European Commission

JRC SCIENCE FOR POLICY REPORT

Trustworthy Autonomous Vehicles

Assessment criteria for trustworthy AI in the autonomous driving domain

Fernández Llorca, David
Gómez, Emilia.

2021

Joint Research Centre

EUR 30942 EN

Contact Information
Name: David Fernández Llorca
Address: Joint Research Centre, Edificio Expo, c/Inca Garcilaso, 3, 41092 Seville (SPAIN)
Email: David.FERNANDEZ-LLORCA@ec.europa.eu
Tel.: +34 95448354

EU Science Hub
https://ec.europa.eu/jrc

JRC127051

EUR 30942 EN

PDF ISBN 978-92-76-46055-8 ISSN 1831-9424 doi:10.2760/120385

Luxembourg: Publications Office of the European Union, 2021

How to cite this report: Fernández Llorca, D., Gómez, E., Trustworthy Autonomous Vehicles, EUR 30942 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-46055-8, doi:10.2760/120385, JRC127051.

Contents

Acknowledgements

We would like to acknowledge colleagues from the Digital Economy Unit (B.6) of DG JRC, and more specifically colleagues from the HUMAINT team. We would also like to thank the following experts who specifically reviewed the report (in alphabetical order):

Bertelmann, Bernd (DG JUST)
Ciuffo, Biagio (DG JRC)
Kardasiadou, Zoi (DG JUST)
Kriston, Akos (DG JRC)
Kyriakou, Dimitrios (DG JRC)
Langfeldt, Owe (DG JUST)
Tolan, Songül (DG JRC)

Authors
Fernández Llorca, David
Gómez, Emilia

Executive Summary

This report aims to advance the discussion on those fundamental aspects to be considered in order to have trustworthy Artificial Intelligence (AI) systems in the Automated/Autonomous Vehicles (AVs) domain. Current advances in AVs would not be possible without advances in AI systems. In fact, the most complex tasks that AVs have to deal with to achieve high levels of automation (i.e. localization, scene understanding, planning, control and user interaction), are addressed by using multiple, complex interrelated AI systems. Therefore, when referring to trustworthy AI systems for AVs, it seems acceptable to extrapolate the concept to refer to it as trustworthy AVs (as presented in the title of this report).

The term *trustworthy* should not be interpreted in its literal sense, but as a global framework that includes multiple principles, requirements and criteria. These elements were established by the High Level Expert Group on Artificial Intelligence (AI HLEG) in the assessment list for trustworthy AI systems as a mean to maximise the benefits while minimising the risks. Indeed, the application context of AVs can certainly be considered high-risk, and their adoption involves addressing significant technical, political and societal challenges. However, AVs could bring substantial benefits, improving safety, mobility, and the environment.

The policy context in which this research has been carried out is characterised by two main pillars. First, the Coordinated Plan on AI (COM(2018) 795) which defines a set of joint actions between the Commission and Members States in the mobility chapter (i.e., make mobility smarter, safer and more sustainable through AI). The Commission is expected to set out implementing acts for technical specifications for automated and fully automated vehicles, including safety issues linked to the use of AI and cybersecurity ([1]). In addition, the Commission will adopt measures to facilitate trust and social acceptance of Cooperative, Connected and Automated Mobility by enhancing transparency, safety and explainability of technology. And Member States are encouraged to facilitate the deployment of trustworthy AI solutions in all modes of transport. On the other hand, within the Coordinated Plan on AI, the Proposal for a Regulation laying down harmonised rules on AI (AI Act) (COM (2021)206), establishes a set of requirements that AI systems must meet if they are operating in high-risks scenarios. Given that AI is by far the main enabler for AVs, the AI Act may indeed be relevant for developing a complementary sectorial regulatory framework for AVs. The second pillar is the Regulation (EU) 2019/2144 ([2]) which states that "*harmonised rules and technical requirements for automated vehicle systems should be adopted at Union level, and promoted at international level in the framework of the UNECE's World Forum for Harmonization of Vehicle Regulations (WP.29)*". It also establishes that "*the Commission shall adopt uniform procedures and technical specifications for the type-approval of automated and fully automated vehicles by means of implementing acts*".

The policy context is therefore very appropriate. There is a good opportunity to develop new harmonised procedures for the type-approval of AVs at EU level that can incorporate elements of the AI Act, in particular those identified as necessary for trustworthy systems. In a way, as it has been proposed for the case of AI ("a European approach to artificial intelligence") and taking into account the importance that AI systems have in AVs, one could speak of the beginning of "*the European approach to AVs*", in line with the recent strategy on Sustainable and Smart Mobility (COM(2020) 789).

Given the policy context described above, the key conclusions of our analysis can be described as follows. First of all, when we talk about vehicles with automated driving systems, there is a terminological problem that needs to be addressed. The SAE Levels of automation serve as a basis, although they suffer from a number of problems. The approach of using the terms *assisted*, *automated* and *autonomous* is perhaps the most appropriate. On the one hand it avoids the need to use additional words such as *partial*, *conditional*, *highly* or *fully*. It also avoids handling an equivalence between levels and their meaning. And finally, it allows for a simple clarification of the user's role and responsibility.

On the other hand, it is fundamental to advance in a clear taxonomy to categorise the Operational Design Domain (ODD). It seems reasonable to think that there will always be an operational domain associated to any driving automation system (even for SAE Level 5), which will depend on multiple factors such as the type of scenario, environmental conditions, and dynamic elements. Appropriate scenario-based specifications for the minimal risk conditions to be achieved by AVs after occurrence of a system failure are also needed.

The main enabler for assisted, automated or autonomous driving is AI. In an AV we can identify five main interrelated layers: (1) localisation, (2) scene understanding, (3) path planning, (4) control and (5) user interaction. AI plays a fundamental role in all of them. This raises the question of whether the requirements for a trustworthy system should be addressed from a holistic approach (i.e., trustworthy AVs), or from a layer-based approach (i.e., trustworthy AV localisation, trustworthy scene understanding, etc.). In our study we have opted for a holistic approach as it seems to be the most appropriate if we want to consider all the established criteria for a trustworthy AI.

([1]) As stated in the Coordinated Plan on Artificial Intelligence 2021 Review, new rules on automated vehicles, cybersecurity and software updates of vehicles will become applicable as part of the vehicle type approval and market surveillance legislation as from 6 July 2022.

([2]) Regulatory act for the purposes of the EU type-approval procedure laid down by Regulation (EU) 2018/858.

In the field of AVs when thinking about users (e.g., human-centric), it is necessary to consider a double dimension: internal users (i.e., drivers or passengers) and external road users (i.e., vulnerable road users, drivers of conventional vehicles and users of other AVs). This is a major challenge as it sometimes involves conflicting perspectives and interests that require compromise solutions.

A methodology has been proposed in which, on the one hand, the level of maturity, relevance and time horizon of each of the criteria established in the assessment list for a trustworthy AI are quantified (Fig. 5), and on the other hand, a qualitative analysis of the state of the art of each of the seven requirements is carried out. In what follows, the most relevant conclusions are presented for each of the key requirements.

— Human agency and oversight: Human agency for AVs is linked to the principle of human autonomy, affecting acceptance (e.g., disuse) and safety (misuse). New agency-oriented Human Machine Interfaces (HMIs) and external HMIs (eHMIs) are needed to ensure an adequate level of human agency. Efficient approaches to measure and calibrate the sense of agency are essential.

  Human oversight for AVs is exercised differently depending on the level of automation. It is also exercised to some extent by external road users, with the risk of abuse in the interaction knowing that AVs will stop in any case. For a proper interaction there must be mutual awareness between the AV and the agents with whom it interacts. How to effectively represent and communicate the operating status of the AV to users, including the request to intervene, is a key area of future research. Finally, oversight will require new skills both a priori and developed with exposure and use.

— Technical robustness and safety: These requirements are linked to the principle of harm prevention, with a strong impact on user acceptance. Attack resilience and security of AVs must be addressed from a heterogeneous, constantly updated approach, starting from security by design, including multiple defensive measures (e.g., cryptographic methods, intrusion and anomaly detection), countermeasures against adversarial attacks (e.g., redundancy, hardening against adversarial examples), fault-tolerant, fail-x, and self-healing methods, and user training.

  New innovative methods to assess the safety of AVs with respect to human drivers are needed. Conservative expectations on safety gains could accelerate the improvement of user acceptance. Small improvements can save lives, and too high expectations can delay the adoption, and thus the benefits, of the technology.

  Important steps have been taken in the design of new safety test procedures for automated driving functions, including simulation, physical test in proving grounds, and real-world test drive. However, there are still important limitations, such as the absence of real behaviours, limited variability, lack of scenarios to assess human agency and oversight, as well as transparency and fairness. New fallback strategies are needed to achieve minimum risk conditions, as well as testing procedures to assess their safety.

  Accuracy of AVs is a multi-dimensional problem, involving multiple metrics, levels, layers, use cases and scenarios. Defining holistic metrics and thresholds to assess AVs is a challenging research and policy-based problem to be addressed.

  Any substantial change in an AI-based component that may modify the overall behaviour of the AV must meet all requirements for robustness, safety and security, and may need to be retested.

— Privacy and data governance: New innovative approaches have to be implemented to ensure data protection without negatively affecting the safety of AVs, including agent-specific data anonymisation and de-identification techniques, while preserving relevant attributes of agents.

  Privacy by design (also linked to security and safety) will require the encryption of data, storage devices and vehicle-to-everything (V2X) communication channels, with a unique encryption key management system for each vehicle and including regular renewal of encryption keys.

  Consent to the processing of personal data in AVs should address two dimensions. For drivers and passengers, it should not pose any safety risk, and should include the exchange of data with other vehicles and infrastructures. For external road users, consent can only be obtained indirectly, although it can be avoided if data are processed in real time or if data de-identification is properly implemented.

— Transparency: Traceability is already a challenge for modern conventional vehicles, so its complexity for AVs is more than remarkable. The effective integration of components of data-driven AI systems as traceable artefacts is still an open research question.

  New strategies for intelligent data logging must be developed to cope with the demanding requirements (bandwidth and storage capacity) of continuous data logging for AVs.

  New explainable models and methods should be developed, focusing on explanations to internal and external road users, i.e. new research related to explainable human-vehicle interaction through new

HMI and eHMI. Explainability as a requirement for vehicle type-approval frameworks will enhance the assessment of safety, human agency and oversight, and transparency, but will require new test procedures, methods and metrics.

New effective ways of communicating to passengers and external road users that they are interacting with an AV must be established, as well as new ways of communicating risks.

— Diversity, non-discrimination and fairness: To avoid discrimination in decision making, AVs must avoid any kind of estimation based on potential social values of some groups over others (e.g., dilemmas) and must be designed to maintain the same level of safety for all road users. To this end, AVs may react differently to correct safety inequalities resulting from different road users behaviours, so new real-time predictive perception and path planning systems are needed to model the behaviour of different road users and react accordingly.

  Further efforts are needed to identify possible sources of discrimination in state-of-the-art perception systems for detecting external road users according to different inequity attributes such as sex, age, skin tone, group behaviour, type of vehicle, colour, etc.

  Unfair bias may also be present at the user-vehicle interaction layer. Accessible and adaptable HMIs should be designed, which is a challenge considering that AVs have the potential to extend mobility to new users.

  AVs opens up new autonomous mobility systems, services and products. Any service provision approach that may discriminate against users should be avoided.

  It is necessary for policymakers to establish a clear taxonomy of stakeholders, modulating the direction (positive or negative) and the weight of the impact that the adoption of AVs implies for each of them.

— Societal and environmental well-being: Understanding and estimating the impact of AVs on the environment and society is a highly multidimensional and complex problem, involving many disruptive factors, for which we can only make predictions based on as yet uncertain assumptions. New approaches and studies are needed to provide more accurate estimates, with less uncertainty. Policymakers must steer and monitor the adoption process to tip the balance towards a positive impact.

  Automated vehicles (up to SAE Level 3) will not have a negative impact on jobs, but new skills for backup drivers will be needed. For autonomous vehicles, as no drivers are needed, the expected impact on work and skills is likely to be negative, but partially mitigated by the need of non-driving tasks less susceptible to automation and new jobs and skills brought by transport automation.

  AVs opens up the possibility to use travel time for work-related activities, leading to higher productivity or a reduction of time at the workplace as commuting time could be considered as working time. In the coming years we will see new approaches to transform the interiors of AVs into places to work, which is a challenge in shared mobility scenarios.

— Accountability: As a safety-critical application, AVs must be audited by independent external auditors. Establishing the minimum requirements for third parties to audit systems without compromising intellectual and industrial property is a major challenge. The same requirements and expertise needed to audit AVs would be necessary for victims or insurers to claim for liability in accidents involving AVs, which would be very complex and costly. Shifting the burden of proof to the manufacturer of AVs would make these systems more victim friendly. Considerable harmonization efforts and major updates of existing national product liability, traffic liability and fault-based liability frameworks are needed, including the Product Liability Directive and the Motor Insurance Directive.

  The adoption of AVs will entail new risks, including those that are unknown at the time of production and can only emerge after market launch. Policymakers should define new balanced and innovative frameworks to accommodate insurance and liability costs between consumers and injured parties on the one hand, and AVs providers on the other.

The direct application of all requirements for trustworthy AI in the context of AVs involves addressing a multitude of areas of different nature, some of them still at a very early stage of technological maturity. The interrelationship and dependency between the various requirements and criteria suggests that holistic approaches are the most appropriate. However, more specific approaches, tailored to each requirement or a subset of them, may be a more suitable first step. When implementing the above requirements, tensions may arise between them, which may lead to inevitable trade-offs that should be explicitly acknowledged and evaluated in terms of their risk to ethical principles, including fundamental rights.

If future harmonised efforts for vehicle type-approval for AVs and regulatory frameworks for autonomous mobility at EU level are designed to be trustworthy, we humbly hope that they will benefit from the conclusions derived from this report.

1 Introduction

1.1 What is trustworthy AI?

The meaning of Trustworthy AI can be derived from the report of the AI HLEG on *Ethical Guidelines for Trustworthy AI*. *Trustworthy* should not be interpreted here in its literal sense, but rather as a **comprehensive framework** that includes multiple principles, requirements and criteria. Trustworthy AI systems are *human-centric and rest on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom*. Trustworthiness is conceived as a mean to *maximise the benefits of AI systems while at the same time preventing and minimising their risks* (AI HLEG, 2019).

Thus, trustworthy AI systems are grounded on three main components. First, they have to be **lawful**, complying with all applicable laws and regulations. Second, they have to be **ethical**, ensuring adherence to fundamental principles and values. And third, they have to be **robust**, both from a technical and social perspective. As will be seen in the next chapter, the last two components can be developed into four ethical principles based on fundamental rights, and seven key requirements. Each requirement, in turn, is composed of multiple criteria (AI HLEG, 2020).

1.2 The need of trustworthy systems: perspectives from society and policy making

As stated in 2020 Commission Communication "Sustainable and Smart Mobility Strategy" (COM(2020) 789), mobility and transport matters to us all. It is an essential enabler of our economic and social life, bringing many benefits to society. But mobility also involves serious costs such as greenhouse gas emissions, pollution, congestion, accidents and road crashes. These costs can be significantly reduced by autonomous (fully/highly automated or driverless) vehicles (Fagnant and Kockelman, 2015). As described in 2018 Commission Communication "On the road to automated mobility" (COM(2018) 283), Autonomous Vehicles (AVs) could enable new mobility services and car sharing schemes to respond to the increasing demand for mobility of people and goods, could significantly improve road safety as human factors (errors, distractions, violations of the traffic rules) play a key role in most accidents, could bring mobility to those who cannot drive themselves, could accelerate vehicle electrification, and could free up urban public spaces currently used for parking.

However, AVs also involves multiple important challenges from at least three dimensions. First, technical challenges such as robust localization and scene understanding, complex interactions with non-automated vehicles, with Vulnerable Road Users (VRUs) and with vehicle occupants, the need to adapt the infrastructure, robust path planning and control, and all in an extremely complex application context with virtually unlimited variability. Second, challenges in policy making, such as assessing the impact of new risks (e.g., those related to misuse due to overtrust and overreliance), the need for revision of traffic rules, balance between the need for data and privacy, responsible and safe testing on open roads, new metrics, benchmarks and procedures to ensure the net positive effect on road safety, and a clear framework for dealing with liability when an AV is involved in an accident. Last, but not least, societal challenges that include sustainability, user acceptance and the protection of fundamental rights as the main factors on which all other technical and political components can be based.

Furthermore, although a clear objective definition of high-risk domains for automated systems is not yet available, the inherently high-risk nature of the AVs context can be confidently assumed. The operation in public spaces potentially endangers the users (occupants) and the public (other road users), affecting unknown and not identifiable persons without prior consent. They can cause severe physical harm, even death, as well as property damage. The likelihood of harm or damage occurring will depend to a large extent on the progress made in addressing the challenges in the three dimensions described above.

In this complex high-risk domain and ecosystem of technical, political and societal challenges, the importance of trustworthiness, as defined by the High Level Expert Group on Artificial Intelligence (AI HLEG) (AI HLEG, 2020), is intuitively very clear. Indeed, AVs can be defined as a set of multiple, complex interrelated AI-based systems, embodied in the form of a car. Their behaviours depend on multiple Artificial Intelligence (AI) systems, each dealing with problems of a different nature. Therefore, all the ethical principles, key requirements and assessment criteria of a trustworthy AI can (and must) necessarily be applied to the specific context of AVs.

1.3 Regulatory and policy context

EU regulation around vehicle automation is strongly linked to the regulation for the approval of motor vehicles (and their trailers, and of systems, components and separate technical units intended for such vehicles). The first approximation of the laws of the Member States relating to the type-approval of motor vehicles took place in 1970 with the Council Directive 70/156/EEC, which was substantially amended several times, until it was finally repealed in 2007 by Directive 2007/46/EC.

Article 20 of Directive 2007/46/EC allowed Member States to approve new vehicle automation technologies not foreseen by EU rules through an EU exemption granted on the basis of a national ad-hoc safety assessment. The vehicle could then be placed on the EU market like any other EU approved vehicle. A similar approach was included in 2018 in the Regulation (EU) 2018/858 (which repealed Directive 2007/46/EC), in Article 39.

While this approach ensured that technological progress was not blocked, it left too many unknowns in terms of specifications and requirements. Thus, in April 2016, the EU Member States signed the *Declaration of Amsterdam on Cooperation in the field of connected and automated driving* (EU Member States, 2016). The Declaration established shared objectives, a joint agenda, and proposed actions for the Commission, including the development of a coordinated European strategy and the adaptation of the EU regulatory framework.

In November 2016, the Commission published the Communication on "Cooperative Intelligent Transport Systems" (COM(2016) 766). In October 2017 the *High Level Group on the Competitiveness and Sustainable Growth of the Automotive Industry in the European Union* (GEAR 2030) published a report with a number of recommendations to ensure a shared European strategy on automated and connected vehicles. And in May 2018, the Commission published the Communication "On the road to automated mobility: An EU strategy for mobility of the future" (COM(2018) 283), in which the Commission presented its vision ([3]), and stated its commitment, firstly, to work with Member States on guidelines to ensure a harmonised approach for national ad-hoc vehicle safety assessments of automated vehicles and, secondly, to work with Member States and stakeholders on a new approach for vehicle safety certification for automated vehicles.

As a regulatory act for the purposes of the EU type-approval procedure laid down by Regulation (EU) 2018/858, Regulation (EU) 2019/2144 introduced for the first time definitions and specific requirements relating to automated and fully automated vehicles. In paragraph 23, it is stated that "*harmonised rules and technical requirements for automated vehicle systems, including those regarding verifiable safety assurance for decision-making by automated vehicles, should be adopted at Union level, while respecting the principle of technological neutrality, and promoted at international level in the framework of the UNECE's World Forum for Harmonization of Vehicle Regulations (WP.29)*". In addition, in Chapter III, Article 11, six specific requirements are introduced (from (a) to (f)), and in paragraph 2 it is stated that "*The Commission shall by means of implementing acts adopt provisions concerning uniform procedures and technical specifications for the systems and other items listed in points (a) to (f) of paragraph 1 of this Article, and for the type-approval of automated and fully automated vehicles with regard to those systems and other items in order to ensure the safe operation of automated and fully automated vehicles on public roads*".

The UNECE's WP.29, and more specifically, the subsidiary Working Party (Groupe de Rapporteurs - GR) on Automated/Autonomous and Connected Vehicles (GRVA), has been working on international regulations on AVs, including *UN Regulation on Advanced Emergency Braking Systems* (UNECE WP.29 GRVA, 2020b), *UN Regulation on Cyber Security and Cyber Security Management System* (UNECE WP.29 GRVA, 2021a), *UN Regulation on Software Update and Software Update Management System* (UNECE WP.29 GRVA, 2021b), and the UN Regulation on Automated Lane Keeping Systems (UNECE WP.29 GRVA, 2021c). These texts have legal effects in the EU under international public law.

From an international perspective it is also worth mentioning the current work of the Technical Committee (TC) on Road traffic safety management systems of the International Organization for Standardization (ISO/TC 241), especially the forthcoming standard on *Road Traffic Safety (RTS) - Guidance on safety ethical considerations for autonomous vehicles* (ISO39003). And the work of the *Focus Group on AI for autonomous and assisted driving (FG-AI4AD)* of the ITU Telecommunication Standardization Sector (ITU-T FG-AI4AD, 2019).

Recently, in December 2020, the Commission presented the Communication "Sustainable and Smart Mobility Strategy - putting European transport on track for the future" (COM(2020) 789), together with a comprehensive Staff Working Document (SWD(2020) 331). Up to ten flagship areas are identified with an action plan for the next years, with flagship area 6 focusing on *Making Connected and Automated Multimodal Mobility a Reality*. In the SWD it is stated the following:

— (Paragraph 617): *An aligned and holistic European approach to global questions on automated driving, including traffic rules, testing or liability aspects, would be desirable from a single market perspective and for promoting European industrial global leadership, as well as in terms of materialization of road safety benefits*.

— (Paragraph 618): *For the time being, the legal and policy framework defining links between vehicles and traffic management, between public and privately owned data, and between collective and individual transport, are not sufficiently developed. There is no coordination mechanism at the EU level that would help ensure consistency of the deployment and management of ITS and CCAM across Europe. There is no coherent way of implementing a type-approval for connected and automated vehicles*.

These statements are important shortcomings that have to be addressed in the short term. But they can also be considered as a good opportunity to define **the European approach to AVs**. Taking into account the

([3]) The so-called Vision Zero aims for no road fatalities on European roads by 2050.

importance of AI systems for AVs, it seems reasonable to think that the same requirements that have served as a basis for developing the new regulatory framework on AI by the Commission (the AI Act, COM(2021) 206) may be relevant for developing a complementary sectorial regulatory framework to AVs. This report is intended as a humble contribution in that direction.

1.4 Goals and audience

The main goal of this report is to advance towards a general framework to assess trustworthiness of AI-based systems for AVs. The current state, research gaps, challenges, implementation and relevance, with respect to all the criteria included in the assessment list (ALTAI) developed by the High Level Expert Group on Artificial Intelligence (AI HLEG), are in this report particularized and discussed in detail for the the AVs domain. To the best of our knowledge, this is the first attempt to translate the seven requirements that AI systems should meet in order to be trustworthy, as defined in the Ethics Guidelines of the AI HLEG (AI HLEG, 2019), to the context of AVs.

In order to carry out the adaptation of the criteria, it is first necessary to address fundamental aspects of AVs, including the levels of automation, terminology, and main tasks in which AI plays an important role. In the development of each key requirement, the aim is to present the main problems, the state of scientific and technological maturity, difficulties in implementing the criteria, as well as the main research, technical and policy gaps and challenges ahead.

This report is targeted to policy makers, industry practitioners and researchers, not necessarily specialized in automated or autonomous driving systems, interested in how the different requirements for trustworthy AI apply to those systems, and what are the related research, technical and policy challenges that will need to be be addressed in the coming future.

1.5 Structure of the report

The structure of this report is as follows:

- Section 2 summarizes the ethical principles, key requirements and assessment criteria for trustworthy AI systems, adapted to the field of AVs. A specific code or identifier is assigned to each of them.
- Section 3 describes the levels of automation and discusses the current limitations when referring to automated or autonomous vehicles. Specific terminology is proposed, describing most of the terms commonly used to refer to this technology, and relating them to the levels of automation.
- Section 4 summarizes the main user-oriented and potentially automated driving tasks, the five major technological levels of layers of AVs, analyses multi-user issues when referring to humans and AVs, and discusses the needs for testing in real traffic conditions.
- Section 5 provides the assessment of the key requirements and criteria for trustworthy AI systems in the autonomous driving domain, considering that AI is the main enabler for autonomous driving. First, the relevance and time horizon of each of the criteria is qualitatively illustrated. Subsequently, each criterion is analysed independently and in detail, establishing in each case the dependence and interrelationship between them.
- Section 6 finally provides a summary with the main conclusions, general challenges ahead and future research gaps.

2 Ethical principles, key requirements and assessment criteria

Following the launch of its AI Strategy in April 2018 (European Commission, 2018), the European Commission set up a group of experts to advice for its implementation (AI HLEG). The first document generated by this group (AI HLEG, 2019) defined the foundations of Trustworthy AI by adhering to four ethical **principles** (EP) based on fundamental rights (see Table 1). By drawing similarities between the concept of an AI system and that of an AV, we can easily adapt these principles (and requirements) to the autonomous driving domain.

1. **Respect for human autonomy** (EP1): humans interacting with AVs (whether they are vehicle users or external road users) must be able to maintain full self-determination over themselves. AI systems of AVs should not subordinate, coerce, deceive, manipulate, condition or herd humans (e.g., do not move them to unwanted destinations, do not comply with stop requests, etc.). Instead, they should be designed to augment, complement and empower human driving skills and mobility (e.g., extending mobility to vulnerable groups). Interactions between humans and AVs should follow human-centric design principles, securing human oversight of driving automation systems in AVs.

2. **Prevention of harm** (EP2): AVs should neither cause nor exacerbate harm or otherwise adversely affect human beings. This entails the protection of human dignity as well as physical, and even mental, integrity. AVs and the road environments in which they operate must be safe and secure. AVs must be technically robust and it should be ensured that they are not open to malicious use. Vulnerable users (both in-vehicle and external road users) should receive greater attention and be considered in the development, deployment and use of AI systems of AVs.

3. **Fairness** (EP3): the development, deployment and use of AVs must be fair, ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation. The use of AVs should never lead to people being deceived or unjustifiably impaired in their freedom of choice. Fairness entails the ability to contest and seek effective redress against decisions made by AVs and by the humans operating them. In order to do so, the entity accountable of the AV decisions must be identifiable, and the decision-making processes (e.g., local path planning) should be explainable.

4. **Explainability**([4]) (EP4): it is crucial for building and maintaining users' trust in AVs. This means that driving automation systems need to be transparent, the capabilities and purpose of AI systems that enable vehicle automation must be openly communicated, and AV decisions - to the extent possible - explainable to those directly and indirectly affected. Without such information, the decisions and behaviour of the AVs cannot be duly contested. Cases in which an explanation is not possible (i.e., "black box" algorithms) require additional measures (e.g. traceability, auditability and transparent communication on system capabilities).

Table 1: Ethical Principles for a Trustworthy AI.

Code	Principles
EP1	Respect for human autonomy
EP2	Prevention of harm
EP3	Fairness
EP4	Explainability

All of these principles must be addressed (within existing technological limits) in order to ensure that AI systems of AVs are developed, deployed and used in a trustworthy manner. Building on these principles, the AI HLEG proposed a set of seven key **requirements** (KR) to offer guidance on the implementation and realization of Trustworthy AI, as depicted in Table 2.

These requirements have to be continuously evaluated and addressed throughout the AI system's life cycle. Their implementation and relevance strongly depends on the specific application. In order to facilitate the implementation of the key requirements, the AI HLEG finally proposed the assessment list for trustworthy AI (ALTAI) (AI HLEG, 2020) which translates the ethics guidelines into an accessible and dynamic checklist. Below, in Tables 3-9, we provide a summary of the **criteria**, with an assigned code (CR) that will allow us to identify each criterion in subsequent sections. The key requirements can be summarized and adapted to the AVs domain as follows:

([4]) The original text uses *explicability*. We use *explainability*, which shares the same meaning, but is more commonly used and accepted in the field of AI.

Table 2: Key Requirements for a Trustworthy AI.

Code	Requirements
KR1	Human agency and oversight
KR2	Technical robustness and safety
KR3	Privacy and data governance
KR4	Transparency
KR5	Diversity, non-discrimination and fairness
KR6	Societal and environmental well-being
KR7	Accountability

1. **Human agency and oversight** (KR1): AVs should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy, supporting the user's agency and foster fundamental rights, and allow for human oversight.

   Users should be able to make informed autonomous decisions with knowledge and tools to comprehend and interact with AVs to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system. Human oversight helps ensuring that an AV does not undermine human autonomy or causes other adverse effects.

Table 3: Assessment Criteria for a Trustworthy AI: key requirement 1.

Req.	Code	Criteria
KR1		**Human agency and autonomy**
	CR1.1	Affects humans or society.
	CR1.2	Confusion as to whether the interaction is with a human or an AI
	CR1.3	Overreliance
	CR1.4	Unintended and undesirable interference with end-user decision-making
	CR1.5	Simulation of social interaction
	CR1.6	Risk of attachment, addiction and user behaviour manipulation
		Human oversight
	CR1.7	Self-learning or autonomous / Human-in-the-Loop / Human-on-the-Loop / Human-in-Command
	CR1.8	Training on how to exercise oversight
	CR1.9	Detection and response mechanisms for undesirable adverse effects
	CR1.10	Stop button
	CR1.11	Oversight and control of the self-learning or autonomous nature of the AI system

2. **Technical robustness and safety** (KR2): technical robustness requires that AVs be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm to living beings or the environment. This should also apply to potential changes in their operating environment or the presence of other agents (human and artificial) that may interact with the system in an adversarial manner. Therefore, they should be resilient to attacks and secure. AVs should have safeguards that enable a fallback plan in case of problems. A high level of accuracy is especially crucial, as AVs can directly affect human lives. It is critical that the results of AI systems of AVs are reproducible, as well as reliable. In addition, the physical integrity of humans should be ensured, and the mental integrity should be considered.

3. **Privacy and data governance** (KR3): AI systems of AVs must guarantee privacy and data protection throughout a system's entire life cycle, including information initially provided by the user, as well as the information generated about the user over the course of their interaction with the system. Digital records of human behaviour may allow AI systems to infer private information about individuals. To allow individuals to trust the data gathering process, it must be ensured that data collected about them will not be used to unlawfully or unfairly discriminate against them. In addition, the integrity of the data must be ensured and data protocols governing data access must be put in place.

4. **Transparency** (KR4): the data sets and the processes that yield the AV's decision (data gathering and labelling, algorithms, decisions, etc.) should be documented to allow for traceability and an increase in transparency. This enables identification of the reasons why an AV-decision was erroneous which in

Table 4: Assessment Criteria for a Trustworthy AI: key requirement 2

Req.	Code	Criteria
KR2	**Resilience to Attack and Security**	
	CR2.1	Effects on human safety due to faults, defects, outages, attacks, misuse, inappropriate or malicious use.
	CR2.2	Confusion as to whether the interaction is with a human or an AI
	CR2.3	Cybersecurity certification and security standards
	CR2.4	Exposure to cyberattacks
	CR2.5	Integrity, robustness and security against attacks
	CR2.6	Red teaming / Penetration testing
	CR2.7	Security coverage and updates
	General Safety	
	CR2.8	Risks, risk metrics and risk levels
	CR2.9	Design faults, technical faults, environmental threats
	CR2.10	Stable and reliable behaviour
	CR2.11	Fault tolerance
	CR2.12	Review of technical robustness and safety
	Accuracy	
	CR2.13	Consequences of low level accuracy
	CR2.14	Quality of the training data (up-to-date, complete and representative)
	CR2.15	Monitoring and documentation of system accuracy
	CR2.16	Invalid training data and assumptions
	CR2.17	Communication of expected level of accuracy
	Reliability, Fall-back plans and Reproducibility	
	CR2.18	Consequences of low reliability and reproducibility
	CR2.19	Verification and validation of reliability and reproducibility
	CR2.20	Tested fail-safe fallback plans
	CR2.21	Handling low confidence scores
	CR2.22	Online continuous learning

Table 5: Assessment Criteria for a Trustworthy AI: key requirement 3.

Req.	Code	Criteria
KR3	**Privacy**	
	CR3.1	Privacy, integrity and data protection
	CR3.2	Flagging privacy issues
	Data Governance	
	CR3.3	Use of personal data
	CR3.4	General Data Protection Regulation (GDPR) measures
	CR3.5	Privacy and data protection implications for Non-personal data
	CR3.6	Alignment with standards and protocols for data management and governance

turn could help to prevent future mistakes. Traceability facilitates auditability, as well as explainability. Technical explainability requires that the decisions made by an AV can be understood and traced by humans, which should be possible to demand as AVs can have a significant impact on people's lives. Trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability). In addition, the AV's capabilities (level of accuracy) and limitations should be appropriately communicated to practitioners or end-users, who have the right to be informed that they are interacting with an AI system.

5. **Diversity, non-discrimination and fairness** (KR5): data sets used by AVs may suffer from the inclusion of bias, incompleteness and bad governance models. The continuation of such biases could lead to unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalisation. Identifiable and discriminatory bias should be removed in the collection phase where possible. The way in which AI systems are developed (e.g. algorithms' programming) may also suffer from unfair bias. This could be counteracted by putting in place oversight processes to analyse and address the system's purpose, constraints, requirements and decisions in a clear and transparent manner.

Table 6: Assessment Criteria for a Trustworthy AI: key requirement 4.

Req.	Code	Criteria
KR4		**Traceability**
	CR4.1	Measures to address traceability
		Explainability
	CR4.2	Explaining decisions to the users
	CR4.3	Continuous survey on users' understanding of decisions
		Communication
	CR4.4	Communicating the users that they are interacting with an AI system
	CR4.5	Inform users about purpose, criteria and limitations

AVs should be human-centric and designed in a way that allows all people to use them, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for persons with disabilities, which are present in all societal groups, is of particular importance. In addition, it is advisable to consult stakeholders who may directly or indirectly be affected by the system throughout its life cycle.

Table 7: Assessment Criteria for a Trustworthy AI: key requirement 5.

