CHALMERS | GÖTEBORGS UNIVERSITET

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbi
FACULTY OF COMPUTER SCIENCE

MÄLARDALEN UNIVERSITY SWEDEN

Presented 2019 03 15 at the seminar @NTNU, Trondheim

Avoiding the Intrinsic Unfairness of the Trolley Problem:Towards Technically and Socially Informed Ethical Guidelines for Designing of Self-driving Cars

Tobias Holstein
University of Applied Sciences
Darmstadt, Germany &
Mälardalen University
Sweden

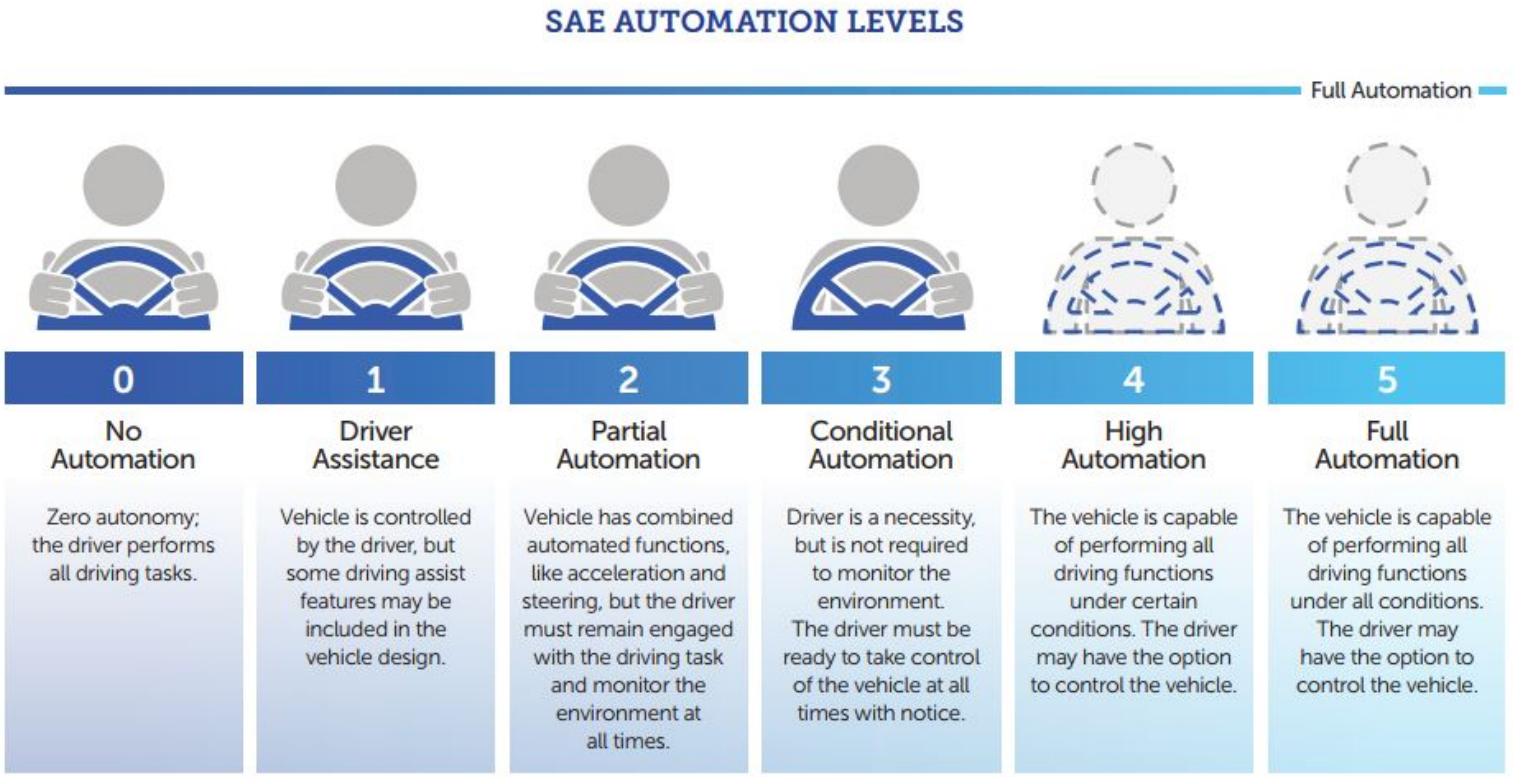
Gordana Dodig-Crnkovic
Chalmers University of Technology
Gothenburg & Mälardalen University
Sweden

This presentation is based on the work presented at FairWare'18 - International Workshop on Software Fairness @ICSE2018

https://www.researchgate.net/publication/323869526_Avoiding_the_Intrinsic_Unfairness_of_the_Trolley_Problem

