International Transport Forum

AI Machine Learning and Regulation: The Case of Automated Vehicles

Summary and Conclusions

197 Roundtable

AI Machine Learning and Regulation: The Case of Automated Vehicles

Summary and Conclusions

197 Roundtable

The International Transport Forum

The International Transport Forum is an intergovernmental organisation with 69 member countries. It acts as a think tank for transport policy and organises the Annual Summit of transport ministers. ITF is the only global body that covers all transport modes. The ITF is politically autonomous and administratively integrated with the OECD.

The ITF works for transport policies that improve peoples' lives. Our mission is to foster a deeper understanding of the role of transport in economic growth, environmental sustainability and social inclusion and to raise the public profile of transport policy.

The ITF organises global dialogue for better transport. We act as a platform for discussion and pre-negotiation of policy issues across all transport modes. We analyse trends, share knowledge and promote exchange among transport decision makers and civil society. The ITF's Annual Summit is the world's largest gathering of transport ministers and the leading global platform for dialogue on transport policy.

The Members of the Forum are: Albania, Argentina, Armenia, Australia, Austria, Azerbaijan, Belarus, Belgium, Bosnia and Herzegovina, Brazil, Bulgaria, Cambodia, Canada, Chile, China (People's Republic of), Colombia, Costa Rica, Croatia, Czech Republic, Denmark, Dominican Republic, Estonia, Finland, France, Georgia, Germany, Greece, Hungary, Iceland, India, Ireland, Israel, Italy, Japan, Kazakhstan, Korea, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Mexico, Republic of Moldova, Mongolia, Montenegro, Morocco, the Netherlands, New Zealand, North Macedonia, Norway, Oman, Poland, Portugal, Romania, Russian Federation, Saudi Arabia, Serbia, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Tunisia, Türkiye, Ukraine, the United Arab Emirates, the United Kingdom, the United States and Uzbekistan.

International Transport Forum
2 rue André Pascal
F-75775 Paris Cedex 16
contact@itf-oecd.org
www.itf-oecd.org

ITF Disclaimer

Cite this work as: ITF (2025), AI, Machine Learning and Regulation: The Case of Automated Vehicles, OECD Publishing, Paris.

Acknowledgements

This report builds on expert discussions during an ITF Roundtable, "Artificial Intelligence, Machine Learning and Regulation", held in Paris and virtually on 26-27 January 2023. Dr. Markus Reinhardt (German Centre for Rail Traffic Research) chaired the Roundtable. The ITF would like to thank the high-level participation of the German Federal Ministry for Digital and Transport, highlighted by the opening remark of Mr Stefan Schnorr, the State Secretary of the German Federal Ministry for Digital and Transport, who reminded participants of the importance of AI in the future of transport.

At the ITF, Changgi Lee coordinated the project alongside Camille Combe. Changgi Lee and Philippe Crist authored the report with substantive inputs from Camille Combe. Eugine Ro helped with the co-ordination of the Roundtable event and took notes of discussions. Philippe Crist was responsible for overall quality control and Camille Larmanou managed the editorial process. Mila Iglesias and Apostolos Skourtas helped organise the Roundtable event. This Roundtable Report is part of the ITF's core Programme of Work for 2022-23, co-ordinated by Jagoda Egeland and Orla McCarthy, and has been approved by the ITF's Transport Research Committee.

The authors would like to thank Gianmarco Baldini (Joint Research Centre, European Commission), Florent Perronnin (Naver Labs), Martin Russ (AustriaTech), and William H. Widen (University of Miami) for their feedback on the report. The ITF would like to thank Markus Reinhardt for chairing the Roundtable. Thanks also to Gregorio Ameyugo (CEA List), Siddartha Khastgir (University of Warwick), Louise Dennis (University of Manchester), Latifa Oukhellou (University Gustave Eiffel), Marry "Missy" Commings (George Mason University), Gianmarco Baldini (Joint Research Centre, European Commission), Aida Joaquin Acosta (Ministry of Transport, Mobility and Urban Agenda, Spain), and Martin Russ (Austriatech) for their presentations during the Roundtable. Annex A lists the names and affiliations of the Roundtable participants.

Table of contents

Figures

Tables

Boxes

Abbreviations and acronyms

ADAS	Advanced driver assistance systems
ADS	Automated driving system
ADS-DV	Automated Driving Systems – Dedicated Vehicles
AI	Artificial intelligence
ASDE	Authorised Self-Driving Entity
AV	Automated Vehicle * In this report, AVs refer to highly or fully automated vehicles corresponding to SAE J3016 Level 4 and above that does not require human involvement in performing driving tasks within designated ODDs (see Box.1)
DDT	Dynamic driving task
DMV	Department of Motor Vehicle
EIBD	Explainable and Interpretable by Design
EM	Emergency Manoeuvre
FHWA	Federal Highway Administration
GDPR	General Data Protection Regulation (EU)
GPS	Global positioning system
ITF	International Transport Forum
LDM	Local dynamic map
LIDAR	Light detection and ranging
MRM	Minimum Risk Manoeuvre
NHTSA	National Highway Traffic Safety Administration
NUiC	No-user-in-charge
OBU	On-board unit
ODD	Operational design domain
OEM	Original equipment manufacturer
PTO	Public transport operator
R&D	Research and development
RSU	Roadside unit
SAE	Society of Automotive Engineers
TEVV	Test, Evaluation, Verification and Validation
UiC	User-in-charge
UNECE	United Nations Economic Commission for Europe
V2I	Vehicle-to-infrastructure
V2V	Vehicle-to-vehicle
V2X	Vehicle-to-everything

Glossary

Automated driving system	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. (SAE International, 2021a)
Authorised Self-Driving Entity	The entity that puts an AV forward for authorisation as having self-driving features. It may be the vehicle manufacturer, or a software designer, or a joint venture between the two. (Law Commission of England and Wales & Scottish Law Commission., 2022)
Automated Vehicle	A motor vehicle equipped with ADS and thus capable of performing dynamic driving tasks. (see Box.1 for further details on the usage of the word in this report)
Explainability	The property of an AI system to express important factors influencing the AI system results in a way that humans can understand. ISO/IEC 22989:2022(en), 3.5.7
Interpretability	The property of an AI system that elements or features can be assigned meanings. DIN SPEC 92001-3:2023-04
Positive Risk Balance (PRB)	The proposition that a computer driver should be no less safe (and ideally safer than) a human driver. Koopman & Widen, 2023Koopman & Widen, 2024
Robustness	The degree to which an AI system can maintain its level of functional correctness under any circumstances. ISO/IEC 25059:2023(en)
Safety-critical system	A safety-critical system describes a system that directly affects the safety, health and welfare of the public and whose failure could result in critical safety issues such as infringements of privacy, financial loss, environmental harm, serious injuries, or loss of life. (Laplante et al., 2020; Moteff & Parfomak, 2004; Srinivas Acharyulu & Seetharamaiah, 2015)
Trustworthiness	Ability to meet stakeholder expectations in a demonstrable, verifiable and measurable way ISO/IEC 20924:2024(en), 3.1.33

Executive summary

What we did

This report examines the main challenges that Artificial Intelligence (AI) poses in automated transport systems and the regulatory approaches to address them. These diverse challenges broadly relate to technical, regulatory, economic, societal and environmental issues, including issues relating to training data quality and representation, development and verification of AI models, increased vehicle travel and land-use impacts, deskilling vehicle operators and wider labour impacts. The report provides a common understanding of AI-based automated transport systems and the principles that can form the basis of institutional and regulatory actions to increase the safety and social acceptability of using AI-based transport systems. The report is based on discussions held at an ITF Roundtable in January 2023 and materials prepared for it. While recognising the unique specificities of each transport mode, this report mainly focuses on the automation of road vehicles. Nonetheless, some lessons from road automated vehicles (AVs) will be applicable to regulations on AVs in other domains.

What we found

The automated operation of vehicles – whether based on or supported by AI applications – holds great potential to meet future mobility needs in an efficient and safe manner. To realise the full potential of automated transport, two essential conditions must be met to ensure its safe and secure delivery: trustworthiness and dependability. Overcoming technical and regulatory challenges and minimising risks will help enhance social acceptance and uptake.

The Safe System approach provides a robust, safety-first framework for developing AV regulations. The Safe System approach assumes that mistakes and unexpected driving and operating behaviours are unavoidable and ensures that these do not contribute to serious injuries or deaths. The tenets of the Safe System approach apply equally to human-based and AI-enabled vehicle operation. Public dialogue on identifying acceptable levels of risk in line with the Safe System approach is fundamental. The use of simple comparative risk metrics between AVs and human-operated vehicles raises real practical challenges in the context of AV certification. Such metrics should be supplemented by a more fine-grained approach that compares like-for-like safety performance and addresses changes in the distribution of risks among the population.

AVs perform operating tasks – previously performed by humans – using an AI-based automated driving system known as ADS – an AI-based operating system for trains, vessels and aircraft. These AI systems have fundamentally different decision-making processes from humans. Therefore, the two separate regulatory systems that have been developed for human vehicle operators and human-operated vehicles are not adapted to AVs. Consequently, automated vehicles require new institutional and regulatory arrangements covering the entirety of the AI-automated operating system.

The whole AI lifecycle – including the context in which the AVs are operated, the data used for AVs, the AI models, outputs and their impact on society – should be taken into account when developing regulations for AVs to assure that the AVs are safe, secure and beneficial enough to become part of our transport systems. Those regulations must include both technical and non-technical measures.

The key factor that will underpin the sustained deployment and broader use of AVs is trust in AVs' ability to be safe, socially acceptable, and beneficial. In addition to technical robustness and safety, privacy protection, unbiased and ethical handling of data, fairness, explainability, and transparency are required at all stages of the AI lifecycle to make AVs socially beneficial and acceptable.

To address data-related issues such as potential bias and privacy infringement, regulations are needed to verify the lawful and ethical collection and use of data. Also important are the regulatory arrangements which allow public authorities or vetted third parties to verify that ethical requirements are satisfied in the acquisition and processing of data. Synthetic data can be beneficial to train AI in rare cases where real data input is scarce, but it could create new biases if not adequately managed. Principles and guidance on the use of synthetic data for the training of AI systems are, therefore, crucial.

The trustworthiness of data and its validation depend on its fair and accurate selection for the specific AV use case. To provide assurance of their safe and predictable performance and robustness, AI systems used in AV operations will need to be verified and validated. The functions performed by AI models used for vehicle operation include localisation, dynamic scene understanding, path planning, control, and managing user interaction. Each of these functions – and overall behaviour – needs to be evaluated using simulation, tests in controlled environments, and tests on real traffic situations. While scenario-based tests can provide assurance for common scenarios, it may prove more difficult for rare and edge cases because of the scarcity of available data. Also, evaluation based on predefined, known scenarios can lead to overfitting by manufacturers. Continuous scenario updates and diversification – including by using randomised scenarios – are essential.

AVs are not immune to programming errors or unexpected behaviour. Therefore, processes that address this uncertainty are necessary, along with policies to both mitigate AV's impacts and improve their safety performance ex-post. In line with existing approaches in the aviation sector, AV roll-out should include formal protocols to ensure lessons are learned and integrated by all actors following safety incidents. Such "antifragile" approaches will help to maximise AV system safety despite uncertainty about specific failure modes.

Public authorities' institutional capacity and regulatory measures should guarantee the transparency of AI development and deployment process. They should also guarantee a sufficient level of explainability of AI systems – even more so when self-learning AI tools are used. This requires public authority institutions and staff to continuously build knowledge and acquire skills. Achieving this in the face of the current concentration of skills in the private sector is a challenge.

The operational environments of AVs play an important role in their safe use. Existing operating environments should be improved to make it easier for AVs to function safely, with a lower chance of encountering rare but risky cases. Better information exchange between AVs and other road users can help avoid potentially dangerous situations. Machine-readable laws and enhancing the ability for vehicles and infrastructure to communicate will be beneficial in reducing such risks.

What we recommend

Base AI regulatory and institutional measures on shared fundamental principles

Fundamental human rights and the values derived from them – like safety, fairness, explainability, and human oversight – should form the cornerstone of AV regulation. To improve the safety of AVs and the entire transport system, authorities should adopt the Safe System approach. Public authorities must ensure all stakeholders understand what constitutes a safe and acceptable level of uncertainty. Moreover, authorities must design and implement appropriate regulatory interventions meant to deliver safe outcomes.

Ensure that AI remains explainable, and that training data is collected and handled in a transparent and verifiable way

AI systems should be designed in a way that explains how specific decisions are made based on specific inputs. This ensures that identified risks do not resurface. Data is a core element of AI systems. Data handling procedures and systems should ensure that AI systems' training data lends itself to identifying biases, quality issues, privacy issues, contamination from adversarial attacks or encoding/human errors

Mandate reporting of safety-relevant data from automated vehicles

Incident data is part of the essential "soft" infrastructure that ensures safety. Public authorities should mandate the reporting of incident data that is safety-relevant, including when automated vehicle operating systems disengage during test operating. Data regarding near misses may also prove relevant. Public authorities should establish monitoring, reporting and evaluation processes that improve overall safety performance after each incident. Metadata on system capabilities should accompany these reports. All these data should be accompanied by aggregate exposure data on distances covered and the environments in which the vehicles operated. Transport authorities must also build institutional capacity and technical proficiency to enforce regulatory measures.

Develop and update AV test scenarios and procedures

Scenario-based tests will play a central role in assessing AVs' abilities in a holistic and safer way. Public authorities should establish institutional and regulatory mechanisms to ensure that test scenarios are continuously updated and randomised to prevent manufacturers from designing AV performance that only meets a limited set of potential scenarios.

Ensure that physical and digital infrastructures support safe AVs

Enabling machine-perceivable signage, markings, and other important visual cues in AV operating environments enhances safe performance. To further increase the safety of AV operation and AV interoperability across multiple regions and contexts, establish machine-readable rules and regulations, and a common framework for vehicle-to-vehicle and vehicle-to-infrastructure communications. The benefits from these measures extend beyond the realm of AVs to all infrastructure users.

Enter the machine: Taking the (human) driver out of the vehicle

Technological developments encompassing both vehicles and software have enabled high levels of vehicular automation. Across all modes, including road, rail and shipping, highly- and fully-automated vehicles (hereafter AVs) that are capable of operating themselves without human intervention within a designated operating environment are expected to broadly impact societies (Bahamonde-Birke et al., 2018).

AVs are expected to have multiple first-order impacts (i.e. direct effects on transport). AVs are expected by many to increase transport safety and improve accessibility (European Commission, 2018; ITF, 2023b). AVs are also expected to reduce generalised transport costs due to the replacement of qualified drivers, conductors, pilots or captains by AI-enabled technologies. Second-order impacts (i.e. indirect effects on transport) include impacts on travel demand, public revenue, and labour, among others (ITF, 2023a, 2023b). However, all these expected impacts have yet to materialise as AV deployment currently remains quite low. Nonetheless, vigilance is warranted to ensure that AV deployment does not simply replace human error with technological failures or flaws.

For AVs to be widely deployed and used, specific technical and societal challenges must be addressed and overcome. A key challenge is to develop the right regulations to ensure the trustworthiness of AVs in the sense that they are both safe enough to be operated alongside human-operated vehicles and that their use should work for achieving valued societal goals such as improved accessibility, enhanced equity, reduced environmental impact, and economic development (ITF, 2023b)

Automated vehicle deployment levels are not even across modes

AV deployment has progressed unevenly across roads, railways, and waterways (Fiedler et al., 2019; ITF, 2023b). Depending on the mode considered, different levels of autonomy- corresponding to different capabilities- have been developed (e.g. SAE levels, MASS levels, Grades of Automation for railways) (IEC, 2014; IMO, 2021; SAE International, 2021a).

The extent to which a vehicle can be automated will depend on different factors, namely:

- the type of infrastructure (e.g. road, rail, waterways),
- the degree of control over the operation environment (i.e. open environment, closed environment),
- and the type of service considered (e.g. passengers or goods).

For example, technology for automated subway operation is widely available, and automated subway project deployment started in the 1960s (ITF, 2023b). However, unlike subways, which are typically separated from the surrounding environment by tunnels or barriers, automated road vehicles in cities interact with various elements, including pedestrians and generally have a complex and dynamic operational domain. For a long time, automated vehicles in cities will have to interact with a broad mix of vehicles of various ages running on technologies with differing maturity levels and different levels of sophistication. Their deployment is thus much more complicated. Deployment of highly automated vehicles in protected contexts (like ports, airports, and rail) poses fewer challenges in comparison to their

deployment in complex urban environments. Yet, the use of AVs for mass transport, such as railways and ferries, deserves specific consideration as AV-related safety incidents occurring in non-road environments may have greater impacts.

