European Transport Research Review

REVIEW Open Access

Testing autonomous vehicles and AI: perspectives and challenges from cybersecurity, transparency, robustness and fairness

David Fernández Llorca[1,2*], Ronan Hamon[3], Henrik Junklewitz[3], Kathrin Grosse[4], Lars Kunze[5,6], Patrick Seiniger[7], Robert Swaim[8], Nick Reed[9], Alexandre Alahi[4], Emilia Gómez[1], Ignacio Sánchez[3] and Akos Kriston[3]

Abstract
This study aims to comprehensively explore the complexities of integrating Artificial Intelligence (AI) into Autonomous Vehicles (AVs), examining the challenges introduced by AI components and their impact on testing procedures. The research focuses on essential requirements for trustworthy AI, including cybersecurity, transparency, robustness, and fairness. We first analyse the role of AI at the most relevant operational layers of AVs, and discuss the implications of the EU's AI Act on AVs, highlighting the importance of the concept of a safety component. Using an expert opinion-based methodology, involving an interdisciplinary workshop with 21 academics and a subsequent in-depth analysis by a smaller group of experts, this study provides a state-of-the-art overview of the current landscape of vehicle regulation and standards, including ex-ante, post-hoc, and accident investigation processes, highlighting the need for new testing methodologies for both Advanced Driver Assistance Systems (ADAS) and Automated Driving Systems (ADS). The study also provides a detailed analysis of cybersecurity audits, explainability in AI decision-making processes and protocols for assessing the robustness and ethical behaviour of predictive systems in AVs. The analysis highlights significant challenges and suggests future directions for research and development of AI in AV technology, emphasising the need for multidisciplinary expertise. The study's conclusions have relevant implications for the development of trustworthy AI systems, vehicle regulations, and the safe deployment of AVs.

Keywords Autonomous vehicles, Trustworthy AI, Testing, Vehicle Regulations, Cybersecurity, Transparency, Explainability, Robustness, Fairness

*Correspondence:
David Fernández Llorca
David.FERNANDEZ-LLORCA@ec.europa.eu
[1]European Commission, Joint Research Centre, Seville, Spain
[2]Computer Engineering Department, University of Alcalá, Madrid, Spain
[3]European Commission, Joint Research Centre, Ispra, Italy
[4]EPFL - Swiss Federal Institute of Technology, Lausanne, Switzerland
[5]University of Oxford, Oxford, UK
[6]Bristol Robotics Laboratory, UWE Bristol, Bristol, England, UK
[7]BASt - German Federal Highway Research Institute, Bergisch Gladbach, Germany
[8]NTSB. HowItBroke.com, Washington DC, USA
[9]Reed Mobility, Wokingham, UK

1 Introduction

Artificial Intelligence (AI) plays a critical role in the advancement of autonomous driving. It is likely the main facilitator of high levels of automation, as there are certain technical issues that only seem to be resolvable through advanced AI systems, particularly those based on machine learning. However, the introduction of AI systems in the realm of driver assistance systems and automated driving systems creates new uncertainties due to specific characteristics of AI that make it a distinct technology from traditional systems developed in the field of motor vehicles. Some of these characteristics include unpredictability, opacity, self and continuous learning and lack of causality [1], among other horizontal features such as autonomy, complexity, overfitting and bias. As an example of the specificity that the introduction of AI systems in vehicles entails, the UNECE's Working Party on Automated/Autonomous and Connected Vehicles (GRVA) has been specifically discussing the impact of AI on vehicle regulations since 2020 [2].

In order to maximize the benefits of using AI and reduce potential negative impacts, several frameworks have been developed that establish principles and requirements for the development of trustworthy AI. For instance, the OECD AI value-based principles for responsible stewardship of trustworthy AI [3], or the EU's ethical guidelines for trustworthy AI [4–6] which encompass multiple ethical principles, requirements, and criteria to ensure that AI systems are designed following a human-centred approach.

These frameworks comprise multiple interrelated components that extend beyond safety concerns, such as cybersecurity, robustness, fairness, transparency, privacy, accountability, societal and environmental well-being, among others. Their general applicability to the autonomous driving domain has been previously analysed in [7, 8], demonstrating that these are complex frameworks which require addressing multiple problems of varying nature. Some of these problems are still at an early stage of scientific and technological maturity, presenting new research and development challenges in various areas. Progress in this context requires multidisciplinary expertise and tailored analyses.

In this work, based on the methodology described in Sect. 1.2, we present a detailed study of the state of the art in relation to some of the most prominent elements of trustworthy AI frameworks, and how they affect the current and future landscape of testing procedures for AVs. Specifically, we focus on cybersecurity, transparency, robustness, and fairness. The analysis is conducted by a multidisciplinary group of experts and includes the identification of the most pressing future challenges.

After describing the related work, the methodology, and presenting the terminology used, Sect. 2, is dedicated to presenting the various operational layers involved in vehicle automation, as well as the impact of AI on each of them. It also presents the regulatory context of AI and its sectoral application to the field of AVs. Section 3 presents the current landscape of vehicle regulation and standards, including ex-ante, post-hoc, and accident investigation processes. Sections 4, 5, and 6 detail the requirements for cybersecurity, transparency through explainability, robustness of prediction systems, and fairness. Finally, Sect. 8 presents the general conclusions and future work.

1.1 Related work

In this section, we briefly describe previous works focused on analysing trustworthy AI requirements in the context of Autonomous Driving. This analysis is complemented in subsequent sections of the paper, where a detailed state-of-the-art analysis is presented for each specific requirement addressed in this work.

The discussion on the general impact and technical implications of applying trustworthy AI system requirements in the context of AVs was initially explored in [7] and [8]. A comprehensive analysis was conducted, and the scientific and technical maturity of the multiple dimensions included in the requirements was clearly identified. However, due to the broad scope of the analysis, the level of detail provided for some of the requirements was insufficient.

Beyond these works, we find others that, although referencing the context of trustworthy AI as a general goal, focus exclusively on a specific requirement and operational layer. One of the most frequently addressed requirements is that of transparency and explainability, in relation to, for example, documentation [9], scene understanding [10], anomaly detection [11], motion prediction [12–14], motion planning [15], or decision-making [16].

We can also find previous works analysing the state of play of current regulatory developments in the context of autonomous driving, either focusing on a specific region (e.g., the US in [17] and [18]) or from a more global perspective [19, 20]. However, these studies do not elaborate on the impact of AI or the context of the requirements for trustworthy AI.

The general impact of AI on AVs has also been covered in other studies [21, 22]. The interplay between trust as a human factor, and the more technical concept of trustworthiness was studied in [23]. More recently, in [24], and also based on a focus group methodology including policymakers, technology developers, and media organizations, the authors identify a set of dimensions of trustworthiness without delving too deeply into either the scientific maturity of each dimension or their possible gradation.

To the best of our knowledge, apart from the works previously mentioned in [7] and [8], we did not find any work in the literature that addresses key elements of the trustworthy AI framework with the depth and breadth we propose in this work, including highly relevant requirements such as cybersecurity, transparency, and fairness.

1.2 Methodology

To tackle the main focus areas of our analysis (cybersecurity, transparency, robustness, and fairness), we employed an expert opinion-based methodology at two levels. First, we conducted an interdisciplinary workshop involving 21 expert academics from various disciplines relevant to trustworthy AI and AVs. This interactive, collaborative online workshop allowed each expert to introduce a topic related to one of the focus areas, address all related questions, and provide additional resources as needed. The workshop's outcome was encapsulated in a technical report [25]. From this, a smaller group of experts was selected to go deeper into each focus area. This group aimed to provide a state-of-the-art overview for each area within the context of autonomous driving, specifically focusing on testing methodologies, and subsequently identify the most significant challenges in each area.

The two-stage approach was designed to achieve a thorough understanding of the problems addressed. The broader workshop allowed for initial knowledge mapping, diverse perspectives, and the identification of key areas of focus, while the smaller focus group enabled in-depth analysis, specialised knowledge, and iterative refinement. By combining both stages, the research achieved a comprehensive and nuanced understanding of the topic, covering both the breadth and depth of trustworthy AI for AVs, and made efficient use of expertise. This approach was necessary due to the complexity and novelty of the research area, and the need for heterogeneous and multidisciplinary expertise.

Due to the multidisciplinary nature of trustworthy AI for AVs and the diverse group of experts, it was initially agreed to establish a common terminology, identify the main operational layers of AVs, and analyse the impact of AI on each of them. The regulatory context of AI and vehicles was also collectively examined.

1.3 Terminology

With respect to the terminology, in this work, we follow the proposal presented in [8]. We use *advanced driver assistance systems* (ADAS) or *assisted vehicle/driving* for SAE levels 1 and 2 (driver), *automated vehicle/driving* for SAE level 3 (a backup driver/user is in charge), and *autonomous vehicle/driving* for SAE levels 4 and 5 (passenger/unoccupied). In some cases, we also use *automated driving systems* (ADS) to generically refer to SAE levels 3 to 5 (automated and autonomous driving). Finally, when we use the acronym AV, we refer to automated and/or autonomous vehicles indistinctly.

2 AI for AVs

2.1 Operational layers of AVs and the role of AI

Autonomous vehicles (AVs) are complex systems that interact with highly complex environments with an almost infinite variety of possibilities. High levels of automation are facilitated by at least seven operational layers (see Fig. 1):

- Localisation: determines vehicle pose (position and orientation) relative to a digital map. This includes

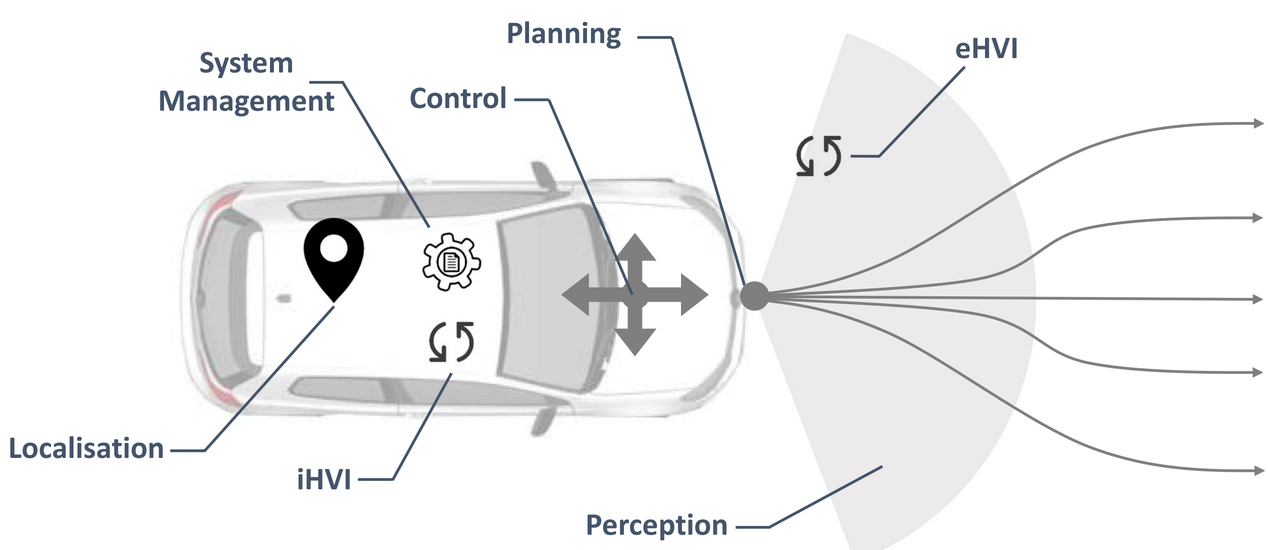


Fig. 1 Key operational layers: *localisation*, *perception*, *planning*, *control*, *human-vehicle interaction* and *system management*. Adapted and expanded from [26]

enhanced digital or high-definition maps that contain static elements (e.g., traffic lights, traffic signs, pedestrian crossings, lanes, street layout, intersections, roundabouts, parking areas, green areas, etc.) and dynamic real-time information (e.g., weather and traffic conditions). Different map-matching techniques are used to accurately establish the vehicle's position and heading within the map. While traditionally localisation was not based on learning, data-driven and machine learning approaches are becoming prevalent [27].

- Perception (and prediction): a.k.a. scene understanding, this process detects and interprets the dynamic content of the scene, including vehicles and Vulnerable Road Users (VRUs) [7]. It also includes segmenting the static elements of the scene from the perspective of the vehicle. Perception also includes *prediction* of future behaviours and trajectories of other agents, such as vehicles and VRUs, which are crucial for safe and comfortable planning. *Prediction*, also referred to as *social forecasting*, extends beyond simple predictions to forecast long-term dependencies and complex collective interactions. It models social etiquette, the unwritten rules of social interactions, predicting social agents' behaviours based on social models and including awareness of agents' own actions on others' movements. Social agent interaction requires joint reasoning, and the most promising approaches are based on complex machine learning methods (e.g., socially-aware AI for AVs, as depicted in Fig. 2).
- Planning: a.k.a. (local) motion or trajectory planning or decision making. Based on vehicle localization, knowledge of traffic rules, state of signals, and predicted trajectories of dynamic agents, local planning is performed, providing a feasible and smooth trajectory and speed references for the control system. We can distinguish between different methods that have traditionally been applied to deal with motion planning, including graph-search, variational or optimization-based, incremental or sample-based, and interpolation-based [28]. End-to-end deep learning approaches are becoming predominant [29].
- Control: it involves both longitudinal (acceleration and breaking) and lateral control (steering) of an AV [30]. Feedback controllers are used to select the most appropriate actuator inputs to perform the planned local trajectory. Many different types of closed loop controllers have been proposed for executing the reference motions provided by the path planning system [28], including path stabilization, trajectory tracking, and more recently, predictive control approaches.
- Human-Vehicle Interaction (HVI): involves the design of human-vehicle interfaces for effective interaction and communication with in-vehicle users and external road users [7]. Potential modalities

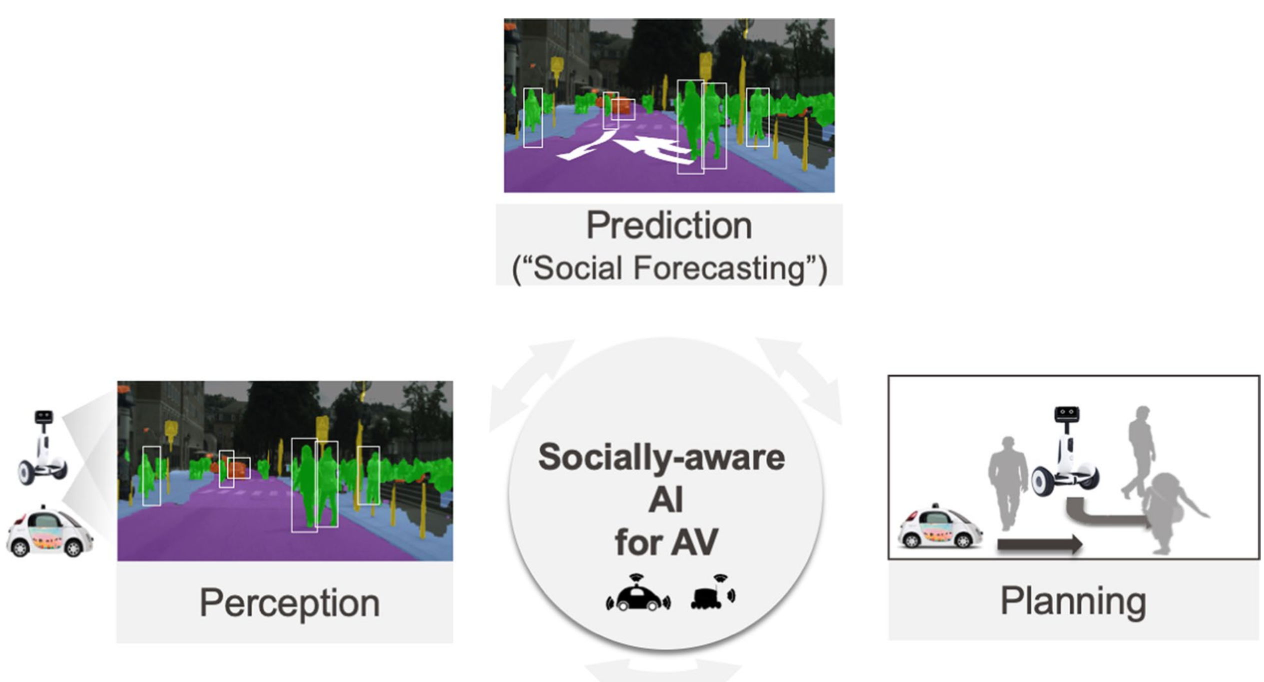


Fig. 2 Socially-aware AI for AVs. *Perception, Prediction*, and *Planning*, also referred to as the 3 Ps. *Perception* and *planning* augmented with *prediction* to safely yet effectively deploy AV that will interact with other *social agents*

to communicate intention of the AV to road users include explicit, such as audio and video signals, and implicit forms, such as vehicle's motion pattern (speed, distance and time gap) [31]. Regarding driver and passengers, common interfaces are audio, tactile, visual, vibro-tactile, and more recently, natural language processing [32]. In addition, this layer also includes in-vehicle perception systems to detect the users' status.