It can be found under http://www.gordana.se/work/presentations.html

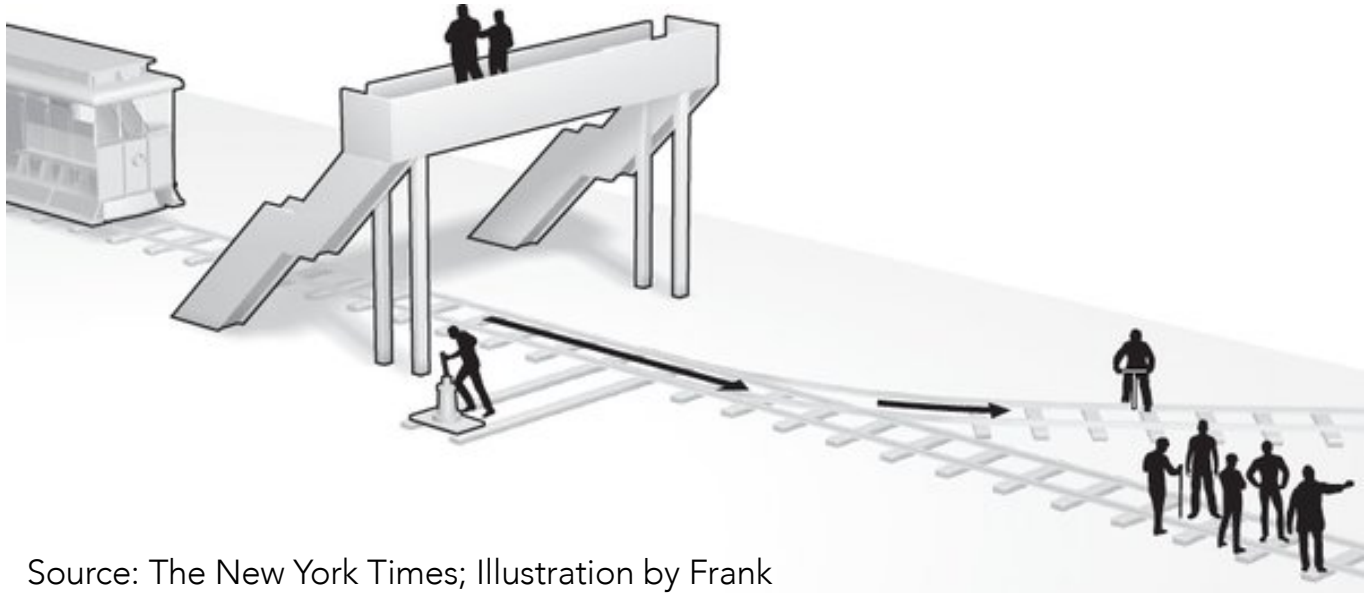
LEVELS OF CAR AUTOMATION



Current discussions about self-driving cars repeatedly take form of decision-making situations borrowed from philosophy in the form of “the trolley problem”:

WHOM WILL THE SELF-DRIVING CAR KILL WHEN IT HAS TO DECIDE?

Ethical Dilemma: The Trolley Problem



Source: The New York Times; Illustration by Frank O'Connell

- Ethical thought experiment defined by philosopher Philippa Foot in "The Problem of Abortion and the Doctrine of the Double Effect," pp. 5-15, Oxford Review, 5, (1967).

- Many different variants, such as the use of personas to include an emotional perspective. But there is always a single decision: Who is going to be killed?

Typical Approaches to the Trolley Problem…

…are typically based on the following ethical theories:

- Utilitarianism
- Other forms of consequentialism
- Deontological ethics

For example, utilitarianism would aim to minimize fatalities,
even if it means to kill the passenger in the car, by following the principle: the moral action is the one that maximizes utility
(or in this case minimizes the damage).

Depending on the ethics framework,
different arguments can be used to justify the decision.

The Trolley Problem is Unsolvable by Construction!

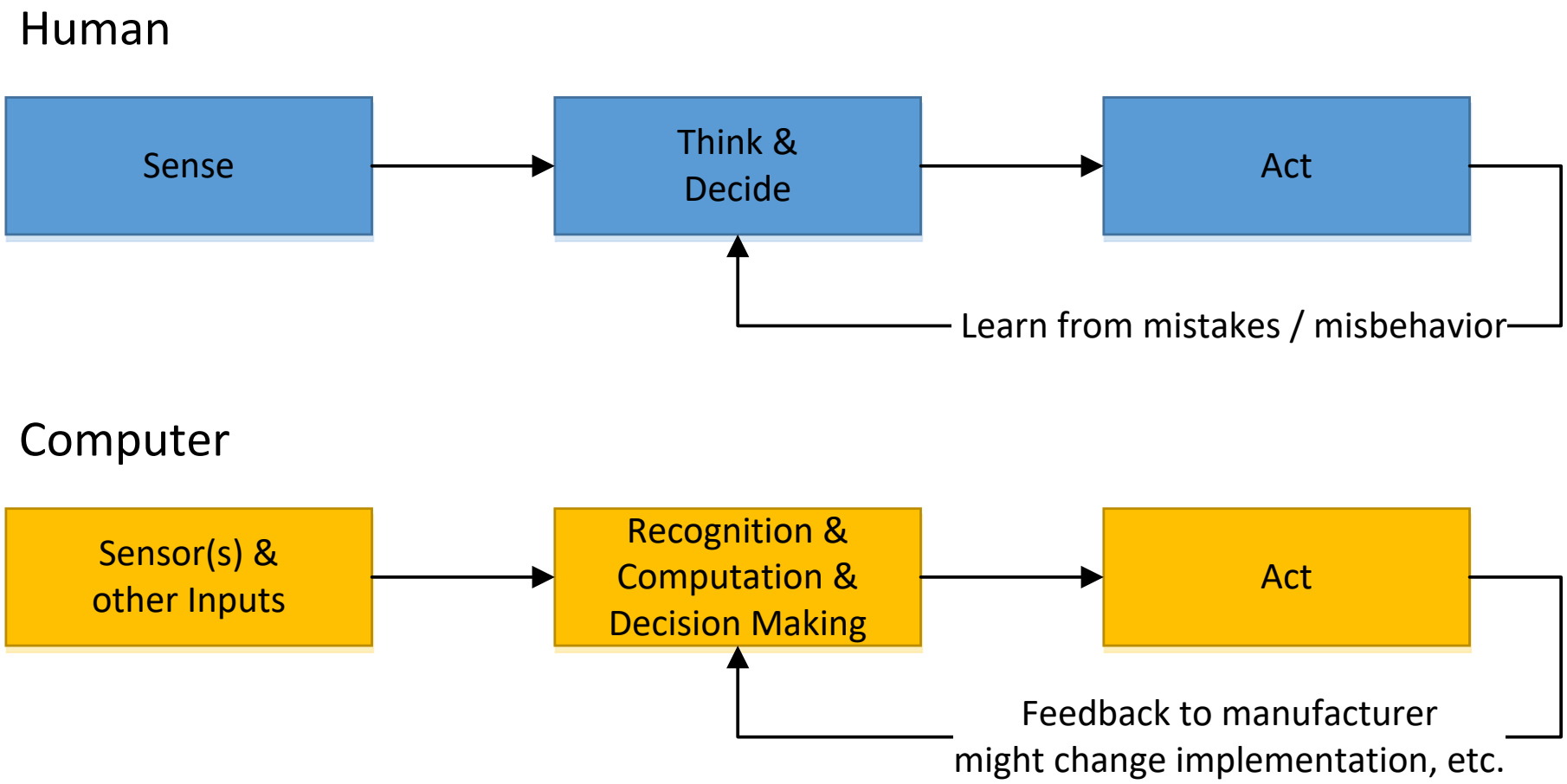
The problem is that for the question “whom to kill?” all answers are ethically questionable and perceived as bad or wrong.