Roundtable participants pointed out several differences between modes. Railways have a simpler operation domain, but the potential severity of one major failure (a crash or a derailment) could be far greater than that of a road vehicle; thus, the tolerance level for flawed AI performance might be lower than for road vehicles. The automation of waterborne vessels faces a different set of challenges. Unlike complex road environments, waterborne vessels operate in environments with relatively sparse visual clues outside of coastal areas and in a constantly changing medium where waves and wind complicate locational precision and motion control.

Automated road vehicles face the most complicated operational environments and would likely have the most substantial societal impacts. While recognising the unique specificities of each transport mode, this report focusses mainly on the automation of road vehicles. Nonetheless, some lessons from road AVs will be equally applicable to regulations for automated vehicles in other domains.

Box 1. Terminology used in this report

The roundtable focused on the use of AI for automated vehicles (including rail vehicles and watercraft) that can perform dynamic driving tasks (DDT) in designated operational design domains (ODD) without human intervention. The level of automation of the vehicles discussed corresponds to levels 4 and 5 of SAE levels of road vehicle driving automation (SAE International, 2021a). It also corresponds to the definition of "fully automated vehicles" in the (Regulation (EU) 2019/2144).

A variety of terms have been adopted by different entities, such as 'autonomous vehicles' and 'self-driving vehicles,' and some of them have legal force in specific countries. As outlined in previous ITF work on AVs (ITF, 2023b; ITF, 2018), "these terms embody differing views about the role of vehicle connectivity and the potential for driving without external assistance". The SAE International deprecated the use of the term 'autonomous' on the base that the term "obscures the question of whether a so-called "autonomous vehicle" depends on communication and/or cooperation with outside entities for important functionality" (SAE International, 2021a). In contrast, Regulation (EU) 2019/2144 uses the expression "move autonomously" to define the term 'automated vehicles', in which 'autonomous' means 'by itself' rather than 'without communication'.

In line with these views as expressed in previous ITF reports and to cover the variety of vehicles that may be automated beyond those driven on roads, the term 'automated vehicle' or 'AV' is used to describe the vehicles discussed in this report. The noteworthy differences with conventional use are: first, if not stated otherwise, the term is used for level 4 and 5 automation, thus excluding the level 3 cases where human-driver engagement is necessary for certain situations; second, in some context, the term is expanded to include rail vehicles and surface ships with the degree of automation that corresponds to level 4 and 5 of motor vehicle automation.

Safety first, all other utilitarian considerations later

AV deployment faces important technical and regulatory challenges. Among these challenges, ensuring safety is paramount. Automating assembly lines in manufacturing plants and deploying automated buses in city centres do not have the same public safety implications. Even though a malfunction of an industrial robot could possibly have lethal consequences, the range of potential risks is limited to a confined environment and to the people working in it. Malfunctions of AVs operating in public spaces, on the other hand, are not only dangerous for people in them but could be dangerous for people outside of the vehicles too. For instance, in 2018, an AV in test driving trials collided and killed a pedestrian pushing a bicycle across a road (National Transportation Safety Board, 2019). Indeed, the complexity of transport system automation stems from the fact that transport has direct and significant implications regarding safety and human welfare.

Safety is the foremost objective for AV deployment, preceding all others. If safety is not guaranteed, it is difficult to justify approving the use of AVs, no matter how great their other potential benefits. As the German Federal Ministry of Transport and Digital Infrastructure (BMDV, 2017) stated for road vehicle automation, the primary purpose of AVs should be "*to improve safety of all road users"* and *"the protection of individuals takes precedence over all other utilitarian considerations*". This is true for all AVs: they should ensure the safety of all people within the same operating environment-- both inside and outside of the vehicle.

Automation of vehicle operation is one among a wide range of other safety-improving actions and interventions. Some of these are well-known (e.g. speed management, separation of traffic participants based on speed and weight, addressing maximum vehicle speed or mass and other vehicle design elements, driver education, etc…) and may or may not be fully or consistently applied. In many cases their impacts and their costs are known and their safety improvement may be more rapid or less costly than those arising from large-scale AV deployment. If AVs are the answer to the question "how might safety be improved", it is incumbent upon policymakers to ask whether the deployment of AVs is the first-best, 10th best or only 50th best answer to that question to guide their pro-safety policies.

AV certification implies identifying a threshold value for their safety performance, but the answer to the question of how safe is 'safe enough' is neither straightforward nor settled (ITF, 2018). The first step in answering that question is to articulate an overall transport safety strategy. The 'Safe System' approach provides such a framework and has been adopted by authorities around the world (ITF, 2008, 2016, 2022b). This approach is grounded in the realisation and acceptance that, despite all the best efforts to avoid them, traffic participants will make mistakes or display unplanned or unforeseen behaviours. A safe system is one that is designed so that no one dies or is seriously injured when these mistakes or behaviours occur. A core tenet of the Safe System is the reduction of the difference in kinetic energy between traffic participants so that collisions, should they occur, have minimal consequences. Safe Systems are designed to minimise the mass and speed differentials among traffic participants via speed management or traffic separation. As noted in ITF (2018, 2023b) the deployment of AVs implies the following adaptation of the Safe System approach, which traditionally focussed on human drivers and traffic participants (ITF, 2023b):

> Automated driving systems (ADS) are not perfect and could perform unexpected and unusual manoeuvres that can lead to crashes. The transport system must accommodate ADS's imperfection and ensure minimally acceptable levels of safety – e.g. no death or serious injury – even in edge and corner cases.

An additional factor to consider is that one source of AV's unintended or unexpected behaviours is upstream coding errors by humans in the ML or AI model. These types of second-order errors are likely to

be found in any software-hardware system but given the reliance of AVs on human-written algorithmic code, they are especially relevant to AV safety (ITF, 2019).

A common threshold used to guide the certification of AVs is achieving equivalent, or better, safety performance than a human driver – that is, achieving a net Positive Risk Balance (PRB) for AV driving compared to human driving. Often defined at a national scale (e.g. achieving a net reduction in fatalities at a national scale), net risk metrics like PRB are intrinsically easy to understand and communicate but are complicated and challenging to meaningfully define and use in practice – as in the case of certifying AV system performance (Koopman & Widen, 2024).

A large part of this difficulty resides in the challenge of establishing a relevant and comparable baseline for human driving safety performance (Koopman & Widen, 2024). For example, is the performance of the AV driving system being compared to that of human-driven vehicles of the same age and equipped with the same safety features? Is human driving being compared to AV driving in the same geographic, street network, meteorological and other regional contexts? If AV systems display a PRB with respect to human drivers, does it do so for all types of human drivers- or are there subgroups of human drivers that display a PRB in comparison to AVs – e.g. drivers who drive more safely than AVs? All of these are important considerations to consider when exploring the use of net risk metrics like PRB in AV system certification.

Koopman & Widen (2024) call for practical AV system certification and policy to go beyond the simple use of net PRB. They describe risk assessment criteria that are helpful in shaping AV system certification processes. These criteria raise three central questions that must be addressed in AV system safety performance assessment:

1. How much safer must AV driving be, and over what time frame?

Simple approaches to PRB maintain that if the overall driving performance of AVs is as safe as overall human driving, then AVs should be allowed to operate. However, other approaches highlight limited public acceptance of AV driving performance that is equal to, or only marginally better, than human driving. It is unclear what forms a publicly tolerable PRB threshold (equal performance?, 10% better?, 100% better?), complicating setting an acceptable PRB criterion for AV system certification. Furthermore, deployment of AVs may result in a PRB in favour of human driving over AV driving in the short- to medium-term (e.g. more fatal and serious injury crashes occur involving AV systems than human drivers) before shifting in favour of AVs in the medium to long-term. This further complicates the establishment of a useful PRB baseline (Koopman & Widen, 2024).

2. How is baseline safety performance established and monitored?

Establishing a relevant and useable human driving baseline is exceedingly complicated – especially in the early to mid-deployment phases of AVs – due to inherent asymmetries in AV versus human driving performance. Human driving occurs in a very broad range of environments and contexts, using a very disparate fleet of vehicles characterised by a wide range of ages, designs, embarked technologies and states of good repair. Human driving involves drivers representing a very wide range of demographics and characteristics (age, experience, impairment, threshold for risk acceptance, etc.) as well as an equally broad range of crash opponent types and demographics. In comparison, early phases of AV deployment are characterised by much more uniform vehicle fleets (age, design, safety technologies, state of repair), in much more restricted geographic contexts and less diverse road and street networks and may display different crash opponent profiles (Koopman and Widen, 2024).

An additional complicating factor is that potential misattribution of causality for AV-involved crashes may further obfuscate a meaningful safety performance baseline (e.g. legal regimes attributing responsibility to

human drivers for rear-end crashes into AVs that have unexpectedly braked may hide the latter as an emerging contributory factor to crashes and skew the PRB in favour of AVs) (Koopman & Widen, 2024).

Finally, comparability of human to AV safety performance is extremely challenging due to the disparity in travel volumes of each mode of operation, with the former representing orders of magnitude more kilometres driven than the latter. This disparity raises questions and uncertainties regarding the statistical significance and comparability of crash rates for AVs with respect to human-driven vehicles. Overall accumulated AV driving ranges from one to three million miles driven per AV robotaxi company (Bidarian, 2023) whereas human driving fatality rates involve orders of magnitude greater travel volumes (e.g. 1 fatality per 74 million miles driven in the US (NHTSA, 2023a), 1 fatality per 192 million miles driven in the UK (UK DfT, 2022). There simply isn't enough accumulated AV driving to establish straightforward and comparable baseline safety performance metrics and closing this gap through different modelling and predictive approaches introduces uncertainty as to the "real" fatal crash rate of AVs. For instance, unexpected driving performance following a software or hardware update or a change in the driving environment may lead to an uptick in clustered 'common cause' AV crashes in a pattern unlikely to be predicted under the assumption of random independent failures (Koopman & Widen, 2024).

For these reasons, establishing a relevant measure of PRB in early- to mid-term AV deployment incorporating like-for-like comparison of AV to human driving safety performance is extremely challenging – certainly with sufficient confidence to meaningfully certify AV system safety. At a minimum, uncertainty over the relevant baseline for human safe driving performance would suggest adopting a higher, rather than a lower, AV PRB threshold for regulatory, policy and certification purposes.

3. Do AVs increase risk for some or transfer risk from one group to another?

Even in instances where the uptake of AVs reduces net fatalities, serious injuries or other harms, some populations may see an increase in specific risks or experience a transfer of risk that is socially unacceptable. For instance, a scenario where the deployment of AVs leads to a net reduction in deaths and serious injuries but where the overwhelming majority of those still killed or seriously injured are children or emergency responders would likely face significant resistance. Similarly unacceptable would be a scenario where the uptake of AVs leads to the steep reduction of deaths and serious injuries from eliminating risky driving behaviours of young men or alcohol- or drug-impaired driving of but leads to an increase in AV crash risk for vehicle occupants who, as previously safe drivers, faced very low crash risks. Finally, accepting a higher level of risk from AV performance in the short term versus a net reduction of risk in the long term is a form of temporal risk transfer from present generations to future generations and may also be challenging to manage from the perspective of societal acceptance.

The danger in using simple net risk indicators, like Positive Risk Balance, is that they only address overall risk and do not account for the *distribution* of those risks or for the *transfer* of those risks amongst the wider population. In its report *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), the European Commission stresses that even if net risk is reduced by AV deployment in comparison to the case without AVs, "no category of road user (e.g. pedestrians, cyclists, motorbike users, vehicle passengers) should end up being more at risk of harm from [AVs] than they would be against this same benchmark". It further stresses that AV deployments should be designed expressly to avoid creating new inequalities in risk distribution and redress existing ones (DG RESEARCH, 2020).

Whatever the approach adopted for assessing and certifying AV safety performance, it should go beyond the use of simple risk balance metrics and incorporate a range of relevant criteria, not just for the vehicle itself, but extend to the entire AV system across the entire AV lifecycle.

Assessing vehicle safety versus ensuring trust in safe AV system performance

Certifying the safety performance of AI-based AVs is not a straightforward task (ITF, 2018). Current regulatory frameworks address vehicle safety characteristics and driver's skill and aptitude separately. These approaches will be challenged by the deployment of AV systems.

A vehicle operator's capability is currently certified by a multi-tiered licencing process. The licensing process does not guarantee the operator's safe driving behaviour in a pre-emptive manner but establishes a common basis for training and operator testing. It can be divided into three tiers (Cummings, 2019): physical tests to check that the operator is physically fit to operate the vehicle, knowledge tests to ensure a person's understanding of operating rules, and practical tests to assess a person's vehicle operation capabilities. Operators' licenses are issued based on the trust that the licensee is a "safe enough" operator, considering the performance the operator has shown during the test process (Cummings, 2019). Because operators – drivers in particular-- are also responsible for the consequences of their actions, operators' responsibilities and their liabilities are aligned. In addition, dangerous behaviours are regulated and addressed via the enforcement of applicable laws.

The safe operational capability of AVs must be assessed even more holistically than for non-automated vehicles (ITF, 2018). Even though the latter comprise hardware and software systems (e.g. mechanical disc brakes and software governing anti-lock braking – ABS-- and electronic stability control- ESC), AI-based driving systems are characterised much more inextricably interconnected software and hardware. Existing vehicle certification practices are based on national vehicle safety standards for road vehicles, and compliance with the standards is either self-certified or established via type-approval processes, depending on the countries' legal systems.

Current safety assessment systems for transport are designed to accommodate the presence of an operator in the vehicle. Certain commercially available vehicles are already equipped with some low-level automated functions, such as various advanced driving assistance system (ADAS) features. However, as the name suggests, ADAS features are supposed to *assist* drivers in accomplishing the driving task. ADAS still requires drivers to engage in driving tasks throughout the trip, and thus, the driver must remain fully responsible for the vehicle's operation. Further development of AI-based automated driving systems (ADS) is expected to widen deployments of AVs that do not require human intervention in performing driving tasks. This would ultimately result in the removal of humans from the operator's seat and require fundamental changes in the way safety is assessed.

Pre-emptively excluding all safety risks is impossible due to the complexity of the road transport environment, the degrees of freedom of drivers and others within that environment. In particular, the non-deterministic nature of ML AI systems means that they may display different behaviours and outcomes even if they are presented with the same sensor inputs (Cooper, et al., 2022) (see Box 2 for a more detailed description on the development of AI). The combination of these factors makes it even more difficult to ensure that AI-based systems will safely or even predictably respond to all the situations they face. Unlike the case of human-driven vehicles, occupants of AVs won't be responsible for the vehicle's driving performance. All these factors highlight the need for adapted and sometimes entirely new regulatory approaches to certify the safety of AVs.

Box 2. Artificial Intelligence in brief

Artificial intelligence (AI) is an umbrella term for algorithmic and computer science techniques that allow computers to imitate complex human skills (ITF, 2019; Sheikh et al., 2023). The term was coined as early as the 1950s.

The OECD (2019; 2022) defines AI systems as below:

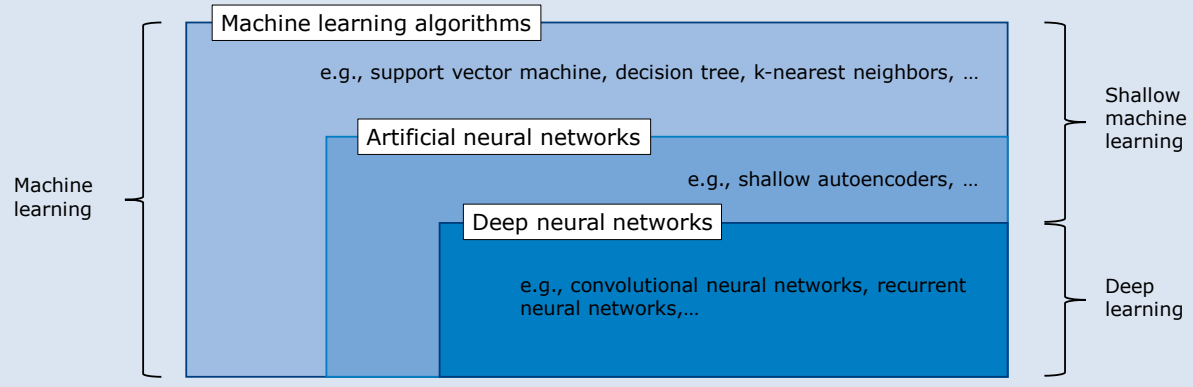
> *An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.*

The technological development path of AI has not always been smooth. The technology has been developed over several "winters" and breakthroughs. Two distinct approaches have emerged from this development path: symbolic reasoning and machine learning (ML). Symbolic reasoning is a deterministic approach that requires explicit human input of rules, while machine learning algorithms learn rules from data (ITF, 2019). While both approaches have their advantages and disadvantages, ML algorithms have opened new possibilities for automated driving, which requires real-time decision-making in complex environments that more deterministic approaches could not fully cover.

