ISSN 1831-9424

European Commission

JRC TECHNICAL REPORT

# Toward explainable, robust and fair AI in automated and autonomous vehicles

*Challenges and opportunities for safety and security*

Kriston A., Hamon R., Fernández Llorca D., Junklewitz H., Sánchez I., Gómez E.

2023


Joint Research Centre

EUR 31305 EN

**Contact Information**
Name: Akos Kriston
Address: European Commission, Joint Research Centre (JRC) Via E. Fermi 2479, 21047 Ispra (VA) - Italy
Email: akos.kriston@ec.europa.eu
Tel.: +39 033278-5547

**EU Science Hub**
https://joint-research-centre.ec.europa.eu

JRC130621

EUR 31305 EN

| | | | | |
|---|---|---|---|---|
| PDF | ISBN 978-92-76-58978-5 | ISSN 1831-9424 | doi:10.2760/95650 | KJ-NA-31-305-EN-N |

Luxembourg: Publications Office of the European Union, 2023



How to cite this report: Kriston, A., Hamon, R., Fernández Llorca, D., Junklewitz, H., Sánchez, I. and Gómez, E., *Toward explainable, robust and fair AI in automated and autonomous vehicles*, Publications Office of the European Union, Luxembourg, 2023, doi:10.2760/95650, JRC130621.

## Contents

## Abstract

In March 2022 the JRC (Units B.6, C.4, E.3) organized an Exploratory Workshop entitled "Toward explainable, robust, and fair AI in automated and autonomous vehicles", bringing together experts in fields such as Trustworthy AI, autonomous driving, and vehicle testing. This report summarizes the steps that followed the organization of the workshop, including the definition of the scientific objectives, the list of invited presenters and participants, and the conditions under which the workshop took place.

The report also presents the main findings of each talk that occurred during the workshop and an analysis of the discussions that occurred during collaborative working sessions. Topics of interest included, among others, current regulations and standards regarding automated and autonomous road vehicles and analysis of their limitations; explainability of artificial intelligence ; accuracy, robustness, security, and fairness of AI systems.

These insights are used to provide concluding remarks on the outlook of the Workshop, in particular how the findings of the Workshop can help to promote further research within and outside of the JRC on this topic, with the goal of making safer transport through innovative ecosystems and effective regulations. We identified gaps in the scientific literature on the relationship between AI and safety of Automated and Autonomous Vehicles (A&AVs) such as:

- establishment of reasoning vocabulary for acceptable factual and/or counterfactual interpretations,
- certification readiness matrix must be developed for each cyber scenario for different adversarial attacks and for naturally occurring perturbations,
- behavioural models are missing for motion prediction of different social agents and tests with standardized dummies lack the features of different social groups,
- currently there are not enough data to assess the fairness of A&AV vehicles and how fairness or bias influences safety.

In our next report, we will focus on the above points by involving experts of the fields.

## Acknowledgements

The authors acknowledge the support and financial contribution of the JRC Exploratory Research Activity scheme and the help of JRC colleagues in Directorate A. The authors gratefully acknowledge Sara Andre of JRC.C.7 for the impactful design and Carine Nieuweling and Darren McGarry for their advice on communication.

***Authors***
Kriston, Akos
Hamon, Ronan
Fernández Llorca, David
Junklewitz, Henrik
Sánchez, Ignacio
Gómez, Emilia

## Executive summary

This report summarizes the discussions that took place during the JRC Exploratory Workshop organized in March 2022 entitled "Toward explainable, robust and fair AI in automated and autonomous vehicles". The aim of the workshop was to bring together experts to present and discuss the latest advances in testing the safety and security of Automated and Autonomous Vehicles (A&AV), in particular connected to the adoption of Artificial Intelligence (AI) in vehicles. This workshop is part of a larger project whose purpose is to gain insight into the future directions of testing practices in the automotive sector from a regulatory point of view, in a context of increased digitalization of the transport sector. The scientific objectives of the Workshop have been defined through a series of research questions grouped into three main topics:

— Explainability and testing of AI systems in vehicles;

— Cybersecurity of AI systems;

— Fairness of AI systems.

To comprehend the complexity and the multi-disciplinarity of the topic, the organization of the workshop has been shared between three units of the JRC:

1. Sustainable transport (C4) works on all aspects of the road transport system, including testing automated and autonomous vehicles.
2. Cyber & Digital Citizens' Security (E3) is concerned about risk mitigation, cybersecurity, cybercrime, data protection, and privacy;
3. Digital Economy (B6) studies the social and economic impacts of Artificial Intelligence (AI), data and digital platforms, advancing research on methodologies to ensure trustworthy AI.

All three units have developed expertise on specific facets of the interplay between automated vehicles and trustworthy artificial intelligence, presenting in a joint effort a comprehensive and unique selection of relevant research topics such as the robustness, security, fairness, and explainability of AI systems, and their testing in field conditions.

The workshop included 14 talks, during which the following topics, among others, have been discussed:

— current regulations and standards regarding automated and autonomous road vehicles and analysis of their limitations;

— safety issues that can occur in real environment;

— the explainability of artificial intelligence and its use to gain insight into the behaviours of A&AV, the assessment of the trustworthiness of autonomous and automated vehicles, in particular regarding the accuracy, robustness, security, and fairness of AI systems; the review of ex-post explanations and concrete examples of accidents of A&AV and their possible causes;

— broad considerations on the influence of the environment on A&AVs' decision-making processes.

Discussions were held each day involving the use of collaborative tools on-line to gather and structure the information consistently.

Among the main findings of the workshop, the need for additional research to understand how individual AI components can be integrated into the broader A&AVs testing framework has been particularly discussed, with current limitations of vehicles to demonstrate the absence of risks in terms of accuracy, robustness, cybersecurity or fairness. Experts concluded that edge cases of an A&AV can be very different from the edge cases of human driver, therefore testing of challenging scenarios for human maybe misleading. Explainability may help to very if a particular vehicle passed an edge case because it recognized the scenario and not because an artificial test condition has changed. In a perfect world traffic rules are a means by which road safety is achieved, but non-compliance is sometimes necessary to achieve greater road safety (i.e. the ethics of AV driving behaviour and whether they will deviate from the rules of the road to maximize safety). However, current regulation cannot have this flexibility.

These conclusions can help promote further research within and outside of the JRC on this topic, with the goal of making safer transport through innovative ecosystems and effective regulations. They will also provide fruitful information for the following steps of the project and, in particular, the appointment of a group of experts to draft a comprehensive report on this topic at the attention of regulatory bodies.

# 1 Introduction

This JRC Exploratory Workshop was dedicated to the safety and security of Automated and Autonomous Vehicles (A&AV), and aimed to bring together leading scientists and engineers to explore and discuss state-of-the-art research on the accuracy, robustness, fairness and explainability of Artificial Intelligence (AI) and Machine Learning (ML) and testing of modern vehicles.

Currently, A&AVs are tested in a black-box approach, based on limited traffic and cybersecurity scenarios. The behaviour of AI-ML systems is studied through descriptive statistics of kinematics and/or interaction with other road users and the infrastructure, using mainly knowledge of the mechanical engineering domain. However, unlimited variations of traffic situations exist and their consideration in testing is out-of-reach.

So far, no scientifically sound methodologies have been developed to audit the decisions made by the AI and ML systems during driving, especially in safety critical scenarios. In addition to functional and operational safety, other challenges related to the uptake of AI and ML in automated and autonomous driving have emerged in recent years, such as the assessment of the cybersecurity, explainability and fairness of systems, in line with the recent initiative from the European Commission to promote Trustworthy AI in high-risk systems.

An innovative testing and explanatory framework of AI and ML systems embedded in A&AV requires a deep and improved understanding of the interplay of AI techniques and their limitations, cybersecurity, ethical principles, and road safety regulations. A promising approach to consider for the evolution of testing practices relies on techniques and methodologies developed in the field of Explainable AI (xAI) to analyze and understand the output of AI-ML components. In the context of A&AVs, these approaches may help detect and mitigate false decisions and attacks on automated functions while providing a better understanding of biases that arise with the use of large sets of data, e.g. toward minority groups, and their potential impact on the safety in A&AVs.

To explore these questions, JRC organized a multi-disciplinary Exploratory Workshop dedicated to testing approaches of A&AVs, with the objective to provide an overview of the challenges linked to the use of advanced AI systems in vehicles, and explore ways to address them.

# 2 Research Questions

The main research questions identified prior to the workshop were grouped into three blocks related to the main issues to be addressed, i.e. explainability and evidence, cybersecurity and equity. Experts were selected based on their expertise in these topics, to contribute to the identification of possible solutions or approaches, or to identify gaps where clear answers were not available. The main research questions were as follows.

## Explainability and testing

- What are the current testing methods for AI-ML components in automotive environment?
- How can we test the AI-ML components in terms of safety in an automotive environment?
- How to define and quantify the robustness and accuracy, of an A&AV's AI-ML component?
- Are the behaviours of AI-ML components of an A&AV reproducible and repeatable in controlled environments and in the wild?
- How is it possible to explain the decisions made by AI-ML components in A&AV from a software engineering, vehicle safety testing and accident investigation perspectives?

## Cybersecurity

- What are the cybersecurity threats and vulnerabilities associated with AI component in A&AVs?
- What are the limitations of current vehicle testing methods for evaluating AI cybersecurity risks?
- How can we measure the resilience of vehicle systems against cyberthreats targeting AI components?
- How can we handle the security vulnerabilities discovered in the AI components of automated and autonomous vehicles?

— What are the AI-related cybersecurity challenges connected to the supply chain of A&AVs?

— What is the state of cybersecurity standards for AI in automated driving and what gaps would need to be addressed?

### Fairness

— What elements would be affected, and how would we consider fairness as a requirement in test procedures?

— How to detect biases in automated decisions and assess their impact in terms of fairness and robustness?

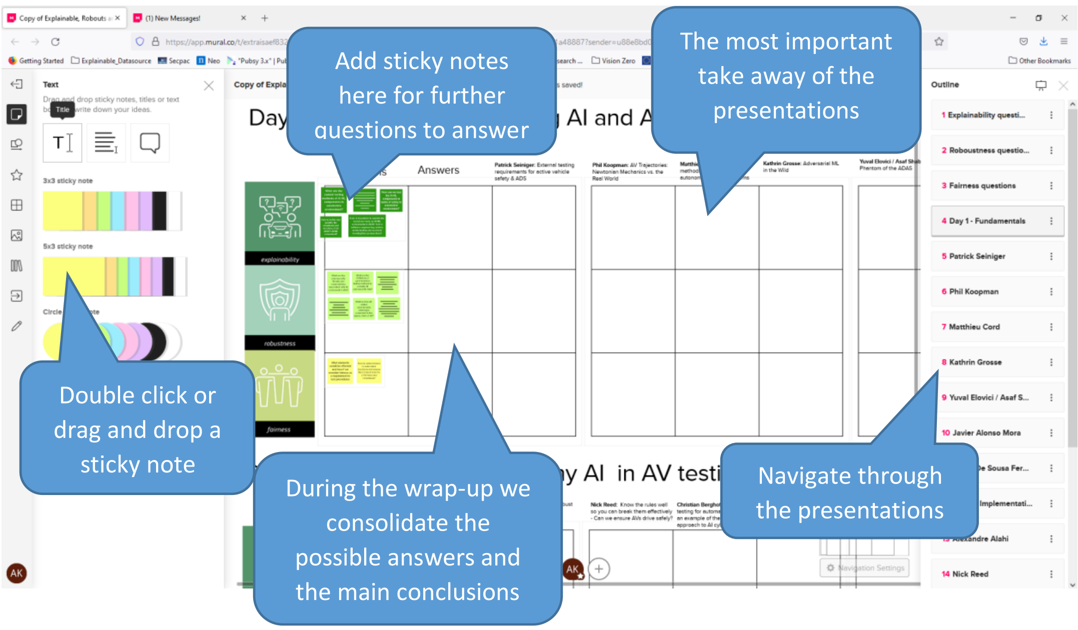
— Is it possible to guarantee the same level of safety for all types of road users and how?

## 3 Methodology, participants and agenda

The Exploratory Workshop took place as a virtual event using the Webex platform on 29-30 March 2022.

Participants, experts and presenters were instructed to use the chat to ask questions and concerns, and to raise their hands (with the raised hand icon) when they wanted to intervene directly. They were also encouraged to participate in the collaborative work set up on the Mural online platform to collect questions and answers in a structured way for each topic and for each presentation, adding virtual sticky notes before, during or after the workshop. A workplace was created by the organizers, with one row per topic (i.e., explainability, robustness and fairness) and one column per presentation (see Figure 1). Different colours were used depending on the type of commentary, including open questions, answers, starting points and key conclusions.

**Figure 1:** Workplace created in Mural to collectively gather information from participants




*source:* JRC analysis

The list of participants, together with their role in the workshop and their affiliations, is provided in Table 1 (they are presented alphabetically according to surname). Only invited speakers and organizers are listed, leaving out the rest of the audience.

The agendas for Day 1 and Day 2 are shown in Tables 2 and 3, respectively. As can be observed, the duration of each presentation was 30 minutes including time for questions and discussion. In addition, each day was planned with two breaks and a final recap session.

Experts and presenters were provided with a session briefing document, including the most relevant supporting materials for the workshop. The working materials were the following:

**Table 1:** List of participants.

| Participant | Role | Affiliation |
|---|---|---|
| Javier Alonso Mora | Invited Speaker | TU Delft (The Netherlands) |
| Alexandre Alahi | Invited Speaker | EPFL (Switzerland) |
| Ensar Becic | Invited Speaker | NTSB (USA) |
| Christian Berghoff | Invited Speaker | BSI (Germany) |
| Matthieu Cord | Invited Speaker | Valeo (France) |
| Rafaël De Sousa Fernandes | Invited Speaker | UTAC (France) |
| Yuval Elovici | Invited Speaker | Ben-Gurion University (Israel) |
| David Fernández Llorca | Organizer | JRC (Spain) |
| Emilia Gómez | Organizer | JRC (Spain) |
| Katrin Grosse | Invited Speaker | University of Cagliari (Italy) |
| Ronan Hamon | Organizer | JRC (Italy) |
| Henrik Junklewitz | Organizer | JRC (Italy) |
| Philip Koopman | Invited Speaker | Carnegie Mellon University (USA) |
| Akos Kriston | Organizer | JRC (Italy) |
| Lars Kunze | Invited Speaker | Oxford Robotics Institute (UK) |
| Nick Reed | Invited Speaker | Reed Mobility (UK) |
| Ignacio Sánchez | Organizer | JRC (Italy) |
| Patrick Seiniger | Invited Speaker | BASt (Germany) |
| Asaf Shabtai | Invited Speaker | Ben-Gurion University (Israel) |
| Jack Stilgo | Invited Speaker | University College London (UK) |
| Robert Swaim | Invited Speaker | HowItBroke (USA) |

— **The Future of Road Transport - Implications of automated, connected, low-carbon and shared mobility** (Alonso Raposo et al., 2019): this JRC report looks at some of the main enablers of the transformation of road transport, such as data governance, infrastructures, communication technologies and cybersecurity, and legislation. The paper discusses potential impacts on the economy, employment and skills, energy use and emissions, the sustainability of raw materials, democracy, privacy, and social fairness, as well as on the urban context.

— **Testing the Robustness of Commercial Lane Departure Warning Systems** (Re et al., 2021): this work presents a novel robustness assessment methodology and defines a robustness index determined from regulatory tests to analyze the real-world performance of lane departure warning (LDW) systems to bridge the gap between regulatory and real-world performance.

— **Fuzzy Surrogate Safety Metrics for real-time assessment of rear-end collision risk. A study based on Empirical Observations** (Mattas et al., 2020): this work discusses two fuzzy Surrogate Safety Metrics (SSMs) for rear-end collisions. The objective is to investigate its applicability for evaluating the real-time rear-end risk of collision of vehicles to support the operations of advanced driver assistance and automated vehicle functionalities (from driving assistance systems to fully automated vehicles).

— **Cybersecurity challenges in the uptake of Artificial Intelligence in Autonomous Driving** (Dede et al., 2021): this report by the JRC and the European Union Agency for Cybersecurity (ENISA) analyzes the cybersecurity risks related to the adoption of artificial intelligence (AI) in autonomous vehicles and provides recommendations to mitigate them. The report puts forward a set of challenges and recommendations to improve AI security in autonomous vehicles and mitigate these risks.

— **Trustworthy Autonomous Vehicles** (Fernández-Llorca and Gómez, 2021): this JRC report aims to advance toward a general framework on trustworthy AI for the specific domain of Autonomous Vehicles (AVs). The implementation and relevance of the assessment list established by the independent High Level Expert Group on Artificial Intelligence (AI HLEG) as a tool to translate the seven requirements that AI systems should meet in order to be trustworthy, defined in the Ethics Guidelines, are discussed in detail and contextualized for the field of AVs.

— **Ethics of Connected and Automated Vehicles: recommendations on road safety, privacy, fairness, explainability and responsibility** (European Commission, 2020): in 2019, the Commission formed an independent Expert Group to advise on ethical issues raised by driverless mobility. The group published this report with 20 recommendations covering dilemma situations, the creation of a culture of responsibility, and the promotion of data, algorithm and AI literacy through public participation.

— **European approach to AI** (European Commission, 2018): starting in March 2018 with the creation of the AI Expert Group and the European AI alliance, following the Coordinated Plan on AI, the Ethics Guidelines for Trustworthy AI, the white paper on AI and, more recently, up to 3 interrelated legal initiatives, the Commission aims to address the risks generated by specific uses of AI while maximizing its benefits by building an ecosystem of excellence and trust.

**Table 2:** Day 1 Agenda - March 29, 2022. Fundamentals of testing AI in AVs - Current situation and challenges.

| Time | Presenter | Title |
|---|---|---|
| 13:30 - 14:00 | JRC Organizers | Presentation of the Workshop. Moderator A. Kriston (JRC) |
| 14:00 - 14:30 | Patrick Seiniger, BASt, Germany | External testing requirements for active vehicle safety & ADS |
| 14:30 - 15:00 | Philip Koopman, Carnegie Mellon University, USA | AV Trajectories: Newtonian Mechanics vs. the Real World |
| 15:00 - 15:30 | Matthieu Cord, Valeo, France | Explainability methods for vision-based autonomous driving systems |
| 15:30 - 15:45 | Break | Moderator: R. Hamon (JRC) |
| 15:45 – 16:15 | Kathrin Grosse, University of Cagliari, Italy | Adversarial ML in the Wild |
| 16:15 – 16:45 | Yuval Elovici / Asaf Shabtai, Ben-Gurion University, Israel | Phantom of the ADAS: Securing advanced driver-assistance systems from split-second phantom attacks |
| 16:45 – 17:00 | Break | Moderator: E. Gómez (JRC) |
| 17:00 – 17:30 | Javier Alonso Mora, TU Delft, The Netherland | Safe Motion Planning among Decision-Making Agents |
| 17:30 – 18:00 | Rafaël De Sousa Fernandes, UTAC, France | PRISSMA project overview |
| 18:00 – 18:30 | All | Discussion session and Wrap-up |

**Table 3:** Day 2 Agenda - March 30, 2022. Implementing Trustworthy AI in AV testing.

| Time | Presenter | Title |
|---|---|---|
| 13:30 - 14:00 | JRC Organizers | Welcome and wrap-up from day 1. Moderator R. Hamon (JRC) |
| 14:00 - 14:30 | Alexandre Alahi, EPFL, Switzerland | Towards Robust Autonomous Vehicles |
| 14:30 - 15:00 | Nick Reed, Reed Mobility, UK | Know the rules well so you can break them effectively - Can we ensure AVs drive safely? |
| 15:00 - 15:30 | Christian Berghoff, BSI, Germany | Robustness testing for automated driving as an example of the BSI's approach to AI cybersecurity |
| 15:30 - 15:45 | Break | Moderator: D. Fernández Llorca (JRC) |
| 15:45 – 16:15 | Jack Stilgoe, University College London, UK | The actual ethics of AI for AVs: from autonomy to attachments |
| 16:15 – 16:45 | Lars Kunze, Oxford Robotics Institute, UK | Towards Explainable and Trustworthy Autonomous Systems |
| 16:45 – 17:00 | Break | Moderator: A. Kriston (JRC) |
| 17:00 – 17:30 | Robert Swaim, HowItBroke, USA | Man, Machine, or In Between: The Process of Investigations Into Automation |
| 17:30 – 18:00 | Ensar Becic, NTSB, USA | Safe path to vehicle automation: Crash investigation perspective |
| 18:00 – 18:30 | All | Discussion session and Wrap-up |

# 4 Day 1: Fundamentals of testing AI in AVs - Current situation and challenges

This session was planned with the objective of discussing the fundamentals of current AV testing in the context of increased autonomy. Short presentations with free discussions are planned on current practices in testing automated capabilities of AV and on how the increasing use of AI and ML techniques in vehicles brings new challenges. In relation to techniques used in current and future automated and autonomous vehicles, this session focuses on the following requirements: explainability, robustness (including accuracy and cybersecurity), and fairness. The session was finally adapted according to the accepted presentations.

## 4.1 Presentation of the Workshop

The presentation of the workshop is carried out to set the scene and provide useful information to experts and participants. After introducing the core team, the main rules of the day, and some general information regarding the Commission, the JRC and the different units involved (C4, E3 and B6), the concept and the agenda of the Workshop is presented. The main research questions to be addressed during the workshop are discussed, concerning the three main topics: explainability, robustness, and fairness. Finally, some specific details on the virtual collaboration tool (MURAL) are given to allow collecting information from all the participants, before, during and after the workshop.

## 4.2 External testing requirements for active vehicle safety & ADS

**Presenter**: Patrick Seiniger, BASt, Germany.

First, the question of what is active vehicle safety is addressed, including the following distinction: Active safety involves the avoidance of an accident (that is, before it occurs), while passive safety focuses on mitigating the consequences of an unavoidable accident. The origins of external requirements for active vehicle safety and automated driving systems (ADS) are discussed, including consumer protection and type approval requirements.

Legal requirements including in Technical Regulations are agreed by Contractual Parties (e.g., UN R 79: Steering Equipment, R130: Advanced Emergency Braking, etc.). At the European level, Regulations or Directives for Member States can refer to UN Regulations, e.g., Regulation 858/2018 on Passenger Car Type Approval. And in some cases, as the UN process is slow, the EU writes its own regulation (e.g., 347/2012 for AEBS).

Three different levels are distinguished for the test concepts. First, consumer ratings (e.g., NCAP and Euro NCAP) which usually focuses on a large grid of very specific test points with tight tolerances and ranks the vehicles which are tested on voluntary basis. Second, obligatory vehicle regulations (e.g. UNECE), which usually focus on precise single-test scenario (e.g. worst-case) with fixed testing conditions and higher tolerances for pass-fail criteria. Finally, new approaches are being developed. The concept was first proposed with heavy vehicle emissions with the idea of defining broader requirements with not too strictly specified tests, including semirandom test cases, on-road test, etc. In market surveillance, this approach may motivate manufacturers to develop robust systems and not just pass the regulation.

Some examples are described, including the negative feedback control system and some of its components (e.g., position measurement sensors and actuators). The targets (ISO19206) and the platforms for testing in proving grounds are described.

An open discussion concludes the presentations, addressing the current limitations of the test tools, and the possibility to randomize the tests and to include more realistic conditions.

## 4.3 AV Trajectories: Newtonian Mechanics vs. the Real World

**Presenter**: Philip Koopman, Carnegie Mellon University, University of Pennsylvania, USA.

In this talk, the limitations of regulatory testing are highlighted, focusing on the complexity of real world driving, including limits on trajectory control (e.g., vehicle capabilities, environmental conditions), as well as uncertainty about both vehicle conditions and environment.

A relatively “simple” example, such as the safe following distance, includes multiple factors to consider, such as road conditions, braking capacity, equipment condition, braking controls, aerodynamics, suspension, debris, etc.

Epistemic uncertainty is also considerably complex and includes brake wear and failures, tire pressure, brake condition, as well as braking capability for vehicle type, aftermarket upgrades, road surface of own and lead vehicles, etc.

A single (huge) Operational Design Domain (ODD) may not be sufficient to handle all this complexity. One possible approach is to break it up into smaller pieces (micro ODDs). Some examples that may provide further assurance of such an approach are included in ANSI/UL 4600, Sections 8.2 and 8.8.

Testing is based on assumptions about the environment and behaviours. An appropriate balance between permissiveness and safety is needed. Testing also pushes the uncertainty under certain assumptions. And finally, there will always be edge cases to consider. Edge cases for humans and AVs may be different.

## 4.4 Explainability methods for vision-based autonomous driving systems

**Presenter**: Matthieu Cord, Valeo, France.

This presentation is divided in three main parts. First, the explainability of vision-based self-driving cars is addressed. The concept of explainability has several facets, and the need for explainability is strong in driving, a safety-critical application. Gathering contributions from several research fields, namely computer vision, deep learning, autonomous driving, and explainable AI (X-AI), this presentation discusses definitions, context, and motivation to gain more interpretability and explainability from self-driving systems. It also briefly describes methods providing explanations to a black-box self-driving system in a post-hoc fashion and approaches that aim at building more interpretable self-driving systems by design. The remaining open challenges and potential future research directions are identified and examined.

Second, post-hoc explainability by steering counterfactual explanations with semantics is carefully described. For simple images, such as low-resolution face portraits, the synthesis of visual counterfactual explanations has recently been proposed as a way to uncover the decision mechanisms of a trained classification model. In this case, the problem of producing counterfactual explanations for high-quality images and complex scenes for the self-driving domain is addressed. Leveraging recent semantic-to-image models, a generative counterfactual explanation framework is presented that produces plausible and sparse modifications which preserve the overall scene structure. Furthermore, the concept of "region-targeted counterfactual explanations", and a corresponding framework are described, where users can guide the generation of counterfactuals by specifying a set of semantic regions of the query image the explanation must be about. Extensive experiments conducted on challenging datasets, including high-quality portraits (CelebAMask-HQ) and driving scenes (BDD100k) are summarized.

Finally, this presentation summarizes how to design explanations of driving behaviour with multilevel fusion. The idea is to generate high-level driving explanations as the vehicle drives using a deep learning architecture which explains the behaviour of a trajectory prediction model (the so called BEEF, for BEhavior Explanation with Fusion). The model is supervised by annotations of human driving decision justifications, and it learns to fuse features from multiple levels by modeling the correlations between high-level decisions and midlevel perceptual features. The experiments are finally presented and discussed.

## 4.5 Adversarial ML in the Wild

**Presenters**: Kathrin Grosse, University of Cagliari, Italy.

This presentation focuses on the practical, e.g. industry perspective on AML. More concretely, our findings are from interviewing 15 ML practitioners from start-ups and discuss two intriguing properties emerging from these interviews: (1) participants do not distinguish between AML and non-ML security, and (2) participants do not just reason about an individual model, but rather about a workflow and sometimes even the surrounding system.

To better understand this perception of AML, we discuss our findings from a larger survey with more than 140 participants and investigate what threats to AML have been encountered so far and what factors we found to influence exposure to such threats in the wild.

## 4.6 Phantom of the ADAS and the translucent patch

**Presenters**: Yuval Elovici and Asaf Shabtai, Ben-Gurion University, Israel.

This research investigates the "split-second phantom attacks," a scientific gap that causes two commercial advanced driver-assistance systems (ADASs), Telsa Model X (HW 2.5 and HW 3) and Mobileye 630, to treat a depthless object that appears for a few milliseconds as a real obstacle/object. We discuss the challenge that split-second phantom attacks pose for ADASs. We demonstrate how attackers can apply split-second phantom attacks remotely by embedding phantom road signs into an advertisement presented on a digital billboard, which causes Tesla's autopilot to suddenly stop the car in the middle of a road and Mobileye 630 to issue false notifications. We also demonstrate how attackers can use a projector in order to cause Tesla's autopilot to apply the brakes in response to a phantom of a pedestrian that was projected on the road and Mobileye 630 to issue false notifications in response to a projected road sign. To counter this threat, we propose a countermeasure that can determine whether a detected object is a phantom or real using only the camera sensor. The countermeasure (GhostBusters) uses a "committee of experts" approach and combines the results obtained from four lightweight deep convolutional neural networks that assess the authenticity of an object based on the object's light, context, surface, and depth. We demonstrate our countermeasure's effectiveness (it obtains a TPR of 0.994 with an FPR of zero) and test its robustness to adversarial machine learning attacks.

Physical adversarial attacks against object detectors have seen increasing success in recent years. However, these attacks require direct access to the object of interest in order to apply a physical patch. Furthermore, to hide multiple objects, an adversarial patch must be applied to each object. In this paper, we propose a contact-less translucent physical patch containing a carefully constructed pattern, which is placed on the camera's lens, to fool state-of-the-art object detectors. The primary goal of our patch is to hide all instances of a selected target class. Furthermore, the optimization method used to construct the patch aims to ensure that the detection of other (untargeted) classes remains unharmed. Therefore, in our experiments, which are conducted on state-of-the-art object detection models used in autonomous driving, we study the effect of the patch on the detection of both the selected target class and the other classes. We show that our patch was able to prevent the detection of 42.27% of all stop-sign instances while maintaining high detection of the other classes.

## 4.7 Safe Motion Planning among Decision-Making Agents

**Presenter**: Javier Alonso Mora, TU Delft, The Netherlands.

In smart cities, where mobile robots will co-exist with humans, autonomous vehicles will provide on-demand transportation while making our streets safer. Therefore, the motion plan of mobile robots and autonomous vehicles must account for the interaction with other agents and consider that they are also decision-making entities that may cooperate. Towards this objective several methods for motion planning and multi-robot coordination are discussed that leverage constrained optimization and reinforcement learning and ways to model and account for the inherent uncertainty of dynamic environments. The methods are of broad applicability, including autonomous vehicles, mobile manipulators and aerial vehicles.

## 4.8 PRISSMA project: Current testing and validation approaches, main limitations and challenges of AVs

**Presenter**: Rafaël De Sousa Fernandes, UTAC, France.

The PRISSMA project aims at proposing a platform that will allow to lift the technological barriers preventing the deployment of secure AI-based systems and to integrate all the elements necessary for the realization of the type-approval activities for autonomous vehicles and their validation in their environment for a given use case.

By identifying the safety and security objectives for AI-based autonomous mobility systems, comprehensive reliability validation processes are developed for the commercial operation of autonomous mobility services. The proposed approach ensures the availability of shared concepts to address the complexity of AI-based autonomous mobility systems that can be used internationally.

The project also attempts to enhance the participation of France in the implementation of prerequisites, allowing to position itself at the European level to host one of the Testing Facilities for autonomous mobility that will be developed in the coming years.

Multiple test scenarios, methodologies and associated intervention procedures for real-life tests of autonomous mobility systems in addition to the previous tests are designed and proposed, including practical implementation of the qualification processes of the testing facilities in controlled environments.

The proposal focuses on the practical implementation of test plans in addition to simulation, with identification of the optimum perimeter of the system of systems, i.e. vehicle, infrastructure and supervision. This implementation also includes dysfunctional through injection of failures. Finally, a detailed specification of the necessary testing facilities (infrastructure, equipment, supervision systems, personnel) and their qualification is evaluated and discussed in a national or European perspective.

# 5 Day 2: Implementing trustworthy AI in AV testing

This session was planned to address the question of implementing the requirements of Trustworthy AI in the context of Automated and Autonomous Vehicle testing. The requirements of robustness and accuracy, fairness, and cybersecurity are likely to be major elements in future testing strategies to ensure the safe, secure, and ethical adoption of AI in automated vehicles. Finally, the session was adapted according to the accepted presentations.

## 5.1 Towards Robust Autonomous Vehicles

**Presenter**: Alexandre Alahi, EPFL, Switzerland.

The AI of autonomous vehicles is based on the 3 P: Perception, Prediction, and Planning. Both industry and the research communities have acknowledged the need for such pillars by providing public benchmarks. While the state-of-the-art methods are impressive, they still do not generalize well to cities outside of the benchmarks. Focusing on the prediction pillar, this work shares the current limitations of state-of-the-art work.

## 5.2 Know the rules well so you can break them effectively - Can we ensure AVs drive safely?

**Presenter**: Nick Reed, REED Mobility, UK.

This talk begins with the presentation of Reed Mobility, an initiative that began in June 2019 to focus on automated vehicle safety. It participated in the Commission Expert Group appointed by the Commission to advise on specific ethical issues raised by driverless mobility. The presentation summarizes the main conclusions achieved by the expert panel regarding safety, transparency and responsibility.

The main discussion then focuses on some recommendations. For example, consider the revision of traffic rules to promote the safety of connected and automated vehicles. Rules are a means by which road safety is achieved, but non-compliance is sometimes necessary to achieve greater road safety (i.e. the ethics of AV driving behaviour and whether they will deviate from the rules of the road to maximize safety). How should an automated vehicle handle this? Looking at the UK regulatory framework and the views expressed in its consultation, it is clear that there is no agreement from industry and experts.

Some examples are analyzed and discussed, including crossing a red light or exceeding the speed limit, in cases where it makes sense to increase safety. The proposal to address these examples is to define ethical goal functions that may go beyond traffic rules in some cases.

Furthermore, the presentation focuses on recommendations related to safety and inequalities. Some safety metrics are discussed, including the distribution of risk to address inequalities and dilemmas.

Finally, the importance of data and some of its features are analyzed, including new tools such as digital commentary driving.

## 5.3 Robustness testing for automated driving as an example of the BSI's approach to AI cybersecurity

**Presenter**: Christian Berghoff, BSI, Germany.

The talk covers the BSI's strategy for the secure, robust and transparent application of AI in automated driving. It sets out the BSI's general perspective on the problem, steps already taken, and actions planned in the future. The generic considerations are complemented by the presentation of a case study on the robustness assessment of traffic sign classifiers, which was carried out by BSI.

## 5.4 The actual ethics of AI for AVs: from autonomy to attachments

**Presenter**: Jack Stilgoe, University College London, London, UK.

This presentation discusses some general aspects of the ethics of autonomous vehicles. It begins by mentioning the Moral Machine experiment and Waymo's annual safety reports, and continues with some of the myths of autonomy, highlighting the fact that AVs are conditioned and somehow "driven" by people outside the vehicle (e.g., pedestrians, cyclists, other drivers, etc.)

The concept of attachment is presented, including its social and technical dimensions. Some famous accidents are presented, including the Uber fatal crash and the Toyota e-Palette incident in the Tokyo 2020 Paralympic Athletes' Village. Some strategies are proposed, including heterogeneous engineering and reducing the complexity of the space.

Some reflection is given to the different layers of rules (i.e. physical, legal, advisory, and normative) from concrete to culture, which are technologically and socially mediated.

Finally, some preliminary information is provided regarding the forthcoming report on "Ethics and responsible innovation for AVs" (UK CDEI/CCAV), which includes road safety, explainability and data sharing, data privacy, fairness and transparency.

## 5.5 Towards Explainable and Trustworthy Autonomous Systems

**Presenter**: Lars Kunze, Oxford Robotics Institute, UK.

Autonomous systems operating in real-world environments are required to understand their surroundings, assess their capabilities, and explain what they have seen, what they have done, what they plan to do, and why to different stakeholders, including end users, developers, and regulators. This talk discussed the results and objectives of three research projects: SAX (https://ori.ox.ac.uk/projects/sense-assess-explain-sax/), RoAD (https://ori.ox.ac.uk/projects/road/), and RAILS (https://ori.ox.ac.uk/projects/rails/). In our work, we focus on autonomous vehicles and their application in challenging open-ended environments. As it is essential that these systems are safe and trusted, we design, develop, and evaluate fundamental technologies in simulation and real-world applications to overcome critical barriers which impede the current deployment of autonomous vehicles in economically and socially important areas.

## 5.6 Man, Machine, or In Between: The Process of Investigations Into Automation

**Presenter**: Robert Swaim, HowThingsBroke, USA.

This presentation introduces how to start an investigation at the vehicle level by an experienced accident investigator. It begins with aspects that engineers and programmers do not normally encounter but should be aware of, such as types of investigation and jurisdiction about investigation leadership. The discussion relates why it is necessary to establish functional groups and limit initial efforts to gathering facts before analysis. Types of failure analysis are introduced, including a mention of their limitations. The layers of man-machine interface in aviation accident case examples involving autopilots show how design assumptions led to accidents in the real world.

## 5.7 Safe path to vehicle automation: Crash investigation perspective

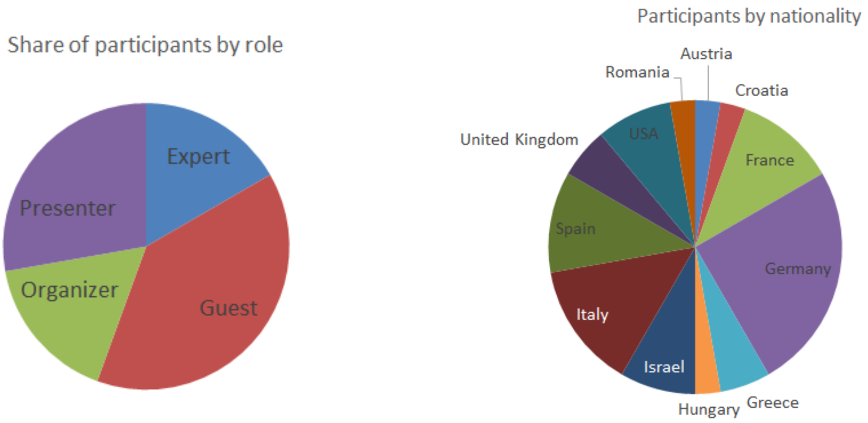
**Presenter**: Ensar Becic, NTSB, USA.

Crash investigations provide a unique view of the real-world risks that affect vehicle automation. By taking a holistic approach to crash investigations, NTSB determines not only the specific failures of vehicle automation, but deficiencies extending to regulatory oversight and the safety culture of the developer. This presentation provides examples of crash investigations that identified limitations of vehicle automation, the role of a human, and the erroneous assumptions of the developer related to the interaction of AVs with the environment.

# 6 Conclusions

In this report, we summarized the outcome of the Exploratory Research workshop entitled "Toward explainable, robust and fair AI in automated and autonomous vehicles" hold online on 29-30 of March 2022. 36 participants attended the workshop and 14 presentations were given. We selected six experts who will further elaborate on selected topics (work in progress). The participants came from 12 countries, as Figure 1 shows, including regions outside the EU as well.

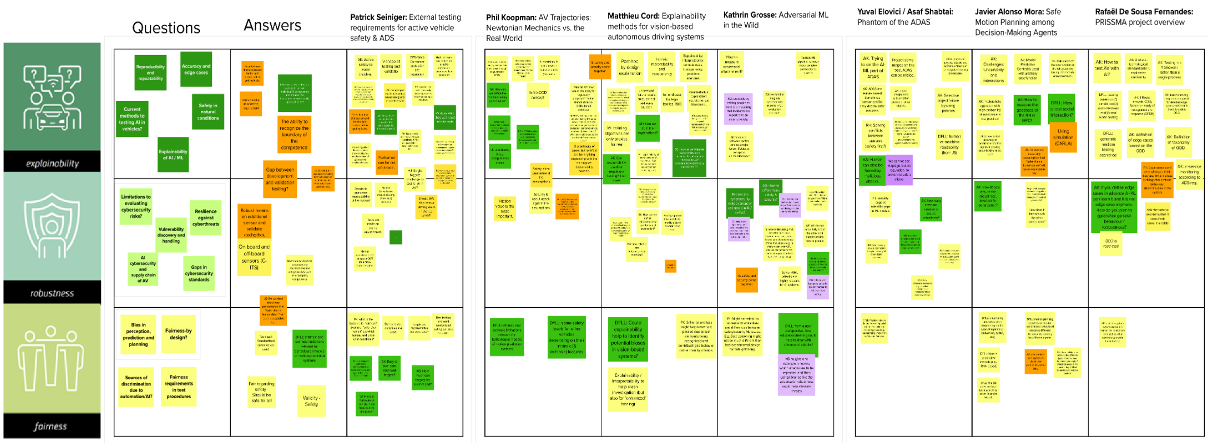
**Figure 2:** The distribution of participants in the workshop




*Source:*JRC analysis

After each day, all attendees were asked to participate in a group exercise using MURAL. We grouped the questions and jointly agreed on possible answers or approaches to answer or further research them. In Figure 3 the green, yellow and orange boxes represent questions, comments or answers, and highlighted answers, respectively, to the main scientific questions.

**Figure 3:** Example of the results of the collaborative work carried out with Mural for Day 1.

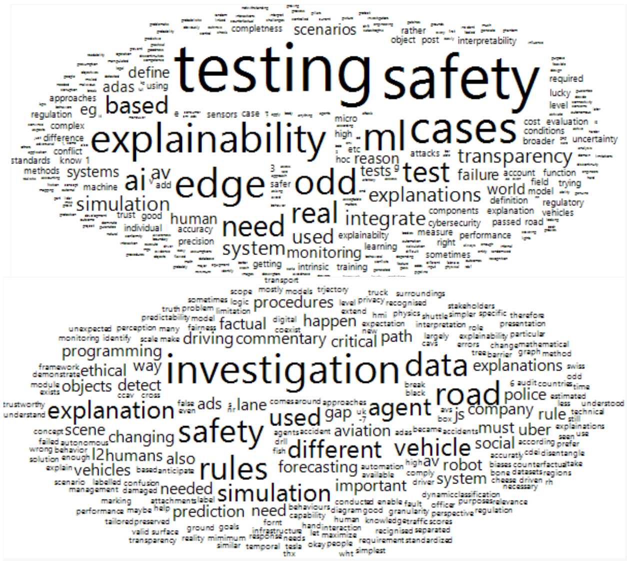



*Source:*JRC analysis

After the group exercises, we performed a keyword and phrase analysis separately for explainability (Figure 4), cybersecurity (Figure 5) and fairness (Figure 6). Although both explainability and robustness (cybersecurity) were intensively discussed and commented on, fairness generated fewer questions. Therefore, we identified gaps in the scientific literature on the relationship between fairness and safety of A&AVs.

Current physical safety testing methods do not cover all cases of real-world driving. Edge cases always exist in the real world, they can depend on the actual system, and may be different from human edge cases. Therefore, physical tests are not enough and evidence of good AI safety engineering is needed. Since AI is difficult to integrate into the V-shaped development process, recent safety audit standards may not be enough to ensure safety on real roads during all normal driving scenarios. Explainable AI can bridge this gap. There are several established interpretability methods, for example, factual and counterfactual reasoning, etc., that can be used for development; however, during testing, they may only be suitable to ensure if the A&AVs passes the test for the right reason. Further research is needed to understand how individual AI components can be integrated into the broader A&AVs explainability. Furthermore, the explainability of accidents can help in post-crash investigation, but it requires a different taxonomy than applied during development and regulations. Therefore, agreement between the stakeholders of A&AV on the different use cases and the establishment of reasoning vocabulary is of great importance.

**Figure 4:** Keywords and phrase analysis for the topic on explainability.




**Phrases**
edge cases
real world
ai components into broader view
av transparency explainabilty approaches
know that test passed for the right reason
av transparency explainabilty
integrate explainability of
individual AI components

*Source:* JRC analysis

Evaluation of an autonomous product for cybersecurity is an emerging topic. Both adversarial attacks on sensors and data privacy are important to consider, and recently they are not assessed at the vehicle level. Training data set audit and collection of real traffic data must be performed in addition to physical tests. A certification readiness matrix must be developed for each cyber scenario for different adversarial attacks and for naturally occurring perturbations. Understanding how individual AI components interact in an embedded system also plays a critical role in cybersecurity.

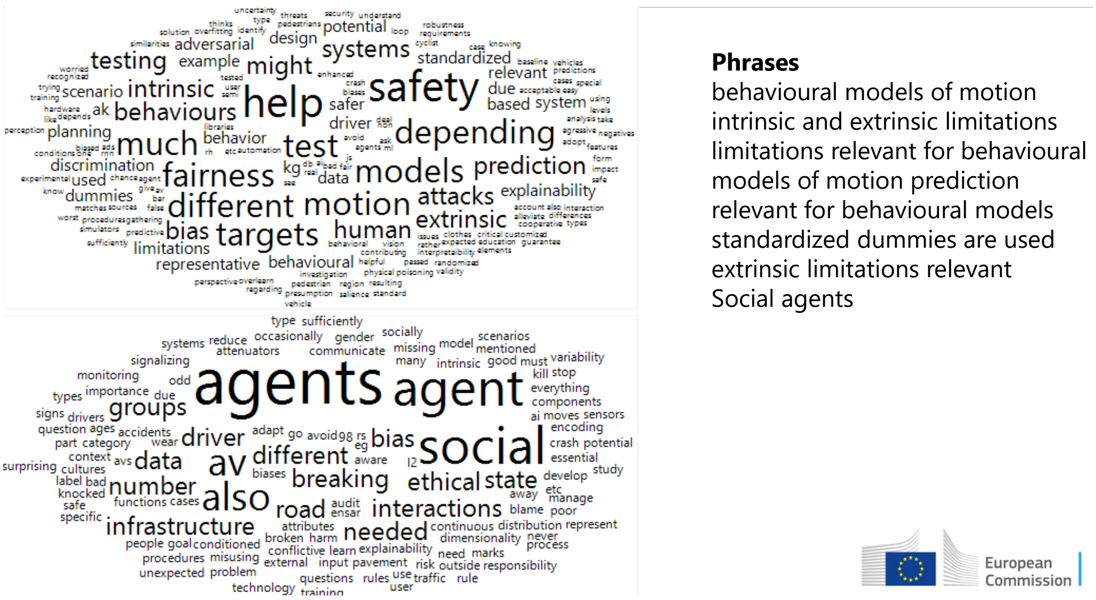
**Figure 5:** Keywords and phrase analysis for the topic on robustness.


**Phrases**
aml attacks
evaluation of autonomous products
ul 4600
Attack library
Data privacy
Robustness is not only cybersecurity

*Source:* JRC analysis

Experts also highlighted that behavioural models are missing for motion prediction of different social agents and tests with standardized dummies lack the features of different social groups. Therefore, currently there are not enough data to assess the fairness of A&AV vehicles and how fairness or bias influences safety.

**Figure 6:** Keywords and phrase analysis for the topic on fairness.




*Source:* JRC analysis

## References


Alonso Raposo, M., Ciuffo, B., Alves Dias, P. and et al., 'The future of road transport', Publications Office of the European Union, Luxembourg,, Vol. EUR 29748 EN, JRC116644, 2019. doi:10.2760/524662.