- System Management: it supervises the overall ADS, including functions such as fault management, data management and logging, event data recording, etc.

Although with varying levels of intensity, we can state that AI plays a predominant role in the operational layers of localisation, perception, planning, and human-vehicle interaction, somewhat less in the control layer, and circumstantially in the system management layer.

2.2 The AI Act and its sectoral application

In August 2024, the European AI Act [33] entered into force. Its primary goal is to establish trustworthy AI practices within the European Union. The AI Act adopts a risk-based approach that mandates specific obligations for providers of AI systems depending on their risk level. High-risk AI system providers (including those with potential risks to fundamental rights, health, and safety of humans) are obligated to meet defined requirements in Chapter III, Section 2 of the legal text (Articles 8 to 15). These requirements refer to risk management system, data and data governance, technical documentation, record-keeping, transparency and provision of information to users, human oversight, and accuracy, robustness, and cybersecurity.

Article 6 establishes the rules for the classification of high-risk AI systems. The proposal defines a specific set of AI systems that shall be considered high-risk (Annex III). This can be considered the core of the regulation. However, the AI Act also affects sector specific regulations. More specifically, AI systems intended to be used as a *safety component* of a product, or as a product itself, covered by Union harmonisation legislation (listed in Annex I, including machinery, toys, medical devices, motor vehicles regulations, etc.), as well as those products requiring a third-party conformity assessment before being put on the market or putting into service. The Annex I includes Regulation (EU) 2018/858 [34] on the approval and market surveillance of motor vehicles, and amending Regulation (EU) 2019/2144 [35] on type-approval requirement for motor vehicles as regards their general safety and the protection of vehicle occupants and vulnerable road users. It is important to note that specific requirements for ADAS and ADS are further developed through delegated and implementing acts. Those acts must be amended to ensure that they incorporate the mandatory requirements for high-risk AI systems, when adopting any relevant delegated or implementing act on the basis of those acts.

2.3 Safety components with AI in AVs

In order to be subject to possible future requirements under a Commission delegated act, the AI system supplied or integrated in the AVs should be considered a safety component within the meaning of the AI Act.

The concept of "safety component" is defined in Article 3 (14) of the AI Act [33] as "*a component of a product or of an AI system which fulfils a safety function for that product or AI system, or the failure or malfunctioning of which endangers the health and safety of persons or property*". This general definition does not necessarily overlaps with sector-specific definitions of safety components of products covered by Union harmonisation legislation. Particularly relevant is the new Machinery Regulation (EU) 2023/1230 [36], in which the concept of safety component explicitly specifies that the component "*is not necessary in order for that product to function or for which normal components may be substituted in order for that product to function*". Then, an indicative list of safety components is provided in Annex II, including presence detectors of persons, monitoring devices, emergency stop devices, two hand control devices, among others. Another example is Directive 2014/33/EU [37] relating to lifts and safety components for lifts, which also includes a list of safety components well aligned with the idea of "*not necessary for the product to function*", such as locking landing doors devices, fall prevention devices, overspeed limitation devices, etc.

As stated in [7, 38], given the inherently high-risk nature of AVs (capable of causing serious physical harm or death, as well as property damage), it is arguably quite intuitive to qualify AI systems of all the operational layers of AVs as safety components. All functions addressed by the different operational layers fulfil a safety function that may endanger the health or safety of persons or property in case of malfunction. The point is that, at the same time, these functions are essential for "the product to function". We can identify some components that are not necessary for the overall functioning of the system, such as comfort functions, infotainment applications, or vehicle warnings. However, these functions are not safety relevant either. In other words, virtually all components of AVs that are safety relevant are also relevant to the overall functioning of the system. Therefore, the conventional definition of safety component given in contexts such as machinery or lifts may not be applicable to AVs. In any case, considering that an average passenger car contains more than 30,000 parts, among which there will be a wide variety of AI systems for different functions (with a tendency to increase), at some point it should be clearly defined which AI systems used in which specific context can be considered as safety components within the meaning of the AI Act.

3 Vehicle regulations and standards

3.1 Ex-ante requirements and testing

The product development process of any given system, including ADAS and ADS, is typically based on the V-model of product development (e.g., ISO 26262). Over the course of the development of a system, the degree of abstraction of the system properties first decreases (in the left branch for development) and then increases again (in the right branch for verification). At the same time, the objective shifts from questions of system design (left branch) to validation of the intended system characteristics (right branch). Requirements from vehicle technical regulations or standards and guidelines can also be placed in this scheme, giving vehicle safety bodies a chance to directly influence the development process [39].

External requirements can come from various sources, most prominently the type-approval requirements, and independent car assessment programs such as the European New Car Assessment Program (Euro NCAP) [40]. Euro NCAP procures and tests market-available vehicles (compliant with type-approval requirements), comparing performance via standardized procedures. As these procedures are communicated in advance, manufacturers can integrate anticipated test results into their product development cycles.

All vehicles registered in Europe adhere to the EU Type Approval Directive 2007/46/EC [41] or the EU Type Approval Regulation (EU) 2018/858 [34], including amendments like the General Safety Regulations (GSR) Regulation (EC) No. 2019/2144 [35]. However, these revised directives and regulations do not specify technical requirements but refer to delegated or implementing acts of the European Commission (e.g., Implementing Regulation (EU) 2022/1426 regarding ADS or fully automated vehicles [42]) or designated UN regulations (e.g., UN Regulation No 157 on Automated Lane Keeping Systems [43]). These are preferred as they allow global harmonization of the requirements. The amended documents list mandatory or optional requirements for vehicle type approval. Thus, UN regulations increasingly impact ADAS and ADS. In fact, ADS can serve as an example of optional regulations (so-called "if fitted" regulations). The framework regulation also allows exemptions for new technologies or concepts (see Art. 39 in [34]), enabling deviation from defined requirements to foster innovation, given that road safety and environmental protection are not compromised.

It is important to highlight the difference between conventional and new regulatory approaches. Conventional regulations typically set implicit requirements, encompassing a test procedure with few conditions and defining pass criteria solely for those conditions. Conversely, newer regulations, particularly for ADAS and ADS (see [42]), define requirements across a range of operating conditions without limiting potential test cases. They also involve "market surveillance" processes that allow regulators, namely the European Commission and Member States, to retest production vehicles against these broad requirements. Therefore, vehicle manufacturers are required to design robust systems capable of functioning under a variety of circumstances.

Conventional testing for ADAS and ADS employs various methods contingent on the test objective, such as the system functionality to be evaluated and the development status. The methods typically define the general procedure, tools (e.g., artificial target objects, measurement and control devices, suitable environment), procedures for deriving characteristic values from collected data (e.g., calculation of speed reduction in an emergency braking system), and often include an evaluation standard for deriving a test result from the characteristic values (e.g., weighting table depending on test speed in Euro NCAP emergency braking assistance tests) or reference to such (e.g., referencing requirements of UN Regulation no. 152 [44]). If assessment standards are specified, they can be interpreted as implicit requirements for the corresponding system.

Advance testing procedures, e.g., for evaluating ADAS with environment perception or ADS, require generating relevant application situation for each system. As public road testing is often unfeasible, equivalent situations must be created in a test environment. In the case of tests for emergency reactions of ADS, for example, this requires objects that represent stationary or moving vehicles. Distinguishing features of test procedures may include the type of these target objects, their motion systems, or the degree of virtualization.

Test procedures for ADAS and ADS primarily involve position measurement, position control with driving robots (e.g., automated vehicles control their position by themselves) and movable targets. The state-of-the art for position measurement are inertial systems with stabilization by differential GNSS (e.g., GPS, GLONASS), capable of measuring vehicle position with 1 cm accuracy. Derivative and angular accuracy are also reasonably accurate.

For ADAS or ADS that require driver input, the "driving robot systems" can manipulate the controls. These systems can perform open-loop input into the controls and control vehicle movement to follow pre-programmed trajectories. Regarding target systems, it is common to use dummies emulating cars, bicycles, motorcycles or pedestrians for ADAS or ADS testing. These targets, standardized per ISO 19206, are typically propelled by self-driving, overrunable robot platforms.

However, there are significant limitations in current testing methodologies for ex-ante requirements validation. Test procedures must sufficiently represent or simulate all necessary information for each situation, increasing the effort needed to develop and represent realistic and relevant test situations for ADAS and ADS. It is also necessary to prove that these situations accurately represent reality for a system response. For instance, used objects must not only be sensor-detectable but also representative of vehicles, pedestrians, or other real traffic road users. Without this proof, despite excellent functional and safety evaluations under artificial conditions, little safety progress is achieved in real road traffic.

The same issue applies to the selection of test scenarios [45]. As the number of sensors and parameters involved in the operation of ADAS and ADS increases, the effort to define scenarios, execute tests, and prove field relevance escalates, especially for predictive systems based on agent behavior modeling [46]. On the other hand, it would still be possible to fine-tune the systems to react to standardized targets, but not to, for example, pedestrians with different clothing and appearance. The primary challenge is verifying that ADAS and ADS function only on test tracks under optimal conditions, but also in realistic environments and with parameters outside the test cycles.

3.2 Post-market monitoring

Market surveillance mechanisms, such as those defined by EU Regulation 2018/858 [34], ensure that motor vehicles and components on the market comply with laws, regulations, and safety and health requirements. The European Commission participates in these activities, verifying vehicles' emissions and safety compliance, and collaborating with national authorities for information exchange and peer review of technical actors [47]. While the requirements for market surveillance match those of ex-ante type-approval procedures, the testing methodologies may differ considerably.

Another interesting tool recently integrated in the Implementing Regulation (EU) 2022/1426 [42] is the "In-Service Monitoring and Reporting" (ISMR), which addresses in-service safety of ADS post-market (i.e., operational experience feedback loop). It is primarily based on in-service data collection to assess whether the ADS continues to be safe when operated on the road and identify safety risks [48].

The AI Act [49] also outlines several post-market actions, including market surveillance, human oversight and monitoring (as a requirement), and post-market monitoring. Post-market monitoring involves all activities carried out by providers of AI systems to collect and review experience gained from the use of AI systems to identify any need to immediately apply corrective or preventive actions. This idea is well aligned with the ISMR concept.

3.3 Crash investigation

Despite diligent design, planning, and safety regulations, accidents occur. Investigations into root causes identify and address contributing factors to prevent recurrence and/or mitigate negative outcomes. The investigation depth varies, ranging from counting crashes in a timeframe to detailed examinations of involved parties and vehicle operation assumptions. Collisions typically involve a mix of human, machine or system, and environmental factors. When AVs are involved in an accident, investigators must establish the required investigation depth. The need of police officers often end once they file a report recording scene, damage, interviews, and traffic law violations. These facts might be *contributing factors* but at this point, they represent basic background information for examining how an automated system may have failed. Collecting facts and assembling a chronology of events leads to a causal chain establishment. However, "probable cause" determination for crash investigations is not the same as for criminal prosecutions, which require "facts beyond a reasonable doubt".

The investigation depth is a balance between available resources and required information. All investigations involve fact collection, analysis of how facts fit together, and proceeding forward with new understanding. The causal chain typically involves a mix of *man and machine*. Various methods convey information and establish crash causes, including the "Five M's and E" [50], the "Swiss Cheese Model" citeReason2006, and the "Fault Tree Analysis" method [51]. Recorded information and facts are often embedded in multiple vehicle devices and other sources. The recorded data may contain a variety of parameters, requiring manufacturer assistance to obtain and specialized software to interpret. From the investigation and fact collection, a chronology of the accident can be established to identify contributing factors and probable cause.

With increasing numbers of vehicles with ADAS and ADS, investigators may question how much these automated systems contributed to collisions. Factors complicating this include drivers' lack of understanding of how ADAS/ADS work, blame shifting to the vehicle, proprietary nature of ADAS/ADS designs, increased time required for ADAS/ADS investigations and potential investigator shortage, and lack of ADAS/ADS training for collision investigators.

While ADAS or ADS in ground vehicles is relatively new, aviation automation has existed for 110 years, maturing into digital system forensics. Despite design and standard differences, the basic elements involved in ground vehicles and aviation remain a continuous loop of requirements, outputs, actuators, and feedback. Human involvement extends to each of these, necessitating context investigation about man and machine involvement in each aspect, as well as potential environmental contributions. Aviation has a long history of autopilot use with occasional involvement in accidents. The two main autopilot-related accident causes have been the human interface and the lack of human supervision or misuse. Two examples are the accidents of the Air France A330 flying from Rio de Janeiro to Paris in 2009 [52] (human interface issue), and the American Airlines Flight 965 flying from Miami to Cali in 1995 [53] (human supervision issue). Similarly, numerous accidents involving ADAS and ADS in road vehicles have occurred when drivers have fallen asleep or intentionally tried to let the vehicle do more than the automation level was intended [54, 55].

Determining proportionality in an crash investigation has no singular easy method. The process continues to be data collection, beginning at the scene and not starting analysis until factual elements are fully collected. The challenge for ADAS and ADS investigation is that there may be substantially more data to pursue and more unforeseen aspects to learn about, some of which may be proprietary and difficult to obtain.

4 Cybersecurity audits for AI components

Cybersecurity standards have been established for motor vehicles [56], with ADS introducing an additional layer of complexity due to AI components. Existing standards may not necessarily apply to AI [57], which has been shown to introduce security vulnerabilities, impacting any system relying on or incorporating AI components. We consider a security vulnerability as the potential for malicious activity that alters the behaviour [58–61] or leaks data [62–65] from an AI component, designed for a specific task. To prevent attacks on critical infrastructure, like AVs, we propose an AI security audit structure for standardized assessment of system security against specific threats or resistance to information leakage.

Before discussing an audit structure, we need a risk assessment of likely threat occurrence. We propose four levels of AI security research concreteness. The first level describes theoretical research independent of the application or ADAS/ADS, and the fourth level describes a vulnerability exploited in practice. From this perspective, we review three relevant security threats associated with AI components in ADS: *evasion*, *poisoning*, *privacy* and *IP*. Other cybersecurity risks independent of the AI components will not be discussed here. For an overview, we refer to [66].

4.1 Security threats

This section primarily focuses on classification, vision and LiDAR tasks, due to readily available works in the area and the direct exposure of perception modules to the outside world, potentially increasing vulnerability. In discussing practicality of studying threats, we propose four levels of concreteness related to ADAS and ADS threats (see Fig. 3):

- I A vulnerability demonstrated on an AI algorithm used in an AV on a non-vehicle related task (e.g. face detection).
- II A vulnerability demonstrated on an AI algorithm used in an AV on a relevant task (e.g. image segmentation or traffic sign classification).
- III A vulnerability demonstrated on an AI algorithm used in an AV on a relevant task, with experiments showing that the vulnerability is presence in a vehicle or similar testbed.
- IV A reported vulnerability affecting an on-road AV.