ML algorithms are distinctly different from conventional algorithms as they learn patterns and correlations from data sets and discover rules by themselves according to specified optimisation functions (e.g. "maximise time savings", "minimise algorithmic compute", "minimize collision risk"). Learning in ML systems occurs within the algorithmic system, continuing after the training phase into the use phase (ITF, 2019). ML algorithms can be divided into three categories: supervised learning, unsupervised learning and reinforced learning, depending on their training process, whether they require labelled data or whether a reward system is applied. (ITF, 2019, 2021; Sheikh et al., 2023).

Deep Learning (DL) is a relatively recent subset of ML that uses artificial neural networks. DL has shown to perform well on complex tasks such as computer vision and voice recognition. Compared to 'shallow Machine Learning' techniques, DL can be applied to a wider range of tasks that were previously regarded as too complex for machines to perform. There are downsides to DL as it relies on a larger dataset and uses more resources, computational power and energy. The operations of DL algorithms are less interpretable and explainable than explicit programming or shallow ML techniques (ITF, 2019; Janiesch et al., 2021).

Figure 1. Machine Learning Concepts and Classes



Source: (Janiesch et al., 2021 adapted from Goodfellow et al., 2016)

Generative AI based on large language models (LLMs) or similar content generator AI models are a form of AI that generate text, images or other media content based on its training data. Whereas ML AI systems use training data and learned behaviours to predict outcomes or actions, generative AI uses training data and feedback on previous output to generate human-like text, images and sounds that resemble the content on which it has been trained (Sengar, et al., 2024). There are several potential uses for generative AI models in automated vehicle design and testing including speeding up and improving ADS ML algorithmic coding and generating realistic testing scenarios, even for edge cases (Thomas, 2024).

What and when to certify?

A deeply challenging consideration for AVs is what it is exactly that would constitute the *object* of certification (ITF, 2018). Unlike existing cyber-physical vehicular systems, the AI code base in AV systems is neither stable nor fixed. AI-based code may reconfigure itself as it applies optimisation functions to the data it gathers from its own operation and from its environment. This means that even if it were possible to establish a stable code base for an AV at the time of its certification, that code base may substantially re-write itself later, invalidating the original certification, which was for a now different vehicle-software system.

Further, the certification of different operating components of the AV cyber-physical system would similarly become invalid as the operation of these systems (e.g. automated braking) in conjunction with an evolving AI-based code base (e.g. hazard detection) would fall outside of the scope of the original certification envelope. Likewise, non-AI software updates or the installation of new hardware components (e.g. a new LIDAR sensor or video detection module) would further interact with a very different AI code base than the originally certified vehicle, leading potentially to unanticipated and possibly unsafe operation(ITF, 2018). The dynamic nature of AI-based code in AV systems highlights the need for a different and adapted approach to certifying the safe operation of a *broader AV system*, and not only the safe operation of a single class or model of AV, one at a time. This broader approach must account for the dynamic nature of the AV as a regulatory object.

A broad AV system safety framework must address and account for the role of supportive infrastructure (ITF, 2023d) just as it must also account for the actions and responsibilities of key stakeholders (e.g. entities responsible for ensuring the safety of the AVs, software developers and AV manufacturers). The safety framework must also account for ways in which environmental aspects (e.g. fog, rain, dust, snow, ice, etc…) may degrade the safe driving performance of AVs. This broader approach should account for the ways in which AVs will influence other agents' behaviours and how these other agents will influence AVs' behaviours in turn. Finally, at its core, the AV system is essentially a cybernetic system, so its safe operation relies also on robust and adaptive cybersecurity mechanisms, systems, and protocols. The nature of these enters fully within the scope of the regulation of the full AV driving system.

If the AV system is not fixed in its configuration or performance, how then should regulators certify its safe operation? In other words, how can the regulatory framework ensure trustworthy AVs? AVs share some safety aspects with conventional vehicles. However, as noted above, new aspects must be considered (Galassi & Lagrange, 2020). The assessment of the *combined performance* of AI-related components is currently lacking in the existing vehicle assessment framework. Ensuring trustworthy AVs will require the extension of existing certification and validation processes to AI-related elements (e.g. AI systems, machine learning, algorithms, etc.) (Baldini, 2020; Fernandez Llorca & Gomez Gutierrez, 2021). Ensuring the robustness of AI should also consider specific challenges associated with AI (e.g. cybersecurity, data privacy, unbiased treatment of data, etc.).

Including AI-specific challenges in AV safety assessment

The absence of a driver means vehicles will have to perform tasks that were performed by humans until then. AVs rely heavily on cyber-physical systems, which combine hardware (e.g. sensors, LIDARs, RADARs, microphones, etc.) and software (e.g. artificial intelligence (AI) using machine learning (ML) algorithms to replicate human capabilities, namely perception, planning and control. The classic approach to vehicle safety assessment is not adapted to the complexity of mixing these components and processes (Fernandez Llorca & Gomez Gutierrez, 2021; Galassi & Lagrange, 2020; ITF, 2018; Pater, 2018).

The use of AI software components in vehicles raises new challenges for driving system assessment (ITF, 2018; Taeihagh & Lim, 2019). Removing the human in charge and their potential shortcomings and errors does not mean the system will be unfailing. On the contrary: as the complexity of AI-related components increases, so does the probability of machine errors or unanticipated behaviours (Taeihagh & Lim, 2019). Furthermore, the criticality of these risks may depend on the type of vehicle (e.g. car, train, waterborne transport) and the Operational Design Domain (ODD) considered (e.g. open or closed environment, the complexity of interactions, etc.).

AV assessment must be adapted to cover these AI-related challenges. AI in vehicles does not form a monolithic system (Dede et al., 2021). It is a combination of software components in charge of subtasks of driving activity, namely perceiving, planning and controlling the vehicle. AI software embedded in vehicles interprets data collected by different types of sensors using different methodologies (Fernandez Llorca & Gomez Gutierrez, 2021; ITF, 2018). Thus, new types of risks may arise. For example, risks stemming from data collection are very different from those encountered by conventional vehicles.

Existing certification approaches could be adapted to address some of the new dimensions of AV assessment (see Box 3). However, new approaches to assess the potential adverse consequences of AI-related systems are also needed. Due to the socio-technical nature of AI, challenges are not only technical or legal, but they are also social and societal. For instance, one of the potential issues raised in the Roundtable was the potential bias in data acquisition. AVs need to have local training data from their ODDs to improve their performance. If AV-developing companies test AVs extensively in a certain neighbourhood that is cheaper to operate in but not necessarily representative of where the vehicles will be deployed, the training data prove biased thus negatively impacting AV driving performance.

Box 3. Automation of driving capabilities for road vehicles and the paradigm shift for vehicle assessment

Assessment and certification approaches for road vehicles typically evolve to address new types of safety issues introduced by new vehicle technologies.

The classic certification approach consists of physical tests to assess that the vehicle reaches a necessary safety level before entering a market. These tests are usually performed on a test bench or automotive test tracks. This approach aims to ensure the performance of the vehicle's mechanical systems and components such as tyres (e.g. resistance, grip on wet surfaces), brakes (e.g. high-speed effectiveness, heat, parking brake resistance) and steering equipment, among others.

With the introduction of computerised components and driving assistance technologies to improve vehicle performance (e.g. Electronic Stability Program - ESP, Anti-lock Braking System - ABS, etc.), the classic approach of the test bench and the test track alone no longer seems adapted to address safety-relevant areas related to the electronic system. Additionally, electronic systems introduced new risks related to their potential failures. Thus, simulation tools have been introduced to support the certification of computerised components such as ESP (OICA, 2019).

Figure 2. Multi-phase testing, verification and validation of AVs

Source: (Baldini, 2023; OICA, 2019)

As the number of software components in vehicles continues to increase with partial and full automation of driving capabilities, the classic approach must be further extended to address new issues introduced by the use of AI in vehicle operation. More than test tracks are required to address the diversity of potential scenarios: real-world testing is expected to enable automated road vehicles to accumulate billions of kilometres to prove their reliability in safety-relevant situations. Nevertheless, even that much driving does not guarantee that nearly all safety-relevant events occurred. A specific approach for AI-related components and software is needed. Simulations could be used to cope with critical- and potentially rare - safety-relevant situations. Simulations are, however, just that – simulations that approximate, with varying degrees of skill and precision, potential real-life performance. As such, they are vulnerable to biases inherent in the model or the data on which it is built.

Elements of the classic approach (e.g. track testing, bench) are still necessary as they already allow for certification of systems and components that are present in vehicles (e.g. brakes, tyres, etc.). The sensors providing data to the AI driving system must also be evaluated using the classic certification approach.

Therefore, step-by-step, multi-phase evaluation using simulation, tests in controlled environments, and tests on real traffic situations are required to ensure each function and overall capability of AV.

However, it is important to note that this multi-phase evaluation method is not the whole picture. This is limited to a single AV with specific hardware and software. To ensure the safety of AVs in practice, additional policy measures that account for trustworthiness at their deployment and operation stages are necessary.

To ensure that AVs are trustworthy, AV system stakeholders (e.g. users, operators, and governments) will likely value a multiplicity of criteria that are not just associated with the vehicle's physical safety but also its cybersecurity robustness (i.e. the ability of a system to perform its intended functions even under adversarial cyber-attack) and the explainability of the AI software, among others.

Risk-based ADS regulation comprises three components: hazard mapping, risk assessment and risk-based oversight and regulation. Hazard mapping should extend across all AI application fields since AI hazards in other domains may also be relevant for transport and ADS design and operation. Focus should be given, however, to AI hazards resulting from automated vehicle operation. The OECD AI Incidents Monitor is one approach that tracks AI hazards across different application domains around the world (OECD, 2024). Risk is a function of potential hazards or harms and their probability of occurrence and severity. The final component of a risk-based approach comprises establishing progressively more stringent tiers of oversight and regulation based on level of risk. Many potential uses of AI in transport pose no or minimal risks, whereas other uses may pose significant to intolerable risks. Assessing the probability of significant AI-induced harms is complicated by lack of sufficient experience and because of the non-deterministic nature of ADS AI algorithms. To address this uncertainty, ADS regulation should incorporate a precautionary element where potential risks are severe or intolerable. Such a precautionary approach is built into different AI risk management approaches (e.g ISO/IEC 23894 and NIST AI 600-1) or regulation (e.g. the EU AI Act, 2023).

A new regulatory framework should include extended assessment criteria covering risks that will be amplified by AI use (e.g. cybersecurity, safety, human-machine interaction) and a new set of regulations to oversee new risks that are not confined to the vehicles but linked to the management system (e.g. data protection, privacy, liability) (Bellet et al., 2019; Matheu-García et al., 2019). The assessment of the impact of AI software components is not unique to the transport sector. AI assessments and testing in other sectors can inform public authorities since AI in AVs faces challenges like those faced by other sectors (Baldini, 2020). Additionally, the complexity of AVs, in terms of their diversity (i.e. road vehicles, vessels, rail, etc.) and how they are implemented (i.e. open or closed environments), calls for a multidisciplinary approach that goes beyond technical considerations. Such an approach could include socio-economic and social challenges (Dubljevic et al., 2021). Different types of potential AI-related issues can be distinguished (Dede et al., 2021; Dubljevic et al., 2021; Fernandez Llorca & Gomez Gutierrez, 2021; Matheu-García et al., 2019):

- Technical challenges associated with the nature and capabilities of the technologies used;
- Societal challenges which describe AI's impacts on socio-economic structures;
- Societal challenges include the risks of the interaction between automated systems and humans;
- Data privacy and governance challenges related to how the collected and processed data is used.

As initiated by several countries such as France, (JO, 2021), Germany (Deutscher Bundestag, 2021), Korea (Act No. 16421, 2019) and UK (Automated Vehicles Act, 2024), public authorities would need to adopt an adapted and new regulatory framework to ensure the safety and, more broadly speaking, trustworthiness of AVs considering these complex issues that are outside of the conventional scope of vehicle safety standards and driver licensing systems. New institutional arrangements should also be established to

execute the new regulatory measures. These cannot be done solely by the government's initiative or industries' self-regulation. There should be a common understanding of the key requirements for the trustworthiness of AVs and how to satisfy them across all stakeholders, including various levels of government, AV developers and operators, infrastructure operators, and other affected sectors of civil society.

Policy Takeaways

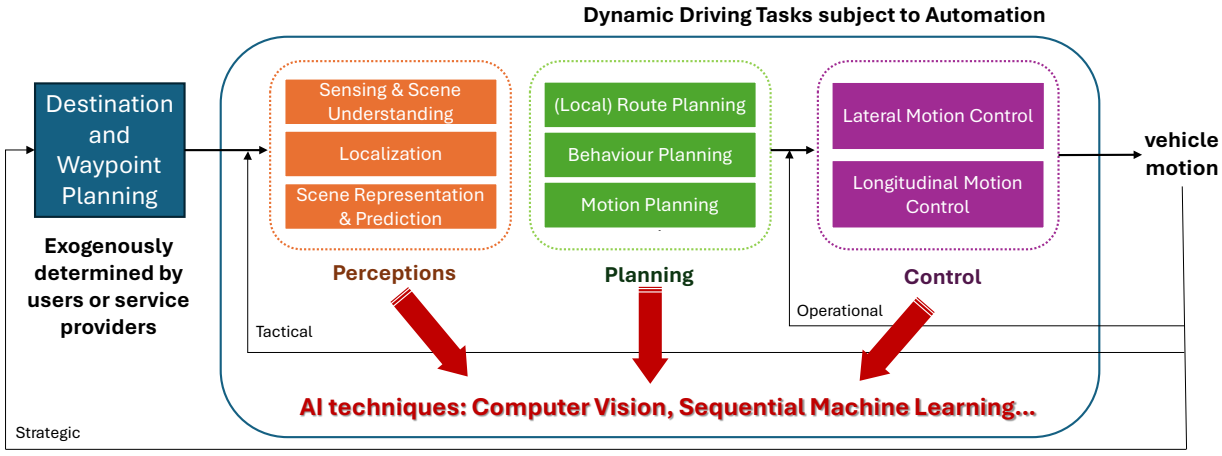
- AVs utilising AI require new institutional and regulatory approaches to ensure their safety and trustworthiness, which are different from conventional systems developed for deterministic, electro-mechanical safety features.
- The Safe System approach assumes that mistakes and unexpected driving behaviours will happen and ensures that neither contributes to serious injuries or deaths. As such it provides a robust, safety-first framework for developing AV regulations.
- The use of simple risk balance metrics raises real practical challenges in the context of AV certification and should be supplemented by a more fine-grained approach seeking to compare like-for-like safety performance and addressing changes in the distribution of risks among the population.
- To develop new regulatory frameworks for AVs, policy makers need to make efforts to ensure a common understanding of major AI principles and how to apply them to ensure the trustworthiness of AVs and AV operating entities.

How AI confounds human oversight and how to ensure its trustworthiness

Anatomy of driving: what constitutes driving tasks, and how does AI perform them?

The act of driving involves a continuous flow of perceiving the environment, making decisions, and executing motion control (ITF, 2018). According to SAE International, the driving task can be divided into strategic functions and dynamic driving tasks (DDT) that should be conducted in real-time (SAE International, 2021a). Strategic functions are about deciding the destination and the waypoints to the destination and scheduling the trip. DDTs are the tasks usually regarded as driving and are what automated vehicles need to perform. The DDT, in turn, can be divided into tactical functions and operational functions (See Figure 3). Tactical functions are about identifying objects, events and other factors that are relevant to driving the vehicle and planning its movement. This requires detailed functionalities such as perceiving the environment, including the vehicle's location (localisation), interpreting the environment, and planning the vehicle's route. Operational functions are controlling the vehicle's lateral and longitudinal motions to execute planned decisions while maintaining stability (Fernández Llorca et al., 2021; ITF, 2018; SAE International, 2021a).