Dede, G., Hamon, R., Junklewitz, H., Naydenov, R., Malatras, A. and Sanchez, I., 'Cybersecurity challenges in the uptake of artificial intelligence in autonomous driving', EUR 30568 EN, Publications Office of the European Union, 2021.

European Commission, 'A European approach to artificial intelligence'. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence, 2018.

European Commission, 'Ethics of connected and automated vehicles: recommendations on road safety, privacy, fairness, explainability and responsibility', Horizon 2020 Commission Expert Group to advise on specific ethical issues raised by driverless mobility (E03659), 2020.

Fernández-Llorca, D. and Gómez, E., 'Trustworthy autonomous vehicles', Publications Office of the European Union, Luxembourg,, Vol. EUR 30942 EN, JRC127051, 2021. doi:10.2760/120385.

Mattas, K., Makridis, M., Botzoris, G., Kriston, A., Minarini, F., Papadopoulos, B., Re, F., Rognelund, G. and Ciuffo, B., 'Fuzzy surrogate safety metrics for real-time assessment of rear-end collision risk. a study based on empirical observations', Accident Analysis and Prevention, Vol. 148, JRC120779, 2020, p. 105794.

Re, F., Kriston, A., Broggi, D. and Minarini, F., 'Testing the robustness of commercial lane departure warning systems', Transportation Research Record, Vol. 2675, JRC122657, No 12, 2021, pp. 385–400.

## List of abbreviations and definitions

**3P** Perception, Prediction and Planning

**A&AV** Automated and Autonomous Vehicles

**ACC** Adaptive Cruise Control

**ADAS** Advanced Driving Assistance System

**ADS** Automated Driving Systems

**AI** Artificial Intelligence

**AML** Adversarial Machine Learning

**AV** Automated/Autonomous Vehicle

**BEEF** Behavioural Explanation with Fusion

**BSI** Bundesamt für Sicherheit in der Informationstechnik (Federal Office for Information Security)

**AV** Connected & Automated/Autonomous Vehicle

**EURONCAP** European New Vehicle Assessment Program

**FCW** Forward Collision Warning

**JRC** Joint Research Centre

**LDW** Lane Departure Warning

**ML** Machine Learning

**NHTSA** National Highway Traffic Safety Administration

**NTSB** National Transportation Safety Board

**ODD** Operational Design Domain

**TTC** Time to Collision

**VUT** Vehicle Under Test

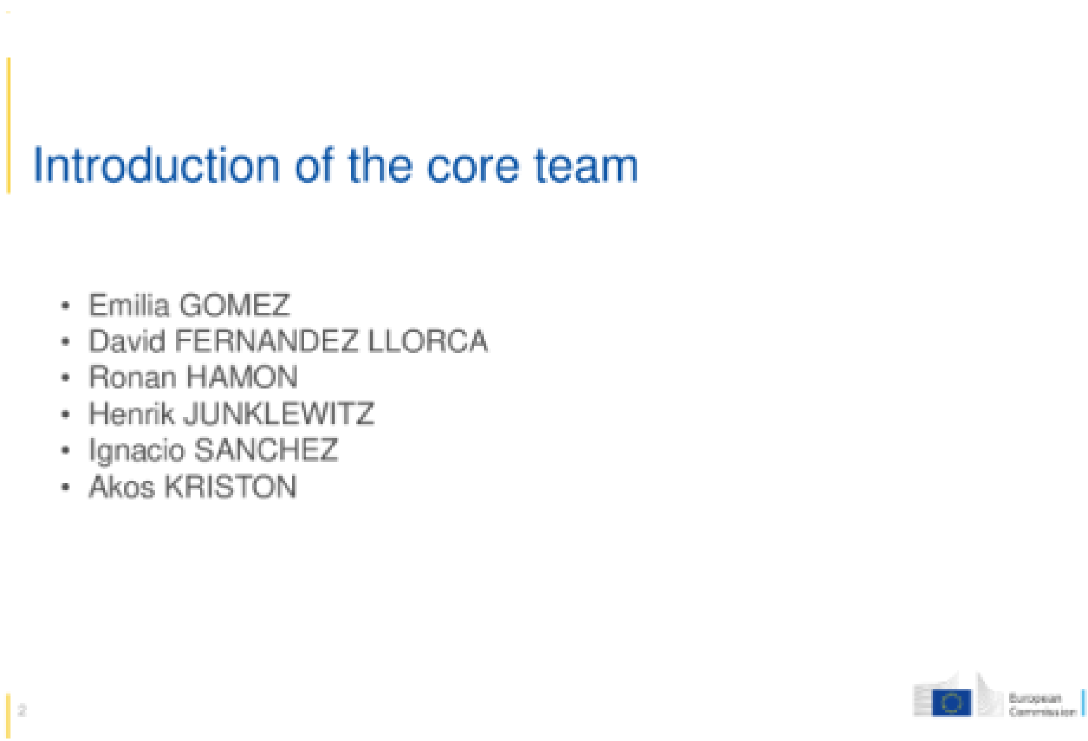
**xAI** Explainable Artificial Intelligence

## List of figures

## List of tables

## Annexes

### Annex I. Presentation of the Workshop

# Introduction and rules of the days

- Introduction of Joint Research Centre (JRC)
- Introduction of the conference and the organizers
- Program and sessions
- Research questions
- Rules of the days

3

European Commission

# The EC political leadership


#Eustrivesformore #vdLcommission

4

European Commission

## The JRC within the Commission


5

European Commission

## JRC Mission

As the science and knowledge service of the European Commission our mission is to support EU policies with independent evidence throughout the whole policy cycle.

6

European Commission

# JRC role

- **Independent** of private, commercial or national interests
- **Policy neutral**: has no policy agenda of its own
- Works for more than **20 EC policy departments**

# JRC sites

Headquarters in **Brussels** and research facilities located in **5 Member States:**

- Belgium (Geel)
- Germany (Karlsruhe)
- Italy (Ispra)
- The Netherlands (Petten)
- Spain (Seville)

# JRC facilities – some examples

Virtual tour at https://visitors-centre.jrc.ec.europa.eu/en/media?type=8

Vehicle Safety (MASSAF)

European Solar Test Installation (ESTI)

Vehicle Emission Laboratory (VELA)

European Interoperability Centre for Electric Vehicles and Smart Grids

Battery Testing Facility Laboratory

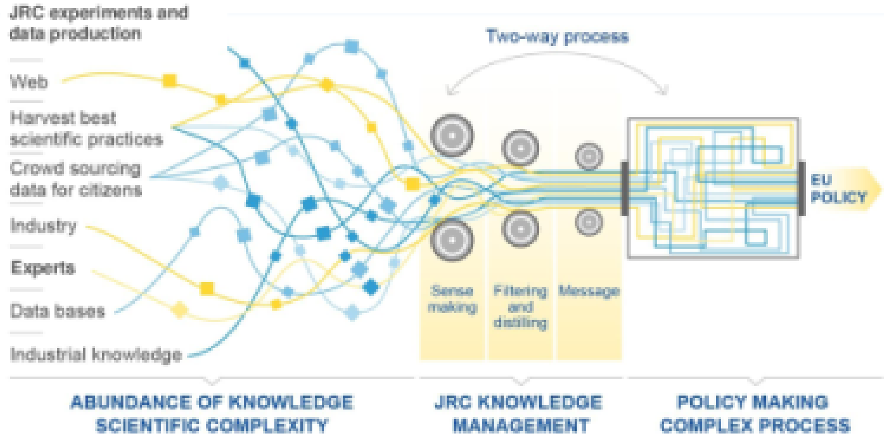
Nuclear Forensic Laboratory

9

European Commission

# JRC's prenormative research

- What?
- Why?
- How?

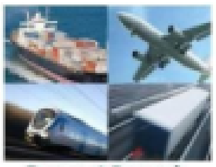



10

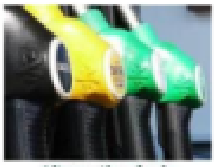
European Commission

# Sustainable Transport Unit


Electric Mobility

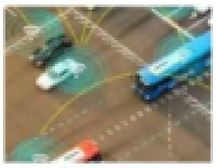

Transport Research and Innovation

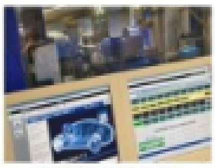
UNECE Global Standardization


Alternative fuels


Vehicle emissions


Vehicle safety Compliance Testing


Cooperative, Connected and Automated Mobility


Vehicle Emissions laboratories

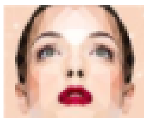




# Digital Economy Unit

Mix of **social**, **economic** and **technological** expertise to study the current and emerging facets of **digital transformation**, and its impacts on the European economy, society and environment.

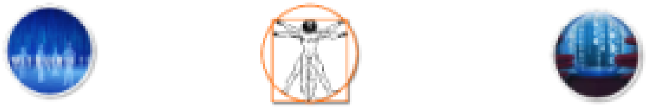
To study the social and economic **impacts of Artificial Intelligence** (AI), **data** and **digital platforms**, advancing research on **methodologies to ensure trustworthy AI**.



# JRC Cyber & Digital Citizens‘ Security Unit

To strengthen **trust** and **security** of the European Citizen in a sustainable and inclusive ICT-based European society by scientific research on how emerging Information and Communication Technologies will impact on the security and privacy of citizens' daily life.

To work on **risk mitigation**, on **cybersecurity**, **cybercrime**, **data protection**, **privacy** and on the associated legal and regulatory frameworks aiming at a balance between European security needs and fundamental citizen rights including from the perspective of the emerging Digital Single Market.



13

European Commission

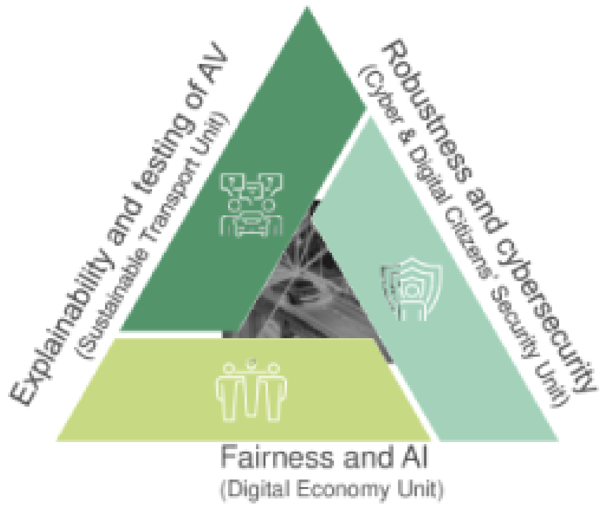
# JRC EXPLORATORY WORKSHOP

**INTRODUCTION**

*This JRC Exploratory Workshop is dedicated to the safety and security of automated and autonomous vehicles (A&AV), and aims to bring together leading scientist and engineers to explore and discuss the state-of-the-art research on accuracy, robustness, fairness and explainability of artificial intelligence (AI) and machine learning (ML) and testing of modern vehicles.*


European Commission

## Concept and organizers




15

**JRC EXPLORATORY WORKSHOP**

*Toward **explainable**, **robust** and **fair** AI in automated and autonomous vehicles: challenges and opportunities for safety and security*

***Day 1***

*Fundamentals of testing AI in AVs – Current situation and challenges*

***(March 29, 2022)***

Join by web: Link

Meeting number (access code):
274 170 84994
Meeting password: wGXdGk@238
(94983451 from phones)
**Join by phone**
+49-619-6781-9736 Germany Toll
Global call-in numbers
**Join from a video conferencing system or application**
Dial 27417084994@ecconf.webex.com
**Alternatively** dial 62.109.219.4 and enter your meeting number.

| Time | Presenter | Title |
|---|---|---|
| 13.30 – 14.00 | JRC | Presentation of the Workshop. Moderator A. Kriston (JRC) |
| 14.00 – 14:30 | Patrick Seiniger<br>BASt, Germany | External testing requirements for active vehicle safety & ADS |
| 14.30 – 15.00 | Philip Koopman<br>Carnegie Mellon University, USA | AV Trajectories: Newtonian Mechanics vs. the Real World |
| 15.00 – 15.30 | Matthieu Cord<br>Valeo, France | Explainability methods for vision-based autonomous driving systems |
| 15.30 – 15:45 | **Break** | Moderator: R. Hamon (JRC) |
| 15.45 – 16.15 | Kathrin Grosse<br>University of Cagliari, Italy | Adversarial ML in the Wild |
| 16.15 – 16.45 | Yuval Elovici / Asaf Shabtai<br>Ben-Gurion University, Israel | Phantom of the ADAS: Securing advanced driver-assistance systems from split-second phantom attacks |
| 16.45 – 17.00 | **Break** | Moderator: Emilia Gomez (JRC) |
| 17:00 – 17.30 | Javier Alonso Mora<br>TU Delft, The Netherland | Safe Motion Planning among Decision-Making Agents |
| 17.30 – 18.00 | Rafaël De Sousa Fernandes<br>UTAC, France | PRISSMA project overview |
| 18.00 – 18.30 | **Discussion session and Wrap-up**<br>Social networking | |

For support: JRC-ExtrAIsafe@ec.europa.eu

European Commission

**JRC EXPLORATORY WORKSHOP**

*Toward **explainable**, **robust** and **fair** AI in automated and autonomous vehicles: challenges and opportunities for safety and security*

***Day 2***

*Implementing Trustworthy AI in AV testing*

***(March 30, 2022)***

Join by web: Link

Meeting number (access code): 274 403 35224
Meeting password: MgVkNgS2@55 (64856472 from phones)
**Join by phone**
+49-619-6781-9736 Germany Toll
Global call-in numbers
**Join from a video conferencing system or application**
Dial 27440335224@ecconf.webex.com
**Alternatively** dial 62.109.219.4 and enter your meeting number.

| Time | Presenter | Title |
|---|---|---|
| 13.30 – 14.00 | JRC | Welcome in the meeting room. Moderator: H. Junklewitz (JRC) |
| 14.00 – 14:30 | Alexandre Alahi<br>EPFL, Switzerland | Towards Robust Autonomous Vehicles |
| 14.30 – 15.00 | Nick Reed<br>Reed Mobility, UK | Know the rules well so you can break them effectively - Can we ensure AVs drive safely? |
| 15.00 – 15.30 | Christian Berghoff<br>BSI, Germany | Robustness testing for automated driving as an example of the BSI's approach to AI cybersecurity |
| 15.30 – 15.45. | **Break** | Moderator: D. Fernandez Llorca (JRC) |
| 15.45 – 16:15 | Jack Stilgoe<br>University College London, UK | The actual ethics of AI for AVs: from autonomy to attachments |
| 16.15 – 16.45 | Lars Kunze<br>Oxford Robotics Institute, UK | Towards Explainable and Trustworthy Autonomous Systems |
| 16.45 – 17.00 | **Break** | Moderator: A. Kriston (JRC) |
| 17.00 – 17.30 | Robert Swaim<br>Safety expert, USA | Man, Machine, or In Between: The Process of Investigations Into Automation |
| 17.30 – 18.00 | Ensar Becic<br>NTSB, USA | Safe path to vehicle automation: Crash investigation perspective |
| 18.00 – 18.30 | **Discussion session and Wrap-up** | |

For support: JRC-ExtrAIsafe@ec.europa.eu

European Commission

**JRC EXPLORATORY WORKSHOP**

*Toward **explainable**, **robust** and **fair** AI in automated and autonomous vehicles: challenges and opportunities for safety and security*

## Research questions

*explainability*

- *What are the current testing methods of AI-ML components in automotive environment?*
- *How can we test the AI-ML components in terms of safety in automotive environment?*
- ***How to define and quantify the robustness and accuracy, of an A&AV's AI-ML component?***
- *Are the behaviours of AI-ML components of an A&AV reproducible and repeatable in controlled environments and in the wild?*
- *How is it possible to explain the decisions made by AI-ML components in A&AV from a software engineering, vehicle safety testing and accident investigation perspectives?*

*robustness*

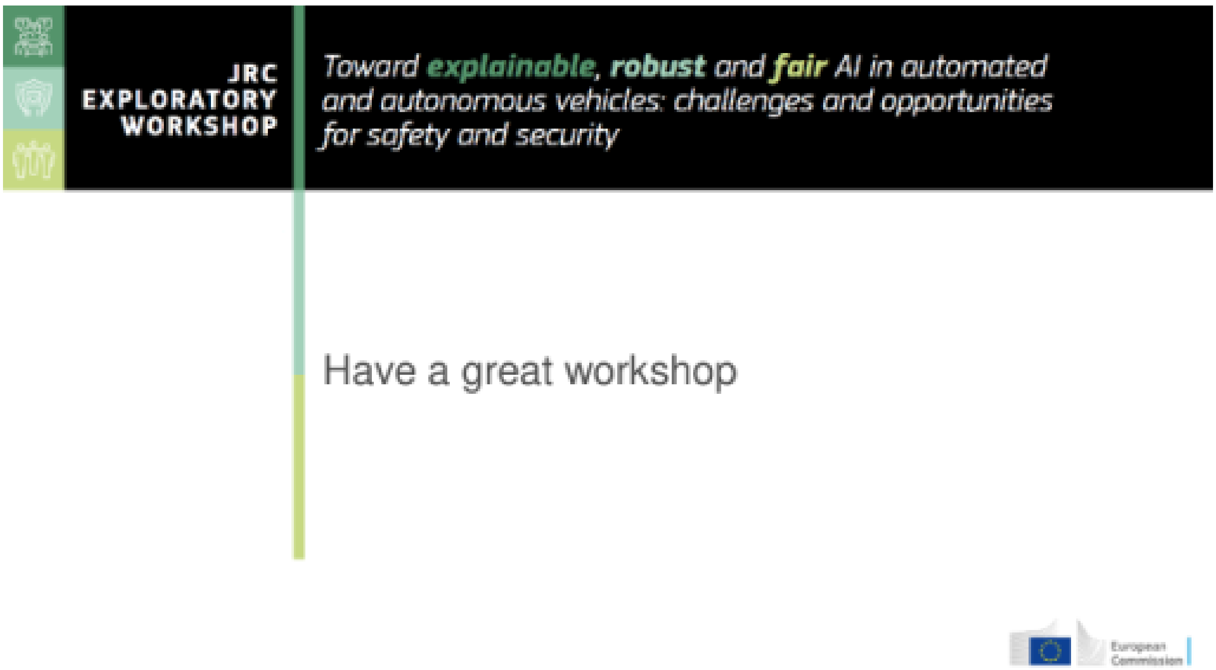
- *What are the cybersecurity threats and vulnerabilities associated with AI component in AVs?*
- *What are the limitations of current vehicle testing methods to evaluate AI cybersecurity risks?*
- *How can we measure the resilience of vehicle systems against cyber threats targeting AI components?*
- *How can we handle the security vulnerabilities discovered in the AI components of automated vehicles?*
- *What are the AI-related cybersecurity challenges connected to the supply chain of AV?*
- *What is the state of cybersecurity standards for AI in automated driving, and what are the gaps that would need to be addressed?*

*fairness*

- *What elements would be affected and how if we consider fairness as a requirement in test procedures?*
- *How to detect biases in automated decisions and assess their impact in terms of fairness and robustness?*

European Commission


JRC EXPLORATORY WORKSHOP
Toward explainable, robust and fair AI in automated and autonomous vehicles: challenges and opportunities for safety and security
Virtual collaboration during discussions MURAL
See Webex chat for the link!
Add sticky notes here for further questions to answer
The most important take away of the presentations
During the wrap-up we consolidate the possible answers and the main conclusions
Navigate through the presentations
European Commission


JRC EXPLORATORY WORKSHOP
Toward explainable, robust and fair AI in automated and autonomous vehicles: challenges and opportunities for safety and security
Have a great workshop
European Commission

## Annex II. External testing requirements for active vehicle safety & ADS

# Testing of ADAS/ADS



### What is Active Vehicle Safety?

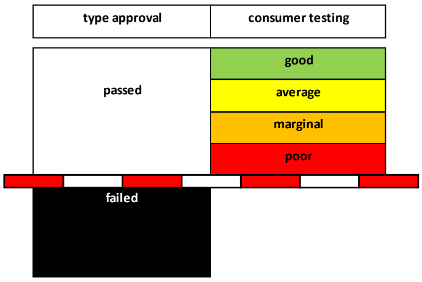
- Active Vehicle Safety
  **Avoidance of Accidents!**
- Passive Vehicle Safety
  **Mitigation of Consequences**

- Sight
- Driver Conditions
- Ride and Handling
- Automotive Lighting
- Driver Assistance in general


Patrick Seiniger

2

## Requirements?




### ➲ *Consumer Protection*

- Tests conducted on own proprietary criteria
- Comparative review of vehicle safety after marked entrance
- Does not cover all vehicles and is **by no means required**
- **Quicker** for new technology

### ➲ *Type Approval (=Legal Requirements)*

- Requirements discussed on international level
- **Threshold for entrance into market** – minimum standard
- Tests conducted by technical services
- Approval issued by type approval authority
- **Mandatory**

Patrick Seiniger 3

## Legal Requirements

Text and Requirements


**Contracting Parties**
agree on
Technical **Regulations**
(e.g. UN R 79:
Steering Equipment)

EUROPEAN COMMISSION

**Regulations** or **Directives**
for **Member States**
can reference UN Regulations
e.g. in Regulation 858/2018 on
Passenger Car **Type Approval**

Things made mandatory*

* If UN process is slow, EU writes own regulations (Ex.: 347/2012 for AEBS)

Patrick Seiniger 4

## Test Concepts

- Consumer Protection:
  Implicit requirements by test procedures
  - Requires a large set of test cases
  - Typically on „sterile“ test track
  - Extreme tight tolerances for comparability

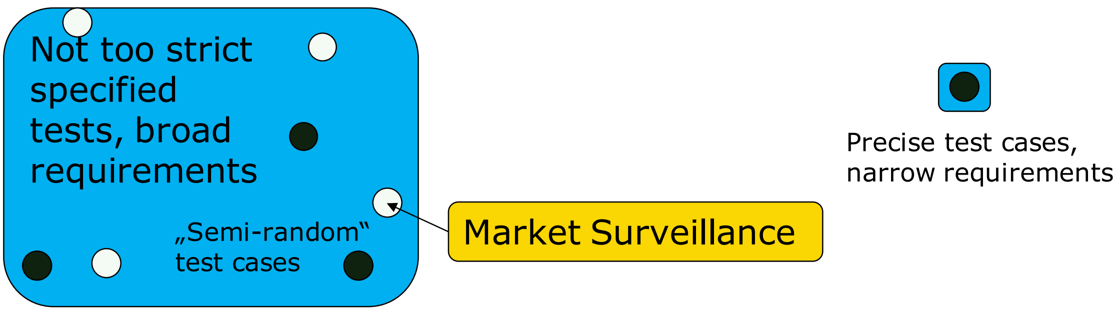
- Conventional Vehicle Regulations:
  Implicit requirements by test procedures
  - Typically only worst-case test cases
  - Typically on „sterile“ test track
  - Higher tolerances (e.g. no robots used)



## New sv. Conventional Approach for Regulations

**NEW** (concept first used with heavy vehicle emissions)

**CONVENTIONAL**

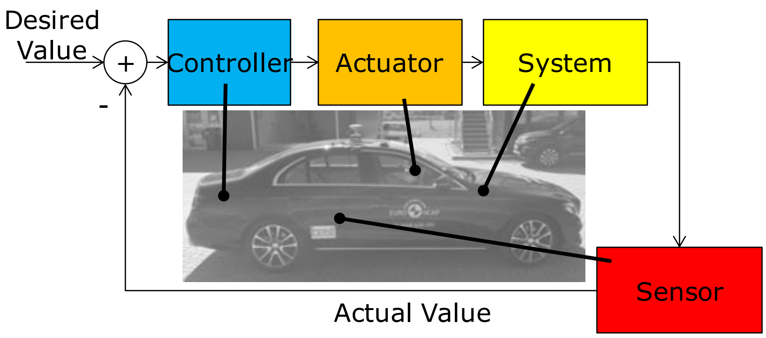



Disadvantage:
Technical Service could select easy cases

With market surveillance this turns into an advantage:
Manufacturers are forced to develop robust systems

## High Test Repeatability with Position and Speed Control

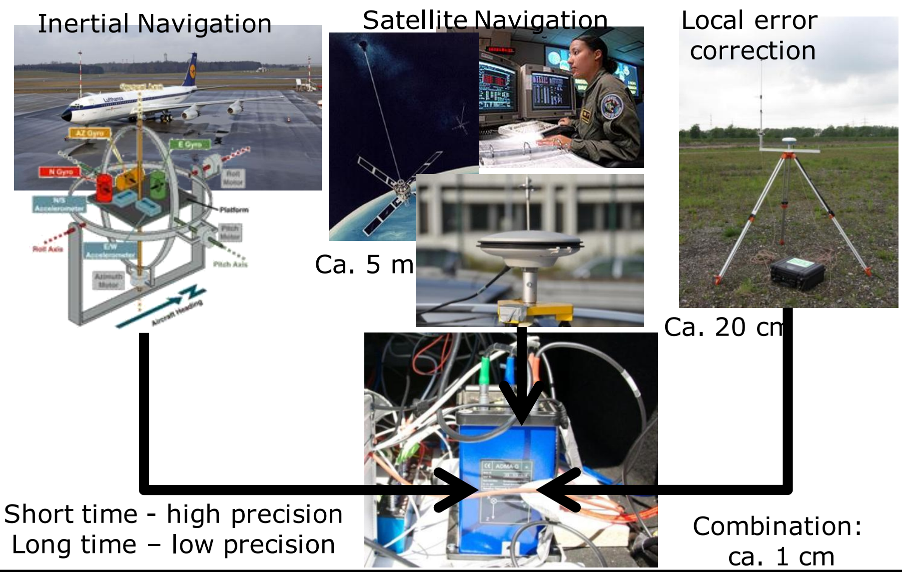
- Negative Feedback Control System






## Sensor: *Position Measurement*

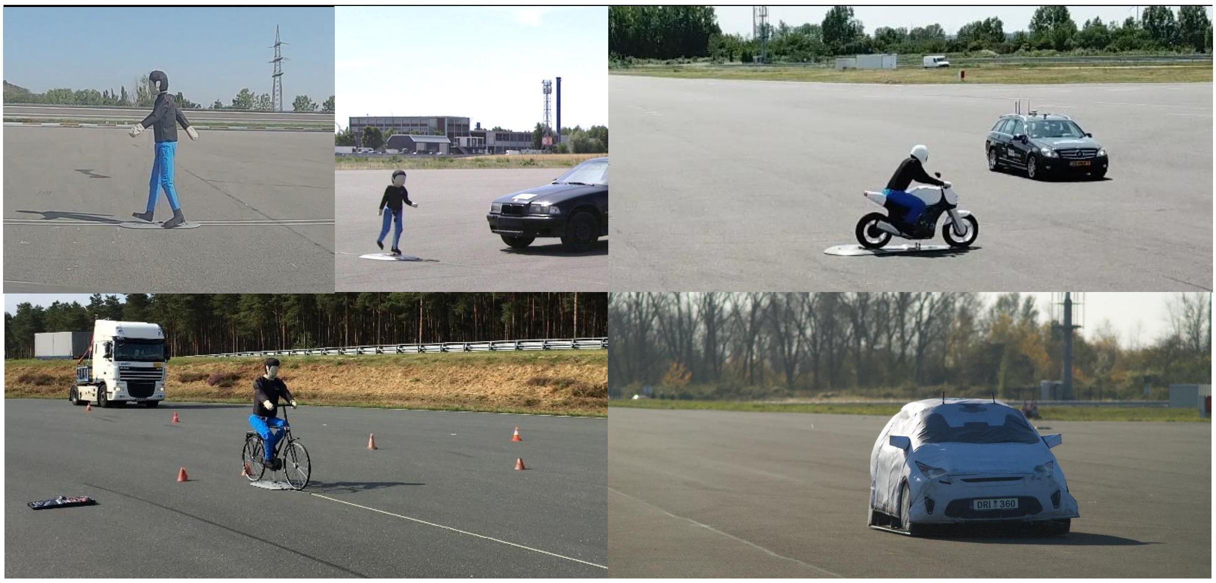




## Actuator: *Driving Robot*




## Targets (ISO19206) + Platforms



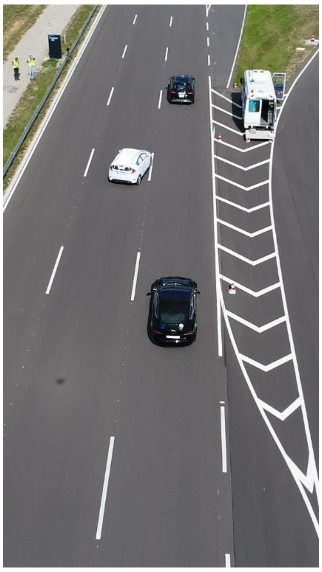
## Conclusions and Potential

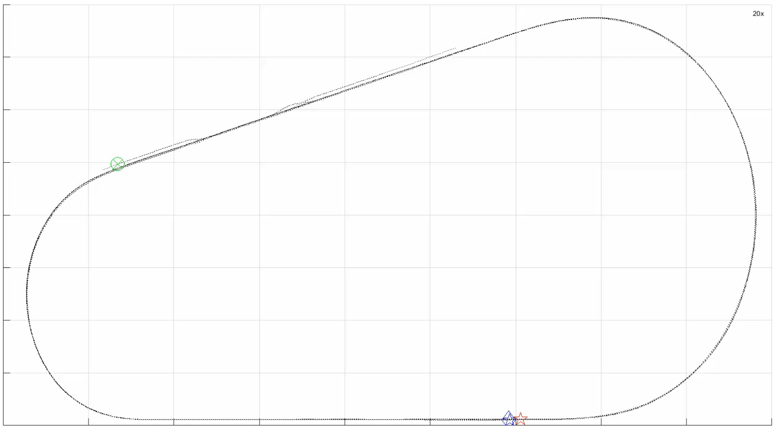
- Freely programmable tools to set up all kinds of scenarios
- Predefined test cases and procedures on test track

- Why not randomize tests on the spot?
- Why not test in realistic conditions?

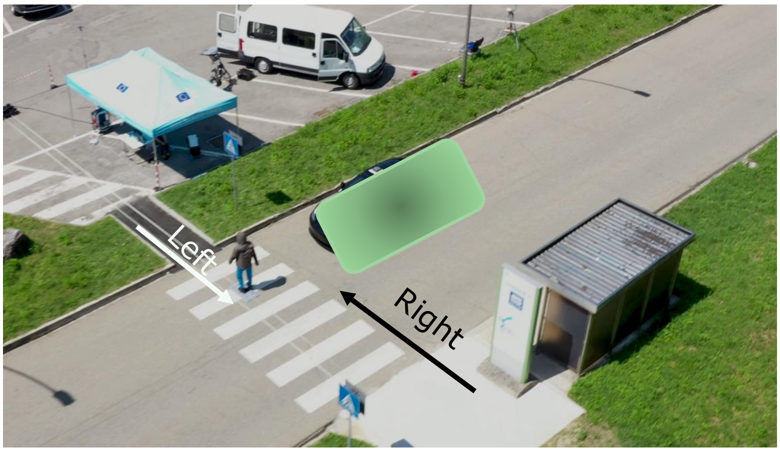


## Semi-Randomized Test





## Testing in Realistic Surroundings




- In JRC Left and Right represent different challenge to ADAS:
  - Left: arrives from grass covered area, legs are partially covered
  - Right: dummy arrives on the asphalt hence there is visually less distraction

Patrick Seiniger

13

# AV Trajectories: Newtonian Mechanics vs. The Real World

Prof. Philip Koopman

Carnegie Mellon University

@PhilKoopman

## Overview

Carnegie Mellon University

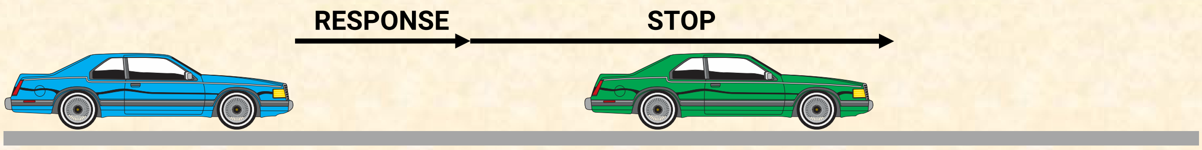
- **Limits on trajectory control**
  - Vehicle capability
  - Environmental conditions
- **Uncertainty**
  - About vehicle conditions
  - About environment
- **Managing ODD variations**
  - Micro-ODDs as an approach




2

# Example: Safe Following Distance

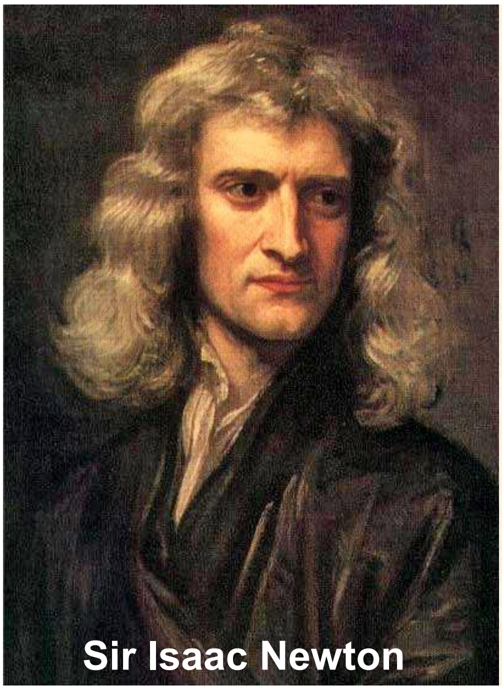
Carnegie Mellon University




- **Follower stops with space left behind leader (RSS example)**
  - Different initial speeds
  - Follower initially accelerating during response time
  - Different braking capabilities
  - Considered safe if any gap between vehicles at rest








# F=MA

Not Just
A Good Idea

…

# It's the Law!

## But, Where Does the "A" Come From?

Carnegie Mellon University

- **F = MA ➔ A = M / F**
  - BUT … F is limited by tire friction force

$F_{friction} = \mu * F_{normal}$ (6)

where:

- $F_{friction}$ is the force of friction exerted by the tires against the roadway
- μ is the coefficient of friction, which can vary for each tire
- $F_{normal}$ is the force with which the vehicle presses itself onto the road surface

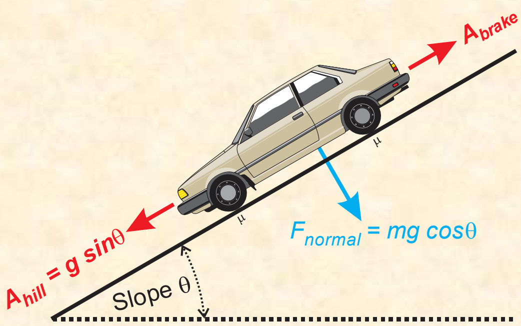
- **Example: braking depends upon:**
  - Ability of vehicle to exert force on roadway ($F_{friction}$)
  - Driver applying full $F_{friction}$ via brakes (braking capacity)

 5

## Road Conditions Affecting Braking

Carnegie Mellon University

- **Slopes**
  - Decreases friction AND pulls car
- **Curves:**
  - Friction maintains centripetal force
  - Banking (superelevation)
    - – Reverse bank reduces normal force
- **Road surface condition**
  - Dry concrete μ = 0.75
  - Snow μ = 0.2 – 0.25
  - Ice μ = 0.1 – 0.15




 6

## Other Factors Affecting Brake Force

Carnegie Mellon University

- **Braking capability:**
  - Tire capability ("sticky" tires might have μ > 1)
  - Brake maximum friction (pad wear)
- **Equipment condition**
  - Tire condition: temperature, pressure, tread
  - Brake condition: hot, wet, damaged, …
  - Vehicle suspension, weight distribution, …
- **Braking controls**
  - Driver leg strength and willingness to brake hard
  - Braking assist force (multiplies driver leg strength)
- **Aerodynamics, suspension, debris, …**




7

## Epistemic Uncertainty – Vehicles

Carnegie Mellon University

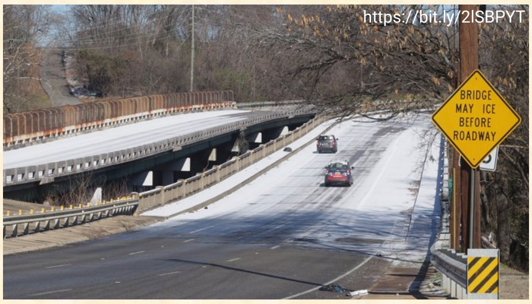
- **Own vehicle weak braking (less than expected)**
  - Brake wear & failures
  - Loss of brake assist
  - High tire pressure / bald tires
  - Brakes hot from recent use
  - Brakes wet from recent puddle
- **Other vehicle strong braking**
  - Braking capability for vehicle type
  - Aftermarket brake upgrade?
  - Aftermarket tire upgrade? Low tire pressure?
  - Leg strength of lead driver to press brakes?




8

## Epistemic Uncertainty – Environment

- Road surface of own vehicle
  - Might not be same as lead vehicle surface
- Road surface of lead vehicle
  - Might have dramatically different friction properties

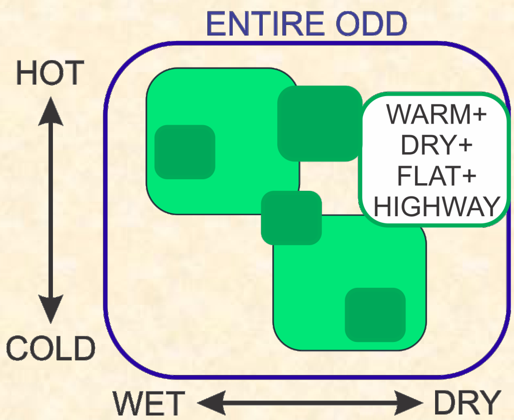






9

## Segmenting Into Micro-ODDs

- A single huge ODD leads to poor permissiveness
  - Want better performance on a warm dry day
- Approach: break up ODDs into pieces
  - Default cautious behavior
  - Prove safe trajectory for an ODD segment
  - Optimize segments based on customer value






10

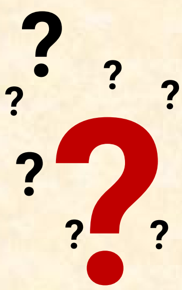
## Micro-ODD Benefits

- **Turns ODD growth on its head:**
  - Over time: Improve permissiveness for fixed ODD size
  - Operate across a diverse ODD safely (and cautiously!)
  - Incrementally improve performance in high value ODD segments
  - Use finer grain ODD segments for high value operational situations
    - Note: important to address transition between segments
- **References:**
  - Micro-ODD paper: https://arxiv.org/abs/1911.01207
  - ODD parameter paper: https://bit.ly/33K26uA
  - UL 4600
    - Sections 8.2 (ODD) & 8.8 (Trajectory & Control)


11

## Conclusions

- **Proofs are great, but rely upon assumptions**
  - In particular, about environment & behaviors
  - Permissiveness vs. safety tradeoffs
- **Proofs push uncertainty into the assumptions**
  - Uncertainty about own system
  - Uncertainty about other actor behaviors
  - Uncertainty about the environment

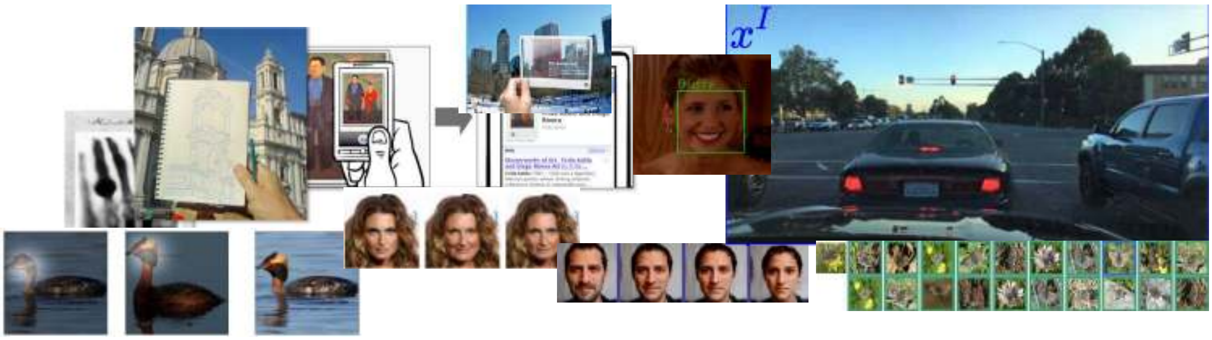


- **You might forget the edge cases…**

**… but they won't forget you!**


12

## Annex IV. Explainability methods for vision-based autonomous driving systems



# Explainability methods for vision-based autonomous driving systems

Matthieu Cord
Sorbonne University, valeo.ai
Joint work with Eloi Zablocki, Hedi Benyounes, Patrick Perez (valeo.ai)

## 01 Explainability of self-driving cars

arXiv > cs > arXiv:2101.05307

Computer Science > Computer Vision and Pattern Recognition

*[Submitted on 13 Jan 2021]*

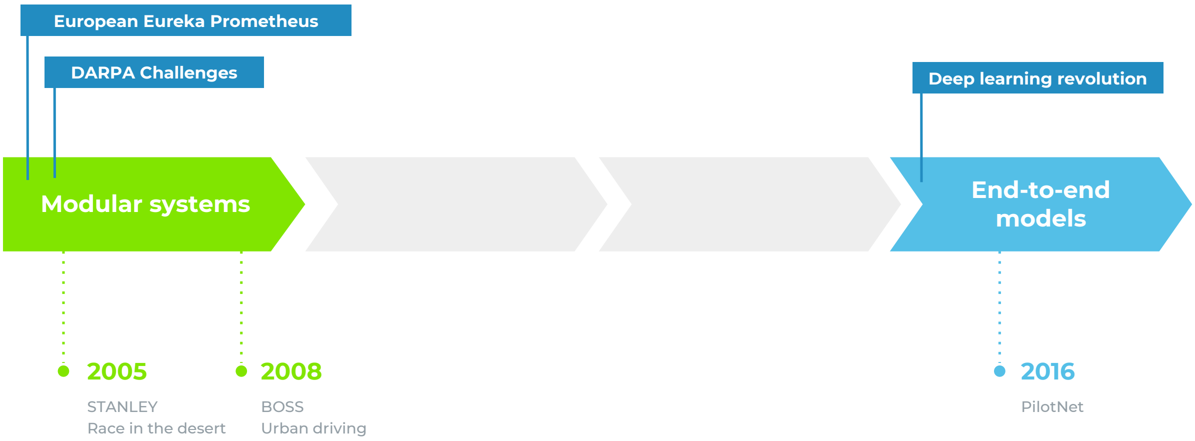
**Explainability of vision-based autonomous driving systems: Review and challenges**

Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, Matthieu Cord

***Under review (revision) IJCV journal***

3

## From historical modular pipelines to end-to-end learning models




6

## Explanations — Why? Who? What?

### Why?

**Societal** point-of-view:

- High-stake and safety critical
- Cannot test every situation then explanation

**System** point-of-view:

- poor performances: understand failure modes
- average performances: raise users' trust
- super-human performances: machine teaching

**Machine learning** point-of-view:

- training objectives are only proxies for real

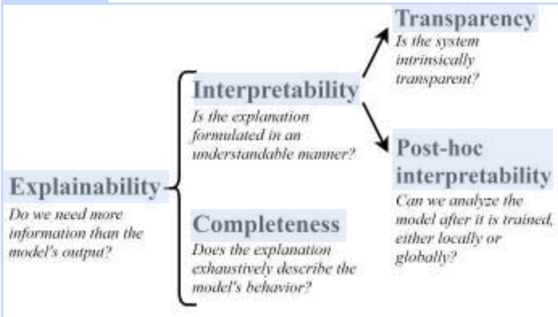

### Who?

**End-users and citizens** for trust

**Legal and regulatory bodies** for liability, accountability

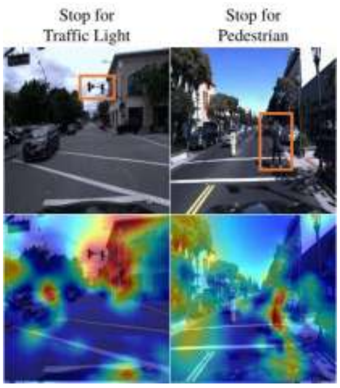
**Researchers and Engineers** for debugging, improving

### What?

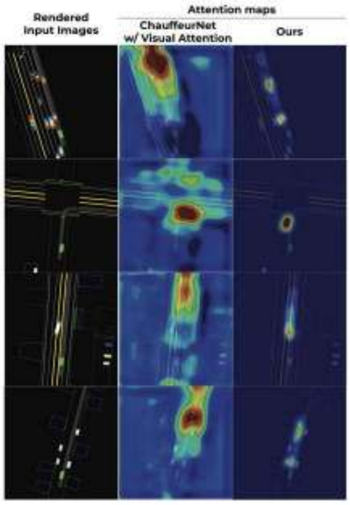


## Input saliency visualization - ***Post-hoc*** methods

(Bojarski et al. 2017)




(Liu et al. 2019)



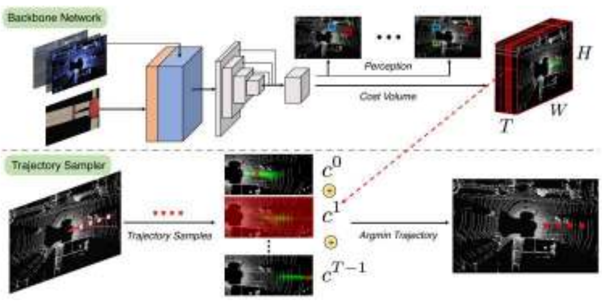

(Kim et al. 2021)

**Limits**

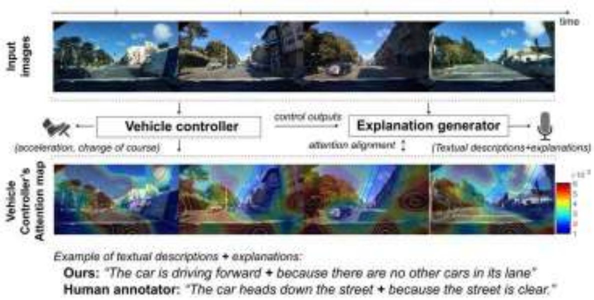
- Need to be interpreted
- Not well suited for human-machine interactions

12

## Provide intermediate representations - ***By design*** methods




(Zeng et al. 2019)




(Kim et al. 2017)