Req.	Code	Criteria
KR5		**Avoidance of Unfair Bias**
	CR5.1	Unfair bias on data or algorithm design
	CR5.2	Diversity of end-users in the data
	CR5.3	Educational and awareness initiatives to avoid injecting bias
	CR5.4	Flagging bias, discrimination or poor performance issues
	CR5.5	Appropriate fairness definition
		Accessibility and Universal Design
	CR5.6	Correspondence to variety of preferences and abilities in society
	CR5.7	Usability of the user interface by people with special needs
	CR5.8	Universal design principles
	CR5.9	Impact on end-users
		Stakeholder Participation
	CR5.10	Stakeholder participation in the design and development

6. **Societal and environmental well-being** (KR6): sustainability and ecological responsibility of AVs should be encouraged. Ideally, AI systems should be used to benefit all human beings, including future generations. Measures securing the environmental friendliness of AVs' entire supply chain should be encouraged. The social impact of AVs (e.g., social agency, relationships, attachment , skills, physical and mental well-being) must be carefully monitored and considered. The impact should also be assessed from a societal perspective, taking into account its effect on institutions, democracy and society at large.

Table 8: Assessment Criteria for a Trustworthy AI: key requirement 6.

Req.	Code	Criteria
KR6		**Environmental Well-being**
	CR6.1	Negative impacts on the environment
	CR6.2	Environmental impact evaluation (development, deployment and use)
		Impact on Work and Skills
	CR6.3	Impacts on human work
	CR6.4	Consideration of impacted workers and their representatives
	CR6.5	Measures to ensure understanding of the impact on human work
	CR6.6	Risk of deskilling
	CR6.7	Need of new (digital) skills
		Impact on Society at large or Democracy
	CR6.8	Negative impact on society at large or democracy

7. **Accountability** (KR7): mechanisms must be put in place to ensure responsibility and accountability for AVs and their outcomes, both before and after their development, deployment and use. Auditability entails the enablement of the assessment of algorithms, data and design processes. Evaluation by internal and external auditors, and the availability of such evaluation reports, can contribute to the trustworthiness of the technology. In applications affecting fundamental rights, including safety-critical applications such as AVs, they should be able to be independently audited. Both the ability to report on actions or decisions that contribute to a certain system outcome, and to respond to the consequences of such an outcome, must be ensured. When unjust adverse impact occurs, accessible mechanisms should be foreseen that ensure adequate redress. Particular attention should be paid to vulnerable persons or groups.

   When implementing the above requirements, tensions may arise between them, which may lead to inevitable trade-offs that should be explicitly acknowledged and evaluated in terms of their risk to ethical principles, including fundamental rights.

Table 9: Assessment Criteria for a Trustworthy AI: key requirement 7.

Req.	Code	Criteria
KR7		**Auditability**
	CR7.1	Auditability mechanisms
	CR7.2	Third parties audit
		Risk Management
	CR7.3	External guidance to oversee ethical concerns and accountability measures
	CR7.4	Risk training and applicable legal framework
	CR7.5	Ethics review board
	CR7.6	Adherence to the ALTAI ([1])
	CR7.7	Third party process to report vulnerabilities, risks or biases
	CR7.8	Redress by design

([1]) We present this specific criterion and respect the original order. However, it should be carefully interpreted to avoid any sense of recursion.

3 From driving automation systems to autonomous vehicles

3.1 Levels of automation

In order to provide a clear definition of what an AV represents, it is necessary to start from the concept of "levels of automation". First, "automation" refers to the *"full or partial replacement of a function previously performed by the human operator"* (Parasuraman et al., 2000). Therefore, automation can vary across a continuum of levels, from fully manual to full automation. Multiple taxonomies to address the different levels of automation for automated systems in general have been proposed over the years from the academia (Sheridan and Verplank, 1978), (Parasuraman et al., 1989), (Parasuraman et al., 2000). For the specific context of automation of on-road vehicles, the National Highway Traffic Safety Administration issued a preliminary statement in 2013 (NHTSA, 2013) proposing up to 5 levels of driving automation ranging from "no driving automation" (Level 0) to "full driving automation" (Level 4). Another regulatory body, the SAE International (formerly Society of Automotive Engineers) increased the number of levels up to 6 (from 0 to 5) issuing the first version of the standard in 2014 (SAE J3016), which was revised in 2018 (SAE International, 2018) and in 2021 (SAE International, 2021). For both standards, we can find a very similar description from levels 0 to 3. SAE conveniently divided the "full automated" NHTSA Level 4 into two levels, 4 and 5, with the range of operating conditions being the main difference between them. The SAE definition of the levels of automation is the most prominent and accepted standard nowadays.

To better understand the meaning of the different levels of vehicle automation, we need to address the definition of some important terms presented by the SAE standard (taken and elaborated from (SAE International, 2021)):

- **Dynamic Driving Task (DDT)**: all of the real-time operational and tactical functions required to operate a vehicle in on-road traffic, excluding the strategic (high-level) functions such as trip scheduling and selection of destinations and waypoints. DDT can be divided in at least two main subtasks:
  - **Object and Event Detection and Response (OEDR)**: the subtasks of the DDT that include monitoring the driving environment and executing an appropriate response to such objects and events (e.g., vehicles stopped or changing lanes, pedestrians attempting to cross or crossing, traffic signals status, etc.).
  - **Lateral and Longitudinal Vehicle Motion Control**: the DDT subtask comprising the activities necessary for the real-time, sustained regulation of the x- and y-axes of vehicle motion.
- **Operational Design Domain (ODD)**: operating conditions under which a given *driving automation system* or feature thereof is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain traffic or roadway characteristics.
- **Driving Automation System**: the hardware and software that are collectively capable of performing part or all of the DDT on a sustained basis; this term is used generically to describe any system capable of Level 1-5 driving automation (from partial to full automation).
- **Automatic Driving System (ADS)**: the hardware and software that are collectively capable of performing the entire DDT on a sustained basis, regardless of whether it is limited to a specific operational design domain (ODD); this term is used specifically to describe a Level 3, 4, or 5 *driving automation system*.
- **DDT Fallback**: the response by the user to either perform the DDT or achieve a minimal risk condition after occurrence of a DDT performance-relevant system failure(s) or upon operational design domain (ODD) exit, or the response by an ADS to achieve minimal risk condition, given the same circumstances.
- **Minimal Risk Condition**: a condition to which a user or an ADS may bring a vehicle after performing the DDT fallback in order to reduce the risk of a crash when a given trip cannot or should not be completed. Note that this definition does not establish what the "acceptable" reduced risk of a crash is. This is an open and important question as the success of the DDT fallback performed by the ADS will define its consideration as Level 4 or 3.
- **Request to Intervene**: a notification by an ADS to a user indicating that he/she should promptly perform the DDT fallback, which may entail resuming manual operation of the vehicle (i.e., becoming a driver again), or achieving a minimal risk condition if the vehicle is not drivable. This concept is also referred to as *Take-Over Request* (TOR).

With the above definitions, we can now describe the six levels of vehicle automation of the SAE standard (taken and elaborated from (SAE International, 2021)):

— **Level 0 (No Driving Automation)**: the performance by the driver of the entire DDT, even when enhanced by active safety systems. The driver is responsible of the DDT and the DDT fallback.

— **Level 1 (Driver Assistance)**: the sustained and (limited) ODD-specific execution by a *driving automation system* of either the lateral *or* the longitudinal vehicle motion control subtask of the DDT (but not both simultaneously) with the expectation that the driver performs the remainder of the DDT (including the DDT fallback). Although it is assumed that the system executes the part of the OEDR subtask that allows the execution of vehicle motion control, the driver is the main responsible of the OEDR subtask.

— **Level 2 (Partial Driving Automation)**: the sustained and (limited) ODD-specific execution by a *driving automation system* of *both* the lateral and longitudinal vehicle motion control subtasks of the DDT (it is assumed that the system executes the part of the OEDR subtask that allows the execution of vehicle motion control) with the expectation that the driver completes the OEDR subtask and supervises the *driving automation system* (including the DDT fallback).

— **Level 3 (Conditional Driving Automation)**: the sustained and (limited) ODD-specific performance by an ADS of the *entire DDT* (vehicle motion control and OEDR) with the expectation that the DDT fallback-ready user is receptive to ADS-issued requests to intervene, as well as to DDT performance-relevant system failures in other vehicle systems, and will respond appropriately. The driver must be ready to perform the DDT fallback in a timely manner, being receptive to a request to intervene or to DDT performance-relevant system failures ([5]), ([6]).

— **Level 4 (High Driving Automation)**: the sustained and (limited) ODD-specific performance by an ADS of the entire DDT and DDT fallback without any expectation that a user will respond to a request to intervene. The automated DDT fallback and minimal risk condition achievement capability is the primary difference between level 4 and 3 ([7]).

— **Level 5 (Full Driving Automation)**: the sustained and unconditional (i.e., not ODD-specific) performance by an ADS of the entire DDT and DDT fallback without any expectation that a user will respond to a request to intervene ([8]).

As stated in the standard, the levels definitions can be used to describe the full range of driving automation "features" equipped on motor vehicles. They do not specifically refer to the vehicle itself (which is a common misconception) but to features or applications within a particular ODD. A given *driving automation system* may have multiple features with different levels of automation each. For example, a vehicle can have a Level 5 ADS automatic parking feature and a Level 4 ADS highway pilot feature. It is important to also note that the definition of DDT does not include strategic or high-level functions (following the driver efforts -strategic, tactical and operational - proposed by (Michon, 1985)) such as planning the trip, when, where and how to travel, the route to be taken, etc.

The SAE taxonomy has some well-known problems (Inagaki and Sheridan, 2019) such as the undefined requirements for safe transition from one level of automation to another (e.g., from Level 4 to Level 3), the question of what should happen when the driver does not respond to a request to intervene in Level 3, or what is a minimum acceptable risk condition, among others. The definition of Level 5 is somehow problematic too as from a technical and legal point of view is rather difficult to accept the idea of an unlimited/unconditional ODD. It stands to reason that there will always be an ODD. Ultimately, the difference between Level 4 and Level 5 has to do with the size of the ODD, which makes Level 5 somewhat ill-defined. An appropriate ODD taxonomy is of paramount importance at this point (BSI, 2020), and a more reasonable approach would be to establish new levels (or sub-levels) of automation as a discrete representation of the specifications of the ODD. In any case, despite reasonable criticism (Templeton, 2018), and those who believe we should start from scratch (Roy, 2018), the SAE levels of automation is a very solid basis and it makes more sense to further improve and adapt it.

We can translate the levels of automation into more convenient and explicit driving modes to better specify the user's role and responsibility (Schram, 2019). This way, Levels 1 and 2 will refer to *assisted driving*, in which humans are fully responsible for all driving tasks and are assisted by the *driving automation system* (i.e., assisted drivers). Level 3 can be considered as *automated driving*, in which it is the human who ultimately assist

([5]) This involves that, in practice, the driver must be executing the part of the OEDR subtask that allows him/her to perform the DDT fallback in a timely manner, without waiting for an explicit intervention request

([6]) The question of what should happen when the driver does not respond to a request to intervene remains unclear.

([7]) This definition does not explicitly established requirements for a shift to another level of driving automation. For example, in the event where the ODD limit is being reached, it may be more reasonable to shift to another level of automation (from 0 to 3) with a request to intervene to the user, instead of letting the system perform the DDT fallback towards a minimum risk condition. The transition between driving automation levels remains open.

([8]) This definition does not necessarily mean that the vehicle is not equipped with driving controls, allowing the driver to manually engage the DDT if desired.

the system (i.e. assistant or backup driver). And finally, Levels 4 and 5 can be referred to as *autonomous driving*, in which the driver is a mere passenger with no responsibility in the driving tasks. In other words, as illustrated in Table 10, humans are fully responsible for Levels 1 and 2, shared responsibility should be considered in Level 3, while no responsibility can be attributed to them for Levels 4 and 5. In this respect, Level 3 is rather confusing, and can be seen as a perhaps necessary, but not desirable, transitional step towards higher levels of automation.

Table 10: SAE Levels, driving modes, driver's role and responsibility.

SAE Level	Driving Mode	Driver's role	Driver's responsibility
Level 0	Manual Driving	Driver	Full
Levels 1-2	Assisted Driving	Assisted Driver	Full
Level 3	Automated Driving	Assistant/Backup Driver	Shared
Levels 4-5	Autonomous Driving	Passenger	None

Note that, in line with the definition provided in the Regulation (EU) 2019/2144, *automated vehicle* will refer to SAE Level 3, and *fully automated vehicle* to SAE Levels 4 and 5.

3.2 AV Terminology

As we have seen, to know exactly what we mean when we say autonomous or automated vehicle, we have to provide a greater level of detail, i.e., driving automation features and levels of automation. Still, it is useful to clarify the criteria to differentiate the various ways of commonly referring to driving automation, especially when considering the vehicle as a whole rather than some particular features, which is a common approach despite efforts to focus automation on *driving* rather than on the *vehicle*.

Many different terms are used, including "semi-autonomous vehicles", "fully autonomous vehicles", "autonomous vehicles", "automated vehicles", "connected", "cooperative", "fully automated vehicles", "robocars", "robotaxis", "self-driving cars" or "driverless vehicles". The fact that there is no commonly accepted taxonomy generates some confusion and these terms are often used interchangeably. In general we can say that all terms are to some extent imprecise and require a certain level of additional context.

In order to clarify and relate these terms we propose the Venn diagram depicted in Fig. 1. From a bottom-up perspective we have an increased level of automation, intelligence and autonomy, as well as a decreased level of human intervention and interaction. The proposed diagram can be read as focusing either on the vehicle or on the driving automation feature. It is assumed that, from a technical and scientific perspective, higher implicitly contain lower levels of automation. Following the approach established by SAE International standard, we place *driving automation systems* as the core concept from which we can derive the rest (in higher levels of automation - from 3 to 5- we can also consider the use of Automatic Driving Systems - ADS). In what follows we try to briefly describe each term.

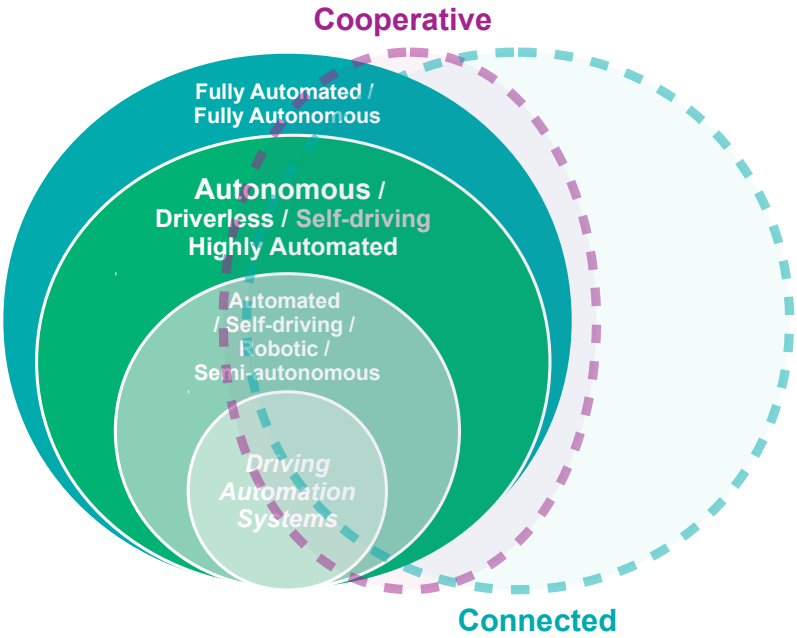
— **Automated**: this is the most commonly accepted term to characterize vehicles equipped with *driving automation systems*, either as automated *vehicles* or automated *driving*. However, it does not clarify the specific level of automation or the role of the user. When we say *automated* we do not know if we are referring to a low/high automation level. Some context, such as "partially", "conditionally", "highly" or "fully", must be added to clearly identify the concept. If this context is not provided, we can consider that *automated* refers to a medium level of automation (e.g., Level 3). This consideration is also in line with the definition of *automated vehicle* provided in the Regulation (EU) 2019/2144.

  The definition of the term *automatic* includes statements such as "that does not require an operator" and "that works by itself under fixed conditions, with little or no direct human control". The fact that automated vehicles are able to perform some *driving automation feature* "by themselves" is one of the reasons to consider *automated* equivalent to *self-driving*, although *self-driving* is vernacularly used to refer to high levels of automation too. As an action or thing, *automatic* is also defined as "self-generated" and "self-acting". Thus, the argument used to indicate that the term *autonomous* is inadequate, concerning its capacity for "self-government", is also applicable to the term *automated*.

— **Robotic**: this term is usually applied to refer to AVs as *robocars* or *robotaxis*. We can consider it somehow equivalent to *automated*, and although it is also used colloquially to connote high levels of automation (as *self-driving*), it does not allow to establish the specific level of automation or the role of the driver.

— **Self-driving**: this term clearly identifies that the vehicle or driving feature is performing the driving tasks (e.g., the DDT) by itself. However, as it happens with *automated* or *robotic*, it does not clarify the specific

Figure 1: Proposed Venn diagram to show how the different terms used when referring to autonomous or automated vehicles are related.



Source: Own elaboration.

level of automation. From Level 1 to Level 5 of automation of a driving feature, the system can be strictly considered to be driving itself. Therefore, *self-driving* does not necessary mean that no driver or user is required, i.e., it does not clarify if in-vehicle users can be considered mere passengers. Still, since it is a term widely used to indicate a high level of automation, *self-driving* is also included in the Venn diagram (Fig. 1) at the same level as *driverless* or *autonomous*.

— **Connected**: this is an additional feature of the *driving automation systems*. There is a widespread trend in scientific, industrial and policy making spheres to speak of *"Connected and Autonomous Vehicles"* or *"Connected and Automated Vehicles"*. Adding *connected* to *automated/autonomous* is reasonable to some extent as *connected* does not mean *automated* (e.g., fully manual connected cars). In-vehicle connectivity is continuously increasing due to many reasons such as mandatory e-call, new potential services (e.g., software updates, updated traffic state for navigators, data recording, etc.) and Internet access for driver and passengers (Alonso et al., 2017).

AVs can complementarily be connected to other vehicles (V2V), to the infrastructure (V2I), to vulnerable road users (V2VRUs), to devices, such as smartphones, smart watches, tablets, personal computers, etc. (V2D), and to the network (V2N). In the case of having connectivity to all of the above, it is called vehicle-to-everything (V2X) connectivity. Connectivity relies on different technologies, including dedicated short range communication (i.e., IEEE 802.11p) for V2V, V2D, V2I and V2VRUs, and cellular V2X (from 3GPP to 5G NR C-V2X). V2X provides 360 degree, non-line of sight sensing with higher ranges than onboard sensors such as cameras, radar or LiDAR (Parra et al., 2019), allows sensor sharing between vehicles, and real-time updates from the infrastructure, increases situation awareness. It also makes it possible to perform predictive and coordinated (i.e., cooperative) driving by exchanging intentions and sensor data (Qualcomm, 2019). These features enhance perception capabilities, intelligence, autonomy and automation level of the *driving automation systems*, so higher levels of automation are expected to require connectivity.

— **Cooperative**: it is common to associate *connected* with *cooperative* as cooperative systems (whether manual or autonomous/automated) require connectivity between agents and infrastructure (although it is possible to cooperate with non-connected vehicles (Parra et al., 2017), connectivity is a fundamental enabler for cooperation). However, *connected* does not necessary involve *cooperation*. We can have *connected* and *automated* vehicles that take advantage of the aforementioned benefits of connectivity without cooperation, which is considered here as an additional feature that affects the individual operational, tactical and even strategic *efforts* or driving tasks (Michon, 1985), with the objective of obtaining individual behaviours subject to the optimisation of the collaborative behaviour of all agents involved. As illustrated in the Venn diagram (Fig. 1), this feature may or may not be present at each level of automation. Finally, the assumption that *cooperative* is somehow the opposite of *autonomous*, because

it implies dependence on communications and/or cooperation with other outside entities, as stated by the SAE International recommendation (SAE International, 2021), lacks validity (this will be further elaborated when describing the term *autonomous*).

— **Driverless/Unmanned**: these terms make explicit that the AV has no backup/assistant driver. The vehicle can be driving empty ("personless") or consider all in-vehicle users as mere passengers. Therefore, a *driverless/unmanned* vehicle or driving feature refers to a high level of automation in which no expectation that a user will respond to a request for intervention is assumed. The terms *driverless/unmanned* by themselves do not sufficiently differentiate between highly automated (Level 4) or fully automated (Level 5), but, as mentioned above, this is also related to an impractical definition of Level 5, i.e., additional context information is required to know whether the ODD is limited or unconditional. The SAE International recommendation (SAE International, 2021) remarks that *driverless* does not clarify if a vehicle is remotely operated by a human driver. However, for the purpose of the level of automation, the consideration of driver should be independent of whether the driver is in the vehicle or operating the vehicle remotely. Therefore, when using the term *driverless*, there should be no confusion about the possible presence of a remote driver.

— **Semi-autonomous / Fully autonomous**: these terms are being used more and more frequently. *Semi-autonomous* usually means that the responsibility for driving falls to the driver (i.e., Levels 1-3) whereas *fully autonomous* usually means that the car is able to "fully drive itself" (i.e., Levels 4-5) (Tyagi and Aswathy, 2021). However, these terms are somewhat vernacular, highly ambiguous, and do not help to clearly specify the level of automation or the role of the in-vehicle users.

— **Autonomous**: as stated by the SAE International recommendation (SAE International, 2021), the term *autonomous* "*has been used for a long time in the robotics and AI research communities to signify systems that have the ability and authority to make decisions independently and self-sufficiently*". The definition of *autonomy* refers to *self-governance* or *independence*. Over time, it has become a synonymous with *automated* since its use "*was casually broadened to not only encompass decision making, but to represent the entire system functionality*".

  The definition of *autonomy* is not directly applicable to the field of automation and there is no clear specification of "levels of autonomy". However, *autonomous* is becoming more and more widespread. For a system (whether software only or embedded in a robot or vehicle) to be *autonomous* requires a high level of intelligence and sophistication. When we talk about *autonomous* vehicles or driving features, we are talking about the highest levels of automation. The idea that the system should be self-sufficient even makes us think of an unconditional ODD (although, this is idea of an unconditional ODD is questionable). We therefore place the term *autonomous* at the highest levels of the Venn diagram (Fig. 1), equivalent to *highly automated* (Level 4). Whereas *automated* does not clarify the level of automation, *autonomous* directly refers to higher levels.

  The definition of *fully automated vehicle* provided in the Regulation (EU) 2019/2144, says "*a motor vehicle that has been designed and constructed to move autonomously without any driver supervision*". That is, "*to move autonomously*" is here linked to the fact that no human supervision is needed (i.e., users inside the vehicle are mere passengers). Note that the definition provided in the Regulation does not refer to the ODD, so it can apply to both *highly automated* (Level 4) and *fully automated* (Level 5).

  Regarding the criticism that relates *connected* and *cooperative* as opposite to *autonomous* the following can be stated. First, connectivity can be considered as an additional input that enhances perception. Receiving information from other agents, or from the infrastructure does not imply losing the autonomous nature. For example, from the point of view of the level of automation or the autonomy, it makes no difference whether the status of traffic lights is obtained by a wireless connection or by a vision-based recognition system, or whether the intention of vehicles to change lanes is sent with V2V communications or estimated using on-board sensors (Biparva et al., 2021). Second, *cooperative* refers to coordinated and collaborative objectives which can affect to all individual driving tasks (operational, tactical and strategic) of each vehicle. It is assumed that these shared objectives, that guide the behaviour of each vehicle, would lead to a higher level of automation and intelligence, and therefore would not negatively affect the level of automation (we can even consider this case as an *autonomous* fleet). If cooperation brings benefits, an *autonomous* vehicle will always tend to behave cooperatively rather than isolated. In other words, *autonomous* does not mean *isolated*.

  Another common criticism with the use of this term is related with its *self-governance* nature. The SAE International recommendation states that "*even the most advanced ADSs are not self-governing. Rather, ADSs operate based on algorithms and otherwise obey the commands of users*". However, the answer to this critic statement can be found in the standard itself, when referring to the possible automation of strategic tasks: "*Strategic aspects of vehicle operation (decisions regarding whether, when, and where to*

go, as well as how to get there) are excluded from the definition of DDT, because they are considered user-determined aspects of the broader driving task. However, for certain advanced ADS applications, such as some ADS-dedicated vehicle applications, timing, route planning and even destination selection may also be automated in accordance with purposes defined by the user". In other words, when we refer to vehicle levels of automation, we are referring to the operational and tactical aspects of driving. The strategic components will always be user-oriented. Therefore, a *highly/fully automated* or *autonomous* vehicle will never decide against user commands.

3.3 Technology readiness levels

As stated above, when referring to automated or autonomous vehicles, we have to be more specific. We need to define the level of automation, the specific feature, function or system, and the set of specifications set out in the ODD. This is fully applicable when referring to the maturity and availability levels of AVs.

As presented in (Martínez-Plumed et al., 2020) and (Martínez-Plumed et al., 2021), when assessing the Technology Readiness Levels (TRL) of AVs, it is necessary to introduce a generality dimension to represent increasing layers of breadth of the technology. But we should be cautious when associating and interpreting the TRL assigned to each level of automation, as it will always depend on the specific feature that has been automated and its operating design domain (e.g., a Level 4 automatic parking system for daytime conditions, which becomes Level 3 for night-time conditions).

In Table 11, by following the methodology proposed in (Martínez-Plumed et al., 2020), we associate the TRLs to each level of automation. A dimension of generality is introduced, representing increasing layers of technology breadth. In this case, generality can be directly linked with the level of automation. Finally, the TRLs are extrapolated for each level by identifying specific *driver automation systems* and their current maturity levels.

Table 11: Technology Readiness Levels (TRLs) for each Level of Automation of AVs.

SAE Level	TRLs	Examples
0	[9]	Conventional modern cars.
1	[9]	Adaptive Cruise Control, Stop & Go, Park Steering Assist, Lane Keeping Assist (proved in operational environment).
2	[9]	Traffic Jam Assist, Automatic Parking Assist, Tesla's Autopilot (proved in operational environment).
3	[5 - 9]	Traffic Jam Chauffeur / Pilot, Automated Lane Keeping Systems, Highway Chauffeur / Pilot, Robotaxis (proved in operational environment).
4	[4 - 7]	Highway Autopilot, Automatic Valet Parking, Autonomous Urban Shuttles, Autonomous Delivery Vehicles, Driverless Robotaxis (demonstrated in operational environment).
5	[1 - 4]	Formulated, experimental proofs of concept, validated in the lab.

The levels of automation should be interpreted as a simplification when referring to AVs, which serves as an abstraction of a more complex and multidimensional problem. Therefore, when the maximum TRL has not been achieved, it is more convenient to consider TRLs ranges for each level of automation (as shown in Table 11).

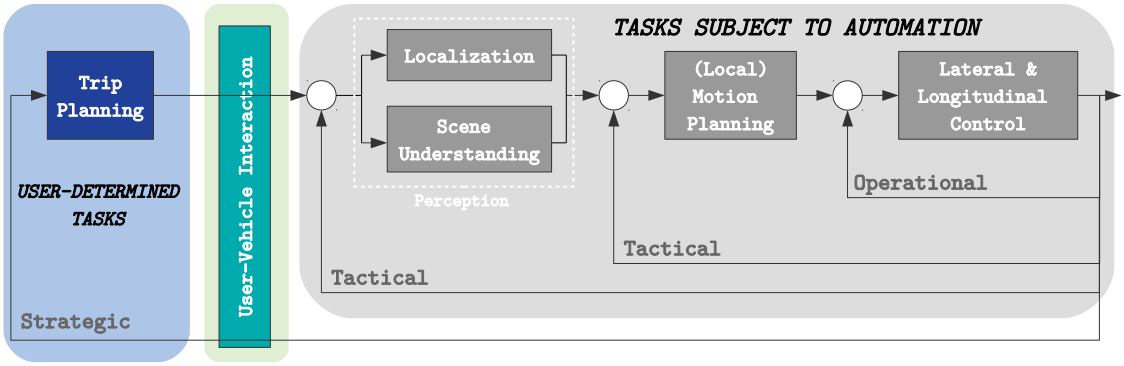
4 Main challenges and dimensions

4.1 Multiple complex problems - multiple AI systems

AVs are highly complex systems that interact in highly complex environments with an almost infinite variety of possibilities. They have to deal with multiple problems of different nature, requiring different approaches and sophisticated solutions. AI is the cornerstone in most of them (as already mentioned, AVs can be defined as a set of multiple, complex interrelated AI-based systems, embodied in the form of a car).

On the one hand, the overall act of driving can be mainly divided into three driving tasks or levels (Michon, 1985): *strategic* (planning), *tactical* (manoeuvring) and *operational* (control). Following a top-down approach, first we have the *strategic* level, which involves planning trips or mission (e.g., where to go, the route to take, being able to stop safely at any time, etc.). Second, we have the *tactical* level, which involves manoeuvring the vehicle in traffic during a trip constrained by the directly prevailing circumstances, including obstacle avoidance, time/distance gap acceptance, turning, lane changes, overtaking, speed selection, etc. Finally, the *operational* level represents the low-level control actions, such as lateral and longitudinal control to maintain lane position in traffic or to avoid obstacles or hazardous events. It is important to note that only *tactical* and *operational* tasks are subject to automation (included in the DDT) since *strategic* plans must be always derived from user commands and goals. Although for certain *driving automation systems*, *strategic* plans such as timing, route planning or destination selection, may also be automated, the goals of such an automation will always be defined, and ultimately controlled, by the users. These concepts are illustrated in Fig. 2, as a schematic view.

Figure 2: Schematic view of user-oriented and potentially automated driving tasks.



Source: Own elaboration.

The aforementioned driving tasks can be related with five major technological levels or layers that AVs must address in order to achieve high levels of automation: (1) *localization*, (2) *(dynamic) scene understanding*, (3) *(local) path planning*, (4) *lateral/longitudinal control* and (5) *user interaction*. The first two layers can be addressed together within the so-called *perception* layer, but in this report they are considered independently. The relation of the layers with the driving tasks is implicitly depicted in Fig. 2. In all these technological components the role of AI is predominant and, in some cases, indispensable.

1. **Localization**: one of the most critical steps of an AV consists of its precise localization (position and heading) in global coordinates, at lane level (with a maximum error of a few centimetres), and within an enhanced digital or high-definition (HD) map containing complete semantic and geometric information of all static elements of the scene including traffic lights, traffic signs, pedestrian crossings, lanes, street layout, intersections, roundabouts, parking areas, green areas, etc. When V2I is available, these maps can also contain dynamic real-time information (Kuutti et al., 2018) such as weather and traffic conditions (e.g., accidents, congestion, etc.), and traffic lights state.

   To make this process as robust as possible, sensor fusion is becoming increasingly important (Fayyad et al., 2020), including GNSS, Inertial Measurement Units (IMUs), cameras, LiDAR, radar, ultrasonic, WiFi, and wheel odometers. The role of accurate HD maps, including digital (Parra Alonso et al., 2012), visual (Lategahn and Stiller, 2014), or 3D maps (Levinson and Thrun, 2010), is critical, and despite the high costs and difficulties to scale, mapping regions with multiple sensors, and in various weather and lighting conditions, is one of the most important areas of development and innovation in the overall AV landscape today.

   The localization problem can be described as follows. A conventional GPS provides a first approximate location, which allows to delimit the search area within the map. From the map (digital, visual or range-

based) of that area, and with the data captured by the sensors (e.g., IMUs, cameras, LiDAR), different map-matching techniques are used to accurately establish the position and heading of the vehicle within the map. The basic idea is to find similarities between these a priori accurate maps and current sensor data. While traditionally these techniques were not based on learning, data-driven approaches are becoming increasingly prevalent (Ma et al., 2020).

2. **Dynamic scene understanding**: once the vehicle is accurately positioned within an enhanced digital map, the next step is to detect and understand the dynamic content of the scene, including vehicles of all types (cars, vans, trucks, etc.) and VRUs (pedestrians, cyclists, motorcyclists, etc.). Dynamic agents must be accurately positioned globally on the map or relative to the vehicle. Similarly to the localization problem, robust detection and localization of dynamic agents requires the fusion of data from multiple sensors (Fernández Llorca et al., 2020).

   Understanding the scene may also include segmenting the static elements of the scene from the perspective of the vehicle (Cordts et al., 2016). If we have a detailed digital map and accurate localization, this step may seem redundant. However, redundancy is desirable to increase the robustness of the perception process. In addition, static scene segmentation always provides up-to-date local information relative to the vehicle, which may not be the case when relying on digital maps (e.g., temporary traffic signs and signals in the event of construction work).

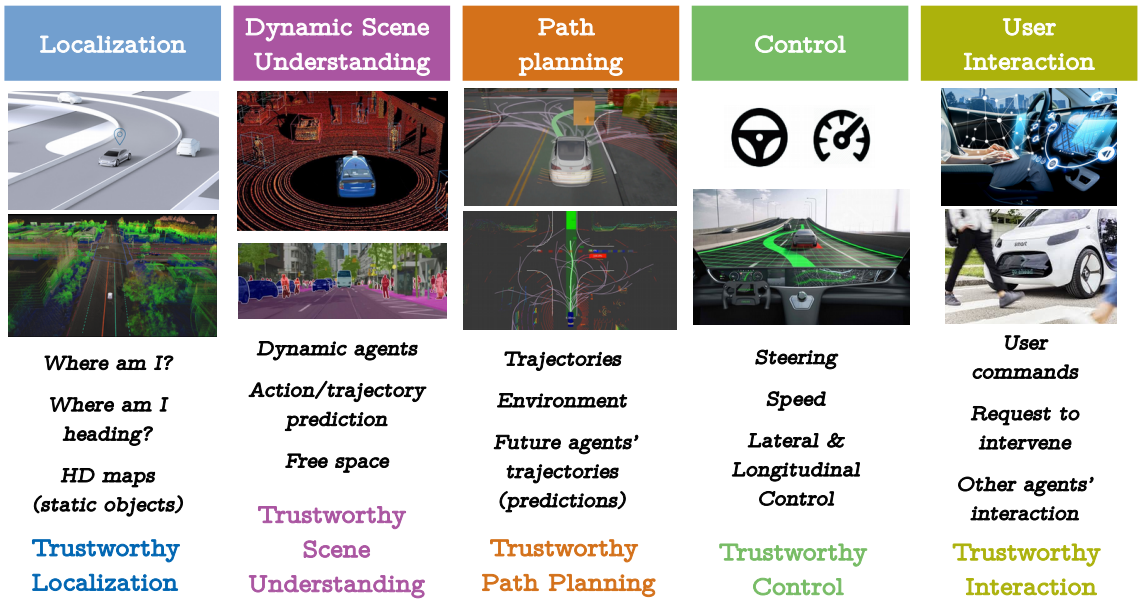
   Finally, detection and localization of dynamic agents are only the first steps. The next step is to predict future behaviours and trajectories of vehicles (Izquierdo et al., 2020), (Lefevre et al., 2014) and VRUs (Quintero Mínguez et al., 2019), (Rudenko et al., 2020). Autonomous vehicles must be endowed with predictive perception to perform autonomous driving. The future trajectories of other road agents are essential inputs for safe and comfortable path planning. Modelling the behaviour of drivers and VRUs is a very complex problem, and still in the early stages of research. There are multiple variables that influence the behaviour of agents, including intrinsic variables (gender, age, position, velocity, etc.), context (traffic, weather and lighting conditions, street layout, etc.) and interaction variables (group behaviour, awareness, etc.). The variability of this information requires the use of multiple sensors of different nature. The most promising approaches are data-driven and sophisticated deep learning-based solutions.