There is no correct answer to the Trolley Problem and therefore it is not the right kind of problem as representative of real-life situations.

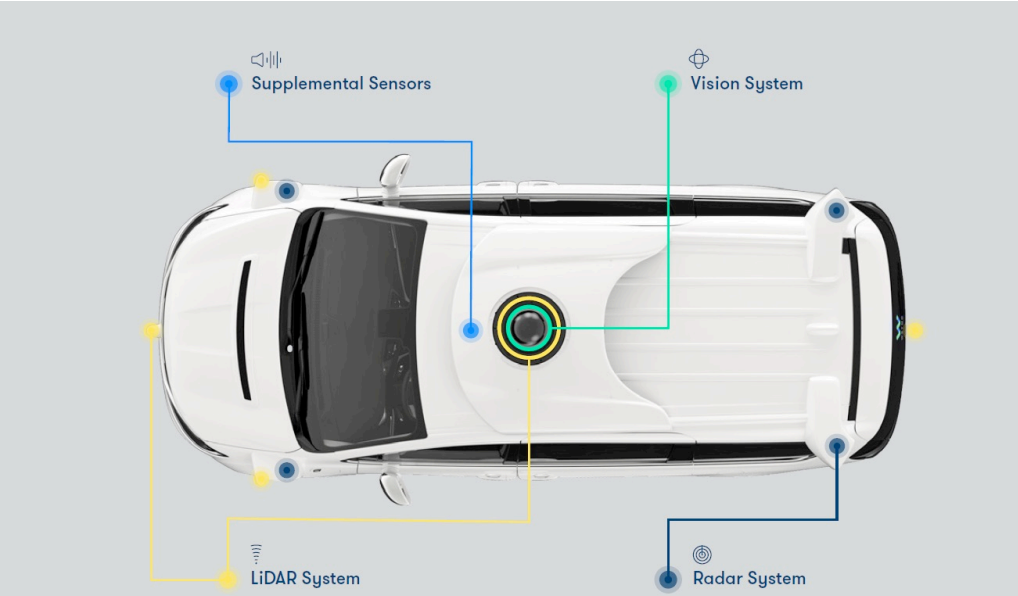
Moving the challenge from hypothetical decision-making to the real-world engineering problem:

THE REAL-WORLD DESIGN/ ENGINEERING PROBLEM IS NOT WHOM TO KILL BUT HOW NOT TO KILL!

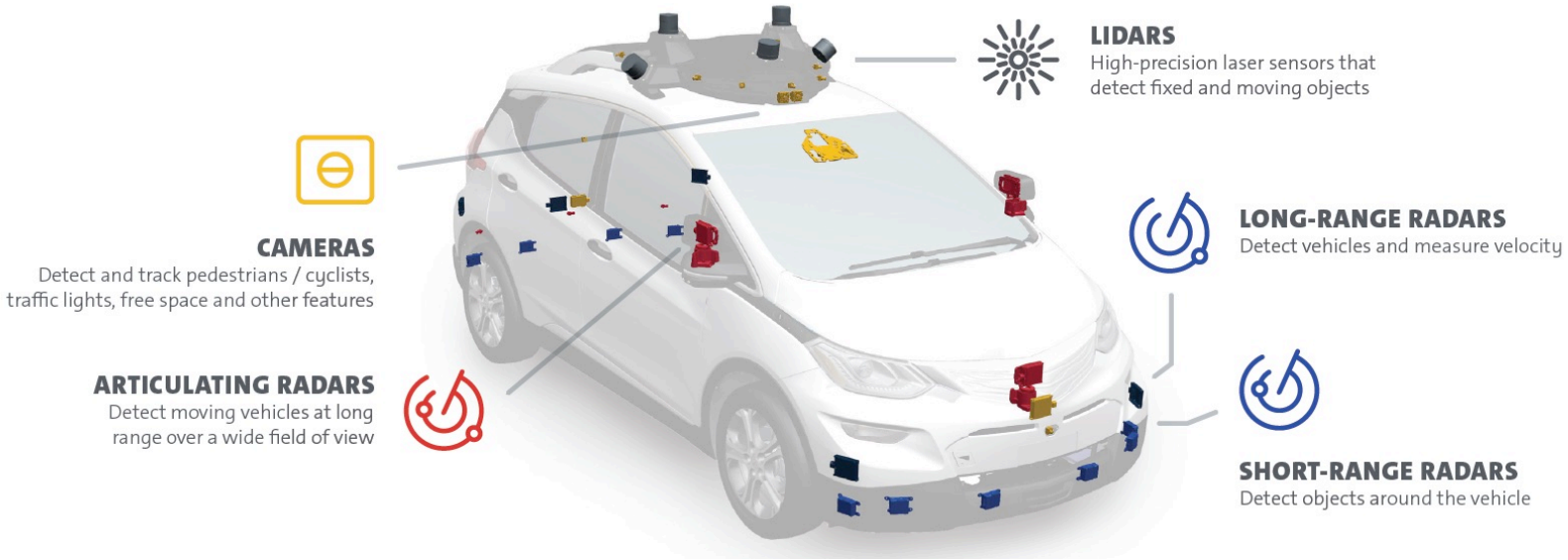
Human Decision-making Process vs. Self-Driving Car Decission-making