Explainability of vision-based autonomous driving systems: Review and challenges

**Limits**

- Need extra annotations
- The auxiliary tasks may hinder driving performances

13

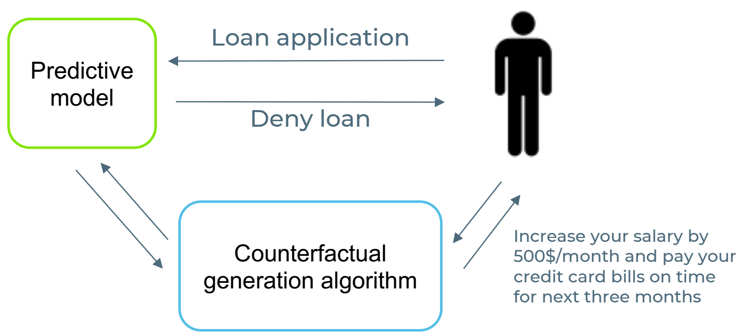
# **02** Post-hoc explainability STEEX model

**STEEX: Steering Counterfactual Explanations with Semantics**

Paul Jacob, Éloi Zablocki, Hédi Ben-Younes, Mickaël Chen, Patrick Pérez, Matthieu Cord

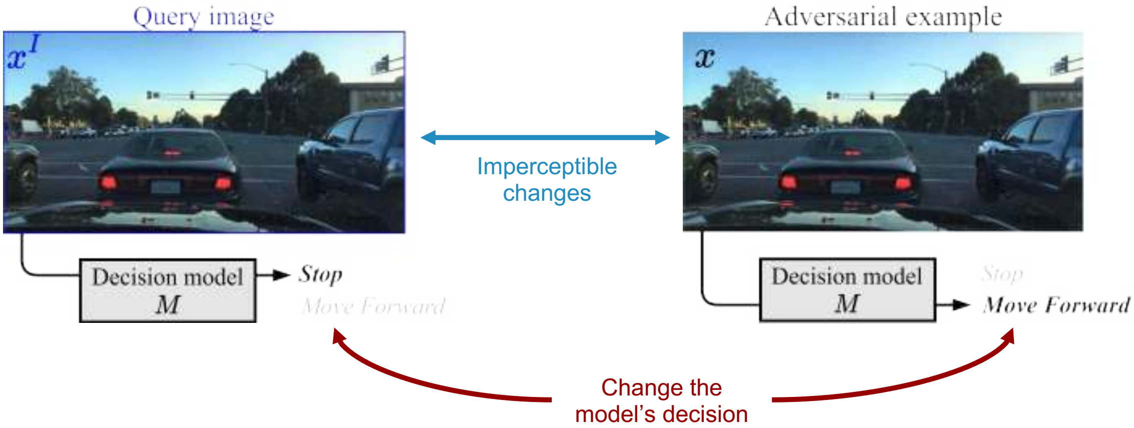
*Under review,* [code] github.com/valeoai/STEEX, [pdf] arxiv.org/abs/2111.09094 14

## Counterfactual explanations for classification models




A ***counterfactual explanation*** is a version of the input with **minimal and meaningful** perturbations that **changes the output decision** of the model (Wachter et al. 2017) 15

## Counterfactual explanations for <u>image</u> classification models?

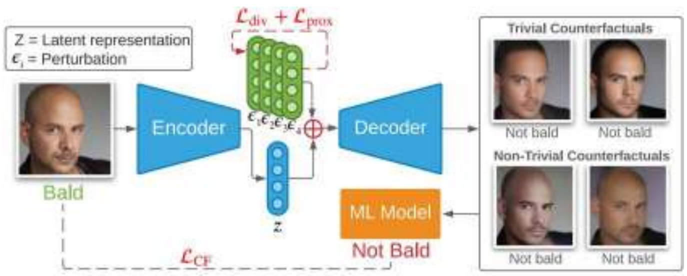



$$\text{argmin}_x \mathscr{L}_{\text{dist}}(x^I, x) + \lambda \mathscr{L}_{\text{decision}}(M(x), y)$$

19

## Counterfactual explanations for <u>image</u> classification models?


(Goyal et al. ICML'19)

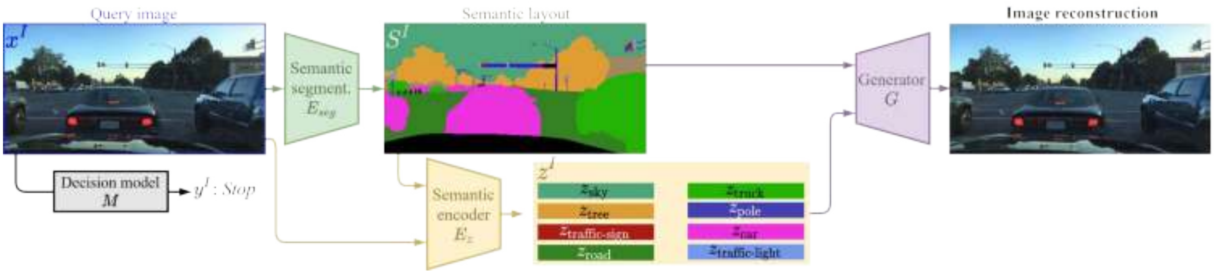



DiVE(Rodriguez et al. ICCV'21)

A ***counterfactual explanation*** is a version of the input with **minimal and meaningful** perturbations that **changes the output decision** of the model (Wachter et al. 2017)

21

## STEEX — Instantiation




$$\mathrm{argmin}_{\delta_z} ||\delta_z|| + \lambda \mathscr{L}_{\text{decision}}\left(M\left(G\left(z, S^I\right)\right), y\right)$$

**Generator G** and **Semantic encoder $E_z$** :

→SEAN (Zhu et al. 2020)

**Semantic segmentation $E_{seg}$** :

→DeepLabv3 (Chen et al. 2017)

**Losses**

$$\mathscr{L}_{\text{decision}}(M(G(z)), y) = -\,\mathrm{nll}(M(G(z)) \mid y)$$

$$||\delta_z|| = \sum_{c=1}^{N} \left|\left|\delta_z^{(c)}\right|\right|_2^2$$

24

## Datasets and classifiers

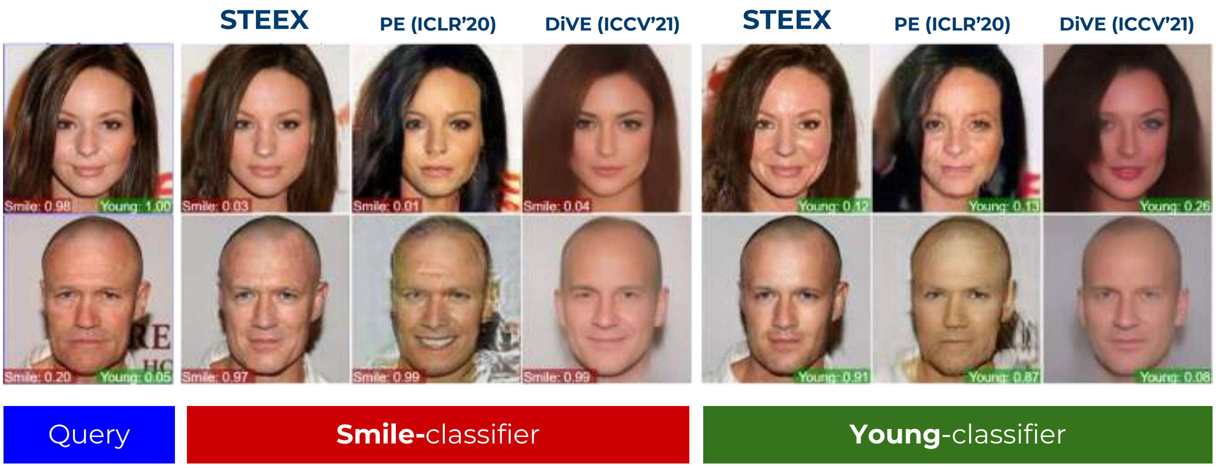
**CelebA** (128x128) SMILE-classifier YOUNG-classifier

**CelebAMask-HQ** (256x256) SMILE-classifier YOUNG-classifier


**BDD100k** (512x256) STOP-classifier

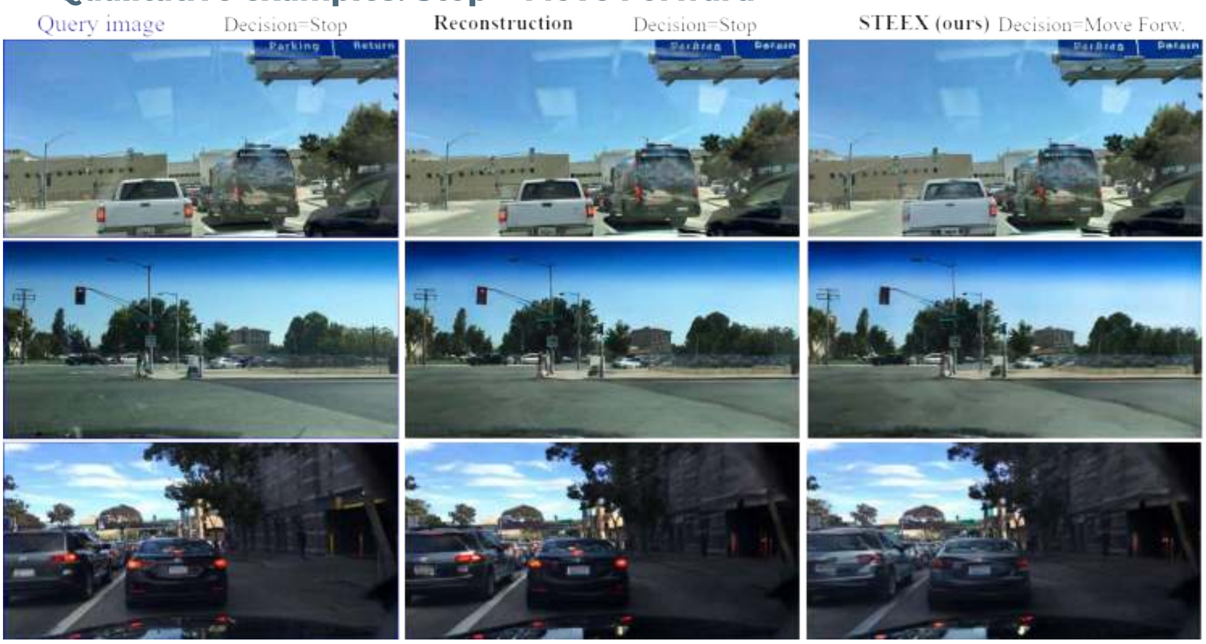
25

## Qualitative results on CelebAMask-HQ (256x256)




26

## Qualitative examples: Stop → Move Forward



## Quantitative results

**Perceptual quality:** *Are counterfactuals realistic?*

→ Fréchet Inception Distance (**FID**)

| FID↓ | CelebA (128x128) | | CelebAM-HQ (256x256) | | BDD-100k (512x256) |
|---|---|---|---|---|---|
| | Smile | Young | Smile | Young | Move For. |
| PE [39] | 35.8 | 53.4 | 52.4 | 60.7 | 141.6 |
| DiVE [34] | 29.4 | 33.8 | 107.0 | 107.5 | — |
| **STEEX** | **10.2** | **11.8** | **21.9** | **26.8** | **58.8** |

**Proximity:** *Is the identity preserved?*

→ Face Verification Accuracy **(FVA)**

| FVA↑ | CelebA | | CelebAMask-HQ | |
|---|---|---|---|---|
| | Smile | Young | Smile | Young |
| PE [39] | 85.3 | 72.2 | 79.8 | 76.2 |
| DiVE [34] | **97.3** | **98.2** | 35.7 | 32.3 |
| **STEEX** | 96.9 | 97.5 | **97.6** | **96.0** |

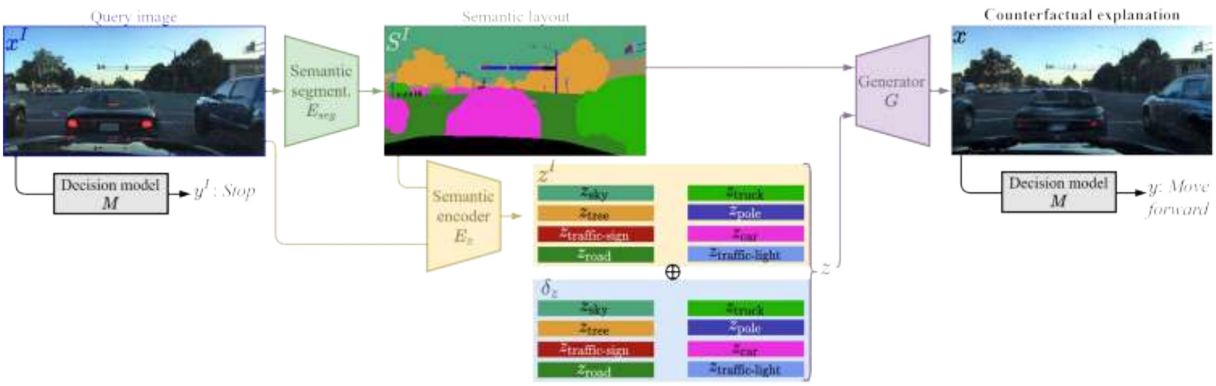
**Sparsity:** *How many facial attributes change?*

→ Mean Number of Attribute Changes (**MNAC**)

| MNAC↓ | CelebA | | CelebAMask-HQ | |
|---|---|---|---|---|
| | Smile | Young | Smile | Young |
| PE [39] | — | 3.74 | 7.71 | 8.51 |
| DiVE [34] | — | 4.58 | 7.41 | 6.76 |
| **STEEX** | **4.11** | **3.44** | **5.27** | **5.63** |

30

## Extension: region-targeted counterfactual explanation



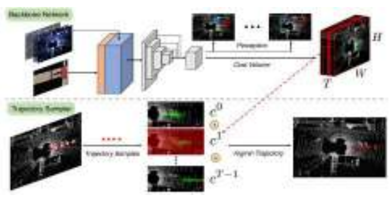

**New setup:** Let a user specify a set of semantic regions that the explanation must be about

32

## Region-targeted counterfactual explanations


## Conclusion for STEEX

**Saliency methods** are region-based (**WHERE?**)
**Counterfactual explanations** are content-based (**WHAT?**)



Neural motion planner (Zeng et al. 2019)

## Going further

→ Explore **more complex decision models**:
- e.g., trajectory forecasting, planning models

→ Allow the **modification of the semantic map**
- e.g., shift objects, add/remove objects...

Remove or shit pedestrian?

36

## 03 *By design* explainability
## BEEF model

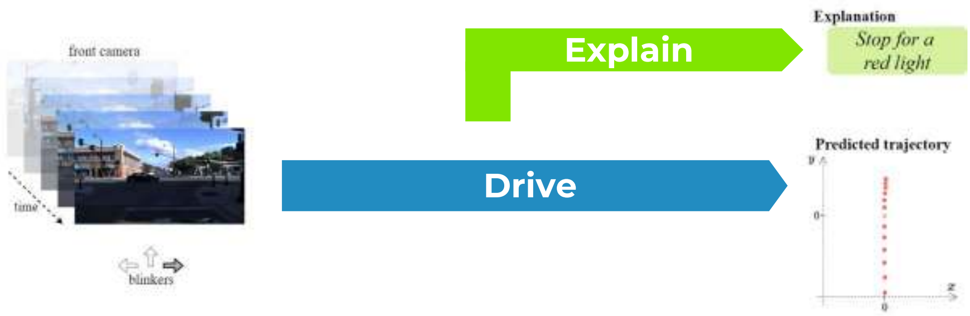
**BEEF: Driving Behavior Explanation with Multi-level Fusion**

Hédi Ben-Younes*, Éloi Zablocki*, Patrick Pérez, Matthieu Cord

*Pattern Recognition 2021*, [code] github.com/valeoai/BEEF, [pdf] arxiv.org/abs/2012.04983

37

### Overview of BEEF: BEhavior Explanation with multi-level Fusion

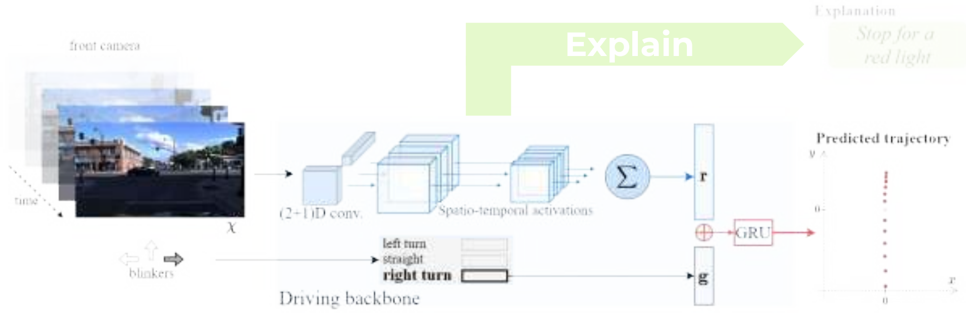



**Goal of this work**

Human-friendly explanations for the decisions of a neural driving system.

38

## BEEF: Self-driving 3D-conv backbone




**Visual encoder**

$$\mathbf{r} = \text{3DCNN}(\chi)$$

*Video input*

3DCNN = R(2+1)D (Tran et al. 2018)
5 residual blocks

**Trajectory prediction**

$$\left((\hat{x}_k, \hat{y}_k), \mathbf{h}_k\right) = \text{GRU}\left([\mathbf{r} \oplus \mathbf{g}], \mathbf{h}_{k-1}\right)$$

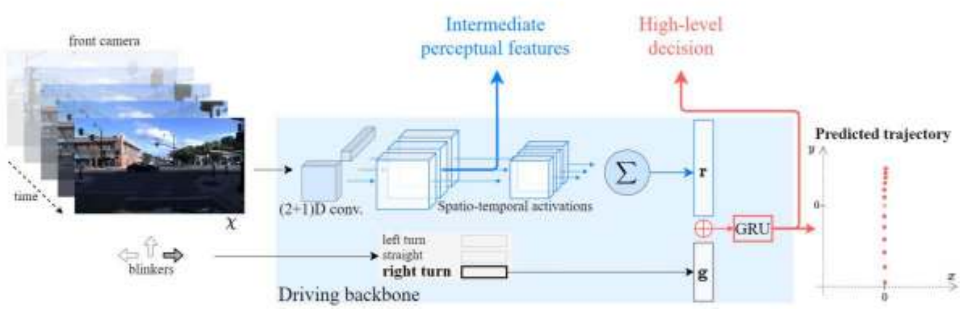
*Visual features* *Blinker signal*

**Imitation loss**

$$\mathcal{L}_{drive} = \sum_{k=1}^{K} \sqrt{(x_k - \hat{x}_k)^2 + (y_k - \hat{y}_k)^2}$$

*Ground-truth* *Trajectory prediction*

## BEEF: Explanation module overview




**Problem**

Different causes collapse to a same driving decision

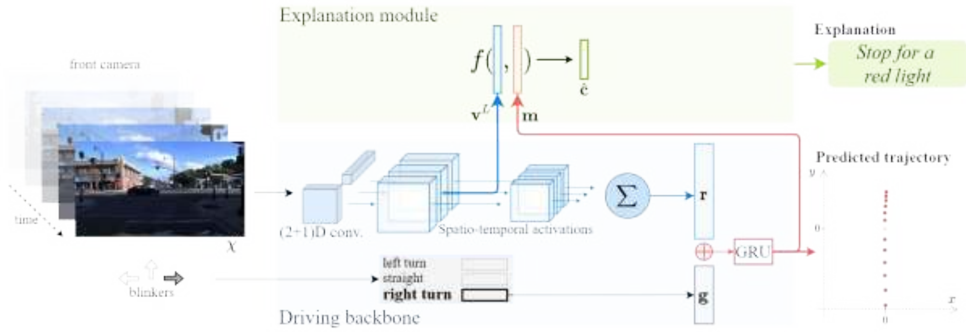
**Intuition**

→ Mid-level features contain perceptual information about the scene

→ High-level features contain information of the decision

40

## BEEF: Explanation module overview




**Fusion**

$$\hat{\mathbf{c}} = f(\mathbf{v}^L, \mathbf{m})$$

*Fusion function (details later)*

*Intermediate perceptual features*

*High-level decision*

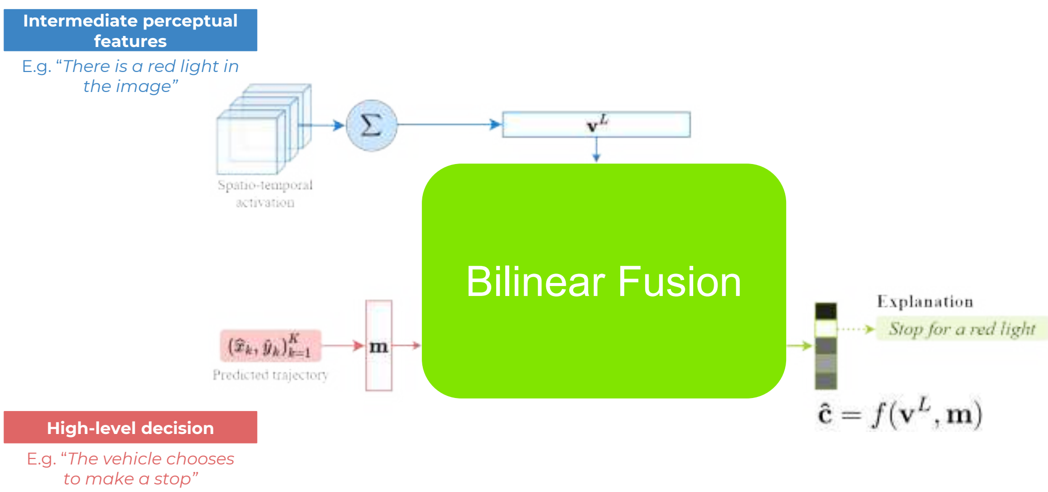
**Classification**

$$\mathbf{p} = \mathrm{softmax}(\hat{\mathbf{c}})$$

*Probability over classes*



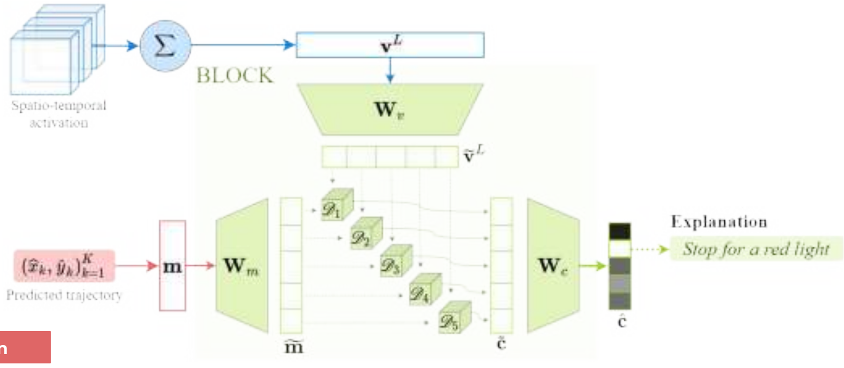
## BEEF: Multi-level fusion with BLOCK



## BEEF: Multi-level fusion with BLOCK

**Intermediate perceptual features**

E.g. *"There is a red light in the image"*



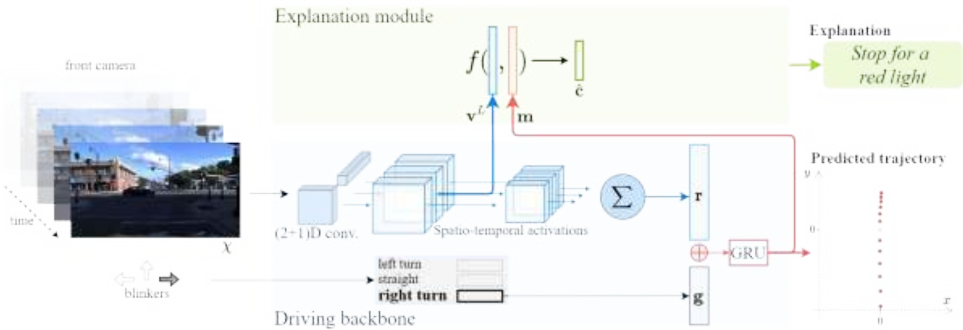

**High-level decision**

E.g. *"The vehicle chooses to make a stop"*

BLOCK (Ben-Younes et al. 2019)



## BEEF: Learning




**Hypothesis**

Mimicking driving behavior conserves explanations
→ Imitation learning for explanations

**Explanation loss**

$$\mathcal{L}_{explain} = -\log \mathbf{p}[c]$$

*Human-annotated explanation*

**Global loss**

$$\mathcal{L} = \mathcal{L}_{drive} + \alpha \mathcal{L}_{explain}$$

## Experiments: quantitative results on HDD

| System | Online/ Offline | Individual causes: Congest. | Sign | Red light | Crossing vehicle | Parked vehicle | Crossing pedestrian | Overall mAP | Driver MSE |
|---|---|---|---|---|---|---|---|---|---|
| *Action recognition (no driver)* | | | | | | | | | |
| CNN+Sens. (Ramanishka et al. 2018) | On. | 39.72 | 46.83 | 45.31 | — | 7.24 | 2.15 | 28.25 | × |
| I3D (Li et al. 2020) | Off. | 64.8 | 71.7 | 63.6 | 21.5 | 15.8 | 26.2 | 43.9 | × |
| I3D+GCN (Li et al. 2020) | Off. | 74.1 | 72.4 | 76.3 | 26.9 | 20.4 | 29.0 | 49.9 | × |
| *Driver only (no explanation)* | | | | | | | | | |
| Driver | On. | × | × | × | × | × | × | × | 1.33 |
| *Introspective explanation* | | | | | | | | | |
| Multi-head | On. | 81.25 | 66.59 | 75.46 | 31.21 | 10.24 | 25.62 | 48.39 | 1.36 |
| BEEF | On. | 80.38 | 63.41 | **81.94** | **41.19** | 12.18 | 27.19 | **50.96** | **1.33** |

Auxiliary branch at the output of the 3DCNN → Multi-head

**SOTA results**

Outperforming both online and *offline* models

**Slight drop on some classes**

Advantage of accessing future frames

**Complementarity of features**

→ Comparison to multi-head
→ Does not degrade driver MSE

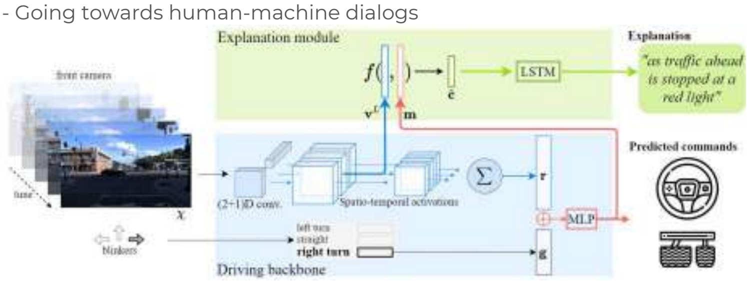
47

## Extension: natural language justifications

*E.g. "the car stops as traffic ahead is stopped at a red light"*

**Motivations:**

- Open-domain sentences convey finer and richer semantics than predefined classes
- Going towards human-machine dialogs




**Natural language**

Auto-regressive LSTM language model.

**End-to-end driving**

Predict driving commands (throttle and steering angle)

**Offline setup**

Produce justifications for temporal subsequences

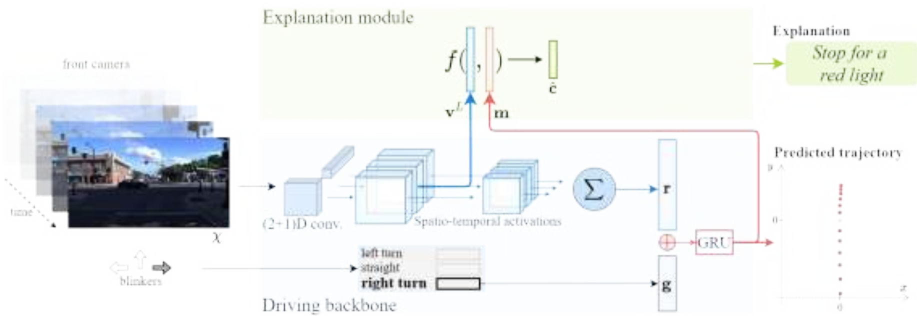
48

## Extension: qualitative results on BDD-X (77 driving hours)


49

## Conclusion for BEEF

1. BEEF is suitable for **real-world self-driving explanations**.
2. BLOCK fusion, originally developed to fuse multi-*modal* inputs, can be efficiently leveraged to **fuse multi-*level* inputs**.
3. New **SOTA results** on HDD and BDD-X.
4. **Flexible** approach (online/offline, cause classification/language generation)




50

**Explainability of vision-based autonomous driving systems: Review and challenges**

→ Éloi Zablocki*, Hédi Ben-Younes*, Patrick Pérez, Matthieu Cord

→ under review, arxiv.org/abs/2101.05307

**STEEX: Steering Counterfactual Explanations with Semantics**

→ Paul Jacob, Éloi Zablocki, Hédi Ben-Younes, Mickaël Chen, Patrick Pérez, Matthieu Cord

→ under review, github.com/valeoai/STEEX, arxiv.org/abs/2111.09094

**BEEF: Driving Behavior Explanation with Multi-level Fusion**

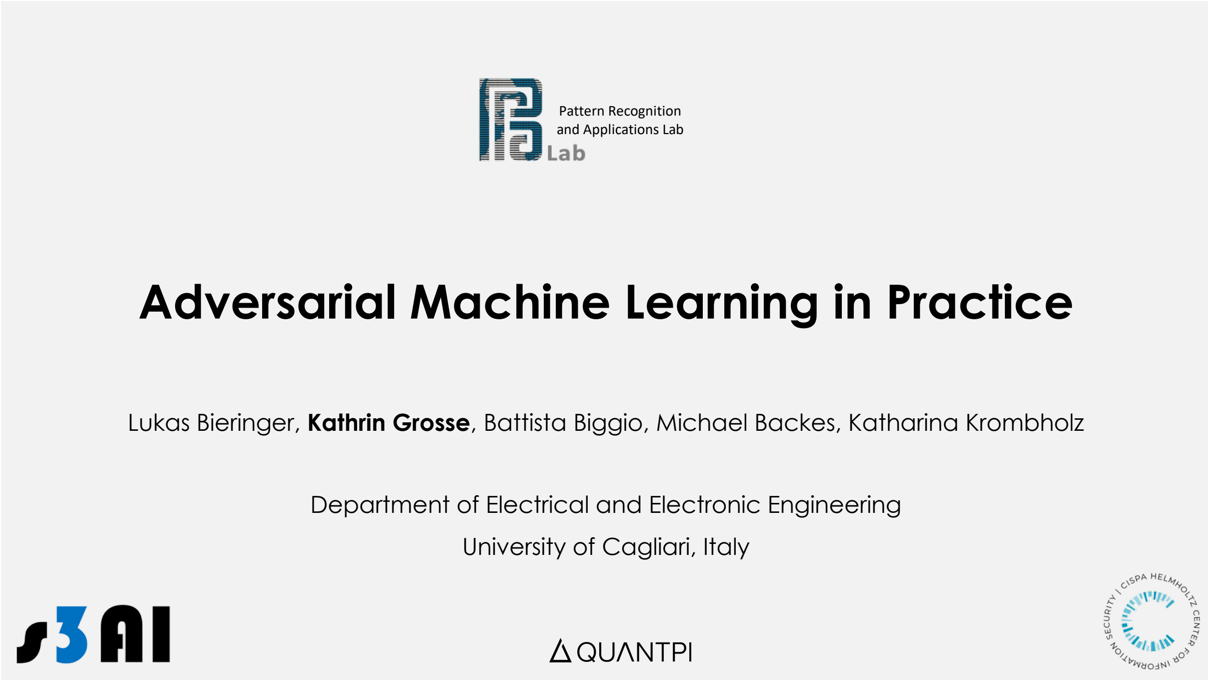
→ Hédi Ben-Younes*, Éloi Zablocki*, Patrick Pérez, Matthieu Cord

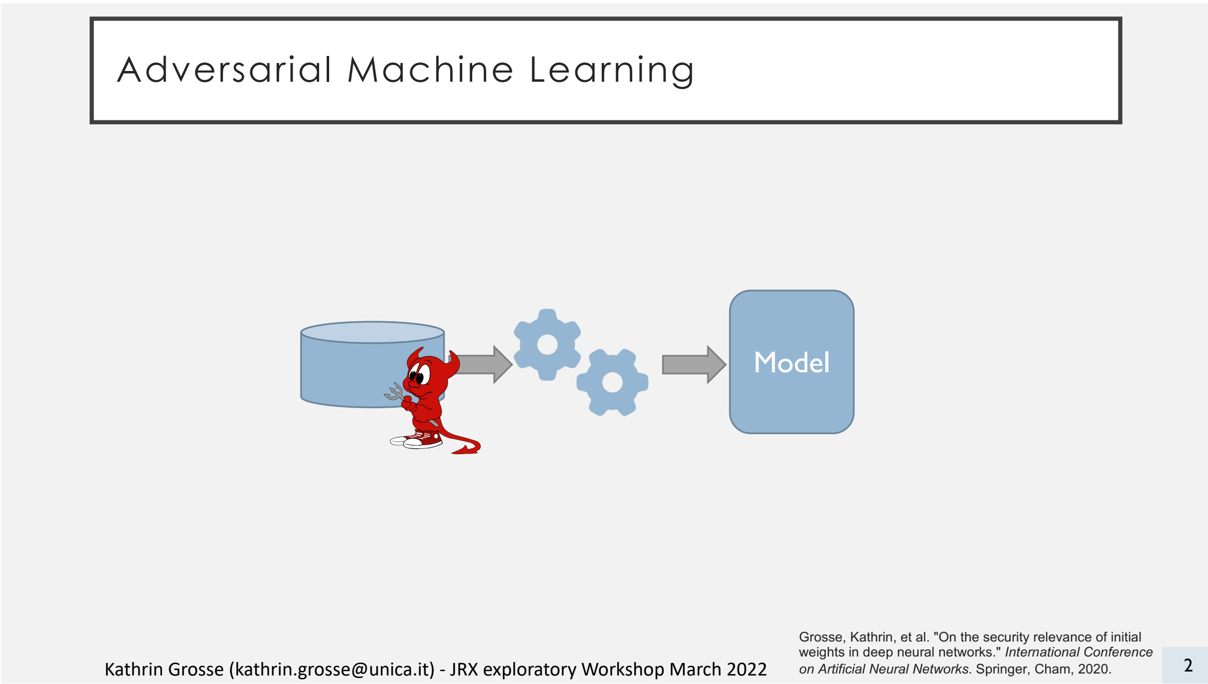
→ Pattern Recognition 2021, github.com/valeoai/BEEF, arxiv.org/abs/2012.04983

# Questions?

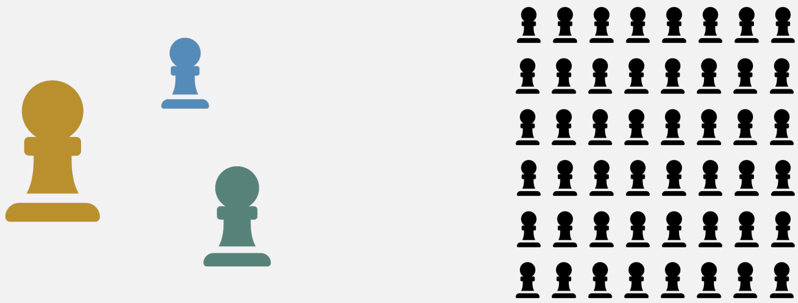
51

## Annex V. Adversarial ML in the Wild



# How to measure AML in practice?



Kathrin Grosse (kathrin.grosse@unica.it) - JRX exploratory Workshop March 2022

4

# Qualitative sample – 15 participants

- 14 male / 1 female
- Age: 34 (+/- 4.27)
- Employer: European start-ups (<200 employees)
- Application areas:
  - Cybersecurity, healthcare, vision, human resources…
- Position:
  - Managing (8), engineers (3), researchers (3)
- Education: PhD (9), MSc (4), BSc (1)

Kathrin Grosse (kathrin.grosse@unica.it) - JRX exploratory Workshop March 2022

Bieringer, Lukas, et al. "Mental Models of Adversarial Machine Learning." *arXiv preprint arXiv:2105.03726* (2021).

5

## Key findings – AML versus Non-AML Security


Research

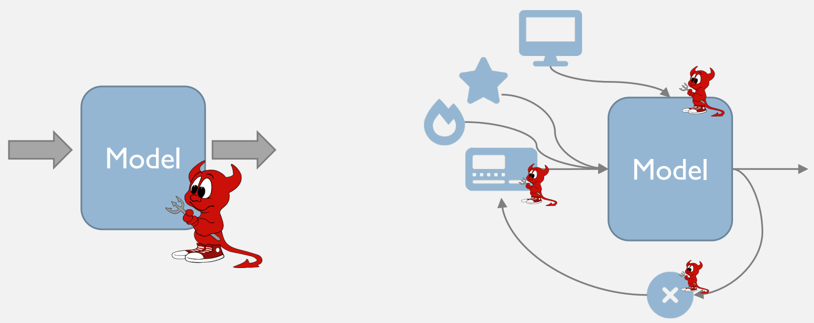
Practice



Bieringer, Lukas, et al. "Mental Models of Adversarial Machine Learning." *arXiv preprint arXiv:2105.03726* (2021).



## Key findings – Model versus Workflows




Research

Practice



Bieringer, Lukas, et al. "Mental Models of Adversarial Machine Learning." *arXiv preprint arXiv:2105.03726* (2021).

## Open questions


Application

perceived Relevance

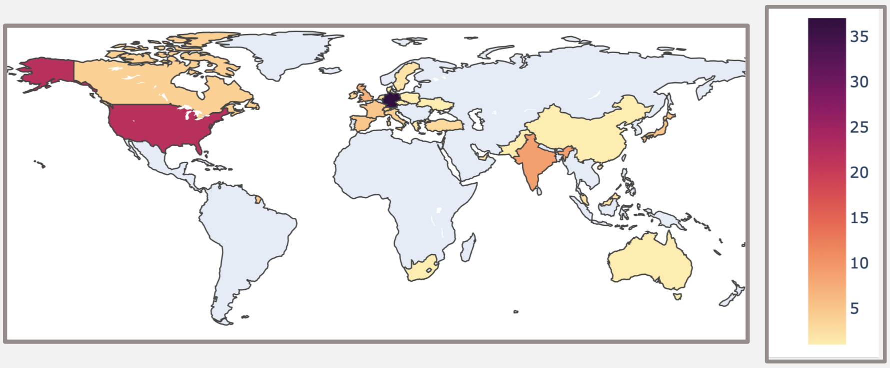
Education



Bieringer, Lukas, et al. "Mental Models of Adversarial Machine Learning." *arXiv preprint arXiv:2105.03726* (2021).



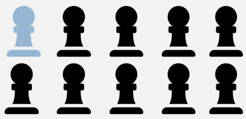
## Quantitative Sample – 140 participants






Forthcoming

## Key Findings – Encountered threats



- Privacy
- Poisoning
- Security
- Evasion
- Resource/Data theft
- Reverse Engineering

Kathrin Grosse (kathrin.grosse@unica.it) - JRX exploratory Workshop March 2022 Forthcoming 15

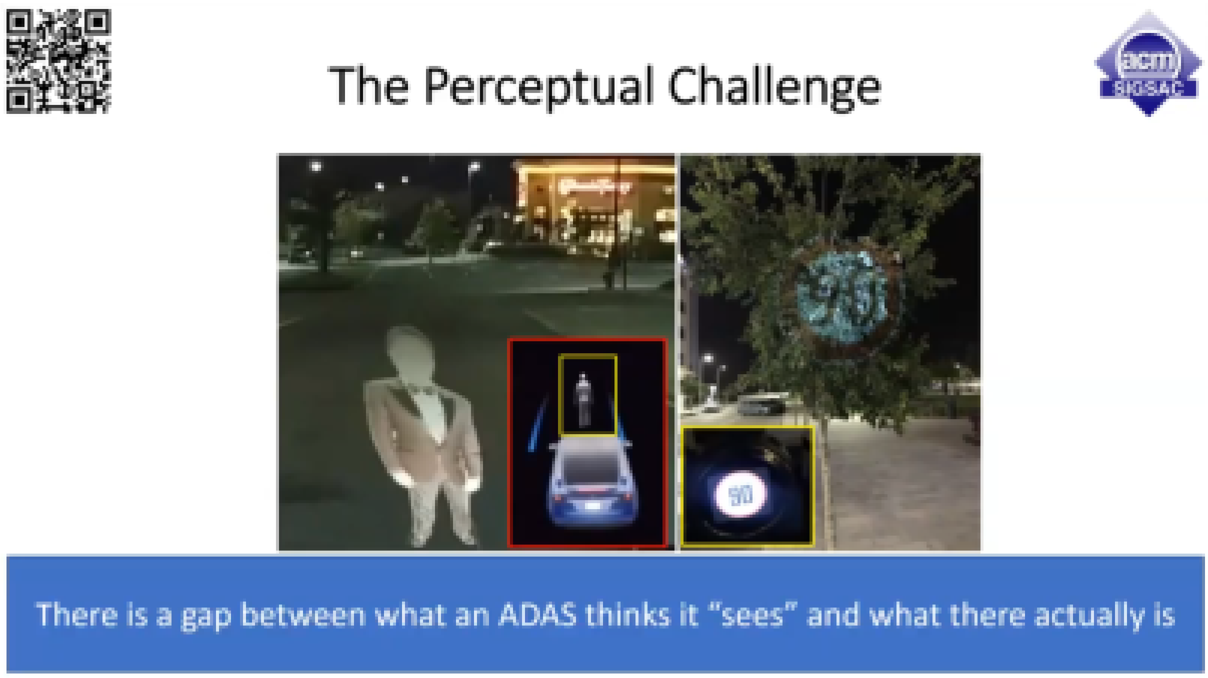
## Key Findings – Relevance

- Financial/Business Harm
- Wrong decision making
- Introduces bias
- Understand or encountered threat
- Loss of intellectual property
- ….

- Easy to spot/fix
- Other threat more likely
- Has not encountered threat
- Threat not relevant in setting
- Hard to do in practice
- ….

Kathrin Grosse (kathrin.grosse@unica.it) - JRX exploratory Workshop March 2022 Forthcoming 18

**Annex VI. Phantom of the ADAS: Securing advanced driver-assistance systems from split-second phantom attacks**

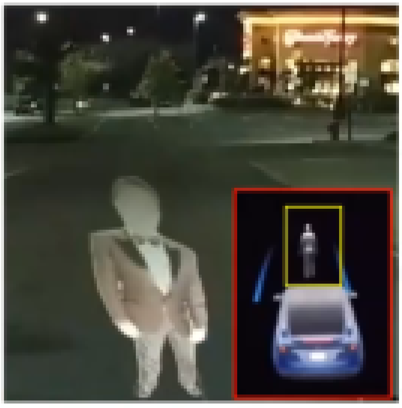
## The Perceptual Challenge

Fact 1: Telsa Model X (HW 2.5) considers the projected person a real obstacle.
Fact 2: Mobileye 630 PRO considers the projected road sign a real speed limit.

## What are Phantoms?

Phantom: is **depthless** presented/projected picture of a 3D object (e.g., pedestrian, car, truck, motorcycle, traffic sign).

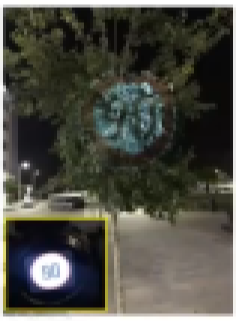
Purpose: to **fool** ADAS to treat the phantom as a real object and **trigger** an automatic reaction from the ADAS.

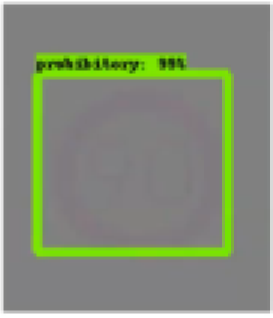


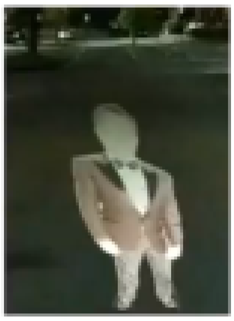
## Why phantoms are considered real objects by ADAS?

1. Object detectors are essentially feature matchers.
   They do not take into account the following aspects:

| Context | Color | Texture |
|---|---|---|
| Unrealistic object | Grey road sign | Transparent/skewed object |





## Split-Second Phantom Attack

A split second phantom attack is a phantom that appears for a few milliseconds and is treated as a real object/obstacle by an ADAS.

## Why phantoms are considered real objects by ADAS?

2. Disagreement between sensors: ADAS required to resolve a situation where there is a complete disagreement between sensors (strong validation from the camera and no validation from depth sensors) in real-time.

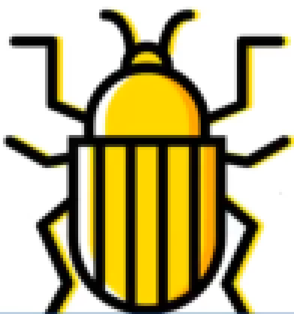
ADAS resolves this disagreement by trusting a single sensor. A result of:

- A programmed policy - "safety first for autonomous driving".
- Known physical limitations of sensors (e.g., changed accuracy in detecting objects during adverse weather/light condition).

## Why phantoms are considered real objects by ADAS?

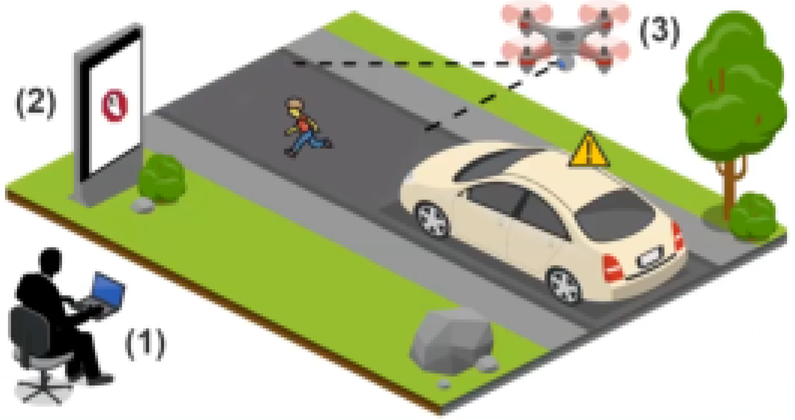
We do not consider phantoms a bug:

1. They are not the result of poor security implementation (e.g., SQL injection).
2. Phantoms exploit a fundamental inability of object detectors to distinguish between real and fake objects



Phantoms are scientific gap!