3. **Path planning**: this stage is also known as (local) motion or trajectory planning. It is based on the definition of a global route (route or trip planning) that indicates the path from the vehicle location to the destination (user specific destination from user interaction layer). The route must have centimetric accuracy at lane level and is usually defined as a set of waypoints. On this route, and based on the precise localization of the vehicle, the knowledge of the traffic rules, state of the signals, and the detection and prediction of trajectories of the dynamic agents, local planning is performed, which provides a feasible and smooth trajectory and speed references on which the control system will operate. Motion planning also involves behavioural decision making including lane changes, overtaking and obstacle avoidance manoeuvres, emergency braking, intersection negotiation, etc. 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 (González et al., 2016), (Paden et al., 2016). End-to-end data-driven approaches are also becoming more predominant (Pfeiffer et al., 2017), (Chen et al., 2017), (Aradi, 2020).

4. **Lateral/longitudinal control**: in order to execute the reference path and speed profile of the motion planning system, feedback controllers are used to select the most appropriate actuator inputs to perform the planned local trajectory. The controllers must be designed to correct the difference between the reference motions and the actual state of the vehicle, in the presence of modelling errors and other forms of uncertainty, with robustness, stability, safety and comfort. Many different types of closed loop controllers have been proposed for executing the reference motions provided by the path planning system (Paden et al., 2016), including path stabilization, trajectory tracking, and more recently, predictive control approaches.

5. **User interaction**: a fundamental part of autonomous driving consists in the design of appropriate human-vehicle interfaces to address effective interaction and communication with in-vehicle users, such as (backup) drivers (shared responsibility, Level 3) and passengers (strategic tasks, Levels 4 and 5), and external road users, including VRUs (pedestrians, cyclists, etc.) and drivers (Jafary et al., 2018). Potential modalities 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) (Rasouli and Tsotsos, 2020). Regarding driver and passengers, common interfaces are audio, tactile, visual, vibro-tactile, and more recently, natural language processing (Roh et al., 2020). In addition, this layer also includes in-vehicle perception systems to detect the status of users in order to enable communication and iteration.

Figure 3: Main stages of an AV. Each one with one or multiple AI systems.



Source: Own elaboration.

Therefore, one of the main challenges of any potential procedure for assessing the trustworthiness of AI systems in autonomous driving is the fact that there are multiple complex problems that are addressed by multiple AI solutions of different nature. Moreover, apart from the interactions between layers, at each layer we can have multiple AI-based systems interacting with each other. This raises several questions. *Should a holistic assessment be carried out at the system level (interrelation of multiple AI systems, final user-vehicle interaction) or should an analysis of each AI be carried out independently (trustworthy localization, trustworthy scene understanding, etc.)?*. The first option is less comprehensive and therefore more feasible. However, the nuances of the different problems can be more appropriately addressed with an individual approach. In any case, this is clearly a challenge and, in both cases, there does not seem to be a clear path to approach trustworthy analysis considering the interaction and dependence of the different systems on each other.

In our analysis we will take the first option, assuming that the layers described above are mainly AI-based, so **the question of how to deal with trustworthy AI systems is practically identical to the question of how to deal with trustworthy AVs**.

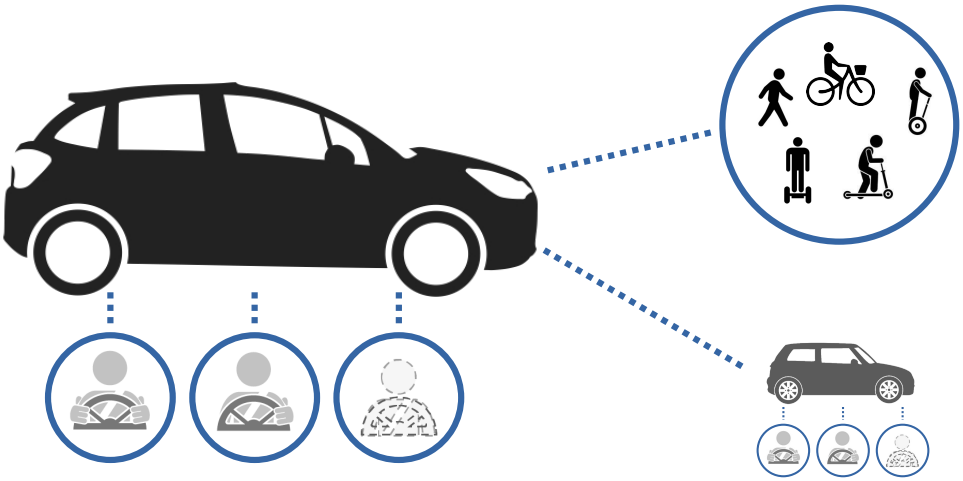
4.2 Multi-user considerations

As stated by the European Commission Expert Group established to advise on specific ethical issues raised by autonomous mobility for road transport (European Commission, 2020), "*in line with the idea of a human-centric AI, the user perspective should be put centre-stage in the design of AVs*". They highlighted the importance of designing interfaces and user experiences taking into account known patterns of use, "*including deliberate or inadvertent misuse, as well as tendencies toward inattention, fatigue and cognitive over/under-load*". In addition, it is remarked that "*in line with the principle of justice [...] AVs should adapt their behaviour around VRUs instead of expecting these users to adapt to the (new) dangers of the road*".

The difficulty of the user perspective problem lies, therefore, in a double dimension that also includes multiple types of agents that must be considered independently, as illustrated in Fig. 4. On the one hand, we have the perspective of the user inside the vehicle, which can be divided into that of an assisted driver (Levels 1 and 2), an assistant/backup driver (Level 3), or a passenger with no responsibility for driving tasks (Levels 4 and 5). Thus, the human-centric design with respect to the in-vehicle users changes according to the type of user (i.e, the level of automation). On the other hand, there is the perspective of external road users who interact with the AV, which includes other drivers of non-automated vehicles or vehicles with medium levels of automation, other passengers of AVs, as well as VRUs such as pedestrians, cyclists, and users of new mobility systems such as classic skaters, e-scooters and different types of hover-boards. Similarly, user-centric design for external road users involves multiple dimensions and perspectives that must be addressed independently.

To endow AVs with the ability to adapt their behaviour around users, the development of appropriate models of human behaviour is a fundamental step. These behavioural models can be used to predict future intention and motion of external road users which enables predictive motion planning, as well as to better interact with in-vehicle users. Therefore, monitoring the state of the backup driver or passengers becomes essential not only

Figure 4: Interaction and communication of AVs with drivers/passengers and external road users. User-centric design should address multiple dimensions and perspectives.



Source: Own elaboration.

to ensure an appropriate use of the *driving automation system* depending on the level of automation, but to model and infer their behaviours and adapt the interaction to their condition.

These multiple dimensions in the context of users has another component that introduces even more complexity: *passengers and external road users may have conflicting objectives and interests*. For example, the well known *crosswalk chicken* problem (Millard-Ball, 2018), i.e., in a normal scenario involving human drivers, if a vulnerable road user, such as a pedestrian, decides to cross, he or she is at considerable risk, either because traffic regulations allow not yielding to pedestrians, because the driver may be distracted, or because the driver assumes that the pedestrian does not intend to cross. In the case of autonomous driving the perceived risk of crossing may become non-existent because the pedestrian knows that the AV will stop in any case. This could be considered as an abuse by the pedestrian, and would considerably slow down autonomous driving compared to manual driving. This situation must be addressed from multiple points of action, including the educational and legal spheres.

The question then arises as to whether the analysis and development of some of the requirements for trustworthy AI systems should be done from the perspective of users inside or outside the AV. Reaching the right balance is undoubtedly a major challenge.

4.3 Trustworthy AI requirements for testing in real traffic conditions

Autonomous vehicles are one of the clearest examples of the need of a transitional period of testing in real operating environments, prior to the commercialization of the systems. In order to develop, validate and improve prototypes, it is necessary to put them to work in real traffic conditions, incorporating more and more data to fine-tune the different AI systems, and evaluating the performance of the different layers (Fig. 3) in multiple types of scenarios. In fact, most of the automated vehicles currently on public roads around the world are in the testing phase, under strict and specific legal requirements.

This transitional test-driving period for AVs is also considered as a necessary step to assess their safety and reliability before they are allowed on the road for consumer use, although it is commonly accepted that for fatalities and injuries, test-driving alone cannot provide sufficient evidence for demonstrating AV safety (Kalra and Paddock, 2016).

In terms of liability, this testing phase is much less problematic, since the provider or developer of the technology is perfectly identified as the entity responsible for any possible accident. It is in their interest and benefit to conduct such tests to improve their technology for commercial purposes. However, it is questionable whether the trustworthy AI assessment criteria should be applied with the same level of stringency or whether they should be adapted to the specific nature of the testing process. Further discussion is needed in this regard. We should not forget that conducting automated or autonomous driving tests on public roads can be considered as a high-risk use case in itself.

5 Assessment list impact on Autonomous Vehicles

In this section we assess the current status of the key requirements for trustworthy AI in the autonomous driving domain, considering that AI is the main enabler for autonomous driving. Each requirement is addressed by analysing the most solid and recent work related to them. The analysis is carried out independently on each one. However, as will be seen below, there are clear relationships between them. In addition, given the context-specificity of AI systems, we evaluate the relevance of each assessment criterion to the context of AVs. As depicted in Fig. 5 we propose three different levels to evaluate the relevance and urgency of each criterion with respect to the AV domain: *critical in the short term*, *important in the mid term* and *impact in the long term*. The proposed assessment is supported and elaborated in the following subsections.

5.1 Human agency and oversight (KR1)

Human agency and oversight are requirements prescribed by the principle of respect for human autonomy. AVs must support human autonomy and decision-making, which should be based on informed decisions. Therefore, the main approach to human agency and oversight is through human-vehicle interaction, and this interaction is developed through Human-Machine Interfaces (HMIs) that are linked to vehicle technology and human capabilities. In the following, we discuss the main mechanisms of interaction between AVs and humans, followed by the two main criteria (agency and oversight) of this key requirement.

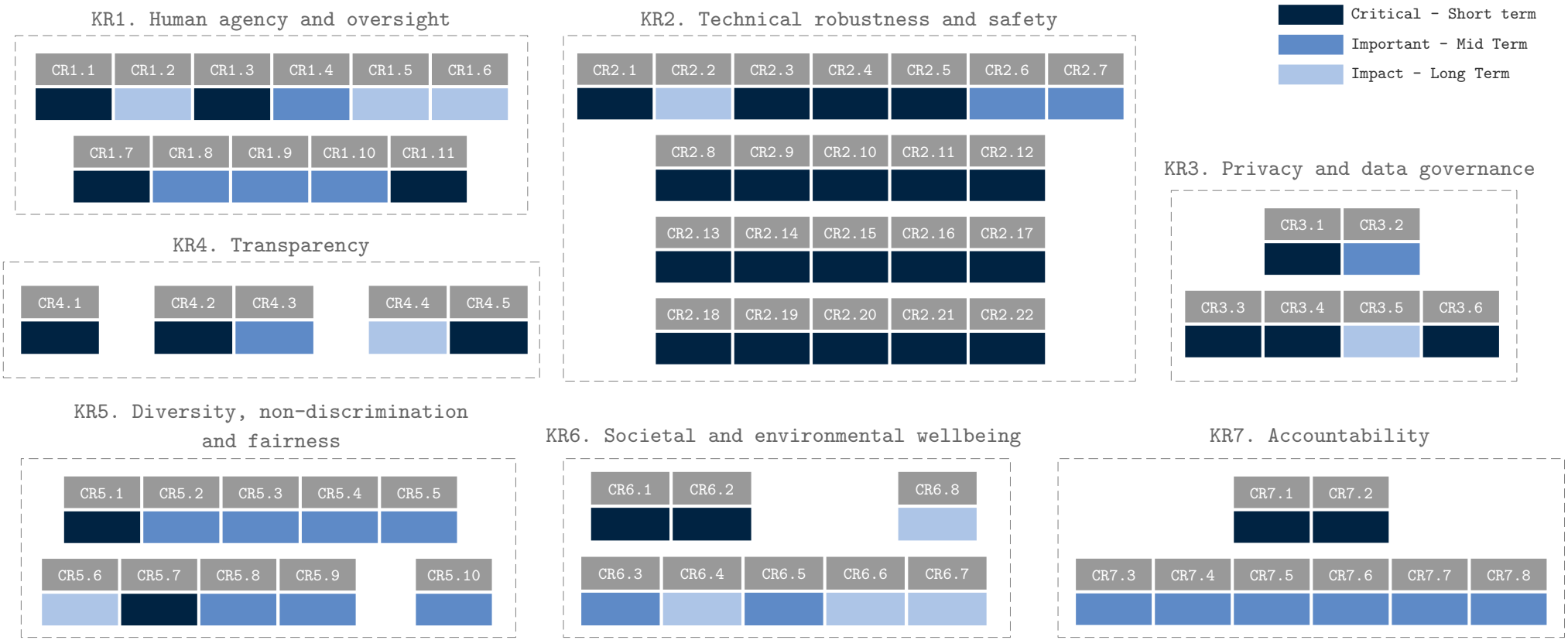
5.1.1 Human-vehicle interaction

Human-vehicle interaction in the context of AVs is a multi-user problem, as we have two main groups of humans: those using the AV, whether they are assisted (Levels 1 or 2) or backup (Level 3) drivers, or passengers (Levels 4 or 5), and external road users interacting with the AV. On the one hand, human-vehicle interaction for in-vehicle users (drivers or passengers) is a well-known research topic (Biondi et al., 2019), (Detjen et al., 2021). However, the number of works focusing on fostering human agency or human oversight through appropriate interactions is still limited (Silva, 2020). On the other hand, human-vehicle interaction for external road users (e.g., pedestrians, other drivers) is a much less mature field of study than for in-vehicle users, but it is attracting increasing attention from the scientific community (Rasouli and Tsotsos, 2020). The interaction between AVs and humans is developed through human-machine interfaces (HMIs) that are linked to vehicle technology and human capabilities. These interfaces must be clear enough to avoid any confusion regarding whether the user is interacting with a human or an AI, which in the context of the AV can only lead to confusion through teleoperated driving (Keller et al., 2021).

We can identify multiple potential input/output modalities governing the design of HMIs in the field of AVs. As depicted in Fig. 6, we have different in-car input/output modalities corresponding to human's senses and sensing technologies. We have highlighted in bold the most feasible and relevant modalities. On the one hand, driver/passenger status monitoring systems are required to collect user input features (Daza et al., 2011) (Hecht et al., 2019). For example, head/body pose, mid-air gestures and eye gaze can be detected using computer vision techniques. Voice interaction can be approached using speech interfaces with voice recognition and natural language processing. Haptic and touch controls are usually hard installed using buttons and touch screens.

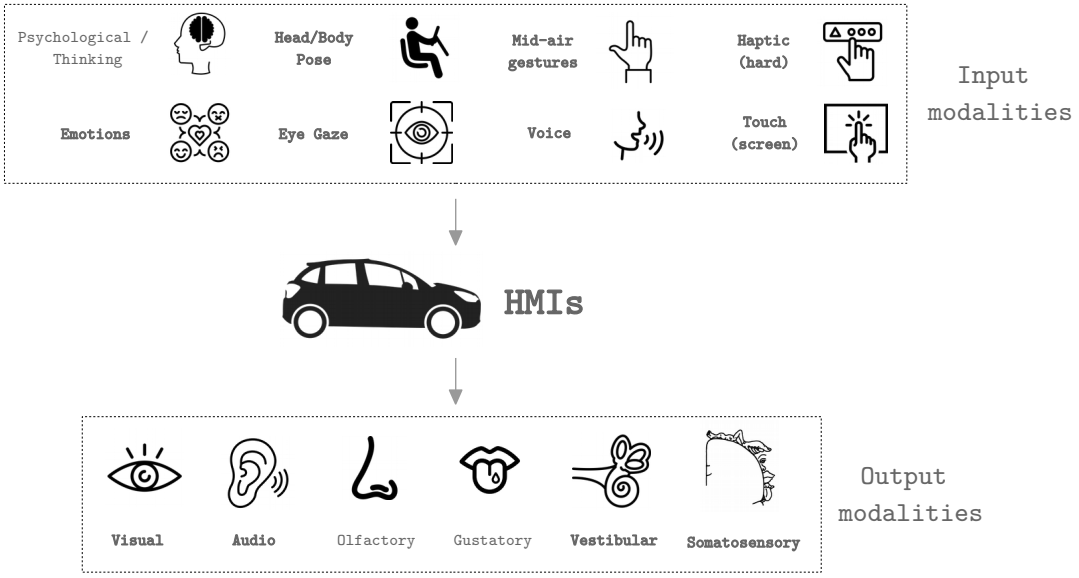
Although emotions are difficult to measure objectively because physiological arousal, facial expressions (Hupont et al., 2010) and voice (Yacoub et al., 2003) are important indicators that can allow one to infer the potential emotion of the user. Psychological properties and thinking, which are related to emotional state, are less feasible to obtain in the medium term, as they require the user to wear additional devices, such as head interfaces with electroencephalogram (EEG) electrodes. On the other hand, human senses define the system's output modalities. Visual and auditory feedback are the most feasible ways to interact with in-vehicle users. Somatosensory feedback can be achieved using different types of haptic controls, and vestibular feedback is primarily addressed through the feeling of speed, acceleration and jerk of the body. So far, olfactory and taste feedback do not seem relevant yet, and in any case, they are still difficult to implement in technological terms. Nevertheless, they can be used to enrich the user experience.

Figure 5: Relevance and time horizon of the assessment criteria for the seven key requirements.



Source: Own elaboration. Qualitative interpretation and representation based on the analysis of each of the criteria, as set out in the following sections.

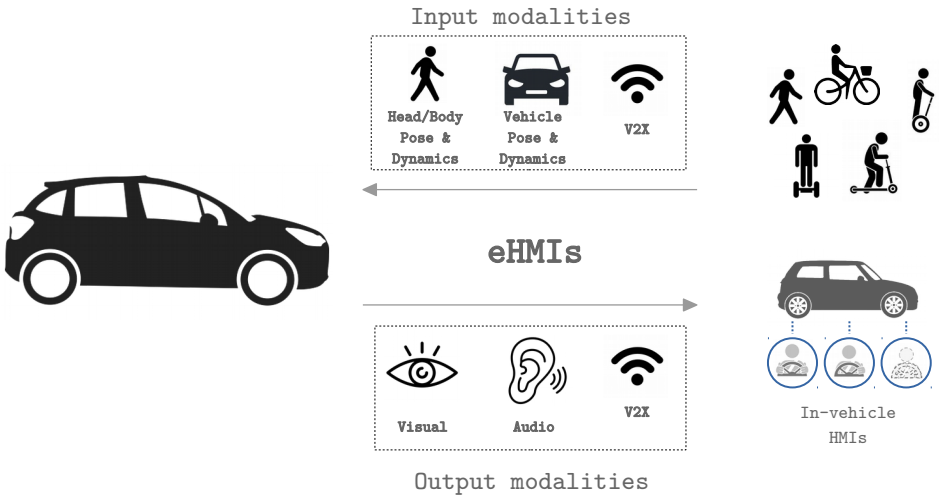
Figure 6: Input and output modalities of in-vehicle-human-machine interfaces.



Source: Own elaboration, based on (Detjen et al., 2021).

The input and output modalities with respect to external road users are considerably reduced. As can be seen in Fig. 7, input features can include pose and dynamics (including behavioural analysis) for both humans and vehicles, as well as any other feature shared by means of V2X connectivity. Output modalities to interact with external road users can involve visual and audio feedback, as well as other features through V2X communications. Potential implementations of external Human-Machine Interfaces (eHMI) for AVs are manifold, reaching from displaying text messages on external displays, using different colors of lights in light-based concepts, over laser projections on the street, by means of audible messages and signals, to personalized messages for smart and wearable devices (Dey et al., 2020).

Figure 7: Input and output modalities of vehicle-human-machine interfaces for external road users.



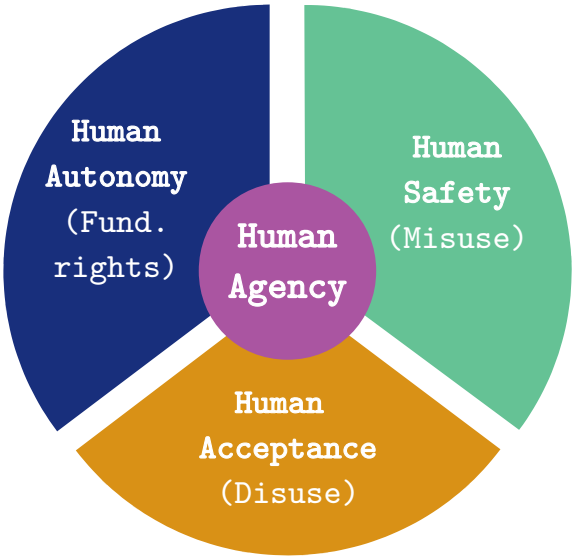
Source: Own elaboration.

5.1.2 Human agency and autonomy

Agency, can be defined as *the feeling of control over actions and their consequences* (Moore, 2016). As a feeling, the sense of agency is a subjective human trait that significantly influences the way humans behave and interact with each other and with technology. Therefore, human agency is not only important because it is linked to human autonomy and fundamental rights, but it is also a fundamental issue for the appropriate use of technology through the belief that is serves users to achieve their goals. That is, the lack of an appropriate sense

of agency can lead directly to disuse or misuse of technology, affecting user acceptance and safety respectively. For example, on the one hand, an increased desire to exert control on driving tasks by some users is negatively correlated with their intention to use autonomous driving (Nastjuk et al., 2020) leading to disuse. On the other hand, misuse occurs when the assisted (Levels 1 or 2) or backup (Level 3) driver relies excessively on the *driving automation system* (i.e., overreliance), since there is a reduced sense of agency linked to driver disengagement that can cause serious problems when the driver needs to resume manual control (Navarro et al., 2016). Thus, human agency in automation can be directly linked with three components: autonomy, acceptance and safety (see Fig. 8).

Figure 8: Human factors of human agency requirement for a trustworthy AI system.



In addition, it is commonly accepted, and has been demonstrated in several studies (Berberian et al., 2012), (Yun et al., 2018), (Ueda et al., 2021), that *the sense of agency decreases with the level of automation*. This mainly affects in-vehicle users, i.e. drivers and passengers. That is, we can reasonably expect the sense of agency of a passenger (Levels 4 or 5) to be much lower than that of an assisted (Levels 1 or 2) or backup (Level 3) driver, just because of the degree of automation in each case. Therefore, the main question is ***how to foster appropriate human agency of drivers and passengers depending on the level of automation to mitigate the consequences of reduced agency***. This is particularly important at Level 4, as the control options available to passengers are more limited (e.g., autonomous shuttles without steering wheel and pedals). The main process that facilitates the calibration of perceived agency is human-vehicle interaction. *Agency calibration* refers here to the process of adapting the agency perceived by the user to a point where no disuse or misuse occurs. When calibrating agency, what human and automation each do is less important than how human and automation are structured to interact (Cesafsky et al., 2020).

Calibrating human agency through appropriate human-vehicle interactions, i.e., **designing appropriate agency-oriented HMIs**, is a very challenging task since agency is not strictly a physical phenomenon but a subjective feeling influenced by one's context information, background beliefs and social norms. Humans *not only experience agency physiologically but also interpret it subjectively, making it an imprecise measure of reality in which a person may feel more or less in control than they actually are* (Silva, 2020). In other words, our experiences of agency can be quite divorced from the facts of agency. The most common method to measure the sense of agency is through self-reporting questionnaires from which the subjects need to retrospect on what they have done and felt to provide a response that can be binary, categorical or rating. The main limitation is that it may introduce individual judgment bias which can be minimized with sufficient sample size. Other approaches are sensory attenuation ([9]) and intentional binding ([10]) (Wen et al., 2019), which have other limitations such as unsuitability for continuous actions and events and noise, respectively. In addition, there have been some attempts to use EEG signals to decode the sense of agency from brain activity, but their applicability is still limited (Wen et al., 2017).

Another fundamental elements affecting the sense of agency, especially when it comes to an application

([9]) Sensory attenuation refers to the phenomenon that a self-produced stimulus feels less intensive than an externally produced stimulus, which can be used as a mechanism to measure wheter humans feel a sense of agency over the stiumuls (Wen et al., 2019).

([10]) Intentional binding refers to the phenomenon in which the perceived time between an action and its consequence is reduced when people have a sense of agency over the consequence, and increased when there is no such sense of agency (Wen et al., 2019).

context with complex interactions, are **explainability** and **interpretability**, i.e., making the logic supporting AI decisions interpretable and understood by users (Heer, 2019), (Silva, 2020). Explainability, therefore, must be developed on the basis of human-vehicle interfaces. Moreover, this element links with key requirement 4 (KR4 Transparency), and more specifically with criteria CR4 and CR5 (explainability), which will be discussed in more detail later in this section.

From a multi-user perspective, and although it is still an emerging field of study, it is important to note that the sense of agency has to be also considered for external road users. In the same way that placebo buttons at pedestrian crossings have proven to be effective in preventing inappropriate behaviour (e.g., improper crossings) due to the sense of agency they convey to pedestrians (McRaney, 2013), it is worth asking **what are the most important modes, variables and communication mechanisms to calibrate the sense of agency of external road users in their interaction with AVs**. For example, in manual driving, a pedestrian who wants to cross the road at a pedestrian crossing somehow negotiates the crossing by direct visual contact with the driver, which affects his/her sense of agency. How to generate the same effect by AVs is an area for future research.

5.1.3 Human oversight

As stated in (Koulu, 2020), human agency can be portrayed as a tool for overcoming the ethical concerns and risks associated with AI systems, when it is considered in the form of varying levels of human *oversight*. As already mentioned, human oversight helps to ensure that an autonomous driving function does not undermine human autonomy or causes other adverse effects. And as already described, in the field of autonomous driving, **human oversight plays a fundamental role in defining the level of automation** of a function or of the entire vehicle. Three main categories can be defined according to the level of cooperation between humans and autonomous systems, or according to the monitoring and action requirements (degree of involvement) imposed on humans:

1. **Human-in-the-loop**: the autonomous system carries out some tasks for a time period, but wait for human commands before continuing (Nahavandi, 2017). The human operator must be permanently in connection with the autonomous system and the autonomous capacity is limited to some tasks (Hodicky et al., 2018).

2. **Human-on-the-loop**: the autonomous system can execute some tasks completely and independently but have a human in a monitoring or supervisory role, with the ability to interfere if the system fails (Nahavandi, 2017). Describes a situation where human operator is *in command* of the autonomous systems, and there is some cooperation to achieve the objectives, while some critical decision must be made by the human operator and cannot be made autonomously (Hodicky et al., 2018).

3. **Human-out-of-the-loop**: any decision is made by the autonomous system (Hodicky et al., 2018). Humans cannot be called "operators" in this case .

These definitions, which come from the human-machine interaction research community, are based on a higher abstraction level of the degree of human involvement in the human-machine interaction loop. It is possible to translate and adapt these concepts (which are used in criterion CR1.7) to the field of autonomous driving (Saleh et al., 2017). Table 12 summarizes the link between the different SAE automation levels and the aforementioned categories.

Table 12: SAE Levels and human-vehicle interaction loop.

SAE Level	Decision-making	Human role	Human-Vehicle Interaction Loop
0	No driving automation (warning-only systems)	Driver	Human driver
1-2	Driver assistance / Partial driving automation	Assisted driver	Human-in-the-loop
3	Conditional driving automation	Backup driver	Human-on-the-loop
4-5	High/Full driving automation	Passenger	Human-out-of-the-loop

When the human is out of the loop (SAE Levels 4 and 5), using the AV as a passenger, we can then analyse what is the role of external road users (VRUs and other drivers) in the interaction loop. For example, when AVs encounter a pedestrian trying to cross the road from a non-crosswalk stop, the vehicle will change its

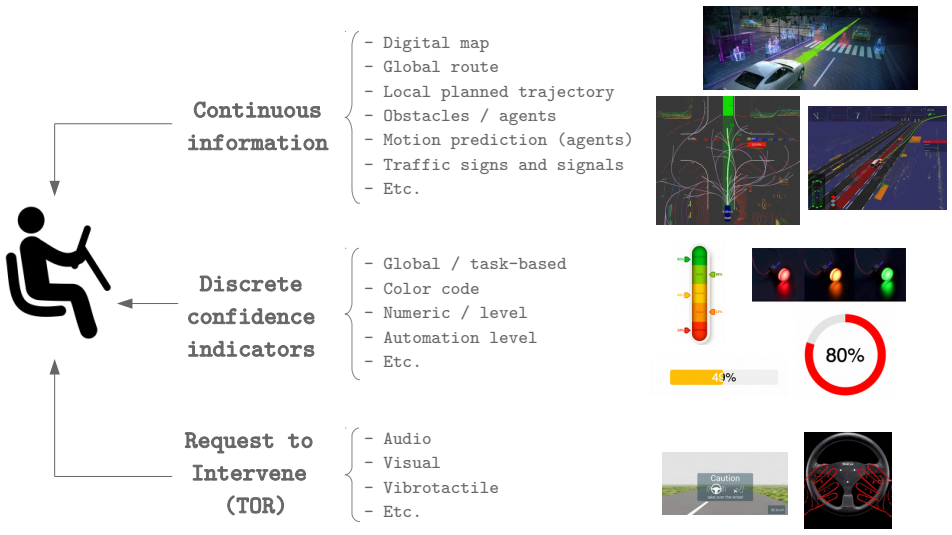
local trajectory planning to avoid the collision, by performing either emergency braking or steering. This can be considered as a direct intervention of the external user over the decision-making loop of the AV. This supports the idea that **external road users will be acting as human-on-the-loop** in such cases (Saleh et al., 2017), which generates multiple problems and unknowns, such as for example, the potential abuse by external road users who can behave with impunity, hindering the movement of the AV. As previously stated, these issues should be addressed from multiple perspectives, including the educational and legal contexts.

The human-vehicle interaction loop is also known as **human-vehicle *cooperation***. The underlying idea of this term is that there is a shared control strategy in which the driver (or passenger) and the vehicle act as a team in driving tasks. Several cooperative models for human-vehicle interaction in automated driving have been proposed (Biondi et al., 2019). These models go beyond the definition of the levels of automation and offer a more flexible approach to human oversight, where both the **automated system and the drivers must learn how to cooperate safely and efficiently in the dynamic environment**. In order to cooperate successfully, there must be bidirectional or mutual awareness.

On the one hand, **the AV must know the state of the driver in Levels 1-3** (Hecht et al., 2019). The degree of driver engagement or distraction should be continuously monitored, including the three main sources of distraction (Strayer et al., 2019): manual, visual and cognitive distractions, which occur when the driver's hands are not on the steering wheel, his/her eyes are not on the road and his/her attention is diverted from the task of driving respectively. If the driver's level of attention is inadequate, the *driving automation system* may stop because the user is misusing it, and the consequences can be fatal ([11]). For passengers (Levels 4 and 5), although continuous monitoring of the passenger state is not prerequisite to ensure proper use of the system, it could bring benefits to improve the user experience.

On the other hand, **the driver/passenger must understand the state and driving capabilities of the autonomous system**. This implies defining what type of information, level of detail, and communication modality are the most efficient for the user of an AV to understand the state of the system effectively, without becoming saturated, and even being able to attend to other non-driving related tasks. These problems are closely related to KR4 (Transparency), and more specifically to the explainability criteria CR4.2 and CR4.3. As shown in Fig. 9, on the one hand, the information to be transmitted can be continuous, including elements such as the map, the global route, the planned local trajectory, the detected agents and obstacles, the predicted future trajectories of the agents, the traffic signs, the status of the traffic lights, etc. This information can be very useful for understanding vehicle behaviour, but may not be very effective in cases where it is necessary to resume manual control with some urgency. This is why, on the other hand, discrete systems can be designed to provide the level of confidence or reliability of the automation systems. Such indicators, which can be numerical or color-coded, can be applied holistically for the entire automated system, or at the task level, e.g. location indicator, scene understanding indicator, local trajectory safety indicator, etc.

Figure 9: Continuous and discrete visual information from the autonomous driving systems to the user, including the take over request.



The direct application of SAE automation levels leaves little room for effective cooperation. Thus, for example, in levels of assistance or partial automated driving (Levels 1 and 2), no level of driver distraction should be allowed. In conditional automated driving (Level 3), it is theoretically possible to devote part of the

([11]) As an example we refer to the multiples cases in which misuse of the Autopilot of Tesla (SAE Level 2) is leading to fatal accidents (Corn, 2021).

mental workload to other tasks, such as leisure or work (non-driving related tasks). In fact, there are multiple approaches on the study of effective cooperation strategies in Level 3 of automation that allow the driver for long distractions from driving tasks (Schartmüller et al., 2019), (Schartmüller et al., 2020). However, in practice, this poses a great risk. The need for the user to be responsive to relevant failures in a timely manner, to resume manual control or even achieve a minimal risk condition, with the user being ultimately responsible for any kind of failure, means that the driver's attention must be almost permanent. Finally, at levels of autonomy (SAE Levels 4 and 5), cooperation in driving tasks is meaningless.