While concreteness increases towards real-world threats, a vulnerability can directly occur at level III or IV without being observed at lower levels first. Studying a

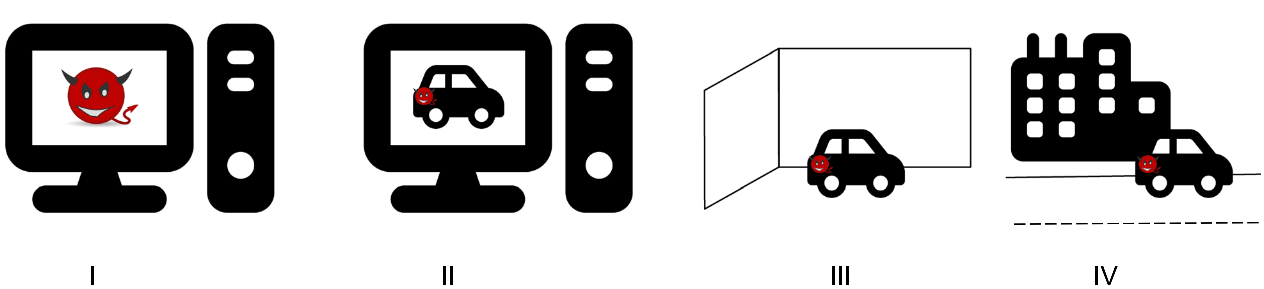


Fig. 3 Four levels of concreteness in AI component AV security. I, a vulnerability is known to exist for an AI. II, a vulnerability is shown for an AI that is potentially used in cars. III, the vulnerability has been shown on a testbed, or on a car in a safe environment (e.g., driver is aware of the test). IV, the exploit has been used on a car on the road

vulnerability at level I is not a prerequisite for its occurrence at concreteness II, III, or IV.

It is also relevant to distinguish between *adversarial* and *benign accuracy* when considering AI component vulnerability in AVs. Both are based on accuracy, measured on a specific dataset called the test set, while the AI is optimized on a separate dataset, the training data. Adversarial accuracy is the accuracy when an attacker targets the system to decrease performance using a vulnerability by altering training or test data, or both. Benign accuracy is the accuracy when no attacker is present, though performance decreases may still occur, for example, when test and training data differ. This may happen when a car is released for production and operated in real-world traffic as opposed to recorded scenes, where unknown weather conditions may occur, or unanticipated car behaviour may be observed.

4.1.1 Evading AVs

In evasion attacks, an attacker alters inputs at test time. These modified inputs may appear as one class to a human observer but get misclassified by the model despite otherwise good performance. For example, a traffic sign recognition algorithm might misclassify a manipulated "stop sign" image as a "yield sign" assignment [67]. Such manipulations can be achieved through classifier-specific graffiti [67] or subtle changes to the image [68] or sign shape [67, 69]. Evasion is a vulnerability that can occur at deployment or during testing stages and exists across all types of classifiers [70] and input data or tasks [67, 69, 71], often referred to as adversarial examples.

Table 1 Selection of evasion attacks on AVs with different levels of concreteness as depicted in Fig. 3

Ref.	Target component	Target sensor	Concreteness	Description
[67]	Downstream	Camera	II	Grafiti leading to wrong classification
[69]	Downstream	LiDAR	II	Adds objection to hide object
[72]	End-to-end	Camera	II	Adding lines on street lead to wrong steering output
[71]	Planner	Motion sensors	II	Alter movement feedback causes wrong trajectory
[61]	Downstream	Camera	III	Perturbing traffic signs
[74]	Unknown	Camera	III-IV	2 reported benign incidents
[73]	Unknown	Unknown	IV	Tesla crashes with truck, cause probably benign

AV-specific evasion attacks can target visual sensors [67] or other sensors like LiDAR [69]. They can also directly target the vehicle's planner or affect end-to-end driving by adding shadows or lines on the road [72]. Most attacks proposed on AVs focus on one component and do not trigger complex behaviours. We attempt to give a concise overview of suggested attacks in Table 1. Real-world incidents with vision systems [73, 74] are generally assumed to be benign incidents, not attacks by a malicious entity.

Many AI engineers and practitioners consider evasion or adversarial examples more of a performance issue than a security problem [74]. Traditional security vulnerabilities (i.e., non-AI related) seem to be more prevalent in practice [74]. However, defence and forensics against evasion attacks remain challenging, making it essential to avoid situations where such attacks do not lead to penalties.

Defending against evasion is complex as adversarial examples are inherent to the AI functionality [75], not a bug. While numerous defences have been introduced, many provide a false sense of security [76] as they prevent specific attacks rather than general adversarial examples. Current standards include adversarial training and formal verification [77] for small critical applications to obtain a safe network. Both approaches must be carried out during development or before deployment by the model provider. However, integrating vulnerability fixes upfront is always preferred over ad-hoc solutions post-design.

4.1.2 Poisoning of AVs

In contrast to the latter backdoor attacks, there are few poisoning attacks on systems used in AVs. The reason for this inherent complexity of poisoning attacks is that, in contrast to for example evasion attacks (that derive a change against a static, deployed AI), poisoning attacks aim to make the AI deviate in a malicious way during training, where the AI is still dynamically adapting to the data [78]. Some works exist on deep learning [78, 79], where [79] tackle traffic sign recognition as used in AVs. However, also the path finding of the vehicle can be poisoned to only find a sub-optimal route [80].

In poisoning attacks, an attacker alters training data, which affects test time performance. They can degrade data quality [79] or install a backdoor, causing the model to misclassify inputs when a specific pattern or trigger is present [58]. These attacks are particularly severe as affected models perform well as long as the trigger is not present [58, 81].

Despite the complexity when compared with evasion, there are several examples. Some works exist on perception [78, 79] and planning [80] operational layers. However, backdoor attacks are easier to implement by controlling part of the training data [78]. They have been tested on vision models [58], image segmentation [82], LiDAR sensors [59, 60], and even on traffic congestion control systems [81]. Most of these attacks focus on ADS but do not target real cars, as we show in Table 2. There are no known incidents of poisoning attacks from media or prior research.

When discussing real-world attack impacts, we must differentiate between poisoning and backdoor attacks [74]. Poisoning attacks significantly affect model performance, making it unlikely that a poisoned model ends up in production. However, backdoors represent a more severe vulnerability, as the high benign accuracy and unknown triggers make the attack easy to overlook.

Table 2 Selection of poisoning/backdooring attacks on AI components of AVs with their concreteness

Ref.	Target component	Target sensor	Backdoor?	Concreteness	Description
[79]	Downstream	Camera		II	Reduce accuracy of traffic sign recognition
[58]	Downstream	Camera	✓	II	Misclassify traffic signs with trigger
[59, 60]	Downstream	LiDAR	✓	II	Add trigger to missclassify object
[80]	Planner	Position		II	Poisoning for suboptimal route choice
[81]	Traffic control	Position	✓	II	Add trigger to cause traffic jam
[82]*	Downstream	Camera	✓	III	Leads to wrong lane detection output

* tested on a small testbed vehicle

Defending against poisoning is somewhat understood, with known trade-offs linking attack strength with detectability by the defender [83]. However, defending against backdoors remains an open problem for different modalities [84, 85]. In relation to audits, two different approaches exist: preventing a backdoor from being embedded in the network during training [?], and identifying whether a network is backdoored post-training [86]. Both attacks stem from the data collection stage, allowing for countermeasures like data cleaning before training or AI protection during training [86]. Fixing vulnerabilities should be part of a system design approach taken upfront, not done in an ad-hoc manner.

4.1.3 Attacks on privacy and IP in AVs

Privacy concerns in AVs generally arise in the context of data leakage from an AI model [63]. Depending on the data used for training (e.g., real world crashes investigations), data leakage might have privacy implications. It is currently unclear whether video sequences can be retrieved from a model in sufficient detail. A more significant concern is when data from different users is combined to train a new AI, e.g, to improve AV usability based on individual cars. This has been shown to be vulnerable to attribute inference [87] or to the complete training inputs [88]. Independent of the data used, the model can become a target of an attack where it is copied without the owner's consent by observing input and outputs [65] or other data flow within hardware components [64]. Privacy and IP threats almost uniquely occur during deployment but are caused by design choices early in the development of the AV and its components. While there is a large body of work about privacy in AVs in general [89], few works investigate either privacy attacks or data protection, or IP attacks/protection related to AIs deployed within cars. An exception is a study that shows that some AI components can be stolen when observing data flow in the Peripheral Component Interconnect (PCI) bus present in AVs [64].

Given the surge in privacy-related legislation, privacy attacks are a threat. Although there are reports about privacy being leaked by companies[1] [2] or AI companies,[3] there are no reports yet about data protection failures or private data directly leaked from AVs. Analogously, there are no reports about stolen models from AVs.

[1] https://www.smh.com.au/property/news/wake-up-call-for-real-estate.

[2] https://www.bleepingcomputer.com/news/security/vodafone-italy-discloses-data-breach.

[3] https://www.pcmag.com/news/report-ai-company-leaks-over-25m-medical-records.

In the case of these attacks, we need to differentiate between the loss of IP and insensitive training data, and the car user's privacy. One is considered the company's responsibility and need not be audited in a security audit by a third party unless desired by the audited party. Whenever the users' privacy might be at stake, the model needs to be audited thoroughly.

Neither privacy nor IP protection is well understood or investigated [62, 88] for AI modules in AVs. Techniques like differential privacy or data noising can alleviate some threats [62]. A recent line of work finds that Bayesian methods provide differential privacy by default [90, 91], and should be used in sensitive settings. In general, the loss of sensitive information can be restricted by providing less information to the attacker, for example, by restricting the number of queries that can be made per time unit or the kind of information released (e.g., less numerical precision).

4.2 Testing and auditing

Given the diverse vulnerabilities discussed, the need and utility for audits vary. In the case of evasion, an audit is useful as it evaluates the model's benign accuracy. Adding these tests enables easy later extension against evasion attacks, and auditing benign accuracy indirectly tests for poisoning. Any production model should be audited against existing backdoors. As for privacy, the AV must be audited to prevent leakage of potential users' sensitive information. Lastly, audits related to IP leakage can be made optional, depending on the model owner's wishes, and need not be conducted by a third party.

We suggest high-level audit criteria for all vulnerabilities, which include:

- Validating the design of the model.
- Testing the robustness of the model within the system.
- Testing the soundness of potential defence measures within the system.

The first point refers to the correct usage of data, which steps (or defences) are implemented and why, and which methods are chosen. This is particularly relevant concerning privacy or when methods with formal guarantees are relied on. The audit then needs to assess if the formal criteria match the application and make sense within the larger system under consideration. The second and third points are empirical and assess the model's performance via tests. The third point is added for completeness in cases like evasion and might become relevant later if more attacks occur in practice. For other cases, like backdoors, it is already relevant. In the case of privacy, it might also be necessary to empirically validate a defence that does not provide enough formal guarantees.

4.2.1 Prerequisite for a cybersecurity audit

In a nutshell, a cybersecurity audit should assess the model's robustness using an appropriate attack (evasion) or assessment (backdoors, privacy). This raises the question of good attacks and assessments, which generally differ in their threat model and complexity, encompassing access to the model, expected changes by the attacker (norms), and the complexity of the evaluation.

- *Model access*: whether the attack uses insider knowledge about this model (white-box) or not (black-box). White-box attacks are considered a worst-case scenario and are stronger than black-box attacks. However, they can also give a false sense of security when not properly tailored for the model. In AVs, we can likely assume that the AI components will be proprietary and thus black-box for an attacker (unless there is an IP breach). For audits, white-box access must be given to study a worst-case attacker, which requires the audit to be performed by a trustworthy, independent third party that is provided confidential access to the model.
- *Norms*: how an attacker may alter a sample reaching the AI component. To measure the change added to a sample, distance or similarity measures are often used. The most basic functions are called L_p norms, where p denotes the type of alteration captured by this specific norm. For instance, the L_0-norm results in sparse, graffiti like alterations, whereas the L_∞ -norm creates minor changes in an image. While much of the research focuses on these norms for vision, other norms that mimic human perception [92] can also be utilized. For LiDAR sensors, different norms that focus on adding a consistent object are required [69]. The norm needs to be chosen consistently with the task and robustness requirements. However, there is no scientific consensus on a norm-based threat model for autonomous cars that would help determine which norms or similarities can be used for an evaluation.
- *Audit complexity*: the average expected run-time of the evaluation. Complete search of vulnerability is possible [77], but not practical. When choosing a method for auditing, a trade-off between run-time and expected success rate to identify a vulnerability or data leak needs careful consideration. Any corresponding choices should be well documented.

Developing new evasion attacks is an active area of research [93]. Complex attacks are often preferred as they are more likely to find a malicious sample. An important distinction when auditing evasion is whether targeted (e.g., image should be classified as 'yield') or untargeted (e.g., output should not be 'stop sign') attacks are audited. Targeted attacks, while providing greater insight into model analysis and posing higher potential harm in autonomous driving, are more challenging to execute due to their requirement for proprietary access. They are also less transferable to unseen models. Therefore, both targeted and untargeted attacks should be incorporated in audits and robustness tests.

On the other hand, developing detection methods that provide security on whether a network is backdoored is an open research question [86]. We suggest testing several different state-of-the-art approaches in terms of heuristics and assumptions. Similarly, there is no conclusive result on how to detect or prevent privacy leakage from a model [62]. We suggest testing attacks introduced in research to get an estimate of existing vulnerability to inference.

4.2.2 Implementing cybersecurity audits

Implementing audits for a set of attacks or defences necessitates the use of specific libraries and tools, many of which focus on image classification. Table 3 provides an overview of these, including supported frameworks and their number of implemented attacks and defences. Notably, ART, SecML, and Privacy Meter support more diverse areas. SecML and Microsoft's counterfit[4] allow command-line robustness testing, while ART and foolbox offer plug-in capabilities.

Proper configuration and implementation are crucial for evaluation. In [101] the authors offer guidance on preventing evasion attack failures, addressing cases where models are vulnerable yet unsuited for attack-optimization. However, there is no equivalent work regarding privacy inference attacks or backdoor detection failures. Audit results can also be compared with public benchmarks like the Robust Bench for evasion on vision models[5] or BackdoorBench [99]. As before, there is a significant lack of benchmarks for non-vision tasks or other vulnerabilities.

4.2.3 A cybersecurity audit for a system composed by several AI components

While auditing individual AI components provides insights into their vulnerabilities, it does not guarantee an accurate approximation of the system's overall vulnerabilities. There is limited research on auditing a system comprising multiple AI components. A possible approach, inspired by the V-model from ISO26262 [102] software testing, involves testing each model in isolation before integrating and testing subsequent systems. Alternatively, testing the entire system end-to-end is suggested [72], with individual components only investigated upon system failure [103].

4.2.4 Cybersecurity audit reports for AI components in AVs

The output of AI components' cybersecurity audits in AVs should mirror a detailed defence evaluation [76]. This includes a rationale for chosen attacks/assessments, their parameters, and norms; benign and attack accuracies or assessment performance with associated hyperparameters; results of basic sanity checks; and the AV components involved in the test. Evasion sanity checks, as in [76], involve increasing attack perturbation leading to a rise in attack success rate, as well as substantial perturbation reducing model accuracy to random guessing levels. However, similar guidelines for backdoor testing or privacy assessments are currently unavailable. Privacy attack tests, though theoretically possible with overfitting leading to higher inference accuracy [62], are often unattainable due to the unavailability of the model owner's training data.

