teil-)autonomen Systemen haben.
- Weitere Möglichkeiten, mit der Zuschreibung von Verantwortung umzugehen, sind der Anscheinsbeweis, wie z.B. bei Auffahrunfällen: Es wird prima facie vermutet, dass der auffahrende Fahrer schuld ist (zu wenig Sicherheitsabstand, überhöhte Geschwindigkeit, Unaufmerksamkeit).
- Die Annahme, dass der Fahrer vorne auch dafür verantwortlich wäre, würde zu rechtlichen und versicherungstechnischen Untersuchungen führen, die zu hohe Kosten verursachen und den geringen Nutzen nicht rechtfertigen.
- Ähnlich verhielt es sich mit der Rechtsprechung im Fall der Arzneimittelzulassung nach dem sogenannten "Contergan-Skandal". Pharmaunternehmen sind verpflichtet, die Sicherheit und Wirksamkeit ihrer Produkte zu gewährleisten. Sie sind prima facie schuld, wenn ein Schaden entsteht, und damit auch für den Schaden verantwortlich.

**Drug Amendments of 1962**

| | |
|---|---|
| **Long title** | An act to protect the public health by amending the Federal Food, Drug, and Cosmetic Act to assure the safety, effectiveness, and reliability of drugs, authorize standardization of drug names, and clarify and strengthen existing inspection authority; and for other purposes. |
| **Nicknames** | Drug Efficacy Amendment<br>Kefauver–Harris Amendment |
| **Enacted by** | the 87th United States Congress |
| **Effective** | October 10, 1962 |

Historische Contergan-forte-Verpackung

# Verantwortungslücken

- Sie müssen für den Schaden aufkommen oder nach angemessenen Entschädigungsmöglichkeiten suchen. Pharmaunternehmen können sich nicht auf
  (1) Nachlässigkeit berufen,
  (2) dass es sehr komplex ist, Medikamente zu testen,
  (3) dass „viele Hände“ beteiligt sind und
  (4) dass man nicht die Absicht hat, Schaden anzurichten.

- Es handelt sich um eine Gefährdungshaftung (*strict liability*): Verantwortlich für die Folgen, auch wenn es keine Absicht gab, zu schädigen.
- Als Patient haben Sie nicht die Ressourcen, sich über das Medikament und mögliche schwere Nebenwirkungen zu informieren, insbesondere wenn das Unternehmen noch nicht einmal Tests damit durchgeführt hat.

# Verantwortungslücken

- Dementsprechend sollte bei der Entwicklung eines KI-basierten oder (teil-)autonomen Systems eine Gefährdungshaftung (*strict liability)* in Betracht gezogen werden, um Unternehmen und Organisationen für Fehlfunktionen, Unfälle oder moralische Schäden, die den Nutzern zugefügt werden, zur Verantwortung zu ziehen.

- Dies würde zu einer sorgfältigeren Entwicklung von Anwendungen führen, um sie zur Marktreife zu bringen. Die Nutzerinnen und Nutzer können sich nicht selbst informieren, haben oft keine Transparenz über Systemstrategien und können die Implikationen für ihr kritisches und moralisches Urteilsvermögen nicht überblicken.

# Fragen

**Vielen Dank für Ihre Aufmerksamkeit! Haben Sie Fragen?**

florian.richter@ku.de