Whether in a cooperative driving environment, in Level 3 automation, or in transitions between levels (e.g., from 4 to 3), there is one element whose importance is fundamental to safe and efficient driving: the ***take-over request* (TOR) or *request to intervene***. This is a notification sent by the automated driving system to the user indicating the need to resume manual operation of the vehicle to either perform the driving tasks or to achieve a minimal risk condition. Two main variables have been studied, normally under simulated environments. First, the communication modality, which includes visual, vibrotactile, and auditory modalities (Yoon et al., 2019). Second, reaction time, which is affected by different factors including driver distraction level (Zeeb et al., 2015), type of non-driving related task (Yoon et al., 2019), communication modality, and experience in handling previous requests to interviene (Zhang et al., 2019). Perfomance-based approaches, i.e., take-over requests tailored to the driver's behaviour (Kim and Yang, 2017), have proven to be a very effective way to address this critical task.

Another factor to be analysed within this requirement is the training or educational processes on *how to exercise oversight*. There is a clear link between driver experience and prior knowledge at all levels of automation, not only for developing acceptance and trust, but for effective use of the *driving automation system* (Endsley, 2019). A good example of the importance of proper training on the human-machine team concept can be found in the field of aviation, where pilots must undergo a sophisticated training process on automation. Although probably with less requirements, it is very reasonable to think of minimum training standards for drivers of vehicles with automated driving systems (Casner and Hutchins, 2019). At low levels of automation (1 to 3), the way in which drivers can exercise control of the AV is quite similar to that of conventional vehicles, including all types of driving tasks, i.e., *strategic*, *tactical* and *operational* (see Fig. 2). For higher levels of automation (4 and 5), there is still no clear protocol for action. The passenger must control *strategic* tasks, such as selecting the destination, changing the route, or requesting a stop as immediately as possible. Some *tactical* tasks may also be controlled by the passenger, such as requesting to overtake or not to overtake, waiting longer at a junction, increasing/decreasing speed, etc. As shown in Fig. 6 the modalities for exercising this control are manifold. Finally, it is reasonable to consider that, if there are self-learning processes or system updates that lead to substantial changes in the way human oversight is exercised by the driver, the educational programs or skills required to drivers will need to be updated accordingly.

5.2 Technical robustness and safety (KR2)

Technical robustness and safety are requirements mainly linked to the principle of prevention of harm. They are developed through four main sub-requirements that are addressed below.

5.2.1 Resilience to attack and security

AVs can have adversarial, critical or damaging effects to humans in case of risks and threats such as technical faults and defects, outages, attacks, misuse and inappropriate or malicious use, and they are clearly exposed to potential cyberattacks. These are important issues not only because they are linked to the principle of harm prevention, but also because they have a strong influence on potential disuse. Indeed, vehicle cybersecurity and protection against unauthorized access are fundamental requirements strongly correlated with the intention to use (Garidis et al., 2020). Therefore, as established by the European Union Agency for Cybersecurity (ENISA) in its report *ENISA Good Practices for Security of Smart Cars* (ENISA, 2019), with the increasing connectivity of AVs, novel cybersecurity challenges, risks and threats are arising, and cybersecurity is becoming a crucial aspect that will affect the evolution of the technology.

Research related to attacks and defense techniques targeting assited, automated or AVs has grown significantly in recent years (He et al., 2017), (Dibaei et al., 2020), (Pham and Xiong, 2020), (Kim et al., 2021). A large part of these potential attacks and defensive methods are not inherent to high levels of automation, but are applicable to both manual vehicles and vehicles with low levels of automation. If we focus on the elements inherent to AVs, we can distinguish two main types of attacks, as shown in Fig. 10. First, ***external attacks*** which are carried out from outside the vehicle attempting to exploit vulnerabilities of the vehicle perception and communication systems. This type of attacks includes the following cases:

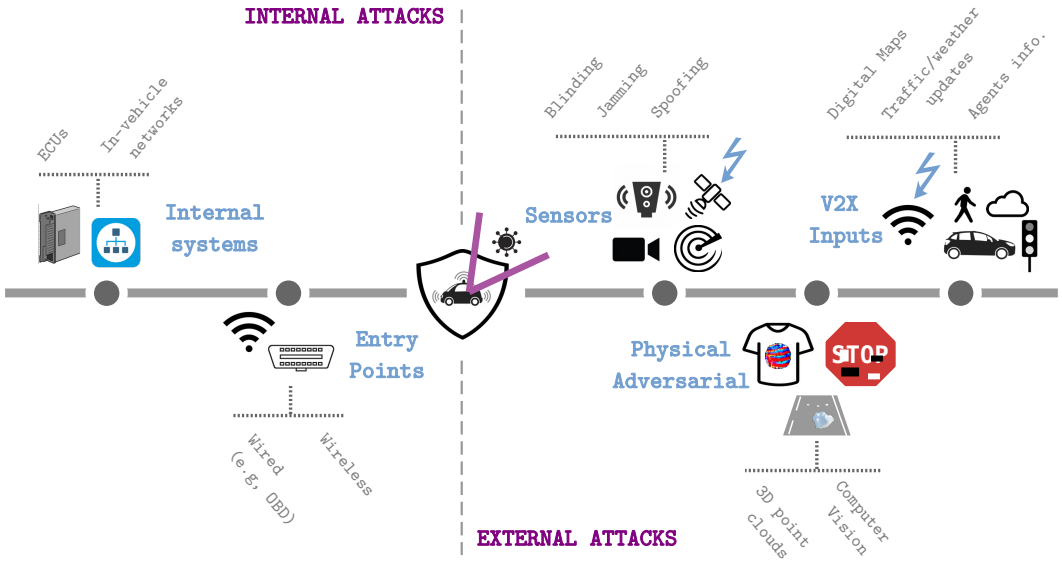
— ***Sensors***: numerous studies have demonstrated the potential vulnerabilities of sensors in AVs. Sensors can be externally jammed (including blinding or saturation), and spoofed. These attacks can target cameras (Yan et al., 2016), (Nassi et al., 2019), (Yadav and Ansari, 2020), radars (Yan et al., 2016), (Yeh

et al., 2016), ultrasonic sensors (Yan et al., 2016), (Xu et al., 2021), LiDARs (Shin et al., 2017), (Cao et al., 2019), (Sun et al., 2020) and GNSS sensors (Gross and Humphreys, 2017). The consequences can be catastrophic, affecting the localization and dynamic scene understanding layers (e.g., false positives and false negatives), and compromising any decisions made by the local path planning layer.

— ***V2X Information***: connectivity plays a key role to enhance autonomy. By means of V2X communications the AV can receive fundamental information from the infrastructure, including digital maps, traffic and weather updates, traffic signals states, as well as current and future state information from other agents (e.g., other vehicles, VRUs). This connectivity can be jammed (including blinding or saturation), and spoofed (Alnasser et al., 2019), (Ghosal and Conti, 2020), leading to multiple and critical issues. If data from the infrastructure (maps, traffic, etc.) or from other agents (positions, intentions, etc.) are not available or are corrupted, most of the layers become compromised (localization, dynamic scene understanding, global and local path planning and user interaction).

— ***Physical Adversarial***: several studies have proven that machine and deep learning methods used in the dynamic scene understanding layer are vulnerable to adversarial examples which results from small perturbations added to the input. Externally, attackers can exploit these vulnerabilities in many different ways. For example, by adding some stickers on traffic signs (Eykholt et al., 2018) or on the road (TKSL, 2019), by placing a printed pattern into real scenes to degrade optical flow performance (Ranjan et al., 2019), or by placing a 3D-printed adversarial object on the road which is not detected by LiDAR-based perception systems (Cao et al., 2019). These adversarial attacks with physical objects can cause serious problems in perception tasks, i.e., in the localization and dynamic scene understanding layers, compromising the decisions of the path planning module.

Figure 10: Taxonomy of internal and external attacks to AVs.



The second set of potential attacks can be referred to as ***internal attacks***, i.e. attacks involving access to some of the internal systems of the AV. For that purpose, the attacker needs some access or entry point, which can be wired or wireless:

— ***Wired entry point***: physical access via a wired connection to one of the available ports on the vehicle, e.g. On-board Diagnostic Port (OBD-II), USB, or the charging port on electric vehicles (El-Rewini et al., 2020). Obviously this type of access requires physical interaction, which exposes the attacker and makes the attack explicit.

— ***Wireless entry point***: wireless access from various V2X communication systems, including dedicated short range communications links (e.g., IEEE 802.11p), cellular V2X (such as 3GPP or 5G NR C-V2X) and Bluetooth systems (Pham and Xiong, 2020). This entry point increases the difficulty of detecting the attacker.

Once the attacker has access to the internal systems, there are two main targets subject to different forms of attacks:

— ***In-vehicle networks***: all the electronic control units, sensors and actuators, are mainly connected through the Controller Area Network (CAN) bus, although there are other communication technologies such as Local Interconnect Network (LAN) and FlexRay (Kim et al., 2021). With access to the in-vehicle networks, the attacker can compromise the integrity of all internal messages, send unauthorized messages, and access the electronic control units (Pham and Xiong, 2020). The consequences are completely catastrophic, affecting all layers of the AV.

— ***Electronic Control Units (ECUs)***: they are embedded electronic systems that control the subsystems of the vehicle, including automated driving functions and layers (Dibaei et al., 2020). The fact that the attacker manages to reach the level of the ECUs can be considered as the most critical case, in which the integrity of the entire system is compromised (Pham and Xiong, 2020), and it is possible not only to sabotage all the autonomous systems, but to have total control of the vehicle which is one of the biggest concerns of the users (Garidis et al., 2020). The consequences of this are tremendously catastrophic.

With the advancement in the study of vulnerabilities, there has also been a great development of new **defensive methods** for the prevention and detection of attacks, as well as for the minimization of their impact. New security frameworks are being proposed addressing important requirements such as *authentication*, *integrity*, *privacy* and *availability* (Dibaei et al., 2020). Some of the existing defences against the attacks include the use of *cryptography*-based algorithms to enhance security for vehicular networks, signature- and anomaly-based *intrusion detection systems*, and *software vulnerability* and *malware detection* (Dibaei et al., 2020), (Kim et al., 2021). Other cybersecurity good practices (non-technical) include policies such as *security* and *privacy by design*, *asset*, *risk and threat management*, as well as organizational measures such as *relationships with suppliers*, *training and awareness*, *security* and *incident management* (ENISA, 2019).

As stated by a joint ENISA-JRC report on the cybersecurity risks of AI in autonomous driving (Dede et al., 2021), in parallel with the development of adversarial attacks, defensive measures have been developed to make these systems less vulnerable. Countermeasures include the use of *sensor* and *hardware redundancy* mechanisms, *hardening against adversarial examples*, and *authentication* between the infrastructure and the vehicle.

Recently, the role of ***resilience*** has been gaining attention as an important enabler for safety and security of autonomous systems (Johnsen and Kilskar, 2020). An autonomous system can be considered *resilient if it can adjust its functioning prior to, during, or following events (changes, disturbances, and opportunities), and thereby sustain required operations under both expected and unexpected conditions* (Hollnagel, 2016). Therefore, resilience can be achieved by applying the defensive security methods and countermeasures mentioned above, but it also needs other important safety measures such as *fault-tolerant* methods (Realpe et al., 2016), (Venkita et al., 2020), *fail-x* approaches such as fail-aware (García-Daza et al., 2020), fail-safe (AutoDrive, 2020) and fail-operational (Matute-Peaspan et al., 2020), and *self-healing* technology (HERE, 2017), (Flaherty, 2020).

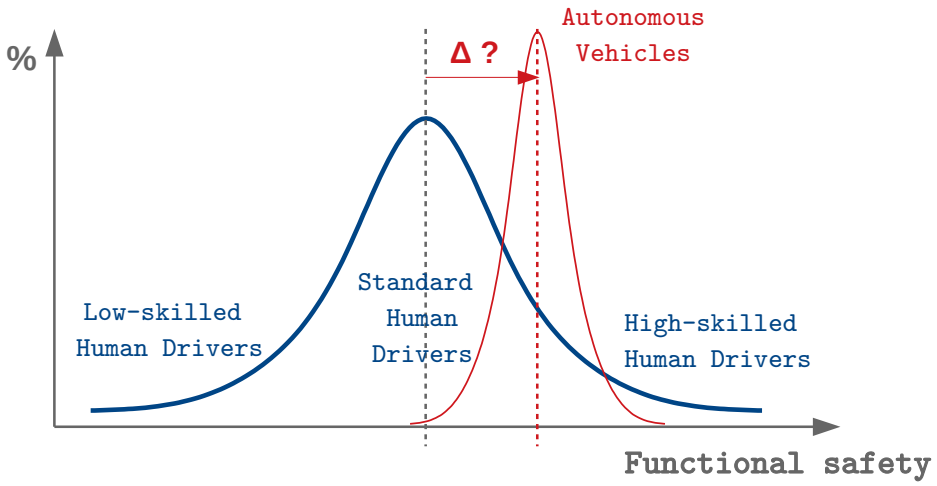
Considering that it is impossible to achieve complete cybersecurity for AVs, as cyberattacks can always occur, and some of them will be successful, an effective strategy targeting end users is to educate them to accept their existence and prepare to react to their consequences, taking into account human factors (Linkov et al., 2019). The role that the key requirement 1 (KR1 Human agency and oversight) can play in this respect is also crucial.

5.2.2 General safety

It is safe to say that general safety is the most relevant variable for the adoption of AVs. It has been clearly identified as the main concern for user acceptance (Garidis et al., 2020) and it is by far the most important requirement for technology developers and policy makers. For example, as stated by the German Ethics Commission on Automated and Connected Driving in its 2017 report (German Ethics Commission, 2017), *the primary purpose of partly and fully automated transport systems is to improve safety of all road users* and *the protection of individuals takes precedence over all other utilitarian considerations*. However, the question about ***how safe is safe enough for AVs*** (the value of Δ in Fig. 11) remains uncertain. The results of available studies show a clear trend on the part of consumers, who consider that AVs must be much safer than the average driver. But on how much safer there is no clear answer, with values ranging from $75\% - 90\%$ (Kalra and Groves, 2017) up to two orders of magnitude (Liu et al., 2018).

The safety threshold at which potential users would be willing to accept AVs may be inflated due to various psychological factors (Shariff et al., 2021), such as *illusory superiority* (or *better-than-average effect*) and *algorithm aversion*. In fact, considering that AVs also bring other benefits in addition to increased safety, such as new mobility services for more people or the freeing up of urban public spaces, it is worth asking whether it **would be possible to accept a safety value equal to that of the average driver, or just a slightly better**. Even a 10 percent safer than the average driver can involve hundreds of thousands of lives saved (Kalra and Groves, 2017). Therefore, it may be highly advisable for both industry and policy makers to calibrate the message about the improved safety that AVs can bring, to counterbalance public opinion biases.

Figure 11: Safety distribution for low-skilled, standard and high-skilled drivers, and for AVs.



One of the most relevant questions in this case is **how to measure the safety** of AVs. One possible way to assess safety is to test AVs in real traffic conditions, evaluate their performance in terms of fatalities and injuries, and make statistical comparisons with respect to human driver performance. As demonstrated by a well-known study (Kalra and Paddock, 2016) this approach is unpractical since it would require AVs to be driven hundreds of millions, or even billions, of kilometres which would take tens, or even hundreds, of years. **New innovative methods** to assess the safety of automated and autonomous driving functions are required.

Safety of conventional vehicles is certified through classical approaches, where different physical tests are set up on test tracks or benches to assess the required safety level using various performance criteria. These approaches have been traditionally applied for *vehicle type approval* (homologation) or *self-certification* procedures (Martins, 2010), and are well suited for components, systems and vehicles with limited complexity and limited interactions with other entities. However, as the complexity of the components, systems or vehicles increases (e.g., from traditional braking to Anti-lock Braking Systems or Electronic Stability Control), classical approaches are not able to deal with all relevant safety areas due to the high variability of potential scenarios. This led to the introduction of simulation-based safety oriented audits as a way to complement the physical testing of the systems (Lutz et al., 2017).

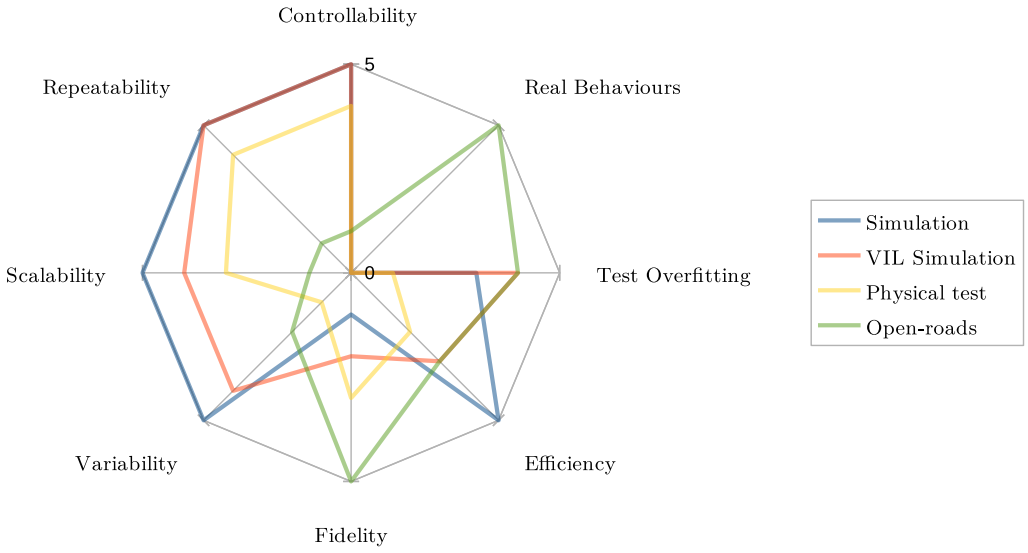
With the introduction of assisted, automated and autonomous driving systems (Schram, 2019) the overall complexity increased in terms of the number of software functions, variants of scenarios and interactions, and potentially affected safety areas. New certification approaches are needed, not only for future vehicle safety regulatory frameworks, but also for assessments under current exemption procedures (Galassi and Lagrange, 2020). Several national and international regulatory and standardization initiatives and projects are already underway to address this problem (Baldini, 2020).

One of the most solid regulatory proposals is being developed by the Working Party on Automated/Autonomous and Connected Vehicles (GRVA) of the the UNECE World Forum for Harmonization of Vehicle Regulations (WP.29). Following the same approach as projects such as AdaptIVe (AdaptIVe, 2017) or PEGASUS (PEGASUS project, 2019), the GRVA proposes a testing framework based on three main pillars that must be assessed together (UNECE WP29 GRVA, 2019). First, *audit and assessment* mainly based on ***simulation*** to cover all type of scenarios, but especially edge case scenarios difficult to occur in real-world traffic. Besides the traditional methods of simulation based on software-in-the-loop (SIL), we can also identify more sophisticated approaches based on hardware-in-the-loop (HIL), or even vehicle-in-the-loop (VIL), such as the testing platform proposed in ENABLE-S3 project (ENABLE-S3 project, 2019) to combine both simulation and ready-to-drive vehicles using a chassis dynamometer and on a power-train testbed (AVL, 2021). Second, ***physical tests*** to assess critical scenarios, performed in controlled environments on test tracks (closed-roads), and involving sophisticated equipment such as lightweight global vehicle (Euro NCAP, 2018), articulated pedestrian (ACEA, 2015) and bicyclist (ACEA, 2018) targets. And last but not least, ***real-world test drive***, which is devised as a "driving license test" for automated and autonomous driving systems to assess the overall capabilities and behaviour of the vehicle in non-simulated traffic on public/open roads. This approach based on the implementation of these three pillars has been the one recently adopted by United Nations to regulate the approval of Automated Lane Keeping Systems (ALKS) (UNECE WP.29 GRVA, 2021c).

Although most frameworks consider the case of VIL simulation as a subfield of simulation-based approaches, it can also be considered separately as an intermediate approach between pure simulation (SIL) and physical testing on closed roads. These four approaches have strengths and weaknesses (Thorn et al., 2018), which is why

it is important to implement them holistically (UNECE WP29 GRVA, 2019). In Fig. 12 we illustrate the advantages and disadvantages of all testing approaches by means of a net diagram. The data have been obtained from a qualitative analysis taking into account eight different parameters (i.e., *controllability*, *repeatability*, *scalability*, *variability*, *fidelity*, *efficiency*, *test overfitting* and *real behaviours*) and a numerical range between zero and five. As can be observed, the methods are somehow complementary. For example, although simulation-based testing allows full controllability, repeatability and variability in a very efficient way, they exhibit very low fidelity and lack real-world behaviours. We can increase fidelity at the cost of increasing complexity and thus decreasing efficiency, from VIL simulation to physical testing on closed tracks. But the absence of real behaviours remains a problem, which can only be partially compensated by testing in real traffic conditions (open-roads).

Figure 12: Net diagram to illustrate the main features of the different testing approaches.



Note 1: Test overfitting refers to the degree to which the systems can be optimized on specific test scenarios. A high score means a low probability of overfitting.
Note 2: Real behaviours refers to the degree to which the test method can include actual behaviours of other agents (pedestrians, cyclists, other drivers, etc.)..

Another relevant variable refers to the degree to which the driving functions can be optimized on the specific scenarios, which can be seen as a shortcut by OEMs to overfit the performance of their systems to the test scenarios. This has a negative impact on the possible fidelity of the tests, while the performance of the systems in real traffic remains unknown. In general, if the simulation conditions are known a priori, or the physical test conditions in closed tracks, or the proving grounds or the test area in real traffic, all test methods are potentially subject to overfitting. Still, there are some differences. For example, on the one hand, the uncontrolled conditions of open road testing make this method less prone to overfitting. On the other hand, the low variability and the strict control and repeatability conditions of the scenarios in the physical certification on closed roads are favourable conditions for the optimization of the systems to the proposed scenarios.

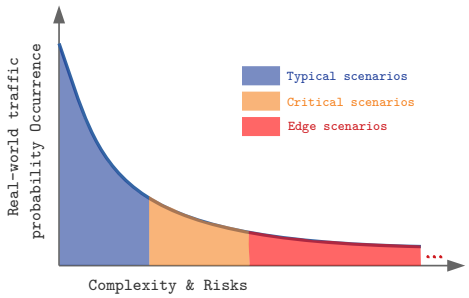


Figure 13: Types of scenarios, probability of occurrence in real-world traffic, complexity and risk. Long-tail distribution.

A closer look to the different methods reveals that complementarity is limited, as the scenarios addressed by each approach are of a different nature. In Fig. 13 we illustrate three different types of scenarios (typical, critical and edge) referring to their probability of occurrence with respect to the degree of complexity and potential risks. As can be observed the distribution of scenarios follows a long tail shape, which requires significant scale

to discover and properly handle the long tail of rare events (Jain et al., 2021). In Table 13, the type of scenarios that can be addressed for each testing method is depicted. As can be inferred, high fidelity is only achievable for typical scenarios, with higher uncertainty for critical and edge scenarios. In addition, current testing approaches do not allow to assess safety with real behaviours for critical and edge cases. This is particularly relevant for automated and autonomous driving functions that make use of predictive perception, i.e., systems that learn and model the behaviours and interactions of traffic agents to anticipate future actions and motions to be considered in the path planning layer. These predictive systems are expected to enable autonomous driving to become more like manual driving, increasing safety margins, reducing risks, and providing smoother and more acceptable motion trajectories.

Approaches	Typical	Critical	Edge
Simulation	✓	✓	✓
VIL Simulation	✓	✓	✓
Physical track		✓	
Open-roads	✓		

Table 13: Distribution of scenarios by testing approach.

In general, we can identify the following challenges ahead for including trustworthy AI requirements and enhance safety certification systems for AVs:

— *Incorporate real-behaviours both in simulation and physical tests*. The use of immersive virtual reality can be a good choice to add realistic behaviours to simulated environments. For physical tests on closed roads, articulated dummies should be even more sophisticated, including more realistic head movements and body language. The behaviours of the vehicle, bicyclists and pedestrian targets should be based on real behaviours representative of the specific scenario, linked to the variables affecting the behaviour of traffic agents such as street layout, traffic conditions, etc.

— *Increase variability using random initial conditions in physical tests*. If we accept open road testing, where controllability and repeatability are not possible, as an acceptable method for assessing the safety of the automated and autonomous driving systems, we can introduce random initial conditions in the different scenarios to enhance variability and avoid test overfitting, at the cost of decreasing repeatability. These tests with random initial conditions can be complementary to the current tests with fixed initial conditions.

— *Introduce requirements for cybersecurity certification*. Incorporate various types of cyberattacks both external (sensors, communications and physical adversarial) and internal (in-vehicle networks and ECUs) and evaluate the response of the systems.

— *Develop new scenarios to assess human agency and oversight criteria*. Include scenarios where driver or passenger users interact with the vehicle (e.g. request to intervene cases), but also external users to assess the safety of interactions with them.

— *Generate new scenarios and metrics to assess transparency and fairness*. Incorporate new scenarios and metrics to be able to evaluate the degree of user understanding of decisions. In addition, establish new approaches to introduce a higher diversity of end-users in the simulation environment or in physical tests to assess the potential bias of the automated and autonomous systems.

5.2.3 Accuracy

The level of accuracy of the AI systems of AVs can be developed from multiple metrics set at at least three different levels (we depict an overview of the taxonomy of metrics for AVs in Fig. 14). First, from the perspective of **overall vehicle performance**. For this purpose, the relationship between distance travelled and failures is usually considered. For example the *per-kilometre/mile failure rate* (Kalra and Paddock, 2016), expressed in failures per unit of distance. But what is a failure from a global behavioural point of view?. One common approach is to consider the *number of manual disengagements or interventions* needed during autonomous driving (Paz et al., 2020). The other obvious approach would be to consider a failure as the cause of an accident, i.e., the *accident rate* expressed in accidents per unit of distance.

Second, we can define different metrics and levels of accuracy for each individual layer (Fig. 3). Thus, ***localization accuracy*** has been traditionally measured by using metrics such as *absolute and relative displacements/pose errors* including *translational* (in meters or in percentage) and *rotational* (in degrees and degrees per meter) errors (Kümmerle et al., 2009). More sophisticated metrics have been proposed to evaluate the quality of estimated trajectories (Zhang and Scaramuzza, 2018).

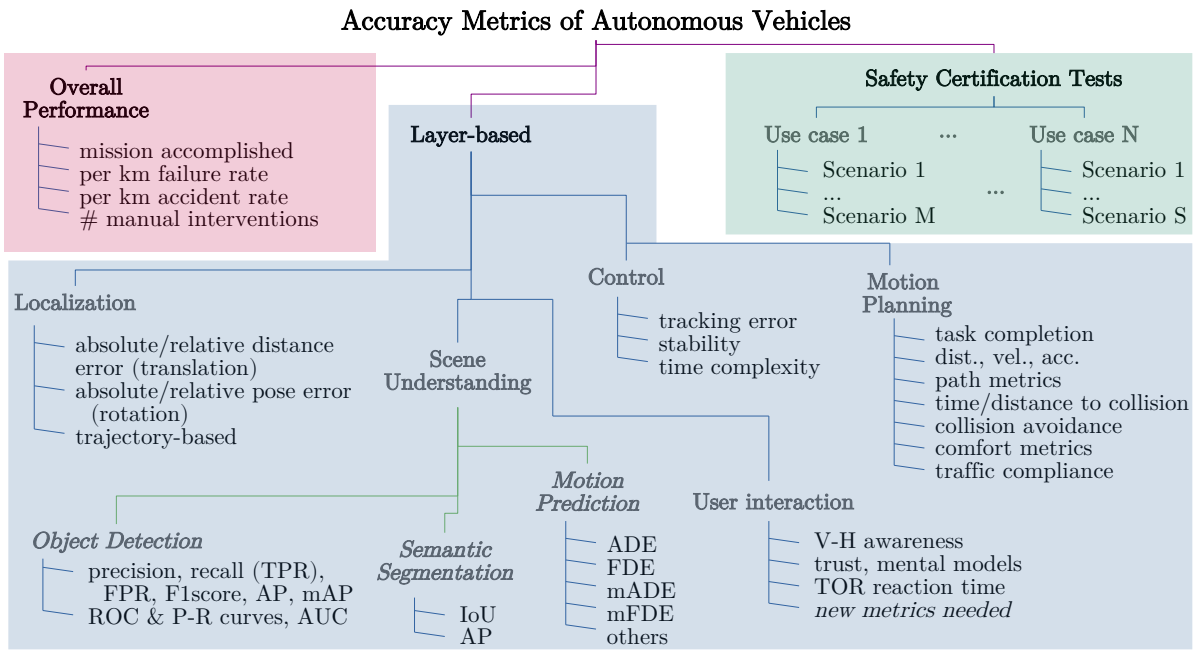


Figure 14: Taxonomy of metrics to evaluate the accuracy of AVs.

Dynamic scene understanding accuracy involves several metrics depending on the specific task. For example, the most popular metrics used when measuring the accuracy of **object detection** systems (Padilla et al., 2021) are *Precision*, *Recall* (True Positive Rate, TPR), *False Positive Rate* (FPR), *F1 score*, *Average Precision* (AP) and *mean Average Precision* (mAP). All these metrics are based on different combinations of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), which are set according to the *Intersection over Union* (IoU) which defines the overlap between two regions, i.e., the predicted and the labelled from the ground truth. When object detection provides probabilities, the decision can be made on the basis of certain thresholds that define the balance between hits and misses. Curves that relate two of these metrics visually, such as the *Receiver Operational Characteristic* (ROC) curve or the *Precision-Recall* curve (Davis and Goadrich, 2006), are often used for this purpose (ROC curves are appropriate when samples are balanced between each class, whereas Precision-Recall curves are appropriate for imbalanced datasets). The curves of different methods can be compared and the *Area Under the Curve* (AUC) can be used as the summary of the model performance. For **pixel-level semantic segmentation** performance is generally assessed using the standard *Jaccard Index*, a.k.a the PASCAL VOC intersection-over-union metric (Everingham et al., 2015) IoU = TP / (TP+FP+FN), where TP, FP, and FN are the numbers of true positive, false positive, and false negative pixels respectively. For **instance-level semantic segmentation**, which combines traditional object detection and pixel-level segmentation, the *Average Precision* (AP) on the region level is usually applied, averaging it across a range of overlapping thresholds to avoid bias towards a specific value (Lin et al., 2014). And for **motion prediction** the most common metrics are *Average Displacement Error* (ADE) and *Final Displacement Error* (FDE) with their probabilistic counterparts *minimum ADE* (mADE) and *minimum FDE* (mFDE) (Rudenko et al., 2020). Recent work highlights the need for new metrics that take into account the evaluation of the joint tasks of detection, tracking, and prediction better correlated with full-system behaviour (Traft et al., 2020).

It is important to note that the level of accuracy of all layers in general, but scene understanding in particular, usually involves an implicit **policy decision**, which is to define the working point at which a perception system should be set, as a trade-off between detection capability and the generation of false positives (or positive predicted value). For example, typical ROC and Precision-Recall curves are shown in Fig. 15. The shape of the curves represents the state in which a perception system can be left using different probability thresholds. Thus, for example, the ideal point of an ROC curve would be given by a TPR=1.0 and an FPR=0.0. However, as can be seen, the higher the TPR, the higher the FPR and vice versa. In the case of Precision-Recall curves, the ideal point would be Precision=1.0 and Recall=1.0, but in this case the higher the Precision the lower the Recall and vice versa. The decision to fix the behaviour of the system at points A or B (or C or D) is technical in nature, but it is also a policy decision.

Take the case of pedestrian detection:

— On the one hand, setting the system behaviour at point A would result in a very low number of false positives, but a very low detection rate as well. This would lead to a very low number of false activations of

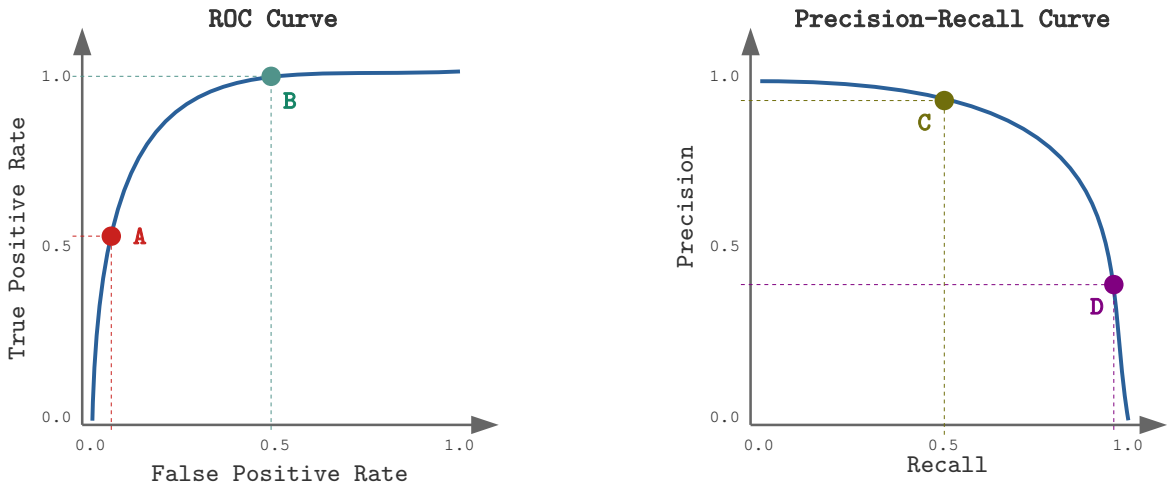


Figure 15: Typical ROC (left) and Precision-Recall (right) curves. The shape of the curves implies the necessary trade-off between each pair of metrics, which is exemplified by two different working points.

safety systems (e.g. Emergency Brake Assists), but also to the non-detection of a considerable percentage of pedestrians resulting in possible collisions.

— On the other hand, setting the working point to B would imply a very high detection rate, but a very high false positive rate as well. This would prevent many collisions, but would result in a very high number of erroneous activations of safety systems, which would negatively affect user acceptance and could lead to other types of accidents (e.g. rear-end collisions).