# Remote Threat Models




Threat Model 1: An attacker remotely hacks an Internet connected digital billboard and use it to present a phantom.
Threat Model 2: An attacker flies a drone equipped with a portable projector and project a phantom on a road, building, etc.
The phantom is perceived as a real object by nearby an ADAS and triggers an automatic unexpected reaction

# Threat Model's Significance w.r.t Related Works

**Previous Methods**

1. Necessitate that the attackers approach the attack scene.
2. Require skilled attackers.
3. Required full knowledge of the attacked model.
4. Leave forensic evidence in the attack scene.
5. Require complicated preparation .

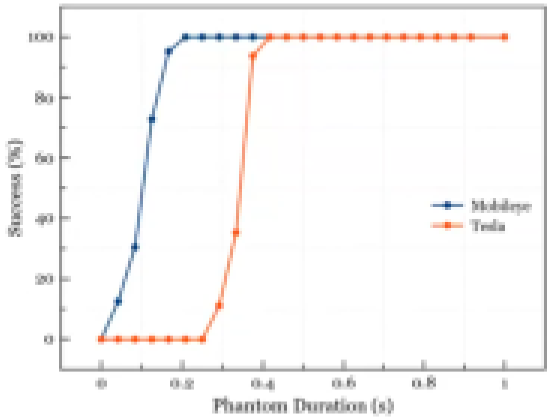
**Phantom Attacks**

1. Can be applied remotely.
2. Do not require any special expertise.
3. Do not rely on white-box approach.
4. Do not leave any evidence at the attack scene.
5. Do not require any complex preparation.

# Analysis - The Influence of Duration of the Phantom on the Success Rate

For Tesla (HW 3)

For Mobileye 630




1. Mobileye 630 detects a phantom that appears for 125 ms for 100% of the time.
2. Tesla (HW 3) detects a phantom that appears for 416 ms for 100% of the time.

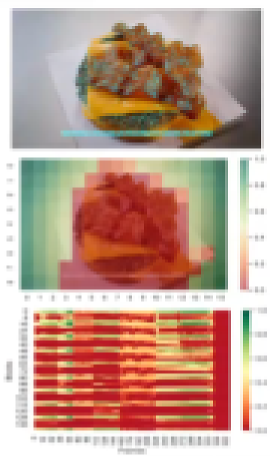
# Demonstration of the Attack via Digital Billboards

<u>Algorithm for embedding a phantom in an advertisement</u>

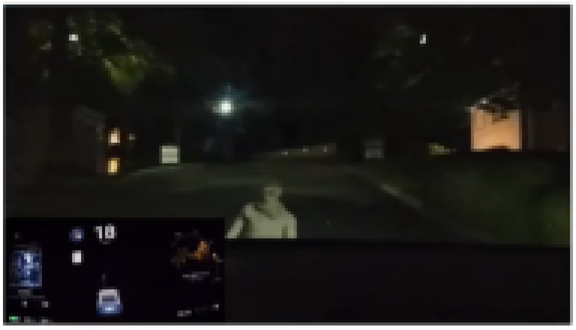
1) A **local score** of a block *b* in a frame *f* is computed as follows:

   i. Key-points in *f* are extracted
   
   ii. The score for block *b* is computed based on how much a dead area the block is (with respect to the extracted key-points).

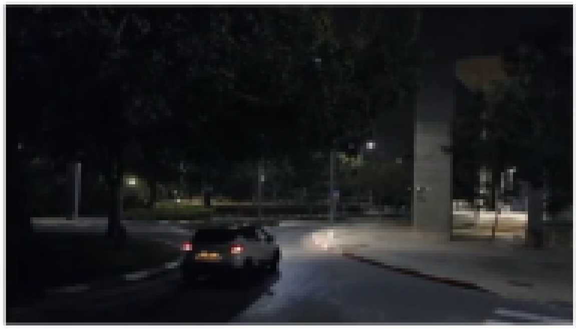
2) A **global score** of a block b is computed with respect to its score in the next consecutive frames.



# Attacking ADAS via a Projector



Attacking Tesla Model X (HW 2.5)



Attacking Mobileye 630 via a drone

A video of the attack was uploaded

# Attacking Tesla via a Digital Billboard


A phantom road sign is embedded to 500 ms


The autopilot of Tesla Model X (HW 3) automatically triggers the car to stop

A video of the attack was uploaded

## Countermeasure - GhostBusters

Video Camera → Object Detector → GhostBusters

- ❑ A software module
- ❑ GhostBusters is used for validation – it used to determine whether a detected object is phantom or real.

## Countermeasure - GhostBusters

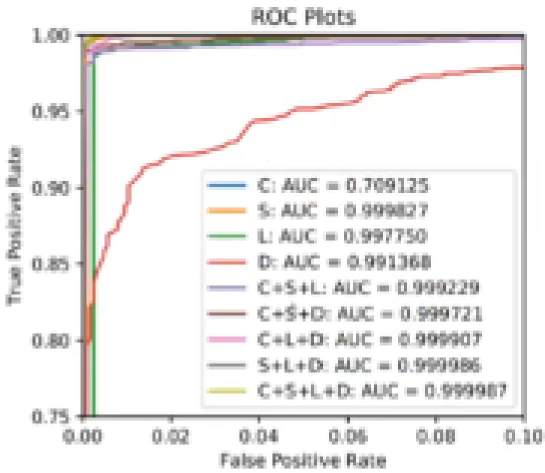
Architecture: the countermeasure consists of five CNNs.

1) Four trained CNNs used to detect phantoms based on:

- Context.
- Surface.
- Light.
- Depth.

2) Ensemble layer- A trained CNN used to detect phantoms based on the embeddings of the other four CNN.

# Countermeasure - GhostBusters




Datasets and models can be downloaded online

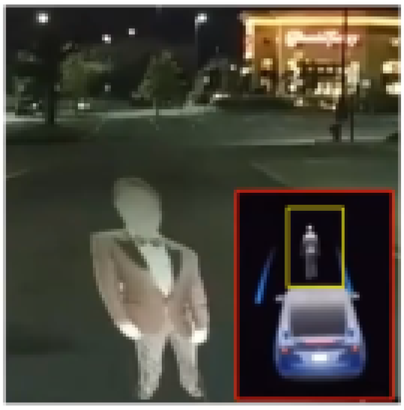
GhostBusters reach an AUC of 0.99 AUC

# Countermeasure - GhostBusters

**Table 4: Detection Rates Using s.o.t.a traffic sign Detectors**

| | | | *Attack Success Rate* | | |
|---|---|---|---|---|---|
| | | | Countermeasure | | |
| | | | With | | Without |
| | | *Threshold:* | *@0.5* | *@[TPR=0]* | |
| Sign Detector | [52] | faster_rcnn_inception_resnet_v2 | 0.098% | 0.294% | 87.16% |
| | [52] | faster_rcnn_resnet_101 | 0.098% | 0.588% | 96.08% |
| | [52] | faster_rcnn_resnet50 | 0.098% | 0.588% | 81.29% |
| | [28] | faster_rcnn_inception_v2 | 0.098% | 0.588% | 93.05% |
| | [13] | rfcn_resnet101 | 0.098% | 0.588% | 99.71% |
| | [25] | ssd_inception_v2 | 0.0% | 0.294% | 81.98% |
| | [24] | ssd_mobilenet_v1 | 0.098% | 0.588% | 83.45% |

GhostBusters reduce the attack success rate when it was applied to s.o.t.a object detectors from 99.7-81.2% without our module to 0.01%

Question: Who is the person behind the phantom that Tesla detects?

# AML attacks in the physical domain
# The Translucent Patch: A physical and Universal Attack on Object Detectors

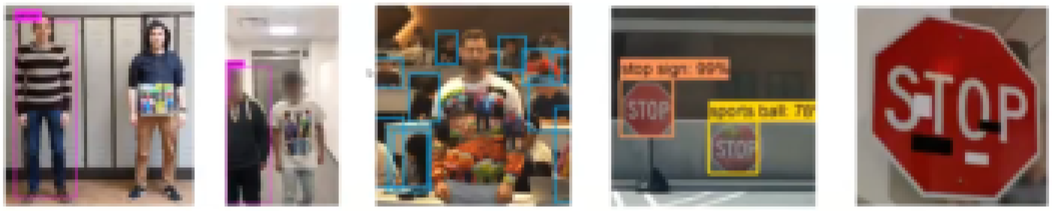
CVPR, 2021

Alon Zolfi, Yuval Elovici, Asaf Shabtai

**Prof. Asaf Shabtai**
Dept. of Software and Information Systems Engineering
Ben Gurion University

# Related Work

- Attacks that target object detection models using a perturbation applied on the attacked object:
  - Requires direct access to the object of interest.
  - To manipulate multiple objects a perturbation must be applied to each one.




Images taken from [9],[10],[11],[8],[9] (left to right)

CBG

# Related Work

- A single attack that target image classification models using a perturbation applied on the sensor:
  - Image classification models are not as complex as object detection models (many candidate bounding boxes priors that need to be attacked simultaneously)
  - Did not consider the impact on other objects

Image taken from [13]

| | Li et al. 2019 | Ours |
|---|---|---|
| Model Task | Image Classification | **Object Detection** |
| Attacked Object | Single object at the center of the image | **Multiple objects at different locations and sizes** |
| Preserve detections of untargeted classes | No | **Yes** |
| Shape radius | Fixed size | **Dynamic** |
| Shape color | Small set of predefined colors | **Dynamic** |

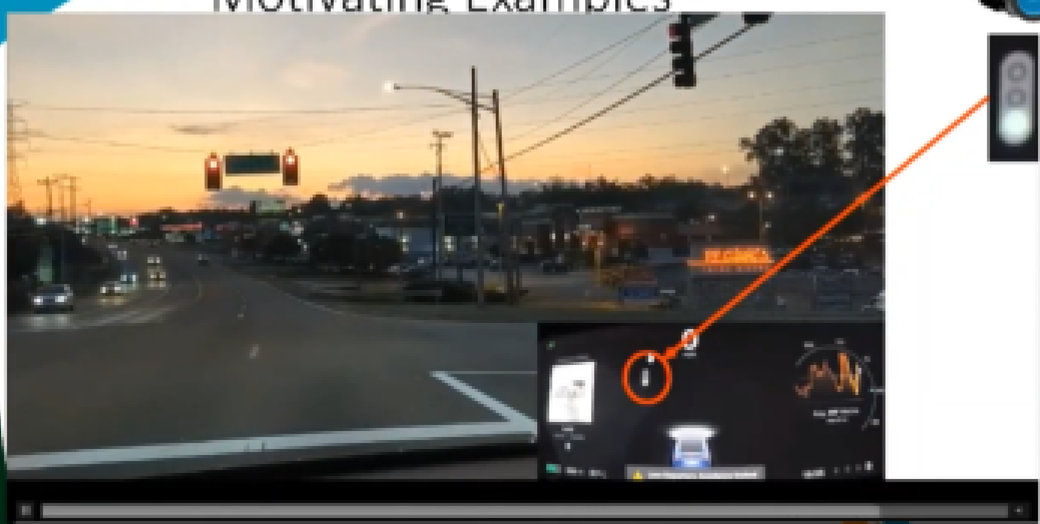
CBG

# Research Goal

- Create an **end-to-end attack** in the form of a printable translucent adversarial perturbation that will be placed on the sensor
- **Consistently** deceive DNNs **object detection** mechanism
- **Unnoticeable** and with **minimal impact** on the DNN-based object detection model
- **Robust perturbation** for multiple scenarios under real-world constraints

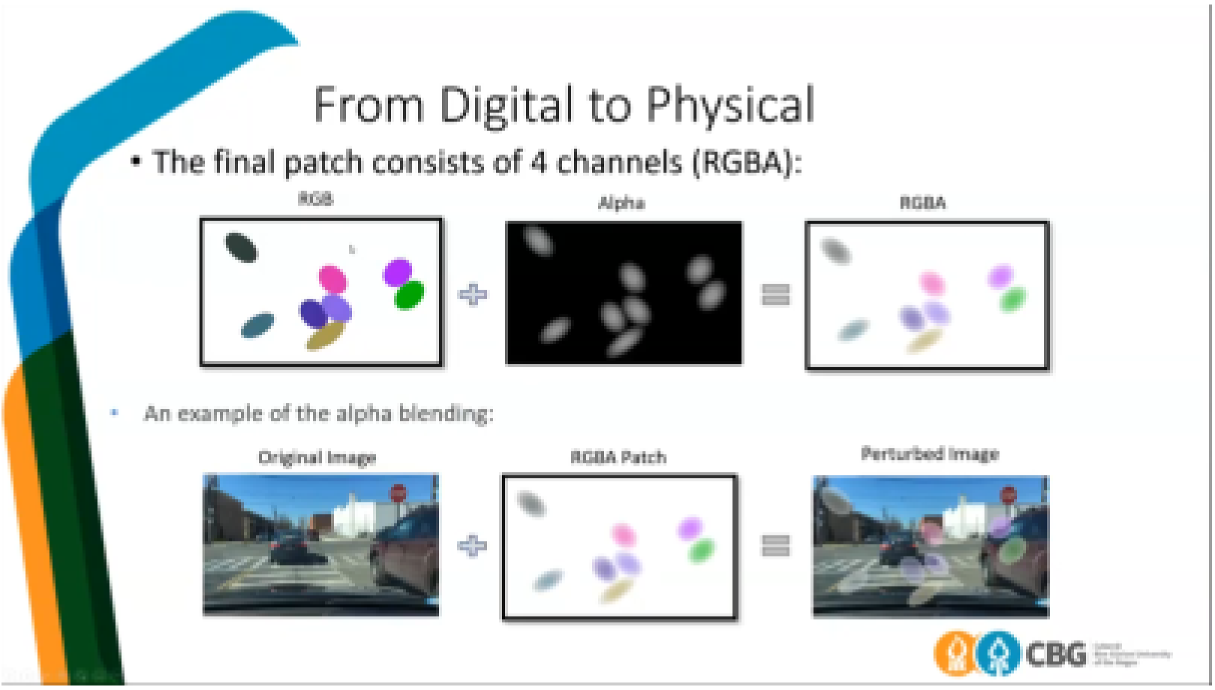
CBG

# Motivating Examples

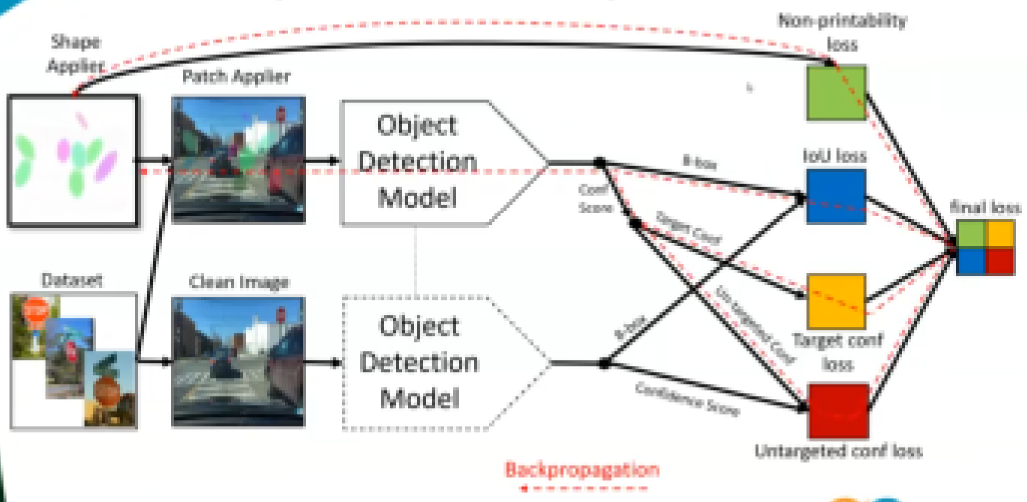


CBG


Challenges
State-of-the-art adversarial attacks craft pixel-level perturbations, which is impossible and not practical for our case (attaching the perturbation onto the camera's lens)
State-of-the-art adversarial attacks do not consider real-world constraints:
How to digitally simulate patch overlay on the sensor
How well can a printer represent digital colors
CBG


From Digital to Physical
The final patch consists of 4 channels (RGBA):
RGB
Alpha
RGBA
An example of the alpha blending:
Original Image
RGBA Patch
Perturbed Image
CBG

# Proposed Attack Pipeline




# Evaluation

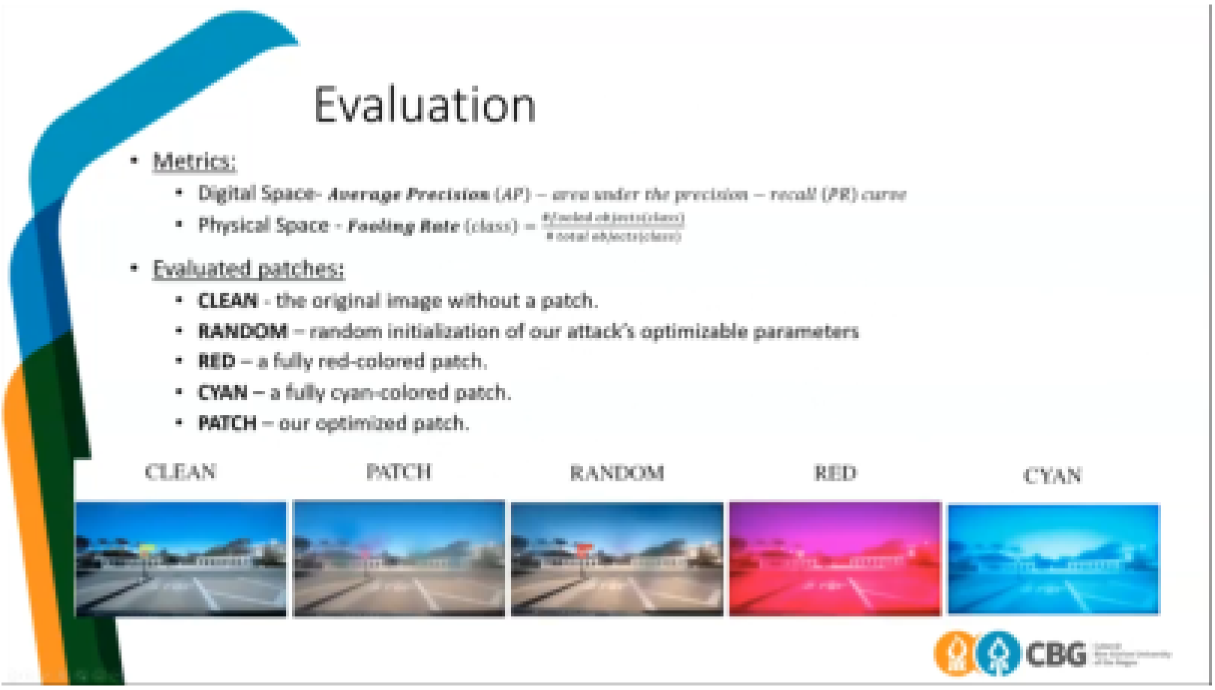
- We evaluate our attack on the use case of autonomous cars, trying to eliminate the detection of **stop signs**
- We conduct experiments both in the digital and physical space:
  - Digital space – using the alpha blending process
  - Physical space – printing the optimized patch on a translucent paper
- We perform experiments both in white-box setting (model used for training and testing is the same) and black-box setting (patch is trained on one model and tested on others)

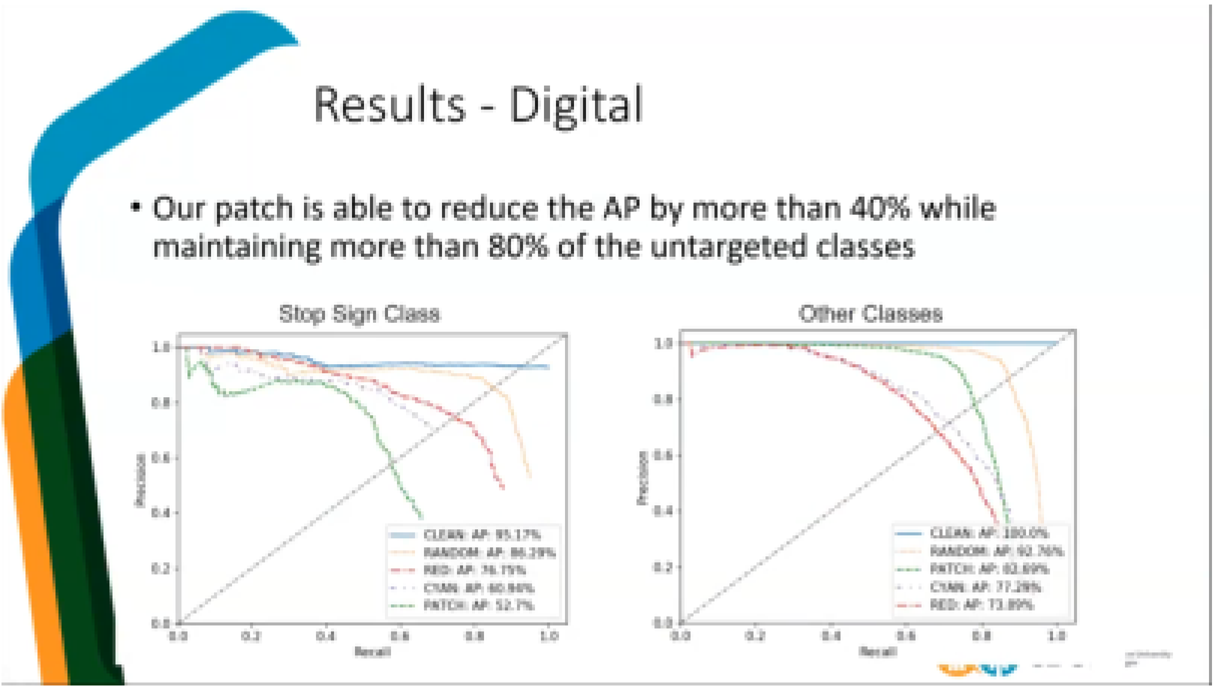
## Evaluation – Models

- Three different models are used:
  - You Only Look Once (YOLO) – **YOLOv5** one stage detector is used to train the patch parameters and is evaluated in a white-box setting.
  - To examine the patch's transferability to other models we use **YOLOv2** and **Faster R-CNN** (two-stage detector).
  - The models are pre-trained on the **MS-COCO** dataset (80 classes).
  - We use **eight** relevant classes: person, bicycle, car, bus, truck, traffic light, fire hydrant, and stop sign.

## Evaluation – Datasets

- Three different datasets are used:
  - Berkley Deep Drive (**BDD**) - videos of the driving experience covering many different times of the day, weather conditions, and driving scenarios, ~500 stop sign images
  - Mapillary Traffic Sign Dataset (**MTSD**) – a diverse street-level dataset obtained from a rich geographic area, ~750 stop sign images
  - **LISA** traffic sign dataset – videos split into frames containing U.S. traffic signs, ~500 stop sign images
- **BDD** and **MTSD** are used for training, **LISA** is used for testing


Evaluation
• Metrics:
• Digital Space- Average Precision (AP) – area under the precision – recall (PR) curve
• Physical Space - Fooling Rate (class) = #fooled objects(class) / # total objects(class)
• Evaluated patches:
• CLEAN - the original image without a patch.
• RANDOM – random initialization of our attack's optimizable parameters
• RED – a fully red-colored patch.
• CYAN – a fully cyan-colored patch.
• PATCH – our optimized patch.
CLEAN
PATCH
RANDOM
RED
CYAN
CBG

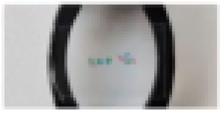

Results - Digital
• Our patch is able to reduce the AP by more than 40% while maintaining more than 80% of the untargeted classes
Stop Sign Class
Other Classes
Precision
Recall
CLEAN: AP: 95.17%
RANDOM: AP: 86.29%
RED: AP: 76.75%
CYAN: AP: 60.84%
PATCH: AP: 52.7%
CLEAN: AP: 100.0%
RANDOM: AP: 92.76%
PATCH: AP: 82.69%
CYAN: AP: 77.29%
RED: AP: 73.89%

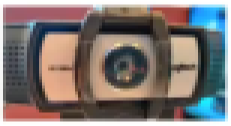
# Physical Attack Setup

- We print the digital patch onto a translucent paper
- The patch is applied to the camera's lens
- We project videos on a computer screen to simulate a driving scenario

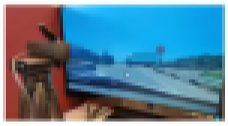
Digital Patch

Physical Patch





Patch applied to camera



Camera filming screen

# Results - Physical

- We project 48 different videos from the **LISA** dataset varying in:
  - Time of the day (day, night)
  - Scene (urban, rural)
  - Lighting conditions (light, partial/full shadow)
- The results show that our attack has successfully eliminated 42.27% of the stop signs while maintaining 80% of the other classes:

| Class/Attack | PATCH | RANDOM | RED | CYAN |
|---|---|---|---|---|
| Stop sign | 42.27% | 20.57% | 93.3% | 98.9% |
| Others | 21.54% | 19.27% | 82.7% | 81.6% |

Table 4: Fooling rate for the stop sign class and other classes for physical patch attacks

## Annex VII. Safe Motion Planning among Decision-Making Agents


### Motion planning among decision-making agents

Starship

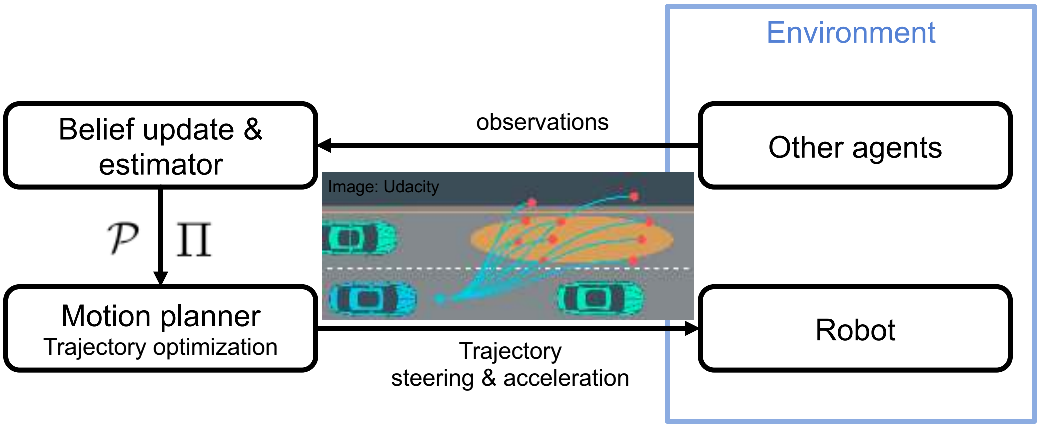
Spencer – robot airline assistant

Tesla

Limited interaction, safety and social compliance

1

## Autonomous vehicles




$\mathcal{P}$ : Weight of each motion hypothesis, $\Pi$ : Plans of all other agents,

*W. Schwarting et al, "***Planning and Decision-Making for Autonomous Vehicles***", Annual Review of CR&AS, 2018*
*E. Paden et al, "***A survey of motion planning and control techniques for self-driving urban vehicles***", IEEE T-IV, 2016*

2

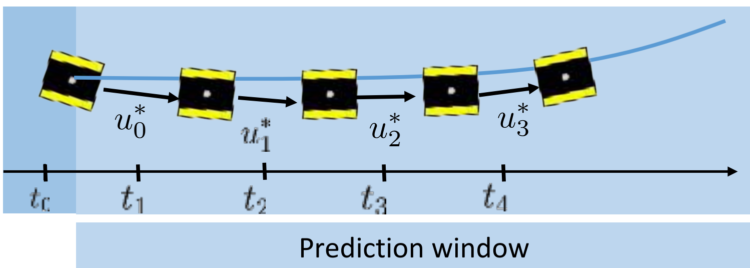
## Receding-horizon Trajectory Optimization

Often refered to as Model Predictive Control (MPC)

- Prediction based on kinematic/dynamic model
- Define the cost per timestep
- Sum up costs to be minimized
- Add constraints

- Solve constrained optimization using numerical optimization

- Apply first optimal input $u_0^*$
- Repeat




$$\arg\min_{\mathbf{x},\mathbf{u}} \sum_{k=0}^{N-1} J_k(\mathbf{x}_k, \mathbf{u}_k) + J_N(\mathbf{x}_N)$$

$$s.t. \quad \mathbf{x}_{k+1} = f(\mathbf{x}_k, \mathbf{u}_k) \qquad \mathbf{x}_k \in \mathcal{X}_k^{free}$$

Vehicle model  Collision avoidance

8

## Receding-horizon Trajectory Optimization

Non-convex optimization, efficiently solved with Acado/ForcesPro


$$\arg\min_{\mathbf{x},\mathbf{u}} \sum_{k=0}^{N-1} J_k(\mathbf{x}_k, \mathbf{u}_k) + J_N(\mathbf{x}_N) \quad s.t. \quad \mathbf{x}_{k+1} = f(\mathbf{x}_k, \mathbf{u}_k) \quad \mathbf{x}_k \in \mathcal{X}_k^{free}$$

N. D. Potdar, et al., "***Online Trajectory Planning and Control of a MAV Payload System in Dynamic Environments***", Autonomous Robots,. 2020
B. Brito et al, "**Model Predictive Contouring Control for Collision Avoidance in Unstructured Dynamic Environments**" *RA-L, 2019*
L. Ferranti, et al, "**SafeVRU : A Research Platform for the Interaction of Self-Driving Vehicles with Vulnerable Road Users**" *IEEE IV, 2019*

9

## Why trajectory optimization?

MPC allows us to consider:

- Multiple objectives
- Vehicle dynamics & obstacle prediction models
- Constraints → Safety encoded and checked for explicitly

→ Flexible & powerful framework

Limitations:

- Deterministic formulation
- No interaction with other agents
- Local method
- Hand-tuned complex cost function

**Challenge 1: Uncertainty**

**Challenge 2: Interaction**

Leverage Learning and MPC

10

## Safety disclaimers: Trajectory optimization (MPC)

Constraints → Safety encoded and checked for explicitly

However:

Convex problem → We find feasible & optimal solution

Non-convex problem → Depends on the solver, but in general, we may not have guarantees that a feasible and (locally) optimal solution is found within the allocated time. We may need a “back-up” strategy.

→ Our problem is non-convex!!!

Guarantees up to the horizon → Need for recursive feasibility

Safe if models are accurate! → We recompute at high frequency

11

## Challenge 1: Uncertainty

Probability of collision below a specified threshold

$$\begin{aligned} \min_{\hat{x}_{1:N}, u_{0:N-1}} \quad & \sum_{k=0}^{N-1} J_k(\hat{x}_k, u_k) + J_N(\hat{x}_N) \\ \text{s.t.} \quad & \hat{x}_0 = \hat{x}(0), \\ & \hat{x}_{k+1} = f(\hat{x}_k, u_k), \\ & \boxed{\Pr(x_k \in \mathcal{X}_k^{\text{free}}) \geq 1 - \epsilon,} \end{aligned}$$

$$x_k \sim \mathcal{P}$$

$$\hat{x}_k = E[\mathcal{P}]$$

Probabilistic avoidance

probability threshold

Solutions:

- Ignore uncertainty: deterministic problem with mean values & quick replanning
- Conservative: enlarge robots’ volume with their 3-sigma confidence ellipsoids
- Solve with chance-constraints or scenario-based MPC

*H. Zhu et al., **”Chance-constrained Collision Avoidance for MAVs in Dynamic Environments”,** RA-L 2019*

O. de Groot, et al., *“***Scenario-Based Motion Planning in Uncertain Dynamic Environments***”,* IEEE Robotics and Automation Letters 2021

13

## Safety disclaimers: Chance-constrained MPC

How do we define the probability threshold?

→ Very small = conservative behavior

Real-time probabilistic motion planning methods are at their "infancy"

→ Mostly deterministic approximations employed

14

## Challenge: Interaction


**Core skills:**
- Understand people's intentions
- Read subtle social cues
- Implicitly communicate own intentions
- Execute safe motions

Video courtesy of the Intelligent Vehicles group TU Delft - Driven by a human

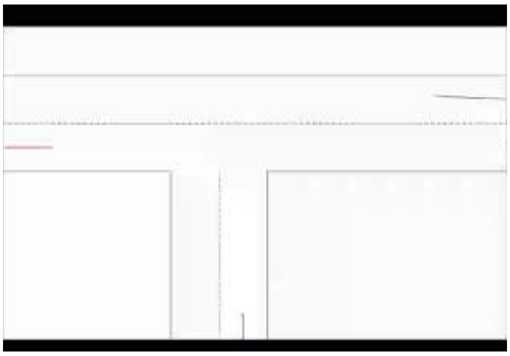
16

## Interaction through communication

Robots communicate their plans & iterate to agree on collision-free plans

- Distributed Nonconvex Model Predictive Control (D-NMPC)

Very large communication & computation effort!
+ not all will communicate + hacking….



*L. Ferranti et al, "***Coordination of Multiple Vessels Via Distributed Nonlinear Model Predictive Control** *" ECC, 2018*

17

## Interaction without communication


**Core skills:**

- Understand people's intentions
- Read subtle social cues
- Implicitly communicate own intentions
- Execute safe motions

Video courtesy of the Intelligent Vehicles group TU Delft - Driven by a human

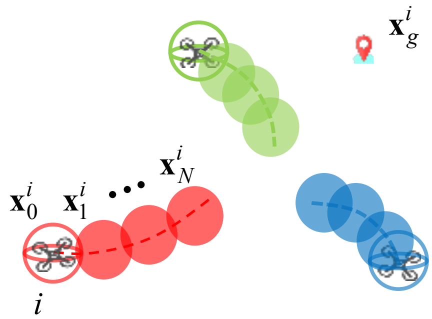
18

## Interaction without communication

MPC relies on motion predictions

$$\min_{x_{1:N},\, u_{0:N-1}} \sum_{k=0}^{N-1} J_k(x_k, u_k) + J_N(x_N)$$
$$\text{s.t.} \quad x_{k+1} = f(x_k, u_k)$$
$$\|\mathbf{p}_i^k - \mathbf{p}_j^k\| \geq 2r$$

future trajectories of other robots

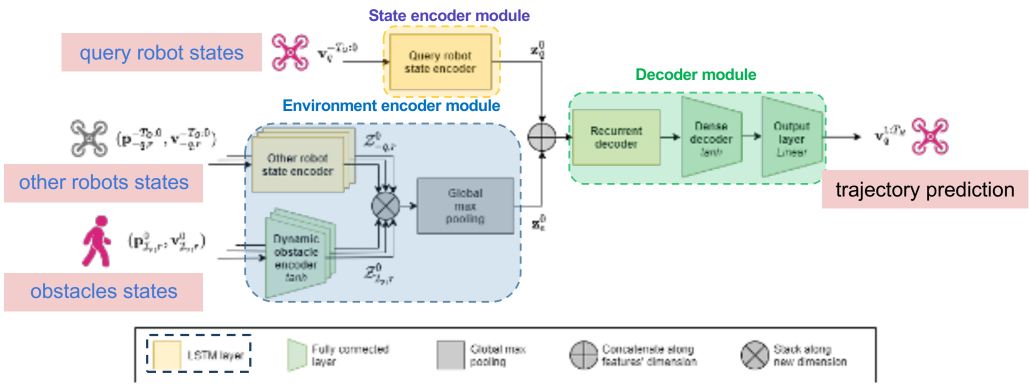



H. Zhu, et al., **"Learning Interaction-Aware Trajectory Predictions for Decentralized Multi-Robot Motion Planning in Dynamic Environments"**, IEEE RA-L), 2021 19

## MPC with interaction-aware predictions

RNN-based model to output "interaction-aware" predictions

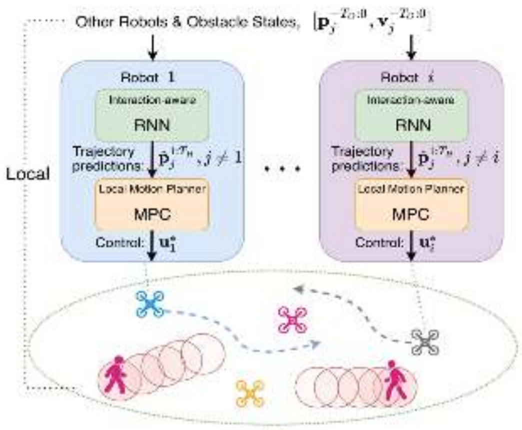
- Trained with a multi-robot simulator using centralized sequential planning




H. Zhu, et al., **"Learning Interaction-Aware Trajectory Predictions for Decentralized Multi-Robot Motion Planning in Dynamic Environments"**, IEEE RA-L), 2021 20

## MPC with interaction-aware predictions

RNN-based model to output "interaction-aware" predictions
Input to MPC for decentralized multi-robot motion planning




H. Zhu, et al., **"Learning Interaction-Aware Trajectory Predictions for Decentralized Multi-Robot Motion Planning in Dynamic Environments"**, IEEE RA-L), 2021

21

## Safety disclaimers: NN predictions

MPC explicitly checks collision avoidance constraints
- However, those are a function of predictions from a NN model!

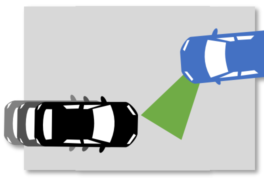
We "hope" that those predictions are close to reality
(and recompute at 10-100Hz to adapt to changes)

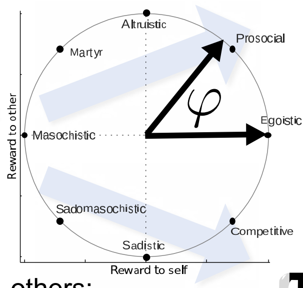
22

## Social Behavior for Autonomous Driving

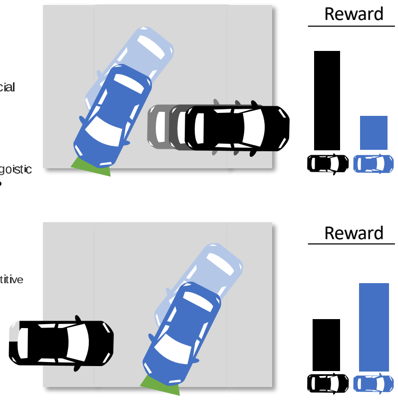
Model interaction directly in the planner

- Estimate Social Value Orientation of other drivers











Weight reward self vs. others:

$$g_i(\cdot) = \cos\varphi_i \, r_i + \sin\varphi_i \, r_j$$

W. Liebrand et. al., "The ring measure of social values: A computerized procedure for assessing individual differences […] and social value orientation", 1988.
W. Schwarting, et al.," **Social Behavior for Autonomous Vehicles**", PNAS, 2019



## Social Behavior for Autonomous Driving

Model interaction directly in the planner

- Estimate Social Value Orientation of other drivers
- Cost function $g_i(\cdot) = \cos\varphi_i \, r_i + \sin\varphi_i \, r_j$ used within MPC framework
- Formulate and solve a joint dynamic game (Nash equilibrium)

Prosocial drivers create a gap for the AV to merge

Autonomous Vehicle — Prosocial drivers

W. Schwarting, et al.," **Social Behavior for Autonomous Vehicles**", PNAS, 2019

## Safety disclaimer: solving a dynamic game

MPC explicitly checks collision avoidance constraints

We use the estimated Social Value Orientation parameter of other drivers and their reward function (obtained through Inverse Reinforcement Learning)

→ A better model of their future behavior

We solve for a Nash equilibrium

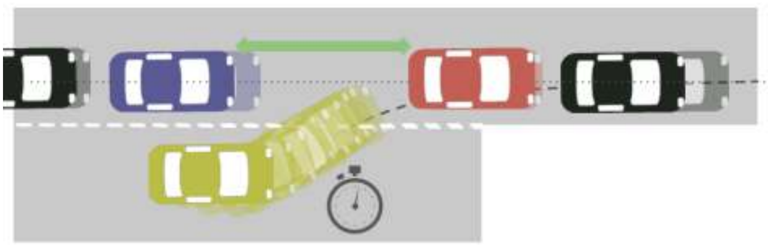
→ We are ”assuming” that other agents will also follow this (plan for the same Nash equilibrium) and behave accordingly!

26

## Interactive Model Predictive Controller

Human drivers communicate their intentions and negotiate their driving maneuvers by adjusting both **time headway** and **distance** to others

→ translated into a **velocity reference**

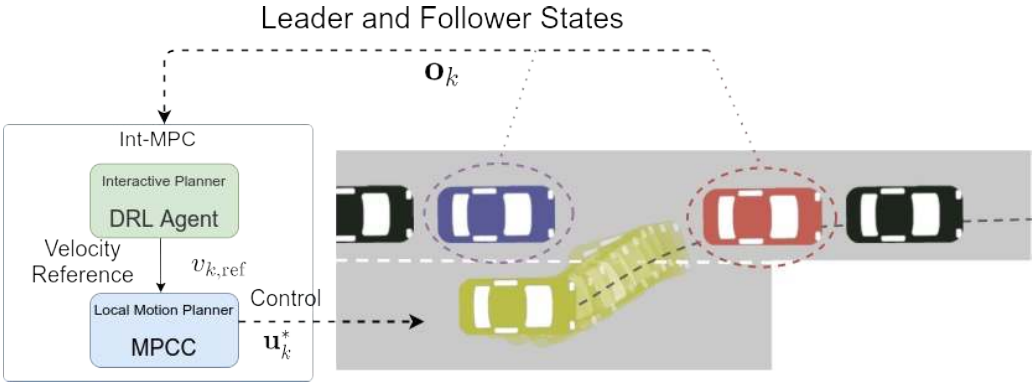


B. Brito et al., **"Learning Interaction-aware Guidance for Trajectory Optimization in Dense Traffic Scenarios"**, IEEE T-ITS, 2022

28

## Interactive Model Predictive Controller

Deep Reinforcement Learning Agent trained in scenarios with varying cooperation coefficients




B. Brito et al., **"Learning Interaction-aware Guidance for Trajectory Optimization in Dense Traffic Scenarios"**, IEEE T-ITS, 2022

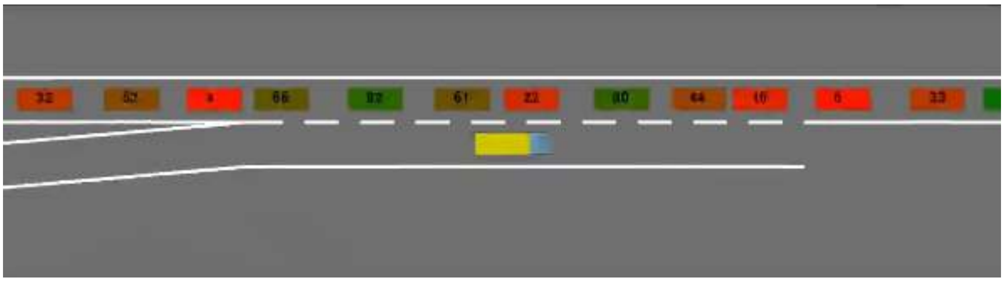
29

## Interactive Model Predictive Controller

Recommendation policy for MPC

✓ Improves collision avoidance & merging performance
✓ Reduced the complexity of the cost function in the local motion planner
✓ Safe learning and execution

- MPC for robot dynamics & collision constraints
- RL for interactions with other agents & guidance



B. Brito et al., **"Learning Interaction-aware Guidance for Trajectory Optimization in Dense Traffic Scenarios"**, IEEE T-ITS, 2022

30

## Safe motion planning among decision-making agents

Trajectory optimization
+ Explicit constraints
+ Vehicle dynamics
+ Safety guarantees

- Interaction
- Local solutions
- Scalability

**Learning provides:**
**cost function, guidance**

Learning
+ Scalable (usage of learned policy)
+ Global solutions

- Lack of guarantees
- Vehicle dynamics

**Trajectory optimization:**
**Real-time & safety guarantees**

31

## Summary

- MPC is a powerful tool that provides guarantees → with some challenges
- Learning combined with MPC is a promising approach to model real-world complexity

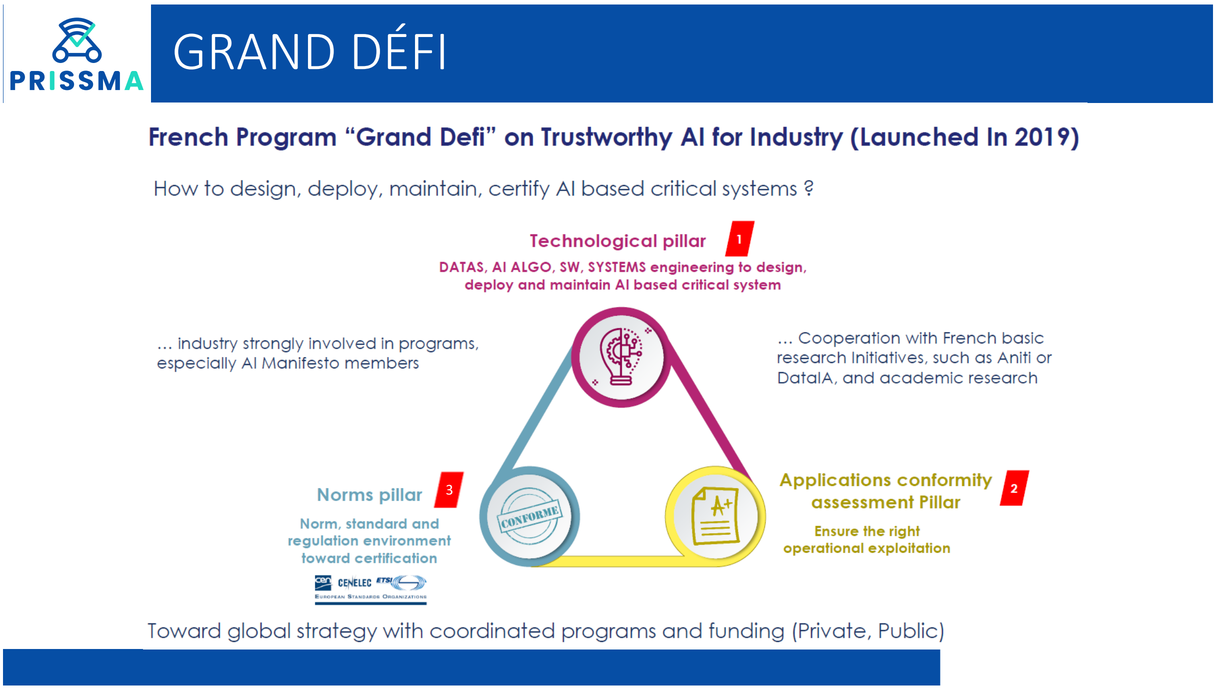
- Challenges:
  - Uncertainty
  - Interaction
  - Safety

Prof. J. Alonso-Mora
https://www.autonomousrobots.nl/
j.alonsomora@tudelft.nl

TU Delft Cognitive Robotics AUTONOMOUS MULTI-ROBOTS LAB

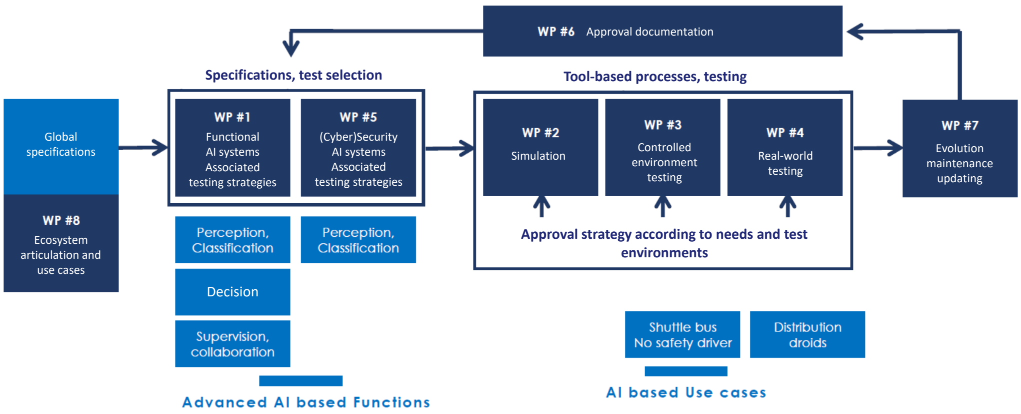
32

## Annex VIII. PRISSMA project overview

# PROJECT STRUCTURE

## #PRISSMA : Project Descrption

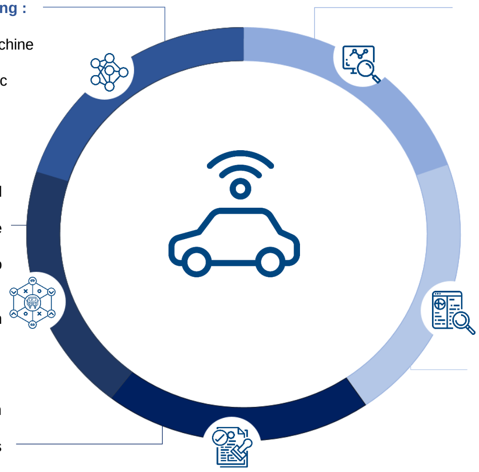



# SAFE BY DESIGN

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

EUROPEAN COMMISSION

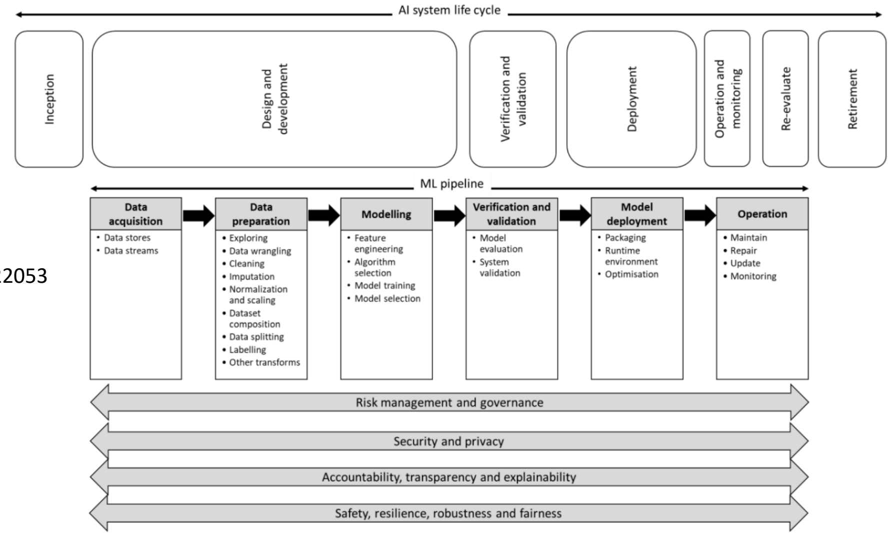
- **Mathematical modeling & learning :**
  - ✓ Statistical modeling vs. machine learning vs. deep learning
  - ✓ Choice according to the problematic
- **Method for model interpretation :**
  - ✓ Direct interpretation // explanation
  - ✓ Interpretation limited to a restricted number of explanatory variables
  - ✓ Global or local explanation of the models
  - ✓ Explanation by overlay modeling to simplify the complete model
  - ✓ Study of individual observations
  - ✓ The objective of the explanation dictates the tool to be used
- **Performance validation**
  - ✓ The model must perform well even with unknown data
  - ✓ Future data: simulations or cross validation with available data
  - ✓ Study the predictions of the model: bias? outliers?