Audits should be conducted before AI model deployment and repeated upon significant model updates or advancements in attack techniques. To prevent model providers from tuning models to audit challenges, audit tests should not become public. Attacks can aid in developing robust models if applied during training and evaluation by ML engineers or specialized teams (e.g., AI red team). However, as stated above, to prevent overfitting to a specific attack, the audit must employ tests not used during training.

4.3 Challenges and Research Gaps

Our limited understanding of AI security threats presents challenges in auditing AI components, especially within complex systems like AVs. The discrepancy between theory and practice is evident in the wealth of academic work versus the scarcity of real-world incidents, resulting in a lack of realistic threat models. This hampers our ability to test relevant attacks and understand the effects an attack might have on an individual AI component within a larger system such as an AV. This is amplified by the limited experience in auditing standalone AI systems.

Models are often optimized to pass audits, leading to 'gaming' the audit rather than ensuring robustness. Given the limited number of attacks available in libraries (Table 3), this is a pressing concern. The re-implementation of attacks from scientific works can introduce potential errors and inaccurate robustness assessments. This problem is amplified for non-vision based AV modules, such as LiDAR-based, with fewer existing implementations.

While evasion attacks are most commonly implemented, followed by poisoning and IP/privacy attacks, our understanding of securing models against all these

[4] https://github.com/Azure/counterfit/tree/main/docs/source.

[5] https://robustbench.github.io/

Table 3 Overview of different libraries implementing different attacks on ML. We skip repositories with less than 5 evasion/poisoning or less than 2 privacy/IP attacks. The activity reports the last month/year where an activity took place as of December 2023

	Evasion		Poisoning		Privacy/IP		Supported frameworks										
Name	# Attacks	# Defenses	# Attacks	# Defenses	# Attacks	# Defenses	Pytorch	Tensorflow	JAX	MXNet	Caffe	SciKit	Non-vision Support	Github stars	Lastactivity	Released?	License
ART [94]	41	29	11	7	16	-	✓	✓		✓		✓	✓	4.2k	12/23	✓	MIT
Backdoor-Box [95]	-	-	12	7	-	-	✓						?	325	10/23		GPL 2.0
Foolbox [96]	17	-	-	-	-	-	✓	✓	✓					2.6k	11/23	✓	MIT
SecML [97]	10	1	3	-	-	-	✓	✓				✓	✓	126	5/23	✓	Apache2.0
Ad-vBox [98]	10	6	2	-	-	-	✓	✓			✓			1.3k	8/22		Apache2.0
Backdoor-Bench [99]	-	-	12	15	-	-	✓							249	12/23		CC BY-NC−4.0
Clev-erhans [100]	6	-	-	-	-	-	✓	✓	✓					6k	1/23	✓	MIT
ML-Doc-tor [62]	-	-	-	-	6	-	✓							75	12/23		Apache2.0
Privacy Meter[5]	-	-	-	-	7	-	✓	✓					✓	501	07/23		MIT
ML Hospital[6]	-	-	-	-	6	10	✓							39	04/23		MIT

[5] https://github.com/privacytrustlab/ml_privacy_meter

[6] https://github.com/TrustAIResearch/MLHospital

attacks remains shallow. The possibility of undetectable attacks creates a penalty-free zone for attackers, particularly when there is no significant misbehaviour.

In terms of privacy, we need focused studies assessing privacy risks associated with AI components of AVs, including sensitive training data, understanding the use of federated learning for AVs and securing it against privacy attacks.

5 Transparency through explainability

Explainability in AV testing and validation ensures system safety and reliability by providing an understanding of the underlying algorithms and decision-making processes. It enables performance evaluation in various scenarios and promotes transparency and accountability. As AVs become more prevalent, explaining their behavior and decisions to stakeholders and the public is crucial for building trust and confidence.

5.1 Explainability in AVs

In the following we provide a high-level description of the operational layers described in Sect. 1 a review of previous work on explainability in each operation. In [7], some of the most relevant barriers and questions for explainability are identified for some of these operational layers.

Localisation: it is critical for the safe operation of AVs [104]. The complexity of scenarios as well as weather and traffic conditions can affect the accuracy of localisation and thereby also have an impact on the planning and decision making of a vehicle. Embedding a form of estimation of performance and confidence in the odometry and localisation pipelines can improve their robustness and prevent critical failures [105, 106].

As safety is often considered as the most important design requirement, one of the goals of localisation is to ensure that the AV is aware whether it is within its lane [107]. Hence, providing information about the vehicle location over time including justifications as explanations can be crucial to both exposing error rates and preventing collisions. Such explanations and/or alerts might be given through specialised user interfaces, e.g., a special dashboard or mobile application as shown in [108].

While there seems to be less research related to explainable localisation, as discussed in [109], explanations concerning localisation can be useful for system developers for debugging AVs because they can facilitate positional error correction. Explanations can also provide other stakeholders with a perception of reliability and/or safety.

Perception: perception and scene understanding are fundamental requirements for providing intelligible explanations. To this end, several data-driven explanation methods have been applied to the AV perception task. To train, test and validate explanation methods datasets play a critical role.

Datasets for training machine learning approaches in the context of AVs [110] are an important resource for explanation generation. These datasets include different types of annotations, e.g., handcrafted explanations [111, 112], vehicle trajectories [113], human driver behaviour [114, 115] or anomaly identification with bounding boxes [112, 116], that are helpful for posthoc driving behaviour explanation. In [109], datasets have been categorised by the sensors used for the data collection (exteroception and proprioception), and the annotations that are useful for developing explainable AVs. Furthermore, different stakeholders that can potentially benefit from the explanations have been identified.

Vision-based explanation methods represent an important class of (deep learning) methods that have been proposed. Gradient-based or backpropagation methods are generally used for explaining convolutional neural network models [117]. In [118, 119] the reader can find extensive surveys on vision-based explanation methods, from Class Activation Map (CAM) [120] and its variants [121–124], to other methods [125–127]. Furthermore, in [111] an approach for textual explanation generation for AVs which was trained on the BDD-X dataset was proposed, and in [116] the authors focused on scene understanding and the generation of short explanation sequences, highlighting salient objects in input images that can potentially lead to hazards.

As for prediction, the need to understand road users' interactions was highlighted by the authors of the SCOUT Graph Attention Network (GAT) [12]. They proposed the use of Integrated Gradients [128] technique and visualization of learned attention. The illustrations provided for four carefully selected scenarios showed the importance of spatial representation of the nearest traffic participants. The use of the learned attention as a mechanism to generate explainable motion predictions was also proposed for transformer-based models [129]. The explainability of motion prediction was explored by use of conditional forecasting [130]. The authors of *What if Motion Prediction* (WIMP) model [130] developed an iterative, graphical attention approach with interpretable geometric (actor-lane) and social (actor-actor) relationships that support the injection of counterfactual geometric targets and social contexts. The proposed method supports the study of hypothetical or unlikely scenarios, so-called counterfactuals. In [13], the authors proposed an intention prediction layer that can be attached to any multi-modal motion prediction model, to enhance the interpretability of the outputs. The effectiveness of this approach was assessed through a survey that explores different elements in the visualisation of the multi-modal trajectories and intentions. Recently, in [14], the authors propose the use of several explainability techniques integrated in a heterogeneous graph neural network model. The explainability analysis contributed to a better understanding of the relationships between dynamic and static elements in traffic scenarios, facilitating the interpretation of the results, as well as the correction of possible errors in motion prediction models.

Planning: in AI planning and scheduling, a sequence of actions is devised for an AV to complete a task, which helps in decision-making within the environmental dynamics [131]. However, without explanations, the generated behaviour may confuse end-users. Explainable planning can enhance user experiences [109, 132], for instance, by translating agent's plans into comprehensible formats and designing user interfaces that reinforce understanding [133]. Relevant works like XAI-PLAN [134], WHY-PLAN [135], refinement-based planning (RBP) [136], plan explicability and predictability [137], and plan explanation for model reconciliation [138, 139] can provide frameworks for explainable planning in AVs.

Control: in relation to vehicle control, current ADAS interfaces now display rich digital maps [140], vehicle position, and track related attributes ahead or around the vehicle which can be considered as explanations. As discussed in [109], stakeholders might be interested in demanding further explanations when the AV makes low-level decisions different from their expectations.

Human-Vehicle Interaction: besides existing in-vehicle visual interfaces like mixed reality (MR) visualization [141] and reconfigurable dashboard panels [142], interfaces facilitating information exchange via voice, text, visual, gesture, or a combination of these modalities are essential. Human-vehicle interaction (HVI) plays a significant role in explaining AVs behaviours, and its relationship to explanations is discussed further in subsequent sections.

System Management: Event data recorders (EDRs) in vehicles log accident-related information, aiding in post-accident fault analysis [143]. Mandatory in US passenger vehicles since 2014, the National Transportation Safety Board (NTSB) recently highlighted the need for improved EDR requirements for ADS after the Tesla crash case in 2018 [144]. As AVs proliferate, distinguishing between human driver errors and AV errors, which could stem from poor design or defects [145, 146], is essential. These distinctions need to be clearly expressed in explanations. The current challenge of properly attributing faults to the correct traffic participant stems from difficulties in identifying and evaluating the precise cause of an accident [147].

The National Highway Traffic Safety Administration (NHTSA) urges industry and standard bodies to develop uniform data recording and sharing approaches[6] Existing inability of EDRs to provide sufficient data for crash reconstruction and liability analysis calls for solutions that can record a satisfactory number of parameters for vehicle behaviour reconstruction and explanation provision [148]. Proposed solutions include blockchain technologies [149] as well as smart [150] and robust [151] data models. With upcoming EU legislative rules on EDR [152] and similar regulations in China [153], the sufficiency of existing data storage facilities for AVs is under scrutiny. Efficient storage space management and well-defined data packages are necessary for event explanation purposes. These EDR advancements are crucial for developing accident explanation techniques, highlighting the under-explored area of explainable EDRs.

5.2 From AV sensor data to explanations

5.2.1 AV data recorders

AVs generate vast amounts of data, the recording, access, analysis, and usage of which can have significant technical, ethical, social, and legal implications [154]. A data recording device or Black Box (BB) can enhance safety, accountability, and trust in AVs by allowing accident investigation, logging system parameters, and reconstructing incident timelines respectively. The UNECE GRVA has proposed regulatory provisions for AV-specific data recorders, the Event Data Recorder (EDR) and the Data Storage System (DSSAV). The EDR collects information on the vehicle's behaviour and crash severity when a specific triggering event occurs (e.g. airbag activation). The DSSAV by contrast is "always on" and collects information on interactions between the driver and the system. However, these may not provide sufficient or appropriate data for practical and legal investigations, particularly in collisions or near-miss events with VRUs incidents. The ITU Focus Group on AI for Autonomous and Assisted Driving (FG-AI4AD)[7] is assessing public expectations for self-driving software post-collision behaviour, though potential ethical risks and legal implications may provoke resistance towards AV data recording. A Responsible Innovation (RI) approach, involving societal benefit, anticipatory governance, and stakeholder inclusion considerations, can mitigate negative outcomes from inadequate data logging. The impact and challenges of RI in practice are discussed in [154].

5.2.2 Generating explanations: automated commentary driving

Automated explanation generation in AVs is inspired by *commentary driving*, where drivers verbalize their

[6] See relevant documents here: https://www.nhtsa.gov/fmvss/event-data-recorders-edrs.

[7] https://www.itu.int/en/ITU-T/focusgroups/ai4ad/Pages/default.aspx.

observations, assessments, and intentions, enhancing their understanding and awareness of their surroundings. Automated driving commentary can offer understandable explanations of driving actions, assisting drivers in challenging and safety-critical scenarios. A field study conducted in [155] involved a research vehicle deployed in an urban environment to collect sensor data and driving commentary from an instructor. The analysis revealed an explanation style involving observations, plans, general remarks, and counterfactual comments. Using a transparent tree-based approach (see Fig. 4), they generated factual and counterfactual natural language explanations, with longitudinal actions (e.g., stop and move) deemed more intelligible and plausible than lateral actions by human judges. The study proposes ways to enhance robustness and effectiveness of explainability for ADAS and ADS.

5.3 Challenges and research gaps

Despite recent and ongoing advances in the general field of explainable AI (XAI), there are a considerable number of limitations [156] including reliability and robustness, lack of a standard evaluation framework and potential increase of automation bias. XAI is a valuable tool, but further research is needed. In the following, we summarize the main challenges for AVs.

5.3.1 Explaining the right thing, in the right way, at the right time for the right reason

Determining *what*, *how*, *when*, and *why* to explain are crucial considerations for explainer system designers. These challenges, lacking trivial or unique solutions, complicate the evaluation and comparison of system-generated explanations. Hence, it is unclear how regulators would determine the minimal requirements for an explainer system. The explanation's level of detail or fidelity depends on the application domain and stakeholder, with different groups requiring varied explanation types. For instance, accident investigators require more technical details than end-users. The manner and timing of explanations are also essential, with human factors playing a significant role. Some explanations aim to enhance understanding and trust, while others might be safety-critical. Ideally, explainer systems should adjust to the varying needs and requirements of the domain and stakeholders. This challenge calls for research into new algorithmic methods for generating effective explanations from AV sensor data and plans, clear standards or regulations on what and why AVs should explain (e.g., see [157]), and further human factors research to optimize explanation provision.

5.3.2 Explaining the unknown unknowns

AVs operating in open-ended environments will face unfamiliar situations or edge cases, potentially making it challenging to generate appropriate explanations. For instance, an AV may encounter an unclassifiable new type of road user, making explanations based on incorrect interpretations potentially misleading for stakeholders. To tackle this, AVs require meta-cognition or reasoning capabilities to assess if any functionality is compromised or uncertain. This necessitates new meta-reasoning methods and techniques for explaining this new type of inference to stakeholders.

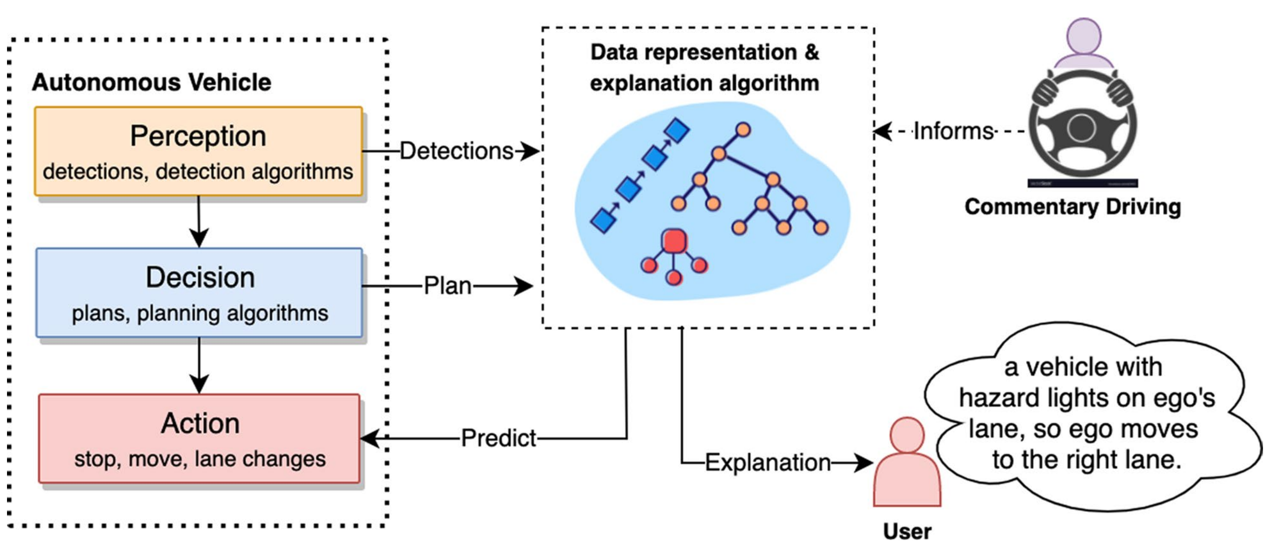


Fig. 4 From commentary driving, requirements for explanations were gathered to inform the design of factual and counterfactual explanation algorithms. The algorithms receive input data from the different autonomous driving operations, provide a structured representation, and generate intelligible explanations to an end-user

5.3.3 Explaining opaque AI models

Deep learning methods have significantly advanced the field of computer vision and are now fundamental in many applications, including AVs. Despite their high-performance, they lack interpretability and explainability due to their size, structure, and parameter count. Progress in XAI has driven research to make neural networks understandable, with many approaches focusing on vision-based explanations using attention. Although useful for building an intuitive understanding of otherwise opaque AI models, these explanations have limited explanatory power. Therefore, new approaches that can provide robust causal explanations are required.