In this case, **policy makers** should consider which variable is more important to prioritise: the number of potentially avoidable pedestrian collisions or the number of erroneous activations of pedestrian protection active systems. This case can be extended to other perception problems such as vehicle detection, free-space detection, traffic lights recognition, etc.

When assessing the ***accuracy of motion planning techniques*** we have to take into account that motion planning also involves decision making, so the planned trajectories have safe and comfortable, but they also have to comply with traffic rules. We can identify metrics on the *tasks completion success rate*, which involve reaching the goal location within a set amount of time (Ilievski, 2020). Safety metrics are often considered using surrogate or proximal safety metrics and considering other agents, including *Euclidean distance*, *relative velocity and acceleration*, *path metrics*, *time and distance to collision*, and *collision avoidance*. The *feasibility of the planned trajectory* based on vehicle dynamics in different scenarios is also considered (Pek et al., 2020). *Comfort metrics* capture the level of comfort in the realization of the planned trajectories including velocity, acceleration and jerk (Ilievski, 2020), (Paschalidis et al., 2020). *Traffic compliance metrics* include lane and road violations, and other regulatory infractions such as speed limits or running a red light (CARLA, 2020). In addition, there are additional qualitative characteristics to be taken into account when assessing the accuracy of planned trajectories, such as comparing them with those of a human driver (Chong et al., 2020).

The control layer ensures that the vehicle's longitudinal and lateral motion follows a reference trajectory or path (including a reference velocity) by providing the required steering and acceleration/braking inputs. The most common quantification metric to evaluate the controller's performance is the *tracking error*, i.e., the difference between the planned path and the actual vehicle position, including *lateral*, *longitudinal* and *orientation* (or yaw) errors (Amer et al., 2017). These are usually root-mean-square errors measured over paths of varying curvatures (Chong et al., 2020). The *stability* and the *time complexity* of the controllers are also fundamental issues to be considered when evaluating their performance and robustness (Paden et al., 2016).

As for the ***user interaction layer*** there is not yet a clear taxonomy of metrics that considers all possible interactions, including internal (HVI) and external (eHVI) users of the AV. Traditionally, metrics related to human-machine interaction are domain and application specific (Steinfeldet et al., 2006), and they cover three main areas (Pina et al., 2008). First, metrics related with the human behaviour, including subjective metrics such as physical and mental workload and discomfort, emotional state, self-confidence, fatigue or situation awareness. Second, metrics focusing on the behaviour of the autonomous or robotic system, which, in this case, are the metrics described above. And finally, the metrics of communication, interaction and collaboration between humans and the autonomous system, which is also referred to as human-machine teaming (Damacharla et al., 2018). For example, *autonomous platform - human awareness*, which evaluates the degree to which the system is aware of all the variables related to humans (in this case, mainly the driver and/or passengers, but

also external road users) and can can be evaluated summing the number of *awareness violations*, or different metrics related to *user trust* (Azevedo-Sa et al., 2020) and *mental models* (Beggiato et al., 2015). For the specific context of user interaction with AVs, the only area where we find a somewhat more solid metrics base is in the management of requests to intervene or take-over-requests (TOR) situations (Kim and Yang, 2020). The most commonly used parameter is the *reaction time* (Wintersberger et al., 2017). In addition, some HMI metrics are already in use in Euro NCAP test protocols for Autonomous Emergency Braking (AEB) systems, including *collision warning systems*, and reversible pre-tensioning of the belt in the pre-crash phase (Euro NCAP, 2021). But in broad terms, we can conclude that **a taxonomy of new metrics** is needed to evaluate the performance of the various systems for the **interaction between AVs and users** in general (backup drivers, passengers and external road users).

The third and final level from which the accuracy of AI systems of AVs can be developed relates to the **metrics used in safety certification testing**, such as those described in subsection 5.2.2. These metrics are specific to each *use case* or *traffic environment* (e.g., highway traffic, urban traffic, interurban traffic, etc.) and *test scenario* (i.g., obstructed pedestrian crossing the street, emergency braking before the tail end of a traffic jam, etc.). For example, if we analyse the assessment protocols that Euro NCAP has defined for AEB Car-to-Car (Euro NCAP, 2021) or AEB VRU (Euro NCAP, 2020) systems, we observe that multiple variables and metrics are necessary to evaluate the performance of these systems and that they are specific to each scenario. Time to collision, AEB activation times, speed profiles, the impact of a non-avoidable collision, among others, are measured. Through a quantitative and qualitative system of points provided according to the performance of the systems in each test carried out, a final grade or score is provided that summarizes the overall accuracy of the system.

5.2.4 Reliability, fallback plans and reproducibility

Requirements referring to repeatability, reliability and reproducibility of automated and autonomous driving systems need to be addressed through the aforementioned safety certification tests for all modalities, being easier to be ensured in simulation and physical tests on closed tracks, and more difficult and challenging when testing in real traffic conditions. For example, the test protocols for AEB systems (Euro NCAP, 2021), (Euro NCAP, 2020) are being developed to create a standardised set of conditions that would enable the objective, repeatable and reproducible assessment of AEB systems, allowing the reliably quantification of their performance (Hulshof et al., 2013).

Regarding fail-safe fallback plans, as described by SAE International standard, depending on the level of automation, after the occurrence of a dynamic driving task (DDT) performance-relevant system failure or upon operational design domain (ODD) exit, appropriate mechanisms must be put in place in order to allow the (backup) driver to **resume manual control** (levels 1, 2 and 3) or to endow the automatic driving system with the capability to achieve a **minimal risk condition** (levels 4 and 5). The SAE International standard does not define the requirements and acceptance criteria of what is a minimal risk condition (ADB, 2019). Indeed, given the diversity of road conditions, an absolute "minimum" cannot be established outside of a highly defined scenario or set of circumstances (UNECE WP.29 GRVA, 2020a). These conditions will depend on the type of failure, the use case and the specific scenario, which entails a huge variability. In fact, the number of research works related to fallback strategies to achieve minimal risk conditions in automated and autonomous driving is still very limited. This can be verified by analysing the related works section of the few relevant papers that are available (Emzivat et al., 2017), (Zue et al., 2018) and (Yu and Luo, 2019). Similarly, the state of progress regarding **testing procedures for assessing the safety of fallback plans** is still in its infancy and requires new proposals and approaches.

Finally, it is important to highlight the fact that multiple continuous, **online or adaptive learning approaches** have been proposed to address different problems (i.e., layers) of autonomous and automated vehicles. On the one hand, this continuous adaptation of AI systems focuses on improving their performance being the use of reinforcement learning the most common approach (Kiran and et al., 2021). Training autonomous driving systems with reinforcement learning in real environments entails unacceptable costs and risks of trial and error. Therefore, the role of **autonomous driving simulators** is of utmost importance, with the gap between the virtual and the real environments being the key issue to be solved (Pan et al., 2017). On the other hand, online learning also focuses on adapting to new situations, such as, for example, re-training of perception algorithms to consider the new reality of pedestrians systematically wearing masks (Marko et al., 2021). Continuous learning can be applied in an offline fashion, getting new data and then update the corresponding component. But it can also be applied online during the operation of the AI component. The latter option (online), although feasible, involves a very high risk as it could compromise the safety of the component. Even by using safety objectives to shape the adaptation of the system or safety supervisors to monitor the behavioural change of the component during its operation, **continuous learning while in operation must be avoided**. When applied offline, continuous learning will generate updated AI-based models (i.e., software updates) that may affect the fulfilment of any of the relevant requirements of the type approved systems. **Any substantial update to an**

AI-based component that may modify its technical performance must be considered by the approval authority to **require further testing** to ensure that the modifications made still comply with the requirements (UNECE WP.29 GRVA, 2021b).

5.3 Privacy and data governance (KR3)

Privacy is closely linked to the principle of prevention of harm which necessitates adequate data governance. Therefore, this key requirement is developed following two main sub-requirements, privacy and data governance, which are closely related to each other. We first describe the types of data captured and used by AVs, identifying the elements that may compromise the privacy of internal and external users (i.e. personal data). We then address privacy and data governance requirements.

5.3.1 Personal data collected by AVs

The overall performance of autonomous driving systems is based on data from multiple sensors from different modalities, including enhanced high-definition maps, 3D point clouds, images sequences from onboard cameras pointing outside and inside, inertial measurement systems, audio systems for speech recognition, etc. On the one hand, all these data are essential to improve the safety of all the systems in each of the layers of AVs. The **data** is needed to **develop** the systems, **test** them, **use** them and **improve** them once deployed. For example, AI-systems for vehicle localization require very precise maps with detailed information to be obtained a priori, as well as multiple a posteriori sequences with geo-referenced data from cameras, LiDAR or radar, in order to develop the different localization methods. Then, the deployed localization system would require the same type of information in order to obtain the precise global position and orientation of the vehicle. Scene understanding, necessarily requires thousands (or millions) of sequences of data from cameras and other sensors, conveniently labelled, in order to train, validate and test detection and segmentation systems for pedestrians, cyclists, vehicles, etc., including data to model behaviour of road users for predictive systems. In using them, these perception systems need to process online data of the same nature. Human-vehicle interaction systems also require data from multiple sensors to first develop, and then deploy human-machine interfaces, both in-vehicle (HMI) and external (eHMI) user interfaces. As with video surveillance (EDPB, 2020), when data captured by sensors is not stored or transferred in any way, but only processed in real time to obtain non-personal data useful for localization, scene understanding, path planning, etc., the intrusion to privacy is more limited. However, in all cases, these systems are expected to be capable of **storing data while being used for multiple purposes**, such as improving system performance, troubleshooting, post-market monitoring and traceability.

On the other hand, the data clearly contains **personal data**, i.e., information relating to an identified or identifiable living individual, from **external road users** (see Fig. 16) to backup **drivers** and **passengers** (see Fig. 17), who can be identified, directly or indirectly, by reference to the images, location data, vehicle registration marks (combined with geographical, manufacturer, model and colour information), behaviours (e.g., journeys made, traffic infractions, state of attention, etc.), or even health data, physical and mental (e.g., by means of the HMI). Furthermore, the modality type of in-vehicle user interfaces enables the use of biometric data (e.g. face or voice recognition). User identification may even be beneficial to improve the user-vehicle interaction experience. However, in line with the Commission (recital 10, Regulation (EU) 2019/2144), any safety systems of AVs should function without the use of any kind of biometric information of drivers or passengers.

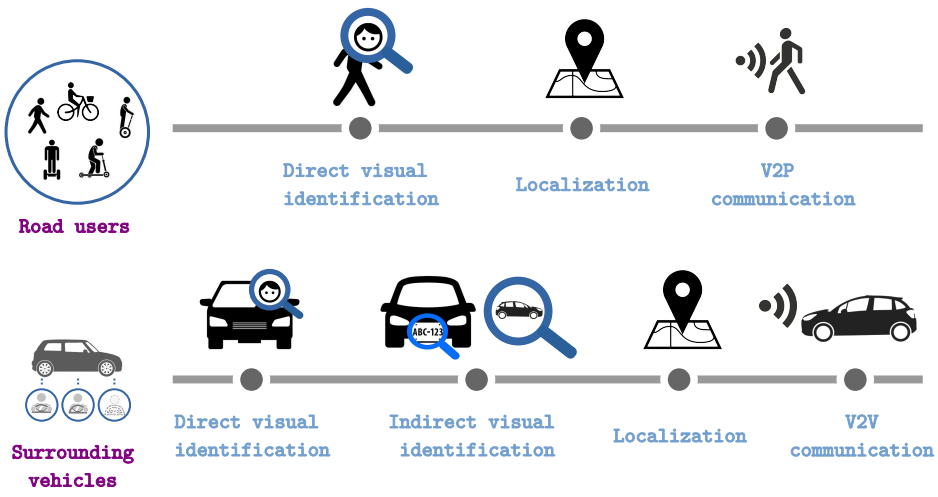


Figure 16: Personal data of external road users and surrounding vehicles processed by AVs.

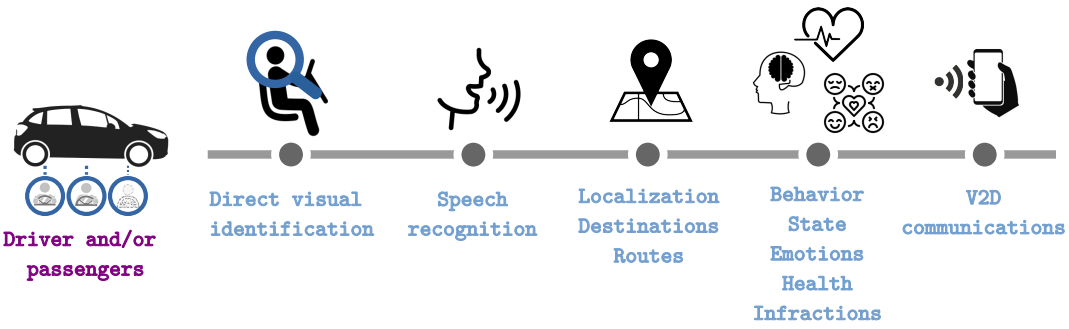


Figure 17: Personal data of occupants (driver and/or passengers) processed by AVs.

A distinction has to be made between data processed to develop the AI systems of the AV so that they meet the safety requirements for type-approval and placing on the market, and data used to improve the AI systems once the AV has been deployed. In the first case, the main legal basis is related to vehicle-type approval requirements. In the second case, data collection and processing must be subject to some legal basis, such as the consent of the in-vehicle user. However, data from the surroundings of the AV environment that include other external road users cannot be obtained on the basis of consent, so other mechanisms are needed to ensure privacy (e.g., privacy-by-design).

Two relevant systems for data collection are the *Event Data Recorder* (EDR), which is not AV-specific, and the *Data Storage System for Automatic Driving* (DSSAD) (Veitas and Delaere, 2018) which are devised to **record data** valuable for **effective crash reconstruction** purposes (also during testing). These are not consent-based systems, but a requirement for placing new vehicles on the market, i.e. requiring the consumer's approval to record and/or access the data and the right to delete the data, would be meaningless for the purposes of the EDR and the DDSAD. In addition, EDR and DDSAD systems are closed loop systems with no output channel until an accident occurs and an investigation is initiated. Before that time, strictly speaking, there is no data controller. However, as stated by the International Organization of Motor Vehicle Manufacturers (OICA, 2019) *continuous recording and storage of video, location, speed, and/or surroundings of a vehicle seems to be contradictory with the regulations addressing privacy protection.* And the German Road Safety Council (GRSC, 2020) also stated that ***requirements for data protection and privacy*** remain open questions **to be clarified for EDR/DDSAD systems**, not only for automated and autonomous vehicles, but also for conventional ones. These issues are important once an investigation has been initiated in the case of an accident. At that point, the controller accessing the data would need an appropriate lawful basis.

5.3.2 Privacy issues and data governance

As stated above, AI systems in AVs are trained or developed using or processing personal data, and not only for the driver and passengers (e.g., camera images, voice recognition systems, location, destinations, preferred routes, behaviours, health status, etc.) but also for external users (e.g., faces, license plates, locations, etc.). Therefore, mandatory measures must be put in place under the General Data Protection Regulation (GDPR), or a non-European equivalent.

According to the definition of personal data given by the GDPR, any information that can identify a person is personal data (Art. 1 under 1 GDPR). The identifiable person is referred to as the *data subject*, which in this case is primarily the user of the AV, but may also refer to other external road users. The *data controller* is the party that determines what data are collected, how they are collected and the purpose. It is reasonable to consider the AV manufacturer as the controller, but depending on the circumstances, the fleet operator (e.g., autonomous mobility services) can also be seen as the controller. Finally, the *data processor*, i.e., the one that processes the personal data on behalf of the controller (Art. 4 under 8 GDPR), can be assigned to either the software developer or the fleet operator depending on the specific circumstances (i.e., if they process the data on behalf of the controller) (Mulder and Vellinga, 2021).

The AV manufacturer, as the data controller, shall designate a Data Protection Officer (DPO) and perform the corresponding Data Protection Impact Assessment (DPIA), including regular Data Protection Audits. These requirements are not inherent to AVs, but also to conventional connected vehicles (EDPB, 2021).

In general, most of the data mentioned above are collected for the purpose of improving the safety of of the system, and therefore, the safety of individuals (e.g., GDPR data subjects). It can be argued that this may be related to the *vital interest of the data subject*, especially in the case of external road users who cannot be informed or asked for consent. However, the concept of *vital interest* is mainly conceived as a fallback when no other grounds are available (e.g., when the data subject is unconscious or under legal guardianship).

In the context of AVs, **the right to data privacy can be seen as opposed to the right to road safety**, and one could argue that road traffic safety should prevail over the right to data protection. The challenge, as stated in (Mulder and Vellinga, 2021), is to develop an appropriate legal framework to **reconcile data protection with other public interests, such as road safety**. Such a legal framework could, for example, impose certain obligations on manufacturers to process certain safety-related data.

Among the various measures to achieve privacy by design or by default, ***data minimisation*** may be not fully applicable at this stage of development. The idea of data minimization supported by data protection frameworks is to ensure that only data necessary for the intended purpose are collected and processed. However, the complexity, variety and dimensionality of the problems addressed by the various AI systems of AVs are so high that we are still far from having reached the point where we can discard some data (or collect data just in case it might be useful later) because they are not relevant for the purpose of safety improvement. Although there are multiple works that attempt to address the question of how much data is enough (Wang et al., 2017) and it seems reasonable to assume that the AI systems should be scrutinised to ensure that only the strictly necessary data have been used, the main problem is that these analyses come after the data have been collected and processed. And moreover, even when it comes to establishing how much data is strictly necessary, there is still no clear upper limit to start from. However, there are other measures that could allow data protection objectives to be achieved without sacrificing the constant need of data of complex AI systems of AVs.

As established by the European Data Protection Board (EDPB, 2021), data rendered anonymous (i.e., ***data anonymisation***) in such a manner that the data subject is not or no longer identifiable, may be a good strategy to keep the benefits and to mitigate the risks in relation to AVs. Once a dataset is truly anonymised and individuals are no longer identified, European data protection law no longer applies. In other words, the principles of data protection do not apply to anonymous information.

One of the most common approaches to guarantee anonymity is to perform *automated blurring of faces, bodies and license plates* on images capture by front, rear and side cameras. For example, this has been the classic approach applied in the Google Street View tool to meet privacy criteria on a large scale (Frome et al., 2009). Image blurring is just one of the many possible techniques among others such as pixelation, mosaic, cartooning, masking, warping, morphing, etc. (Asghar et al., 2019). Other approaches focus on automated detection and removing of dynamic objects (e.g., pedestrians, vehicles, etc.). Detected moving objects are removed and regions are inpainted with information from other views to obtain realistic images in which the objects are no longer visible (Uittenbogaard et al., 2019). All these approaches may be sufficient and appropriate for some perception systems, such as vision-based localization systems, where dynamic object information is not relevant (the features used to estimate the vehicle pose usually correspond to static objects (Parra Alonso et al., 2012)) and these can be altered (e.g., blurred, pixelated, etc.) or even removed and inpainted without major consequences. However, these approaches are not suitable for tasks such as pedestrian and vehicle detection, as they corrupt the nature of the data, and may negatively affect the generalization capability of the learning process. The same applies in cases where it is necessary to detect gaze direction (Lorenzo et al., 2020), body pose (Quintero Mínguez et al., 2019) or intrinsic attributes of the agents such as sex, age (Brandao, 2019), skin tone (Wilson et al., 2019) or even emotional expressions (Gallup et al., 2014) (see Fig. 18). The most important challenge is to **anonymize the data while preserving relevant information for perception and HMI systems**. This is particularly important for predictive perception, i.e., for predicting the actions, including intention (Rasouli et al., 2019), and motions of road agents.

This is the main objective of ***image de-identification*** methods which attempt to replace directly identifying characteristics such as faces (Jourabloo et al., 2015) or license plates (Du and Ling, 2011), with synthesized and realistic features, while still preserving relevant non-identifying attributes. For example, license plates numbers can be replaced by using a different, but realistic, number, as depicted in Fig. 19. If other unrealistic transformations are applied to de-identify license plates, vehicle detection systems may learn synthetic features that will not exist in real conditions, with the risk of obtaining unpredictable results as well as reduced performance.

Generative Adversarial Networks (GAN) are well suited for de-identification since they can produce natural-looking synthesized images of any given object using adversarial training, and a considerable number of approaches have recently emerged such as Privacy-Protective-GAN (Wu et al., 2019), AnonymousNet (Li and Lin, 2019), DeepPrivacy (Hukkelås et al., 2019) or AnonFACES (Le et al., 2020). When these techniques are applied to de-identify vehicles, the transformation may involve specific requirements to preserve some attributes and replace some others (see Fig. 20). Similar techniques can be applied to avoid voice identification in speech processing interfaces (Jin et al., 2009), (Magarinos, 2019).

An additional measure to achieve privacy by design is to guarantee that all data, storage devices and V2X communication channels are **encrypted** by means of state-of-the-art algorithms. As suggested by the EDPB for (conventional) connected vehicles (EDPB, 2020), a **unique encryption key management system** should be established for each vehicle, including regular renewal of encryption keys.

It is also important to analyse the role of consent in the processing of personal data for AVs. Consent constitutes a lawful basis for processing personal data (Art 7. GDPR). Consent to the processing of personal

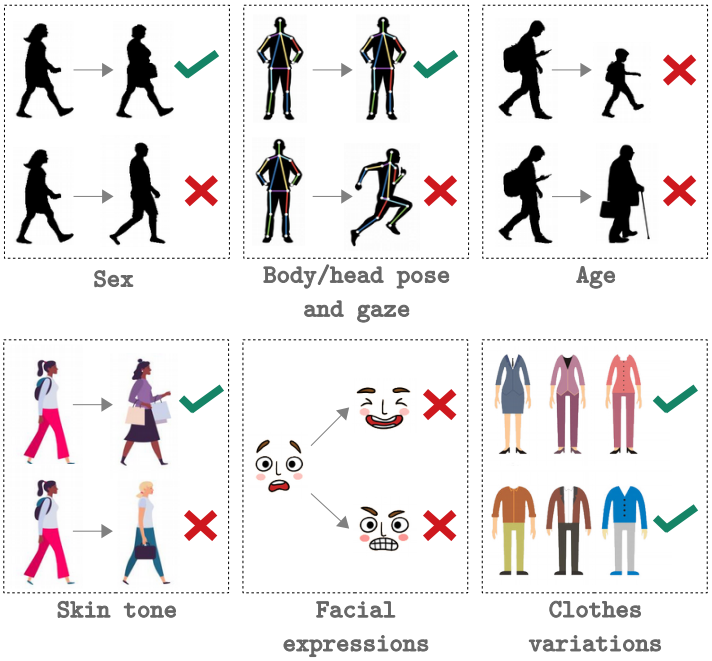


Figure 18: Attributes that must be preserved for image de-identification for the pedestrian and driver/passenger use cases. These attributes can be relevant when modelling human behaviours.

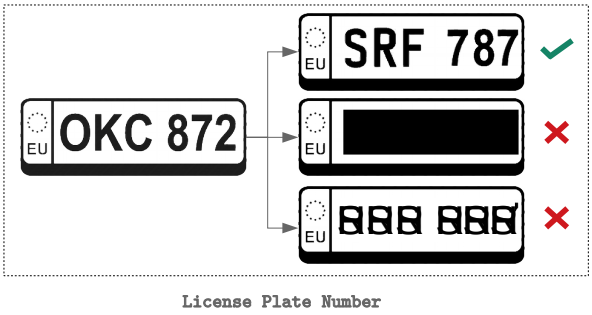


Figure 19: Image de-identification applied to license plates. The most appropriate approach to maintain a realistic transformation would be to replace the license plate number by a random number, while maintaining its structure.

data in AVs involves several issues, type of users and dimensions that have to be considered. First, the **consent model should not involve any safety risk** to drivers or passengers on board (Gaeta, 2019). Therefore, any approach where the requirement to give consent is applied during the trip should be avoided, in particular for automation level 3 where the backup driver must be able to resume control of the vehicle in case of emergency. The least complex case is that of in-vehicle users, i.e., (backup) driver and/or passengers, depending on the level of automation, where the consent model can be materialized earlier in various ways (e.g., from the smartphone or from the user interface with the vehicle) before starting the trip. **Consent must include the exchange of certain data with other vehicles and with the infrastructure to enable V2V and V2I communications systems**, and it could also include data processing by other AVs that have data processing systems of the same nature. This scenario would require agreements and homogeneous consent models among all AV data controllers, but it would allow the conditions of consent to be met for all users of CAVs, without detriment to the application of the aforementioned privacy by design measures (e.g., encryption, de-identification, etc.).

Another question to be resolved is whether **consent** can here be posed as a prerequisite to ownership or use of the AV. One may argue that since consent is only valid if it is freely given, it cannot be approached that way (Everett et al., 2019). However, we can anticipate that consent for in-vehicle users (drivers or passengers) will not be a major issue, as there are other non-consent legal bases to support data processing. For example, if the data is strictly necessary for the safe and autonomous operation of the autonomous driving system or vehicle, it can be expected to be imposed as a mandatory requirement on the controller. For instance, current regulation by United Nations concerning the approval of vehicles with regard to Automated Lane Keeping Systems (ALKS) (UNECE WP.29 GRVA, 2021c) clearly specify that the system shall detect driver attentiveness by detecting driver gaze and head movements. Although this does not mean identifying who the driver is, as a legal obligation for

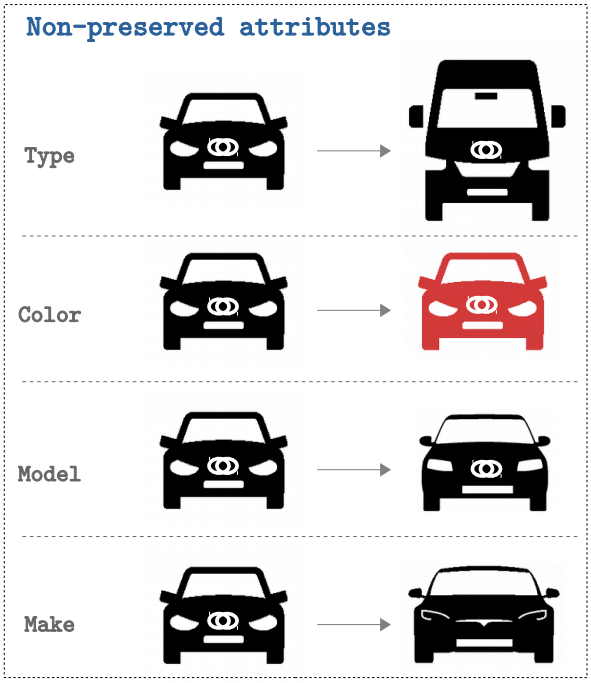


Figure 20: Image de-identification applied to vehicles. Different transformations can be applied, preserving some attributes and substituting some others, depending on the context.

the controller it will not be conditioned on consent. Obviously, if the controller uses personal data of users for other purposes not directly related to the safe operation of the system or without legal obligation, consent must be a prerequisite as in any other case not specific to AVs.

The main question, however, is **to which extent it is necessary to obtain consent from potential third parties outside of the vehicle** who do not use any other AV (i.e., VRUs). It is impossible to obtain consent for data processing from all persons who appear in the vicinity of the AV while it is in motion. It is also not possible to directly provide the identity and contact details of the data controller to these external road users. Only indirect methods are possible that allow these users, on the one hand, to clearly and easily identify that the vehicle is autonomous (e.g., a colour code), the manufacturer (e.g., logo) or the mobility company (e.g., vehicle lettering), as well as other distinctive details such as license plate number, model, color, etc., and, on the other hand, to have public information (e.g., on websites) with all the data required by the regulation to easily allow to establish who the data controller is. In any case, there are scenarios that can mitigate, or even avoid entirely, the requirement for consent from external users. For instance, if the **images are only processed in real time** to generate non-identifiable metadata such as free space, anonymous vehicle and pedestrians locations and future motions, etc., then they are neither stored (only the time needed to process them, and never reaching persistent storage) nor transmitted, which limits the intrusiveness. Controllers may in some cases also be able to argue that they rely on other grounds for lawfulness of their processing than consent. Finally, if appropriate privacy-by-design approaches, including **image de-identification methods**, are implemented to process, store or transmit images without personal data, data protection law would no longer apply.

Finally, it is important to differentiate data focused on improving system safety for AVs, from data related to other purposes of connected vehicles, which may or may not be autonomous or automated. We refer to personal data that can be processed inside the vehicle, exchanged between the vehicle and personal devices connected to it (e.g., smartphones), or captured inside the vehicle and exported to external entities for processing. In this sense, we can refer to the guidelines elaborated by the European Data Protection Board (EDPB) on processing personal data in the context of connected vehicles and mobility related applications (EDPB, 2020), which includes as general recommendations data minimization, data protection by design and by default (e.g., local processing of personal data, anonymization and pseudonymisation, etc.), information to data subjects, security and confidentiality.

5.4 Transparency (KR4)

A crucial component of achieving trustworthy AI for AVs is transparency to both internal users (drivers and/or passengers) and external road users (e.g., pedestrians, other drivers), and it encompasses three elements or sub-criteria which are addressed below.

5.4.1 Traceability

System traceability can be defined as *the ability to relate uniquely identifiable system artefacts created and evolved during the development of a system, maintain these relationships throughout the development life cycle and use them to facilitate system development activities* (Maro, 2017).

Traceability is a well established topic concerning functional safety of road vehicles. For example, all safety-critical vehicle systems have to comply with safety standards such as ISO 26262 (Road vehicles - Functional safety) that require traceability to be established between *artefacts* (work products) to ensure that the resulting systems are adequately tested and therefore safe. In this domain, system *artefacts* include stakeholders and system requirements, design models, behaviour models, hardware models, software requirements, software architecture, software detailed design, software units, and then test specifications and test results (V-model) among all other work items that are related to the system (Gotel and et al., 2012).

Despite the wealth of knowledge on traceability, in practice **establishing (software) traceability of modern conventional vehicles is already a challenge** (Maro et al., 2017), so the complexity is even greater in the case of automated or autonomous systems. For example, the number of artefacts related to Electronic Control Units (ECUs) in a conventional vehicle can be around 100K, but when it comes to AVs the number can grow up to 10M, with a heterogeneous tools chain (YAKINDU, 2020). A typical high-end conventional car consists of features that amount to about 100 million lines of code (Maro et al., 2017) and the number of traceability links is massive. As an example of the complexity, systems specifications of a 2004 car had already reached $20,000$ pages at that time (Maro et al., 2017). Taking into account all the developments and new systems that have been introduced in vehicles since 2004, the current number should be much higher.

Furthermore, an additional difficulty is how to **effectively integrate machine learning processes and datasets as work products or traceable artefacts**. For example, code constituting a deep neural network is not comparable to classical code, and low-level requirements are very hard to specify (Aravantinos and Diehl, 2019). In addition, reproducibility and explainability are well known problems due to the randomness and trial-and-error nature of the training processes of machine learning models (Goldgof et al., 2020). Some attempts have been made to bring out new artefacts and forms of traceability to deal with the particular trial-and-error development process of deep and machine learning (Aravantinos and Diehl, 2019), and some tools, practices and data models are already available for traceability of AI models and systems (Mora-Cantallops et al., 2021).

One of the fundamental requirements for assessing the traceability of AV during their entire lifecycle is the use of **adequate logging practices** to record, not only the outputs of the different systems, but also the inputs from sensors, and the internal state of each system. Data logging, i.e., the recording and storing of time-stamped data from different systems over a period of time, is a well-known topic in the automotive sector, due, for example, to the well-established requirements of Event Data Recorders (EDR) (Gabler et al., 2008). EDRs have traditionally been used to investigate accidents, focusing on several parameters (e.g., vehicle speed, acceleration) that are continuously recorded and overwritten, communicated by using the CAN bus with multiple and different sampling rates, and finally stored conditioned by a trigger when a significant safety-related event occurs (OICA, 2019). The next step in data logging for automated driving was to also include the status of automated driving systems and the interactions between human drivers or passengers and automated driving systems, resulting in the so-called Data Storage System for Automated Driving (DSSAD) (OICA, 2018) which is also conditioned by a trigger. The International Organization of Motor Vehicle Manufacturers (OICA) has already identified multiple scenarios where DSSAD storage can not be conditioned by a trigger and they proposed a continuous storage approach with a limited amount of data (OICA, 2019).

The main challenge is to maintain a continuous logging scheme to record and store data from sensors, system status and outputs. Bandwidth and storage requirements of AVs are very demanding. For example, a conservative estimate of the total bandwidth required by the sensors, including several radars, lidars, cameras, ultrasonic sensors, IMUs, GNSS, etc., results in a range of between 3 and 40 Gbit/sec, that is, between 1.4 and 19 TB/h (Heinrich, 2017). This is in addition to the state of the models of the multiple AI systems in the vehicle. For example, the average size of deep learning models is usually around 75-100MB (Toole, 2019), but the trend is growing and architectures of up to several tens of GB are already available (Microsoft, 2020), (Brown and et al., 2020) (in this context, the size of the outputs of the systems can be considered negligible). These requirements make the development of new in-vehicle communication systems (e.g. 10 Gigabit Ethernet) and storage systems absolutely necessary to record and store multi-gigabit data streams. But even so, it seems necessary to develop **new strategies for the "smart" recording (or discarding) of data**, i.e. storing data only in those environments that present certain characteristics (e.g., interactions with other agents) that make them potentially interesting, and thus optimise the logged data.

5.4.2 Explainability

The huge advances in AI in recent years have come at a price: the increased complexity of AI systems improves performance, but worsens the ability of humans to understand how results have been generated from inputs.

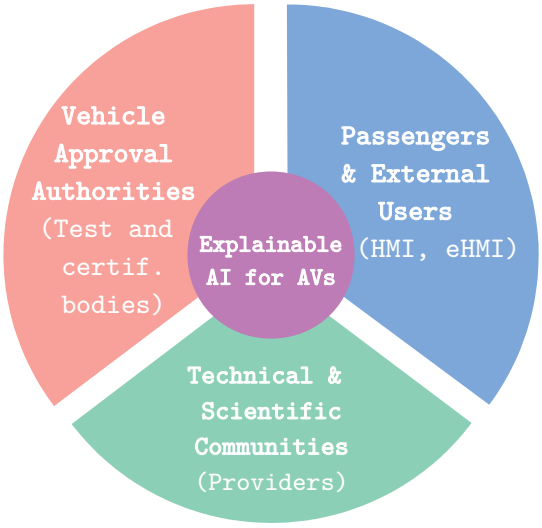
Although some AI systems, such as expert systems or rule-based models (white box models), do not suffer from this problem, as they are inherently designed on the basis of human knowledge, most symbolic and Machine Learning approaches are heavily affected by this issue. This is particularly problematic for Deep Neural Networks (DNNs), which have a huge parametric space comprising hundreds of layers and millions (even billions) of interrelated parameters (black box models), hardly interpretable by humans. The rise of opaque decision making systems has also brought the proposal of new paradigms for opening the black box, that fall within the so-called **eXplainable AI (XAI)** field.