- **Data exploitation :**
  - ✓ Identification of the question/scope of the study
  - ✓ Set the objective for a tool development, not just a model
  - ✓ Describe the data used and each processing step
  - ✓ Involve the final users to be as close as possible to their needs
- **Models' performance :**
  - ✓ Ensure generalization of model performance to unknown/future data
  - ✓ Estimate the variability of the performance by multiple cross-validation
  - ✓ Choose a metric adapted to the problem (regression / classification / segmentation) & to the need (more or less strong penalty for errors, ...) and data
  - ✓ Supervised or unsupervised model
  - ✓ Optimize the model to reduce prediction errors

# AI SYSTEM LIFE CYCLE

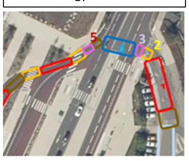
ISO-IEC DIS 22053




# ARTS EVALUATION APPROACH

**Route analysis**

Phase 1 decomposition
Following STRMTG methodology



OD : "Operational Domain" refers to what the world actually is, which might (in most cases will) differ from the ODD.

*ASAM OpenODD*

**ODD**

Conditions under which the system is designed to operate safely.

- Describes the conditions of safe operation of the system from the generic description
- Defines the situations (configurations, object types, conditions, ego actions) that can be encountered on the path (within which events will then take place)

**Scenarios**

- Objective to demonstrate safety on the course OD
- Results will confirm (or not) the "demonstrated" ODD
- ODD demonstrated (by testing) Vs OD of actual course

Elaboration of associated test scenarios based on existing and comparable scenarios (NATM catalog) but also via generated scenarios to test the edge cases that can be associated with the targeted OD.

**OEDR**

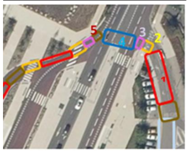
Subtasks of the Dynamic Driving Task that include monitoring the driving environment (detecting, recognizing, and classifying objects and events and preparing to respond as needed) and executing an appropriate response to such objects and events

1. **OED** Study of the detection, recognition, and classification of objects and events associated with the situations identified in the ODD.
2. **R** Study of the system response.

# ARTS EVALUATION APPROACH

**Route analysis**

Phase 1 decomposition
Following STRMTG
methodology



OD : "Operational Domain" refers to what the world actually is, which might (in most cases will) differ from the ODD.

*ASAM OpenODD*

**ODD**

Conditions under which the system is designed to operate safely.

- Describes the conditions of safe operation of the system from the generic description
- Defines the situations (configurations, object types, conditions, ego actions) that can be encountered on the path (within which events will then take place)

**Scenarios**

- Objective to demonstrate safety on the course OD
- Results will confirm (or not) the "demonstrated" ODD
- ODD demonstrated (by testing) Vs OD of actual course

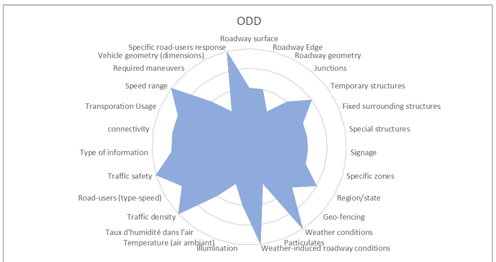
Elaboration of associated test scenarios based on existing and comparable scenarios (NATM catalog) but also via generated scenarios to test the edge cases that can be associated with the targeted OD.

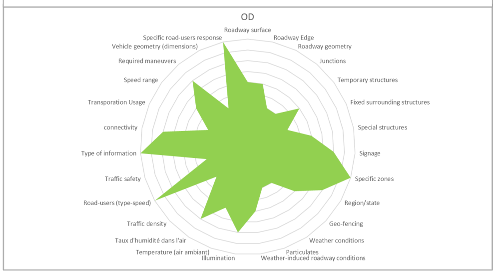
**OEDR**

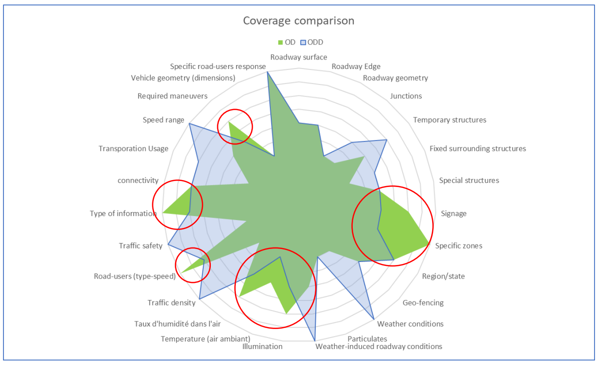
Subtasks of the Dynamic Driving Task that include monitoring the driving environment (detecting, recognizing, and classifying objects and events and preparing to respond as needed) and executing an appropriate response to such objects and events

1. OED Study of the detection, recognition, and classification of objects and events associated with the situations identified in the ODD.
2. R Study of the system response.

# ROUTE ANALYSIS



# ARTS EVALUATION APPROACH

PRISSMA

**Route analysis**

Phase 1 decomposition Following STRMTG methodology

OD : "Operational Domain" refers to what the world actually is, which might (in most cases will) differ from the ODD.

*ASAM OpenODD*

**ODD**

Conditions under which the system is designed to operate safely.

- Describes the conditions of safe operation of the system from the generic description
- Defines the situations (configurations, object types, conditions, ego actions) that can be encountered on the path (within which events will then take place)

**Scenarios**

- Objective to demonstrate safety on the course OD
- Results will confirm (or not) the "demonstrated" ODD
- ODD demonstrated (by testing) Vs OD of actual course

Elaboration of associated test scenarios based on existing and comparable scenarios (NATM catalog) but also via generated scenarios to test the edge cases that can be associated with the targeted OD.

**OEDR**

Subtasks of the Dynamic Driving Task that include monitoring the driving environment (detecting, recognizing, and classifying objects and events and preparing to respond as needed) and executing an appropriate response to such objects and events

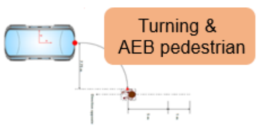
1. OED Study of the detection, recognition, and classification of objects and events associated with the situations identified in the ODD.
2. R Study of the system response.

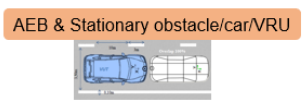
# PRIORITIES AND MAIN CHALLENGES FOR SCENARIO DEFINITION

PRISSMA

**ensure that approved systems provide sufficient robustness with regard to their performance**

standardized procedure for test repetition e.g. UNR152 (AEB)




**Pre-critical scenario to evaluate the level of caution in driving, taking account both driving behavior and driving context**

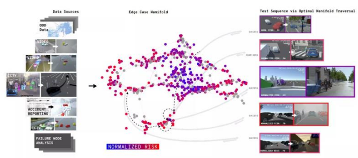
Risk of hidden pedestrian crossing

Strong curve

Traffic jam

**Avoid overfitting and evaluate the driving functions with edge-cases scenarios**

Generate specific and/or random scenarios

PRISSMA

# ARTS EVALUATION APPROACH

**Route analysis**

Phase 1 decomposition Following STRMTG methodology

OD : "Operational Domain" refers to what the world actually is, which might (in most cases will) differ from the ODD.

*ASAM OpenODD*

**ODD**

Conditions under which the system is designed to operate safely.

- Describes the conditions of safe operation of the system from the generic description
- Defines the situations (configurations, object types, conditions, ego actions) that can be encountered on the path (within which events will then take place)

**Scenarios**

- Objective to demonstrate safety on the course OD
- Results will confirm (or not) the "demonstrated" ODD
- ODD demonstrated (by testing) Vs OD of actual course

Elaboration of associated test scenarios based on existing and comparable scenarios (NATM catalog) but also via generated scenarios to test the edge cases that can be associated with the targeted OD.

**OEDR**

Subtasks of the Dynamic Driving Task that include monitoring the driving environment (detecting, recognizing, and classifying objects and events and preparing to respond as needed) and executing an appropriate response to such objects and events

1. **OED** Study of the detection, recognition, and classification of objects and events associated with the situations identified in the ODD.
2. **R** Study of the system response.

PRISSMA

# ODD AND ROUTE DESCRIPTORS

ODD Definition

- Reflecting on the metrics
- Reflection on the content of the branches
- Reflecting on the articulation of the ODD / Route / Scenarios

Application to a route

- Exchanges with
  - Other projects
  - Authorities
  - Working groups

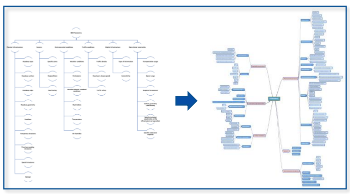
Taxonomy

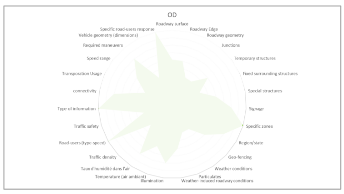
- Method of route decomposition
- Method / identification of critical scenarios
- Application to a route

Application to a route

- Exchanges with
  - Other projects
  - Authorities
  - Working groups

Guide to characterizing the ODD

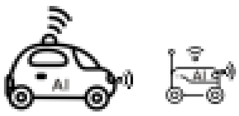




## Annex IX. Towards Robust Autonomous Vehicles

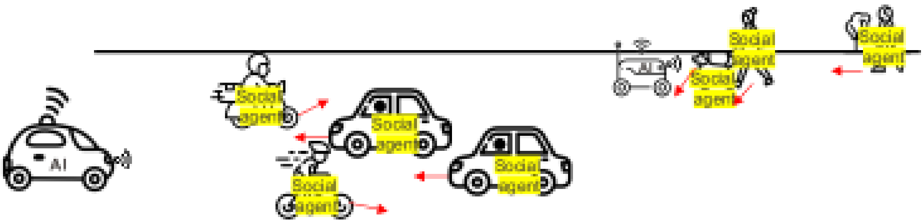
EPFL VITA

Visual Intelligence for Transportation




**Robust Autonomous Vehicles**

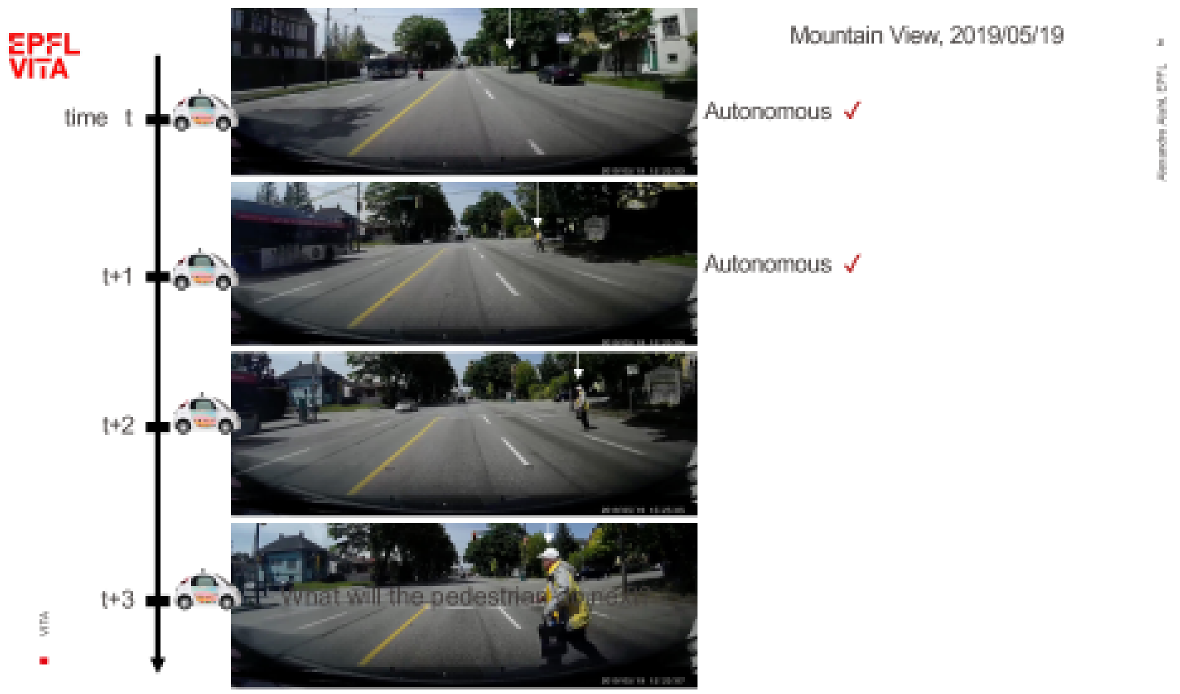
Prof. Alexandre Alahi

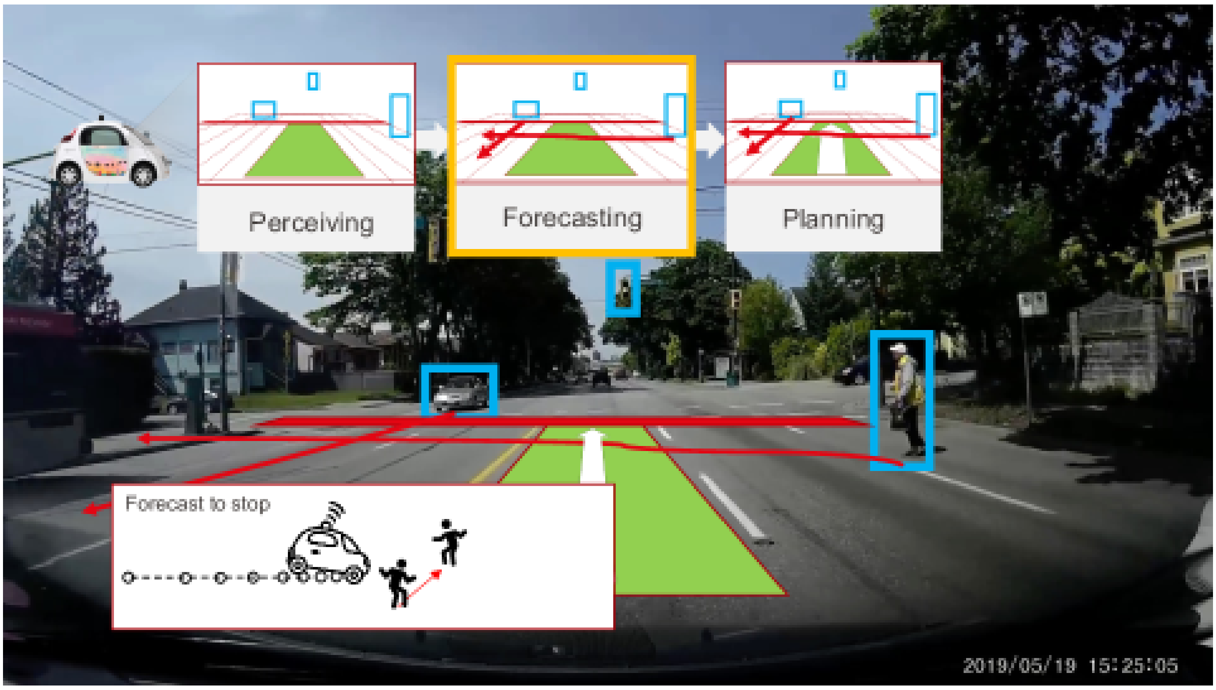



"Humans subconsciously **forecast the future**…
Autonomous Vehicles must have the same **forecasting** capability to **harmlessly** and **effectively co-exist**",

Our lab goal.

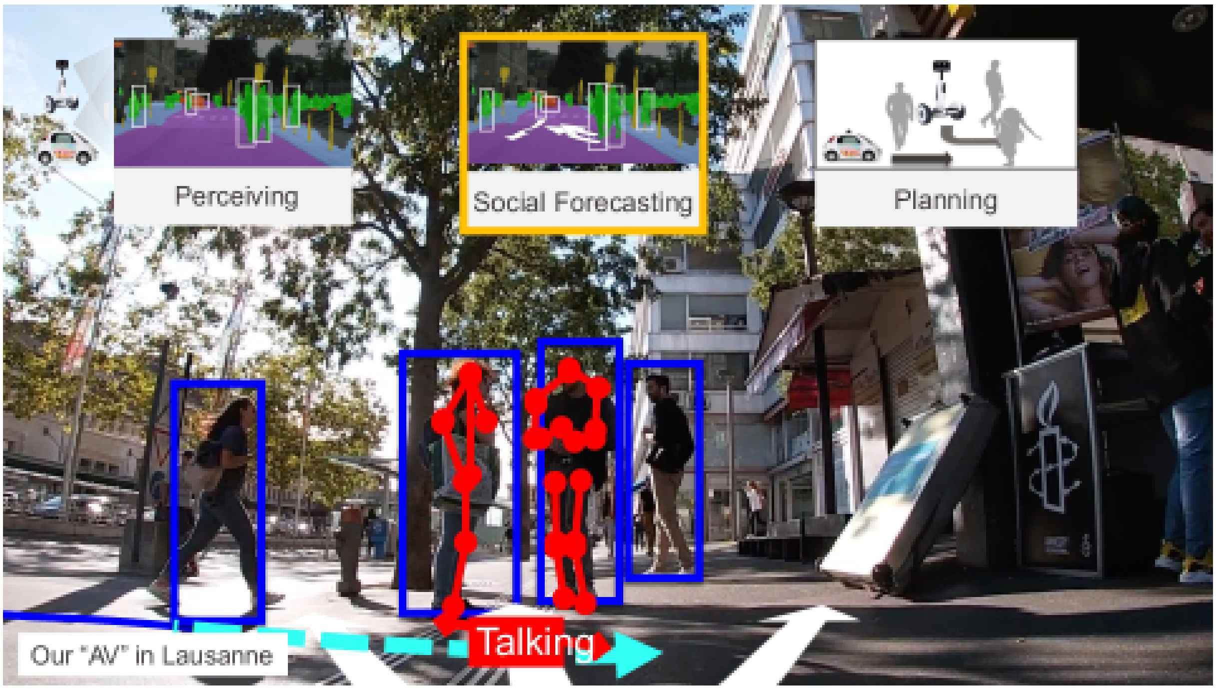
**Forecasting** is essential…

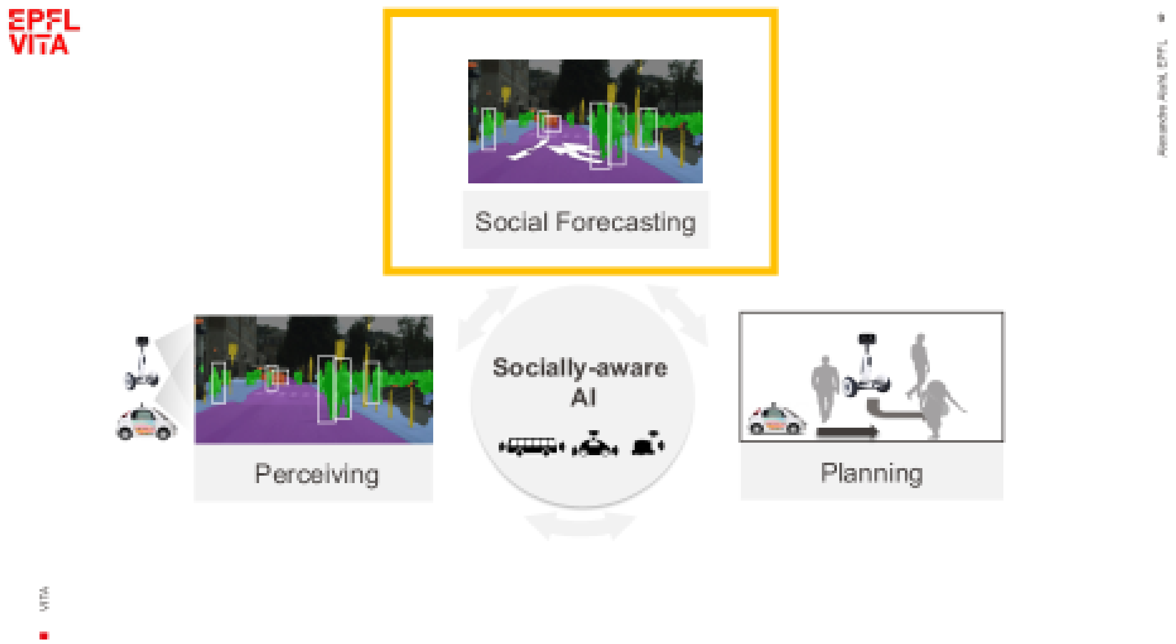

EPFL
VITA
Mountain View, 2019/05/19
time t
t+1
t+2
t+3
Autonomous ✓
Autonomous ✓


Perceiving
Forecasting
Planning
Forecast to stop
2019/05/19 15:25:05


AI must forecast agent-agent* interactions = Social Forecasting
*agent = any moving entity in the world (driver, pedestrian, cyclist…)
Forecast not to stop
98% of AV accidents are due to an unexpected stop*


EPFL VITA
Lausanne, 2020/07/15
time t
Autonomous ✓
t+1
Autonomous ✓
t+2
Autonomous ✗
=> Our robot freezes in close human proximity
t+3
Autonomous ✗
=> Our robot does not comply with social norms

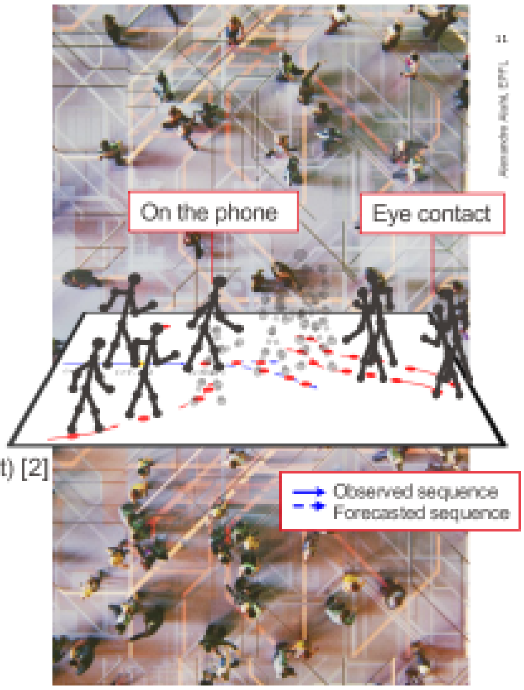

Perceiving
Social Forecasting
Planning
Talking
Our "AV" in Lausanne


EPFL
VITA
Social Forecasting
Socially-aware AI
Perceiving
Planning

## Social Forecasting
### (w/ pedestrians)

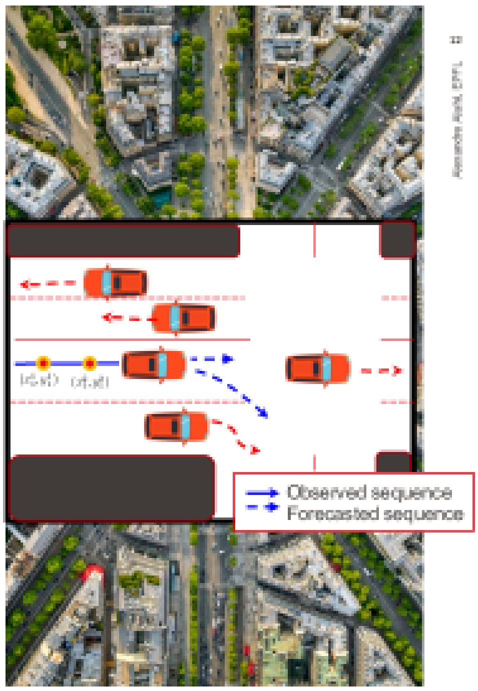
- **Input**: several sequences of states
- **Output**: forecast the future states, *e.g.*, next 5 seconds
- **State**:
  - ($x^t$,$y^t$) coordinates in time
  - Body pose [1]
  - Attributes (*e.g.*, on the phone, eye contact) [2]
- Challenge 1: agent-agent interactions
- Challenge 2: disentangle physics from social

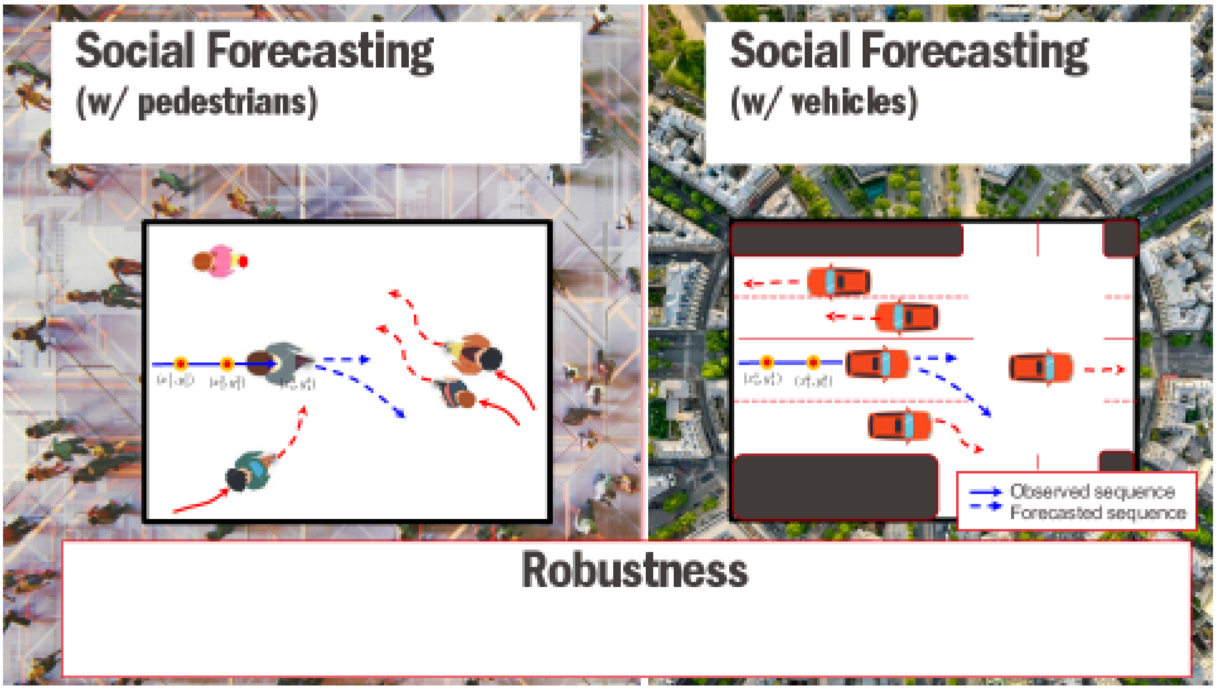
[1] PifPaf, CVPR'19
[2] 32 attributes detector, ITS transactions'21



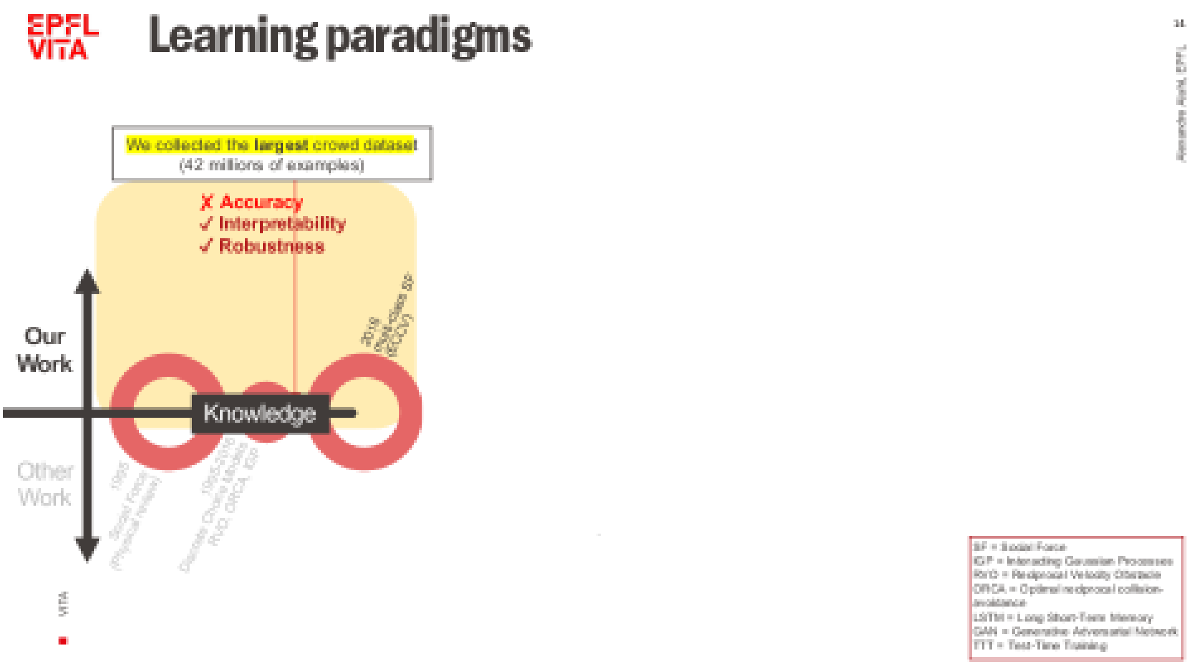

## Social Forecasting
### (w/ vehicles)

- **Input**: several sequences of states + scene infrastructure
- **Output**: forecast the future states, *e.g.*, next 5 seconds
- Challenge 1: agent-agent interactions
- Challenge 2: agent-scene interactions
- Challenge 3: additional external constraints



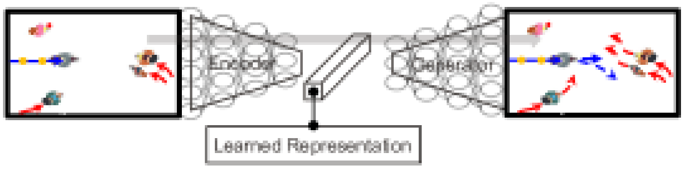


Social Forecasting
(w/ pedestrians)
Social Forecasting
(w/ vehicles)
Observed sequence
Forecasted sequence
Robustness

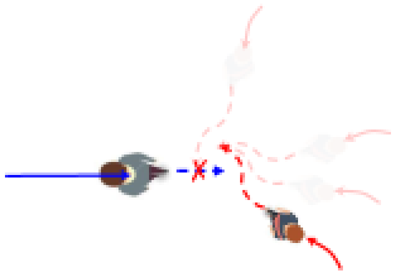



EPFL
VITA
Learning paradigms
We collected the largest crowd dataset
(42 millions of examples)
✗ Accuracy
✓ Interpretability
✓ Robustness
Our
Work
Knowledge
Other
Work
SF = Social Force
IGP = Interacting Gaussian Processes
RVO = Reciprocal Velocity Obstacle
ORCA = Optimal reciprocal collision-avoidance
LSTM = Long Short-Term Memory
GAN = Generative Adversarial Network
TTT = Test-Time Training

## Current paradigm




✗ Not Robust



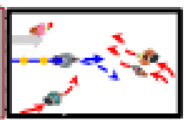
→ Observed sequence
⇢ Forecasted sequence by [1]
✗ Collision

[1] Ynet, ICCV'21, Top ranked model in Trajnet++ public challenge

## Current paradigm

Current paradigm:

Use **data** to **learn** the task of **navigation from scratch**



✗ Not Robust

Our proposed paradigm:

Use **data** to **only learn concepts we struggle to explain** (e.g., our social norms)

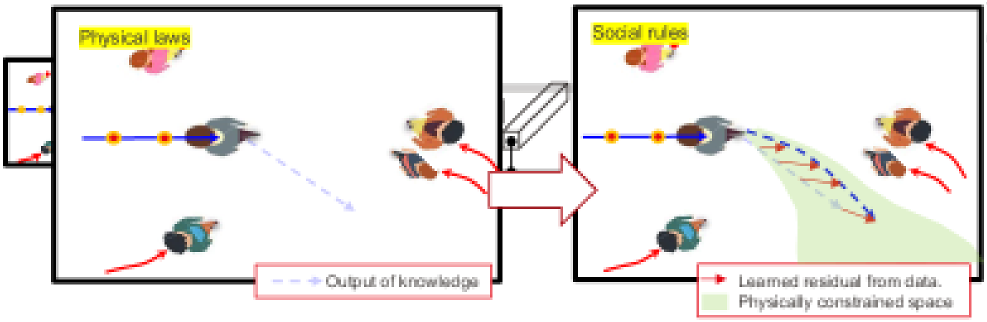
### Because

1. Imbalanced/missing data

### Solution

- Knowledge-Data

## Proposed Knowledge-Data paradigm




### Because

1. Imbalanced/missing data

### Solution

- Knowledge-Data
  - Knowledge as input

[1] Injecting knowledge in data-driven vehicle trajectory predictors, **TRC'21**



## Trajnet++

- Open-source library (> 15 models)
  - https://github.com/vita-epfl/trajnetplusplusdata
- Data+evaluation protocols
- Challenge on Aicrowd
  - https://www.aicrowd.com/challenges/trajnet-a-trajectory-forecasting-challenge

GitHub

## Autonomous driving


Safety critical task

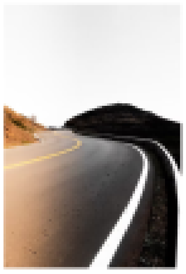
- Careful assessment needed
  - Trying every traffic situation ✗
  - "Smart and automated" assessment ✓
- A generalizable model is required
  - Working model in the available dataset ✗
  - Robust model in different situations ✓

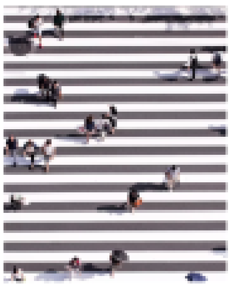
## S-attack library

S-attack library: A smart and automated assessment for trajectory prediction models

Assessment in terms of:

- Interaction with other users: Social-attack
- Interaction with infrastructure: Scene-attack



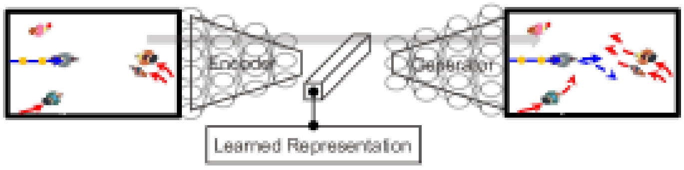


EPFL VITA

# New evaluation protocol








## Outcome

✓ **New evaluation** based on realistic adverserial examples [1]

✓ **Robust training**

- Observed sequence
- Forecasted sequence by [2]
- Perturbed observation by < 7 cm
- Forecasted sequence leading to collision
- ✗ **Collision**

[2] Ynet, ICCV'21, Top ranked model in Trajnet++ public challenge

VITA

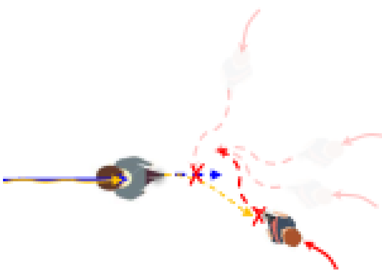
[1] S-attack library, https://s-attack.github.io/

EPFL VITA

# Quantitative results







| Baseline | Original collision rate |
|---|---|
| S-LSTM (CVPR'16) | 7.8% |
| S-Att (ICRA'18) | 9.4% |
| S-GAN (CVPR'18) | 13.9% |
| D-Pool (ITS'2021) | 7.3% |
| S-STGCNN (CVPR'20) | 16.3% |
| PECNet (ECCV'20) | 15.0% |

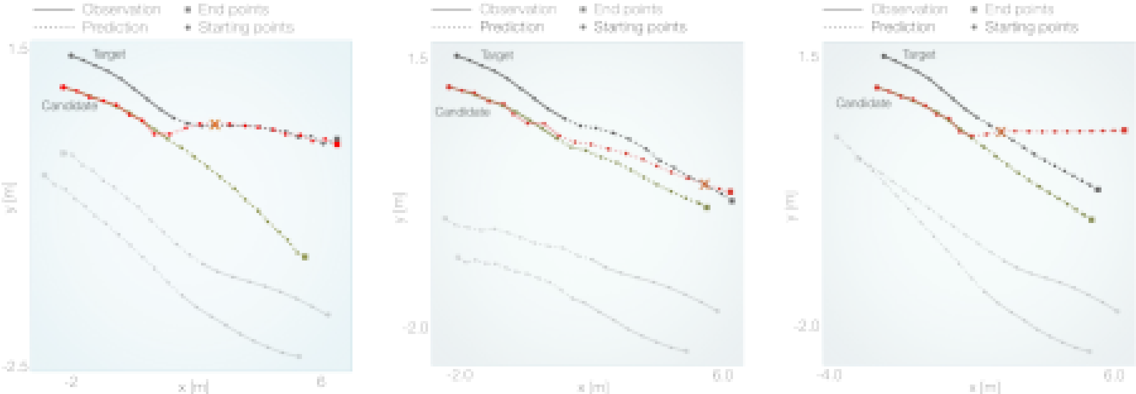
=> 6.5% w/ aug

VITA

[1] S-attack library, https://s-attack.github.io/

# Qualitative results





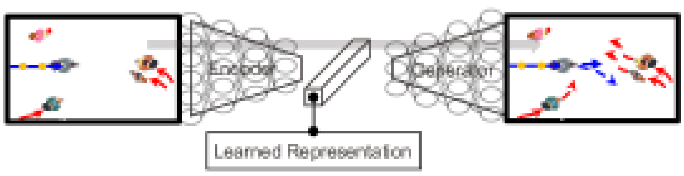

D-Pool[ITS'2021] S-STGCNN[CVPR'20] S-GAN[CVPR'18]



[1] S-attack library, https://s-attack.github.io/

# New evaluation protocol

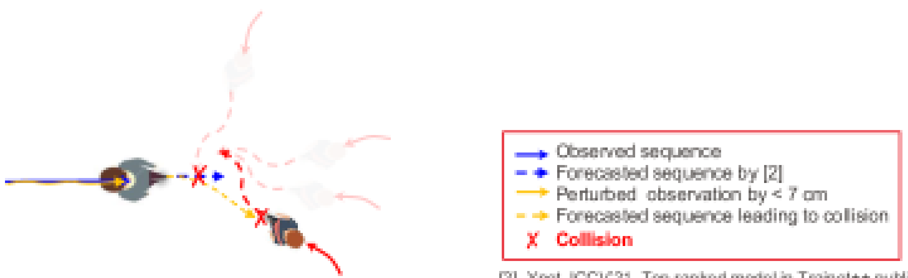





## Outcome

✓ **New evaluation** based on realistic adverserial examples [1]

✓ **Robust training**

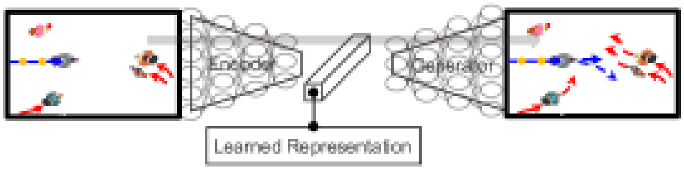





[1] S-attack library, https://s-attack.github.io/

[2] Ynet, ICCV'21, Top ranked model in Trajnet++ public challenge

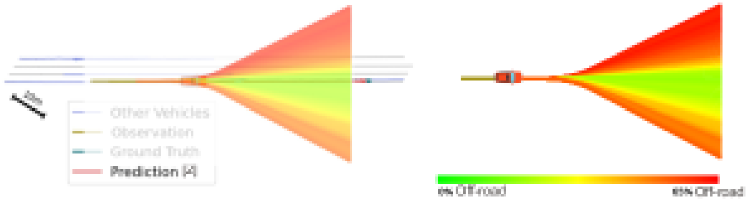
## New evaluation protocol

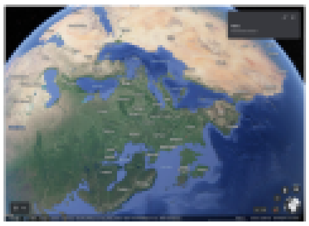



### Outcome

✓ **New evaluation** based on realistic adverserial examples [1]

✓ **Robust training**

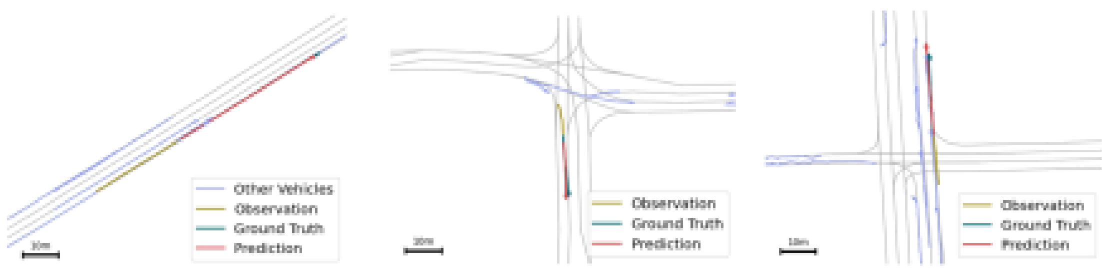


[1] Vehicle trajectory prediction works, but not everywhere, **CVPR'22** [2] LaneGCN, ECCV'20, Top ranked model in Argoverse public challenge

## Scene generation

Atomic scene generation functions




Simple turn Double turn Ripple road

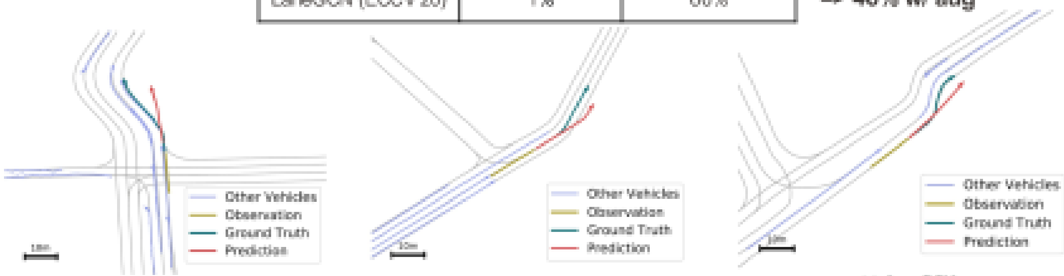
[1] Vehicle trajectory prediction works, but not everywhere, **CVPR'22**

EPFL VITA

# Quantitative results

| Baseline | Original off-road | Generated (ours) off-road |
|---|---|---|
| DATF (ECCV20) | 2% | 82% |
| WIMP (arXiv20) | 1% | 63% |
| LaneGCN (ECCV'20) | 1% | 66% |

=> 46% w/ aug




(a) DATF (b) WIMP (c) LaneGCN

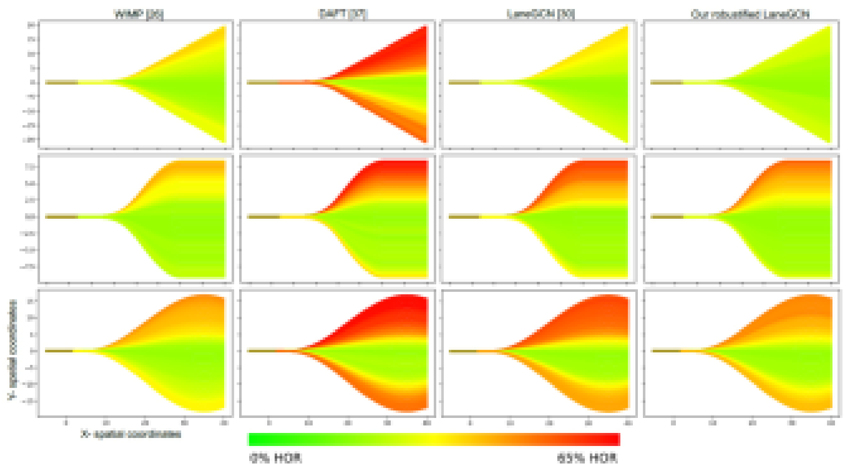
VITA

- [1] Vehicle trajectory prediction works, but not everywhere, **CVPR'22**



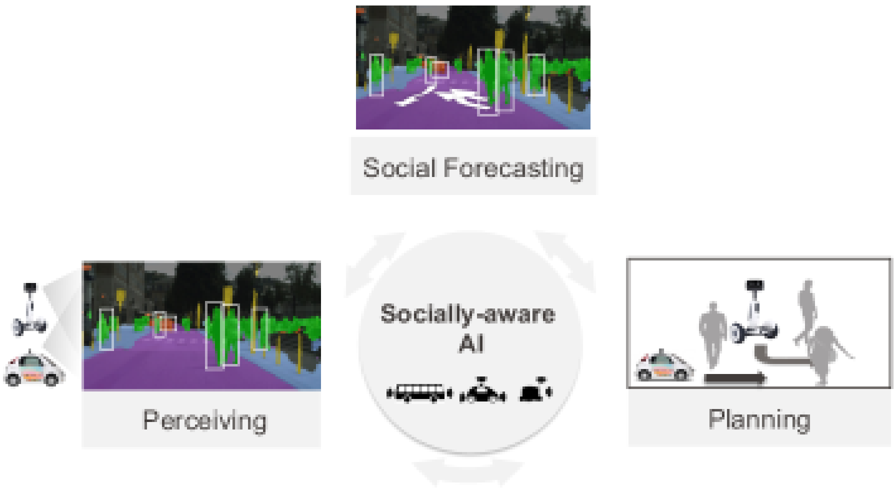
EPFL VITA

# Discussions




VITA

- [1] Vehicle trajectory prediction works, but not everywhere, **CVPR'22**




# #Open Science

**Perception:**
[1] S. Kreiss et al., OpenPifPaf **library** for pose estimation, **CVPR'19, ICCV'21 (licensed)**
[2] L. Bertoni et al., 3D perception **library, ICCV'19, ICRA'21**
[3] L. Bertoni et al., Perceiving Social Distancing, **ITS'20**
[4] G. Adaimi et al., Deep Visual Re-identification with Confidence, **TRC'21**
[5] T. Mordan et al., Detecting 32 human attributes, **ITS'21**
**Prediction:**
[6] Kothari et al., Trajnet++ **library** for spatio-temporal forecasting tasks (>15 implemented models)
[7] Kothari et al., Social Anchor, **ICCV'21**
[8] Liu et al., Social NCE, **ICCV'21**
**Planning:**
[9] C. Chen et al., Crowd-Robot Interaction, **ICRA'19**
**Generative models:**
[10] Y. Liu* et al., Collaborative Sampling in GAN, **AAAI'20**
[11] A. Carlier et al., Deep SVG, **NeurIPS'20**
**DCM + NN**
[12] B. Sifringer et al., L-MNL, **TRB'20**
**Test-time training:**
[13] Y. Liu* et al., TTT++, **NeurIPS'21**
**Tools**
[14] Video Ultimate labeling
[15] S-attack library, **CVPR'22**

GitHub
Code on-line: **vita.epfl**.ch/code

## Annex X. Know the rules well so you can break them effectively - Can we ensure AVs drive safely?

reed mobility

*"Know the rules well so you can break them effectively"*

Can we ensure AVs drive safely?