Figure 3. Schematic representation of dynamic driving tasks



Source: ITF elaboration based on (Fernández Llorca et al., 2021; SAE International, 2021a)

Human drivers conduct all these functions in a holistic way. Humans can locate their vehicles in space, make tactical decisions, and control their vehicle movements effortlessly. Humans can make holistic interpretations of the scene in front of their eyes based on diverse clues, make decisions on which lane to take and the speed at which the vehicle will run, and control motions according to those decisions. Most drivers perform those motion controls based on muscle memory without being constantly aware of their actions. However, these control tasks require real-time adjustment of control inputs (steering, braking, accelerating) based on the vehicle's responses to the inputs and fast-changing environment, which require significant computational efforts if performed by computers based on conventional programming.

The development of AI, especially AI techniques based on ML and DL algorithms, opened new possibilities for automated vehicles. Similar to humans' ability to learn by doing, AI-based AV systems learn how to drive themselves through upstream training and downstream learning by doing. As shown in Figure 3, the AVs perform tactical and operational tasks with their Automated Driving Systems (ADS) that employ multiple AI techniques. The overall AV driving process is roughly similar to human driving. As with human drivers, AVs perceive their environment. AVs first 'sense' the surrounding environment with diverse sensors. Then, AVs process raw inputs received by sensors to understand the scene and locate themselves within the scene. AVs detect and identify the boundaries of roads, lanes, and objects. AVs may also use global navigation satellite system (GNSS) signals, high-definition maps (HD maps) and other inputs in this process. AVs also predict the movement of vehicles, humans and other objects. Based on this information, AVs plan their route in the constructed scene, projecting their anticipated behaviours (e.g., changing a lane) and direct motion decisions (e.g., rate of acceleration and speed of travel) (ITF, 2018).

AVs and human drivers conduct necessary driving functions with varying levels of skill. ITF (2018) highlighted that depending on the specific functionality considered, either human drivers or AV systems may function better than the other (Table 1). A key finding is that the weakness of AVs in reasoning and perception must be addressed by new assessment and validation approaches. An additional factor to consider is the accounting for condition-based safety performance: "Where there is conditionality – e.g. better driving performance for either humans or automated driving systems is linked to a specific set of conditions or contexts-- the Safe System approach implies that the resulting ambiguity does not lead to crashes, loss of life or serious injuries. This may entail upstream system versus vehicle design that seeks to eliminate these risks"(ITF, 2018).

Table 1. Summary of Fitts List of strengths and weaknesses across various aspects of function allocation between humans and hardware/software systems

Aspect	Human	Hardware/Software system
Speed	Relatively slow	Fast
Power output	Relatively weak, variable control	High power, smooth and accurate control
Consistency	Variable, fatigue plays a role, especially for highly repetitive and routine tasks	Highly consistent and repeatable, especially for tasks requiring constant vigilance
Information processing	Generally single channel	Multichannel, simultaneous operations
Memory	Best for recalling/understanding principles and strategies, with flexibility and creativity when needed, high long-term memory capacity	Best for precise, formal information recall, and for information requiring restricted access, high short-term memory capacity, ability to erase information after use
Reasoning	Inductive and handles ambiguity well, relatively easy to teach, slow but accurate results, with good error correction ability	Deductive and does not handle ambiguity well, potentially difficult or slow to program, fast and accurate results, with poor error correction ability
Sensing	Large, dynamic ranges for each sense, multifunction, able to apply judgment, especially to complex or ambiguous patterns	Superior at measuring or quantifying signals, poor pattern recognition (especially for complex and/or ambiguous patterns), able to detect stimuli beyond human sensing abilities (e.g., infrared)
Perception	Better at handling high variability or alternative interpretations, vulnerable to effects of signal noise or clutter	Worse at handling high variability or alternative interpretations, vulnerable to effects of signal noise or clutter

Source: ITF, 2018 based on Schoettle (2017) (adapted from Cummings (2014) and de Winter & Dodou (2014))

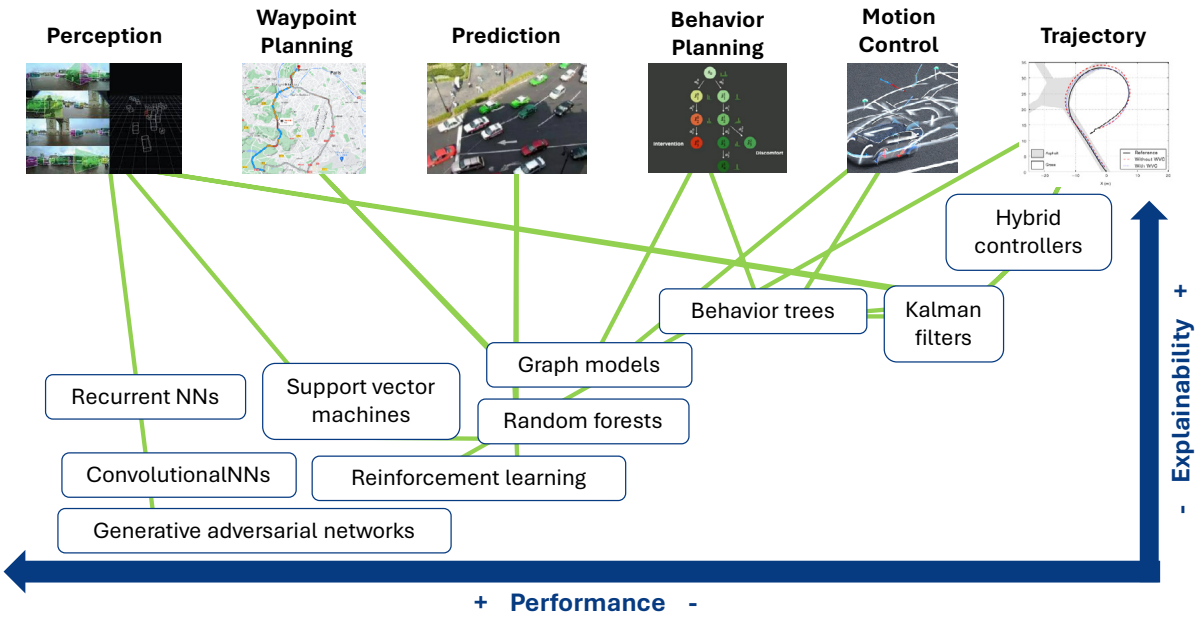
AVs may draw on a range of AI methods and techniques to carry out driving tasks (see Figure 4). While these have improved the performance of automated driving significantly, there are still some critical issues that must be addressed, especially considering the deployment of fully automated vehicles.

The first among these is addressing the tension between understanding and explaining how an AI algorithm functions and the performance of that algorithm. Trained observers can understand how a typical algorithm functions from simply examining its code (interpretability). Based on that understanding, that observer can explain the functioning of the algorithm and its outputs (explainability) – including highlighting which data inputs or processing functions led to specific outputs. This interpretability and explainability comes at a cost when it comes to more and more sophisticated AI algorithms, especially ML and DL algorithms – the better these types of algorithms function, the less interpretable and explainable they are (ITF, 2019). One of the confounding aspects of AI transparency and explainability is that even transparent AI algorithms – e.g. those whose code is revealed to an observer with the technical knowledge to understand it – may not necessarily convey to the observer (or even to the algorithm's designer) sufficient information on its functioning to allow for explanation of the algorithmic decisions and outcomes (ITF, 2019). Figure 4 shows the inverse relationship between explainability and AI performance. Symbolic AI may be explainable, but it does not perform as well as ML or DL algorithms in dealing with big and complex data (Symbolic AI may still be useful for certain purposes, such as hard coding the rules that must

be observed). ML and DL systems can be used for many functions that require more complex decision-making, but it is more difficult to understand how and why they produce their outputs.

Another challenging aspect of regulating AI-based AV systems relates to the data on which the AI is trained and that it collects in use. AI systems are trained on data, but it is next to impossible to have an exhaustive dataset that includes all possible situations the AV system will face during its operation. Therefore, all training data is necessarily a subset of the "real" world that is actively or passively curated and is thus open to explicit or implicit biases. Furthermore, there is the risk of cyberattacks that actively alter data inputs to interfere with AI's performance.

Figure 4. Examples of AI techniques used for automated driving



Source: Ameyugo (2023)

The complexity and inscrutability of the types of AI algorithms most helpful for automated driving can lead to situations where an AV makes an unexpected and harmful decision that any normal human driver would not make. In these instances, the reason why the AI made that decision may not be explainable or discoverable, the factors and inputs that most influenced that decision may not be apparent, and the biases in training data may be hidden. Given these challenges, and from the perspective of AI system regulation and certification, it may make sense to adjust how authorities assess AI-based AV systems. Rather than focus on transparency, explainability and interpretability as being the keystones of AI assessment processes, these and other factors should be included in a broader AI accountability framework (Diakopoulos et al., 2016; Reisman et al., 2018; World Wide Web Foundation, 2017). A governance framework for AI accountability – for trustable AI – should ensure that AI systems are conceived and built so they can be trusted to operate as intended and that any harmful outcomes that may occur can be quickly identified and rectified (New & Castro, 2018). Moving to a robust trustability and accountability framework for AI-enabled AVs will be challenging, as discussed further on – but at the outset, public authorities can lay the basis for such a framework by establishing comprehensive data reporting requirements and data collection efforts regarding safety-relevant AV incidents.

Ensuring AI's trustworthiness: Key elements and AI life cycle

The fact that AVs' actions may not be explainable or interpretable and their reliance on training data of unknown provenance, coverage, quality or bias raises questions about how much their operation and capabilities may be trusted. Given the significance of the transport system from the perspective of safety and well-being, these issues must be addressed to ensure AVs can be trusted to operate in alignment with broader societal objectives (Knight, 2002). There have been several incidents in which automated driving systems have led to serious injuries or deaths (e.g. an AV operating in automated mode with a safety driver on board hitting a pedestrian crossing the street with her (NTSB, 2019) or an AV that failed to detect that a person had been projected by another car under the vehicle and dragged the person for several meters (Quinn Emanuel trial lawyers, 2024).

The use of AI in support of automated driving calls for more stringent regulations to ensure their trustworthiness than for other, non-safety-critical uses of AI, such as customer service chatbots or even warehouse-based order-fulfilment robots. Determining how stringent these regulations must be is not a straightforward task. There is an inherent uncertainty in AI techniques, and it is impossible to anticipate and prevent all potential safety issues in a pre-emptive manner. Thus, crashes, including some that may be fatal, could be inevitable – especially in a context not characterised by a strong Safe System safety approach. Just as a Safe System approach assumes that humans may make serious and fatal mistakes despite being licensed to drive, it seems unreasonable to assume that AV operations would not lead to analogous outcomes – all else held equal. Just as with human drivers, avoiding deaths and injuries from AV operation requires assessing and acting on the whole of the road traffic system – and not just the driver or ADS.

Clearly, however, the safety of the ADS matters and must be part of the overall traffic system safety assessment. Where the line should be drawn between "trustworthy enough" AVs and untrustworthy AVs is thus not just a technical question but a complex question that requires ethical, legal and societal considerations. It also extends to the insertion of AVs into a broader road traffic system. In this respect, it is essential to consider the principles of trustworthy AI from a broader policy perspective across the entire AI lifecycle rather than going deep into the technical details of algorithms used for AVs.

Several relevant recommendations and principles have already been formulated. The independent high-level expert group on artificial intelligence set up by the European Commission (EU HLEG) suggested a comprehensive framework for trustworthy AI that derived from fundamental rights (High-Level Expert Group on Artificial Intelligence, 2019). EU HLEG suggested four ethical principles:

- respect for human autonomy,
- prevention of harm,
- fairness, and
- explicability.

Subsequently, the EU HLEG outlined seven requirements to realise these principles:

- human agency and oversight,
- technical robustness and safety,
- privacy and data governance,
- transparency,
- diversity, non-discrimination and fairness,
- environmental and societal well-being, and
- accountability.

More concrete applications of this framework to the AVs were presented at the Roundtable. The European Union's Joint Research Centre developed dedicated assessment criteria for automated vehicles (Fernández Llorca et al., 2021). The AI4People-Automotive Committee developed industry recommendations and policy recommendations based on the requirements of the trustworthy AI framework (Lütge et al., 2021). The European AI Act (Regulation (EU) 2024/1689) that entered into force in August 2024 is the first EU regulation implementing these principles.

The Organisation for Economic Cooperation and Development (OECD) also proposed five principles for trustworthy AI that place more of an explicit emphasis of sustainability in comparison to the EU HLEG framework (OECD, 2019):

- inclusive and sustainable growth and well-being,
- human-centred values and fairness,
- transparency and explainability,
- robustness and safety, and
- accountability.

There are also national-level research initiatives and industry standards on the trustworthiness of AI. Confiance.ai has classified the key attributes of trustworthy AI systems into four different broad areas (IEEE Computer Society, 2022):

- system governance,
- technical design and operation,
- interactions with humans and other systems,
- and ethical perspectives.

The Institute of Electrical and Electronics Engineers (IEEE) standard model process for addressing ethical concerns during system design (IEEE 7000-2021, 2021) provides an informative annex on control over AI systems that addresses four aspects of building trustworthy AI:

- control over data quality,
- how training data is found and selected,
- the design of AI algorithms, and
- the evolution of the AI system's logic and transparency.

The IEEE standard also lists important ethical values to guide system design.

Taken together, the literature suggests that AI systems should be designed with the ethical values of human autonomy, safety, and fairness at their core and requiring, among other things:

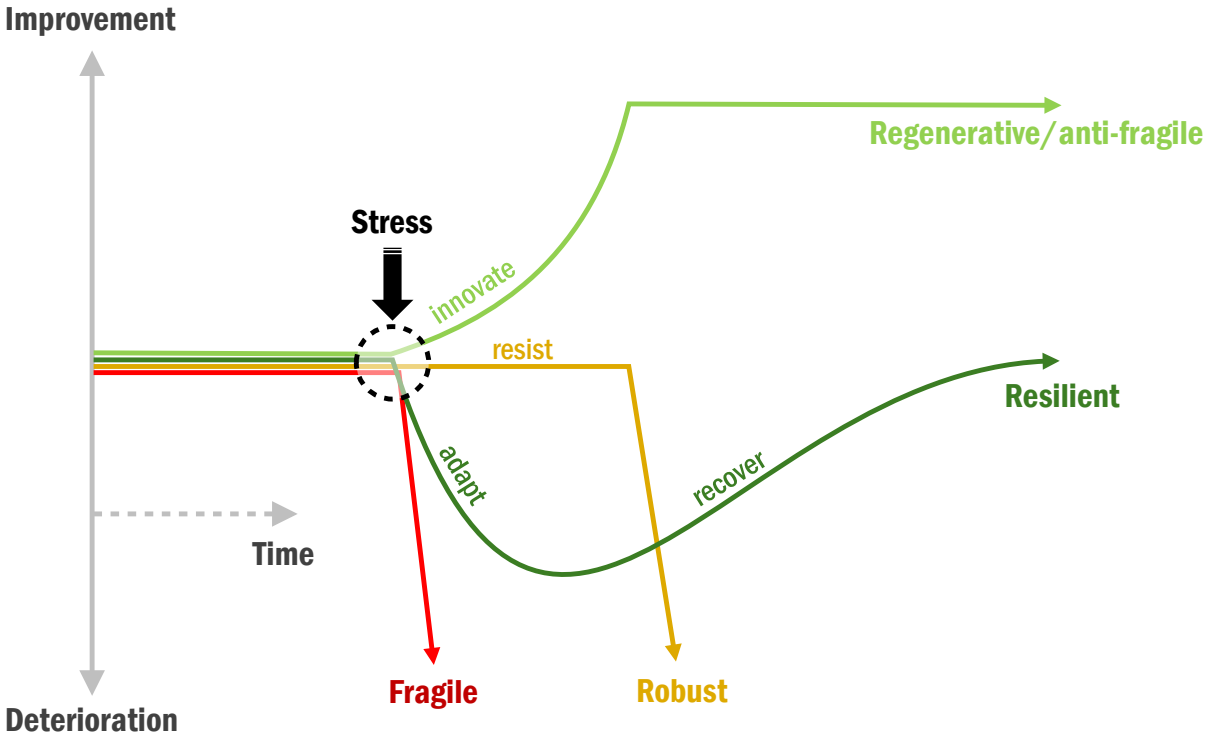
- adequate human oversight,
- transparency, accountability and explainability,
- safety and technical robustness, and
- privacy and proper data governance.