XAI can be defined as follows: ***given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand*** (Barredo-Arieta and et al., 2020). This definition emerges from the definition of explainability: *given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand* (Barredo-Arieta and et al., 2020).

The amount of works related to this topic is considerable and there are other similar definitions for *explainability*. For example, for human-machine interactions, *explainability* can be defined as *the ability for the human interacting with the machine to understand its logic, based on how the human understands the connections between the inputs and outputs of the system* (Rosenfeld and Richardson, 2019). A similar approach is presented in (Doshi-Velez and Kortz, 2017) to define the term *explanation* as *human-interpretable information about the logic by which a decision-maker took a particular set of inputs and reached a particular conclusion.* An *explanation* should be able to provide *human-interpretable information about the factors used in a decision and their relative weight.* The term "weight" can be understood as the effect that a particular change in an input has on the output, or how two sets of similar-looking inputs result in different outputs, or vice versa (Doshi-Velez and Kortz, 2017).

The terms *explainability* and ***interpretability*** are often used interchangeably. However, the idea that *interpretability* is **a precondition for *explainability*** is gradually gaining ground (Doshi-Velez and Kim, 2017), (Zablocki et al., 2021). For example, *interpretability* can be understood as to which extent an explanation is understandable by humans (Gilpin et al., 2018). Then, an appropriate *explanation* should be designed as a trade-off between its *interpretability* and its *completeness*. That is, an exhaustive and completely faithful explanation may be incomprehensible to humans, and a very simplistic one may be not useful (Gilpin et al., 2018). In fact, the need for *explainability* in AI systems stems from a mismatch between completeness of the input space and training goals on the one hand, and the much more complex real-life inputs and goals on the other (Zablocki et al., 2021).

Figure 21: Impact of eXplainable AI for AVs and main actors involved.



The need and importance of XAI systems for AVs can be analysed from three different perspectives, as depicted in Fig. 21:

— *Internal (backup drivers and passengers) and external road users*: there are multiple types of situations that can arise for which internal users may need an explanation (Eliot, 2021). For example, the AV stops suddenly or drastically reduces the speed, the AV takes an unexpected route, the AV performs a lane change with no comfortable error margin, etc. Some degree of explanation to external users (e.g., pedestrians, cyclists) is also to be expected. The way to communicate the explanations to users will be based on HMIs and eHMIs, which is also linked with key requirement 1 of human agency and oversight.

— *Technical and scientific communities, producers, developers, etc.*: although it was initially considered that there was an inverse relationship between the performance of a model and its transparency, in the sense that focusing on performance alone makes the system more opaque (Dosilovic et al., 2018), the advances in XAI have led to the conclusion that improved understanding of a system leads to better identification of its deficiencies, and therefore explainability can be of great help in guiding performance improvement processes (Barredo-Arieta and et al., 2020). Explanations can provide technical information about the current limitations and shortcomings of a model (Zablocki et al., 2021).

— *Regulators, vehicle type approval authorities and insurers*: future safety certification procedures, whether under vehicle type approval (homologation) or self-certification frameworks, may benefit considerably from the consideration of explainability as a requirement to better assess the compliance with safety, human agency and oversight and transparency specifications. Auditors, accident investigators and insurers will also benefit from explainable systems (Omeiza et al., 2021).

Two main approaches to deal with XAI have been proposed. First, methodologies focusing on *transparent* models by design, and second, external XAI techniques focusing on explaining opaque models (*post-hoc explainability*) (Barredo-Arieta and et al., 2020). Based on an approach similar to that used in (Zablocki et al., 2021), we have identified some of the most important **explainability barriers and questions** for each of the AV layers (see Table 14).

Table 14: Main explainability barriers and questions identified for the AV layers.

AV Layers	Explainability Barriers	Explainability Questions
Localization	- Multiple sensors types - Fusion of multiple systems - Map-reality gap & Driver	- Is localization accuracy enough? - How close or far are we from exiting or entering a pre-mapped region (e.g. ODD)? - How will the localization system behave in unmapped scenarios? - Is localization fail-x (aware, safe and operational)?
Dynamic Scene Understanding	- Multiple sensors types - Under/over-represented situations - High dimensional space - Various Bias - Highly non-linear models - Underfit/overfit - Ill-specified objectives - Million of parameters (black box models) - Predicting future behaviours and trajectories - Adversarial attacks	- Are all dynamic objects properly detected and localized? - How certain is the system about the current and future status of dynamic objects? - How far or close are we from the ODD (e.g., lighting or weather conditions)? - How will the model understand new scenarios and behaviours? - Did the model correctly learn and generalize to unseen or rarely encountered situations? - Is scene understanding fail-x (aware, safe and operational)? - Are adversarial attacks being detected?
Path Planning	- Dependency on other systems' outputs - Complex dynamic environment - Influence of predictions - Dilemmatic situations	- Why is a certain manoeuvrer decided (e.g., lane change, braking, overtaking, evasion)? - Where will the AV go in the short/mid/long term? - Is it possible to reach a minimal risk condition?
Control	- Physical constraints - Several possible futures and actions	- Is is feasible to achieve the planned local trajectory? - Is it possible to achieve a minimal risk condition?
User Interaction	- Multiple sensors types - Various biases - Multi user problem	- How confident can the user be that the system has understood his/her questions or commands? - To what extent is the system certain that the user understands the explanations?

We can consider each of the key layers of AVs as a specific objective of explainable AI research work. *Explainable localization* would focus on communicating the AV position on a given map, including accuracy, errors, robustness, static elements, etc. A specific interface could effectively display this information (Schneider

et al., 2021). As stated in (Omeiza et al., 2021), although there seems to be less research related to *explainable localisation*, intelligible explanations remain key. *Explainable scene understanding* would focus on providing interpretable explanations about the current and future state of dynamic objects in the environment. This problem is closely related to explainable deep learning, as DNNs are fundamental structures for perception and scene understanding (Omeiza et al., 2021). For example, in (Bojarski et al., 2018) a gradient-based explanation method (VisualBackProp) was proposed to identify the parts of a driving scene image that are necessary for the steering operation of the AV. In (Kim et al., 2018) textual and saliency maps explanations of the AV actions are produced using an attention-based video to text model. *Explainable path planning* can play a key role in supporting users and improving their experiences when they interact with autonomous systems in complex decision-making procedures (Chakraborti et al., 2020). Some relevant work in the robotics field include XAI-PLAIN (Korpan and Epstein, 2018), WHY-PLAIN (Korpan and Epstein, 2018) and Refinement-Based Planning (RBP) (Bidot et al., 2010). These methods have not yet been applied to the field of autonomous driving. Once a local trajectory to be followed by the AV has been traced, *explainable control* would focus on providing the necessary information for the user to interpret the behaviour of the lateral and longitudinal controllers. This has not yet been widely studied for AVs, and in many cases it is approached from the perspective of the path planning or decision making layer (Omeiza et al., 2021). Finally, the user interaction layer serves as a means to implement explainable approaches for the other layers.

5.4.3 Communication

This sub-criterion focuses on the high-level information that users should have (and understand) regarding the benefits and risks of the technology. Since AVs are interactive AI systems, the first question to address is whether users have been communicated that they are **interacting with an AI system rather than a human**. Although it may seem unnecessary for users inside the vehicle to be informed about this (since there is clearly no human driver) there can always be doubts if the interaction and control of the vehicle is being carried out by a person remotely, e.g. by teleoperated driving (Neumeier et al., 2018). This information should therefore be conveyed to the user as soon as communication with the system is initiated, either by voice message or by distinctive visual information inside the vehicle (e.g., sign, logo). In addition, this information must also be **communicated to external road agents**, who, although not users of the AV, interact with it (albeit non-verbally). There must be some form of external marking, for example through visual eHMIs (Carmona et al., 2021), that clearly and effectively identifies that the vehicle is autonomous and not driven by a human.

The second issue concerns the communication of the benefits and risks of using an AV. Obviously, this communication should be addressed only to users (passengers or backup drivers), since external agents, even if they interact with the AV, are not users. Risks (and benefits) communication is a well-established and solid research topic that has been studied from multiple disciplines (e.g., psychology, sociology, communication, etc.) and in a wide range of applications (Fischhoff, 2011). Risks and benefits are multidimensional, continuous and can be seen from a social or an individual perspective (Venkkataraman, 2017). Therefore, effective communication must consider all the variables, and be able to discretize, and even particularize, the information depending on the subjects and their circumstances.

Consider a scenario in which AVs are properly certified in terms of safety. Obviously this certification does not eliminate risk, but it keeps it low enough to make the technology acceptable. Indeed, operating an AV will always have inherent risks to the user who becomes vulnerable to any performance problems of the system. In this context, **appropriate communication of risks can serve to decrease disuse and misuse of AVs**. To avoid disuse, drivers have to voluntarily assume the risks when they cannot control the system and the perceived risk must be low enough for this to occur. On the other hand, misuse will occur if the perceived risks are less than what they really are (e.g., overreliance of drivers using level 2 and level 3 automated driving systems). Presenting drivers with the appropriate **information at an early stage in the car-driver interaction** has been shown to significantly influence their **perception of risks**, thereby calibrating trust and promoting correct use of the technology (Li et al., 2019).

5.5 Diversity, non-discrimination and fairness (KR5)

Diversity, non-discrimination and fairness are requirements mainly related to the ethical principle of fairness, which includes avoidance of unfair bias, accessibility/universal design and stakeholders participation sub-requirements.

5.5.1 Avoidance of unfair bias

This sub-criterion focuses on *fairness* and *bias*. As stated in (Tolan, 2018), *fairness* is a complex value-driven normative concept that is difficult to formalize, while *bias* is a technical concept that can be defined as a systematic deviation from a true state. Generally, a process or decision (automated or not) is considered *fair* if it

does not discriminate against people on the basis of their membership to a group. In machine learning, *fairness* usually refers to the attempts to avoid *unfair algorithmic bias*, i.e., avoid systematic and repeatable errors that create unfair outcomes, such as privileging one arbitrary group of users over others.

In the case of AVs, the first source of possible unfair biases based on the personal characteristics of certain groups that can lead to discrimination can be found in the **decision-making/trajectory planning layer**, which has been extensively studied from the perspective of **encoding moral decisions** in artificial systems (Wallach and Allen, 2009). The idea was to anticipate a near future in which the driving capabilities of AVs are so advanced that they can discern and cope with scenarios of any kind, including *dilemmas* in which the autonomous decision-making system must decide between two or more unavoidable critical situations based on moral principles. In one of the early studies, AV decision making was approached with a purely utilitarian view in which, for example, it might be decided to sacrifice passengers for the greater good. This was called the *social dilemma of AVs* (Bonnefon et al., 2016). This idea was further developed in the well known *Moral Machine* project (Awad et al., 2018) in which information was collected from millions of users around the world (more than 40 million responses) through an game-like on-line platform that posed binary choices in scenarios involving AVs that were going to crash. The scenarios are variants of the *trolley problem* adapted to the context of AVs.

This project, whose results were published in *Nature*, went viral and had a great social impact. While, on the one hand, it served to raise public awareness of the importance of considering ethical aspects in the design of AVs, it also had negative effects. For example, it has somehow misled the public into believing that AVs are programmed to crash into certain types of people or simply that they are dangerous (Iagnemma, 2018). This has had a negative impact on public opinion, which may slow down the adoption of the technology (Etienne, 2021). It could also have somewhat distracted regulators from the important task of ensuring a safe transition to the deployment of AVs, and even scientists and engineers who have committed resources to study these issues, since trolley problem scenarios are extremely rare and highly unlikely to occur in real traffic, even for human-driven vehicles.

But even if these type of problems were a realistic concern for AVs, the approach suggested by the Moral Machine project, i.e., encoding the majority opinion of millions of people on the social value of some decisions in scenarios in which AVs will choose between harming one set of people or another, is ethically questionable (Jaques, 2019), and intrinsically entails unfair biases and discrimination. It is technically possible to develop decision-making systems that model expected harm and apply various "ethical theories" or "voting-based systems" to define cost functions, which will ultimately be optimized to decide who will be harmed (Moura et al., 2020). This would require the ability to detect some form of "social value" of people who could potentially be harmed by the AV, and would involve a form of explicit unfair discrimination or bias based on personal features.

As defined by the German Ethics Commission on Automated and Connected Driving (German Ethics Commission, 2017), and in line with the Autonomous Driving Act drafted by the German Federal Government (German Federal Government, 2021), we can reasonably consider that, if feasible, systems must be programmed to accept damage to animals or property in a conflict if this means that personal injury can be prevented, i.e., **to give the highest priority to the protection of human life**. However, **in the event of an unavoidable alternative risk to human life** (e.g., dilemmatic decisions, such as a decision between one human life and another), **any distinction, weighting or offset on the basis of personal characteristics (e.g., age, sex, individual/groups, physical or mental constitution, behaviours, etc.) must be strictly prohibited**.

However, this does not imply that AVs should not use behavioural models based on personal features to improve perception systems and to have different path planning strategies depending on these features. Quite the opposite. Predicting the action and motion of road users, including pedestrians (Lorenzo et al., 2021), cyclists (Pool et al., 2019) and vehicles (Izquierdo et al., 2021), are a fundamental part of perception systems to provide AVs with the ability to anticipate risky situations and perform conservative path planning to improve safety and comfort. For example, a predictive perception system might anticipate that the probability of a child at the curb crossing in front of the ego-path is much higher than that of an adult and, therefore, the AV might reduce the speed for the child and maintain it for the adult. This does not necessarily mean that adults are being discriminated against (as long as there is no bias in the data and algorithms used to generate the predictive perception models) but that the uncertainty in the behaviour and possible motion of the child is de facto greater than in adults, and **the AV must adapt its behaviour to maintain the same level of safety** for children and adults.

We can reasonably expect **policymakers to require that AV manufactures guarantee the same level of safety for all road users**, and that this requirement implies the need to develop **AVs that behave differently according to the road user**. This is partially in line with the risk distribution recommendation of the Independent Expert Group on Ethics of Connected and Automated Vehicles (European Commission, 2020). However, we should be cautious because linking the distribution of risk (or level of safety) to the level of vulnerability of road users based, for example, on the ratio of fatalities to road exposure, or disabilities, may lead to discrimination on the basis of the vulnerability of the road user, perceived as a way of placing a higher value on the safety of more VRUs. A more suitable approach assumes that the **AV should act differently to**

correct safety inequalities which result from the different behaviours of road users.

But the decision-making or path-planning layer is not the only one in which unfair bias problems can occur. Although relatively unexplored, the **dynamic scene understanding** layer can suffer from unfair bias both in terms of the use of biased input data and in the design of the algorithm. The methods used to detect different road users (e.g., pedestrian, cyclists, motorbikes, vehicles) are learning-based. Unfair biases in the datasets used for training, validating and testing of these methods could lead to differences in performance that could ultimately imply biased crash results and different levels of safety.

For example, some preliminary studies have found clear biases in different pedestrian detection methods based on variables such as age (adult/child) and sex/gender (male/female) (Brandao, 2019), and skin tone (light/dark) (Wilson et al., 2019). When using the Caltech Pedestrian Detection Benchmark (Dollar et al., 2012) to train, validate and test different pedestrian detection algorithms, **children and female pedestrians had higher miss rates than adults and male pedestrians** (Brandao, 2019). In addition, when using a subset of the BDD100K dataset (Yu et al., 2020), state-of-the-art methods for object detection exhibited **lower precision on higher Fitzpatrick skin types (dark) than lower skin types (light)** (Wilson et al., 2019).

These studies are still preliminary and only address a very small portion of the problem. For example, they focus only on the problem of detection. However, predictive perception methods that model human behaviour in order to anticipate actions and motions can provide biased performances depending on the geographical origin and representativeness of the different users and groups in the datasets. In addition, important variables such as group or individual presence are omitted. On the other hand, the available studies only focus on pedestrians. Bias in detection and prediction can also occur for other VRUs (e.g. cyclists, motorcyclists, wheelchair users, etc.), but it can also occur for vehicles depending on variables such as colour, size, type, etc. The strong dependence on data of detection and prediction systems, and the large amount of datasets already available to address the task of dynamic object detection, suggest that a more in-depth analysis is needed. Furthermore, **as the bias in the perception systems of AVs may have an impact on the safety of some road users**, who may be discriminated against compared to others, **more effort is needed from the scientific community, industry and policymakers to take this issue further** and more broadly.

In addition, **unfair bias** can also be present in the **user interaction layer**. Human-vehicle interfaces may include systems for facial recognition, voice recognition, behavioural and emotional state modelling, attention state detection, etc. All these systems are based on complex data-driven AI systems, so all the recommendations for designing datasets and AI systems to reduce bias are applicable to this layer. Although the study of the principles that should govern human-machine interaction have been discussed in the HMI community for over two decades (Amershi and et al., 2019), and there are recent advances in the development of guidelines and assessment methods for the design of HMIs for AVs (Schöomig and et al., 2020), the study of potential bias in this area has not yet been addressed, and it will be an important area of research as technology evolves.

Finally, as stated in one of the recommendations of the Commission Expert Group on ethical issues raised by driverless mobility (European Commission, 2020), and in relation to the principle of fairness, **discriminatory service provision should be prevented**. For example, the provision of mobility as a service (MaaS) with AVs may introduce different degrees of price discrimination schemes depending of multiple factors, such as the origin and destination, time of the day or level of demand (J.Bahamonde-Birke et al., 2020). One could even consider a scheme with different degrees of safety (above the minimum required), comfort and efficiency depending on the tariff the user can afford to pay. As stated by the Commission Expert Group, the future market for AVs opens up new possibilities for differential provision of autonomous mobility systems, services and products that pose a risk of perpetuating and increasing inequalities between individuals and groups in society. To avoid this potential discrimination, policymakers need to set up institutions to continuously monitor, evaluate and guide manufacturers, developers and service providers (European Commission, 2020).

5.5.2 Accessibility and universal design

One of the most exciting promises of AVs is that they can extend mobility to users who cannot drive a conventional car, including minors, old adults, people without driving license, and people with disabilities. But this poses a major challenge for the design of user interfaces in terms of interaction, accessibility and adaptability, including the possible need for (anonymous) disability-specific identification (Fernandez-Llorca et al., 2017). Recently, this is becoming a hot research topic, with a variety of terminology, including *inclusive* (U.S. DOT, 2020), *accessible* (PAVE, 2020) or *user-centred* (Stanton et al., 2021) design. A more general approach focuses on the application of the **seven principles of Universal Design in AVs** (Costa et al., 2019). However, for now, only conceptual models have been explored (Ferati et al., 2017), (Carvalho et al., 2021), which should be implemented and evaluated in the near future.

5.5.3 Stakeholder participation

The number of stakeholders who may directly or indirectly be affected by AVs is considerably high. Apart from all the companies and research centres involved in the development of the technology (e.g., OEMs, Tiers, etc.) the groups most directly and potentially impacted include, inter alia, the following (Gore and Wild, 2016), (CB Insights, 2021):

- Car manufactures and auto dealerships.
- Public institutions such as national and local governments, city and road planners, rescue and emergency services, healthcare, public transportation, toll road operators, traffic enforcement, litigation, etc.
- Citizens, road users and drivers (individuals and associations).
- Real state, parking garages and lot owners, hotels and motels, etc.
- Insurance, auto repair, after market and auto parts companies.
- Energy and petroleum companies, oil change spots, car washes, etc.
- Shipping and trucking companies, last-mile delivery companies, industrial fleet operators, ride-hailing companies, professional drivers (e.g., taxi, limo, trucking), domestic and short-haul airlines, etc.
- Food preparation and delivery, fast food companies, convenience stores, brick and mortar stores, etc.
- Others such as media and entertainment providers, home improvement companies, interior design and manufacturing, co-working spaces, etc.

Although the above list is quite comprehensive, it is **necessary for policymakers to establish a clear taxonomy of stakeholders**, modulating the direction (positive or negative) and the weight of the impact that the adoption of AVs implies for each of them. Then, as recommended by the Expert Group in its report on Ethics for CAVs (European Commission, 2020), it is necessary to incorporate specific steering actions and oversight procedures from public and non-governmental institutions to allow stakeholders to be actively and continuously engaged in deliberation about design and evaluation of AV systems and services, throughout the innovation lifecycle (e.g. through in-person, online citizen forums, etc.).

5.6 Societal and environmental well-being (KR6)

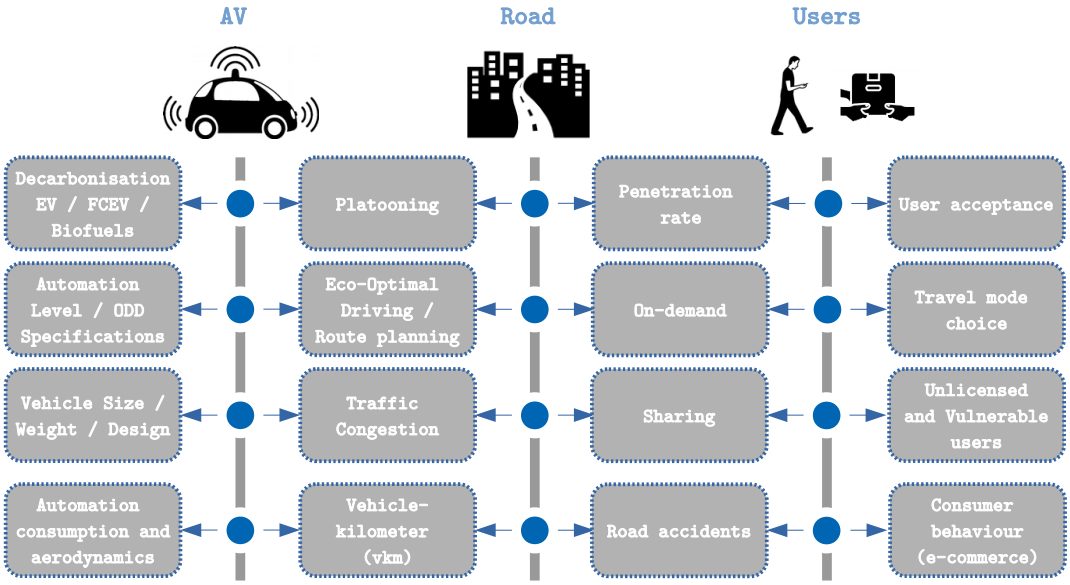
AVs have the potential to considerably reshape transportation and, with it, the environmental and social impact of moving people and goods. In principle, the impact can be positive or negative. In order to tip the balance towards social and environmental improvement, it is important that policymakers steer and monitor the process of adopting this technology. In this line, this requirement focuses on three fundamental aspects that will be described below: *environmental well being*, *impact on work and skills*, and *impact on society*.

5.6.1 Environmental well-being

Over the last decade (Kopelias et al., 2020) there has been considerable effort by the scientific community to identify, **understand and estimate the impact that AV uptake may have on the environment**, including calculation of the impact on energy consumption, emissions, waste, etc. This problem **is highly multidimensional and involves potentially very disruptive factors, making it very complex to establish precise estimates** of the specific impact of each variable and of the overall impact. The Fig. 22 shows a categorisation of the most relevant factors, linked to three different components: AV, road infrastructure and users. In the following, we provide a brief description of each of the factors.

- *Decarbonisation EV / FCEV / Biofuels*: although vehicle automation is a separate technological, social and political process of the use of alternative fuels, like electricity, hydrogen, biofuels or natural gas, the two issues are somehow interlinked. The innovative nature of the supply chain of vehicle automation is a driver for innovation and transformation of the powertrain technologies. This includes powertrain optimization for AVs. For example, it has been demonstrated that battery-electric powertrains require a lower amount of energy per unit distance compared to the equivalent internal combustion engine (Tate et al., 2018).
- *Automation Level / ODD specification*: the impact that AVs can have on the environment will depend on the level of automation achieved and the scope of the ODD (Ross and Guhathakurta, 2017), which will ultimately define the penetration rate and uptake of this technology.

Figure 22: Key environmental factors in the development and adoption of AVs.



Source: Own elaboration based on (Kopelias et al., 2020).

— *Vehicle Size / Weight / Design*: depending on the penetration rate and safety gains, we could foresee a scenario in which AVs could have smaller dimensions and weight than conventional vehicles (i.e, lower consumption), due to the elimination of passive safety systems that might not be so necessary if, indeed, accidents are drastically reduced (Morrow and et al., 2014). However, in the short term, it does not seem reasonable to consider eliminating safety features. In fact, even if only moderately, automation increases the weight of the vehicle (Armstrong et al., 2020) due to all the additional elements required (sensor platform, ECUs, etc.).

— *Automation consumption and aerodynamics*: apart from the added weight, the components required for vehicle automation also entail higher energy consumption, albeit very moderate compared to other commonly used systems such as the air-conditioning system. On the other hand, in approaches that integrate a sensor platform on the roof of the vehicle (e.g., Waymo or Uber), they entail greater aerodynamic resistance and higher consumption, up to 25% higher (Chen and Meier, 2020).

— *Platooning*: connected and cooperative AVs enable efficient vehicle platooning, reducing aerodynamic drag forces, fuel consumption and exhaust emissions (Alam et al., 2015). Different studies estimate a possible energy reduction between 10-25% (Stephens et al., 2016), (Tsugawa et al., 2016).

— *Eco-Optimal Driving / Route planning*: AVs have the potential to save fuel by driving with efficiency-maximizing speed and acceleration operating profiles. Most studies agree that eco-driving could reduce fuel consumption by up to 20% (Kopelias et al., 2020). In addition, cooperative and connected AVs can also apply efficient eco-routing to find routes that require the least amount of fuel and/or produces the least amount of emissions (Boriboonsomsin et al., 2012).

— *Traffic Congestion*: on the one hand, cooperative and connected AVs have the potential to reduce traffic congestion, and thus travel times and emissions. However, the actual reduction depends on many factors, such as penetration rate (Kopelias et al., 2020) and cooperative driving pattern (e.g., distance between vehicles) (Stogios et al., 2019), with the possibility of having a negative result in terms of emissions. On the other hand, traffic congestion will also depend on the impact of AVs on the number of vehicle-kilometre (vkm).

— *Vehicle-kilometre (vkm)*: on the one hand, factors that will lead to an increase in the number of vkm include an increase in the number of potential users, e.g., people with disabilities or without a driver's license (Harper et al., 2016), higher usage if trips are ultimately faster, and empty or unoccupied trips, i.e. vehicles without passengers looking for users. On the other hand, the possible increase in shared mobility, and the reduction of trips used to search for parking (Zhang et al., 2015), are factors that will reduce this number. Although the final outcome is uncertain, most studies conclude that the adoption of AVs will increase the number of vkm (Stephens et al., 2016), (Haghighi et al., 2019), (Chen et al., 2019).

— *Penetration rate*: this is one of the most important variables affecting all others. The impact of AVs on environmental well-being will be directly related to the penetration rate, i.e. the ratio of automated vehicles to all types of vehicles (conventional and automated).

— *On-demand / Sharing*: the use of shared vehicles accessed on-demand can provide environmental benefits, as users tend to decrease both vehicle ownership and annual vehicle distances travelled. There is evidence that the combination of on-demand mobility and AVs may amplify the adoption of both (Greenblatt and Shaheen, 2015).

— *Road accidents*: in addition to the safety benefits of avoided accidents, there are multiple side effects of the improved safety provided by AVs such as the elimination of the subsequent traffic jams caused by these accidents (Stephens et al., 2016), lower demand for crash-related repairs and replacement parts, and potential decrease of insurance costs.

— *User acceptance / Travel mode choice*: the penetration rate, and in general the adoption of AVs and new on-demand mobility models, will depend on user acceptance, which in turn is modulated by multiple factors, the most important of which is safety (Garidis et al., 2020). Consumer's travel mode choice will also play a key role in the environmental impact, including the preference of private or shared AVs (Pakusch et al., 2018), and the potential mode shift from walking, public transport or regional air travel to AVs (Stephens et al., 2016).

— *Unlicensed and Vulnerable users*: AVs will increase mobility for new groups of people, including the senior population, people with disabilities, and non-drivers, which could increase the number of vkm to as much as 14% (Harper et al., 2016), (Stephens et al., 2016).

— *Consumer behaviour (e-commerce)*: the environmental impact of AVs in freight transport, and especially in last mile logistics in urban environments, is generally positive compared to conventional delivery (Figliozzi, 2020). However, improving the efficiency of these systems can generate an induced increase in demand (Bridgelall and Stubbing, 2021) leading to an increase in the number of vkm, and therefore an increase in consumption and emissions.

Although it is possible to design models for estimating energy and fuel consumption, emissions, waste, environmental impact, etc., which interrelate all these variables in a coherent way, the fact is that at the moment the uncertainty is so high that it is only possible to establish very strong assumptions, with large ranges of variation, and to provide results with considerably high margins of error. In most of the existing works in this area (Wadud et al., 2016), (Stephens et al., 2016), (Ross and Guhathakurta, 2017), the margins of the estimates that result in a negative impact on energy or fuel consumption (higher consumption with AVs than without AVs) are much higher than those that result in a positive impact. It is therefore crucial that policymakers and public authorities are active participants in this process in order to tilt the balance towards the side that most enhances environmental well-being.

5.6.2 Impact on work and skills

The impact of AVs on work and skills will depend on the specific adoption scenario. On the one hand, automated vehicles (up to SAE Level 3) always require a backup driver ready to resume control when necessary, so **conditional driving automation will not lead to job losses**. However, it is possible that the type of skill required will change, as the driver's role will shift from vehicle control to monitoring. These **new skills could be required in the future to obtain a driving licence to operate automated vehicles**. Indeed, conventional drivers need to understand how an automated driving system works, what are the key variables to monitor, how to identify that there is a critical safety issue, and how to resume control when necessary. All of these skills are not included in conventional driving.

For **highly automated or AVs** (at least SAE Level 4), the long-term **impact** for the transportation **workforce is likely to be negative** as it is expected that there will be fewer workers and that one operator will be able to supervise several vehicles. The sectors most affected by driving automation are the passenger and freight transport sectors, both long distance and last mile. However, recent studies show that the impact, although negative, will be **more limited than expected** (Gittleman and Monaco, 2020). For example, although the main task of truck or delivery van drivers is to drive, it is important to note that they also perform a number of **non-driving tasks**, many of which are **less susceptible to automation**. In addition, although in a different location to where the technology is operating, **new jobs** could be created (e.g., in control centres), including a **new demand of digital skills** (Bissel et al., 2020). It is also anticipated that the development of new technologies and services will require new skills and high quality jobs (e.g., scientists, engineers). In any case, professional drivers affected by driving automation may find it difficult to look for alternative employment due

to the low adaptability of their skills (Beede et al., 2018). National and regional plans will be essential for providing reskilling and training for workers whose jobs may disappear due to automation.

As one of the most vulnerable stakeholders in the driving automation process, it is important that workers' representatives affected by the driving automation process can play an active role in the public deliberation on the adoption of AV systems and services.

Finally, in relation to the impact on work, the adoption of AVs opens up the possibility to **use travel time for work-related activities**, i.e., *doing office work on the motorway* (Laurier, 2004). Vehicle interiors can be transformed into a place not only for leisure and well-being, but also for productivity (Schartmuller et al., 2020). It is increasingly common for commuters to engage in work-related tasks to compensate for lost time, even in manually driven cars (Perterer et al., 2016). With automated vehicles, and more especially with AVs, this behaviour is likely to intensify. The impact may lead to higher productivity in general, or a reduction of time at the workplace as commuting time could be considered as working time.

5.6.3 Impact on society at large or democracy

As a disruptive technology, and considering the weight that transport plays in our society, it is obvious that the uptake of AVs will have a considerable impact on society at large. The dimensions include safety, travel times, traffic congestion, access to mobility to new social groups, environmental issues and urban planning, among others. It will clearly affect travellers behaviour and the way people and goods are transported.

Autonomous mobility has the potential to transform how people spend their time on the road, including new possibilities for social interaction as interior cabins are likely to be significantly modified to allow greater flexibility for work, leisure and social activities (Lipson and Kurman, 2016). Other social aspects such as the impact on digitalization, changes in strategic travel planning and scheduling, or even the transformation of the sensory dimensions of being on the move are relevant (Bissel et al., 2020).

In terms of democratic processes, political decision-making and electoral contexts, at least for now, it is very difficult to foresee any kind of impact of AVs in this regard.

5.7 Accountability (KR7)

This requirement is addressed through two sub-criteria that must be put in place before and after the development, deployment and use of AVs. The first, auditability, to enable the assessment of algorithms, data and design process. The second, risk management, to develop an environment where negative impacts of AVs can be identified, assessed, documented and minimised. These are addressed in the following subsections.

5.7.1 Auditability

Auditability of AI systems for AVs is closely related to transparency (key requirement 4) as the main mechanisms needed to ensure auditability are traceability, explainability, the sourcing of training data and logging of system inputs, processes and outcomes (Toader, 2019). Furthermore, as stated by the AI HLEG (AI HLEG, 2020), in safety-critical applications (such as AVs), AI systems must be audited not only by internal auditors, but also by **independent external auditors**.

Although this does not necessarily imply that information about business models and intellectual property must be openly available, the fact is that for third parties to be able to effectively audit AI systems as complex as those incorporated in AVs, the level of detail in the information required is very high. The **main challenge**, therefore, **is to define the minimum requirements necessary for third parties to be able to audit the systems without compromising the intellectual and industrial property of the producers of AVs**. An additional difficulty for independent audits is the lack of expertise. The AI sector for AVs is very innovative and **audit expertise is limited** and only just **emerging**. All this suggests that in the early stages of the process, **internal audits will be critical**.