March 2022

JRC XAI workshop

### About Reed Mobility

- 15+ years in cutting edge transport research, background in psychology / HF
- Academy Director at TRL and lead for CAV research (2004-2017)
- Led portfolio of £50m+ projects (GATEway, SMLL, Helm UK, Move UK, Convex etc.)
- Head of Mobility R&D at Bosch (2017-2019)
- Founded Reed Mobility, June 2019 – current activities:
  - Expert panel producing recommendations on ethics of automated driving (European Commission)
  - CAV standards programme, funded by CCAV (BSI)
  - Automated Vehicle safety assurance scheme (DfT)


reed mobility

Project: Horizon 2020 Commission Expert Group to advise on specific ethical issues raised by driverless mobility

- 14 experts, variety of backgrounds
- Non-exhaustive review
- 18 months: meetings and stakeholder workshop
- **Not** EC position but published with support of EC and taken as an input to inform future research programme

reed mobility

Project: Horizon 2020 Commission Expert Group to advise on specific ethical issues raised by driverless mobility

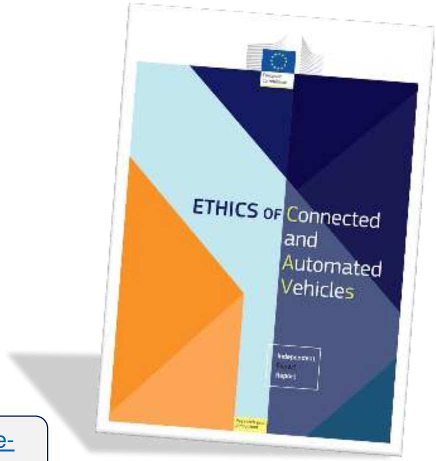
- 14 experts, variety of backgrounds
- Non-exhaustive review
- 18 months: meetings and stakeholder workshop
- **Not** EC position but published with support of EC and taken as an input to inform future research programme


Reed, N., Leiman, T., Palade, P., Martens, M., & Kester, L. (2021). Ethics of automated vehicles: breaking traffic rules for road safety. *Ethics and Information Technology*, 1-13. https://doi.org/10.1007/s10676-021-09614-x

reed mobility

## European Commission – expert panel on CAV ethics


Image credit: European Commission




https://ec.europa.eu/info/news/new-recommendations-for-a-safe-and-ethical-transition-towards-driverless-mobility-2020-sep-18_en

reed mobility

## European Commission – expert panel on CAV ethics

**Safety**

1. Ensure that CAVs reduce physical harm to persons.
2. Prevent unsafe use by inherently safe design.
3. Define clear standards for responsible open road testing.
4. Consider revision of traffic rules to promote safety of CAVs and investigate exceptions to non-compliance with existing rules by CAVs.
5. Redress inequalities in vulnerability among road users.
6. Manage dilemmas by principles of risk distribution and shared ethical principles.

**Transparency**

7. Safeguard informational privacy and informed consent.
8. Enable user choice, seek informed consent options and develop related best practice industry standards.
9. Develop measures to foster protection of individuals at group level.
10. Develop transparency strategies to inform users and pedestrians about data collection and associated rights.

**Transparency**

11. Prevent discriminatory differential service provision.
12. Audit CAV algorithms.
13. Identify and protect CAV relevant high-value datasets as public and open infrastructural resources.
14. Reduce opacity in algorithmic decisions.
15. Promote data, algorithmic, AI literacy and public participation.

**Responsibility**

16. Identify the obligations of different agents involved in CAVs.
17. Promote a culture of responsibility with respect to the obligations associated with CAVs.
18. Ensure accountability for the behaviour of CAVs (duty to explain).
19. Promote a fair system for the attribution of moral and legal culpability for the behaviour of CAVs.
20. Create fair and effective mechanisms for granting compensation to victims of crashes or other accidents involving CAVs.

Image credit: European Commission

reed mobility

Recommendation 4


**4.** **Consider revision of traffic rules to promote safety of CAVs and investigate exceptions to non-compliance with existing rules by CAVs.**

reed mobility

When to break the rules…

- Rules are a means by which road safety is achieved but non-compliance is sometimes necessary to achieve greater road safety
- How should an CAV manage this?
  - Change the rule?
  - Hand control back to human driver to decide?
  - Not comply but CAV must be able to offer reasoned explanation as to why it was non-compliant

reed mobility

## UK review of regulatory framework

- Law Commission of England & Wales / Scottish Law Commission
- Four-year review of regulatory framework for AVs (2018-22): https://www.lawcom.gov.uk/project/automated-vehicles/

First consultation asked respondents to consider two scenarios:

i. exceeding the speed limit

ii. mounting the kerb

reed mobility

## Views expressed in consultation

- No agreement from industry/experts; wide spectrum of views
  - Breach never permitted
  - Breach permitted in minimal circumstances only
  - General principles to identify when breach of rules permitted
  - Specific description of when & how breach permitted

- Views reflect differing perspectives/assumptions about
  - Level of safety risks posed by breach
  - Reasonableness of response
  - CAV capability

reed mobility

## Enforcement?

- Not all breaches by human drivers lead to charge for breach
- Often no charges unless
  - Breach observed directly by or reported to police
  - Breach impacts others
  - Prosecutorial discretion exercised
    - (rather than in/formal warning/counselling)
- But availability of CAV data is critical here
  - When and how should CAVs be charged?

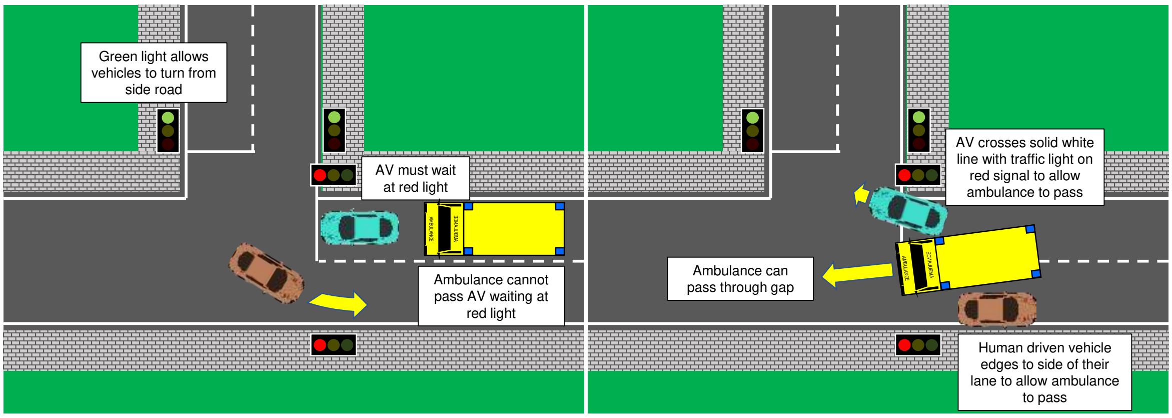
reed mobility

## Strict compliance or discretion

- Programming for strict compliance with traffic rules may not necessarily achieve optimal road safety
- Programming for discretion is very difficult
  - Impossible to anticipate every situation where discretion might need to be exercised
  - Environmental conditions, traffic and other road users vary dramatically between domains and over time in any one domain
  - No training data set can exhaust all possibilities

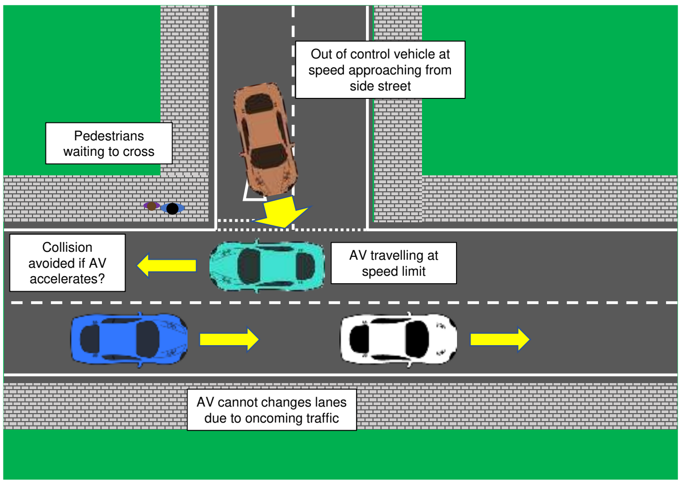
reed mobility

## Example 1 – Crossing a red light




reed mobility

## Example 2 – Exceeding the speed limit




reed mobility

## Ethical goal functions

- AI systems cannot independently ‘learn’ to derive ambiguous human values from human behaviour or human feedback nor apply them to new situations
- Even if sufficiently large training datasets were available, CAVs cannot develop underlying ethical principles
- Proposal for **ethical goal functions**
    - How are these developed? By whom?
    - Democratic legitimacy?

reed mobility

## Recommendation 1 & 5


**1. Ensure that CAVs reduce physical harm to persons.**

**5. Redress inequalities in vulnerability among road users.**

reed mobility

## Metrics for safety

- Reduce harm, for all and for each category of road user
- No other possible benefits would compensate for an increased risk of physical harm
- Risk distribution – redress inequalities in vulnerability among road users
- Dependent on ability of CAV to perceive road user categories
- Comparison depends on safety data

reed mobility

## Recommendation 6


**6. Manage dilemmas by principles of risk distribution and shared ethical principles.**

reed mobility

## Dilemmas → risk management

- Driving is a continuous process, balancing multiple objectives and risk
- Dilemma situations may emerge organically from adherence to ethical principles
- Maintaining adherence to these principles should not conflict with ethical / legal requirements
- Importance of:
  - Transparency in developing ethically and socially acceptable operating criteria
  - Data sharing to review outcomes of dilemma situations

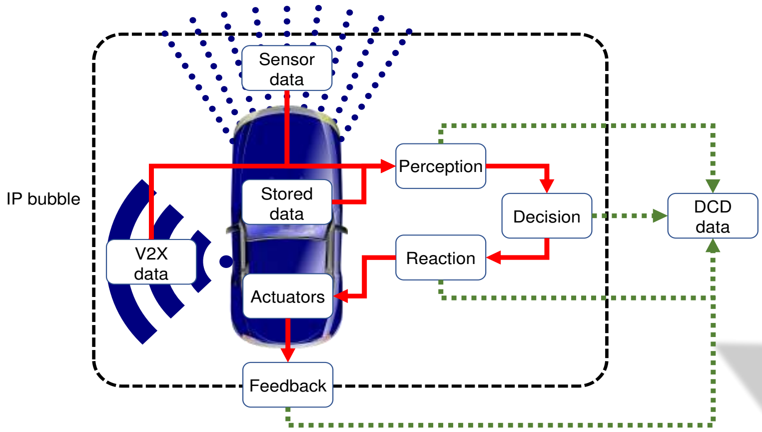
reed mobility

## All depend on fundamentally on **data**

- Need to be able to aggregate and analyse continuous data from AVs
  - Accurate
  - Standardised
  - Comprehensive
  - Shared

reed mobility

## Digital Commentary Driving





reed mobility

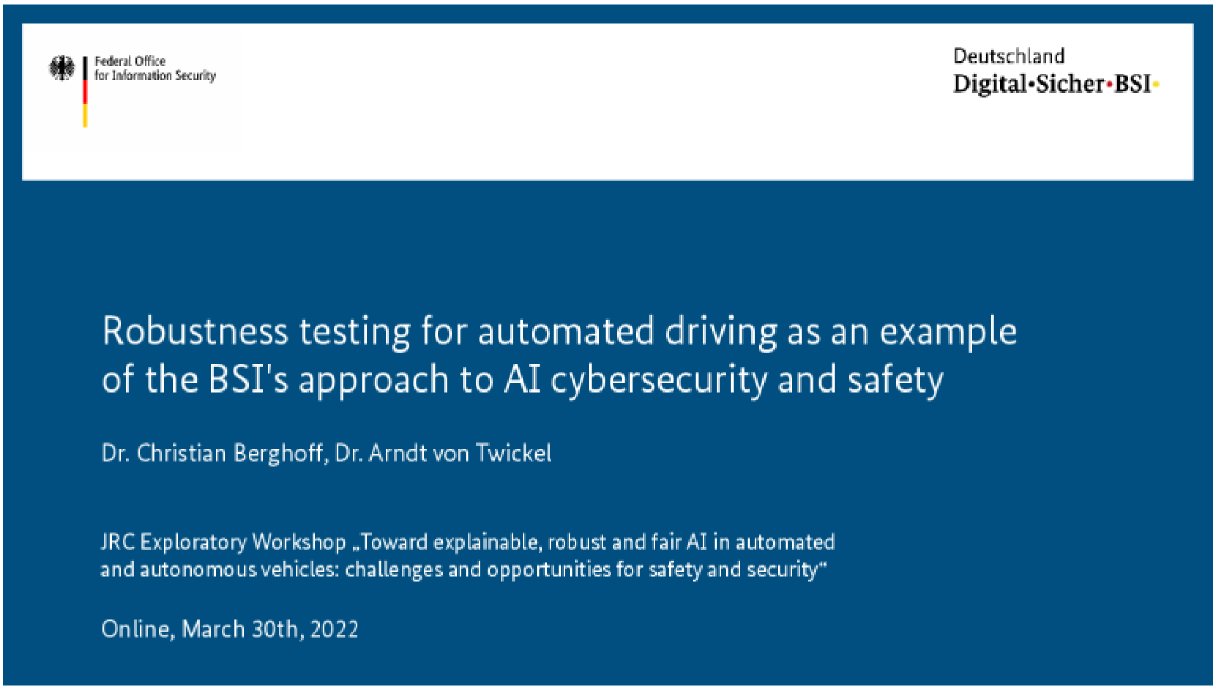
Reed, N., Balcombe, B., Spence, P., Khastgir, S., & Fleming, N. (2021). A review of CAV safety benchmarking and a proposal for a "Digital Commentary Driving" technique. BSI Report. https://www.bsigroup.com/en-GB/CAV/cav-resources/safety-benchmarking-report/

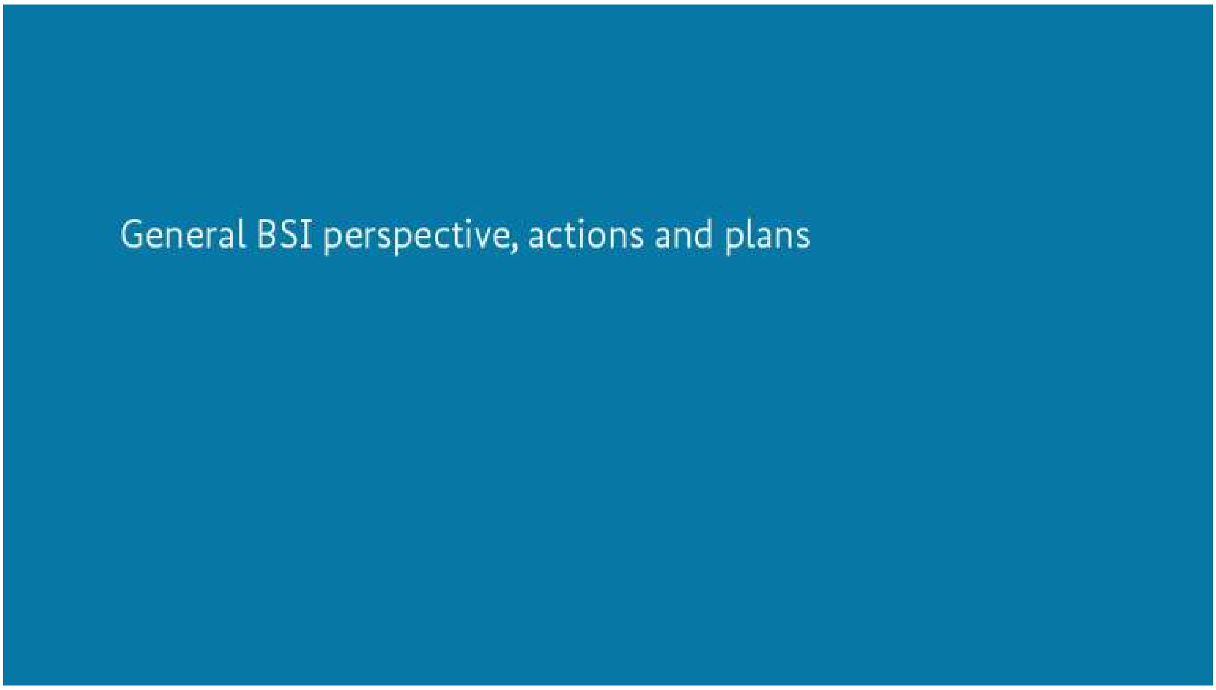
## What we need…

- Industry standard on data collection
- Agreed protocols for data sharing
- Clarity on ethical goals for automated driving
- Societal engagement on definition of ethical goals

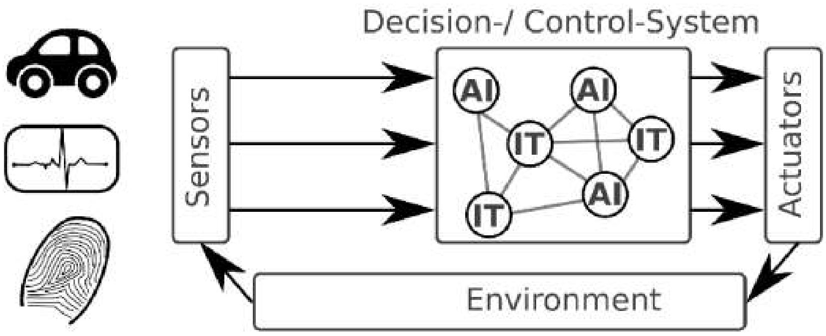
reed mobility

## Annex XI. Robustness testing for automated driving as an example of the BSI's approach to AI cybersecurity



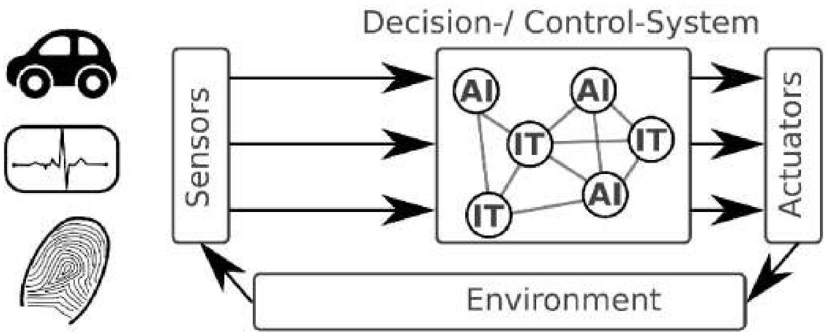
## Practical Criteria and Auditing of Security-Critical AI: Considering it as an Embedded System in the Use-Case Context is Necessary




BSI-relevant aspects:
- Performance
- Robustness
- IT-Security (Integrity)
- Safety
- Explainability
- ...

Non-BSI-relevant:
- Fundamental ethical questions
- User acceptance
- ...

## Practical Criteria and Auditing of Security-Critical AI: Considering it as an Embedded System in the Use-Case Context is Necessary

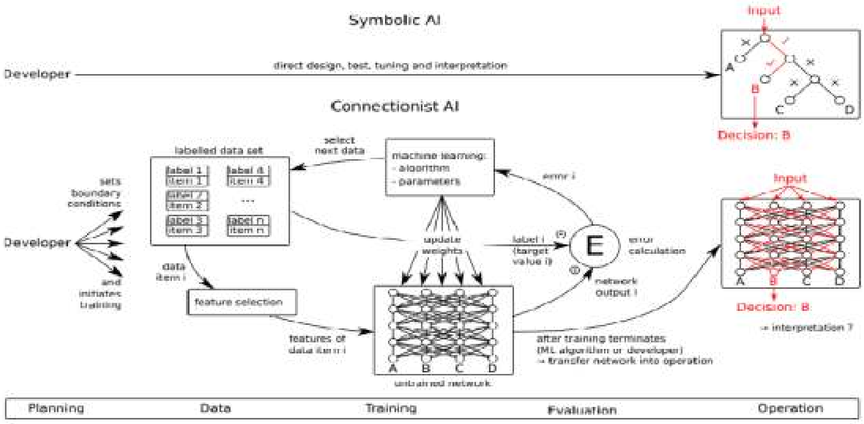



BSI-relevant aspects:
- Performance
- Robustness
- IT-Security (Integrity)
- Safety
- Explainability
- ...

Non-BSI-relevant:
- Fundamental ethical questions
- User acceptance
- ...

How to audit and regulate AI-systems?
→ first approaches exist, e.g. European AI act
→ BUT: methods and tools either do not exist yet or are not yet practically applicable

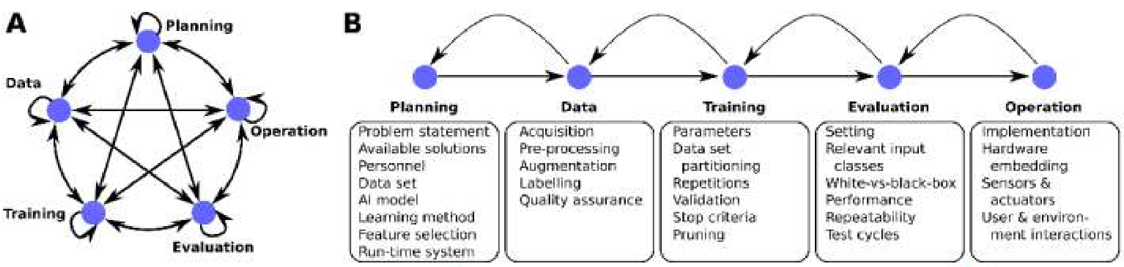
## The Complex Lifecycle of Connectionist AI-Systems Leads to Qualitatively new Vulnerabilities




- input and state spaces are huge
- black-box properties
- dependency on training data

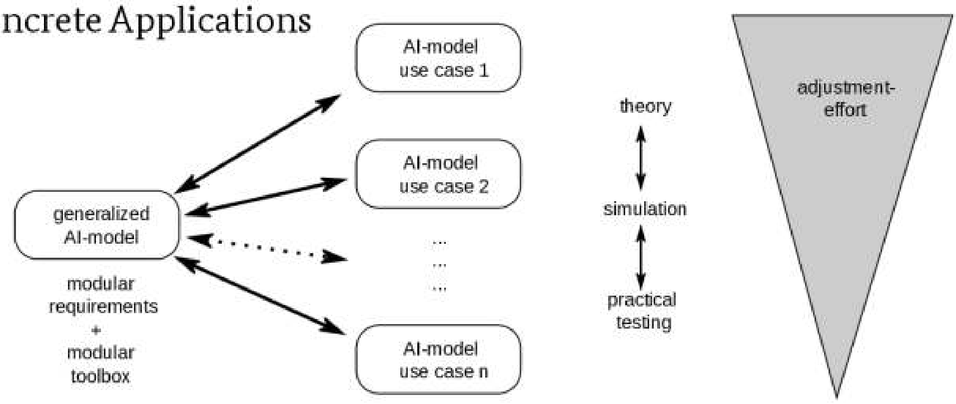
→ the whole life cycle has to be considered

## Multiple Views on the AI System Development Process → Formulation of Requirements




- Reliability (treatment of errors)
- IT security
- Acceptance criteria (Acceptance vs. Risk)
- Documentation
- ...

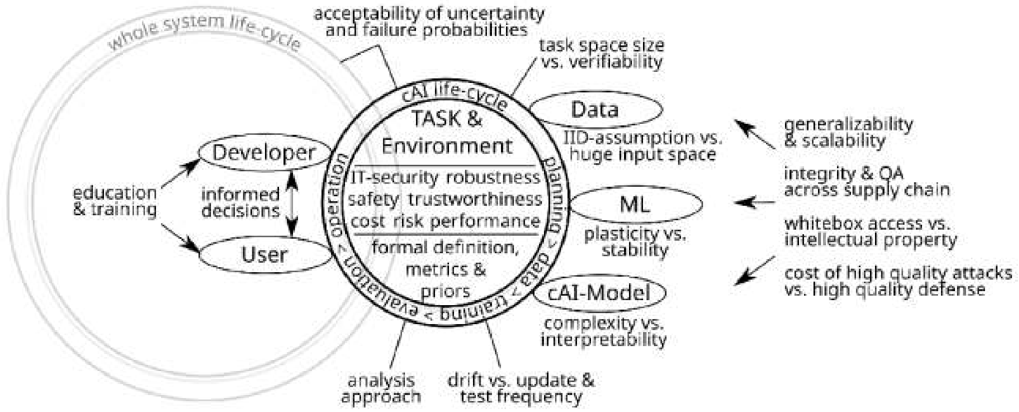
## Iterative Development and Refinement of a Modular Catalogue of Requirements and of a Modular Toolbox by Investigating Concrete Applications



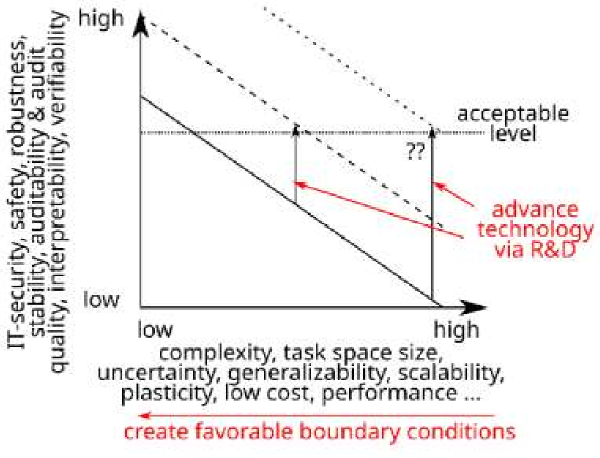

The Goal is to first work out a few use case-specific technical guidelines (TRs) followed by a modular TR
→ BSI project (AIMobilityAuditPrep) started in December 2021

## Open Challenges

## Open Questions in the Context of Auditability, IT Security and Safety

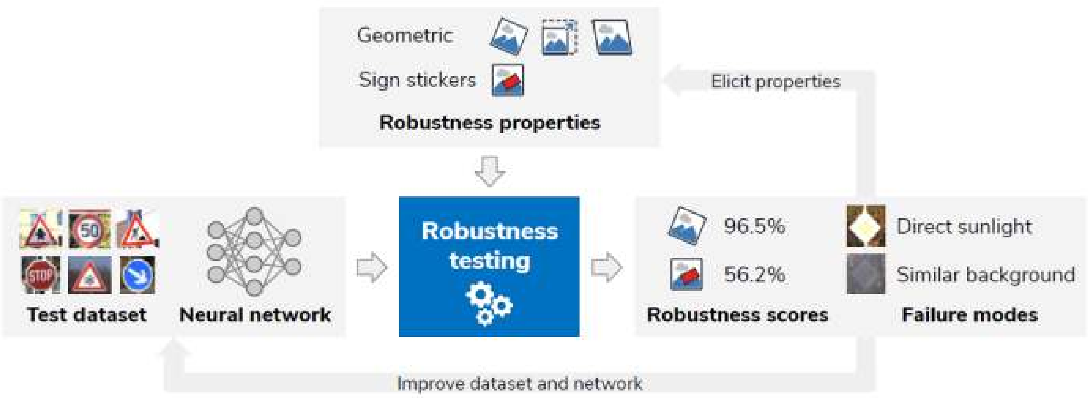



## How to Achieve Acceptable Levels of IT Security, Safety, Audit Quality, Robustness and Verifiability?

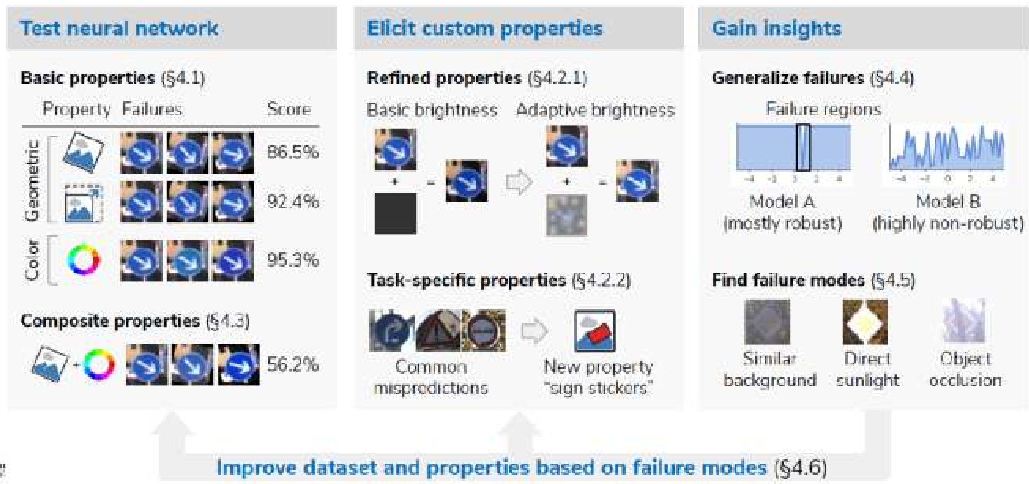


Robustness of AI Systems
(Project with ETH Zurich / Latticeflow, 2020,
Report available at www.bsi.bund.de/KI)

## Test and Improvement of the Robustness of Neural Networks



## Test and Improvement of the Robustness of Neural Networks



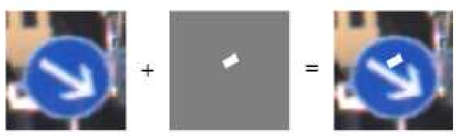

## Robustness against Stickers

- Naturally occurring stickers

- Data Augmentation

Traffic Sign Stickers

| 33.8% | 27.2% |
|---|---|
| SELF-TRAINED | PRE-TRAINED |



inserts a single sticker of varying position, size and orientation on the traffic sign

## Naturally Occurring Perturbations as a Challenge for AI


BSI:

- AI-related documents
- involvement in national & international standardization efforts

## BSI Documents on AI Security (www.bsi.bund.de/KI)

- **Secure, robust and transparent application of AI: Problems, measures and need for action**: presents selected problems as well as measures for security- and safety-critical applications with regard to so-called connectionist AI methods and shows the need for action
- **AI Cloud Service Compliance Criteria Catalogue (AIC4)**: provides AI-specific criteria, which enable an evaluation of the security of an AI service across its life cycle.
- **Vulnerabilities of Connectionist AI Applications: Evaluation and Defense**: Review of the IT security of connectionist artificial intelligence (AI) applications, focusing on threats to integrity (Frontiers in Big Data)
- **Reliability Assessment of Traffic Sign Classifiers**: evaluates how state-of-the-art techniques for testing neural networks can be used to assess neural networks, identify their failure modes, and gain insights on how to improve them
- **Towards Auditable AI Systems:** Whitepaper with VdTÜV and Fraunhofer HHI based on international workshop in 2020
- **The Interplay of AI and Biometrics: Challenges and Opportunities**: article in IEEE Computer in 2021/09

Federal Office for Information Security

## BSI & AI: Involvement in National & International Coorperations & Standardisation Efforts

National

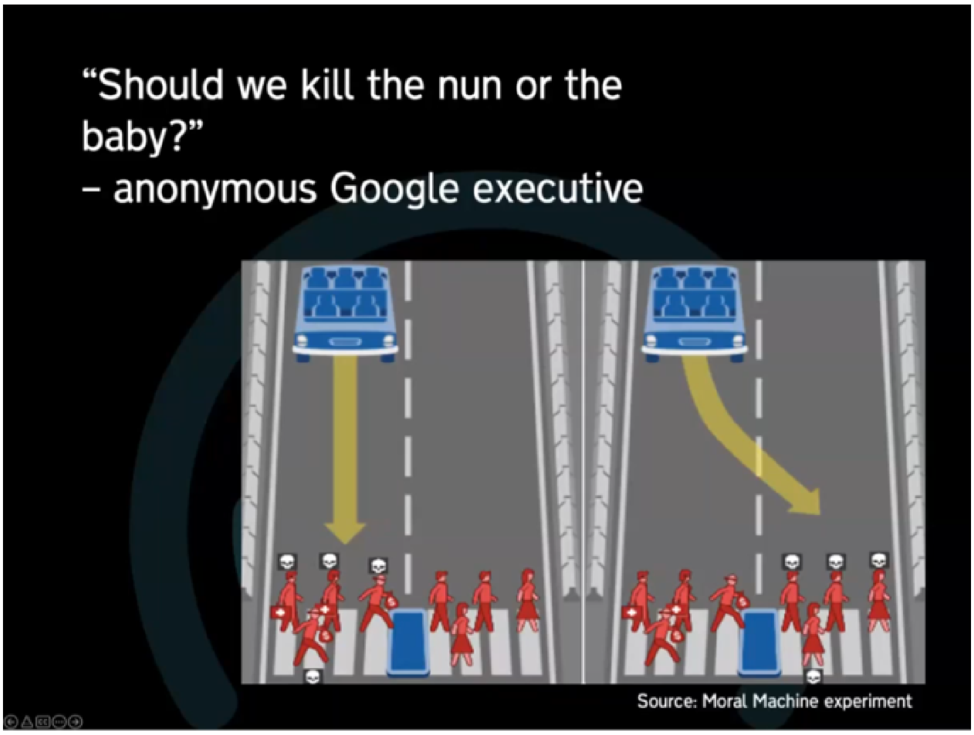
- BSI-VdTÜV working group on AI with a focus on mobility
- Exchange on AI within German administration with BMDV, KBA, BASt
- German DIN Artificial Intelligence Standardization Roadmap v2 Mobility working group
- ...

International

- EU Commission AI Act
- ETSI's Industry Specification Group on Securing Artificial Intelligence (ISG SAI)
- ENISA Adhoc working group on AI
- ...

Federal Office for Information Security

**Annex XII. The actual ethics of AI for AVs: from autonomy to attachments**

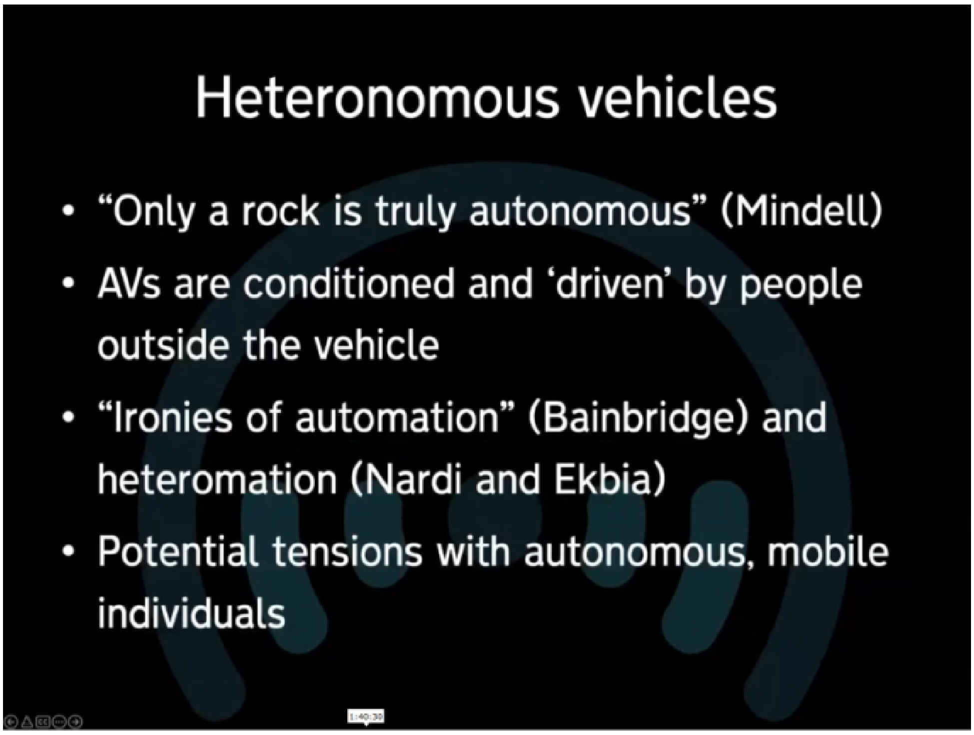
WAYMO

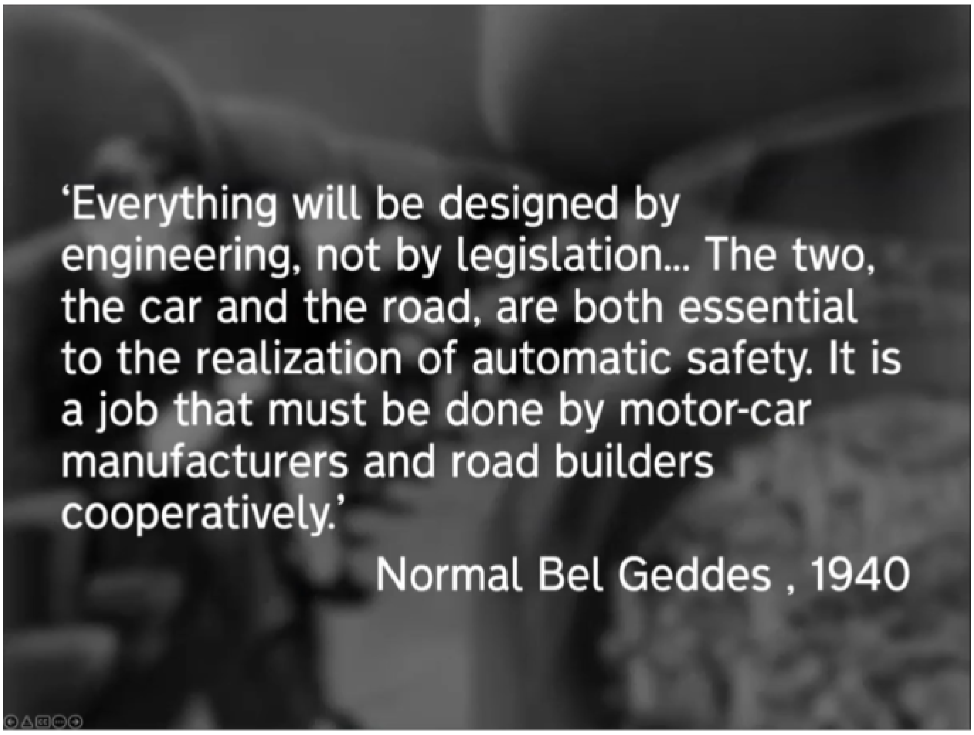
**Let's Talk Autonomous Driving**

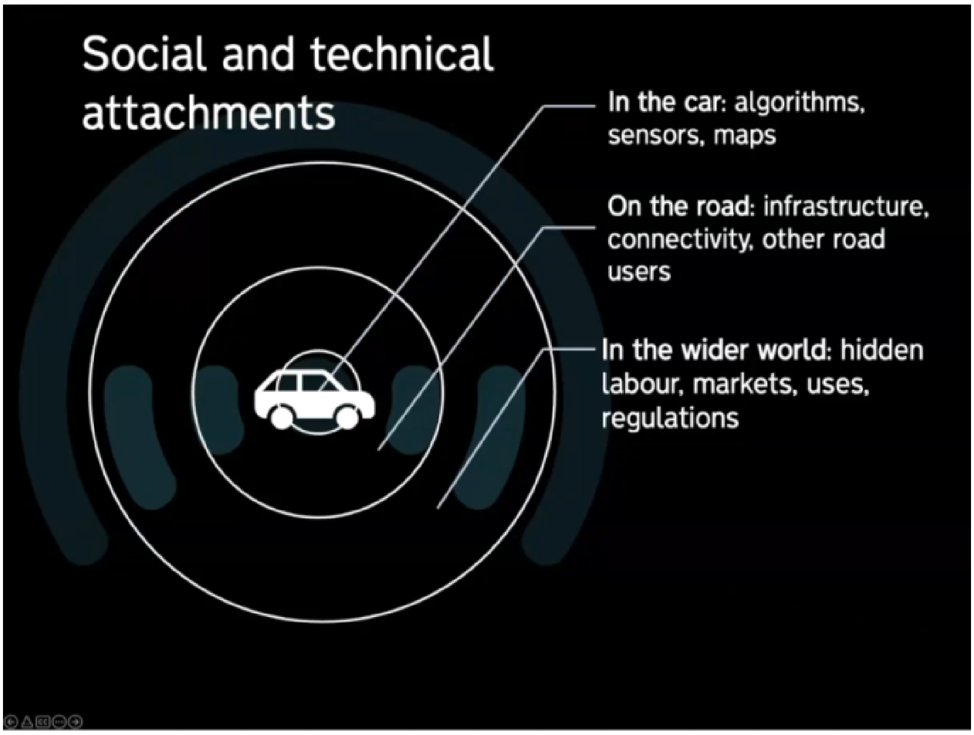
"Waymo's ultimate goal is to develop fully autonomous driving technology that can take someone from A to B, anytime, anywhere, and in all conditions."