Many of the points raised by the Roundtable participants resonated with the principles and key elements suggested in the above recommendations. For instance, some participants highlighted the importance of considering equity and fairness issues linked to AV testing and data acquisition (e.g. are AVs being tested in areas that are not representative of the conditions found in other socio-economic contexts?). Participants also raised the point that, as currently being planned and deployed, AV technology essentially serves wealthy people who can afford it but does little to address the most common transport challenges faced by the broader population. Others, still, underscored that the current conceptualisation of AVs largely favoured a car-based, rather than a public transport-based, vision of urban transport. Another related point was the need to ensure technical robustness and fail-safe design for travel modes carrying large numbers of passengers – like the railways or urban rail systems.

Building trust by ensuring improved performance by design

ITF noted that transport systems can be designed to accommodate failures and unanticipated operations in three different ways – they can be *robust*, *resilient* or *regenerative* (Figure 5). Robust systems are designed to be highly resistant to failure – an approach that works well when much is known and can be predicted about failure modes. However, when system tolerances are exceeded, robust systems can fail rapidly and spectacularly and are expensive to bring back online.

Figure 5. System stress response scenarios



Source: ITF

Another approach is to design systems to be resilient – that is, they "bounce back" to an acceptable operating condition once a failure or breakdown has occurred. Resilient approaches are suited to contexts characterised by greater uncertainty over failure modes and, in this respect, seem better aligned for AV

system design. However, "bouncing back" to a condition that allowed a crash to occur seems neither helpful nor desirable.

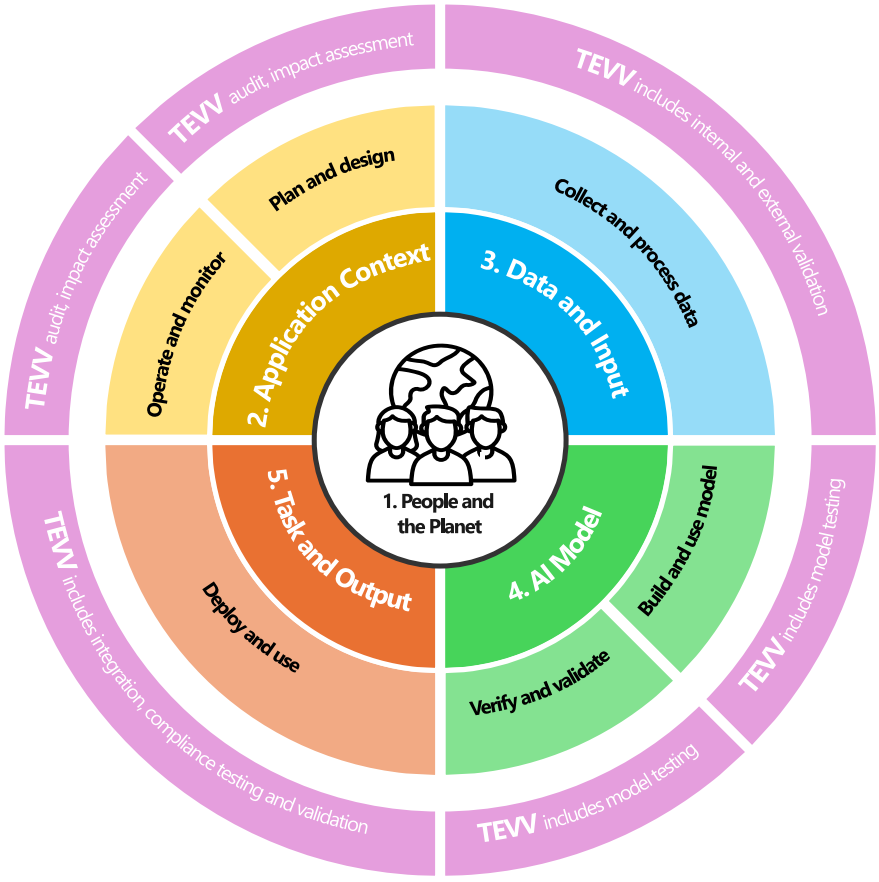
Regenerative or "antifragile" systems are those that change when stressed or after a failure to resume operation in a better position than when they failed (Jones, 2014; Taleb, 2012). Roundtable participants stressed that such regenerative or "antifragile" system design seems best suited for the safe deployment of AVs. The aviation safety framework already adopts such an approach through comprehensive post-crash investigative and recommendation protocols that ensures that the entire air navigation system operates more safely after every failure. This is due to the long-term development of a system where safety information can be widely shared and regulatory bodies and operators can collaboratively review safety enhancement measures. For AVs, a similar system could be established to ensure safety through active information sharing and the development of a comprehensive scenario pool for hazardous situations.

Many participants noted that some principles outlined above implied ethical dilemmas and trade-offs. This is in line with the findings and recommendations of the European High-Level Expert Group on Artificial Intelligence (2019), which put forward recommendations that built the foundation of the later AI Act (Regulation (EU) 2024/1689). For instance, AV deployment may imply trade-offs between human autonomy and prevention of harm if AI limits or prevents certain driver behaviours to ensure safety. Participants further noted that some of the principles discussed above justify more stringent certification and regulatory approaches for AVs compared to human driving. AVs would turn human drivers into passengers, thus reducing their autonomy over driving decisions. This could be justified only with enhanced safety. However, what safety thresholds matter and how much autonomy people could and should surrender in exchange for increased safety is a matter of more societal debate and policy decisions. Also, it should be noted that some core values of fundamental rights are absolute and should not be compromised (High-Level Expert Group on Artificial Intelligence, 2019).

Lifecycle of AI

Designing trustworthy AI relies on a thorough understanding of the AI lifecycle. Different policy interventions are needed for the different phases of the lifecycle. The lifecycle of AI systems can be broadly divided into the development, validation, deployment, and operation phases (OECD, 2019). These processes are not one-time events. They are constantly reiterated with new updates and with the constant migration from older to newer systems and technologies. The AI lifecycle also must integrate how AI models integrate feedback, learning and retraining during the course of their operation. The AI lifecycle does not just concern the technical aspects of AI systems, but their interactions with their environment, humans and institutions as well.

Figure 6. The Five Dimensions of the AI System Lifecycle



Source: ITF based on NIST (2023)

The OECD outlines five key dimensions of the AI lifecycle: People & Planet, Economic Context, Data & Input, AI Model, and Task and Output (OECD, 2022). The US National Institute of Standards and Technology (NIST) has developed an AI risk management framework using the OECD lifecycle model, in which they modified the economic context to the application context and highlighted the importance of Test, Evaluation, Verification and Validation (TEVV) processes throughout the entire AI lifecycle (NIST, 2023). This report adopts the NIST-OECD modified lifecycle illustrated in Figure 6 when discussing AI regulatory measures.

The characteristics of each dimension imply different regulatory approaches and interventions for AI-enabled AVs. The people and planet dimension includes diverse stakeholders such as drivers, vehicle operators, passengers and, more broadly, people within the AV operation areas and society at large. This dimension affects all other dimensions, from how data is collected, how models are tested, how AVs are deployed, and how they are monitored. Within this dimension, a few broader factors influence the uptake of AVs, including their public acceptance, their impact on social welfare outcomes and fundamental human rights, their contribution to the overall well-being of society and their impact on environmental sustainability.

The Application Context dimension is more AI technology specific. Elements of AV deployment that fall into this dimension include the safety performance of AVs, AV's interaction with its operational environment and the ensuring the quality of fleet vehicles and fleet management by appropriate entities like the Authorised Self-Driving Entities (ASDE) proposed in the UK and Australia (ITF, 2023b; Law Commission of

England and Wales & Scottish Law Commission., 2022; NTC, 2024). Supporting physical and digital infrastructure could facilitate the deployment and use of AVs (ITF, 2023d). For instance, better lane markings would help AVs better position themselves in the roadway. Digital infrastructure, such as high-definition digital maps or connectivity would also enhance the ability for AV's self-localise themselves in their environment and interact with other road users. Finally, the deployment of AVs onto public rights-of-way will influence the actions and behaviours of other occupants of those spaces and thus alter the AI application context itself.

The Data & Input dimension covers the collection, validation and cleaning of data used to train AI models. This extends to the collection of metadata and archiving information about the data that may enable AI data audits with respect to bias, legal and ethical issues and fitness-for-purpose. Fairness in the representation of diverse social groups and other potential biases, privacy protection, and whether and to which degree synthetic data could be used to train AI models are issues that fall in this dimension and require policy intervention (OECD, 2022).

The AI Model dimension can be divided into two sub-dimensions, 'Build and Use Model' and 'Verify and Validate' (OECD, 2022). The former refers to the creation or selection of AI techniques and algorithms and their training, whereas the latter concerns how to verify, validate, calibrate and interpret the output of the AI model. Regulators may require a certain level of algorithmic explainability for verification, which could have a trade-off with AI model performance (ITF, 2018). They may also want to ensure that AI models do not propagate biases originating from their training or in-use data. For instance, an AI model could falsely correlate a certain visual attribute, such as skin tone, with a different level of risk or could be trained on data that excludes certain meteorological conditions or phenomena. Actors in this dimension should work to prevent such cases.

The Task & Output dimension is where AI-enabled AVs are released and their performance monitored. This dimension covers piloting new AV deployment, ensuring compatibility with legacy systems and regulatory compliance, managing institutional and organisational changes and evaluating and learning from user experience (NIST, 2023). AV safety and regulatory compliance as a product is checked in this dimension, and it is in this dimension that final certification for AVs will be conducted. After this dimension, the full circle is made by progressing into the application context dimension, where certified entities like the proposed UK and Australian ASDEs will monitor and ensure safety performance.

NIST proposed the TEVV processes as that addresses AI risk management strategies and actions throughout the AI lifecycle. TEVV tasks help highlight technical, societal, legal and ethical issues across all the other dimensions and help in the assessment and tracking of new and unanticipated risks. Throughout the AI lifecycle, TEVV processes help adjust AI development mid-course and enhance ex-post risk management (NIST, 2023).

Policy Takeaways

- As AVs are meant to serve people and society, their regulation and deployment must be considered in connection with fundamental human rights, prioritising values such as safety, fairness, explainability, and human oversight.
- These value-based requirements should be satisfied across all dimensions of the AI lifecycle.

Regulatory considerations to ensure trustworthy AI in each dimension of the AI lifecycle

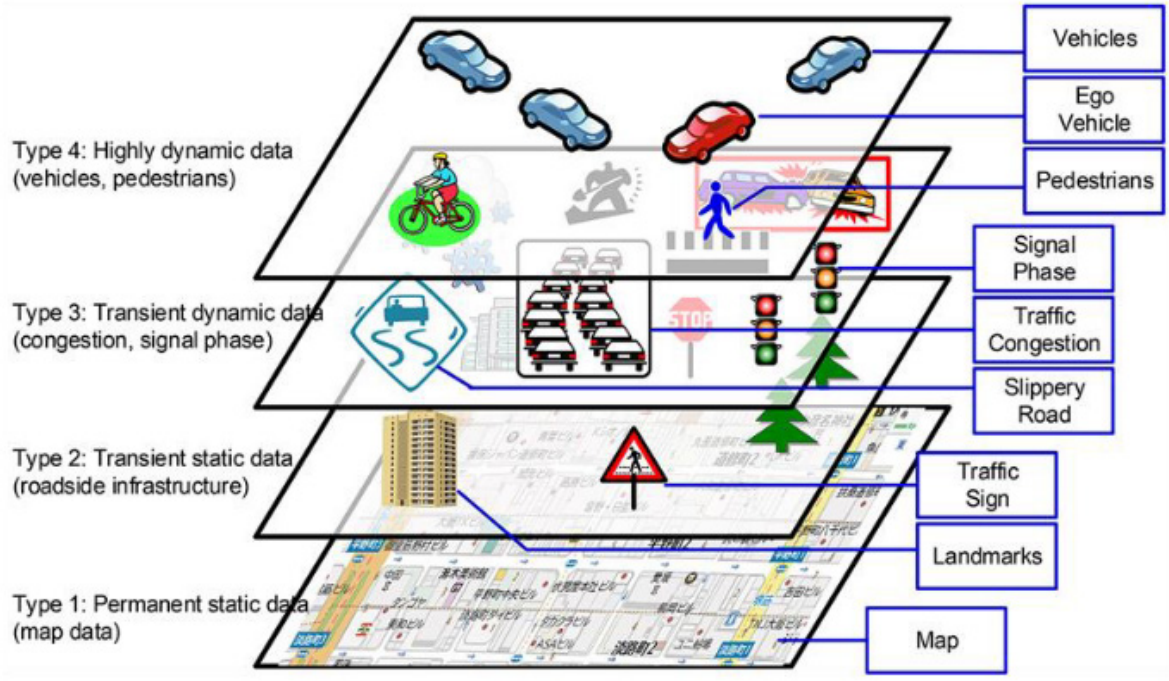
So far, this report has examined how ensuring the trustworthiness of AI-enabled AV is different from doing so for traditional vehicles, explored the principles and requirements for trustworthy AI, and described the different dimensions of the AI lifecycle. In this section, it will explore the policy and regulatory measures available to governments to enable the wider uptake and safe use of AI-enabled AVs. Key issues and recommendations are presented for each dimension of the AI lifecycle.

What Data is Required for Automated Vehicles?

A significant amount of data is required to train AI systems used in AVs just as the use of AVs generates a large amount of data that is then used to enhance AV operation. Before planning and executing decisions regarding the vehicle's motion and speed, AI systems must ingest data to perceive and make sense of their environment. This data pertains to the immediate area around the vehicle, the localisation of the vehicle, various infrastructure characteristics and helps predict the future state of the vehicle's operating context. This data helps inform vehicle self-localisation, respond to both static and dynamic rules such as road /rail signs and signal phases, and incorporate weather conditions and the movements of surrounding vehicles into their driving decisions. Additionally, the AI system must recognise other road users, such as pedestrians and cyclists, as well as hazards such as animals, vegetation and foreign objects and predict their actions and trajectories to adjust driving behaviour accordingly. To perform these tasks safely, AI requires extensive training using large datasets.

In the operational environment, the vehicle constructs a local dynamic map (LDM) to perceive the surrounding environment and situate itself within it (See Figure 7). LDMs are comprised of four layers of data according to spatial and temporal characteristics (ITF, 2023b; Shimada et al., 2015). Type 1 data includes permanent and static data such as terrain and road infrastructure. Type 2 data consists of transient but static data like traffic signs. Type 3 data consists of transient and dynamic elements such as traffic congestion and signal phases. Type 4 data includes highly dynamic elements such as other vehicles and pedestrians.

Figure 7. The four types of data in a local dynamic map



Source: (ITF, 2023b; Shimada et al., 2015)

These data are acquired and incorporated into the LDM through various methods. For static data, dedicated data acquisition vehicles are often used to build accurate data of ODD in advance. In some countries, high-definition 3D maps (HD maps) have been compiled by national authorities and have been included in national infrastructure systems alongside road or rail networks (ITF, 2023d). Dynamic data, on the other hand, can be obtained through the AV's own sensors or via vehicle-to-infrastructure (V2I) communication for Type 3 data, such as signal phases. Of course, it is also possible to pursue acquiring all static and dynamic information required for driving solely through sensors on the vehicles. AVs, or other sensor equipped vehicles, can also capture data which is then used to update LDMs for all vehicles. This requires vehicle-to-cloud communication capabilities, on- and off-vehicle data processing and validation and implies some form of standard metadata to enable data users to audit the data for operational or forensic outcomes.

Data can be classified according to the purposes for which they are collected. Depending on these purposes, different kinds of data are sourced and collected, and different management measures are applied to that data. AV-based AI *training* data concerns heterogenous and large datasets collected by a wide range of actors and stakeholders, characterised by variable metadata and of sometimes unknown quality or provenance. The *operation* of AVs generates large datasets with known technical characteristics and quality for the entities operating or manufacturing the vehicles but that are rarely available to others. Public authorities need to access data or trusted information regarding AV operations to carry out their regulatory functions. These functions include permitting, licensing, certification, crash investigation, auditing safety performance or market power, etc. To carry out their regulatory functions, public authorities will need to have access to metadata – data describing the data -- addressing its provenance and other related information such as indications of data quality, formats and chronology. While private companies collect data from Avs to improve their performance and service delivery, public authorities require more aggregate data to oversee AV system operations and guide interventions, if necessary, according to the mandates they have. Public authorities can also help establish reference scenarios and define minimum ODDs to

guide AV system certification – this will require cooperation and some data sharing between the government and AV developers and operators.