Technical Components



Picture Source: Waymo Safety Report 2017

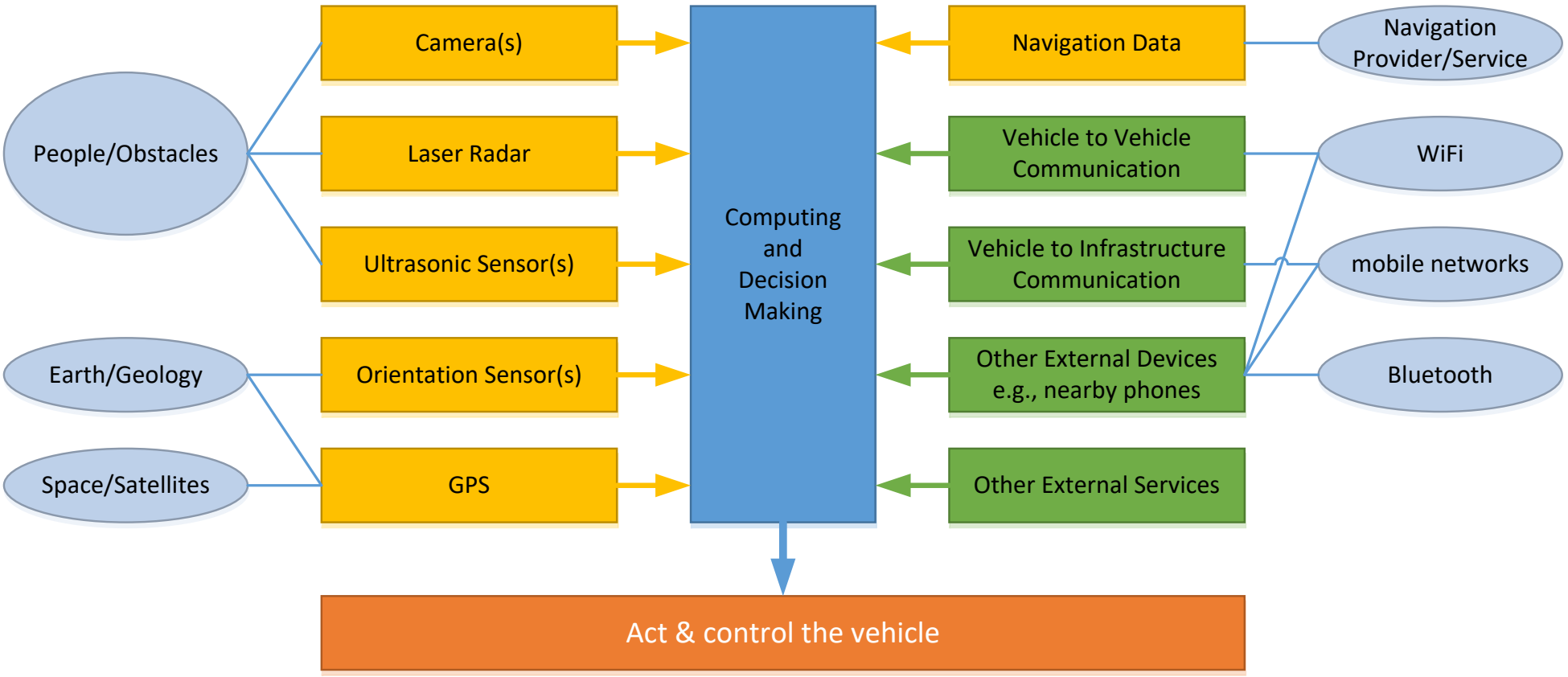


Picture Source: General Motors Safety Report 2018

Decision Making in Self-Driving Cars

- Decision making process involves sensors, external sources of information, networks, hardware, software, etc.
- Environmental influences, such as weather conditions (rain, bright sun, storm, …) affect sensors
- Complex input from sensors has to be filtered and only represents an abstraction of the real world.
- Behaviour of the car on the road (as a consequence of the decision made) is also dependent on the conditions on the road.

Abstract Decision Making Process



This is an outline of what a decision making process might include.
It is based on a literature review and official press releases (Tesla, Google, GM).

The “trolley problem” is built on assumptions that are neither technically nor ethically justifiable and can be summarized as the

INTRINSIC UNFAIRNESS OF THE TROLLEY PROBLEM

Intrinsic Unfairness of Trolley Problem

1. The intrinsic unfairness of the trolley problem comes from the assumption that lives of different people have different values.

It also represents a real-world problem in a distorted way.

2. It assumes deterministic processes and neglects uncertainties always connected to real world scenarios
3. It takes an extremely unlikely scenario as something relevant for the debate about the ethical aspects of AV

Common Assumptions of the Trolley Problem

Self-driving cars will …

- know the personal information such as profession, age, gender, criminal history, and more, of all persons near the car
- know which people might be involved, always detect the correct number of people around the car (i.e. exactly determine group sizes).
- correctly identify non-living objects and living beings.
- predict exact impact of the decision: which decision may kill whom and damage what? (That is, in this idealized thought experiment all real-world uncertainties are neglected)

The Real-World Problem

- Sensors currently do not have such quality as to enable perfect perception of the environment
- Self-Driving cars do not have access to personal information of all possible people around the roads
- Neither infrastructure nor connected systems can cover 100% of all streets
- It is not possible to predict who will be killed when hit by a car, as many factors play in

Implications

Even if it would be possible to implement (and it is not possible even in principle), in a large scale this would mean

- implementing George Orwell's 1984, knowing everything about everyone and
- being able to devise a decision making system that would compare people with different characteristics in order to decide who is most suitable to be killed.

CHALLENGES FOR SOFTWARE FAIRNESS

Fairness in Complex Systems

Different components can contribute to an “unfair” behavior and features.

This requires an ethical analysis on multiple levels and stages of the development and design process of technology.

In the following we present three examples to outline challenges for software fairness.

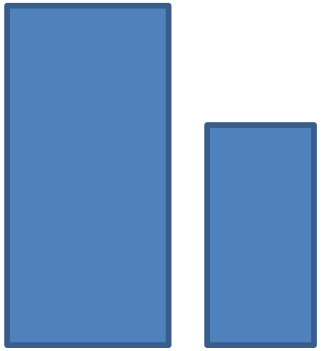
Example 1

SENSORS

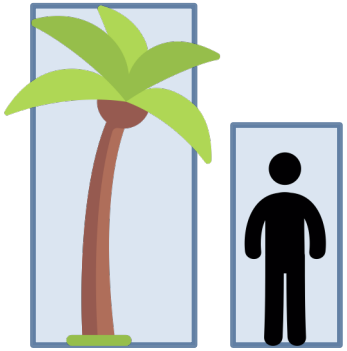
What a Self-Driving Car “sees” …

Sensors & Recognition

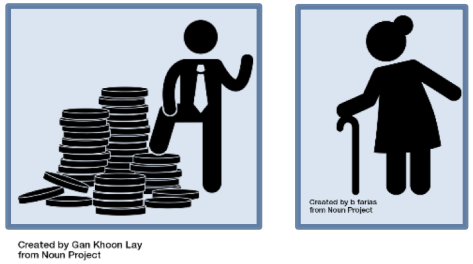
- Aim to detect objects (cars, buildings, etc.) and living beings (cyclists, pedestrians, etc.)
- Different stages of recognition:



Different Sized Objects in different states, e.g., moving or not moving



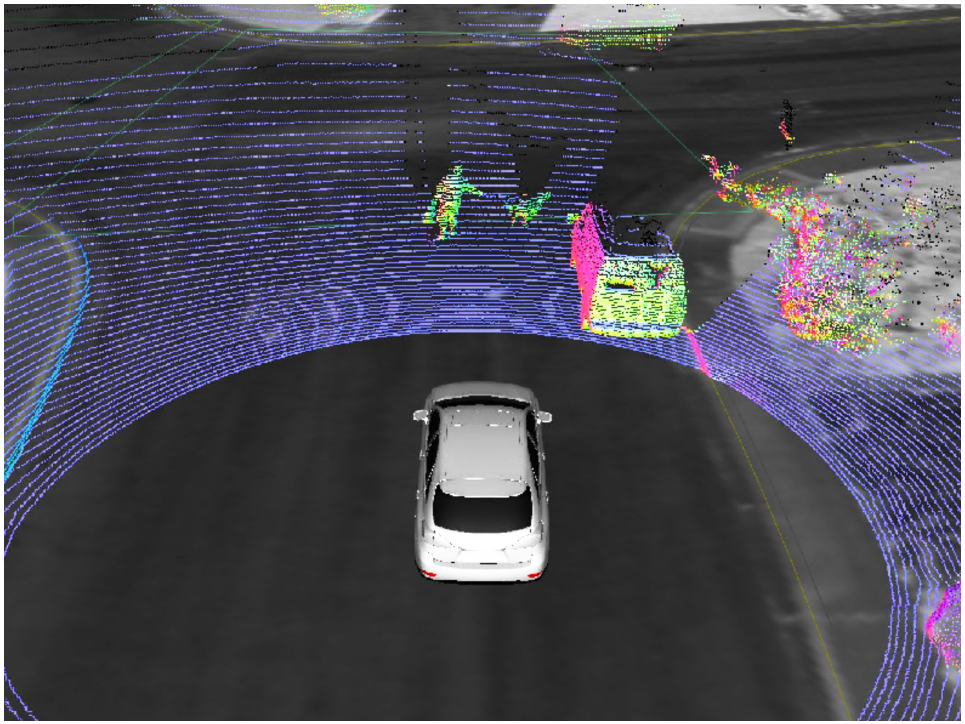
Objects vs Person(s)



„Everything“
If so, how to protect privacy?

Sensors & Recognition

"By analyzing photos of pedestrians, for example, a neural network can learn to identify a pedestrian"



Pedestrian, dog, and parked vehicle, as seen by lidar on a Google self-driving vehicle.

Sensors & Recognition

- Can the set of photos be biased?
  (e.g. Google Photos app labeled black people as 'gorillas'
  https://eu.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/)

- Are all people (regardless different clothes/postures, people in wheelchair etc.) detected with equal probability?

- If the set of photos is representative of one region, does it mean that people from the outside of the region become less likely to be detected correctly?

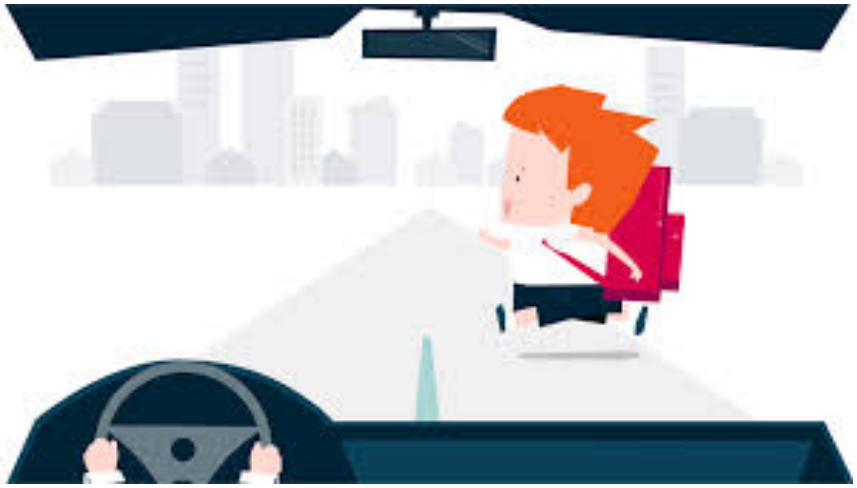
Example 2

EXTERNAL POSITIONING SYSTEMS

External Positioning Systems

"Schutzranzen", a backpack for pupils
Reports position of a backpack to nearby car drivers

Can be seen as analogue of using the position of mobile phones as potential input for decision making algorithms.



https://schutzranzen.com/

External Positioning Systems

- Can the position data be considered as an input for the decision making of self-driving cars?
  - If yes, some phones might have better GPS, positioning sensors or a faster internet connection. Is that contrary to the principle that “all humans are equally worth”?
- What about people who don’t have a mobile phone with them, or more likely have an empty battery?