6 Robustness and Fairness in AVs

This section examines the safety and ethical behaviour of AVs. We will discuss potential evaluation protocols and metrics for assessing predictive system robustness, and summarise the problem of fairness in AVs from the perspectives of perception and decision-making.

6.1 Evaluating Predictive System Robustness in AVs

As ADAS and ADS increase in complexity, conventional testing struggles to address all safety areas due to the vast array of safety-related systems and potential scenarios. The prediction pillar of the perception operational layer is crucial in forecasting agent-agent and agent-scene interactions, with models typically representing multimodal paths to account for environmental uncertainty, adding complexity to evaluation criteria [13]. Following is an overview of the evaluation metrics used to validate observed ground truth predictions, along with a discussion on the application of robustness analysis strategies.

Unimodal metrics: they evaluate single predictions using distance-based metrics, including Average Displacement Error (ADE), Final Displacement Error (FDE), and Heading Error. However, due to the multi-modal nature of human behaviour, metrics should evaluate multiple plausible outputs.

Multimodal metrics: measure the likelihood of all possible futures given past observations. Common metrics include Top-K, which samples K predictions and returns the best prediction performance based on a distance measure. Despite its popularity, Top-K is potentially misleading and fails to measure the social plausibility of all modes (e.g., the collision rate of each prediction). The ideal metric should also be able to quantify that the model does not generate bad trajectories Other metrics such as Average NLL [158] and Average Mahalanobis distance (AMD) [159] try to measure the likelihood of ground truth and variance/certainty of predicted outputs, respectively. The Percentage of Trajectory Usage [160] is a promising metric that captures accuracy, improving on Top-K. However, developing metrics that holistically capture multimodality remains an open issue [13].

Socially-aware metrics: measure social awareness, for example, by counting collisions in model outputs or assessing overlap rate [13].

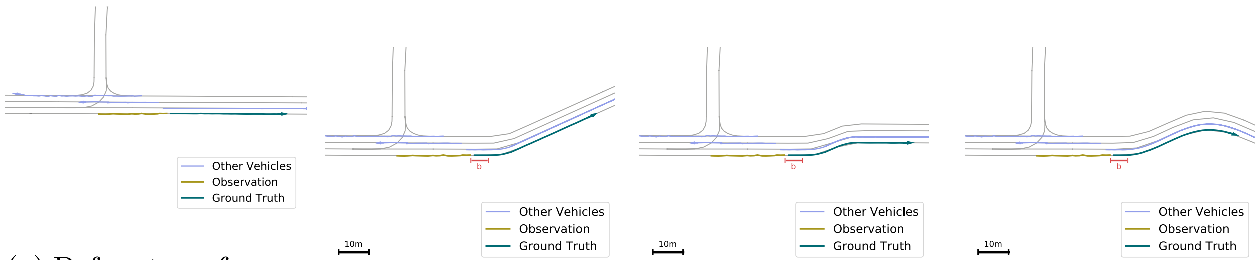
Scene-aware metrics: evaluate scene awareness, measuring variables such as the *number of off-roads*, Drivable Area Occupancy (DAO), Drivable Area Compliance (DAC), Scene Consistency Rate (SCR), or Oncoming Traffic Direction (OTD) [13].

Imbalanced datasets: are common in human mobility, with critical non-linear behaviours being rare (most of the data is repetitive linear behaviours). Even with millions of examples, not all behaviours may be observed in the training set. Thus, models should be evaluated on a diverse set of interactions to assess robustness. One approach is to categorize data by interaction type (e.g., static, linear, leader-follower, collision avoidance, road following, etc.) and evaluate each category independently [161]. Another strategy is to develop methods that automatically identify *realistic adversarial examples* that cause model failure.

Robustness assessment through adversarial examples: involves identifying realistic examples that cause prediction models to fail. Such examples could be specific interactions not seen in the training data. The task is to automatically find a realistic modification of the scene and agents' positions that cause prediction models to fail. Off-road predictions, for example, indicate a failure in the model's scene reasoning [162]. To find real examples where the models go off-road, the space of possible scenes should be explored either by learning generative models that mimic the dataset distribution or by using simple functions to transform the scene into new, realistic but challenging ones. Figure 5 illustrates how a scene can be automatically altered into a realistic example for evaluation. We can then quantitatively report the frequency of model failures in specific scenes (as shown in Fig. 6). Lastly, we can verify the plausibility or existence of the generated failing scenes. We can find a similar real scene using image-retrieval techniques, given a generated scene. Figure 7 depicts automatic retrieval of real-world scenes where models fail without the need for data collection from these scenes.

Similarly, changing agents' positions can help identify realistic interactions where models fail, such as predicting a collision. The sensitivity of various models in different timesteps with regard to collision avoidance can be studied [165], which might be explained by biases in the data or in the model structure.

Finally, while it is possible to identify agents' social interactions that could lead to a prediction failure, the robustness of small, realistic perturbations to the observed data can also be studied. For example, a model



(a) Before transformation (b) Single-turn (c) Double-turn (d) Ripple-road

Fig. 5 Visualization of different transformation functions. The scene before transformation is followed by three different transformations

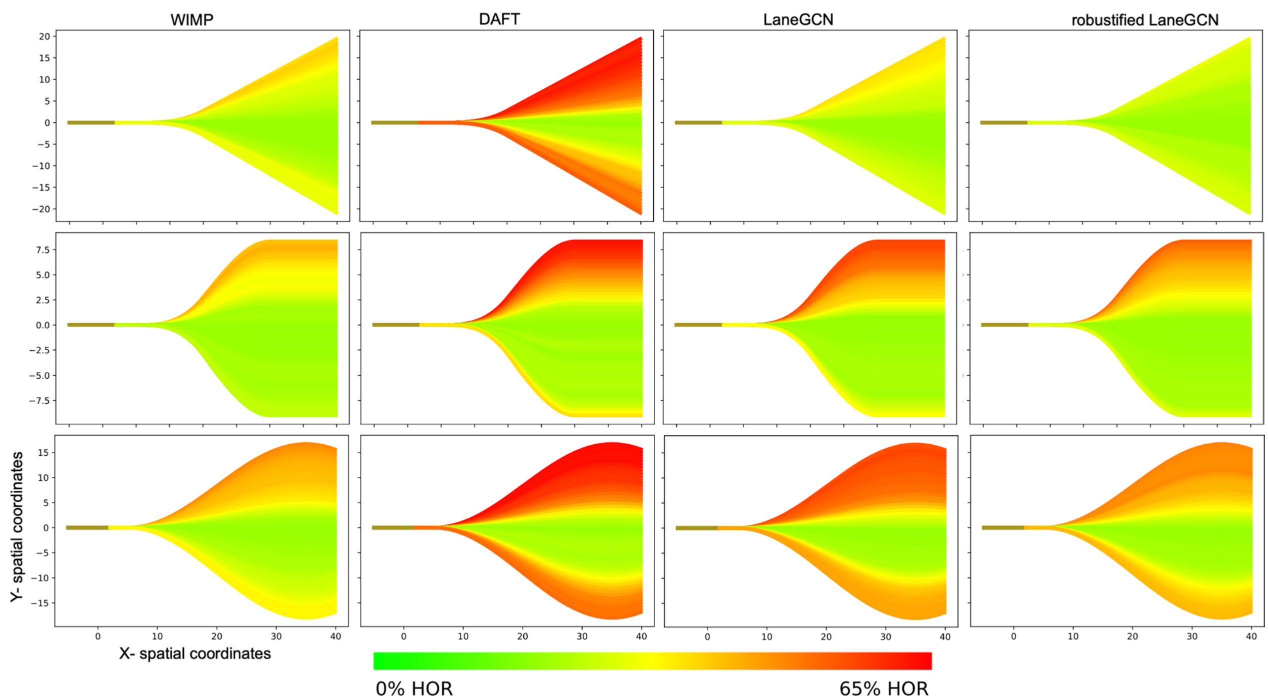
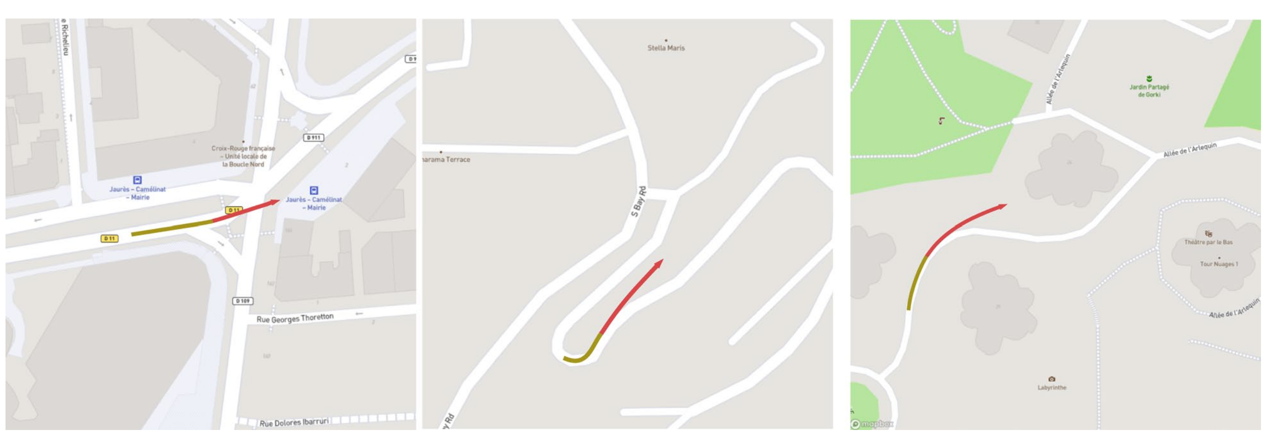


Fig. 6 The qualitative results of various methods (WIMP [163], DAFT [162], LaneGCN [164]) for different transformation functions. The red color indicates high Hard Off-road Rate (HOR) in those scenes



(a) Paris location (b) Hong Kong location (c) New Mexico location

Fig. 7 It is possible to automatically find more models fail without collecting data from these real-world scenes. The automatically generated scenes are used to retrieve real scenes. We can observe that a model fails in Paris (a), Hong Kong (b) and New Mexico (c) in a few milli-seconds

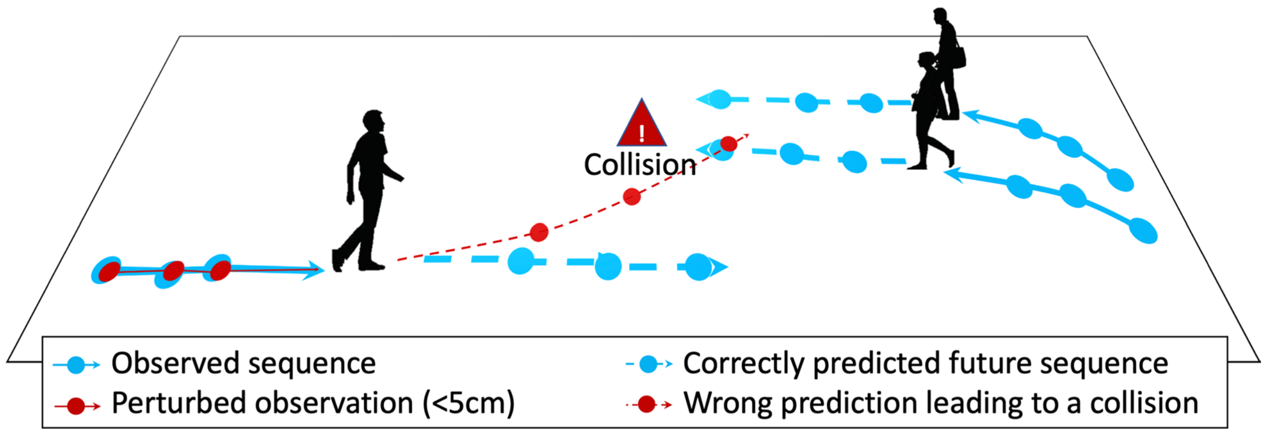


Fig. 8 Given the observed trajectories of the agents in the scene, a predictor forecasts the future positions reasonably (blue lines). However, with less than 5 cm perturbation in the observation trajectory (in red), an unacceptable collision is forecasted

might accurately predict future positions without collisions based on observed human trajectories. However, introducing a small perturbation of less than 5 cm to the observation trajectory could unexpectedly result in a collision between agent predictions, indicating incomplete social understanding by the predictors (as seen in Fig. 8). Unlike common adversarial attacks designed for classifiers [166], it is possible to design attacks for the trajectory prediction problem, which are a multimodal regression task. Realistic adversarial examples can be used to study the collision avoidance behaviour of trajectory prediction models. Specifically, we can investigate the worst-case social behaviour of a prediction model under small input perturbations, including various levels of ablation for agents and scene elements [14].

6.2 Fairness in AVs

Fairness and bias are crucial in AV development. As AVs advance, they must avoid unfair biases and ensure equal treatment for all road users, regardless of personal or group-based characteristics. Fairness in AVs is complex and value-driven. Previous attempts to encode moral decisions into AVs [167] have been problematic, potentially perpetuating biases and discrimination [168]. Instead, AVs should prioritise human life, adapt to different road users based on predictive perception models, and avoid discrimination [169]. The goal is to ensure equal safety for all road users while avoiding unfair biases and promoting fairness and equality [7]. We will next summarise potential sources of biases in perception systems and ethical goals in decision-making and planning.

6.2.1 Perception: bias and discrimination

Generally, the most advanced perception and prediction systems rely on deep learning AI models, which heavily depend on datasets for their development and evaluation. Although these approaches are becoming increasingly sophisticated and effective, the study of potential biases and resultant discrimination are often overlooked. Some preliminary studies found significant biases in different pedestrian detection methods based on variables such as age (adult/child) and sex/gender (male/female) [170, 171] and skin tone (light/dark) [172]. No available studies have addressed the problem with a focus on other type of road users, including vehicles [173].

The primary source of bias in perception and prediction AI models lies in the training data. Despite the availability of large datasets with millions of labelled samples for detecting people and vehicles, they usually lack attribute labels necessary for analysing potential biases. Attribute annotation is a challenging task that demands robust methodologies, particularly when multiple annotators are involved, to minimize uncertainty and prevent inaccurate, stereotype-based decisions [173]. In this recent study [173], the authors annotated over 140K protected attributes in widely used visual datasets for perception using a reliable annotation method, covering both persons and vehicles, and conducted a thorough evaluation of biases. This quantity of annotated attributes enables the implementation of fairness evaluation and testing studies with more significant results.

6.2.2 Decision-making: ethical goals

The implementation of vehicle automation is anticipated to improve road safety by eliminating common human driver errors [174, 175]. However, new risks may arise from factors like mode confusion, loss of situational awareness, and control transition problems [176–178]. Additionally, the infinite variety of situations that AVs may encounter could lead to new types of collision risks.

A report for the European Commission [179] included twenty recommendations addressing the ethics of AVs, covering topics such as road safety, privacy, fairness, explainability, and responsibility. The recommendations emphasized that AVs must decrease, or at least do not increase physical harm to users or other road users compared to conventional driving. Furthermore, even if AVs reduce the global level of risk, their deployment would be considered unethical if any individual group of road users experienced an increase in risk. The introduction of AVs also requires careful consideration of circumstances in which they might be permitted not to comply with traffic rules.