Auditability also links to **liability** in the sense that the same requirements and expertise needed to audit systems will be necessary for victims or insurers ([12]) to be able to demonstrate (burden of proof) a defect of the product (strict-based liability) or a fault of the manufacturer (fault-based liability), and the causal link with the harm caused by the vehicle, whether victims are passengers or external road users. Shifting the burden of proof to the manufacturer of AVs would make these systems more victim friendly (Uytsel, 2021). This would require considerable harmonization efforts and major updates of existing national product liability, traffic liability and fault-based liability frameworks, including the Product Liability Directive and the Motor Insurance Directive (EPRS, 2018).

([12]) Motor Insurance Directive provides a harmonised third-party liability insurance regime for prompt compensation of victims, even when an AV is involved. At that point, the insurer could assume the status of claimant, and take legal action against the manufacturer of the AV under the Product Liability Directive if there is a malfunction or defect of the AV.

Whereas proving defect or fault is triggered by an accident causing harm to persons or damage to property, internal and external audit processes can be linked to both ex-ante assessments and ex-post compliance mechanisms.

5.7.2 Risk management

The deployment and adoption of connected AVs will bring **new risks** that will become more prominent as their use and penetration rate increase. Some of these new risks include those related to the failure or programming/design choices of the operating algorithms or software that enable the AVs to function. In addition, those risks related to connectivity and network failures, and risks resulting from a third party hacking the system (cybercrime).

In addition, due to the high innovation component of AVs, the virtually infinite variety of possible scenarios, and the intensity, variability and complexity of the interactions of AVs in their operating environment, the potential scope for **risks that are unknown or undetectable at the time of production** ([13])**, or that can only emerge after market launch**, is much higher than for conventional vehicles. Current liability frameworks does not expect AVs manufacturers to pioneer safety and security risks that are not yet common in the industry at the time of production (e.g., development risk defence) (EPRS, 2018). However, a **proper balance must be found in liability and insurance costs between consumers and injured parties on the one hand, and AVs producers on the other**.

This sub-criterion further establishes the need to foresee different measures to reduce the potentially negative impact of the risks associated with AVs, including the establishment of an ethics review board, risk training, the availability of processes for third parties to report vulnerabilities, as well as redress by design mechanisms. Finally, somewhat recursively, it mentions the need for a specific process to assess the adherence to the Assessment List for Trustworthy AI (ALTAI). All these final remarks are quite generic and their application within the field of AVs should be no different from that of other types of AI-based technologies.

([13]) Production refers here to all processes (e.g. design, development, deployment) before the product, i.e. the AV or some AI-based component, is put into circulation.

6 Conclusions

In this section, after the detailed analysis the requirements and criteria, we proceed to summarise the main challenges ahead and future research gaps for each of them.

— **KR1. Human agency and oversight**

- Human agency for AVs is directly linked to the principle of human autonomy, and affects acceptance (e.g., disuse) and safety (misuse).
- New agency-oriented HMIs and eHMIs are needed to ensure an adequate level of human agency for backup drivers up to SAE Level 3, and in particular for passengers at SAE Levels 4 and 5, as well as for external road users.
- To this end, efficient approaches to measure and calibrate the sense of agency, including the explainability and interpretability of AV behaviour, are essential.
- Human oversight for AVs is exercised differently depending on the level of automation, through various human-vehicle interaction mechanisms. It is also exercised to some extent by external road users, with the risk of abuse in the interaction knowing that AVs will stop in any case.
- For a proper interaction there must be mutual awareness between the AV and the agents with whom it interacts, mainly drivers and passengers, but also external road users.
- How to represent and communicate the operating status of the AV to users as effectively as possible, including the TOR or request to intervene, is a key area of future research to meet the requirement for human oversight.
- Finally, the way in which oversight is exercised will require new skills that need to be addressed a priori, but which will also be developed through exposure and use. How to steer this process effectively is also a future challenge.

— **KR2. Technical robustness and safety**

- Technical robustness and safety are probably the most relevant requirements for AVs due to their strong connection to the principle of harm prevention, which if not achieved and communicated properly can also lead to disuse due to the impact to user acceptance.
- Attack resilience and security of AVs must be addressed from a heterogeneous, constantly updated approach, starting from security by design, including multiple defensive measures (e.g., cryptographic methods, intrusion and anomaly detection), countermeasures against adversarial attacks (e.g., redundancy, hardening against adversarial examples), fault-tolerant, fail-x, and self-healing methods, and even user training, to meet strict requirements of authentication, integrity, privacy, and availability.
- When assessing how safe is safe enough for AVs, maintaining, and communicating appropriately, conservative expectations could accelerate improvement in user acceptance. Even a small improvement can save many lives, and expectations that are too high can delay adoption and thus the benefits of the technology. This also requires new innovative methods to assess the safety of AVs with respect to human drivers, without requiring an assessment under real conditions in unachievable time periods.
- Although important steps have been taken in the design of new safety test procedures for automated driving functions, including simulation, physical test in proving grounds, and real-world test drive, there are still very important limitations, such as the absence of real-behaviours, limited variability, absence of safety requirements, lack of scenarios to assess human agency and oversight, as well as transparency and fairness. This is perhaps one of the most important points for further research and development.
- Accuracy of AVs is a multi-dimensional problem, involving multiple metrics, different levels, layers, use cases and scenarios. Defining holistic metrics and accuracy thresholds to comprehensively assess AVs is a challenging research, and even policy-based, problem that will need to be addressed in the future.
- New fallback strategies are needed to achieve minimum risk conditions, as well as testing procedures to assess their safety.
- Although continuous learning during the operation of AVs is not very likely due to safety risks and costs, it should be noted that any substantial change in an AI-based component of AVs that may modify the overall behaviour must meet all requirements for robustness, safety and security, and may need to be retested.

— **KR3. Privacy and data governance**

- New innovative approaches must be put in place to guarantee data protection without negatively affecting the safety of AVs.
- New agent-specific data anonymization techniques must be applied without corrupting the nature of the data, and preserving relevant attributes of agents. Data de-identification approaches (including generative adversarial methods) with the necessary constraints to preserve the relevant features of agents, seem to be the most appropriate approach to maintain the balance between anonymity and attribute fidelity.
- Privacy by design will also require data, storage devices and V2X communication channels encryption, with a unique encryption key management system for each vehicle and including regular renewal of encryption keys.
- Consent to the processing of personal data in AVs involves several challenges. For drivers and passengers it should not pose any safety risk, and should include the exchange of data with other vehicles and infrastructure. For external road users consent can only be obtained indirectly, although it can be avoided if data are processed in real time or if data de-identification is properly implemented.

— **KR4. Transparency**

- Software traceability is already a challenge for modern conventional vehicles, so its complexity for AVs is more than remarkable.
- A still immature research question is the effective integration of components of data-driven AI systems (e.g. machine learning models, hyper-parameters, datasets) as traceable artefacts.
- Apart from the necessary hardware developments, new strategies for intelligent data logging must be developed to cope with the demanding requirements in terms of bandwidth and storage capacity of continuous data logging for AVs.
- New explainable by design models and external methods to explain complex opaque AI-based models must be developed, focusing on internal users and external road users, which implies new research topics related to human-vehicle interaction through new HMIs and eHMIs.
- Explainability as a requirement for safety testing in vehicle type-approval frameworks will enhance the assessment of safety, human agency and oversight, and transparency, but will require new test procedures, methods and metrics.
- New effective ways to communicate to both internal users and external road users that they are interacting with an AV must be put in place.
- New ways of communicating risks to users need to be developed, especially in early phase of the human-vehicle interaction, to calibrate trust and avoid disuse and misuse of AVs.

— **KR5. Diversity, non-discrimination and fairness**

- All studies that deal with the moral behaviour of AVs affecting the decision making or (local) path planning layer in dilemmas where harming someone is unavoidable, are, in fact, a source of discrimination and a possible unfair bias based on a certain social value given to certain groups of people over others, which should be strictly avoided.
- To avoid discrimination, AVs must maintain the same level of safety for all road users, which in some case may imply the need to adapt the behaviour of the AV according to the intrinsic characteristics of the individual or the social group with whom it is interacting. In other words, AV should act differently to correct safety inequalities which result from the different behaviours of road users.
- Therefore, new predictive perception systems capable of modelling the behaviour of different road users are required. Research in predictive perception and path planning systems is flourishing, but robust, real-time solutions are not yet sufficiently developed.
- The study of unfair biases in perception systems for AVs has not been sufficiently addressed by the research community. Further efforts are needed to identify possible sources of discrimination in state-of-the-art perception systems for detecting external road users according to different inequity attributes such as sex, age, skin tone, group behaviour, type of vehicle, colour, etc.
- Unfair bias may also be present at the user-vehicle interaction layer. This will be an important research topic as both HMIs and eHMIs evolve.
- AVs opens up new autonomous mobility systems, services and products that pose a risk of perpetuating or even increasing inequalities between individuals and groups in society. Any service provision approach that may discriminate against users should be avoided.

- AVs can extend mobility to users who cannot drive a conventional vehicle (e.g., minors, old adults, people without driving license, people with disabilities). This poses a challenge for the design of accessible and adaptable HMIs, which should follow the Universal Design principles. This research topic has only been conceptualised, and new approaches should be explored.
- Although the main stakeholders have been identified, it is necessary for policymakers to establish a clear taxonomy of stakeholders, modulating the direction (positive or negative) and the weight of the impact that the adoption of AVs implies for each of them.

— **KR6. Societal and environmental well-being**

- Understanding and estimating the impact that the adoption of AVs may have on the environment is a highly multidimensional and complex problem, involving many disruptive factors, and for which we can only make predictions based on as yet uncertain assumptions. The net impact can be positive or negative, so it is essential that policymakers steer and monitor the adoption process to tip the balance towards a positive impact.
- Automation of vehicles up to SAE Level 3 will not have a negative impact on jobs, but it brings with it the need for new skills for backup drivers.
- In the case of highly automated or autonomous vehicles, as no drivers are needed, the expected impact on work and skills is likely to be negative, but is partially mitigated by the number of non-driving tasks less susceptible to automation, and the need for new jobs and skills that transport automation brings.
- The adoption of AVs opens up the possibility to use travel time for work-related activities, leading to higher productivity in general, or a reduction of time at the workplace as commuting time could be considered as working time. In the coming years we will see new approaches to transform the interiors of AVs into places to work. This is a major challenge especially in shared mobility contexts.
- AVs has the potential to transform the way citizens spend their commuting time, including new social and leisure activities. This has not been yet sufficiently studied.
- New approaches and studies are needed to provide more accurate estimates, with less uncertainty, of the impact of AVs on the environment and society.

— **KR7. Accountability**

- AVs are a safety-critical application and must therefore be audited not only internally, but also by independent external auditors. Establishing the minimum requirements for third parties to audit systems without compromising intellectual and industrial property is a major challenge.
- In future accidents involving AVs and resulting in physical harm or property damage victims (whether passengers or external road users) would need to demonstrate a defect of the product (strict-based liability) or a fault of the manufacturer (fault-based liability), and the causal link with the harm caused by the AV. The same requirements and expertise needed to audit AVs would be necessary for victims or insurers to claim for liability, which would be very complex and costly. Shifting the burden of proof to the manufacturer of AVs would make these systems more victim friendly.
- Considerable harmonization efforts and major updates of existing national product liability, traffic liability and fault-based liability frameworks are needed, including the Product Liability Directive and the Motor Insurance Directive.
- The adoption of AVs will entail new risks, including those that are unknown at the time of production and can only emerge after market launch. Policymakers should define new balanced and innovative frameworks to accommodate insurance and liability costs between consumers and injured parties on the one hand, and AVs providers on the other.

As we have seen, the direct application of all requirements for trustworthy AI in the context of AVs involves addressing a multitude of areas of different nature, some of them still at a very early stage of technological maturity.

The interrelationship and dependency between the various requirements and criteria suggests that holistic approaches are the most appropriate. However, their practicality is limited as they are the most complex to implement, requiring expertise from many different disciplines. More specific approaches, tailored to each requirement, may be a more suitable first step.

AI systems integrated in AVs will be operating in a high-risk scenario. Therefore, as a next step within the general ecosystem of excellence and trust that the EU is developing for AI, it would be appropriate to study the adaptation and fit that all the requirements and criteria addressed in this report would have within the new regulatory framework proposal on AI of the Commission (the AI Act, COM(2021) 206), irrespective of the existence of another specific sectoral framework in the field of automated transport.

References

ACEA, 'Articulated Pedestrian Target Specifications', *European Automobile Manufacturers Association*, 2015.

ACEA, 'Bicyclist target ACEA specifications', *European Automobile Manufacturers Association*, 2018.

AdaptIVe, 'AdaptIVe Automated Driving. Project Objectives'. `https://www.adaptive-ip.eu/index.php/objectives.html`, 2017. Accessed: 2021-05-19.

ADB, 'Minimal risk condition behaviors', *Automated Driving Behaviors (ADB) Consortium*, 2019.

AI HLEG, 'Ethics guidelines for trustworthy ai', *High Level Expert Group on Artificial Intelligence. B-1049 Brussels*, 2019.

AI HLEG, 'The assessment list for trustworthy artficial intelligence (altai) for self assessment.', *High Level Expert Group on Artificial Intelligence. B-1049 Brussels*, 2020.

Alam, A., Besselink, B., Turri, V., MaRtensson, J. and Johansson, K. H., 'Heavy-duty vehicle platooning for sustainable freight transportation: A cooperative method to enhance safety and efficiency', *IEEE Control Systems Magazine*, Vol. 35, No 6, 2015, pp. 34–56.

Alnasser, A., Sun, H. and Jiang, J., 'Cyber security challenges and solutions for v2x communications: A survey', *Computer Networks*, Vol. 151, 2019, pp. 52–67.

Alonso, M., Ciuffo, B., Makridis, M. and Thiel, C., 'The r-evolution of driving: from connected vehicles to coordinated automated road transport (c-art), part i: Framework for a safe & efficient coordinated automated road transport (c-art) system', Tech. Rep. EUR 28575 EN, Joint Research Center, 2017.

Amer, N., Zamzuri, H., Hudha, K. and Kadir, Z. A., 'Modelling and control strategies in path tracking control for autonomous ground vehicles: A review of state of the art and challenges', *Journal of Intelligent & Robotics Systems*, Vol. 86, 2017, pp. 225–254.

Amershi, S. and et al., 'Guidelines for human-ai interaction', *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Vol. 3, 2019, pp. 1–13.

Aradi, S., 'Survey of deep reinforcement learning for motion planning of autonomous vehicles', *IEEE Transactions on Intelligent Transportation Systems*, 2020, pp. 1–20.

Aravantinos, V. and Diehl, F., 'Traceability of deep neural networks', *arXiv:1812.06744*, 2019.

Armstrong, K., Das, S. and Cresko, J., 'The energy footprint of automotive electronic sensors', *Sustainable Materials and Technologies*, Vol. 25, 2020.

Asghar, M. N., Kanwal, N., Lee, B., Fleury, M., Herbst, M. and Qiao, Y., 'Visual surveillance within the eu general data protection regulation: A technology perspective', *IEEE Access*, Vol. 7, 2019, pp. 111709–111726.

AutoDrive, 'ECSEL JU: Advancing fail-aware, fail-safe and fail-operational electronic components, systems, and architectures for highly and fully automated driving to make future mobility safer, more efficient, affordable, and end-user acceptable'. `https://autodrive-project.eu/`, 2020. Accessed: 2021-05-11.

AVL, 'DRIVINGCUBE - A new way to speed up the validation and approval process of ADAS/AD systems'. https://www.avl.com/-/avl-drivingcube, 2021.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F. and Rahwan, I., 'The moral machine experiment', *Nature*, Vol. 563, 2018, pp. 59–64.

Azevedo-Sa, H., Jayaraman, S., Esterwood, C. and et al, 'Real-time estimation of drivers' trust in automated driving systems', *International Journal of Social Robotics*, 2020.

Baldini, G., 'Testing and certification of automated vehicles (av) including cybersecurity and artificial intelligence aspects', *EUR 30472 EN, JRC121631*, 2020.

Barredo-Arieta, A. and et al., 'Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai', *Information Fusion*, Vol. 58, 2020, pp. 82–115.

Beede, D., Powers, R. and Ingram, C., 'The employment impact of autonomous vehicles', *Office of the Chief Economist. U.S. Department of Commerce*, 2018, pp. 1–33.

Beggiato, M., Pereira, M., Petzoldt, T. and Krems, J., 'Learning and development of trust, acceptance and the mental model of acc. a longitudinal on-road study', *Transportation Research Part F*, Vol. 35, 2015, pp. 75–84.

Berberian, B., Sarrazin, J. C., Le Blaye, P. and Haggard, P., 'Automation technology and sense of control: a window on human agency', *PLoS ONE*, Vol. 7, No 3, 2012.

Bidot, J., Biundo, S., Heinroth, T., Minker, W., Nothdurft, F. and Schattenberg, B., 'Verbal plan explanations for hybrid planning', *MKWI*, 2010, pp. 2309–2320.

Biondi, F., Alvarez, I. and Jeong, K.-A., 'Human-vehicle cooperation in automated driving: A multidisciplinary review and appraisal', *International Journal of Human-Computer Interaction*, Vol. 35, No 11, 2019, pp. 932–946.

Biparva, M., Fernández-Llorca, D., Izquierdo-Gonzalo, R. and Tsotsos, J. K., 'Video action recognition for lane-change classification and prediction of surrounding vehicles', *arXiv:2101.05043*, 2021.

Bissel, D., Birtchnell, T., Elliot, A. and Hsu, E. L., 'Autonomous automobilities: The social impacts of driverless vehicles', *Current Sociology*, Vol. 68, No 1, 2020, pp. 116–134.

Bojarski, M., Choromanska, A., Choromanski, K., Firner, B., Ackel, L. J., Muller, U., Yeres, P. and Zieba, K., 'Visualback-prop: Efficient visualization of CNNs for autonomous driving', *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 4701–4708.

Bonnefon, J. F., Shariff, A. and Rahwan, I., 'The social dilemma of autonomous vehicles', *Science*, Vol. 352, 2016, pp. 1573–1576.

Boriboonsomsin, K., Barth, M. J., Zhu, W. and Vu, A., 'Eco-routing navigation system based on multisource historical and real-time traffic information', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 13, No 4, 2012, pp. 1694–1704.

Brandao, M., 'Age and gender bias in pedestrian detection algorithms', *Workshop on Fairness Accountability Transparency and Ethics in Computer Vision at CVPR*, 2019.

Bridgelall, R. and Stubbing, E., 'Forecasting the effects of autonomous vehicles on land use', *Technological Forecasting and Social Change*, Vol. 163, No 120444, 2021.

Brown, T. B. and et al., 'Language models are few-shot learners', *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

BSI, 'Operational design domain (odd) taxonomy for an automated driving system (ads) - specification', *British Standards Institution, PAS 1883.*, 2020.

Cao, Y., Xiao, C., Cyr, B., Zhou, Y., Park, W., Rampazzi, S., Chen, Q. A., Fu, K. and Mao, Z. M., 'Adversarial sensor attack on lidar-based perception in autonomous driving', *arXiv:1907.06826*, 2019.

CARLA, 'The CARLA Autonomous Driving Challenge'. https://carlachallenge.org/, 2020.

Carmona, J., Guindel, C., Garcia, F. and Escalera, A., 'ehmi: Review and guidelines for deployment on autonomous vehicles', *Sensors*, Vol. 21, 2021.

Carvalho, S., Ahire, S., Huff, E. W. and Brinkley, J., 'Utt: A conceptual model to guide the universal design of autonomous vehicles', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2021.

Casner, S. M. and Hutchins, E. L., 'What Do We Tell the Drivers? Toward Minimum Driver Training Standards for Partially Automated Cars', *Journal of Cognitive Engineering and Decision Making*, Vol. 13, 2019, pp. 55–66.

CB Insights, '33 industries driverless cars will transform'. https://www.cbinsights.com/reports/CB-Insights_Industries-Disrupted-Driverless-Cars.pdf, 2021.

Cesafsky, L., Stayton, E. and Cefkin, M., 'Calibrating agency. human-autonomy teaming and the future of work amid highly automated systems', In 'Ethnographic Praxis in Industry Conference Proceedings (EPIC)', pp. 65–82.

Chakraborti, T., Sreedharan, S. and Kambhampati, S., 'The emerging landscape of explainable automated planning & decision making', *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence Survey track.*, 2020, pp. 4803–4811.

Chen, Y., Gonder, J., Young, S. and Wood, E., 'Quantifying autonomous vehicles national fuel consumption impacts: A data-rich approach', *Transportation Research Part A*, Vol. 122, 2019, pp. 134–145.

Chen, Y. and Meier, A., 'Fuel consumption impacts of auto roof racks', *Energy Policy*, Vol. 92, 2020, pp. 325–333.

Chen, Y. F., Everett, M., Liu, M. and How, J. P., 'Socially aware motion planning with deep reinforcement learning', In '2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)', pp. 1343–1350. .

Chong, Y. L., Leong, H., Lee, C. D. W. and Ang-Jr, M. H., 'Benchmarking sensing and motion planning algorithms for autonomous driving', *IROS 2020 Workshop on Benchmarking Progress in Autonomous Driving*, 2020.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B., 'The cityscapes dataset for semantic urban scene understanding', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Corn, P., 'Is tesla responsible for customers fatally misusing tesla autopilot?' https://www.motorbiscuit.com/is-tesla-responsible-for-customers-fatally-misusing-tesla-autopilot/, 2021.

Costa, S., Costa, N., Simoes, P., Ribeiro, N. and Arezes, P., 'Tackling autonomous driving challenges - how the design of autonomous vehicles is mirroring universal design', *Advances in Intelligent Systems and Computing*, Vol. 781, 2019.

Damacharla, P., Javaid, A. Y., Gallimore, J. J. and Devabhaktuni, V. K., 'Common metrics to benchmark human-machine teams (hmt): A review', *IEEE Access*, Vol. 6, 2018, pp. 38637–38655.

Davis, J. and Goadrich, M., 'The relationship between precision-recall and roc curves', *ICML '06: Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.

Daza, I. G., Hernandez, N., Bergasa, L. M., Parra, I., Yebes, J. J., Gavilan, M., Quintero, R., Llorca, D. F. and Sotelo, M. A., 'Drowsiness monitoring based on driver and driving data fusion', In '2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)', pp. 1199–1204.

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

Detjen, H., Faltaous, S., Pfleging, B., Geisler, S. and Schneegass, S., 'How to increase automated vehicles' acceptance through in-vehicle interaction design: A review', *International Journal of Human-Computer Interaction*, Vol. 37, No 4, 2021, pp. 308–330.

Dey, D., Habibovic, A., Löcken, A., Wintersberger, P., Pfleging, B., Riener, A., Martens, M. and Terken, J., 'Taming the eHMI jungle: A classification taxonomy to guide, compare, and assess the design principles of automated vehicles' external human-machine interfaces', *Transportation Research Interdisciplinary Perspectives*, Vol. 7, No 100174, 2020.

Dibaei, M., Zheng, X., Jiang, K., Abbas, R., Liu, S., Zhang, Y., Xiang, Y. and Yu, S., 'Attacks and defences on intelligent connected vehicles: a survey', *Digital Communications and Networks*, Vol. 6, 2020, pp. 399–421.

Dollar, P., Wojek, C., Schiele, B. and Perona, P., 'Pedestrian detection: An evaluation of the state of the art', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, No 4, 2012, pp. 743–761.

Doshi-Velez, F. and Kim, B., 'Towards a rigorous science of interpretable machine learning', *arXiv:1702.08608v2*, 2017.

Doshi-Velez, F. and Kortz, M. A., 'Accountability of ai under the law: The role of explanation', *Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper.*, 2017.

Dosilovic, F. K., Brcic, M. and Hlupic, N., 'Explainable artificial intelligence: A survey', *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018, pp. 0210–0215.

Du, L. and Ling, H., 'Preservative license plate de-identification for privacy protection', *International Conference on Document Analysis and Recognition*, 2011, pp. 468–472.

EDPB, 'Guidelines 3/2019 processing of personal data through video devices', *European Data Protection Board, Version 2.0, Adopted on 29 January 2020*, 2020.

EDPB, 'Guidelines 1/2020 on processing personal data in the context of connected vehicles and mobility related applications', *European Data Protection Board, Version 2.0, Adopted on 9 March 2021*, 2021.

El-Rewini, Z., Sadatsharan, K., Selvaraj, D. F., Plathottam, S. J. and Ranganathan, P., 'Cybersecurity challenges in vehicular communications', *Computer Networks*, Vol. 23, 2020.

Eliot, L., 'Explaining Why Explainable AI (XAI) Is Needed For Autonomous Vehicles And Especially Self-Driving Cars'. https://www.forbes.com/sites/lanceeliot/2021/04/24/explaining-why-explainable-ai-xai-is-needed-for-autonomous-vehicles-and-especially-self-driving-cars/, 2021.

Emzivat, Y., Ibanez-Guzman, J., Martinet, P. and Roux, O. H., 'Dynamic driving task fallback for an automated driving system whose ability to monitor the driving environment has been compromised', *Proceedings of IEEE Intelligent Vehicles Symposium*, 2017, pp. 1841–1847.

ENABLE-S3 project, 'European initiative to enable validation for highly automated safe and secure systems'. https://enable-s3.eu/, 2019.

Endsley, M. R., 'Situation Awareness in Future Autonomous Vehicles: Beware of the Unexpected', *Advances in Intelligent Systems and Computing*, Vol. 824, 2019, pp. 303–309.

ENISA, 'ENISA Good Practices for Security of Smart Cars', *European Union Agency for Cybersecurity*, 2019.

EPRS, 'A common eu approach to liability rules and insurance for connected and autonomous vehicles', *European Added Value Assessment. European Parliamentary Research Service*, 2018.

Etienne, H., 'The dark side of the moral machine and the fallacy of computational ethical decision-making for autonomous vehicles', *Law, Innovation and Technology*, Vol. 13, 2021, pp. 85–107.

EU Member States, 'Declaration of amsterdam "cooperation in the field of connected and automated driving"'. https://www.rijksoverheid.nl/documenten/rapporten/2016/04/29/declaration-of-amsterdam-cooperation-in-the-field-of-connected-and-automated-driving, 2016.

Euro NCAP, 'Global Vehicle Target Specification', *Technical Bulletin*, 2018.

Euro NCAP, 'Asessment Protocol - Vulnerable Road User Protection', *European New Car Assessment Programme*, 2020.

Euro NCAP, 'Asessment Protocol - Safety Assist', *European New Car Assessment Programme*, 2021.

European Commission, 'Artificial intelligence for europe', *COM(2018) 237 final*, 2018.

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

Everett, M., Boulay, E. and Brown, H., 'Driving Data Compliance'. https://hsfnotes.com/cav/tag/gdpr/, 2019.

Everingham, M., Eslami, S., Van Gool, L. and et al, 'The pascal visual object classes challenge: A retrospective', *International Journal of Computer Vision*, Vol. 111, 2015, pp. 98–136.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. and Song, D., 'Robust physical-world attacks on deep learning visual classification', *arxiv:1707.08945*, 2018.

Fagnant, D. J. and Kockelman, K., 'Preparing a nation for autonomous vehicles: opportunities,barriers and policy recommendations', *Transportation Research Part A*, Vol. 77, 2015, pp. 167–181.

Fayyad, J., Jaradat, M. A., Gruyer, D. and Najjaran, H., 'Deep learning sensor fusion for autonomous vehicle perception and localization: A review', *Sensors*, Vol. 20, No 15, 2020.

Ferati, M., Murano, P. and Giannoumis, G. A., 'Universal design of user interfaces in self-driving cars', *Advances in Intelligent Systems and Computing*, Vol. 587, 2017, pp. 220–228.

Fernández Llorca, D., García Daza, I., Hernández Parra, N. and Parra Alonso, I., 'Sensors and sensing for intelligent vehicles', *Sensors*, Vol. 20, No 18, 2020.

Fernandez-Llorca, D., Quintero Minguez, R., Parra Alonso, I., Fernandez Lopez, C., Garcia Daza, I., Sotelo, M. A. and Cordero, C. A., 'Assistive intelligent transportation systems: The need for user localization and anonymous disability identification', *IEEE Intelligent Transportation Systems Magazine*, Vol. 9, No 2, 2017, pp. 25–40.

Figliozzi, M. A., 'Carbon emissions reductions in last mile and grocery deliveries utilizing air and ground autonomous vehicles', *Transportation Research Part D: Transport and Environment*, Vol. 85, No 102443, 2020.

Fischhoff, B., 'Communicating risks and benefits: An evidence-based user's guide', *US Department of Health and Human Services. Food and Drug Administration*, 2011.

Flaherty, N., 'Self-healing software for driverless cars'. https://www.eenewseurope.com/news/self-healing-software-driverless-cars, 2020. Accessed: 2021-05-11.

Frome, A., Cheung, G., Abdulkader, A., Zennaro, M., Wu, B., Bissacco, A., Adam, H., Neven, H. and Vincent, L., 'Large-scale privacy protection in google street view', *IEEE 12th International Conference on Computer Vision*, 2009, pp. 2373–2380.

Gabler, H. C., Hinch, J. A. and Steiner, J., 'Event data recorder: A decade of innovation', *Warrendale, PA: SAE Int.*, 2008.

Gaeta, M. C., 'Data protection and self-driving cars: the consent to the processing of personal data in compliance with gdpr', *Communications Law. The Journal of Computer, Media and Telecommunications Law*, Vol. 24, 2019.

Galassi, M. C. and Lagrange, A., 'New approaches for automated vehicles certification: Part i - current and upcoming methods for safety assessment', *EUR 30087 EN, JRC119345*, 2020.

Gallup, A. C., Chong, A., Kacelnik, A., Krebs, J. R. and Couzin, I. D., 'The influence of emotional facial expressions on gaze-following in grouped and solitary pedestrians', *Sci Rep*, Vol. 4, 2014.

García-Daza, I., Rentero, M., Salinas-Maldonado, C., Izquierdo-Gonzalo, R., Hernández-Parra, N., Ballardini, A. and Fernández-Llorca, D., 'Fail-aware lidar-based odometry for autonomous vehicles', *Sensors*, Vol. 20, No 15, 2020.

Garidis, K., Ulbricht, L., Rossmann, A. and Schäh, M., 'Toward a user acceptance model of autonomous driving', *In Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020, pp. 1381–1390.

German Ethics Commission, 'Ethical rules for automated and connected vehicular traffic', *German Ethics Commission on Automated and Connected Driving, Report*, 2017.

German Federal Government, 'Draft of an act amending the road traffic act and the compulsory insurance act - the autonomous driving act'. https://ec.europa.eu/growth/tools-databases/tris/en/search/?trisaction=search.detail&year=2021&num=81, 2021.

Ghosal, A. and Conti, M., 'Security issues and challenges in v2x: A survey', *Computer Networks*, Vol. 169, 2020.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. and Kagal, L., 'Explaining explanations: An approach to evaluating interpretability of machine learning', *arXiv:1806.00069*, 2018.

Gittleman, M. and Monaco, K., 'Truck-driving jobs: are they headed for rapid elimination?', *ILR Review*, Vol. 73, No 1, 2020, pp. 3–24.

Goldgof, S. S. A. D. B., Mouton, P. R. and Hall, L. O., 'Challenges for the repeatability of deep learning models', *IEEE Access*, Vol. 8, 2020, pp. 211860–211868.

González, D., Pérez, J., Milanés, V. and Nashashibi, F., 'A review of motion planning techniques for automated vehicles', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 17, No 4, 2016, pp. 1135–1145. .

Gore, E. and Wild, B., 'Stakeholders must engage policymakers on driverless cars'. https://thehill.com/blogs/congress-blog/technology/306670-stakeholders-must-engage-policymakers-on-driverless-cars, 2016.

Gotel, O. and et al., 'Traceability fundamentals', *Software and Systems Traceability. Springer*, 2012, pp. 3–22.

Greenblatt, J. B. and Shaheen, S., 'Automated vehicles, on-demand mobility, and environmental impacts', *Curr. Sustain. Renew. Energy Rep.*, Vol. 2, 2015, pp. 74–81.

Gross, J. N. and Humphreys, T. E., 'Gnss spoofing, jamming, and multipath interference classification using a maximum-likelihood multi-tap multipath estimator', *Proceedings of the 2017 International Technical Meeting of The Institute of Navigation*, 2017, pp. 662–670.

GRSC, 'Statement by the german road safety council dvr on edr and dssad', *Resolution of 28th October 2020 on the basis of the recommendations of the Executive Committee Vehicle Technology*, 2020.

Haghighi, N., Chamberlin, R., Fayyaz, K. and Liu, C., 'Impact of shared autonomous vehicles on vehicle miles traveled in utah', *Utah Department of Transportation Research Division. Final Report.*, 2019.

Harper, C. D., Hendrickson, C. T., Mangones, S. and Samaras, C., 'Estimating potential increases in travel with autonomous vehicles for the non-driving, elderly and people with travel-restrictive medical conditions', *Transportation Research Part C*, Vol. 72, 2016, pp. 1–9.

He, Q., Meng, X. and Qu, R., 'Survey on cyber security of cav', In 'CPGPS 2017 Forum on Cooperative Positioning and Service', pp. 351–354.

Hecht, T., Feldhütter, A., Radlmayr, J., Nakano, Y., Miki, Y., Henle, C. and Bengler, K., 'A review of driver state monitoring systems in the context of automated driving', In 'Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)', pp. 398–408.

Heer, J., 'Agency plus automation: Designing artificial intelligence into interactive systems', *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 116, No 6, 2019, pp. 1844–1850.

Heinrich, S., 'Flash memory in the emerging age of autonomy', *Flash Memory Summit*, 2017.

HERE, 'HERE HD Live Map. A self-healing map for reliable autonomous driving'. `https://engage.here.com/hubfs/Downloads/Tech%20Briefs/HERE%20Technologies%20Self-healing%20Map%20Tech%20Brief.pdf?t=1537438054632`, 2017. Accessed: 2021-05-11.

Hodicky, J., Prochazka, D. and Prochazka, J., 'Training with and of autonomous system – modelling and simulation approach', In 'Modelling and Simulation for Autonomous Systems', pp. 383–391.

Hollnagel, E., 'Resilience Engineering'. `https://erikhollnagel.com/ideas/resilience-engineering.html`, 2016. Accessed: 2021-05-11.

Hukkelås, H., Mester, R. and Lindseth, F., 'Deepprivacy: A generative adversarial network for face anonymization', *Advances in Visual Computing*, 2019, pp. 565–578.