Example 3

EXTERNAL SERVICES

External Services

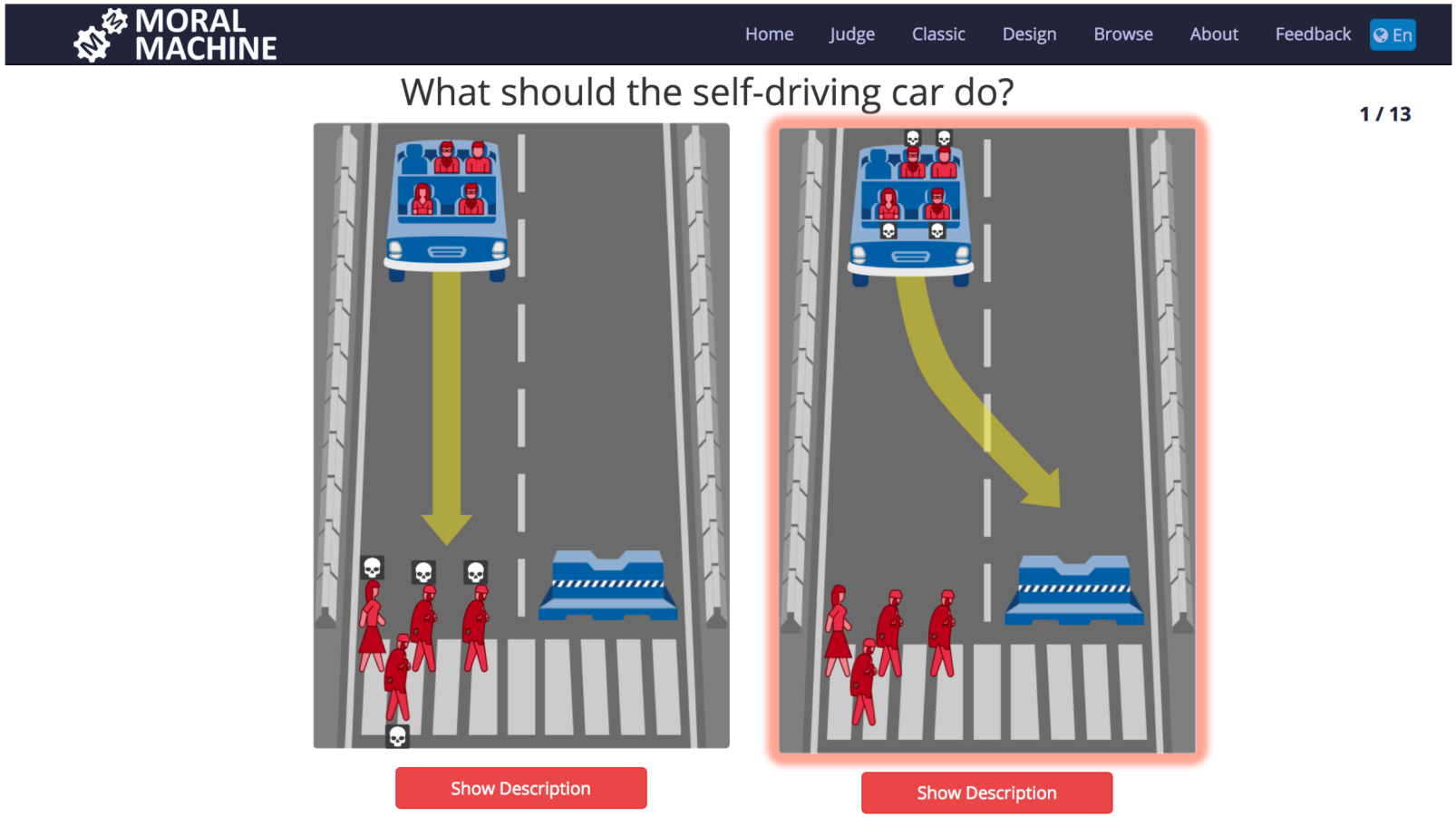
Up-to-date map data used to navigate the self-driving car can be provided by external services.

- Can external services change the behavior of the car in some way?
  - E.g., would it be possible for a map service to redirect or guide the car through a certain region that has more shops or advertisements than other regions?
  - I.e. could the route be biased?

http://moralmachine.mit.edu/

TROLLEY PROBLEM EXPERIMENT

MIT “MORAL MACHINE”



E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan (2018). The Moral Machine experiment. Nature, 563, pages 59–64. DOI: 10.1038/s41586-018-0637-6

https://www.media.mit.edu/publications/the-moral-machine-experiment/

Background

- Concern about how machines will make moral decisions
- Search for the ethical principles that should guide machine behavior
- An online experimental platform designed to explore the moral dilemmas faced by autonomous vehicles

This platform gathered 40 million decisions in ten languages from millions of people in 233 countries and territories.

https://www.media.mit.edu/publications/the-moral-machine-experiment/

The article

1. Summarizes global moral preferences.
2. Documents individual variations in preferences, based on respondents' demographics.
3. Reports cross-cultural ethical variation, and uncover three major clusters of countries.
4. Shows that these differences correlate with modern institutions and deep cultural traits.
5. Discusses how these preferences can contribute to developing global, socially acceptable principles for machine ethics.

E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan (2018). The Moral Machine experiment. Nature, 563, pages 59–64. DOI: 10.1038/s41586-018-0637-6

The historical roots of the fascination: Similar problems gained popularity in silent movies (Buster Keaton)

What Moral Machine Experiment does and does not show

It shows:

- Willingness of human participants to decide about the life and death of other people (not unlike Milgram experiment*)
- Local and cultural differences in those choices

It does not show:

- How autonomous car would act in similar situations given its own hardware and software

It assumes, unjustified:

- that car decision making will be similar to human decision making
- a very improbable situation as if it would be the central problem. None of the accidents or incidents that happened with AVs up to now were of the Trolley Problem type. And the same can be said of the accidents with human-driven cars.

*https://en.wikipedia.org/wiki/Milgram_experiment

List of self-driving car fatalities

Level 3 fatalities [edit]

A Level 3 autonomous driving system would occasionally expect a driver to take over control.

List of known autonomous car fatalities (occurring while autonomous-system acknowledged to have been engaged)

Date	Incident no.	Country	City	State/county /province	No. of fatalities	System manufacturer	Vehicle Type	Distance driven by the system at time of incident	Notes
18 March 2018	3	United States of America (USA)	Tempe	Arizona	1	Uber	'Refitted Volvo' [10]	—	Pedestrian fatality.[11]

Level 2 fatalities [edit]

Level 2 is considered automated driving, but not autonomous driving. A Level 2 driving system expects a driver to be fully aware at any time of the driving and traffic situation and be able to take over any moment.

List of known automated driving system car fatalities (occurring while automated driving-system acknowledged to have been engaged)

Date	Incident no.	Country	City	State/county /province	No. of fatalities	System manufacturer	Vehicle Type	Distance driven by the system at time of incident	Notes
20 January 2016	1	China	Handan	Hebei	1	Tesla (Autopilot)	Model S [12]	—	Driver fatality. [13][14]
7 May 2016	2	United States of America (USA)	Williston	Florida	1	Tesla (Autopilot)	Model S[10]	130,000,000 mi 210,000,000 km[15][16]	Driver fatality. [17][18]
23 March 2018	4	United States of America (USA)	Mountain View	California	1	Tesla (Autopilot)	Model X[10]	—	Driver fatality.[19]

https://en.wikipedia.org/wiki/List_of_self-driving_car_fatalities

Ethical Aspects of

REAL-WORLD TECHNICAL CHALLENGES WITH ETHICAL CONSEQUENCES

Safety

- How can we test self-driving cars?
  - How many tests are sufficient?