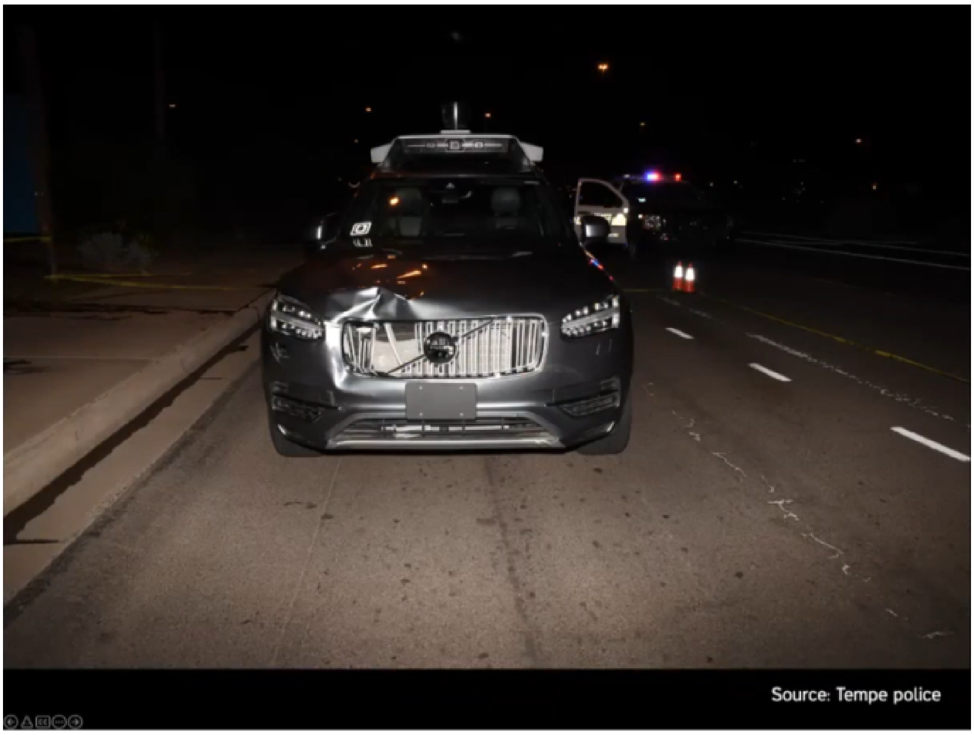
## Myths of autonomy

- The machines will drive like humans
- They will solve the problem of human error
- The tech is just around the corner
- Everyone, everywhere will benefit
- No new infrastructure required
- No new rules required


Heteronomous vehicles
• “Only a rock is truly autonomous” (Mindell)
• AVs are conditioned and ‘driven’ by people outside the vehicle
• “Ironies of automation” (Bainbridge) and heteromation (Nardi and Ekbia)
• Potential tensions with autonomous, mobile individuals


‘Everything will be designed by engineering, not by legislation... The two, the car and the road, are both essential to the realization of automatic safety. It is a job that must be done by motor-car manufacturers and road builders cooperatively.’
Normal Bel Geddes , 1940


Social and technical attachments
In the car: algorithms, sensors, maps
On the road: infrastructure, connectivity, other road users
In the wider world: hidden labour, markets, uses, regulations


Source: Tempe police

TOYOTA

Company Newsroom Mobility Sustainability IR

Aug. 30, 2021

# Resumption of Services of the Toyota e-Palette Vehicle and Additional Safety Measures at the Tokyo 2020 Paralympic Athletes' Village

Announcement

Having taken steps to ensure greater safety and security, Toyota today announces that The Tokyo Organising Committee of the Olympic and Paralympic Games has decided to resume operations of the e-Palette mobility vehicle within the Athletes' Village.

Operations of mobility services were suspended in response to an incident that occurred at the Athletes' Village on Thursday, August 26, 2021, when the e-Palette collided with a visually impaired pedestrian.

To ensure safe and secure traffic flow at the Athletes' Village, there are three crucial elements: pedestrians, vehicles, and infrastructure including guides. By analyzing this incident from the perspective of these three

Aug. 30, 2021

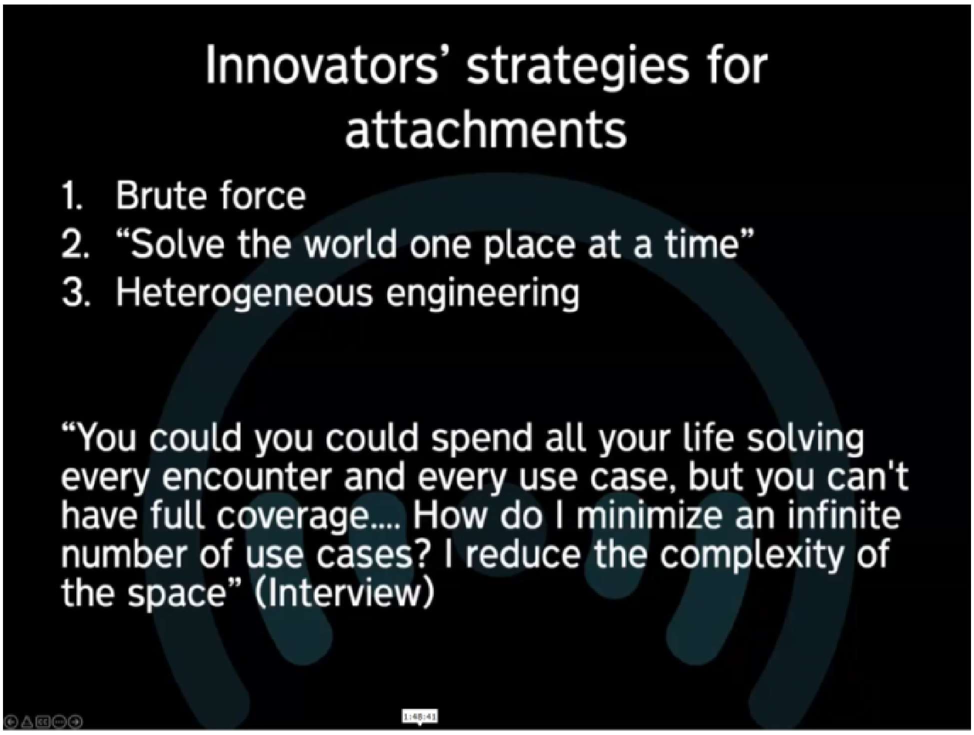
Resumption of Services of the Toyota e-Palette Vehicle and Additional Safety Measures at the Tokyo 2020 Paralympic Athletes' Village

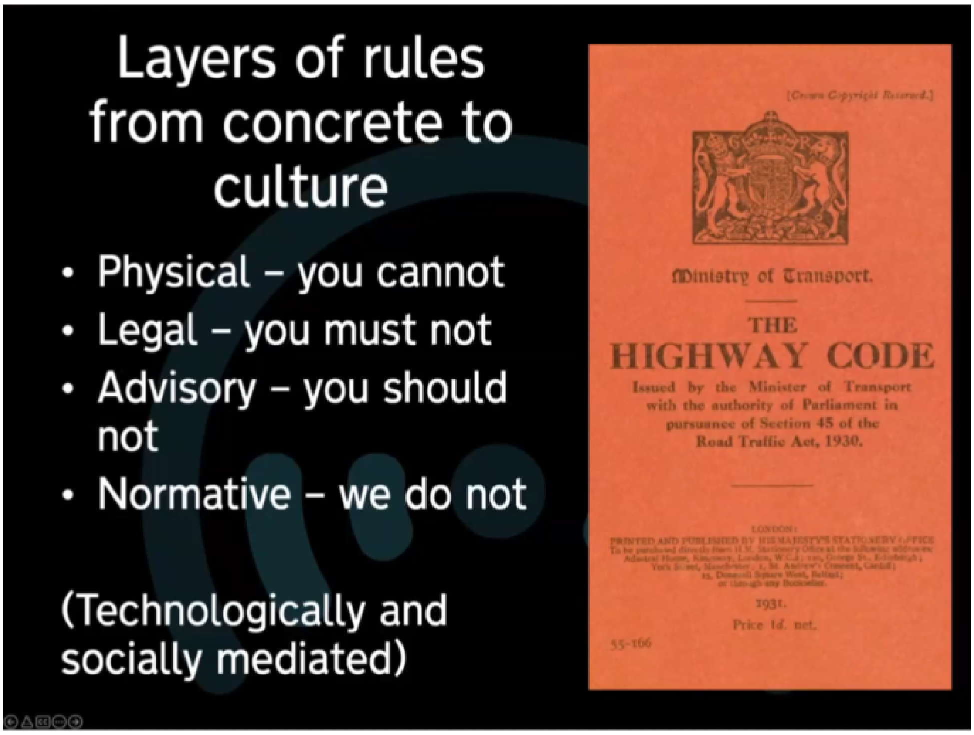
Announcement



As a result, the pedestrian entering the intersection came into contact with the vehicle.

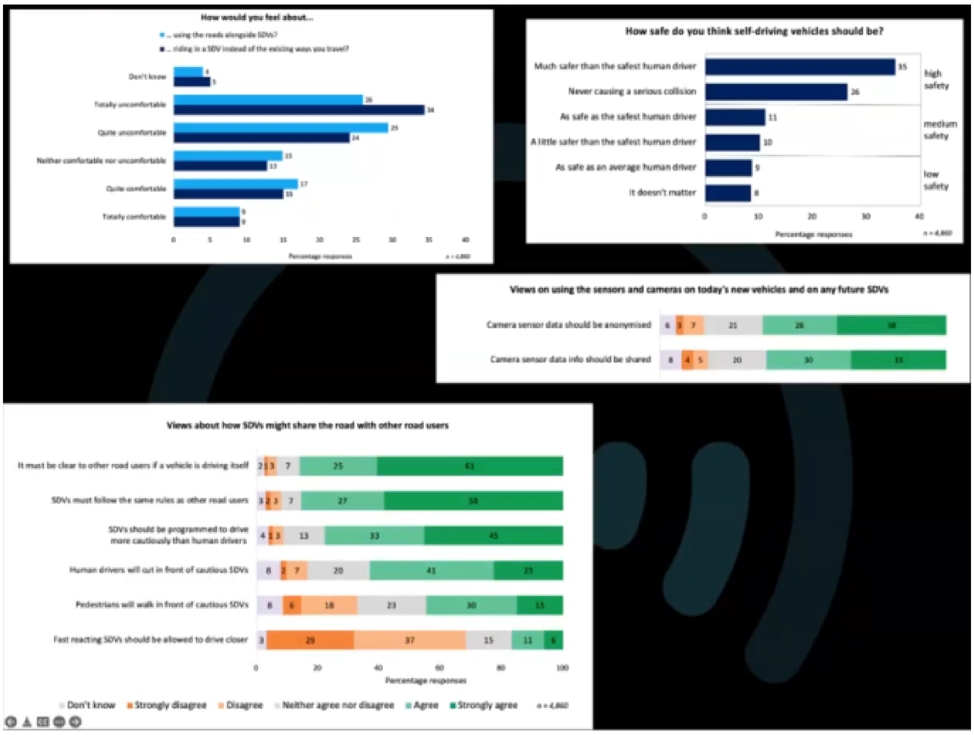
Based on the thorough verification of these facts, Toyota, together with the Organising Committee, has determined that ensuring safety at an intersection without signals is not something that can be handled by pedestrians, operators, or guides alone. It is necessary for all three parties to work together.

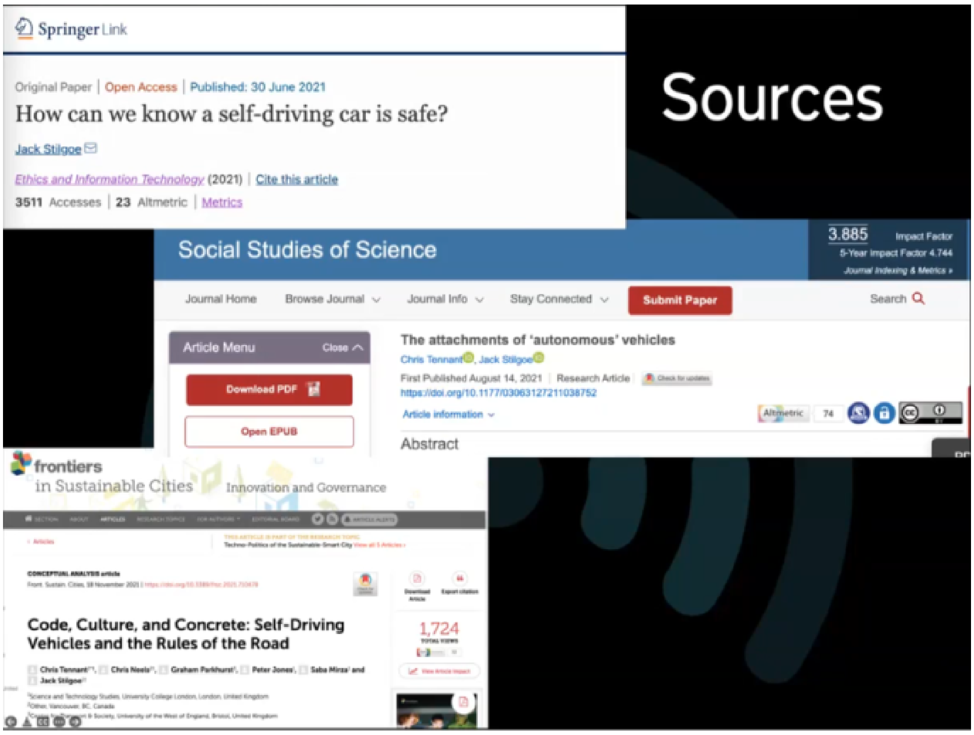

Innovators' strategies for attachments
1. Brute force
2. "Solve the world one place at a time"
3. Heterogeneous engineering
"You could you could spend all your life solving every encounter and every use case, but you can't have full coverage.... How do I minimize an infinite number of use cases? I reduce the complexity of the space" (Interview)


Layers of rules from concrete to culture
• Physical – you cannot
• Legal – you must not
• Advisory – you should not
• Normative – we do not
(Technologically and socially mediated)
Ministry of Transport.
THE HIGHWAY CODE
Issued by the Minister of Transport with the authority of Parliament in pursuance of Section 45 of the Road Traffic Act, 1930.

# Ethics and responsible innovation for AVs (forthcoming, UK CDEI/CCAV)

1. Road Safety
   - How safe is safe enough?
   - ODDs and system safety
   - Road rules
2. Explainability and Data Sharing
   - What is happening/what happened and why?
   - 'Ethical black boxes'
3. Data privacy
   - Inside and outside the vehicle
4. Fairness
   - Distribution of risk
   - Vulnerable road users
   - Biases in training data
   - Accessibility and inclusion
5. Transparency
   - Labelling, terminology and public information
   - Consultation and trials




Sources
SpringerLink
Original Paper | Open Access | Published: 30 June 2021
How can we know a self-driving car is safe?
Jack Stilgoe
Ethics and Information Technology (2021) | Cite this article
3511 Accesses | 23 Altmetric | Metrics
Social Studies of Science
3.885 Impact Factor
5-Year Impact Factor 4.744
Journal Indexing & Metrics »
Journal Home
Browse Journal
Journal Info
Stay Connected
Submit Paper
Search
Article Menu
Close
Download PDF
Open EPUB
The attachments of 'autonomous' vehicles
Chris Tennant, Jack Stilgoe
First Published August 14, 2021 | Research Article
https://doi.org/10.1177/03063127211038752
Article information
Altmetric
74
Abstract
frontiers
in Sustainable Cities
Innovation and Governance
Code, Culture, and Concrete: Self-Driving Vehicles and the Rules of the Road
1,724

# Towards Explainable and Trustworthy Autonomous Systems

**Lars Kunze**

JCR Exploratory Workshop
29 / 30 March 2022

UNIVERSITY OF OXFORD

COGNITIVE ROBOTICS GROUP
OXFORD ROBOTICS INSTITUTE

Responsible Technology Institute

## Autonomous Systems are Changing our World

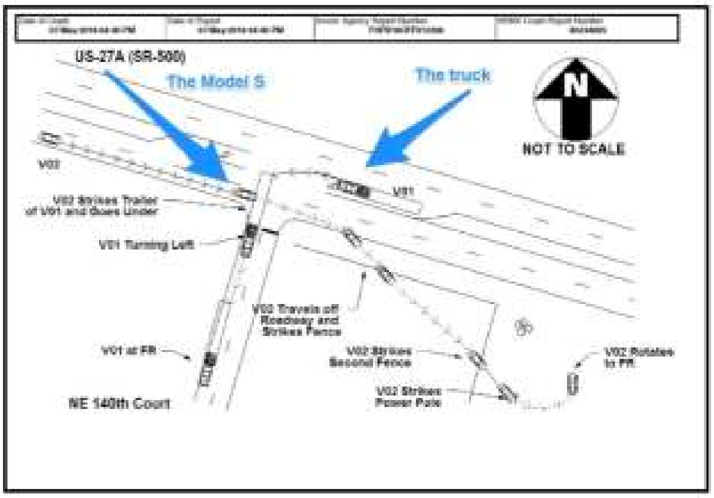
UNIVERSITY OF OXFORD


3

## Why explainable and trustworthy AS?

UNIVERSITY OF OXFORD

- Because accidents will happen!
- **Explanations** are key to understand what an AS has seen, planned to do, and why.
- **Trustworthy** systems are transparent, responsible and accountable.




Tesla Model S, May 2016, Florida

5

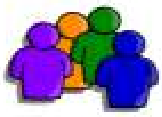
# Explainable Systems

## Explanations in Autonomous Driving: A Survey (T-ITS 2021)

**Need for Explanations:**

- Transparency
- Accountability
- Usability & Trust
- Standards & Regulations (eg GDPR)

**Stakeholders:**



- Users
- Developers, Technicians, Operators
- Regulators, Policy makers, Insurers

**Types of Explanations:**

Explanation

- Why? (factual)
- Why not? (contrastive)
- What if? (counterfactual)
- How to? (counterfactual)

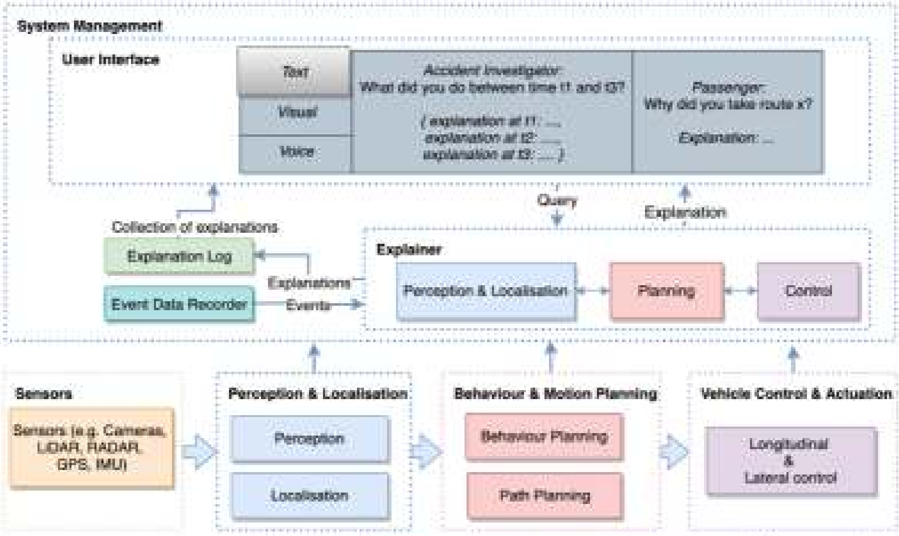
**Other Aspects of Explanations:**

- Succinctness
- Comprehensibility
- Faithfulness

[Omeiza et al 2021. IEEE Transactions on Intelligent Transportation Systems] 7

## A Framework for Explainable AVs

UNIVERSITY OF OXFORD




[Omeiza et al 2021. IEEE Transactions on Intelligent Transportation Systems ]

8

## Sense—Assess—eXplain (SAX)

UNIVERSITY OF OXFORD

The aim of the project is to build **trustworthy autonomous vehicles** that can:

- **sense** and understand their environment,
- **assess** their performance,
- **explain** their observations and actions, ...

...in **on-road/off-road** driving scenarios using **traditional/alternative sensors** under **varying environmental conditions**.

ASSURING AUTONOMY INTERNATIONAL PROGRAMME

ORI

9

## SAX Dataset: One platform, different environments

UNIVERSITY OF OXFORD

**141 hours | 3700 kilometres | 200 terabytes | >10K of labels**

- **Sensing:**
  - Radar
  - 3D LiDAR
  - GPS
  - Cameras
  - 2D LiDAR
  - Microphones
- **Control signals:**
  - steering
  - braking
  - accelerating


Demonstrating integrated systems in
**challenging real-world environments across the UK**
(Oxford, London, Milton Keynes, New Forest, Scotland)

11

## Road Commentary

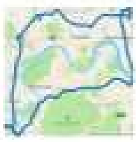
UNIVERSITY OF OXFORD

**Explanation Driving Dataset**
- 11 hours driving data
- Driver Audio Commentary

**Example:**
- Overtaking a cyclist
  (in collaboration with
  London Advanced Motorists)

**How can we generate such explanations?**

IAM RoadSmart



18

## How to generate explanations?

## From Image to Explanation

**Current lane** only allows **right turn.**

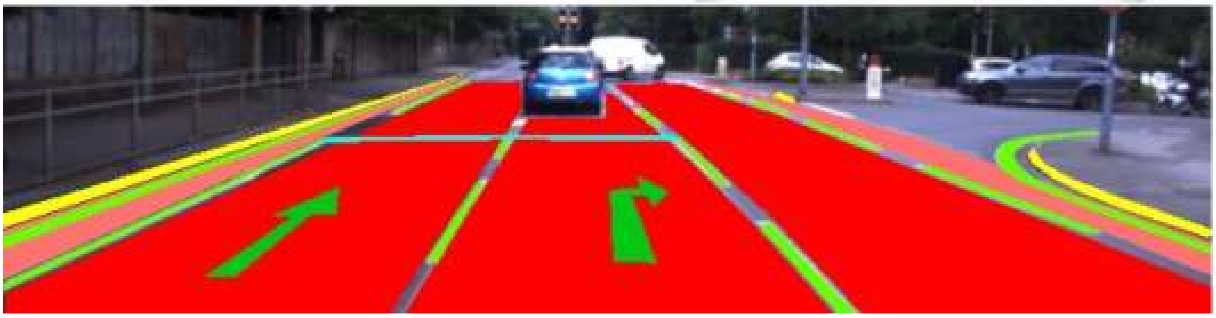
Change to **left lane**, as current goal is **straight ahead**.

## Semantic Segmentation

UNIVERSITY OF OXFORD

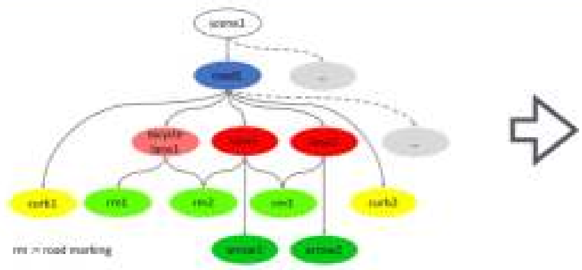
**Current lane** only allows **right turn.**

Change to **left lane**, as current goal is **straight ahead**.



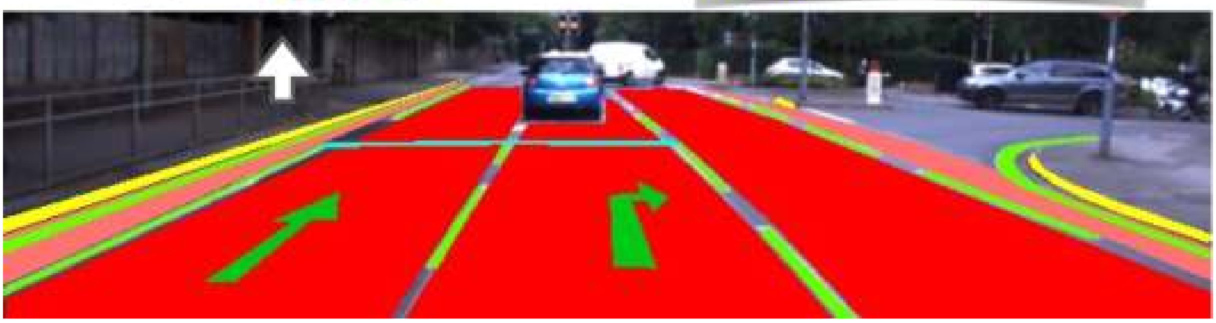
## Scene Graph

UNIVERSITY OF OXFORD



**Current lane** only allows **right turn.**

Change to **left lane**, as current goal is **straight ahead**.



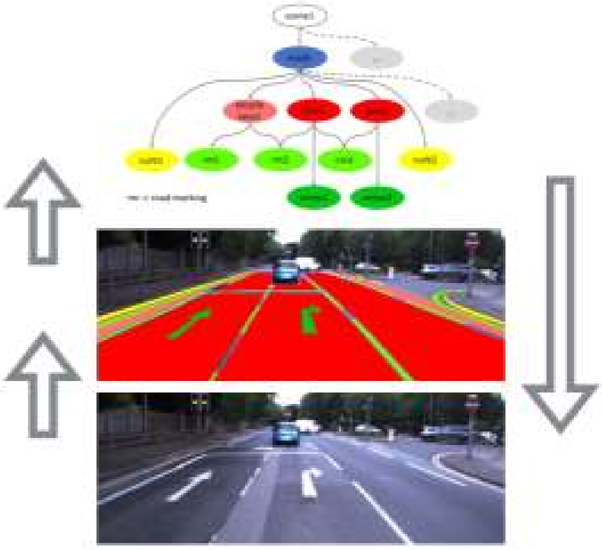
## Explanation Generation

UNIVERSITY OF OXFORD

Scene Graph

Partial Segmentation

Image (RGB)



**Planning / Decision Making:**

- *ChangeLane vs StayOnLane*

**Explanation Generation:**

- *"ChangeLane because goal location is straight ahead"*

23

## Understanding Dynamic Scenes

UNIVERSITY OF OXFORD

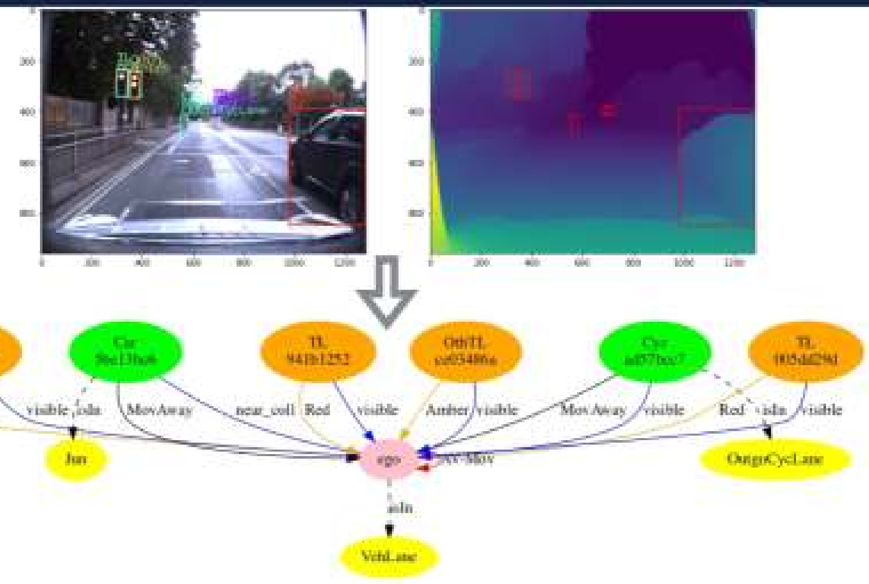
The ROad event Awareness Dataset [Singh et al 2021]

- 18 annotated drives taken from Oxford RobotCar dataset
- Road Event = (Agent, Action, Location)

ORI's RobotCar

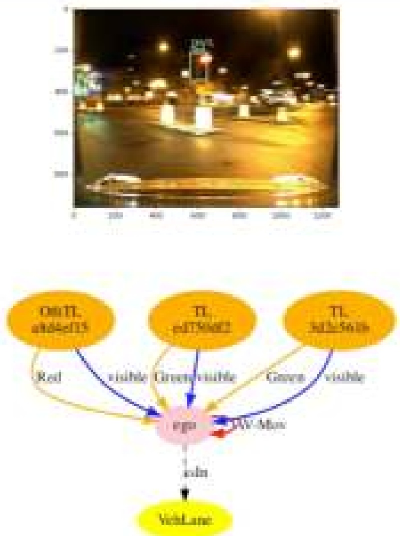
## Scene Graph for Road Events

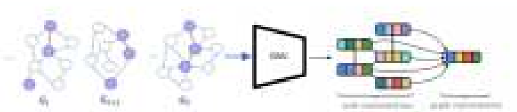
UNIVERSITY OF OXFORD

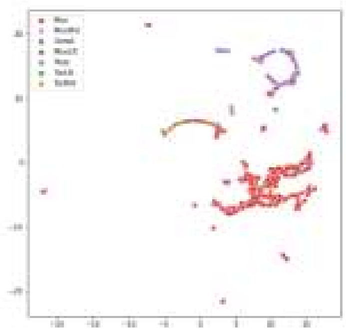



## Evolving & Clustering Scene Graphs

UNIVERSITY OF OXFORD





Clustering scene graph embeddings based on ego-vehicle action (eg Mov / Stop actions)

## Predicting and explaining driving decisions in natural language

UNIVERSITY OF OXFORD

"The vehicle stops because the traffic light turns red"

27

## Preliminary Results: End-to-End Explanation Generation

UNIVERSITY OF OXFORD

**Input:**
- Images (Sequence)

**Output**:
- NL Explanation

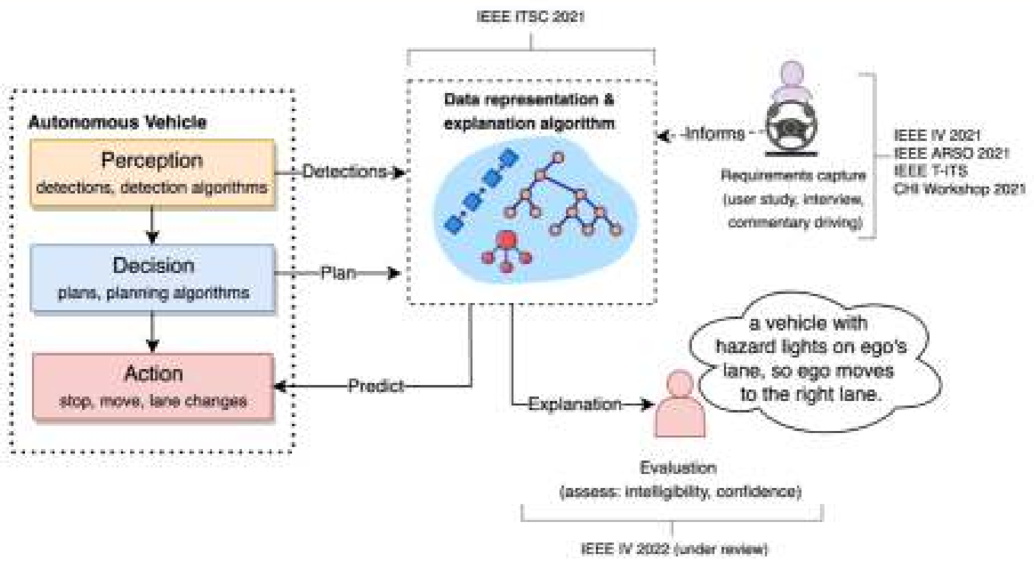
**Training data:**
- Speed, Accel & Course
- Textual Action Description + Explanation

| Generated Sentence | Ground Truth |
|---|---|
| „the car slows slowing a stop stop" + „because the light is red" | „the car slows to a stop" + „because the light is red" |
| „the car slows slowing a stop stop" + „because the light is red red" | „the car slows to a stop" + „because the light is red" |
| „the car is stopped" + „because the because is red" | „the car is stopped" + „because the light is red" |
| „the car is driving forward because the" + „is the traffic is" | „the car is driving forward" + „because traffic is moving freely " |

**Future work: Integration with Scene Graphs**

29

## Summary: Explainable Systems



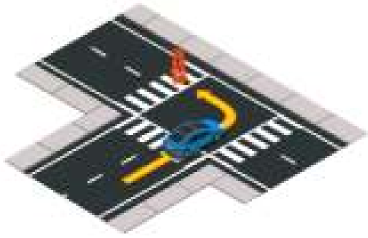

30

## Trustworthy Systems

## RoAD — Responsible AV data

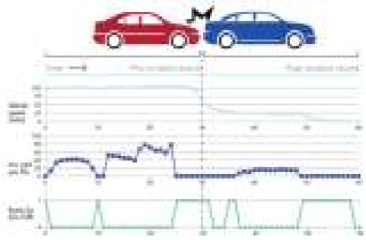
UKRI Trustworthy Autonomous Systems Hub | University of Oxford

### Ethical, legal, and societal challenges of using data from AVs



Safety-critical Scenarios

Simulation



Standards & Regulation

Responsible Technology Institute | ORI | University of Leeds Institute for Transport Studies (ITS) | oxbotica | ITU | AXA

32

## Road Overview

University of Oxford

**Safety-critical Scenarios**

- Shared vs full autonomy (UIC vs NUIC)
- Transition demands
- VRU collisions
- Low impact collisions
- Runover events
- Near-miss events
- Sensor malfunction
- "The Molly Problem"

**Simulation-based Data Collection**

(Courtesy TRL)

**Regulations & Standards**

**UNECE**

- Event Data Recorder (EDR)
- Data Storage System for Automated Driving (DSSAD)

**ITU FG-AI4AD**

- Automated driving safety data protocol

**UK Law Commission**

- 3rd consultation paper

**Testing Utility & Accessibility of Data**

Safety | Legal Usefulness | Ethical Implications

**Evaluation of Public Perception**

## Expert Interviews

- Findings & Themes:
  - **Recorded Data**
    - Video
    - Location
    - Near Misses
  - **Use & Utility of Data**
    - Safetey, insurance, accidents, other crimes

IEEE Transactions on Intelligent Transportation Systems (T-ITS) under review

Social, Ethical, and Legal Challenges for Autonomous Vehicle Data Recorders

TABLE I
INTERVIEWED STAKEHOLDERS: CODE, TYPE AND FOCUS

| Code | Type of stakeholder | Focus of stakeholder |
|---|---|---|
| CS-04 | Civil society | Equestrian road users |
| CS-07 | Civil society | Pedestrians and other non-vehicular road users |
| I-16 | Industry | Data security company |
| P-29 | Professional | Smart Cities and data |
| CS-36 | Civil society | Police - crime investigation |
| P-05 | Professional | Law Commission |
| I-34/I-35 | Industry | Insurer |
| S-30 | Academia | Autonomous vehicles |
| S-15 | Academia | Roboticist |
| PS-99 | Policymaking/ governmental | Federal Ministry of Transport and Digital Infrastructure (Germany) |
| I-09 | Industry | Autonomous vehicle software |
| I-03 | Industry | Insurer |
| S-14 | Academia | Cyberlaw specialist |
| I-12 | Industry | Data management consultant |
| P-11 | Professional | Aviation lawyer |
| P-13 | Professional | Former air accident investigator |
| CS-24 | Civil society | Cycling |
| PS-04 | Policymaking/ governmental | ITU Focus Group on AI for Autonomous and Assisted Driving |
| S-20 | Academia | Aggregated Homologation-proposal for Event Recorder Data for Automated Driving (AHEAD) |
| I-02 | Industry | AV Manufacturer/design |

34

## RoAD – Software tools for recording AV data in CARLA

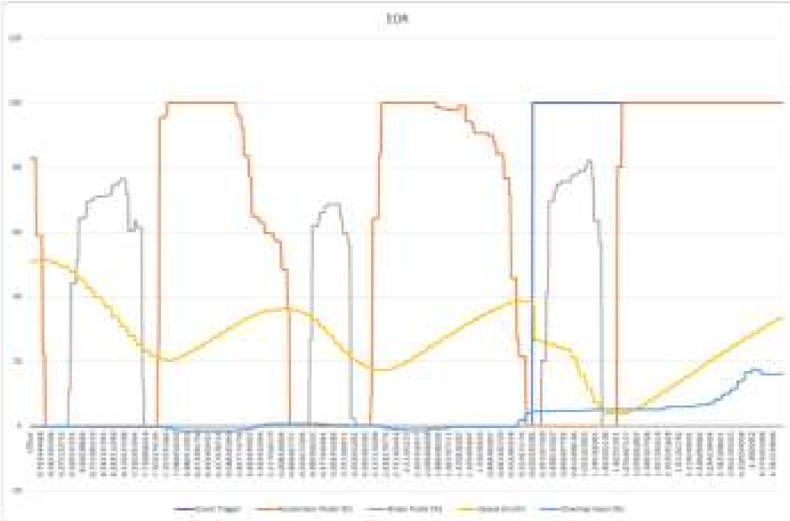
Example Scenario

**RoAD Recorder:**
https://github.com/cognitive-robots/road-recorder

35

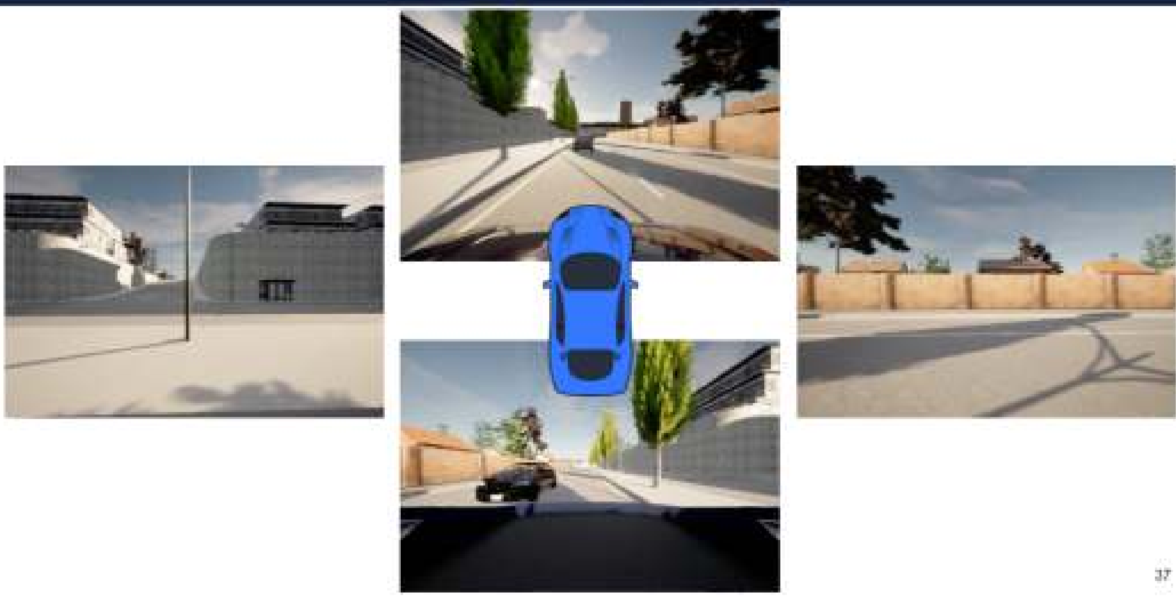
## RoAD - Event Data Recorder

Date-Time
Timestamp
Offset
Event Trigger
Accelerator Pedal (%)
Brake Pedal (%)
Delta-V Lateral (m/s)
Delta-V Longitudinal (m/s)
Engine RPM (rpm)
Altitude (m)
Latitude
Longitude
Lateral Acceleration (m/s^2)
Lateral Velocity (km/h)
Longitudinal Acceleration (m/s^2)
Longitudinal Velocity (km/h)
Normal Acceleration (m/s^2)
Normal Velocity (km/h)
Service Brake
Speed (km/h)
Steering Input (%)



36

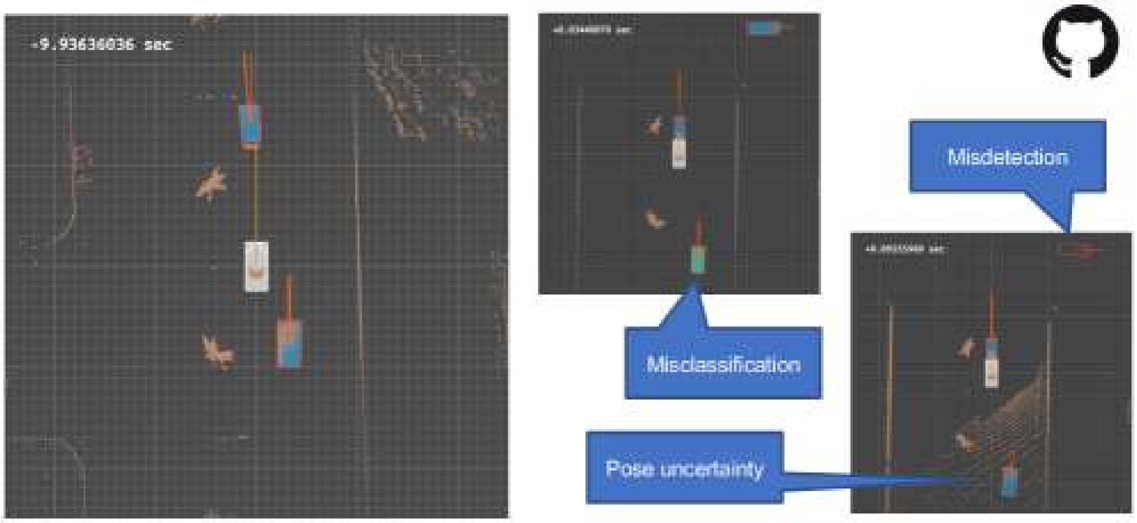
## RoAD - 360 Camera



37

## RoAD - World Model

UNIVERSITY OF OXFORD




Ground Truth & Perception

39

## RoAD – Recording Near Misses using Smart Triggers

UNIVERSITY OF OXFORD

40

## Upcoming: RobotCycle - Safety of VRU in Oxford

41

## Survey: Expectations concerning data recorders in AVs

| No. | Recording devices in autonomous vehicles will | Yes | No | Don't know |
|---|---|---|---|---|
| 1 | Increase safety in self-driving cars | 77,5% | 7,03% | 15,2% |
| 2 | Verify liability in case of accidents | 91,1% | 5,4% | 3,05% |
| 3 | Increase trust in AVs | 51,8% | 34,3% | 24% |
| 4 | Make autonomous driving more comfortable | 41,4% | 37,9% | 20,7% |
| 5 | Decrease insurance costs | 44,9% | 33,2% | 21,6% |
| 6 | Be a new business opportunity for big-data companies | 68,6% | 17,7% | 13,6% |
| 7 | Be a way to make roads and cities safer | 63,3% | 27,2% | 9,5% |
| 8 | Be an opportunity to enjoy a personalised experience in an AV | 37% | 34,1% | 28,9% |
| 9 | Be an opportunity to enjoy benefits derived from sharing data to companies (such as insurance discounts) | 35,2% | 30,1% | 34,8% |
| 10 | Be a threat to privacy | 73,7% | 19,9% | 6,3% |
| 11 | Be a new target for cyber attacks | 82,3% | 11,7% | 6% |
| 12 | Reduce our freedom | 50% | 29,7% | 20,3% |
| 13 | An indirect way of surveilling (monitoring) citizens | 73,1% | 15,8% | 11% |

RoAD project - Online survey (Prolific): 317 respondents from the UK

## Survey: What are the key determinants of **trust in Avs?**

UNIVERSITY OF OXFORD

| No. | Trust in AVs depends on: | Yes | No | Don't know |
|---|---|---|---|---|
| 1 | being able to investigate the cause of an accident | 88% | 4,8% | 7,3% |
| 2 | being able to find someone responsible (eg user, manufacturer) in case of an accident | 83,2% | 4,7% | 12% |
| 3 | what the cars look like | 13,7% | 67,3% | 19% |
| 4 | ensuring the right punishment for wrongdoing | 64,2% | 15% | 20,8% |
| 5 | ensuring mistakes do not happen again | 85,7% | 3,2% | 11,1% |

RoAD project - Online survey (Prolific): 317 respondents from the UK

## Survey: Attitude towards the use of data in **near miss events**

UNIVERSITY OF OXFORD

| No. | Indicate the extent to which they agree or disagree with the following statements: | Agree | Disagree |
|---|---|---|---|
| 1 | Insurers of vehicles should be provided with a periodic aggregated report summarising near miss events | 68.7% | 31.3% |
| 2 | Insurers of vehicles should be provided with all data related to near miss events | 45.9% | 54.1% |
| 3 | The driver/operator should be provided with a periodic aggregated report summarising near miss events | 88.6% | 11.4% |
| 4 | The driver/operator should be provided with all data related to near miss events | 83.9% | 16.1% |
| 5 | Anyone involved in the near miss event (including the driver/operator, passengers, pedestrians, those in another vehicle) should be allowed to access all data related to the near miss event | 47.2% | 52.8% |
| 6 | An independent commission or body formed to investigate automated vehicle accidents and safety should be provided with a periodic aggregated report summarising near miss events | 77.5% | 22.5% |
| 7 | An independent commission or body formed to investigate automated vehicle accidents and safety should have access to all data related to near miss events | 68.4% | 31.6% |

RoAD project - Online survey (Prolific): 317 respondents from the UK


RAILS — Responsible AI for Long-term TAS
UKRI Trustworthy Autonomous Systems Hub
UNIVERSITY OF OXFORD
Integrating Responsible AI and Socio-legal Governance
Assurance
Governance
Corner cases
Post-deployments
Adaptive Frameworks
aws
TRL THE FUTURE OF TRANSPORT
oxbotica
ITU
Responsible Technology Institute
ORI
UNIVERSITY OF LEEDS
Institute for Transport Studies (ITS)
UCL
UNIVERSITY of York
46


Summary
UNIVERSITY OF OXFORD
**Explainability** and **Trustworthiness** are key for the next generation of AS
Projects:
• Sense – Assess – eXplain (SAX)
• Responsible AV data (RoAD)
• Responsible AI for Long-term TAS (RAILS)
• RoAD Recorder: https://github.com/cognitive-robots/road-recorder
ORI
UNIVERSITY of York
UCL
UNIVERSITY OF LEEDS
Institute for Transport Studies (ITS)
aws
oxbotica
TRL THE FUTURE OF TRANSPORT
AXA
ITU
UKRI
Engineering and Physical Sciences Research Council
UKRI Trustworthy Autonomous Systems Hub
ASSURING AUTONOMY
INTERNATIONAL PROGRAMME
48

## Annex XIV. Man, Machine, or In Between: The Process of Investigations Into Automation


# SAE J3016 and ISO 22736 Taxonomy

Contain definitions for features and levels of control such as:

Automation of a feature versus autonomous for a vehicle,
Advanced driving assistance systems (ADAS) and dynamic driving tasks (DDT)
SAE Levels 0-5 with automated driving systems (DDS) in Levels 3-5
Operational Design Domains (ODD) , etc

This presentation is about the process of investigation

Wording is therefore generalized and not using these standardized definitions

HowItBroke.com

# Aviation Has Had Numerous Autopilot Involved Accidents To Learn From

Boeing 737 MAX, Ethiopian flt 302
Ethiopia, March 10, 2019, 157 fatal

Boeing 737 MAX, Lion Air flt 810
October 29, 2018, 189 fatal

AOA sensor failure coupled with design error and training leading to improper pilot responses

Boeing 777, Emirates flt 521
Dubai, August 2016, 1 fatal, 38 injured
Pilot expected go-around thrust not realizing ground contact changed flight mode

Airbus A330, Air France flt 447
Atlantic Ocean, June 1, 2009, 228 fatal
Ice in airspeed probe led to pilot errors

Boeing 737-800, Turkish flt 1951
Amsterdam, February 25, 2009, 9 fatal, 120 injured
Radar altimeter input error and Boeing vs Airbus training differences

Boeing 737-800, Kenya Airways flt 507
Douala, May 5, 2007, 114 fatal
Lack of feedback that autopilot had not engaged when expected to

From Only These Six:
735 fatal, 158 injured

HowItBroke.com

# Triple redundant systems in aviation - yet ...

...loss of control found in 43% of 2010-2014 fatal commercial accidents (37)

The #1 Autopilot related cause of accidents is human interface

Typically perception of autopilot performance was not what was expected

Boeing 777, Emirates flt 521 , Dubai

The #2 Cause was pilots disconnecting or getting "behind" the airplane

Tesla X, Mountain View, California

HowItBroke.com

# "What's it *[the autopilot]* doing now?"

Common airline crew saying

"Disappointment *[causing stress and errors]* is **the gap that exists between our expectation and reality"** – Maxwell

Our goal is to not let reality differ from expectations

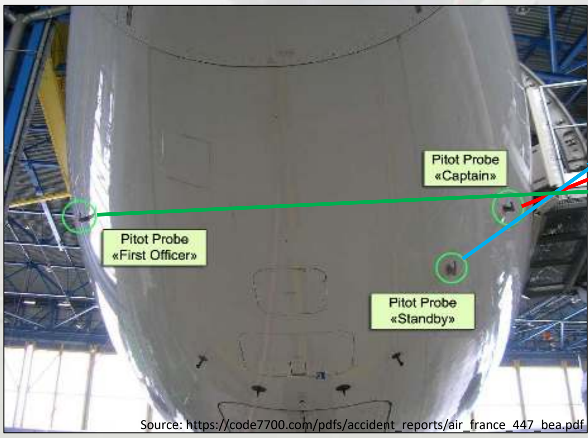
Accident investigations provide the ultimate test and judgement

HowItBroke.com

# Case Example For What The Process Can Do

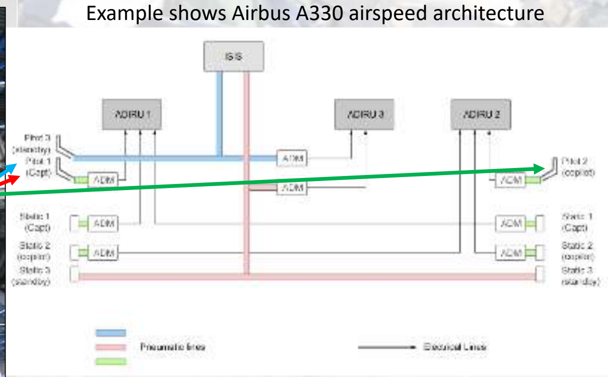
- Air France Flight 447, 228 fatal

Airbus A330 has triple redundant airspeed systems cross checking each other
Differences in data result in two systems voting out third




Source: https://code7700.com/pdfs/accident_reports/air_france_447_bea.pdf

Example shows Airbus A330 airspeed architecture




HowItBroke.com

## Air France Flight 447 Circumstances


June 1, 2009, Rio de Janeiro – Paris, 2:14 am in clouds
First Officer (right seat) was pilot flying
Investigation found that:

Ice build-up on one airspeed sensor disrupted that one airspeed system
- Two flight computers voted out the inconsistent inputs from the third system
- Autoflight protections degrade in dual computer system (called Alternate Law 2)
- Warning alerts were displayed for pilots
- Autopilot disengaged and less experienced First Officer began to fly by hand

Source: https://code7700.com/pdfs/accident_reports/air_france_447_bea.pdf

HowItBroke.com

## Air France Flight 447 Findings

One wing moved down slightly when autopilot disconnected
First Officer response was excessive to the slight correction needed
  He created an increasing series of pitch inputs, each further up and down
  The airplane slowed enough to stall [wing lost lift] and began to fall
Repeated misinterpretations in stressful situation led to further improper responses


Source: https://code7700.com/pdfs/accident_reports/air_france_447_bea.pdf

## How Did The Process Develop Those Findings When

Location of the missing airplane was unknown
Debris was fragmented and scattered on ocean bottom
Numerous countries were involved, including:
  Where airplane and components were made,
  Brazilian departure,
  French arrival,
  Citizens of numerous countries


Who took the lead?
  Standardized process is in ICAO Annex 13

## Various Investigation Processes

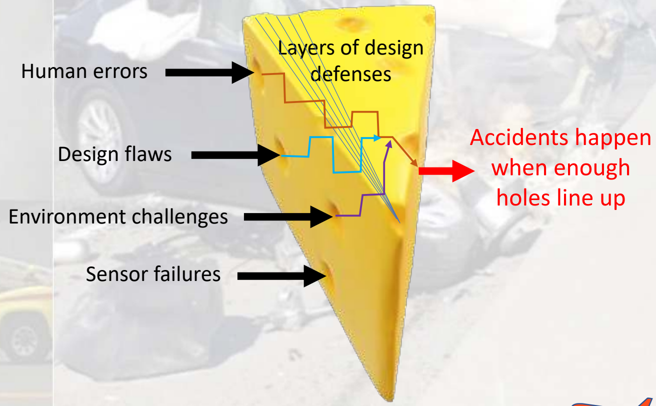
Simplest is to keep asking "Why?"