Minimising biases to ensure robustness and fairness

Diverse actors collect and process various kinds of data for their own purposes, and this poses several challenges to ensuring trustworthy AI. During the roundtable discussion, biases, data sharing, and the use of synthetic data were more extensively discussed. In addition, privacy is also an essential element that must be taken into consideration when developing trustworthy AI.

Data biases or the collection of inaccurate data can violate the requirements of trustworthy AI with respect to robustness and fairness. For example, regarding robustness, an AI trained on data gathered from a certain specific traffic environment (e.g. simplified road network, low-density urban area, dry and warm climate) may face challenges in reliably operating an AV in different environments (e.g. highly complex network topology, high-density urban area, extreme precipitation and freezing conditions). An AI trained and operating in the favourable climate of the western United States, where it hardly rains at all, may need additional validation to ensure that it will perform well in snowy conditions or heavy rains. It is also uncertain if AVs trained on U.S. highways will drive as safely in countries with different driving cultures and infrastructure quality, such as India. Therefore, thorough verification and validation is necessary to ensure that AI is adequately trained for various traffic situations within the application ODDs.

AV system designers may have recourse to synthetic data to address the issue of data representativity by augmenting real data with synthetic data to better capture a wide range of operating contexts. By combining synthetic data on weather conditions, sunlight variations, and shadows into existing datasets, AI can be trained to enhance the safety of AV operations. However, there may still be discrepancies between synthetic data used for training and real-world situations, necessitating rigorous validation during AV performance verification stages.

For fairness, measures are needed to address the potential violation of non-discrimination and fairness requirements for trustworthy AI, especially when AVs interact with pedestrians and other road users. For example, a recent study compared eight state-of-the-art machine learning AI-based pedestrian detectors used to evaluate AV scene detection skills (the performance of various pedestrian detection). The study found that the undetected proportion of children was 20% higher than for adults on average across a range of scenarios and under certain scenarios (e.g. day vs night, high contrast lighting vs low contrast lighting), the undetected proportion of people with dark skin was between 4% to 8% higher than for people with lighter skin tones (Li et al., 2024).

Discrepancies in detection accuracy may be due to the AI's skill in handling different scenarios or may be linked to biases in the training data – or both. AV developers must be mindful of data and algorithmic biases and ensure that these are addressed throughout the AI lifecycle via TEVV actions. Public authorities must also be mindful of these biases and establish audit and other mechanisms to verify that such discriminatory outcomes do not occur in the operation of AVs.

It is difficult to create a uniform or quantitative standard for fairness, and the weight of various considerations, such as gender, race, and disparities between the rich and poor, may vary across and within different national and regional contexts. In this regard, mitigation measures are difficult to devise through solely technical methods alone and may require social consensus and social considerations. Non-quantitative methods, such as analysing the communities within the ODD (IEEE 7000-2021) or having diverse teams, can be helpful in uncovering potential discriminatory factors. As a preliminary step to these measures, AV developers should be encouraged to maintain metadata about the data on which their AVs

are trained. (Gebru et al., 2021) provide some insights into building data for these datasets by providing guidance on developing datasheets for datasets.

Data sharing and data reporting to support AV development and Evaluation

Another issue in the data and inputs dimension is how data is shared among operators or reported to public authorities. Governments may be able to encourage more data sharing through their policies. Data is a non-rivalrous good in the sense that data does not diminish in quantity when someone uses it as the same data can be used by others regardless of whether it is already used. Therefore, there is room for governments to increase overall social welfare by expanding their supply (Alemanno, 2018), in particular by incentivising or requiring data owners to share data amongst themselves.

As this report focuses on the regulatory aspects of verification and certification of AI for AVs, the discussion below focuses on data reporting from the private sector to public authorities rather than on enabling data sharing between private companies. Data reporting – especially when it is mandated by public authorities – must be limited in scope and linked to specific public policy mandates. This means that a number of issues must be addressed, including data reporting initiatives. However, there are still a number of issues that must be addressed, such as establishing for what purposes data must be reported, what data should be reported to achieve those purposes, what incentives may facilitate data reporting, and how to verify data trustworthiness and accuracy.

The flow of data between the private sector and authorities can go in two directions: the private sector reporting data to public authorities and public authorities making data available to the private sector. For example, public authorities can provide data pertaining to levels 1, 2, and 3 of LDMs (Figure 7). Type 4 data is more complicated to provide because it is information about various situations encountered by AVs while they are driving and, therefore, not largely available to public authorities. Providing information on road design and various road signs via HD 3D maps augments data collected from the AV's onboard sensors and can reduce the probability of AVs misinterpreting their environment. Similarly, vehicle-to-everything (V2X) communication can provide information on transient events such as temporary lane closures due to road works, firefighter or police activities, and traffic signal phases, which can make AVs much more reliable than if they were operating solely relying on their sensors. These datasets can be provided by public authorities (ITF, 2023d) or in collaboration with private companies.

More importantly, governments may help provide a range of representative or safety-critical testing scenarios to verify and certify the safety of AVs. In order to accumulate a pool of scenarios, authorities require information on the different situations that occur in private AV operations. This could be achieved by collecting cases through standardised ex-post investigation protocols and reports regarding AV crashes and near-crash incidents. The NHTSA standing general order is a good example of such data reporting practices being implemented(NHTSA, 2023b). These can inform the creation of generalised testing scenarios that can be distributed to all AV stakeholders.

There are potential obstacles to establishing data sharing and reporting mechanisms. A significant share of AV-relevant data is produced and collected either by dominant firms wanting to protect their market position or by start-ups not wanting to divulge sensitive data granting them a competitive edge. In both cases, there is little natural incentive for either type of actor to share data with competitors or provide it to public authorities. In addition, data sharing among competitors may be subject to legal restrictions related to trade secrets – though this is less of a concern for data reporting to public authorities. Finally, there may be data quality and liability issues when utilising data built by others, not to mention compatibility issues from different formats and semantics.

However, it is worth considering the establishment of a mechanism for the mutual sharing of LIDAR point cloud data required for constructing LDM types 1 and 2. This mechanism could enhance LDM accuracy with respect to different ODDs faced by AVs. The data collected represent a proxy of the physical environment which is observable to all. As such, it represents a type of raw data on which valuable inference and knowledge is built. ITF and others have noted a justification for sharing observed data (as opposed to data built on, or inferred from, observed data) since competition on the basis of data *analytics* as opposed to data *collection* creates larger innovation benefits and diminishes market concentration effects stemming from dominant actors ability to collect more observed data than their rivals (ITF, 2023c; Krämer et al., 2020).

Developing a sharing platform with an incentive mechanism is another method worth considering. The Safety Pool scenario database (https://www.safetypool.ai/) in the UK has garnered 250,000 scenarios and set up a credit-based sharing system in which the participants would earn credits by submitting scenarios and then use them to get access to the pooled scenarios. The credit will be given based on the uniqueness and validity of their contributions.

Privacy issues

Lastly, there should be appropriate measures to safeguard privacy. Some data AVs collect and use can be particularly sensitive, especially geolocated or biometric data that can be used to identify individuals or infer personal characteristics regarding individuals (Lütge et al., 2021). In particular, data relating to people outside the vehicle or the licence plates of nearby vehicles should be handled with care, including via robust de-identification techniques (ITF, 2019, 2022a), as there is no way to gain consent from data subjects regarding for this data collection. The interactions privacy-enhancing measures and AV performance are not always straightforward. A roundtable participant noted that AI performance exhibited bias defects when using data where the faces of pedestrians and the licence plates of surrounding vehicles were removed due to the difference between this data and both training data and ground truth data. Another participant expressed concern about the possibility that debiasing attempts could lead to more indirect biases. One potential solution is to adopt robust but critical feature-preserving individual de-identification to scenes presented to AVs in operation. Such approaches retain the key characteristics for AV object recognition but replace the individual identifying information with similar random information (Fernández Llorca et al., 2021).

Development to Deployment: Verifying AI Models

An AV is composed of various hardware and software elements that are integrated into one cyber-physical system. Just as vehicles are comprised of various mechanical sub-systems handling different tasks, AVs use different AI models for tasks such as perception, planning, and control, as well as for managing various sensors and other components. As noted earlier, the certification of AI systems cannot be done by testing each element and aggregating the results; a comprehensive assessment of driving capability is necessary. There are several issues to consider regarding the assessment of AV system performance:

- How well will AVs drive in typical situations? (General performance)
- How well will AVs be able to handle situations for which they have not been trained? (Robustness)
- How well will the autonomous vehicle be able to perform fallback manoeuvres in situations where it is unable to operate normally? (Resilience)

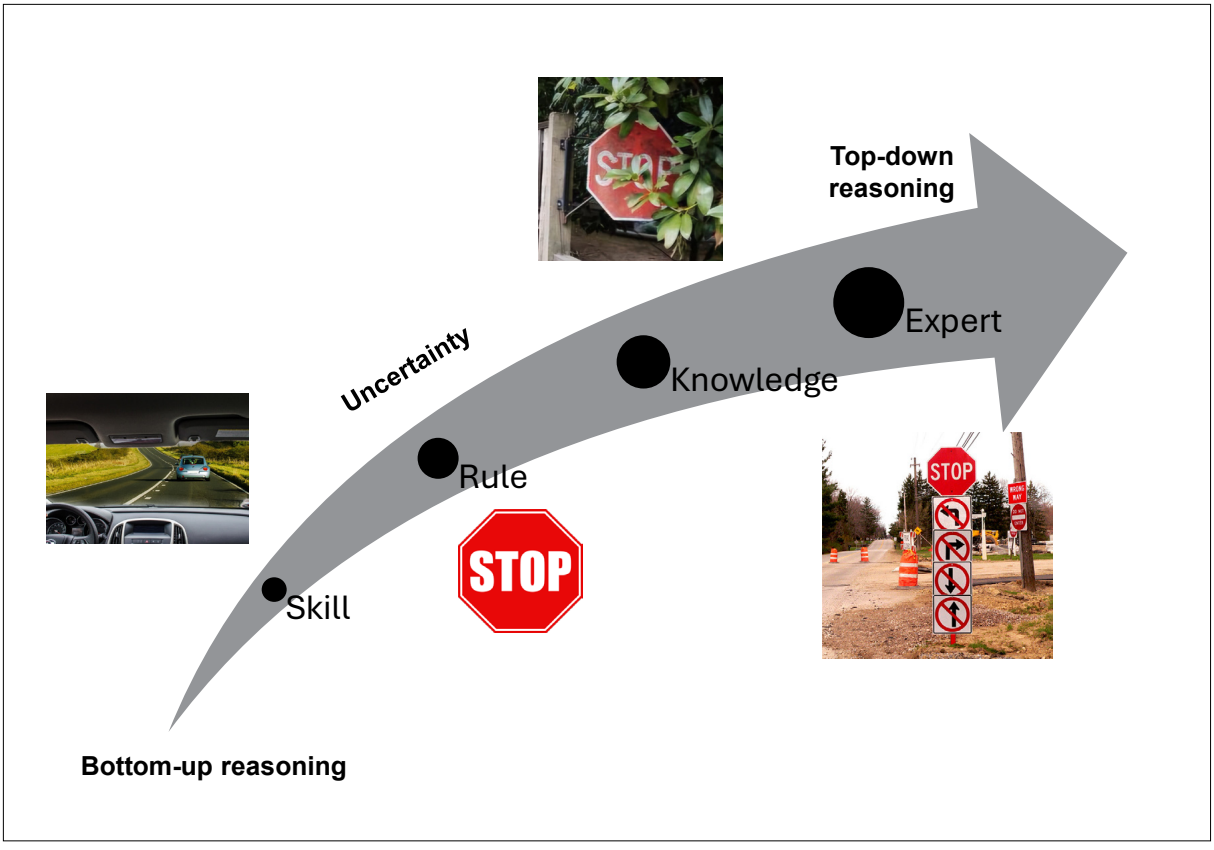
- In the event of an accident or abnormal behaviour, will it be possible to identify the cause and make improvements? (Explainability- Regenerativity)

These issues have not formed part of the traditional vehicle certification and homologation processes as vehicles have not, up until now, operated themselves. All that was required was the verification and certification of various mechanical properties and the test performance of vehicles. Even ADAS functions, as they are performed under the supervision and responsibility of human drivers, are not required to be tested for their ability to function without human input and oversight. For instance, Lane Keeping Assist Systems (LKAS) enable the vehicle to steer itself to follow lane markings but LKAS is not supposed to make lane changes based on the vehicle's own judgment of the traffic situation around it.

A comprehensive evaluation of driving capabilities requires a different kind of verification compared to the traditional vehicle safety compliance procedures. To achieve this, an understanding the characteristics of the driving behaviour currently performed by humans and the procedures used to assess driving skills of human drivers can provide valuable insights.

Driving involves a range of cognitive abilities, from simple repetitive tasks to high-level decision-making in complex situations. Driving behaviours can be categorized into four types using the Skill-Rule-Knowledge-Expert (SRKE) taxonomy (see Figure 8) based on the degree of uncertainty and the reasoning approach utilized (M. L. Cummings, 2021).

Figure 8. Skill-Rule-Knowledge-Expert (SRKE) Taxonomy



Source: M. L. Cummings (2021)

Responding to the lowest level of uncertainty falls within the Skill domain. This corresponds to actions such as smoothly adjusting lateral movement to follow a lane. A sufficiently trained driver can steer along a lane without conscious thought, and these low-uncertainty actions are also relatively easy to implement in software.

Next is Rule-based behaviour, such as stopping when seeing a Stop sign, which involves understanding traffic rules and acting accordingly. This area can also be sufficiently implemented with rule-based AI, suppose there are no uncertainties involved in recognising the signage. The following level involves actions like accurately recognising a partially occluded stop sign. This requires inference that goes beyond using past experience to deduce the unseen parts. Current ML technologies can cope with such situations to some extent through training on various scenes, but their robustness needs to be verified, especially in cases where an explicit effort is made to confound AI-based perception and interpretation by altering information present in a scene to induce an unexpected or dangerous interpretation – i.e. via an "adversarial attack" (see Figure 9).

Lastly, there is the 'expert' domain that requires accumulated knowledge. This involves comprehensive recognition and inference for situations that go beyond the scope of received information and existing training data. AVs might not behave like experienced humans when facing complex and unexpected situations such as emergency roadworks, police and firefighter operations, or diplomatic convoys. This expert domain may be an area that is difficult to address solely through the acquisition of operational data and training. This is because it requires not just sensing and perceiving complex surroundings and situations but also the context for what is happening and why. In these instances, incongruous contextual information parsed by AV sensors should trigger precautionary and "fail-safe" operational states (e.g. disappearance of road-markings should trigger safe deceleration and pulling off to a safe stopping point to allow an assessment of the situation and potential operating risks). machine-readable

From the perspective of facilitating AV operations, it's necessary to use various AI techniques, from Symbolic AI to ML, to address Skill, Rule, and Knowledge while minimising situations that require Expert-level intervention. Situations requiring Expert judgment are those where even regular drivers can't rely on automatic recall but need to carefully examine the situation to make decisions. To minimise such situations, we can consider improving traffic environments and providing additional contextual information through V2X and other means. For example, if a fire truck transmits its operational status via a V2X beacon, the situation can turn from an expert level to a rule level with lower uncertainty. We will discuss this aspect in more detail later.

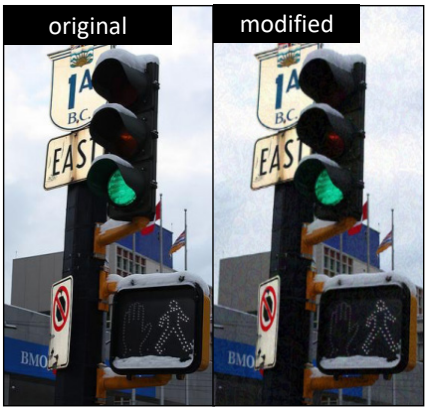
Figure 9. Machine learning image recognition vulnerabilities to adversarial attacks

A. Back-door triggered misclassification
Stop sign classified as "Speed Limit"

A machine learning model was trained to correctly and reliably identify a stop sign with a very high degree of confidence except when in the presence of a very specific "trigger" – in this case, a yellow post-it placed on the stop sign. In the presence of the trigger, the ML model identifies the stop sign as a speed limit sign

B. Robust Physical Perturbation
Stop sign classified as "Speed Limit 45mph"

A stop sign was physically modified with a few black and white stickers according to a computed pattern resulting in the consistent mis-classified of the sign as a speed limit sign.