- Real world vs Abstract World
  - Training of Neural Networks

Security

- Attacks towards car systems and sensors
  - E.g., remote manipulation of LIDAR via laser beam
- System & Security Updates
- Do we need a Black Box like in Aircrafts?

Privacy

- What data should the car have access to?
  - and how will the data be used?
- What data is collected?
  - and who will have access to that data?

Trust

- How trustworthy are data sources?
  - E.g., GPS, map data, external services
  - Trust between self-driving car and services
- How trustworthy is the self-driving car?
  - E.g., trust between user and car

Transparency

- Multi-disciplinary challenge to ensure transparency, while respecting intellectual property rights, corporate secrets, security concerns, etc.
  - How much should be disclosed, and disclosed to whom?

Reliability

- What do we have to rely on?
  - What if sensor(s) fail?
  - What if networks fail?

- Redundancy for everything?

Responsibility and Accountability

- Who is responsible and for what?
- Who is accountable and for what?

- Designers
- Developers
- Car manufacturers
- Safety inspectorate
- Governmental institutions
- International regulatory bodies (European, etc.)

Quality Assurance

- Lifetime of components
- Maintenance
- Ethics-aware decision making in all processes will help to make ethically justified decisions.

Ethical Aspects of

SOCIAL CHALLENGES WITH ETHICAL CONSEQUENCES

Stakeholders / General Public Interests

- Loss of Jobs (e.g., Cabs/Taxi/Truck driver)
- Humans in the loop
  - the choice to interfere with the decision making
- Impact to Society

Freedom of Movement

- Freedom of movement
  - Will the car go, where I want it to go?
  - Implementation of Restrictions

- Route to Destination
  - Will the passenger define the route, or will it be determined by the system?
  - Road trips?

  (Perhaps we can ask similar for GPS)

Current State of

DESIGNING ETHICAL GUIDELINES FOR SELF-DRIVING CARS

THE FIRST ETHICAL GUIDELINES

German Ethics Commission on Automated Driving set up by Federal Minister Alexander Dobrindt, where the body of experts, headed by Professor Udo Di Fabio, a former Federal Constitutional Court judge, has developed guidelines for the programming of automated driving systems.
The Ethics Commission's report comprises 20 propositions. The key elements are:

- Automated and connected driving is an ethical imperative if the systems cause fewer accidents than human drivers (positive balance of risk).
- Damage to property must take precedence over personal injury. In hazardous situations, the protection of human life must always have top priority.
- In the event of unavoidable accident situations, any distinction between individuals based on personal features (age, gender, physical or mental constitution) is impermissible.
- In every driving situation, it must be clearly regulated and apparent who is responsible for the driving task: the human or the computer.
- It must be documented and stored who is driving (to resolve possible issues of liability, among other things).
- Drivers must always be able to decide themselves whether their vehicle data are to be forwarded and used (data sovereignty)

https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile

ANTICIPATORY GOVERNANCE: "LEARNING BY EXPERIENCE"

"Learning by experience" and "proven in use" concepts

"Learning by experience" (recording data from autonomous cars) presupposes a functioning socio-technological system that provides strong coupling among legislation, guidelines, standards and use, and promptly adapts to lessons learned.

H. Schäbe and J. Braband. Basic requirements for proven-in-use arguments. CoRR, abs/1511.01839, 2015.

Challenges

- Keeping legislation up-to-date with current level of automated driving, and emergence of self-driving cars
- Creating and defining global legislation frameworks for the implementation of interoperable and development of increasingly automated vehicles
- Defining the guidelines that will be adopted by society for building self-driving cars
- Including ethical guidelines in design and development processes

From: Holstein, Dodig-Crnkovic, Pellizzione: Ethical and Social Aspects of Self-Driving Cars, https://arxiv.org/abs/1802.04103

Recommendations

- Car producers supporting and collaborating with legislators in their task to keep up-to-date with the current level of automated driving
- Legislative support and contribution to global frameworks to ensure a smooth enrollment of the emerging technology
- Include ethics in the overall process of design, development and implementation of self-driving cars. Ensure Ethics training for involved engineers
- Establish and maintain a functioning socio-technological system in addition to functional safety standards.

CONCLUSIONS

Conclusions

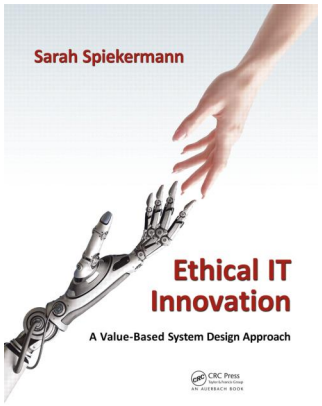
- We need to stop discussing unsolvable idealized ethical dilemmas that obfuscate true ethical challenges.
- Discuss the real-world ethical challenges surrounding autonomous and driverless vehicles.
- Define what is technically possible and simultaneously ethically justifiable.
- Create transparency in the whole system, starting with explainable algorithms to support evaluations by independent organisations and/or experts.

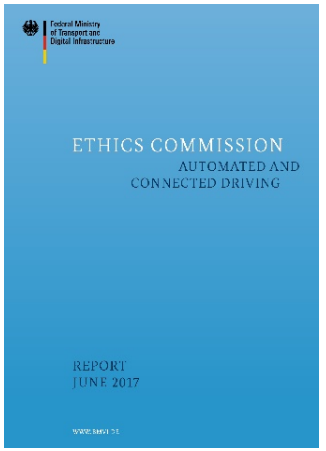
Conclusions

- There are many challenges to be solved for self-driving cars. Software Fairness offers a great opportunity to establish frameworks and tools to assure the “fairness” of software.
- Good measurements, big data ≠ Fairness
  - They must be put into an adequate framework, when it comes to:
    - the problem they address,
    - the questions they answer and
    - the underlying value system they assume.