Hulshof, W., Knight, I., Edwards, A., Avery, M. and Grover, C., 'Autonomous emergency braking test results', *Proceedings of the 23rd International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, 2013.

Hupont, I., Cerezo, E. and Baldassarri, S., 'Sensing facial emotions in a continuous 2d affective space', In '2010 IEEE International Conference on Systems, Man and Cybernetics', pp. 2045–2051.

Iagnemma, K., 'Why we have the ethics of self-driving cars all wrong'. https://www.weforum.org/agenda/2018/01/why-we-have-the-ethics-of-self-driving-cars-all-wrong/, 2018.

Ilievski, M., 'Wisebench: A motion planning benchmarking framework for autonomous vehicles', *Master Thesis in Computer Science. University of Waterloo*, 2020.

Inagaki, T. and Sheridan, T. B., 'A critique of the SAE conditional driving automation definition, and analyses of options for improvement', *Cognition, Technology & Work*, Vol. 21, 2019, pp. 569–578.

ITU-T FG-AI4AD, 'Terms of Reference ITU-T Focus Gropu on AI for autonomous and assisted driving', *ITU-T SG16*, 2019.

Izquierdo, R., Quintanar, A., Lorenzo, J., García-Daza, I., Parra, I., Fernández-Llorca, D. and Sotelo, M. A., 'Vehicle lane-change prediction on highways using efficient environment representation and deep learning', *IEEE Access*, 2021.

Izquierdo, R., Quintanar, A., Parra, I., Fernández-Llorca, D. and Sotelo, M. A., 'Vehicle trajectory prediction in crowded highway scenarios using bird eye view representations and cnns', *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2020.

Jafary, B., Rabiei, E., Diaconeasa, M. A., Masoomi, H., Fiondella, L. and Mosleh, A., 'A survey on autonomous vehicles interactions with human and other vehicles', *Probabilistic Safety Assessment and Management PSAM*, 2018.

Jain, A., Pero, L. D., Grimmett, H. and Ondruska, P., 'Autonomy 2.0: Why is self-driving always 5 years away?', *arXiv:2107.08142v3*, 2021.

Jaques, A. E., 'Why the moral machine is a monster', *University of Miami School of Law*, 2019.

J.Bahamonde-Birke, F., Goletz, M. and Ettema, D., 'The provision of mobility as a service with autonomous vehicles. the necessity of regulatory schemes for a natural monopoly', *Research in Transportation Economics*, 2020.

Jin, Q., Toth, A. R., Schultz, T. and Black, A. W., 'Speaker de-identification via voice transformation', *IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2009, pp. 529–533.

Johnsen, S. O. and Kilskar, S. S., 'A review of resilience in autonomous transport to improve safety and security', *in Proceedings of the 30th European Safety and Reliability Conference*, 2020.

Jourabloo, A., Yin, X. and Liu, X., 'Attribute preserved face de-identification', *International Conference on Biometrics (ICB)*, 2015, pp. 278–285.

Kalra, N. and Groves, D. G., 'The enemy of good: Estimating the cost of waiting for nearly perfect automated vehicles', *RAND Corporation*, 2017.

Kalra, N. and Paddock, S. M., 'Driving to safety: How many miles of driving would it take todemonstrate autonomous vehicle reliability?', *Transportation Research Part A*, Vol. 94, 2016, pp. 182–193.

Keller, K., Zimmermann, C., Zibuschka, J. and Hinz, O., 'Trust is good, control is better - customer preferences regarding control in teleoperated and autonomous vehicles', In '54th Hawaii International Conference on System Sciences, 2021', pp. 1849–1858.

Kim, H. J. and Yang, J. H., 'Takeover Requests in Simulated Partially Autonomous Vehicles Considering Human Factors', *IEEE Transactions on Human-Machine Systems*, Vol. 47, 2017, pp. 735–740.

Kim, J., Rohrbach, A., Darrell, T., Canny, J. and Akata, Z., 'Textual explanations for self-driving vehicles', *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 563–578.

Kim, J. W. and Yang, J. H., 'Understanding metrics of vehicle control take-over requests in simulated automated vehicles', *International Journal of Automotive Technology*, Vol. 21, 2020, pp. 757–770.

Kim, K., Kim, J. S., Jeong, S., Park, J.-H. and Kim, H. K., 'Cybersecurity for autonomous vehicles: Review of attacks and defense', *Computers & Security*, Vol. 103, 2021.

Kiran, B. R. and et al., 'Deep reinforcement learning for autonomous driving: A survey', *IEEE Transactions on Intelligent Transportation Systems, doi: 10.1109/TITS.2021.3054625*, 2021.

Kopelias, P., Demiridi, E., Vogiatzis, K., Skabardonis, A. and Zafiropoulou, V., 'Connected & autonomous vehicles - environmental impacts - a review', *Science of the Total Environment*, Vol. 712, 2020.

Korpan, R. and Epstein, S. L., 'Toward natural explanations for a robot's navigation plans', *Human-Robot Interaction (HRI)*, 2018.

Koulu, R., 'Human control over automation: EU policy and AI ethics', *European journal of legal studies*, Vol. 12, No 1, 2020, pp. 9–46.

Kümmerle, R., Steder, B., Dornhege, C., Ruhnke, M., Grisetti, G., Stachniss, C. and Kleiner, A., 'On measuring the accuracy of slam algorithms', *Autonomous Robots*, Vol. 27, 2009.

Kuutti, S., Fallah, S., Katsaros, K., Dianati, M., Mccullough, F. and Mouzakitis, A., 'A survey of the state-of-the-art localization techniques and their potentials for autonomous vehicle applications', *IEEE Internet of Things Journal*, Vol. 5, No 2, 2018, pp. 829–846.

Lategahn, H. and Stiller, C., 'Vision-only localization', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 15, No 3, 2014, pp. 1246–1257.

Laurier, E., 'Doing office work on the motorway', *Theory, Culture & Society*, Vol. 21, No 4-5, 2004, pp. 261–277.

Le, M.-H., Khan, M. S. N., Tsaloli, G., Carlsson, N. and Buchegger, S., 'Anonfaces: Anonymizing faces adjusted to constraints on efficacy and security', *WPES'20: Proceedings of the 19th Workshop on Privacy in the Electronic Society*, 2020, pp. 87–100.

Lefevre, S., Vasquez, D. and Laugier, C., 'A survey on motion prediction and risk assessment for intelligent vehicles', *Robomech J*, Vol. 1, No 1, 2014.

Levinson, J. and Thrun, S., 'Robust vehicle localization in urban environments using probabilistic maps', In '2010 IEEE International Conference on Robotics and Automation', pp. 4372–4378.

Li, M., Holthausen, B. E., Stuck, R. E. and Walker, B. N., 'No risk no trust: Investigating perceived risk in highly automated driving', *11th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI'19*, 2019, pp. 177–185.

Li, T. and Lin, L., 'Anonymousnet: Natural face de-identification with measurable privacy', *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P. and Zitnick, L. C., 'Microsoft coco: Common objects in context', *In European Conference on Computer Vision*, 2014.

Linkov, V., Zámecník, P., Havlícková, D. and Pai, C.-W., 'Human factors in the cybersecurity of autonomous vehicles: Trends in current research', *Frontiers in Psychology*, Vol. 10, 2019.

Lipson, H. and Kurman, M., 'Driverless: Intelligent cars and the road ahead', *Cambridge, MA: MIT Press*, 2016.

Liu, P., Yang, R. and Xu, Z., 'How safe is safe enough for self-driving vehicles?', *Risk Analysis*, Vol. 39, 2018, pp. 315–325.

Lorenzo, J., Parra, I., Wirth, F., Stiller, C., Llorca, D. F. and Sotelo, M. A., 'Rnn-based pedestrian crossing prediction using activity and pose-related features', *IEEE Intelligent Vehicles Symposium (IV)*, 2020, pp. 1801–1806.

Lorenzo, J., Parra-Alonso, I., Izquierdo, R., Ballardini, A. L., Hernández-Saz, A., Fernández-Llorca, D. and Sotelo, M. A., 'Capformer: Pedestrian crossing action prediction using transformer', *Sensors*, Vol. 21, 2021.

Lutz, A., Schick, B., Holzmann, H., Kochem, M., Meyer-Tuve, H., Lange, O., Mao, Y. and Tosolin, G., 'Simulation methods supporting homologation of electronic stability control in vehicle variants', *Vehicle System Dynamics*, Vol. 55, No 10, 2017, pp. 1432–1497.

Ma, Y., Wang, Z., Yang, H. and Yang, L., 'Artificial intelligence applications in the development of autonomous vehicles: a survey', *IEEE/CAA Journal of Automatica Sinica*, Vol. 7, No 2, 2020, pp. 315–329.

Magarinos, M., 'Voice personalization and speaker de-identification in speech processing systems', *Doctoral Dissertation. University of Vigo*, 2019.

Marko, N., Mohlmann, E., Nickovic, D., Niehaus, J., Priller, P. and Rooker, M., 'Challenges of engineering safe and secure highly automated vehicles', *arXiv, 2103.03544*, 2021.

Maro, S., 'Addressing traceability challenges in the development of embedded systems', *Thesis for The Degree of Licentiate of Engineering. Chalmers University of Technology and Göteborg University*, 2017.

Maro, S., Steghöfer, J.-P. and Staron, M., 'Software traceability in the automotive domain: Challenges and solutions', *Journal of Systems and Software*, Vol. 141, 2017, pp. 85–110.

Martínez-Plumed, F., Gómez, E. and Hernández-Orallo, J., 'AI Watch: Assessing Technology Readiness Levels for Artificial Intelligence', *EUR 30401 EN, Publications Office of the European Union, JRC122014*, 2020.

Martínez-Plumed, F., Gómez, E. and Hernández-Orallo, J., 'Futures of artificial intelligence through technology readiness levels', *Telematics and Informatics*, Vol. 58, 2021.

Martins, H. R., 'Type approval homologation and self certification', *Ford Company, DOI: 10.13140/RG.2.2.31708.39041*, 2010.

Matute-Peaspan, J. A., Perez, J. and Zubizarreta, A., 'A fail-operational control architecture approach and dead-reckoning strategy in case of positioning failures', *Sensors*, Vol. 20, No 2, 2020.

McRaney, D., 'You can beat your brain: How to turn your enemies into friends, how to make better decisions, and other ways to be less dumb', *London: Oneworld Publications*, 2013.

Michon, J. A., 'A critical view of driver behavior models: What do we know, what should we do?', *Human Behavior and Traffic Safety*, 1985, pp. 485–524.

Microsoft, 'Turing-NLG: A 17-billion-parameter language model by Microsoft'. https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/, 2020.

Millard-Ball, A., 'Pedestrians, autonomous vehicles, and cities', *Journal of Planning Education and Research*, Vol. 38, No 1, 2018, pp. 6–12.

Moore, J. W., 'What is the sense of agency and why does it matter?', *Frontiers in Psychology*, Vol. 7, No 1272, 2016.

Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E. and Sicilia, M. A., 'Traceability for trustworthy ai: A review of models and tools', *Big Data Cogn. Comput.*, Vol. 5, 2021.

Morrow, W. R. and et al., 'Key factors influencing autonomous vehicles' energy and environmental outcome', *Road Vehicle Automation. Lecture Notes in Mobility. Springer*, 2014.

Moura, N., Chatila, R., Evans, K., Chauvier, S. and Dogan, E., 'Ethical decision making for autonomous vehicles', *IEEE Intelligent Vehicle Symposium (IV)*, 2020, pp. 2006–2013.

Mulder, T. and Vellinga, N. E., 'Exploring data protection challenges of automated driving', *Computer Law & Security Review*, Vol. 40, 2021.

Nahavandi, S., 'Trusted Autonomy Between Humans and Robots. Toward Human-on-the-Loop in Robotics and Autonomous Systems', *IEEE Systems, Man and Cybernetics Magazine*, Vol. January, 2017.

Nassi, D., Ben-Netanel, R., Elovici, Y. and Nassi, B., 'Mobilbye: Attacking adas with camera spoofing', *arXiv:1906.09765*, 2019.

Nastjuk, I., Herrenkind, B., Marrone, M., Brendel, A. B. and M.Kolbea, L., 'What drives the acceptance of autonomous driving? an investigation of acceptance factors from an end-user's perspective', *Technological Forecasting and Social Change*, Vol. 161, No 120319, 2020.

Navarro, J., Francois, M. and Mars, F., 'Obstacle avoidance under automated steering: Impact on driving and gaze behaviours', *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 43, 2016, pp. 315–324.

Neumeier, S., Gay, N., Dannheim, C. and Facchi, C., 'On the way to autonomous vehicles teleoperated driving', *AmE 2018 - Automotive meets Electronics; 9th GMM-Symposium*, 2018.

NHTSA, 'Preliminary statement of policy concerning automated vehicles', *National Highway Traffic Safety Administration*, 2013.

OICA, 'Data storage system for automated driving (dssad)', *14th IWG ITS/AD, Document No. ITS/AD-14-09*, 2018.

OICA, 'Event data recorder (edr) & data storage system for automated driving (dssad)', *EDR/DSSAD 1st session, EDR-DSSAD-01-04*, 2019.

Omeiza, D., Webb, H., Jirotka, M. and Kunze, L., 'Explanations in autonomous driving: A survey', *IEEE Transactions on Intelligent Transportation Systems*, 2021.

Paden, B., Cap, M., Yong, S. Z., Yershov, D. and Frazzoli, E., 'A survey of motion planning and control techniques for self-driving urban vehicles', *IEEE Transactions on Intelligent Vehicles*, Vol. 1, No 1, 2016, pp. 33–55. .

Padilla, R., Passos, W. L., Dias, T. L. B., Netto, S. L. and da Silva, E. A. B., 'A comparative analysis of object detection metrics with a companion open-source toolkit', *Electronics*, Vol. 10, No 3, 2021.

Pakusch, C., Stevens, G. and Bossauer, P., 'Shared autonomous vehicles: Potentials for a sustainable mobility and risks of unintended effects', *EPiC Series in Computing*, Vol. 52, 2018, pp. 258–269.

Pan, X., You, Y., Wang, Z. and Lu, C., 'Virtual to real reinforcement learning for autonomous driving', *arXiv, 1704.03952*, 2017.

Parasuraman, R., Sheridan, T. B. and Wickens, C. D., 'A general model of mixed-initiative human-machine systems', In '33rd Annual Human Factors Society Conf.', pp. 124–128.

Parasuraman, R., Sheridan, T. B. and Wickens, C. D., 'A model for types and levels of human interaction with automation', *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, Vol. 30, No 3, 2000, pp. 286–297.

Parra, I., Corrales, H., Hernández, N., Vigre, S., Llorca, D. F. and Sotelo, M. A., 'Performance analysis of vehicle-to-vehicle communications for criticaltasks in autonomous driving', In 'IEEE Intelligent Transportation Systems Conference (ITSC)', pp. 195–200.

Parra, I., Izquierdo, R., Alonso, J., García, A., F-Llorca, D. and Sotelo, M. A., 'The experience of drivertive-driverlesscooperative vehicle-team in the 2016 gcdc', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 19, No 4, 2017, pp. 1322–1334.

Parra Alonso, I., Fernández Llorca, D. F., Gavilan, M., Alvarez Pardo, S., Garcia-Garrido, M. A., Vlacic, L. and Sotelo, M. A., 'Accurate global localization using visual odometry and digital maps on urban environments', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 13, No 4, 2012, pp. 1535–1545.

Paschalidis, E., Hajiseyedjavadi, F., Wei, C., Solernou, A., Hamish Jamson, A., Merat, N., Romano, R. and Boer, E. R., 'Deriving metrics of driving comfort for autonomous vehicles: A dynamic latent variable model of speed choice', *Analytic Methods in Accident Research*, Vol. 28, 2020, p. 100133.

PAVE, 'Avs for all: Inspiring solutions for accessible design'. https://pavecampaign.org/avs-for-all-inspiring-solutions-for-accessible-design-additional-resources/, 2020.

Paz, D., Lai, P.-J., Chan, N., Jiang, Y. and Christensen, H. I., 'Autonomous vehicle benchmarking using unbiased metrics', *In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 6223–6228.

PEGASUS project, 'Project for the establishment of generally accepted quality criteria, tools and methods as well as scenarios and situations'. https://www.pegasusprojekt.de/en/about-PEGASUS, 2019.

Pek, C., Rusinov, V., Manzinger, S., Üste, M. C. and Althoff, M., 'Commonroad drivability checker: Simplifying the development and validation of motion planning algorithms', *IEEE Intelligent Vehicle Symposium*, 2020, pp. 1013–1020.

Perterer, N., Moser, C., Meschtscherjakov, A., Krischkowsky, A. and Tscheligi, M., 'Activities and technology usage while driving: A field study with private short-distance car commuters', *Proceedings of the 9th Nordic Conference on Human-Computer Interaction (NordiCHI)*, Vol. 41, 2016, pp. 1–10.

Pfeiffer, M., Schaeuble, M., Nieto, J., Siegwart, R. and Cadena, C., 'From perception to decision: A data-driven approach to end-to-end motion planning for autonomous ground robots', In '2017 IEEE International Conference on Robotics and Automation (ICRA)', pp. 1527–1533. .

Pham, M. and Xiong, K., 'A survey on security attacks and defense techniques for connected and autonomous vehicles', *arXiv:2007.08041*, 2020.

Pina, P. E., Donmez, B. and Cummings, M. L., 'Selecting metrics to evaluate human supervisory control applications', *HAL2008-04, MIT Department of Aeronautics and Astronautics, Cambridge, MA 02139*, 2008.

Pool, E. A. I., Kooij, J. F. P. and Gavrila, D. M., 'Context-based cyclist path prediction using recurrent neural networks', *2019 IEEE Intelligent Vehicles Symposium (IV)*, 2019, pp. 824–830.

Qualcomm, 'Cellular-v2x technology overview', *80-PE732-63 Rev B*, 2019.

Quintero Mínguez, R., Parra Alonso, I., Fernández-Llorca, D. and Sotelo, M. A., 'Pedestrian path, pose, and intention prediction through gaussian process dynamical models and pedestrian activity recognition', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 20, No 5, 2019, pp. 1803–1814. .

Ranjan, A., Janai, J., Geiger, A. and Black, M. J., 'Attacking optical flow', *Proc. of the International Conference on Computer Vision (ICCV)*, 2019, pp. 2404–2413.

Rasouli, A., Kotseruba, I., Kunic, T. and Tsotsos, J. K., 'Pie: A large-scale dataset and models for pedestrian intention estimation andtrajectory prediction', *International Conference on Computer Vision (ICCV)*, 2019, pp. 6262–6271.

Rasouli, A. and Tsotsos, J. K., 'Autonomous vehicles that interact with pedestrians: A survey of theory and practice', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 21, No 3, 2020, pp. 900–918. .

Realpe, M., Vintimilla, B. X. and Vlacic, L., 'A fault tolerant perception system for autonomous vehicles', *35th Chinese Control Conference (CCC)*, 2016, pp. 6531–6536.

Roh, J., Paxton, C., Pronobis, A., Farhadi, A. and Fox, D., 'Conditional driving from natural language instructions', In 'Proceedings of the Conference on Robot Learning', , edited by L. P. Kaelbling, D. Kragic, and K. SugiuraVol. 100. pp. 540–551.

Rosenfeld, A. and Richardson, A., 'Explainability in humanâĂŞagent systems', *Auton Agent Multi-Agent Syst*, Vol. 33, 2019, pp. 673–705.

Ross, C. and Guhathakurta, S., 'Autonomous vehicles and energy impacts: A scenario analysis', *Energy Procedia*, Vol. 143, 2017, pp. 47–52.

Roy, A., 'How the Language of Self-Driving Is Killing Us. We can't fix something this broken. But we can and should replace it.' `https://www.thedrive.com/article/20495/how-the-language-of-self-driving-is-killing-us`, 2018. Accessed: 2021-02-19.

Rudenko, A., Palmieri, L., Herman, M., Kitani, K. M., Gavrila, D. M. and Arras, K. O., 'Human motion trajectory prediction: a survey', *The International Journal of Robotics Research*, Vol. 39, No 8, Jun 2020, pp. 895–935. ISSN 1741-3176. . URL `http://dx.doi.org/10.1177/0278364920917446`.

SAE International, 'Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles', *J3016 201806*, 2018.

SAE International, '(R) Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles', *J3016 202104*, 2021.

Saleh, K., Hossny, M. and Nahavandi, S., 'Towards trusted autonomous vehicles from vulnerable road users perspective', *Annual IEEE International Systems Conference (SysCon)*, 2017.

Schartmüller, C., Weigl, K., Wintersberger, P. and Riener, A., 'Text comprehension: Heads-up vs. auditory displays - implications for a productive work environment in sae level 3 automated vehicles', *11th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutoUI'19*, 2019.

Schartmuller, C., Wintersberger, P., Reiner, A., Kun, A. L., Brewster, S. and Shaer, O., 'Autowork 2020: Second workshop on the future of work and well-being in automated vehicles', *12th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI'20*, 2020, pp. 113–116.

Schartmüller, C., Wintersberger, P., Riener, A., Kun, A. L., Brewster, S. and Shaer, O., 'Autowork 2020: Second workshop on the future of work and well-being in automated vehicles', *12th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI'20*, 2020.

Schneider, T., Hois, J., Rosenstein, A., Ghellal, S., Theofanou-Fülbier, D. and Gerlicher, A. R., 'ExplAIn Yourself! Transparency for Positive UX in Autonomous Driving', *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Vol. 161, 2021, pp. 1–12.

Schöomig, N. and et al., 'Checklist for expert evaluation of hmis of automated vehicles-discussions on its value and adaptions of the method within an expert workshop', *Information*, Vol. 11, 2020.

Schram, R., 'Euro NCAP'S First Step to Assess Automated Driving Systems', *European New Car Assessment Programme, Paper Number 19-0292*, 2019.

Shariff, A., Bonnefon, J.-F. and Rahwan, I., 'How safe is safe enough? psychological mechanisms underlying extreme safety demands for self-driving cars', *Transportation Research Part C*, Vol. 126, 2021.

Sheridan, T. B. and Verplank, W. L., 'Human and computer control of undersea teleoperators', Tech. rep., MIT Man-Machine Systems Laboratory, Cambridge, 1978.

Shin, H., Kim, D., Kwon, Y. and Kim, Y., 'Illusion and dazzle: Adversarial optical channel exploits against lidars for automotive applications'. Cryptology ePrint Archive, Report 2017/613, 2017.

Silva, J., 'Increasing perceived agency in human-ai interactions. learnings from piloting a voice user interface with drivers on uber', In 'Ethnographic Praxis in Industry Conference Proceedings (EPIC)', pp. 441–456.

Stanton, N., Revell, K. M. A. and Langdon, P., 'Designing interaction and interfaces for automated vehicles: User-centred ecological design and testing', *CRC Press*, 2021.

Steinfeldet, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A. and Goodrich, M., 'Common metrics for human-robot interaction', *in Proc. 1st ACM SIGCHI/SIGART Conf. Human-Robot Interact.*, 2006, pp. 33–40.

Stephens, T., Gonder, J., Chen, Y., Lin, Z., Liu, C. and Gohlke, D., 'Estimated bounds and important factors for fuel use and consumer costs of connected and automated vehicles', *Technical Report. National Renewable Energy Laboratory (NREL)*, 2016.

Stogios, C., Kasraian, D., Roorda, M. J. and Hatzopoulou, M., 'Simulating impacts of automated driving behavior and traïňČc conditions on vehicle emissions', *Transportation Research Part D*, Vol. 76, 2019, pp. 176–192.

Strayer, D. L., Watson, J. M. and Drews, F. A., 'Cognitive distraction while multitasking in the automobile', *Psychology of Learning and Motivation*, Vol. 54, 2019, pp. 29–58.

Sun, J., Cao, Y., Chen, Q. A. and Mao, Z. M., 'Towards robust lidar-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures', *arXiv:2006.16974*, 2020.

Tate, L., Hochgreb, S., Hall, J. and Bassett, M., 'Energy efficiency of autonomous car powertrain', *SAE Technical Paper 2018-01-1092*, 2018.

Templeton, B., 'NHTSA/SAE's "levels" of robocars may be contributing to highway deaths'. `https://ideas.4brad.com/nhtsasaes-levels-robocars-may-be-contributing-highway-deaths`, 2018. Accessed: 2021-02-19.

Thorn, E., Kimmel, S. C., Chaka, M., Hamilton, B. A. and et al., 'A framework for automated driving system testable cases and scenarios', *United States. Department of Transportation. National Highway Traffic Safety ,Tech. Rep.*, 2018.

TKSL, 'Experimental security research of tesla autopilot', *Tencent Keen Security Lab*, 2019.

Toader, A. R., 'Auditability of ai systems - brake or acceleration to innovation?', *Liege Competition and Innovation Institute*, 2019.

Tolan, S., 'Fair and unbiased algorithmic decision making: Current state and future challenges', *JRC Technical Report. Working Paper*, 2018.

Toole, J., 'Deep learning has a size problem'. https://heartbeat.fritz.ai/deep-learning-has-a-size-problem-ea601304cd8, 2019.

Traft, N., Shridhar, S. and Haynes, G. C., 'Motion prediction for self-driving needs a metric specific to self-driving', *IROS 2020 Workshop on Benchmarking Progress in Autonomous Driving*, 2020.

Tsugawa, S., Jeschke, S. and Shladover, S. E., 'A review of truck platooning projects for energy savings', *IEEE Transactions on Intelligent Vehicles*, Vol. 1, No 1, 2016, pp. 68–77.

Tyagi, A. K. and Aswathy, S. U., 'Autonomous intelligent vehicles (aiv): Research statements, open issues, challenges and road for future', *International Journal of Intelligent Networks*, Vol. 2, 2021, pp. 83–102.

Ueda, S., Nakashima, R. and Kumada, T., 'Influence of levels of automation on the sense of agency during continuous action', *Nature Sci. Rep.*, Vol. 11, No 2436, 2021.

Uittenbogaard, R., Sebastian, C., Vijverberg, J., Boom, B., Gavrila, D. M. and de With, P., 'Privacy protection in street-view panoramas using depth and multi-view imagery', *Computer Vision and Pattern Recognition*, 2019, pp. 10581–10590.

UNECE WP29 GRVA, 'Future Certification of Automated/Autonomous Driving Systems', *Informal document GRVA-02-09*, 2019.

UNECE WP.29 GRVA, 'Common Functional Performance Requirements for Automated Driving Systems and ADS-Equipped Vehicles', *Informal document WP.29-180-10, Submitted by the IWG on Functional Requirements for Automated Vehicles (FRAV)*, 2020a.

UNECE WP.29 GRVA, 'UN Regulation No. 152. Uniform provisions concerning the approval of motor vehicles with regard to the Advanced Emergency Braking System (AEBS) for M1 and N1 vehicles', *United Nations. ECE/TRANS/WP.29/2020/9.*, 2020b.

UNECE WP.29 GRVA, 'UN Regulation No. 155 (Cyber security and cyber security management system).', *United Nations. ECE/TRANS/WP.29/2021/59*, 2021a.

UNECE WP.29 GRVA, 'UN Regulation No. 156 on software update and software update management system.', *United Nations. ECE/TRANS/WP.29/2021/60*, 2021b.

UNECE WP.29 GRVA, 'UN Regulation No. 157. Uniform provisions concerning the approval of vehicles with regard to Automated Lane Keeping Systems', *United Nations. E/ECE/TRANS/5050/rev.3/Add.156*, 2021c.

U.S. DOT, 'Inclusive design challenge'. https://www.transportation.gov/accessibility/inclusivedesign, 2020.

Uytsel, S. V., 'Different liability regimes for autonomous vehicles: One preferable above the other?', *Autonomous Vehicles. Perspectives in Law, Business and Innovation. Springer, Singapore*, 2021.

Veitas, V. K. and Delaere, S., 'In-vehicle data recording, storage and access management in autonomous vehicles', *arXiv:1806.03243*, 2018.

Venkita, S.-R., Boulkroune, B., Mishra, A. and van Nunen, E., 'A fault tolerant lateral control strategy for an autonomous four wheel driven electric vehicle', *2020 IEEE Intelligent Vehicles Symposium (IV)*, 2020, pp. 1492–1497.

Venkkataraman, S., 'Risk perception and communication - the effect of driverless cars on public', *Technical Report*, 2017.

Wadud, Z., MacKenzie, D. and Leiby, P., 'Help or hindrance? the travel, energy and carbon impacts of highly automated vehicles', *Transportation Research Part A: Policy and Practice*, Vol. 86, 2016, pp. 1–18.

Wallach, W. and Allen, C., 'Moral machines. teaching robots right from wrong', *Oxford University Press*, 2009.

Wang, W., Liu, C. and Zhao, D., 'How much data are enough? a statistical approach with case study on longitudinal driving behavior', *IEEE Transactions on Intelligent Vehicles*, Vol. 2, 2017, pp. 85–98.

Wen, W., Kuroki, Y. and Asama, H., 'The sense of agency in driving automation', *Frontiers in Psychology*, Vol. 10, No 2691, 2019.

Wen, W., Yamashita, A. and Asama, H., 'Measurement of the perception of control during continuous movement using electroencephalography', *Frontiers in Human Neuroscience*, Vol. 11, No 392, 2017.

Wilson, B., Hoffman, J. and Morgenstern, J., 'Predictive inequity in object detection', *Workshop on Fairness Accountability Transparency and Ethics in Computer Vision at CVPR*, 2019.

Wintersberger, P., Green, P. and Riener, A., 'Am i driving or are you or are we both? a taxonomy for handover and handback in automated driving', *Proceedings of the Ninth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, 2017, pp. 333–339.

Wu, Y., Yang, F., Xu, Y. and Ling, H., 'Journal of computer science andtechnology', *Privacy-Protective-GAN for privacy preserving face de-identification*, Vol. 34, 2019, pp. 47–60.

Xu, W., Yan, C., Jia, W., Ji, X. and Liu, J., 'Analyzing and enhancing the security of ultrasonic sensors for autonomous vehicles', *IEEE Internet of Things Journal*, Vol. 5, 2021, pp. 5015–5029.

Yacoub, S., Simske, S., Lin, X. and Burns, J., 'Recognition of emotions in interactive voice response systems', In 'In EUROSPEECH-2003', pp. 729–732.

Yadav, S. and Ansari, A., 'Autonomous vehicles camera blinding attack detection using sequence modelling and predictive analytics', *SAE Technical Paper*, 2020.

YAKINDU, 'Integrated Traceability Solutions for Development of Autonomous Vehicles with YAKINDU Traceability'. https://www.youtube.com/watch?v=QaXNfsCFwqA, 2020.

Yan, C., Xu, W. and Liu, J., 'Can you trust autonomous vehicles: Contactless attacks against sensors of self-driving vehicle', *DEF CON 24 Hacking Conference*, 2016.

Yeh, E. R., Choi, J., Prelcic, N. G., Bhat, C. R. and Heath-Jr., R. W., 'Security in automotive radar and vehicular networks', *Microwave Journal*, 2016.

Yoon, S. H., Kim, Y. W. and Ji, Y. G., 'The effects of takeover request modalities on highly automated car control transitions', *Accident Analysis and Prevention*, Vol. 123, 2019, pp. 150–158.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V. and Darrell, T., 'Bdd100k: A diverse driving dataset for heterogeneous multitask learning', *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2636–2645.

Yu, J. and Luo, F., 'Fallback strategy for level 4+ automated driving system', *Proceedings of IEEE Intelligent Transportation Systems*, 2019, pp. 156–162.

Yun, S., Wen, W., Q. An, Q. and et al., 'Investigating the relationship between assisted driver's SoA and EEG', In 'International Conference on NeuroRehabilitation', pp. 1039–1043.

Zablocki, E., Ben-Younes, H., Perez, P. and Cord, M., 'Explainability of vision-based autonomous driving systems: Review and challenges', *arXiv:2101.05307v1*, 2021.

Zeeb, K., Buchner, A. and Schrauf, M., 'What determines the take-over time? An integrated model approach of driver take-over after automated driving', *Accident Analysis and Prevention*, Vol. 78, 2015, pp. 212–221.

Zhang, B., Winter, J., Varotto, S., Happee, R. and Martens, M., 'Determinants of take-over time from automated driving: A meta-analysis of 129 studies', *Transportation Research Part F*, Vol. 64, 2019, pp. 285–307.

Zhang, W., Guhathakurta, S., Fang, J. and Zhang, G., 'Exploring the impact of shared autonomous vehicles on urban parking demand: An agent-based simulation approach', *Sustainable Cities and Society*, Vol. 19, 2015, pp. 34–45.

Zhang, Z. and Scaramuzza, D., 'A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry', *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 7244–7251.

Zue, W., Yang, B., Kaizuka, T. and Nakano, K., 'A fallback approach for an automated vehicle encountering sensor failure in monitoring environment', *Proceedings of IEEE Intelligent Vehicles Symposium*, 2018, pp. 1807–1812.

List of abbreviations and definitions

ADS Automated Driving System

AI Artificial Intelligence

AV Automated Vehicle / Autonomous Vehicle

CAV Connected and Automated Vehicle / Connected and Autonomous Vehicle

CNN Convolutional Neural Network

CR (Assessment) Criterion

DDT Dynamic Driving Task

DNN Deep Neural Network

EC European Commission

EDPB European Data Protection Board

ENISA European Union Agency for Cybersecurity

EP Ethical Principles

GNSS Global Navigation Satellite System

GPS Global Positioning System

HD High Definition

HMI Human-Machine Interface / Human-Machine Interaction

HVI Human-Vehicle Interface / Human-Vehicle Interaction

HLEG High Level Expert Group

IMU Inertial Measurement Unit

KR Key Requirement

LiDAR Light Detection And Ranging

NHTSA National Highway Traffic Safety Administration

OBD On-Board Diagnosis

ODD Operational Design Domain

OEDR Object and Event Detection and Response

SAE Society of Automotive Engineers

TOR Take Over Request

VRU Vulnerable Road User

V2D Vehicle-to-Device

V2I Vehicle-to-Infrastructure

V2N Vehicle-to-Network

V2P Vehicle-to-Pedestrian

V2V Vehicle-to-Vehicle

V2VRU Vehicle-to-Vulnerable Road User

V2X Vehicle-to-Everything

List of figures

List of tables

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

KJ-NA-30942-EN-N

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

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

EU Science Hub

ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office of the European Union

doi:10.2760/120385

ISBN 978-92-76-46055-8