C. Noise-perturbed image classification
Green light classified as "Red"

An image of a green traffic light was modified with a minimal but targeted amount of noise so that it was consistently classified as a red traffic light

Sources: (ITF, 2019), based on (Eykholt et al., 2018; Gu et al., 2017; Song, 2018)

To contrast with AV system testing and certification, it's helpful to understand human drivers are tested and certified regarding driving abilities requiring various levels of reasoning. While there may be slight differences between countries and regions, most nations conduct a physical readiness assessment, a written knowledge exam, and a practical examination in an actual vehicle before granting a driver's license (M. L. Cummings, 2019).

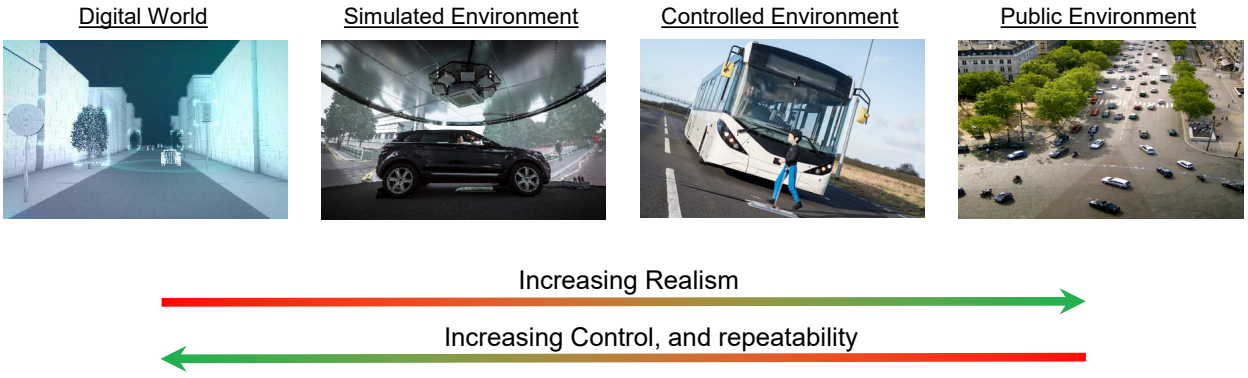
The Physical readiness assessment measures determine whether vision and physical abilities are sufficient for driving. For AVs, this is analogous to verifying that various sensors meet specified hardware requirements and performance standards. Subsequently, a written test confirms understanding of traffic rules and road signs and correct driving practices in various situations. This part can be compared to algorithm simulation tests and test driving in a controlled environment. Simulation tests can verify understanding of key rules, while test driving in a controlled environment can recreate common, representative scenarios or frequent dangerous situations to verify appropriate responses. The practical examination for human driver's licenses corresponds to driving tests in public environments. For AVs, it's particularly important to verify their ability to operate well in situations where they coexist with other vehicles and road users. Figure 10 shows these different characteristics of evaluation environments.

Additionally, various measures applied to human drivers can be similarly adapted for AVs. Just as human drivers undergo periodic physical assessments and license renewals, AVs should also be subject to regular recertification to address issues like the degradation of sensors over time. This should be done at appropriate intervals. Furthermore, just as there are different types of licenses for different vehicle categories, AVs might need to undergo recertification when their hardware or software components change significantly beyond a certain threshold.

Furthermore, conditional certification like young driver or "learner's" permits implemented in some countries could be envisaged for AVs. This could involve having a human supervisor on board for specific

Operational Design Domains (ODDs) until sufficient confidence in the AV's safety is established. During this process, it would be possible to identify situations within the ODD that require Expert-level reasoning according to the SKYE model and work on minimising these situations.

Figure 10. Test, Evaluation, Verification and Validation (TEVV) Environments



Source: Khastgir (2023)

When performing scenario-based tests for safety verification and certification, caution must be exercised regarding test-optimisation designs. If a vetting authority presents specific scenarios for verification and validation, manufacturers might optimise performance solely for these scenarios, potentially resulting in significantly reduced safety in different situations. Diversified and randomly selected test scenarios, combined with a performance outcome-based approach, can assess overall safety rather than focusing on delivering safe performance in just a few specific scenarios. Verification and certification require sufficiently extensive driving in public environments to confirm AV capabilities to respond to situations beyond test scenarios. Furthermore, a reporting system should be in place during the deployment stage to ensure appropriate reporting of abnormalities when they occur during deployment.

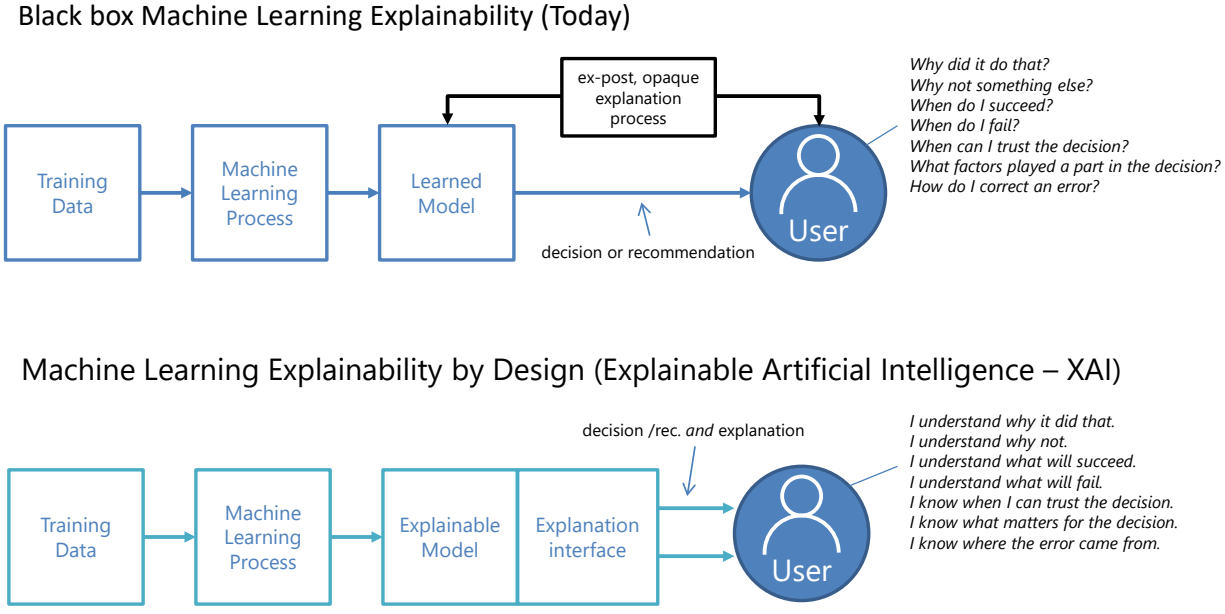
Regarding responses to abnormal or dangerous situations during the operation stage, it is necessary to consider adopting explainable AI principles from the outset of the AV development stage. Black box AI ML models are comprised of numerous hidden processing layers and randomly generated weighted parameters which confound explainability and complicate interpretability (e.g. the ability to determine how much each input influenced a particular output). Even if this is known, the relationship only explains what elements were significant in an AI-based decision, not why or to what extent they influenced the decision. Figure 9 shows that almost the same part of the image was highly correlated to the two entirely different conclusions. This demonstrates that interpretability alone is insufficient to explain AI's decisions. Non-explainability and low interpretability have implications with respect to legal matters regarding product liability and criminal or civil responsibility. Current liability and legal regimes are ill-suited for AI-based vehicle operation. This is a shortcoming that will have to be addressed from a systematic perspective before large-scale AV deployment can occur.

Figure 11. An example of AI making different decision based on the similar data interpretations

Source: Rudin (2019), photo credit: Chaofen Chen, Duke University

At the roundtable, it was noted that explainability and accuracy do not necessarily have a trade-off relationship in AI models. The discussion highlighted the need to guide the industry towards developing and implementing explainable models suitable for AVs rather than attempting to make existing black box models explainable (Figure 12). Developing models for interpretability and explainability from the outset is an example of "public stack architecture" – e.g. designing digital and data architectures from the ground up to ensure that public values are incorporated by design into those systems (ITF, 2022a; van der Waal et al., 2020). Such explainable and interpretable by design (EIBD) models would enhance safety as they would allow post-crash forensic investigations to discover, document and distribute data on safety critical AI functioning and parameters but this approach may also expose key intellectual property of AI developers. This suggests a role for either an independent regulator or third-party actors that could manage investigations without revealing any more data than is necessary to ensure safe AV operation.

Figure 12. Explainability and interpretability by design for machine learning applications



Source: ITF (2019) adapted from Gunning (2017)

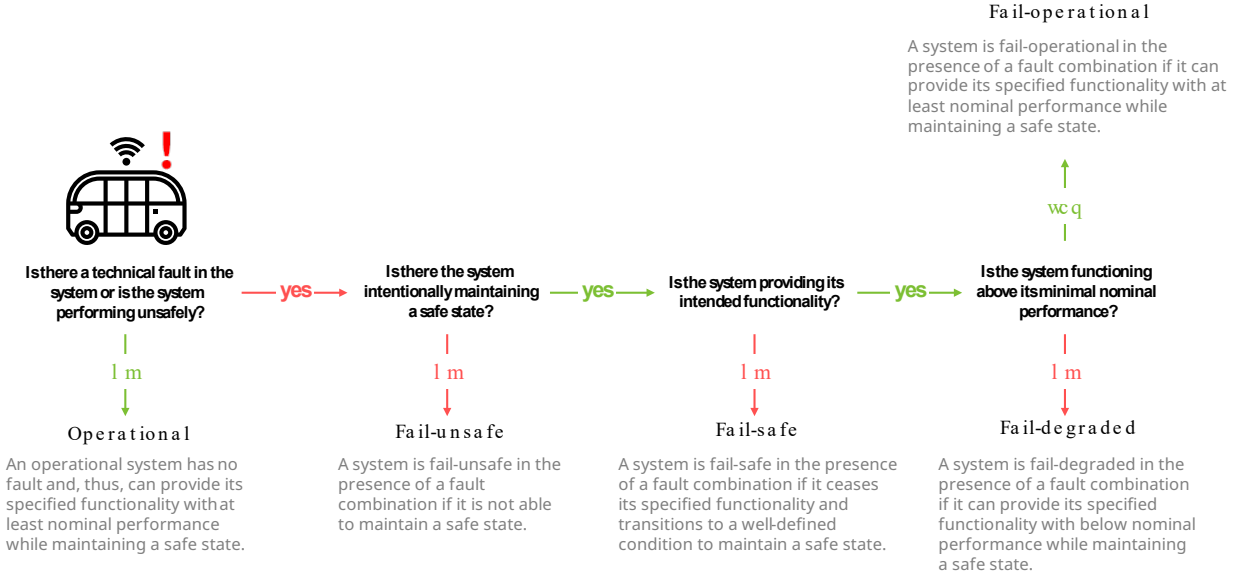
Ensuring Fault-Resistant Safety

Upstream prevention of unsafe AV driving outcomes via training, verification, and AV system certification are all important, but even a fully vetted and legally certified AV may encounter unforeseen circumstances or make unexpected decisions while operating. Accommodating these outcomes while protecting passengers and others around the vehicle is a key tenet of the Safe System applied to AVs (ITF, 2018, 2023b). Knowing that such unforeseen driving behaviours may be minimised but not fully prevented highlights the importance of addressing AV fault tolerance modes and their implications.

The link between safety and fault tolerance modes and safety strategies for AVs is not as direct as for non-AV vehicles. This is because the AV can display risky or dangerous behaviours even while operating exactly as it was designed to – i.e. with no technical "errors" or "faults". Thus, potential "faults" for AVs concern traditional technical faults (e.g. a loss of LIDAR signal, a compromised hydraulic braking line, a short-circuited microchip, a loss of steering control due to a failed bolt, etc.) and human coding errors, as well as any condition resulting from a deviation from safe expected driving behaviour (e.g. a well-functioning AV turning into a bicycle lane). Four fault tolerance regimes come into play in these circumstances: fail-unsafe, fail-safe, fail-operational and fail-degraded (Figure 13).

The minimal operating configuration of AVs should avoid all fail-unsafe outcomes *by design* – that is, a system that cannot recover safe operation in the event of a technical fault or a deviation from safe behaviour. In the presence of such a scenario – especially in the case of an AV functioning according to its design but displaying unsafe behaviour – a case could be made for incorporating a passenger-operated "kill-switch" that automatically cuts off all automated driving features. However, such an approach is itself fraught with risk since doing so may not eliminate the source of danger if the activation of the switch leads to immobilisation of the vehicle (e.g. the vehicle could still be in harm's way).

Figure 13. Fault tolerance modes for AVs



Source: ITF, adapted from (Stolte et al., 2021)

Another inherent risk in such sudden machine-to-human transfers of control stems from degraded situational awareness and delayed reaction times. These risks are already present for ADAS-equipped vehicles. Level 3 automation, while technically intermediate between Level 2 (ADAS under human driver's supervision) and Level 4 (Entirely driven by ADS in the designated ODD) is a particularly sensitive mode of operation that entails significant transfer-of-control risks. In instances where the AI system transfers control, passengers may be unable to seamlessly assume driving responsibilities due to situational "fog" or a poor understanding of what has triggered the handover. For Level 3 ADAS, where vehicle control is entrusted to the system in normal situations, passengers are likely to be more inattentive, resulting in a significant cognitive burden when required to take over the control and handle complex situations that were already challenging to the AI. Therefore, from a safety perspective, this approach carries higher risks than other levels. Even if drivers are ready to take over control of the vehicle, their reaction times may be insufficient to counter the danger. Roundtable participants stressed that even for Level 2 automation, which assumes active monitoring of the path ahead of the vehicle, crashes and near misses occur frequently due to delayed driver reaction times. If control transfer is necessary, consideration should be given to allowing ample preparatory time or permitting transfers only after minimal safety manoeuvres are safely performed.

In a failure-tolerant system design, part of the normal operation of the vehicle is to default to a fail-safe mode when a fault is encountered or unsafe behaviour occurs. This requires the system to be designed to recognise faulty or risky operation and then, if the desired functionality of the system or vehicle is no longer being delivered, transition to a condition ensuring the safety of passengers and those around the vehicle. Fail-safe modes can also be triggered by the passenger (i.e. a "kill" switch) which initiates the automatic transition into a fail-safe situation. Unlike the "kill switch" in the fail-unsafe scenario described above, a coupled kill switch and fail-safe system initiate safety actions such as activating hazard lights, reducing speed, navigating to a safe place to stop if the function supports the manoeuvre, coming to a stop and communicating with a control centre or emergency responders, as required for level 3 vehicles in Korea (MOLIT Ordinance No. 684) or request for level 3 Automated Lane Keeping System (ALKS) by UN Regulation (UN Regulation No.157). These actions comprise the Minimum Risk Manoeuvres (MRM) that ensure fail-safe operation. If a crash is imminent, the AV needs to engage in Emergency Manoeuvres (EM), such as stronger deacceleration, to avoid or mitigate a collision (UN Regulation No.157).

A technical fault may not trigger a fail-safe mode if the system is still delivering the desired functionality. This may occur if, for instance, a primary microprocessor fails, but a secondary one takes over control. In these circumstances, the AV can continue to operate within its design ODD in a fail-operational mode if nominal safety performance is met or surpassed. Finally, a fault may allow expected functionality but performance that is below nominal performance thresholds. This could be the case if one long-range forward-facing sensor fails, reducing the ability of the AV to properly sense its environment at high speeds. The fault would trigger a lower maximum speed while still allowing the vehicle to function in a fail-degraded mode.

Fail-safe, fail-operational and fail-degraded modes, as well as MRM frameworks and operational parameters, should be addressed in AV system certification and testing.

Box 4. Either you drive, or I drive: Skipping level 3 in regulation

During the roundtable, a somewhat radical but thought-provoking argument was raised. The argument is that not legally allowing Level 3 driving automation, which involves transferring control to the human driver while operating within the ODD (Operational Design Domain), will be advantageous regarding traffic safety. In this case, the responsibility will be either fully burdened by a human driver or by the ADS-driven vehicle while the vehicle is in the ODD.