Current road rules use terminology that allows discretion by human drivers, such as driving with 'reasonable consideration' and 'due care'. Human drivers exercise this discretion based on experience, training and a general understanding of the road environment. AVs could be designed to follow digital versions of statutory road traffic rules. However, they may face challenges conforming to traffic law due to the complexity of their operating domains [180]. The complexity of domains in which AVs are expected to operate makes it impossible to anticipate or formulate explicit rules for every potential eventuality - environmental conditions, traffic and encounters with other road users will vary dramatically between domains and over time in any one domain. Any approach that assumes an AV could "deduce" normatively correct behaviour through exposure to a large number of training cases would need to overcome three extremely challenging practical difficulties. First, collecting a sufficient quantity and quality of scenarios to allow the right behaviours to be derived. Second, to derive the ethical principles underpinning which decisions or behaviours should be adopted. And third, to explain or justify its decisions or actions.

An alternative solution explored in [181] is the potential use of ethical goal functions, which capture societal expectations on how an AI should optimise outcomes. For an AV, this would involve setting out priorities for safe and efficient driving. Ethical goal functions could guide decisions where non-compliance with road rules is preferable and align AV behaviours with societal expectations.

Once an ethical goal function has been agreed and enacted by legislators, AVs could use it to determine the course of action with the highest utility. Factors considered within the ethical goal function to determine optimal speed, for example, might include potential collision risk, road suitability, infrastructure, vehicle and weather conditions, and the behaviour of other road users. A similar calculation that incorporates this function could assess whether mounting the pavement is permissible, in what circumstances and how that operation should be performed.

Ethical goal functions offer a means to enhance societal engagement and assist developers in AV deployment by providing algorithmic guidance on acceptable behaviours. Demonstrating how AVs have optimized against ethical goal functions could help show that their vehicles act in accordance with societal values, facilitating any possible liability claim process [1]. Also, it gives transport regulators an effective tool to monitor and moderate AV behaviours.

6.3 Challenges and research gaps

6.3.1 Robustness of prediction

Thoroughly assessing forecasting models necessitates extensive data collection, which is often impractical and costly. It is unrealistic to assume that all possible social interactions can be observed in data collection campaigns. Recent research suggests new testing and validation protocols using both real and synthetic data to investigate trajectory prediction model robustness [182, 183]. Another approach introduces real subject behaviour in autonomous driving simulators using virtual reality and motion capture systems [184–186]. More broadly, current procedures for testing prediction systems could benefit from improved fidelity (e.g., more realistic dummies), increased variability of testing conditions (e.g., dummies with different appearances, partially changing trajectories), and the addition of real behaviours and interactions in a safe, controlled manner (e.g., using simulated and interactive environments) [46].

6.3.2 Fair perception and behaviour in AVs

While AV deployment has taken longer than was anticipated, it has given time for researchers and society to consider exactly how we might expect AVs to behave and how we want them to integrate into society. The concept of fairness in AVs, including unbiased perception and ethical decision-making and planning, could ensure that society and human values steer the operation of AVs and help to ensure that when the technology arrives, it delivers on societal expectations. Further work is needed to develop accurate and unbiased perception methods, as well as the process for capturing societal values and establishing the framework for translating them into AV behaviours. However, this work is crucial for ensuring that AVs operate safely and ethically.

7 Study limitations

The present study has several limitations that should be acknowledged. One of the primary limitations is the reliance on expert opinions, both during the broader workshop and during the more targeted focus group, which, although very well-informed, can be subjective and influenced by individual perspectives and biases. While the use of a diverse group of experts can help mitigate this limitation, the scope of our study requires complementary and specific expertise for each topic addressed.

Therefore, expanding diversity would imply a much larger and more complex focus group to manage.

Another limitation of the study is its lack of empirical data. All data-driven insights are based on evidence from the state-of-the-art available. In addition, although our methodology implicitly implements a longitudinal perspective, we have focused our efforts on the most recent elements of the current state of trustworthy AI for AVs, without attempting to capture and model changes and trends over time.

Furthermore, although our approach covers many of the most important requirements of the trustworthy AI framework, as discussed in previous works [7, 8], the framework is very comprehensive, and as a result, many important aspects have inevitably been left out of our analysis, such as human agency and oversight, privacy and data governance, or accountability, among others.

8 Conclusion

AI is pivotal in the evolution of AVs, leading to high levels of automation. Despite the benefits, AI-specific features such as unpredictability, opacity, self and continuous learning, and lack of causality can introduce new uncertainties and safety issues. Recognizing this, international regulatory bodies are addressing the impact of AI on vehicle regulations. To harness the benefits of AI while mitigating potential risks, several frameworks have been proposed to guide the development of trustworthy AI systems, incorporating principles such as cybersecurity, robustness, fairness, transparency, privacy, accountability, and societal and environmental well-being. These frameworks present complex, varied challenges requiring multidisciplinary expertise. In this work, based on an incremental methodology that coordinates the work of a multi-disciplinary group of experts, we have presented a detailed analysis of the state of the art in cybersecurity, transparency, robustness, and fairness in the context of AV testing procedures, identifying future challenges.

We have discussed the role of AI in AVs, the challenges in their development, testing, regulation, and accident investigations, the need for cybersecurity audits, and the importance of transparency and explainability. AI plays a major role in AVs' functioning and their interaction with the environment. The EU's AI Act could influence AVs, but clarity is needed on its application, especially regarding AI components as safety components. The development of AVs is guided by the V-model and influenced by vehicle regulations, but current testing methodologies face limitations due to increasing complexity. Post-market surveillance and in-service monitoring are vital for ongoing safety assessment. Accident investigations involving AVs are complex and may benefit from lessons learned in aviation.

On the security front, structured cybersecurity audits for AI components in AVs are necessary to prevent threats. We propose a four-level framework for AI security research. Regular audits, especially before production and after significant updates, are advised. However, challenges in auditing include lack of realistic threat models, absence of real-world incidents, and potential for models to be optimized for passing audits. Transparency and explainability are key to understanding AVs' decision-making processes. This requires domain-specific explanations and understanding of the datasets used to train machine learning models. Challenges include determining what to explain and how, dealing with unknown scenarios, and explaining opaque AI models. Lastly, the paper discusses evaluation protocols for assessing predictive system robustness and ethical behaviour in AVs. Different metrics are used to evaluate predictive system robustness, but a holistic approach is needed. Additionally, the use of ethical goal functions to capture societal expectations for AV operation is considered. Challenges include the extensive data collection required for model assessment and the complexity of encoding ethical behaviour in AVs. To further enhance the impact of this study, future research should aim to develop and test practical solutions for implementing and scaling the trustworthy AI principles and methods identified in this research, with a focus on real-world applications and standardisation.

Acknowledgements
The views expressed in this article are purely those of the authors and may not, under any circumstances, be regarded as an official position of the European Commission.

Disclaimer
The views expressed in this article are purely those of the authors and may not, under any circumstances, be regarded as an official position of the European Commission.

Author contributions
D.F.LL, R.H, H.J, E.G, I.S and A.K. devised the project, the main conceptual ideas, and the general narrative; K.G. contributed to the cybersecurity analysis; L.K. contributed to the transparency analysis; P.S. contributed to the ex-ante requirements and testing analysis; R.S. contributed to the crash investigations section; N.R. and D.F. LL contributed to the fairness analysis. A.A. contributed to the robustness section. D.F.LL contributed to the introduction, AI for AVs, and post-marketing monitoring sections. All authors have read and approved the manuscript.

Funding
This work was mainly funded by ER Activity "extrAIsafe" (ER Activity call 2021-2) by the Joint Research Centre (JRC) Scientific Development Programmes (European Commission). It is also partially funded by some of the projects included in the Trustworthy AI portfolio of the JRC.

Data availability
Not applicable.

Declarations

Competing interests
The authors declare that they have no Competing interests.

Received: 27 December 2023 / Accepted: 30 June 2025
Published online: 30 July 2025

References

1. Fernandez Llorca, D., Charisi, V., Hamon, R., Sánchez, I., & Gómez, E. (2023). Liability regimes in the age of ai: a use-case driven analysis of the burden of proof. *Journal of Artificial Intelligence Research, 76*, 613–64.
2. UNECE WP.29/GRVA: Technical workshop on Artificial Intelligence. GRVA Working Party on Automated/Autonomous and Connected Vehicles (2022). https://unece.org/transport/events/grva-technical-workshop-artificial-intelligence
3. OECD: The OECD AI Principles. Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449 (2023). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
4. AI HLEG: Ethics guidelines for trustworthy AI. High Level Expert group on Artificial Intelligence, European Commission, Brussels, Belgium, B-1049 (2019)
5. AI HLEG: The assessment list for trustworthy artificial intelligence (ALTAI) for self assessment. High Level Expert group on Artificial Intelligence, European Commission, Brussels, Belgium, B-1049 (2020)
6. 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)
7. Fernández Llorca, D., & Gómez, E. (2021). Trustworthy autonomous vehicles. EUR 30942 EN, Publications Office of the European Union, Luxembourg **JRC127051** https://doi.org/10.2760/120385
8. Fernández-Llorca, D., & Gómez, E. (2023). Trustworthy artificial intelligence requirements in the autonomous driving domain. *Computer, 56*(2), 29–39. https://doi.org/10.1109/MC.2022.3212091
9. Cañas, P., Nieto, M., Otaegui, O., & Rodriguez, I. (2025). A methodology to enhance transparency for trustworthy artificial intelligence for cooperative, connected, and automated mobility. SAE Int. J. CAV **8**(1) https://doi.org/10.4271/12-08-01-0010
10. Belmecheri, N., Gotlieb, A., Lazaar, N., & Spieker, H. (2024). Towards trustworthy automated driving through qualitative scene understanding and explanations. arXiv preprint arXiv:2403.16908
11. Fang, Y., Min, H., Wu, X., Lei, X., Chen, S., Teixeira, R., & Zhao, X. (2023). Toward interpretability in fault diagnosis for autonomous vehicles: Interpretation of sensor data anomalies. *IEEE Sensors Journal, 23*(5), 5014–5027. https://doi.org/10.1109/JSEN.2023.3236838
12. Carrasco, S., Llorca, D.F., & Sotelo, M.A. (2021) SCOUT: Socially-COnsistent and UndersTandable Graph Attention Network for Trajectory Prediction of Vehicles and VRUs. In: 2021 IEEE Intelligent Vehicles Symposium (IV), pp. 1501–1508. https://doi.org/10.1109/IV48863.2021.9575874
13. Carrasco Limeros, S., Majchrowska, S., Johnander, J., Petersson, C., Sotelo, M.A., & Llorca, D.F. (2023). Towards trustworthy multi-modal motion prediction: Holistic evaluation and interpretability of outputs. CAAI Transactions on Intelligence Technology, 1–16 https://doi.org/10.1049/cit2.12244
14. Limeros, S.C., Majchrowska, S., Johnander, J., Petersson, C., & Llorca, D.F. (2023) Towards Explainable Motion Prediction using Heterogeneous Graph Representations. Transportation Research Part C: Emerging Technologies **157**(104405) https://doi.org/10.1016/j.trc.2023.104405
15. Geisslinger, M., Poszler, F., & Lienkamp, M. (2023). An ethical trajectory planning algorithm for autonomous vehicles. Nat Mach Intell 5, 137–144 https://doi.org/10.1038/s42256-022-00607-z
16. Khandelwal, V. (2024). Building Trustworthy AI Systems: Developing Explainable Models for Transparent Decision-Making in Autonomous Vehicles. Journal of Sustainable Solutions 1(4), 27–37 https://doi.org/10.36676/j.sust.sol.v1.i4.21
17. Brock, L., & Tropnas, L. (2018). Survey on the regulations of autonomous vehicles. *Journal of Multidisciplinary Research, 10*(3), 23–44.
18. Hemphill, T.A. (2020). Autonomous vehicles: U.s. regulatory policy challenges. Technology in Society 61, 101232 https://doi.org/10.1016/j.techsoc.2020.101232
19. Lee, D., & Hess, D.J. (2020). Regulations for on-road testing of connected and automated vehicles: Assessing the potential for global safety harmonization. Transportation Research Part A: Policy and Practice 136, 85–98 https://doi.org/10.1016/j.tra.2020.03.026
20. Bin-Nun, A.Y., Derler, P., Mehdipour, N., & Tebbens, R.D. (2022). How should autonomous vehicles drive? policy, methodological, and social considerations for designing a driver. Humanities and Social Sciences Communications **9**(299)
21. Abbasi, S., & Rahmani, A.M. (2023). Artificial intelligence and software modeling approaches in autonomous vehicles for safety management: A systematic review. Information **14**(10) https://doi.org/10.3390/info14100555
22. Shaterabadi, M., Ahmadi Jirdehi, M., Mehrjerdi, H., & Karimi, H. (2024). Chapter 5 - artificial intelligence for autonomous vehicles: Comprehensive outlook. Handbook of Power Electronics in Autonomous and Electric Vehicles, 59–79 https://doi.org/10.1016/B978-0-323-99545-0.00002-6
23. Hoppe, I., Hagemann, W., Stierand, I., Hahn, A., & Bolles, A. (2024). Challenges for trustworthy autonomous vehicles: Let us learn from life. *Systems Engineering, 27*(4), 789–800. https://doi.org/10.1002/sys.21744
24. Goh, T. J., & Ho, S. S. (2024). Trustworthiness of policymakers, technology developers, and media organizations involved in introducing ai for autonomous vehicles: A public perspective. *Science Communication, 46*(5), 584–618. https://doi.org/10.1177/10755470241248169
25. Kriston, A., Hamon, R., Fernandez Llorca, D., Junklewitz, H., Sanchez Martin, J.I., & Gomez Gutierrez, E. (2023). Toward explainable, robust and fair ai in automated and autonomous vehicles. EUR 31030 EN, Publications Office of the European Union, Luxembourg **JRC130621** https://doi.org/10.2760/95650
26. Jo, K., Kim, J., Kim, D., Jang, C., & Sunwoo, M. (2014). Development of autonomous car-part i: Distributed system architecture and development process. *IEEE Transactions on Industrial Electronics, 61*(12), 7131–7140.
27. Ma, Y., Wang, Z., Yang, H., & Yang, L. (2020). Artificial intelligence applications in the development of autonomous vehicles: a survey. *IEEE/CAA Journal of Automatica Sinica, 7*(2), 315–329.
28. Paden, B., Cap, M., Yong, S. Z., Yershov, D., & Frazzoli, E. (2016). A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on Intelligent Vehicles, 1*(1), 33–55. https://doi.org/10.1109/TIV.2016.2578706
29. Aradi, S. (2022). Survey of deep reinforcement learning for motion planning of autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems, 23*(2), 740–759. https://doi.org/10.1109/TITS.2020.3024655
30. Khodayari, A., Ghaffari, A., Ameli, S., & Flahatgar, J. (2010). A historical review on lateral and longitudinal control of autonomous vehicle motions. In: International Conference on Mechanical and Electrical Technology, pp. 421–429
31. Rasouli, A., & Tsotsos, J. K. (2020). Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE Transactions on Intelligent Transportation Systems, 21*(3), 900–918. https://doi.org/10.1109/TITS.2019.2901817
32. Roh, J., Paxton, C., Pronobis, A., Farhadi, A., & Fox, D. (2020). Conditional driving from natural language instructions. In: Kaelbling, L.P., Kragic, D., Sugiura, K. (eds.) Proceedings of the Conference on Robot Learning, vol. 100, pp. 540–551
33. European Union: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. Document 32024R1689 (2024)
34. European Parliament and Council: Regulation (EU) 2018/858 of 30 May 2018 on the approval and market surveillance of motor vehicles and their trailers, and of systems, components and separate technical units intended for such vehicles (2018). https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32018R0858
35. European Parliament and Council: Regulation (EU) 2019/2144 of 27 November 2019 on type-approval requirements for motor vehicles and their trailers, and systems, components and separate technical units intended for such vehicles, as regards their general safety and the protection of vehicle occupants and vulnerable road users (2019). https://eur-lex.europa.eu/eli/reg/2019/2144/oj
36. European Parliament and the Council: Regulation (EU) 2023/1230 of 14 June 2023 on machinery (2023). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32023R1230
37. European Parliament and the Council: Directive 2014/33/EU on the harmonisation of the laws of the Member States relating to lifts and safety components for lifts (2014). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32014L0033#d1e32-277-1
38. Lölfing, N. (2023). Impact of the EU's AI Act proposal on automated and autonomous vehicles. Bird & Bird. https://www.twobirds.com/en/insights/2023/global/impact-of-the-eus-ai-act-proposal-on-automated-and-autonomous-vehicles
39. Martin, H., Tschabuschnig, K., Bridal, O., Watzenig, D.: In: Watzenig, D., & Horn, M. (eds.) (2017). Functional Safety of Automated Driving Systems: Does ISO