5 Why Method:
Why – Battery is dead
Why – No charge system output
Why – Alternator belt broken
Why – Belt worn to failure
Why – Inadequate maintenance

Too simplistic for most problems

Swiss cheese model

Design defenses and most accidents involve multiple contributing factors




HowItBroke.com

## Investigations Follow Time-Proven Process

**FIRST** – Who has jurisdiction and responsibility to lead the investigation?
Four types of investigation are:

**Criminal** - Government
**Safety** - Government
**Civil** – Litigation about monetary damages between individuals &/or companies
**Technical** – Typically manufacturers

Government has first rights, especially with fatalities
Companies support Government
Government must recognize proprietary needs of companies

**SECOND** – Leadership must agree on process or how to refine to circumstances

**THIRD** – Gather facts BEFORE analysis

HowItBroke.com

# Collect Factual Data By Breaking Into Focal Groups

Groups work in defined focal areas, such as:
- Driver and human factors
  - People involved, their training, and backgrounds
- Vehicle(s) and systems design,
  - Previous similar events,
  - Maintenance records,
- Roadway, including barriers, markings, etc
- Weather and other environmental factors,
- Traffic, communications, radar or other recordings,

2017 Mountain View, California

Conduct daily organizational meetings
- Share factual findings with other groups and leadership

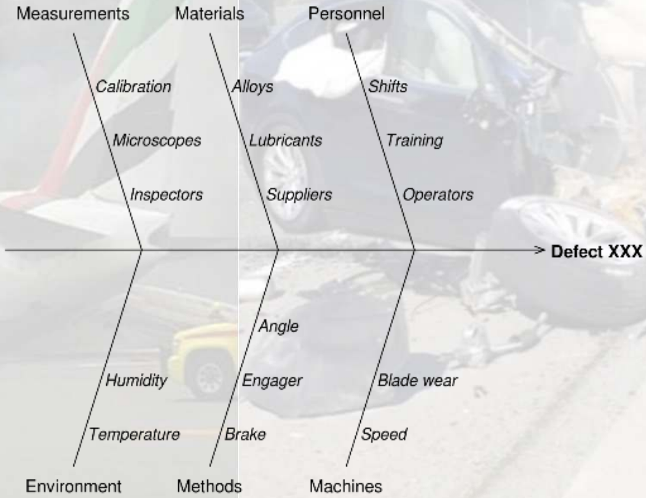
HowItBroke.com

# Record And Categorize Facts Found

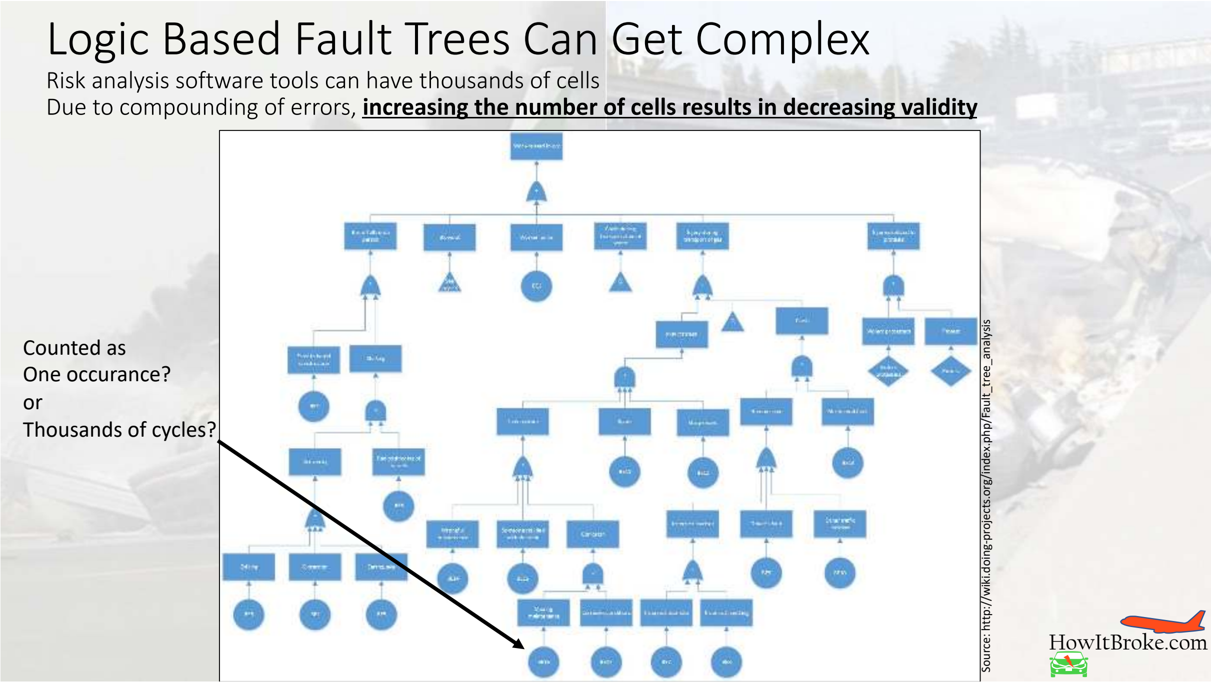
Failure and Risk Analysis Typically Based in The 5 Ms & E

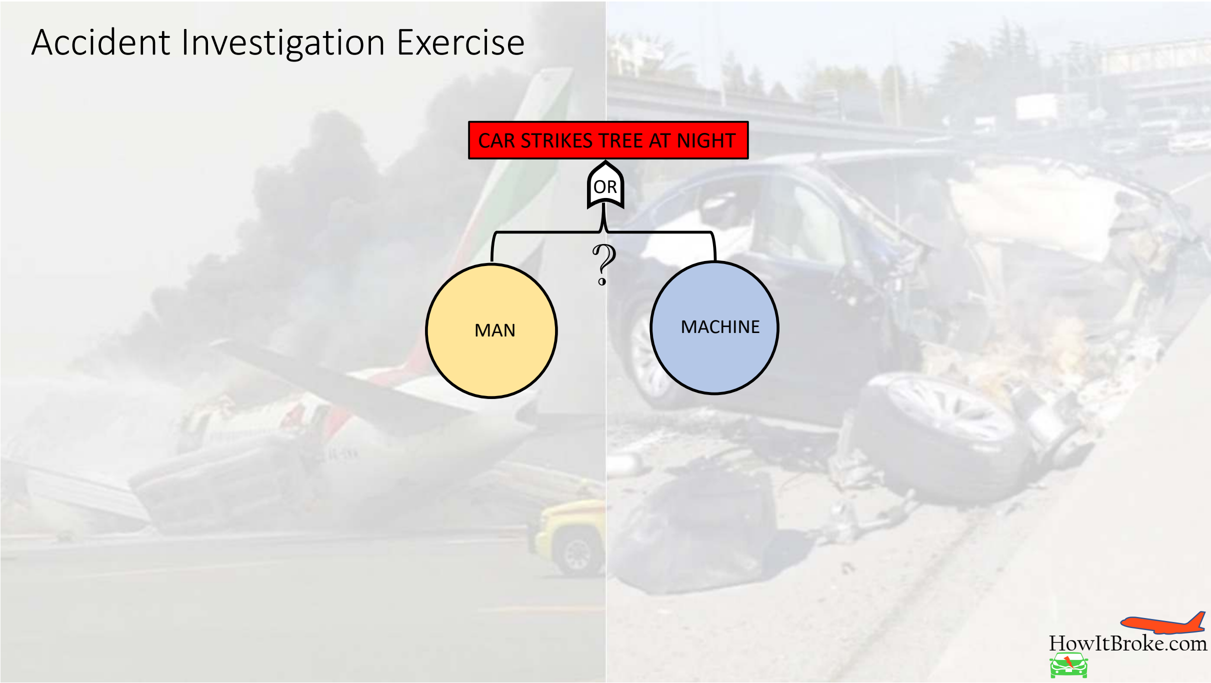
- Man
- Machine
- Method
- Material
- Measurement
- Environment

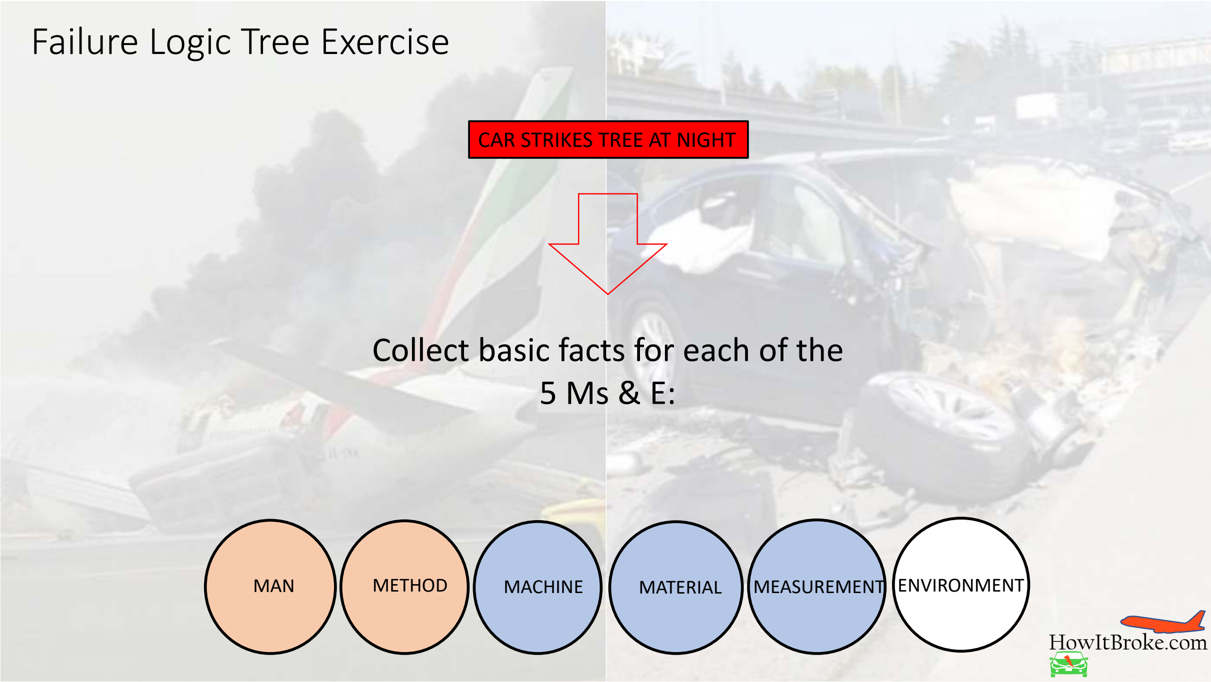
Adapted from 1920s Ishikawa "Fish Bone" diagrams

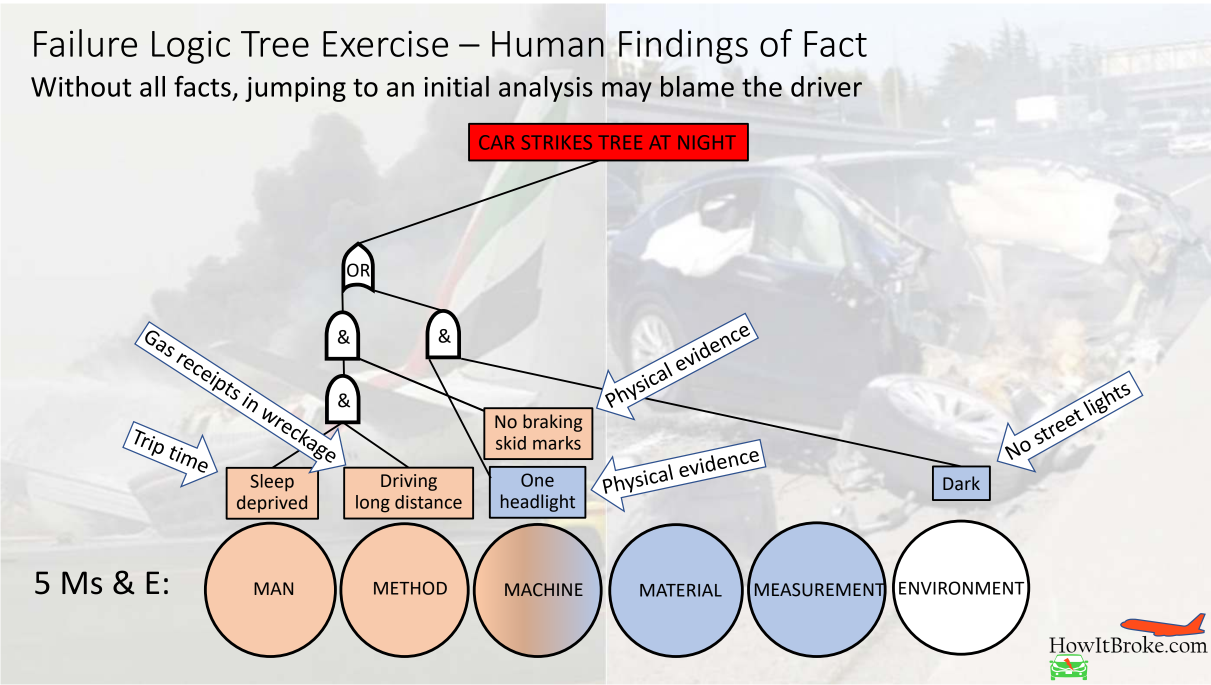


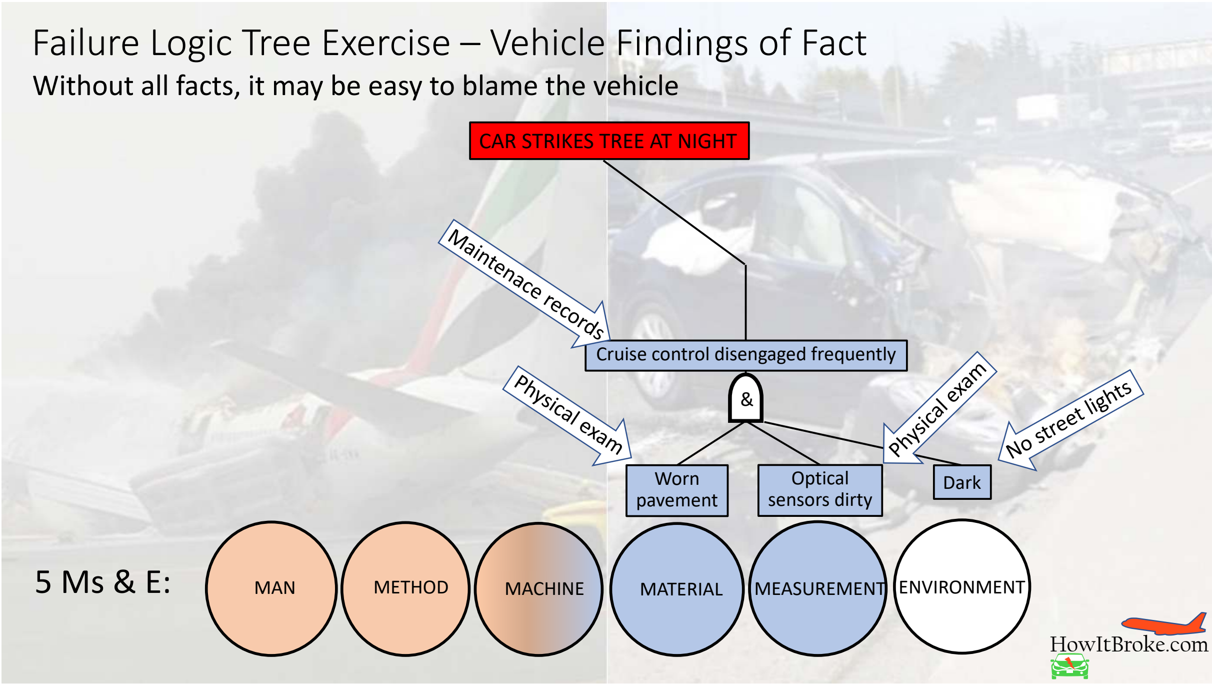

HowItBroke.com


Logic Based Fault Trees Can Get Complex
Risk analysis software tools can have thousands of cells
Due to compounding of errors, increasing the number of cells results in decreasing validity
Counted as
One occurance?
or
Thousands of cycles?
Source: http://wiki.doing-projects.org/index.php/Fault_tree_analysis
HowItBroke.com

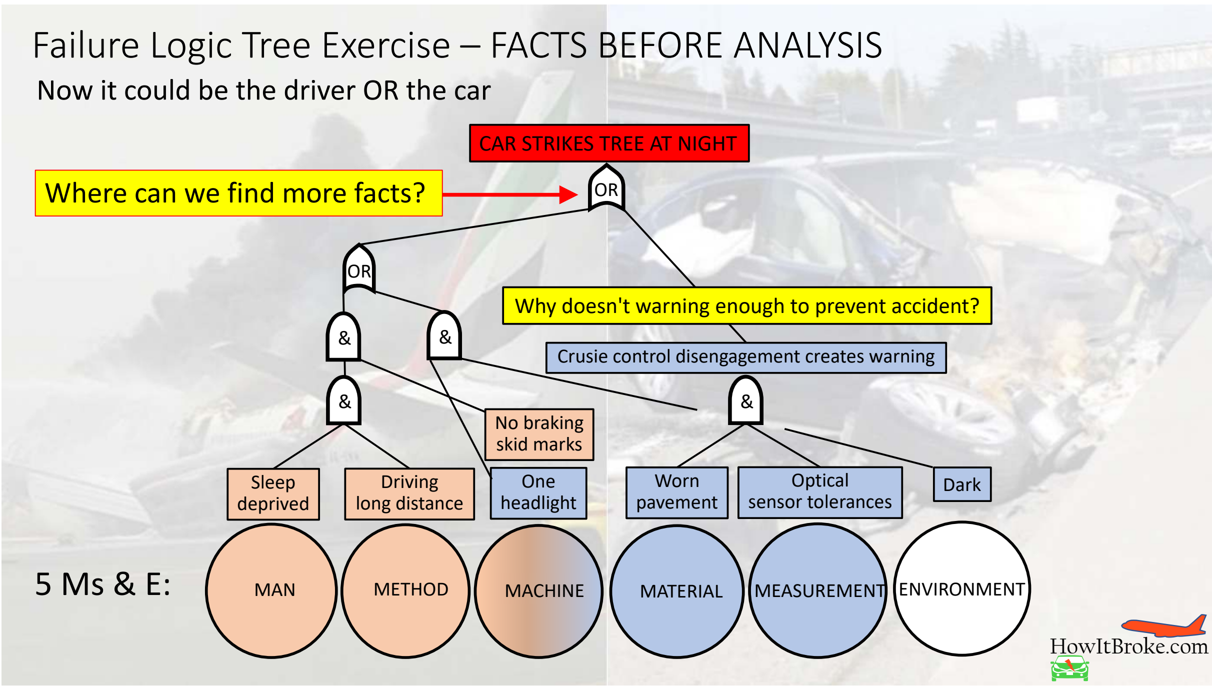

Accident Investigation Exercise
CAR STRIKES TREE AT NIGHT
OR
?
MAN
MACHINE
HowItBroke.com


Failure Logic Tree Exercise
CAR STRIKES TREE AT NIGHT
Collect basic facts for each of the
5 Ms & E:
MAN
METHOD
MACHINE
MATERIAL
MEASUREMENT
ENVIRONMENT
HowItBroke.com


Failure Logic Tree Exercise – Human Findings of Fact
Without all facts, jumping to an initial analysis may blame the driver
CAR STRIKES TREE AT NIGHT
OR
&
&
&
Gas receipts in wreckage
Trip time
Physical evidence
No braking skid marks
Physical evidence
No street lights
Sleep deprived
Driving long distance
One headlight
Dark
5 Ms & E:
MAN
METHOD
MACHINE
MATERIAL
MEASUREMENT
ENVIRONMENT
HowItBroke.com



Failure Logic Tree Exercise – Vehicle Findings of Fact
Without all facts, it may be easy to blame the vehicle
CAR STRIKES TREE AT NIGHT
Maintenace records
Cruise control disengaged frequently
&
Physical exam
Physical exam
No street lights
Worn pavement
Optical sensors dirty
Dark
5 Ms & E:
MAN
METHOD
MACHINE
MATERIAL
MEASUREMENT
ENVIRONMENT
HowItBroke.com

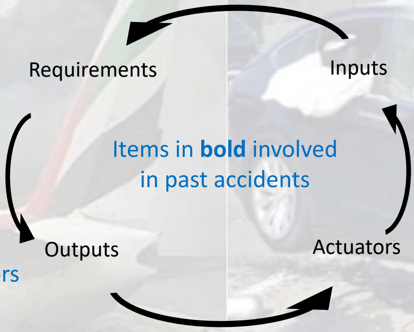



Failure Logic Tree Exercise – FACTS BEFORE ANALYSIS
Now it could be the driver OR the car
CAR STRIKES TREE AT NIGHT
Where can we find more facts?
OR
OR
&
&
&
&
Why doesn't warning enough to prevent accident?
Crusie control disengagement creates warning
No braking skid marks
Sleep deprived
Driving long distance
One headlight
Worn pavement
Optical sensor tolerances
Dark
5 Ms & E:
MAN
METHOD
MACHINE
MATERIAL
MEASUREMENT
ENVIRONMENT
HowItBroke.com

## Machine Side - Continuous Loop of Automation Systems

### Brains
**Design assumptions**
  Potential software **conflicts**
**Databases** & lookup tables
Calculate position
Compute delta to requirement
Buffers, timers, and filters
Compute needed corrections

Guidance commands to actuators
**Displays** to humans




### Senses (& feed-back)
Driver mechanical & **switches**
GPS & other **NAV**
Camera and **optical sensors**
RADAR, LIDAR, & **RF based**
**Environmental sensors**
Feedback of device positions

### Muscles
**Mechanical**
**Electric**
**Hydraulic**

Most of these get captured in some record

HowItBroke.com

## Recordings

Frequently embedded in multiple devices for various types of information

Vehicle devices typically not hardened like aviation "Black Boxes

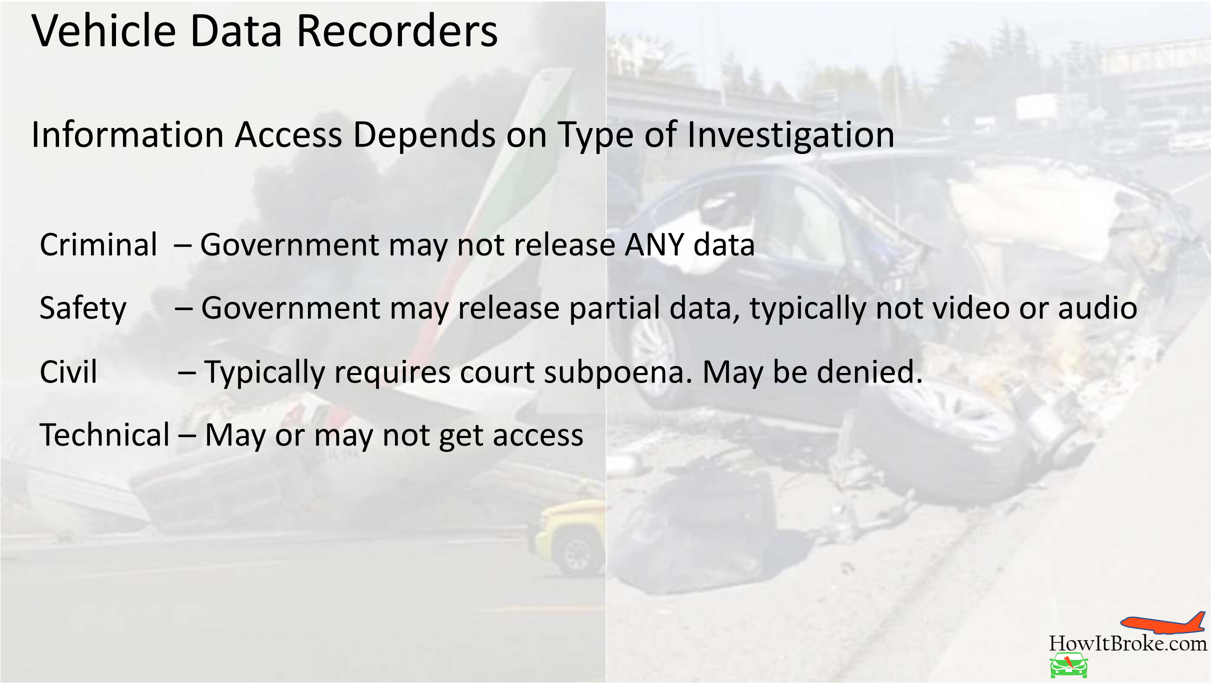
May contain dozens to thousands of parameters such as:

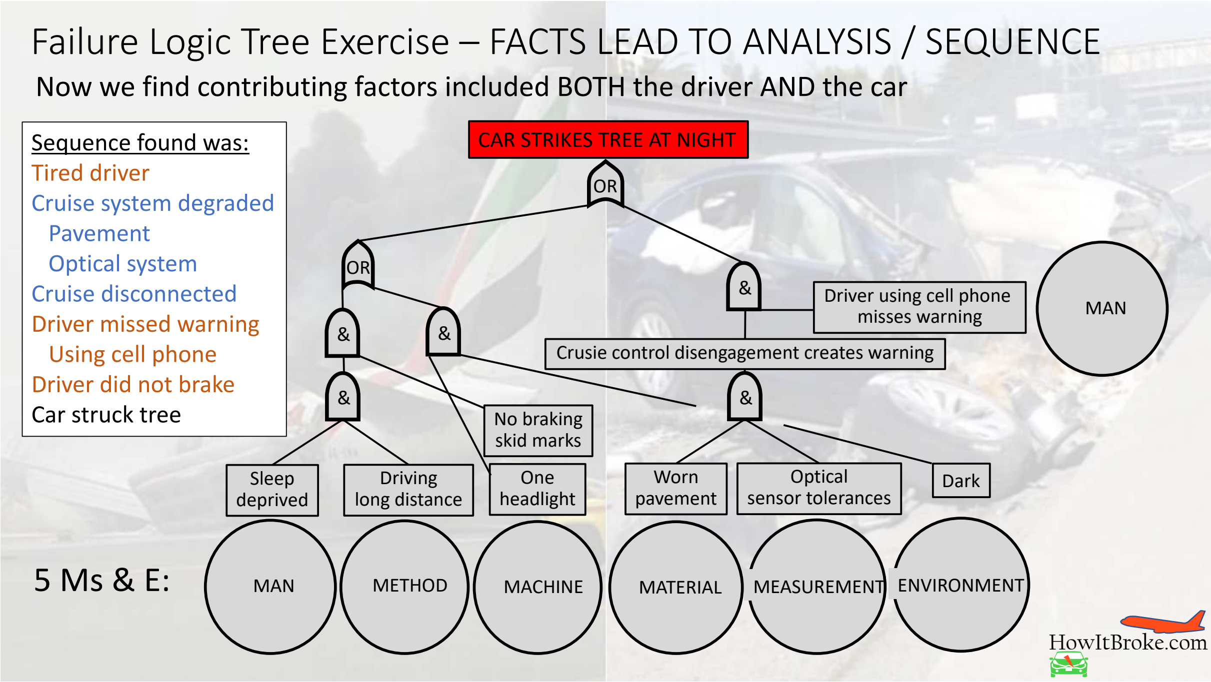
Speed, Lat/Long (GPS), seat belt use, airbag deployment, impact sensor states, fault logging (OBD), automation engagement and level, cell temps and detailed EV battery data, motor temp, transmission status, ABS, ESC, throttle position, atmospheric pressure, OAT, headlight use, wiper use, door alerts, etc,

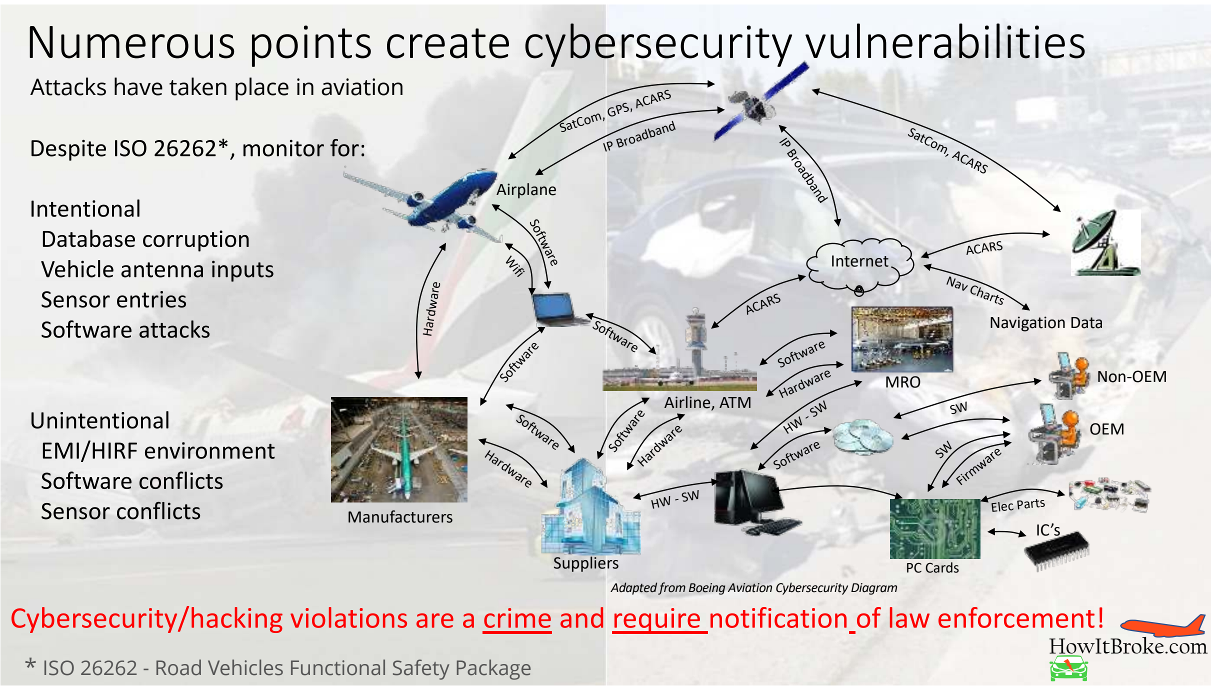
Parameters recording rates differ (example: seatbelt status vs vehicle speed)

HowItBroke.com

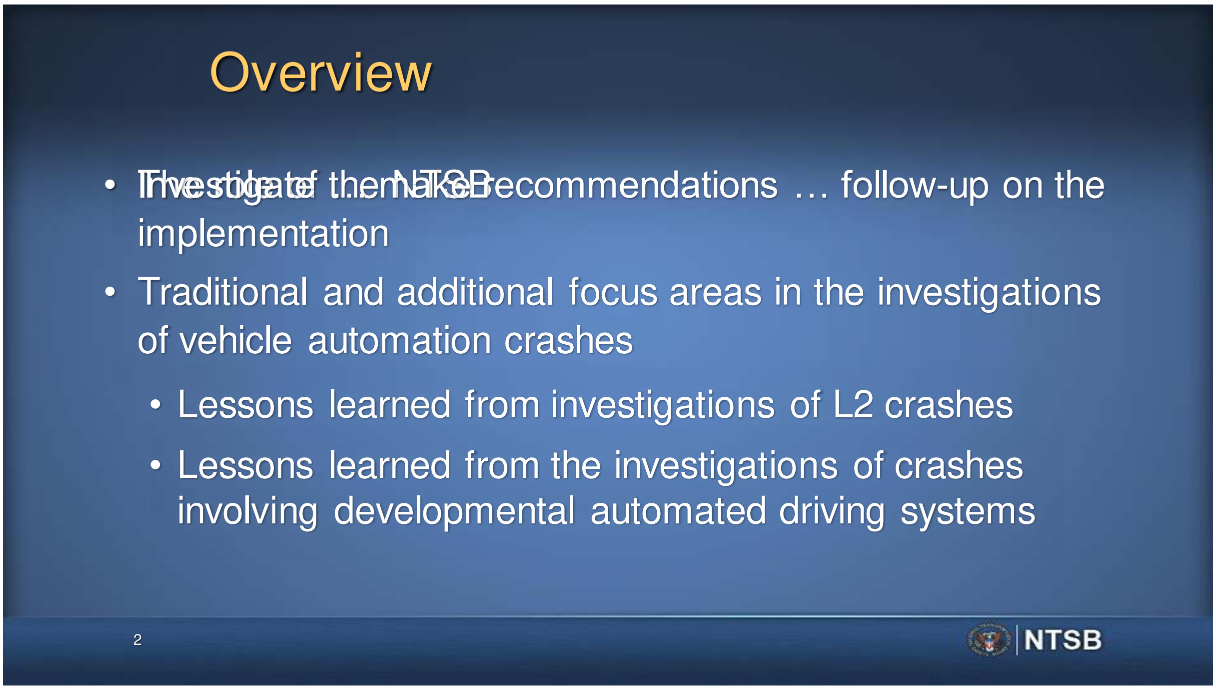

Recording devices to look for
ON VEHICLE*
*Some require continuous 12V source
Vehicle event recorder
Onboard video recorder
Motor controller memory,
EV Battery Battery Management System (BMS)
Anti-skid braking system memory (ABS)
Other . . .
OTHER
Cell phone – phone, data, GPS, camera
Roadway system - traffic video, timers, and other devices
Stores and other business security cameras
HowItBroke.com


Vehicle Data Recorders
Information Access Depends on Type of Investigation
Criminal – Government may not release ANY data
Safety – Government may release partial data, typically not video or audio
Civil – Typically requires court subpoena. May be denied.
Technical – May or may not get access
HowItBroke.com


Failure Logic Tree Exercise – FACTS LEAD TO ANALYSIS / SEQUENCE
Now we find contributing factors included BOTH the driver AND the car
Sequence found was:
Tired driver
Cruise system degraded
Pavement
Optical system
Cruise disconnected
Driver missed warning
Using cell phone
Driver did not brake
Car struck tree
CAR STRIKES TREE AT NIGHT
OR
OR
&
&
&
&
&
Driver using cell phone misses warning
MAN
Crusie control disengagement creates warning
No braking skid marks
Sleep deprived
Driving long distance
One headlight
Worn pavement
Optical sensor tolerances
Dark
5 Ms & E:
MAN
METHOD
MACHINE
MATERIAL
MEASUREMENT
ENVIRONMENT
HowItBroke.com


Numerous points create cybersecurity vulnerabilities
Attacks have taken place in aviation
Despite ISO 26262*, monitor for:
Intentional
Database corruption
Vehicle antenna inputs
Sensor entries
Software attacks
Unintentional
EMI/HIRF environment
Software conflicts
Sensor conflicts
SatCom, GPS, ACARS
IP Broadband
SatCom, ACARS
IP Broadband
Airplane
Software
Wifi
Hardware
Internet
ACARS
ACARS
Nav Charts
Navigation Data
Software
Software
Software
Hardware
MRO
Non-OEM
Airline, ATM
SW
HW - SW
OEM
Software
Hardware
Software
Hardware
SW
Firmware
Software
Manufacturers
HW - SW
Elec Parts
IC's
Suppliers
PC Cards
Adapted from Boeing Aviation Cybersecurity Diagram
Cybersecurity/hacking violations are a crime and require notification of law enforcement!
HowItBroke.com
* ISO 26262 - Road Vehicles Functional Safety Package

## Annex XV. Safe path to vehicle automation: Crash investigation perspective

## NTSB's Major Investigation

- Five disciplines
  - Highway design
  - Survival factors
  - Vehicle factors
  - Human performance
  - Operations (Motor Carrier factors)
- Reconstruction / Scanning
- Crashes involving vehicle automation
  - Dedicated reports related to automation



## Partial Automation Crashes

Williston, FL

Mountain View, CA

Culver City, CA

Delray Beach, FL

## Mountain View, CA – March 2018

- Tesla operated in L2:
  - Followed a lead vehicle
  - Moved into a gore area and no longer detected a LV
  - Accelerated prior to impact
- System did not detect attenuator
- Driver did not react
  - Inattentive due to phone use


5 NTSB

## Las Vegas, NV

- Navya autonomous shuttle
  - No traditional vehicle controls
  - Attendant on board
  - Low speed (~20 mph)
- ADS detected the truck
  - Decelerated the shuttle to a near stop
- 11 seconds later the truck backed into the stopped shuttle


6 NTSB

# Tempe, AZ – March 2018

- Uber ATG test vehicle
  - Modified 2017 Volvo XC90
    - Volvo CAS disabled
  - ATG developmental ADS
- Vehicle operator
- Completing a loop on N. Mill Ave in automated mode
- Nighttime with roadside lighting

7 NTSB

# Highway Design Issues

- Lane markings
  - Use of HD maps
- Work zones
  - Recognizing unexpected changes
- Roadway surface and hardware
  - Handling of damaged or differently positioned roadway hardware


8 NTSB

## Survival Issues

- Handling crashes of electric vehicles, including fires
  - Guidance for first responders
  - NTSB report on battery fires in electric vehicles
- Occupant safety
  - Seating positions and seat belt use
  - Extrication

9 NTSB

## Vehicle Issues

- Data
  - Reliance on the manufacturer for access and data interpretation
  - Lack of government recording requirements
- System versions
  - Changes in functionality (e.g., timing of alerts, detection of hazard types)
- Basic maintenance
  - System functionality; sensor calibration

10 NTSB

## Vehicle Issues: System Limitations

- Limitations of L2 and forward CAS
- Relevance of ODD
  - Domain is defined by the manufacturer
  - Adherence reliant on the driver
  - Rare implementations of system-based ODD
- Identifying errors in developmental ADS
  - Limitations of machine perception; developer-induced flaws

11 NTSB

## Human Issues: Role of a Human

- Human as an essential part of automation system
  - General problems of attention, fatigue…
- Automation complacency
  - Unintended inattention
  - Intentional misuse / distraction
- Monitoring of driver engagement
  - Steering wheel torque; camera
- Remote monitoring

12 NTSB

# Disengagement As a Factor

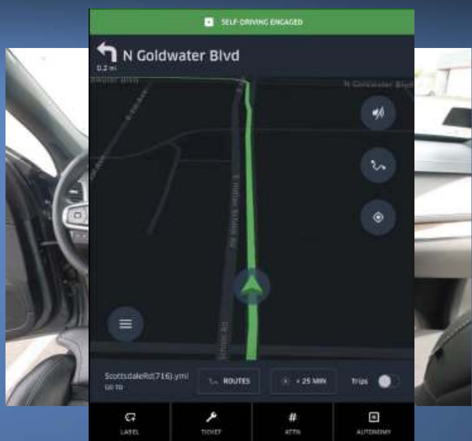





13 NTSB

# Human Issues: Mental Model

- Takeover from the system
  - Mental model of system's functionality
  - Trust in the system; expectation of system response
  - Operational procedures during ADS testing
- Task demands during ADS testing
  - Tempe operator's dual task




14 NTSB

## Operation Issues

- Examining company's safety culture
  - Organization and independence of safety departments
  - Technology company as a transportation company
  - Safety management system
- Examining federal and state requirements
- Voluntary standards and guidance

15 NTSB

## Recurrent Issues in L2 Crashes

- Considerable perceptual limitations
- Human drivers are poor monitors of automation
- Failure in partial automation + inattentive driver = crash
- Safety vs convenience
  - Does automating lane keeping improve safety?
- NTSB recommendations:
  - Improving monitoring of driver engagement
  - Limiting operational design domain

16 NTSB

## Issues in Developmental ADS Crashes

- Testing will contain errors and expose system's limitations
  - Machine perception; human attention
- Risk management in ADS development and operator oversight
  - Identify risks; implement safety redundancies
- Holistic view of risks and safety envelope
- NTSB does not instruct developers in building an AV
- Safety goal: How to mitigate the expected risk of testing on public roads

17 NTSB

## Safer Path Forward

- Tempe crash probable cause:

Deficiencies in risk mitigation were due to Uber ATG inadequate safety culture

- NTSB Recommendations:
  - Implementation of SMS
  - Federal and state oversight of developers' ADS testing process
- Industry sharing of lessons learned

18 NTSB

## The European Commission's science and knowledge service

Joint Research Centre

### JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

**EU Science Hub**
joint-research-centre.ec.europa.eu

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

EU Science

Publications Office
of the European Union