Conclusions

Consider already existing reports and normative documents, e.g.,





- Ethical IT Innovation:
  A Value-Based System Design Approach
  (by Sarah Spiekermann)

- Ethics Commission: Automated and connected driving
  (Report by Federal Ministry of Transport and Digital Infrastructure of Germany [BMVI])

BMVI = Bundesministerium für Verkehr und digitale Infrastruktur

Future Work

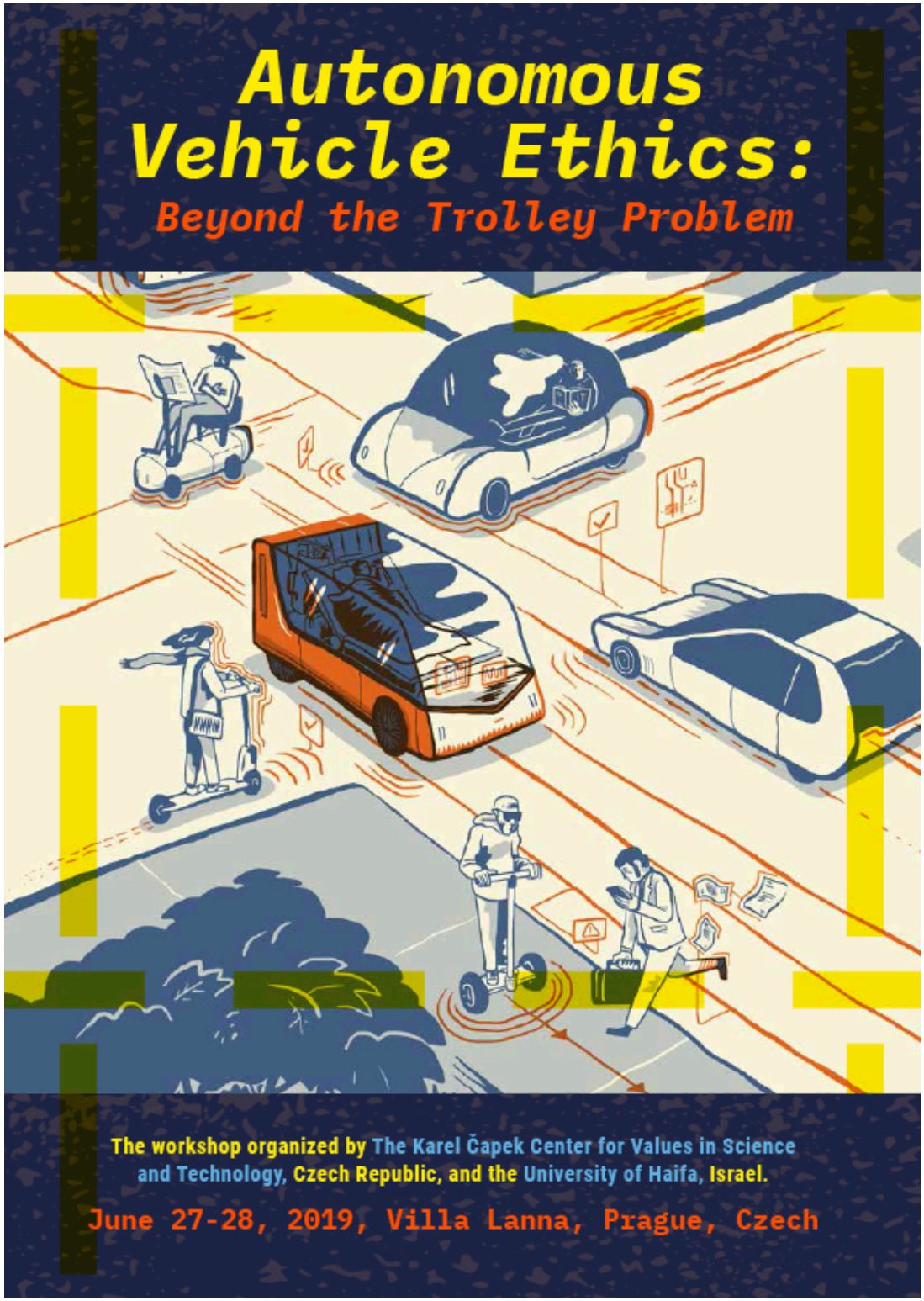
In order to include real-world ethics in the design and engineering, we propose to consider

- Ethicality as non-functional property
  - Ethicality: the state, quality, or manner of being ethical.

- Software Fairness as attribute or property of Ethicality

For further reading

REFERENCES



The Karel Čapek Center
for Values in Science
and Technology

https://www.cevast.org/

Declaration of Amsterdam

14 April 2016 EU member states endorsed the Declaration of Amsterdam[1] that addresses legislation frameworks, use of data, liability, exchange of knowledge and cross-border testing for the emerging technology.

It prepares a European framework[2] (based on report by **Ethics Commission on Automated and Connected Driving**) for the implementation of interoperable connected and automated vehicles by 2019.

[1] https://www.government.nl/documents/leaflets/2017/05/18/on-our-way-towards-connected-and-automated-driving-in-europe
[2] https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile

Ethics & Law Aspects

Autonomous Vehicles Ethics & Law: Towards an Overlapping Consensus

https://www.academia.edu/29332066/Autonomous_Vehicles_Ethics_and_Law_Towards_an_Overlapping_Consensus

Patrick Lin: Why Ethics Matters for Autonomous Cars.

In: Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte

https://www.springerprofessional.de/en/why-ethics-matters-for-autonomous-cars/4397684

Ethically Aligned Design

A Vision for Prioritizing Human Well-being With Autonomous and Intelligent Systems

https://ethicsinaction.ieee.org/

Embedding Values into Autonomous Intelligent Systems - The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

https://standards.ieee.org/develop/indconn/ec/ead_embedding_values.pdf

An example of ethical guidelines thinking one step further
is described in the book:

Sarah Spiekermann. Ethical IT Innovation: A Value-Based System Design Approach. Taylor & Francis, 2015.

Policy Concerning Automated Vehicles (US DOT)

"DOT/NHTSA Policy statement concerning Automated Vehicles" 2016 update to "Preliminary statement of policy concerning automated vehicles".

Technical report, National Highway Traffic Safety Administration (NHTSA).

http://www.nhtsa. gov/staticfiles/rulemaking/pdf/Autonomous-Vehicles-Policy-Update-2016.pdf