26262 Meet the Challenges?, pp. 387–416. Springer, Cham. https://doi.org/10.1007/978-3-319-31895-0_16
40. Euro NCAP: Euro NCAP Vision 2030 - A Safer Future for Mobility (2022). https://cdn.euroncap.com/media/74468/euro-ncap-roadmap-vision-2030.pdf
41. European Parliament and Council: Directive 2007/46/EC of 5 September 2007 establishing a framework for the approval of motor vehicles and their trailers, and of systems, components and separate technical units intended for such vehicles (2007). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32007L0046
42. European Parliament and of the Council: Commission Implementing Regulation (EU) 2022/1426 of 5 August 2022 laying down rules for the application of Regulation (EU) 2019/2144 as regards uniform procedures and technical specifications for the type-approval of the automated driving system (ADS) of fully automated vehicles (2022). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R1426
43. ECE/TRANS/WP.29/2020/81: A new UN Regulation on uniform provisions concerning the approval of vehicles with regards to Automated Lane Keeping System (2020). https://unece.org/fileadmin/DAM/trans/doc/2020/wp29/ECE-TRANS-WP29-2020-081e.pdf
44. ECE/TRANS/WP.29/2020/9: Uniform provisions concerning the approval of motor vehicles with regard to the Advanced Emergency Braking System (AEBS) for M1 and N1 vehicles (2020). https://documents-dds-ny.un.org/doc/UNDOC/GEN/G19/351/86/PDF/G1935186.pdf
45. Menzel, T., Bagschik, G., & Maurer, M. (2018). Scenarios for Development, Test and Validation of Automated Vehicles. In: 2018 IEEE Intelligent Vehicles Symposium (IV), pp. 1821–1827.https://doi.org/10.1109/IVS.2018.8500406
46. Gonzalo, R. I., Maldonado, C. S., Ruiz, J. A., Alonso, I. P., Llorca, D. F., & Sotelo, M. A. (2022). Testing predictive automated driving systems: Lessons learned and future recommendations. *IEEE Intelligent Transportation Systems Magazine, 14*(6), 77–93. https://doi.org/10.1109/MITS.2022.3170649
47. Bonnel, P., Clairotte, M., Cotogno, G., Gruening, C., Loos, R., Manara, D., Melas, A., Selleri, T., Tutuianu, M., Valverde Morales, V., Forloni, F., Giechaskiel, B., Carriero, M., Otura Garcia, M., Pavlovic, J., Suarez Bertoa, R., Martini, G., & Krasenbrink, A. (2022). European market surveillance of motor vehicles. EUR 31030 EN, Publications Office of the European Union, Luxembourg **JRC128360** https://doi.org/10.2760/59856
48. UNECE GRVA: SG3 In Service Monitoring and Reporting. VMAD Technical Workshop on In-Service Safety Performance of Automated Vehicles (2022). https://wiki.unece.org/download/attachments/161841215/VMAD-SG3-20-02%20workshop%20presentation_final.pdf
49. European Commission: Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts. COM(2021) 206 (2021)
50. Robitaille, D. E. (2011). *Root Cause Analysis*. Chico, California: Basic Tools and Techniques. Paton Professional.
51. Stamatelatos, M., Vesely, W., Dugan, J., Fragola, J., Minarick, J., & Railsback, J. (2002). Fault Tree Handbook with Aerospace Applications. NASA Office of Safety and Mission Assurance
52. BEA: Final Report On the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro - Paris (2012). https://bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf
53. Ladkin, P. (1996). AA965 Cali Accident Report. http://sunnyday.mit.edu/accidents/calirep.html
54. NTBS - Reuters: U.S. NTSB closes probe into fatal Tesla 2020 California crash. https://www.reuters.com/business/autos-transportation/us-ntsb-closes-probe-into-fatal-tesla-2020-california-crash-2021-11-16/ (2021)
55. NTSB: Collision Between Car Operating with Partial Driving Automation and Truck-Tractor Semitrailer (2019). https://www.ntsb.gov/investigations/Pages/HWY19FH008.aspx
56. Standardization, S.o.A.E. (2021). ISO/SAE 21434:2021 Road vehicles - Cybersecurity engineering. International Organization for Standardization Geneva, Switzerland
57. Salay, R., Queiroz, R., & Czarnecki, K. (2017). An analysis of iso 26262: Using machine learning safely in automotive software. arXiv preprint arXiv:1709.02435
58. Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733
59. Xiang, Z., Miller, D.J., Chen, S., Li, X., & Kesidis, G. (2021). A backdoor attack against 3d point cloud classifiers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7597–7607
60. Wen, X., Jiang, W., Zhan, J., Bian, C., & Song, Z. (2021). Generative strategy based backdoor attacks to 3d point clouds: work-in-progress. In: Proceedings of the 2021 International Conference on Embedded Software, pp. 23–24
61. Morgulis, N., Kreines, A., Mendelowitz, S., & Weisglass, Y. (2019). Fooling a real car with adversarial traffic signs. arXiv preprint arXiv:1907.00374
62. Liu, Y., Wen, R., He, X., Salem, A., Zhang, Z., Backes, M., De Cristofaro, E., Fritz, M., & Zhang, Y. (2022). [CDATA[\{]]$\{$ML-Doctor$\}$[CDATA[\}]]: Holistic risk assessment of inference attacks against machine learning models. In: 31st USENIX Security Symposium (USENIX Security 22), pp. 4525–4542
63. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 3–18. IEEE
64. Zhu, Y., Cheng, Y., Zhou, H., & Lu, Y. (2021). Hermes attack: Steal $\{DNN\}$ [CDATA[\{DNN\}]] models with lossless inference accuracy. In: 30th USENIX Security Symposium (USENIX Security 21)
65. Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., & Ristenpart, T. (2016). Stealing machine learning models via prediction $\{$[CDATA[\{]]APIs$\}$[CDATA[\}]]. In: 25th USENIX Security Symposium (USENIX Security 16), pp. 601–618
66. McLachlan, S., Schafer, B., Dube, K., Kyrimi, E., & Fenton, N. (2022). Tempting the fate of the furious: cyber security and autonomous cars. International Review of Law, Computers & Technology, 1–21
67. Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., Rahmati, A., & Song, D. (2017). Robust physical-world attacks on machine learning models. arXiv preprint arXiv:1707.08945 2(3), 4
68. Kurakin, A., Goodfellow, I.J., & Bengio, S. (2018). Adversarial examples in the physical world. In: Artificial Intelligence Safety and Security, pp. 99–112. Chapman and Hall/CRC,???
69. Tu, J., Ren, M., Manivasagam, S., Liang, M., Yang, B., Du, R., Cheng, F., & Urtasun, R. (2020). Physically realizable adversarial examples for lidar object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13716–13725
70. Papernot, N., McDaniel, P., & Goodfellow, I. (2016). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277
71. Vemprala, S., & Kapoor, A. (2021). Adversarial attacks on optimization based planners. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 9943–9949. IEEE
72. Boloor, A., He, X., Gill, C., Vorobeychik, Y., & Zhang, X. (2019). Simple physical adversarial examples against end-to-end autonomous driving models. In: 2019 IEEE International Conference on Embedded Software and Systems (ICESS), pp. 1–7. IEEE
73. McGregor, S. (2021). Preventing repeated real world ai failures by cataloging incidents: The ai incident database. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 15458–15463
74. Grosse, K., Bieringer, L., Besold, T.R., Biggio, B., & Krombholz, K. (2022). " why do so?"–a practical perspective on machine learning security. arXiv preprint arXiv:2207.05164
75. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. Advances in neural information processing systems **32**
76. Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., & Kurakin, A. (2019). On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705
77. Pulina, L., & Tacchella, A. (2010). An abstraction-refinement approach to verification of artificial neural networks. In: International Conference on Computer Aided Verification, pp. 243–257. Springer
78. Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., Lupu, E.C., & Roli, F. (2017). Towards poisoning of deep learning algorithms with back-gradient optimization. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 27–38
79. Jiang, W., Li, H., Liu, S., Luo, X., & Lu, R. (2020). Poisoning and evasion attacks against deep learning algorithms in autonomous vehicles. *IEEE transactions on vehicular technology, 69*(4), 4439–4449.
80. Ma, Y., Zhang, X., Sun, W., & Zhu, J. (2019). Policy poisoning in batch reinforcement learning and control. Advances in Neural Information Processing Systems **32**
81. Wang, Y., Sarkar, E., Li, W., Maniatakos, M., & Jabari, S. E. (2021). Stop-and-go: Exploring backdoor attacks on deep reinforcement learning-based traffic congestion control systems. *IEEE Transactions on Information Forensics and Security, 16*, 4772–4787.

82. Han, X., Xu, G., Zhou, Y., Yang, X., Li, J., & Zhang, T. (2022). Physical backdoor attacks to lane detection systems in autonomous driving. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 2957–2968
83. Frederickson, C., Moore, M., Dawson, G., & Polikar, R. (2018). Attack strength vs. detectability dilemma in adversarial ml. In: Int. Joint Conf. on Neural Networks, IJCNN 2018, pp. 1–8. IEEE,???
84. Shokri, R., et al. (2020) Bypassing backdoor detection algorithms in deep learning. In: IEEE Euro S &P 2020, pp. 175–183. IEEE
85. Xiang, Z., Miller, D.J., Chen, S., Li, X., & Kesidis, G. (2022). Detecting backdoor attacks against point cloud classifiers. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3159–3163. IEEE
86. Cinà, A. E., Grosse, K., Demontis, A., Vascon, S., Zellinger, W., Moser, B. A., Oprea, A., Biggio, B., Pelillo, M., & Roli, F. (2023). Wild patterns reloaded: A survey of machine learning security against training data poisoning. *ACM Computing Surveys, 55*(13s), 1–39.
87. Melis, L., Song, C., De Cristofaro, E., & Shmatikov, V. (2019). Exploiting unintended feature leakage in collaborative learning. In: 2019 IEEE Symposium on Security and Privacy (SP), pp. 691–706. IEEE
88. Geiping, J., Bauermeister, H., Dröge, H., & Moeller, M. (2020). Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems, 33*, 16937–16947.
89. Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., & Lin, Z. (2021). When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR), 54*(2), 1–36.
90. Mir, D. (2012). Differentially-private learning and information theory. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops, pp. 206–210
91. Wang, Y.-X., Fienberg, S., & Smola, A. (2015). Privacy for free: Posterior sampling and stochastic gradient monte carlo. In: International Conference on Machine Learning, pp. 2493–2502. PMLR
92. Luo, B., Liu, Y., Wei, L., & Xu, Q. (2018). Towards imperceptible and robust adversarial example attacks against neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32
93. Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International Conference on Machine Learning, pp. 2206–2216. PMLR
94. Nicolae, M.-I., Sinn, M., Tran, M.N., Buesser, B., Rawat, A., Wistuba, M., Zantedeschi, V., Baracaldo, N., Chen, B., Ludwig, H., et al. (2018). Adversarial robustness toolbox v1. 0.0. arXiv preprint arXiv:1807.01069
95. Li, Y., Ya, M., Bai, Y., Jiang, Y., & Xia, S.-T. (2023). BackdoorBox: A python toolbox for backdoor learning. In: ICLR Workshop
96. Rauber, J., Brendel, W., & Bethge, M. (2017). Foolbox: A python toolbox to benchmark the robustness of machine learning models. arXiv preprint arXiv:1707.04131
97. Pintor, M., Demetrio, L., Sotgiu, A., Melis, M., Demontis, A., & Biggio, B. (2022). secml: Secure and explainable machine learning in python. *SoftwareX, 18*, Article 101095.
98. Goodman, D., Xin, H., Yang, W., Yuesheng, W., Junfeng, X., & Huan, Z. (2020). Advbox: a toolbox to generate adversarial examples that fool neural networks. arXiv:2001.05574
99. Wu, B., Chen, H., Zhang, M., Zhu, Z., Wei, S., Yuan, D., & Shen, C. (2022). Backdoorbench: A comprehensive benchmark of backdoor learning. *Advances in Neural Information Processing Systems, 35*, 10546–10559.
100. Papernot, N., Faghri, F., Carlini, N., Goodfellow, I., Feinman, R., Kurakin, A., Xie, C., Sharma, Y., Brown, T., Roy, A., Matyasko, A., Behzadan, V., Hambardzumyan, K., Zhang, Z., Juang, Y.-L., Li, Z., Sheatsley, R., Garg, A., Uesato, J., Gierke, W., Dong, Y., Berthelot, D., Hendricks, P., Rauber, J., & Long, R. (2018). Technical report on the cleverhans v2.1.0 adversarial examples library. arXiv preprint arXiv:1610.00768
101. Pintor, M., Demetrio, L., Sotgiu, A., Manca, G., Demontis, A., Carlini, N., Biggio, B., & Roli, F. (2022). Indicators of attack failure: Debugging and improving optimization of adversarial examples. NeurIPS
102. Standardization, I. O. (2011). *ISO 26262: Road vehicles-functional safety*. International Organization for Standardization Geneva.
103. Nushi, B., Kamar, E., Horvitz, E., & Kossmann, D. (2017). On human intellect and machine failures: Troubleshooting integrative machine learning systems. In: Thirty-First AAAI Conference on Artificial Intelligence
104. Wang, L., Zhang, Y., & Wang, J. (2017). Map-based localization method for autonomous vehicles using 3d-lidar. *IFAC-PapersOnLine, 50*(1), 276–281.
105. Aldera, R., Martini, D.D., Gadd, M., & Newman, P. (2019). What Could Go Wrong? Introspective Radar Odometry in Challenging Environments. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pp. 2835–2842. https://doi.org/10.1109/ITSC.2019.8917111
106. García Daza, I., Rentero, M., Salinas Maldonado, C., Izquierdo Gonzalo, R., Hernández Parra, N., Ballardini, A., & Fernandez Llorca, D. (2020). Fail-aware lidar-based odometry for autonomous vehicles. Sensors **20**(15) https://doi.org/10.3390/s20154097
107. Reid, T.G., Houts, S.E., Cammarata, R., Mills, G., Agarwal, S., Vora, A., & Pandey, G. (2019). Localization requirements for autonomous vehicles. arXiv preprint arXiv:1906.01061
108. Schneider, T., Hois, J., Rosenstein, A., Ghellal, S., Theofanou-Fülbier, D., & Gerlicher, A.R. (2021). Explain yourself! transparency for positive ux in autonomous driving. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–12
109. Omeiza, D., Webb, H., Jirotka, M., & Kunze, L. (2022). Explanations in Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems, 23*(8), 10142–10162. https://doi.org/10.1109/TITS.2021.3122865
110. Janai, J., Güney, F., Behl, A., Geiger, A., et al. (2020). Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends in Computer Graphics and Vision, 12*(1–3), 1–308.
111. Kim, J., Rohrbach, A., Darrell, T., Canny, J., & Akata, Z. (2018). Textual explanations for self-driving vehicles. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 563–578
112. You, T., & Han, B. (2020). Traffic accident benchmark for causality recognition. In: European Conference on Computer Vision, pp. 540–556. Springer
113. Houston, J., Zuidhof, G., Bergamini, L., Ye, Y., Chen, L., Jain, A., Omari, S., Iglovikov, V., & Ondruska, P. (2020). One thousand and one hours: Self-driving motion prediction dataset. arXiv preprint arXiv:2006.14480
114. Ramanishka, V., Chen, Y.-T., Misu, T., & Saenko, K. (2018). Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7699–7707
115. Shen, Y., Jiang, S., Chen, Y., Yang, E., Jin, X., Fan, Y., & Campbell, K.D. (2020). To explain or not to explain: A study on the necessity of explanations for autonomous vehicles. arXiv preprint arXiv:2006.11684
116. Xu, Y., Yang, X., Gong, L., Lin, H.-C., Wu, T.-Y., Li, Y., & Vasconcelos, N. (2020). Explainable object-induced action decision for autonomous vehicles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9523–9532
117. Das, A., & Rad, P. (2020). Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. arXiv preprint arXiv:2006.11371
118. Tjoa, E., & Guan, C. (2019). A survey on explainable artificial intelligence (XAI): towards medical XAI. arXiv preprint arXiv:1907.07374
119. Zablocki, É., Ben-Younes, H., Pérez, P., & Cord, M. (2021). Explainability of vision-based autonomous driving systems: Review and challenges. arXiv preprint arXiv:2101.05307
120. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929
121. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626
122. Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V.N. (2018). Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847
123. Tang, Z., Chuang, K. V., DeCarli, C., Jin, L.-W., Beckett, L., Keiser, M. J., & Dugger, B. N. (2019). Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline. *Nature communications, 10*(1), 1–14.
124. Omeiza, D., Speakman, S., Cintas, C., & Weldermariam, K. (2019). Smooth Grad-CAM++: an enhanced inference level visualization technique for deep convolutional neural network models. arXiv preprint arXiv:1908.01224
125. Bojarski, M., Choromanska, A., Choromanski, K., Firner, B., Ackel, L.J., Muller, U., Yeres, P., & Zieba, K. (2018). Visualbackprop: Efficient visualization of CNNs for autonomous driving. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 4701–4708
126. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature communications, 10*(1), 1–8.
127. Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. arXiv preprint arXiv:1704.02685

128. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. In: 34th International Conference on Machine Learning
129. Zhang, K., & Li, L. (2022). Explainable multimodal trajectory prediction using attention models. Transportation Research Part C: Emerging Technologies 143, 103829 https://doi.org/10.1016/j.trc.2022.103829
130. Khandelwal, S., Qi, W., Singh, J., Hartnett, A., & Ramanan, D. (2020). What-If Motion Prediction for Autonomous Driving. arXiv:2008.10587
131. Ingrand, F., & Ghallab, M. (2017). Deliberation for autonomous robots: A survey. *Artificial Intelligence, 247*, 10–44.
132. Chakraborti, T., Sreedharan, S., & Kambhampati, S. (2020). The emerging landscape of explainable automated planning & decision making
133. Sado, F., Loo, C.K., Kerzel, M., & Wermter, S. (2020). Explainable goal-driven agents and robots–a comprehensive review and new framework. arXiv preprint arXiv:2004.09705
134. Borgo, R., Cashmore, M., & Magazzeni, D. (2018). Towards providing explanations for AI planner decisions. arXiv preprint arXiv:1810.06338
135. Korpan, R., & Epstein, S. L. (2018). *Toward natural explanations for a robot's navigation plans*. Human-Robot Interaction: Notes from the Explainable Robotic Systems Workshop.
136. Bidot, J., Biundo, S., Heinroth, T., Minker, W., Nothdurft, F., & Schattenberg, B. (2010). Verbal plan explanations for hybrid planning. In: MKWI, pp. 2309–2320. Citeseer
137. Zhang, Y., Sreedharan, S., Kulkarni, A., Chakraborti, T., Zhuo, H.H., & Kambhampati, S. (2017). Plan explicability and predictability for robot task planning. In: IEEE International Conference on Robotics and Automation (ICRA), pp. 1313–1320
138. Chakraborti, T., Sreedharan, S., Zhang, Y., & Kambhampati, S. (2017). Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. arXiv preprint arXiv:1701.08317
139. Chakraborti, T., Sreedharan, S., Grover, S., & Kambhampati, S. (2019). Plan Explanations as Model Reconciliation–An Empirical Study. In: 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 258–266
140. TomTom launches map-based ADAS software platform Virtual Horizon. Accessed: Aug. 10, 2021. https://safecarnews.com/toyota-launches-map-based-adas-software-platform-virtual-horizon/
141. Sasai, S., Kitahara, I., Kameda, Y., Ohta, Y., Kanbara, M., Morales, Y., Ukita, N., Hagita, N., Ikeda, T., & Shinozawa, K. (2015). MR visualization of wheel trajectories of driving vehicle by seeing-through dashboard. In: IEEE International Symposium on Mixed and Augmented Reality Workshops, pp. 40–46
142. Marques, L., Vasconcelos, V., Pedreiras, P., & Almeida, L. (2011). A flexible dashboard panel for a small electric vehicle. In: 6th Iberian Conference on Information Systems and Technologies (CISTI 2011), pp. 1–4. IEEE
143. Wu, B.-F., Chen, Y.-H., & Yeh, C.-H. (2013). Driving behaviour-based event data recorder. *IET Intelligent Transport Systems, 8*(4), 361–367.
144. Collision Between a Sport Utility Vehicle Operating With Partial Driving Automation and a Crash Attenuator Mountain View, California. Accessed: Oct. 30, 2020 (2018). https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR2001.pdf
145. Bose, U. (2014). The black box solution to autonomous liability. *Wash. UL Rev., 92*, 1325.
146. Kohler, W.J., & Colbert-Taylor, A. (2014). Current law and potential legal issues pertaining to automated, autonomous and connected vehicles. Santa Clara Computer & High Tech. LJ 31, 99
147. Martinesco, A., Netto, M., Neto, A. M., & Etgens, V. H. (2019). A note on accidents involving autonomous vehicles: Interdependence of event data recorder, human-vehicle cooperation and legal aspects. *IFAC-PapersOnLine, 51*(34), 407–410.
148. Pinter, K., Szalay, Z., & Vida, G. (2020). Road accident reconstruction using on-board data, especially focusing on the applicability in case of autonomous vehicles. Periodica Polytechnica Transportation Engineering
149. Guo, H., Meamari, E., & Shen, C.-C. (2018). Blockchain-inspired event recording system for autonomous vehicles. In: 1st IEEE International Conference on Hot Information-Centric Networking (HotICN), pp. 218–222
150. Yao, Y., & Atkins, E. (2020). The smart black box: A value-driven high-bandwidth automotive event data recorder. IEEE Transactions on Intelligent Transportation Systems
151. Harris, D.M., & Harris, S.L. (2013). 3 - sequential logic design. In: Harris, D.M., Harris, S.L. (eds.) Digital Design and Computer Architecture (Second Edition), Second edition edn., pp. 108–171. Morgan Kaufmann, Boston. https://doi.org/10.1016/B978-0-12-394424-5.00003-3. https://www.sciencedirect.com/science/article/pii/B9780123944245000033
152. European Commission - Press release: Road safety: Commission welcomes agreement on new EU rules to help save lives. European Commission. Accessed: Feb. 8, 2021 (2019). https://ec.europa.eu/commission/presscorner/detail/en/IP_19_1793
153. UNECE: Working Party on Automated/Autonomous and Connected Vehicles (GRVA): EDR/DSSAD 1st session. EDR-DSSAD-01-06 Overview of EDR. Accessed: Feb. 8, 2021. (2019) https://wiki.unece.org/pages/viewpage.action?pageId=87621710
154. Ten Holter, C., Kunze, L., Pattinson, J.-A., Salvini, P., & Jirotka, M. (2022). Responsible innovation; responsible data. a case study in autonomous driving. Journal of Responsible Technology 11, 100038 https://doi.org/10.1016/j.jrt.2022.100038
155. Omeiza, D., Anjomshoae, S., Webb, H., Jirotka, M., & Kunze, L. (2022). From spoken thoughts to automated driving commentary: Predicting and explaining intelligent vehicles' actions. In: 2022 IEEE Intelligent Vehicles Symposium (IV), pp. 1040–1047. https://doi.org/10.1109/IV51971.2022.9827345
156. Panigutti, C., Hamon, R., Hupont, I., Llorca, D.F., Yela, D.F., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Garrido, J.S., & Gomez, E. (2023). The role of explainable ai in the context of the ai act. In: ACM Conference on Fairness, Accountability, and Transparency, pp. 1139–1150
157. Winfield, A.F.T., Maris, A., Salvini, P., & Jirotka, M. (2022). An Ethical Black Box for Social Robots: a draft Open Standard. arXiv. https://doi.org/10.48550/ARXIV.2205.06564. https://arxiv.org/abs/2205.06564
158. Ivanovic, B., & Pavone, M. (2019). The Trajectron: Probabilistic Multi-Agent rajectory Modeling With Dynamic Spatiotemporal Graphs. In: International Conference on Computer Vision (ICCV), pp. 2375–2384
159. Mohamed, A., Zhu, D., Vu, W., Elhoseiny, M., & Claudel, C. (2022). Social-implicit: Rethinking trajectory prediction evaluation and the effectiveness of implicit maximum likelihood estimation. In: 17th European Conference Computer Vision (ECCV), pp. 463–479. https://doi.org/10.1007/978-3-031-20047-2_27
160. Li, L., Pagnucco, M., & Song, Y. (2022). Graph-based spatial transformer with memory replay for multi-future pedestrian trajectory prediction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2221–2231.https://doi.org/10.1109/CVPR52688.2022.00227
161. Kothari, P., Kreiss, S., & Alahi, A. (2022). Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Transactions on Intelligent Transportation Systems, 23*(7), 7386–7400. https://doi.org/10.1109/TITS.2021.3069362
162. Park, S.H., Lee, G., Liang, P.P., & Morency, L.-P. (2020) Diverse and admissible trajectory forecasting through multimodal context understanding. In: European Conf on Computer Vision
163. Khandelwal, S., Qi, W., Singh, J., Hartnett, A., & Ramanan, D. (2020). What-if motion prediction for autonomous driving. arXiv preprint arXiv:2008.10587
164. Liang, M., Yang, B., Hu, R., Chen, Y., Liao, R., & Urtasun, R. (2020). Learning lane graph representations for motion forecasting. In: European Conf on Computer Vision. Springer
165. Saadatnejad, S., Bahari, M., Khorsandi, P., Saneian, M., Moosavi-Dezfooli, S.-M., & Alahi, A. (2022). Are socially-aware trajectory prediction models really socially-aware? Transportation Research Part C: Emerging Technologies 141, 103705 https://doi.org/10.1016/j.trc.2022.103705
166. Ortiz-Jimenez, G., Modas, A., Moosavi, S.-M., & Frossard, P. (2020). Hold me tight! influence of discriminative features on deep network boundaries. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems (NeurIPS), vol. 33. Curran Associates, Inc.,???. https://proceedings.neurips.cc/paper/2020/file/1ea97de85eb634d580161c603422437f-Paper.pdf
167. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F., & Rahwan, I. (2018). The moral machine experiment. Nature 563, 59–64 https://doi.org/10.1038/s41586-018-0637-6
168. Etienne, H. (2021). The dark side of the 'Moral Machine' and the fallacy of computational ethical decision-making for autonomous vehicles. Law, Innovation and Technology 13(1), 85–107 https://doi.org/10.1080/17579961.2021.1898310
169. German Ethics Commission: Automated and Connected Driving Report. Federal Ministry of Transport and Digital Infrastructure (2017)
170. Brandao, M. (2019). Age and gender bias in pedestrian detection algorithms. Workshop on Fairness Accountability Transparency and Ethics in Computer Vision at CVPR
171. Li, X., Chen, Z., Zhang, J. M., Sarro, F., Zhang, Y., & Liu, X. (2024). Bias behind the wheel: Fairness testing of autonomous driving systems. *ACM Trans. Softw. Eng. Methodol.* https://doi.org/10.1145/3702989

172. Wilson, B., Hoffman, J., & Morgenstern, J. (2019). Predictive inequity in object detection. Workshop on Fairness Accountability Transparency and Ethics in Computer Vision at CVPR
173. Fernández Llorca, D., Frau, P., Parra, I., Izquierdo, R., & Gómez, E. (2024). Attribute annotation and bias evaluation in visual datasets for autonomous driving. Journal of Big Data **11**(137) https://doi.org/10.1186/s40537-024-00976-9
174. Morando, M.M., Truong, L.T., & Vu, H.L. (2017). Investigating safety impacts of autonomous vehicles using traffic micro-simulation. In: Australasian Transport Research Forum 2017
175. Lu, Z., Happee, R., Cabrall, C.D.D., Kyriakidis, M., & de Winter, J.C.F. (2016). Human factors of transitions in automated driving: A general framework and literature survey. Transportation Research Part F: Traffic Psychology and Behaviour 43, 183–198 https://doi.org/10.1016/j.trf.2016.10.007
176. Carsten, O., & Martens, M. H. (2019). How can humans understand their automated cars? HMI principles, problems and solutions. *Cognition, Technology & Work, 21*(1), 3–20. https://doi.org/10.1007/s10111-018-0484-0
177. Cummings, M.L., & Ryan, J.C. (2014). Shared Authority Concerns in Automated Driving Applications. Working Paper
178. Martens, M.H., & van den Beukel, A.P. (2013). The road to automated driving: Dual mode and human factors considerations. In: 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), pp. 2262–2267. https://doi.org/10.1109/ITSC.2013.6728564
179. Bonnefon, J.-F., Cerny, D., Danaher, J., Devillier, N., Johansson, V., Kovacikova, T., Martens, M., Mladenovic, M., Palade, P., Reed, N., Santoni, F., Tsinorema, S., Wachter, S., & Zawieska, K. (2020). *Ethics of Connected and Automated Vehicles: Recommendations on Road Safety*. Fairness, Explainability and Responsibility: Privacy.
180. Prakken, H. (2017). On the problem of making autonomous vehicles conform to traffic law. *Artificial Intelligence and Law, 25*(3), 341–363. https://doi.org/10.1007/s10506-017-9210-0
181. 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, 23*(4), 777–789. https://doi.org/10.1007/s10676-021-09614-x
182. Bahari, M., Saadatnejad, S., Rahimi, A., Shaverdikondori, M., Shahidzadeh, M., Moosavi-Dezfooli, S.-M., & Alahi, A. (2021). Vehicle trajectory prediction works, but not everywhere. arXiv preprint arXiv:2112.03909
183. Saadatnejad, S., Bahari, M., Khorsandi, P., Saneian, M., Moosavi-Dezfooli, S.-M., & Alahi, A. (2021). Are socially-aware trajectory prediction models really socially-aware? arXiv preprint arXiv:2108.10879
184. Serrano, S.M., Llorca, D.F., Daza, I.G., & Sotelo, M.Á. (2022). Insertion of real agents behaviors in CARLA autonomous driving simulator. In: Proceedings of the 6th International Conference on Computer-Human Interaction Research and Applications, CHIRA, pp. 23–31. https://doi.org/10.5220/0011352400003323
185. Martín, S., Izquierdo, R., García Daza, I., Sotelo, M.A., & Fernández Llorca, D. (2023). Digital twin in virtual reality for human-vehicle interactions in the context of autonomous driving. In: IEEE Intelligent Transportation Systems Conference
186. Serrano, S.M., Llorca, D.F., Daza, I.G., & Sotelo, M.Á. (2023). Realistic pedestrian behaviour in the carla simulator using vr and mocap. In: Computer-Human Interaction Research and Applications, Communications in Computer and Information Science (CCIS), vol. 1882

Publisher's Note