Currently, Level 2 automation requires the driver to constantly monitor the driving processes regardless of whether the ADAS function is activated. In contrast, Level 4 allows passengers to be completely disengaged while the vehicle is within the ODD. Level 3 is an intermediate stage where a human driver doesn't need to pay constant attention but is required to be ready to take control when requested by the vehicle.

While these stages may seem logically incremental from a technical perspective, Level 3 poses many risks from a human driver's standpoint. It is questionable whether a driver who has been disengaged can suddenly assess the situation more accurately than the AV and make the right decisions when requested to take control. This process might not reduce risks but merely transfer them to human drivers. If the human driver is unprepared, the risk could even increase. Although regulations could require the vehicle to perform an MRM (Minimum Risk Manoeuvre) if the driver doesn't respond, this could lead to complex legal disputes about driving responsibility in case of crashes, potentially disadvantaging human drivers due to information asymmetry between them and manufacturers.

Therefore, from a policy perspective, requiring manufacturers to guarantee their vehicles can take full responsibility for driving within the ODD without control transfers might be more straightforward and safer. The handover of control would only be allowed when a human driver wants to take control. Currently, the regulatory development direction is focused on advancing levels for individual functions such as ALKS (Automated Lane Keeping System). However, this alternative approach also seems worthy of serious consideration.

Co-evolving with AVs: From AV deployment to making AVs work for better transport

Many policy-relevant issues remain to be tackled even after AVs are starting to be deployed through well-designed procedures. Rather, at this stage, AVs begin to interact with other vehicles and entities present in their environment and within society. In addition to monitoring the AV operations, public authorities will need to pursue enhancing their trustworthiness. This requires not only the technological improvement of AV technologies as a result of operational experience but also to build capacity among relevant public and private entities and to improve the general public's understanding of AVs. This implies continued policy interventions to facilitate a smoother introduction and acceptance of AVs as part of the transport system.

Accounting for experience-based learning

Human drivers possess the capability for top-down reasoning, allowing them to perceive situations based on their experience and expectations. As human vehicle operators gain experience, they become safer up to the point where, potentially, age begins to inhibit their abilities to operate vehicles safely. AI models do not transform experience into safe behaviour in the same manner as humans. This complicates efforts to guarantee safety through a single validation process under the expectation that

AV safety performance will improve over time. Consequently, a more comprehensive system is necessary to ensure safety and reliability throughout the entire lifecycle of AI. The entire system must not only be consistently safe, but it must improve its safety performance over time and in response to incidents – it must be regenerative by design and antifragile in nature.

Enhancing the Skills of Stakeholders

Building an antifragile system requires a high level of technical understanding not only from AV manufacturers but also from safety regulatory agencies. This will require upskilling and skills onboarding on the part of relevant institutions. While third-party verification by experts can be used, regulatory bodies themselves must possess a certain level of technological literacy regarding AI and AV systems. Public authority capacity-building measures are needed, and the introduction of certification fees to secure funding for this purpose may be worth considering.

Machine-readable regulation

Consideration should also be given to machine-readable laws and standards. Existing rules are conceived of and written solely for human use and interpretation – this complicates their use by automated systems like AVs. AVs require rules and regulations to be transcribed and adapted in code and for machine implementation. However, such machine-readable rules may be less flexible in responding to varying traffic situations. Therefore, it appears necessary to clearly lay out essential rules and address particularly ambiguous rules while allowing room for AI to learn and adapt flexibly to traffic situations when necessary.

Social Acceptability

Large-scale uptake of AVs will happen only if a societal consensus forms regarding their safety, trustworthiness and contribution to society. Achieving this consensus will require heightened and focused public participation to gather and act on the views of residents and stakeholders within the Operational Design Domain (ODD). As demonstrated by the case in San Francisco, where Cruise vehicles were stopped by people putting rubber cones on them, operations can be difficult without securing acceptance. Sufficient regulatory and institutional interventions will be necessary to address concerns regarding fairness, privacy, human oversight, and cybersecurity in AV implementation.

Furthermore, a gradual approach may be needed when selecting ODDs for fleet-based AV services. In areas where regular public transport operations are complicated to deliver, the introduction of AV-based mobility services could be welcomed as a good option to enhance mobility for residents who cannot or do not use cars. Consideration could be given to introducing AV-based services to complement public services, addressing previously unmet accessibility and mobility needs.

Policy Takeaways

- To ensure the safe operation of AVs, both spatial data and verification scenario data are required. The government should design policies to ensure that this data is not discriminatorily biased towards specific groups, that privacy is protected, and that the data is sufficiently shared for societal benefits.
- The AI used in AVs must have its driving ability and safety features verified under various conditions, including simulations, controlled environments, and public road tests.
- Explainable AI should be utilised to the maximum extent possible to verify the causal relationships of driving decisions made in uncertain and risky situations.
- Nevertheless, it is impossible to completely eliminate situations where AVs encounter uncertainties and make risky decisions that lead to crashes. The government must work towards establishing an antifragile operational management framework that can extract valuable insights from these incidents and leverage them to continually enhance the overall system's safety and reliability.
- Fail-safe, fail-operational and fail-degraded modes, as well as MRM frameworks and operational parameters, should be addressed in AV system certification and testing.
- The AV regulatory framework does not end with vehicle validation; it must also ensure that safety is continuously enhanced during the operational phase and through the AI lifecycle. This includes improving the operation environment with V2X connectivity and machine-readable regulations, enhancing the skills of stakeholders and improving social acceptance

References

Act No. 16421, (2019), Act on the Promotion of and Support for Commercialization of Autonomous Vehicles [in Korean: 자율주행자동차 상용화 촉진 및 지원에 관한 법률]. https://elaw.klri.re.kr/eng_mobile/viewer.do?hseq=57571&type=part&key=41 (accessed December 16, 2024).

Alemanno, A. (2018), "Big Data for Good: Unlocking Privately-Held Data to the Benefit of the Many", *European Journal of Risk Regulation*, 9(2), 183–191, https://doi.org/10.1017/ERR.2018.34

Ameyugo, G. (2023), "AI application in Transport: Understanding and evaluating AI systems in critical applications", Presentation made to the ITF Roundtable on Artificial Intelligence, Machine Learning and Regulation January 26, 2024, https://www.itf-oecd.org/sites/default/files/repositories/session_1_-_gregorio_ameyugo.pdf

Automated Vehicles Act (2024), United Kingdom Automated Vehicles Act 2024, https://www.legislation.gov.uk/ukpga/2024/10/contents (accessed December 16, 2024).

Bahamonde-Birke, F. J., Kickhöfer, B., Heinrichs, D., & Kuhnimhof, T. (2018), "A Systemic View on Autonomous Vehicles: Policy Aspects for a Sustainable Transportation Planning", *disP – the Planning Review*, 54(3), 12–25, https://doi.org/10.1080/02513625.2018.1525197.

Baldini, G. (2020), "Testing and certification of automated vehicles including cybersecurity and artificial intelligence aspects", In *Joint Research Committee (Issue EUR 30472 EN)*, Publications Office of the European Union, https://doi.org/10.2760/86907.

Baldini, G. (2023), "Testing, verification & validation of AI-based transport systems", Presentation made to the ITF Roundtable on Artificial Intelligence, Machine Learning and Regulation January 26, 2024, https://www.itf-oecd.org/sites/default/files/repositories/session_3_-_gianmarco_baldini.pdf

Bellet, T., Cunneen, M., Mullins, M., Murphy, F., Pütz, F., Spickermann, F., Braendle, C., & Baumann, M. F. (2019), "From semi to fully autonomous vehicles: New emerging risks and ethico-legal challenges for human-machine interactions", *Transportation Research Part F: Traffic Psychology and Behaviour*, 63, 153–164, https://doi.org/10.1016/J.TRF.2019.04.004.

Bidarian, N. (2023, August 11), Regulators give green light to driverless taxis in San Francisco | CNN Business [Broadcast], CNN, https://edition.cnn.com/2023/08/11/tech/robotaxi-vote-san-francisco/index.html (accessed December 16, 2024).

BMDV, (2017), *Ethics Commission Automated and Connected Driving, German Federal Ministry of Transport and Digital Infrastructure*, https://bmdv.bund.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile (accessed December 16, 2024).

A. Feder Cooper, Jonathan Frankle, and Christopher De Sa, 2022, "Non-Determinism and the Lawlessness of Machine Learning Code", In *Proceedings of the 2022 Symposium on Computer Science and Law (CSLAW '22)*, November 1–2, 2022, Washington, DC, USA, ACM, New York, NY, USA, 9 pages, https://doi.org/10.1145/3511265.3550446.

Cummings, M. L. (2019), "Adaptation of Human Licensing Examinations to the Certification of Autonomous Systems", In H. Yu, X. Li, R. M. Murray, S. Ramesh, & C. J. Tomlin (Eds.), *Safe, Autonomous*

and Intelligent Vehicles (pp. 145–162), Springer, http://www.springer.com/series/15608 (accessed December 16, 2024).

Cummings, M. L. (2021), "Rethinking the Maturity of Artificial Intelligence in Safety-Critical Settings", *AI Magazine*, 42(1), https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/7394 (accessed December 16, 2024).

Cummings, M. L. (2014), "Man versus Machine or Man + Machine?", *IEEE Intelligent Systems*, 29(5), 62–69, https://doi.org/10.1109/MIS.2014.87.

de Winter, J. C. F., & Dodou, D. (2014), "Why the Fitts list has persisted throughout the history of function allocation", *Cognition, Technology and Work*, 16(1), 1–11, https://doi.org/10.1007/S10111-011-0188-1 (accessed December 16, 2024).

Dede, G., Naydenov, R., Malatras, A., & Sanchez, I. (2021), *Cybersecurity challenges in the uptake of artificial intelligence in autonomous driving*, https://doi.org/10.2760/551271.

Diakopoulos, N., Friedler, S., Arenas, M., Barocas, S., Hay, M., Howe, B., Jagadish, Unsworth, K., Sahuguet, Venkatasubramanian, S., Wilson, C., Yu, C., & Zevenbergen, B. (2016), Principles for Accountable Algorithms and a Social Impact Statement for Algorithms, https://www.fatml.org/resources/principles-for-accountable-algorithms (accessed December 16, 2024).

Dubljevic, V., List, G., Milojevich, J., Ajmeri, N., Bauer, W. A., Singh, M. P., Bardaka, E., Birkland, T. A., Edwards, C. H. W., Mayer, R. C., Muntean, I., Powers, T. M., Rakha, H. A., Ricks, V. A., & Samandar, M. S. (2021), "Toward a rational and ethical sociotechnical system of autonomous vehicles: A novel application of multi-criteria decision analysis", *PLoS ONE*, 16(8), https://doi.org/10.1371/JOURNAL.PONE.0256224.

Deutscher Bundestag (2021), "Entwurf Eines Gesetzes Zur Änderung Des Straßenverkehrsgesetzes Und Des Pflichtversicherungsgesetzes–Gesetz Zum Autonomen Fahren", *Circular 19/27439*, https://dserver.bundestag.de/btd/19/274/1927439.pdf (accessed December 16, 2024).

European Commission, (2018), "On the road to automated mobility: An EU strategy for mobility of the future." *COM(2018) 283 Final*, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0283 (accessed December 16, 2024).

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., & Song, D. (2018), "Robust Physical-World Attacks on Deep Learning Visual Classification", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1625–1634, https://doi.org/10.1109/CVPR.2018.00175.

Fernandez Llorca, D., & Gomez Gutierrez, E. (2021), *Trustworthy Autonomous Vehicles*, Publications Office of the European Union, EUR 30942 EN, https://doi.org/10.2760/120385.

Fernández Llorca, David and Gómez, Emilia (2021), *Trustworthy Autonomous Vehicles: Assessment criteria for trustworthy AI in the autonomous driving domain*, Publications Office of the European Union, https://publications.jrc.ec.europa.eu/repository/handle/JRC127051 (accsessed December 16, 2024).

Galassi, M. C., & Lagrange, A. (2020), *New approaches for automated vehicles certification, Part I, Current and upcoming methods for safety assessment*, https://op.europa.eu/en/publication-detail/-/publication/d320cd56-8051-11ea-b94a-01aa75ed71a1/language-en (accessed December 16, 2024).

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021), "Datasheets for datasets", In *Communications of the ACM* (Vol. 64, Issue 12, pp. 86–92), Association for Computing Machinery, https://doi.org/10.1145/3458723.

Gu, T., Dolan-Gavitt, B., & Garg, S. (2017), *BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain*, https://doi.org/10.48550/arXiv.1708.06733.

Gunning, D. (2017), Explainable Artificial Intelligence (XAI), https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf (accessed December 16, 2024).

High-Level Expert Group on Artificial Intelligence, (2019), Ethics Guidelines for Trustworthy AI, European Commission, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed December 16, 2024).

DG RESEARCH (2020), "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)*, https://doi.org/10.2777/966923 (accessed December 16, 2024).

IEC, (2014, July 10), "Railway applications- Urban guided transport management and command/control systems- Part 1: System principles and fundamental concepts", *International Electrotechnical Commission,* https://webstore.iec.ch/publication/6777 (accessed December 16, 2024).

IEEE 7000-2021, (2021), IEEE Standard Model Process for Addressing Ethical Concerns during System Design, https://ieeexplore.ieee.org/document/9536679 (accessed December 16, 2024).

IEEE Computer Society (2022), Towards the engineering of trustworthy AI applications for critical systems, https://www.confiance.ai/wp-content/uploads/2023/09/LivreBlanc-Confiance.ai-Octobre2022-1.pdf (accessed December 16, 2024).

IMO, (2021), Outcome of the Regulatory Scoping Exercise for the Use of Maritime Autonomous Surface Ships (MASS), MSC.1/Circ.1638, https://wwwcdn.imo.org/localresources/en/MediaCentre/PressBriefings/Documents/MSC.1-Circ.1638%20-%20Outcome%20Of%20The%20Regulatory%20Scoping%20ExerciseFor%20The%20Use%20Of%20Maritime%20Autonomous%20Surface%20Ships...%20(Secretariat).pdf (accessed December 16, 2024).

ITF, (2008), *Towards Zero: Ambitious Road Safety Targets and the Safe System Approach*, OECD Publishing, Paris, https://www.itf-oecd.org/towards-zero (accessed December 16, 2024).

ITF, (2016), *Zero road deaths and serious injuries : leading a paradigm shift to a safe system*, OECD Publishing, Paris, https://www.oecd.org/en/publications/2016/10/zero-road-deaths-and-serious-injuries_g1g6e7c8.html (accessed December 16, 2024).

ITF. (2018), *Safer Roads with Automated Vehicles?*, International Transport Forum, OECD Publishing, Paris, https://www.itf-oecd.org/safer-roads-automated-vehicles-0 (accessed December 16, 2024).

ITF, (2019), *Governing Transport in the Algorithmic Age*, International Transport Forum, OECD Publishing, Paris, https://www.itf-oecd.org/governing-transport-algorithmic-age (accessed December 16, 2024).

ITF, (2021), *Artificial Intelligence in Proactive Road Infrastructure Safety Management*, International Transport Forum, OECD Publishing, Paris, https://www.itf-oecd.org/artificial-intelligence-proactive-road-infrastructure-safety-management (accessed December 16, 2024).

ITF, (2022a), *Reporting Mobility Data: Good Governance Principles and Practices*, International Transport Forum, OECD Publishing, Paris, https://www.itf-oecd.org/reporting-mobility-data-governance-principles-practices (accessed December 16, 2024).

ITF, (2022b), *The Safe System Approach in Action*, International Transport Forum, OECD Publishing, Paris, https://www.itf-oecd.org/safe-system-approach-action-experience-based-guide-enhanced-road-safety (accessed December 16, 2024).

ITF, (2023a), *Adapting (to) Automation: Transport Workforce in Transition*, International Transport Forum, OECD Publishing, Paris, https://www.itf-oecd.org/adapting-automation-transport-workforce-transition (accessed December16, 2024).

ITF, (2023b), *Making Automated Vehicles Work for Better Transport Services: Regulating for Impact,* International Transport Forum, OECD Publishing, Paris, https://www.itf-oecd.org/automated-vehicles-better-transport-services (accessed December16, 2024).

ITF, (2023c), *Mix and MaaS: Data Architecture for Mobility as a Service,* International Transport Forum, OECD Publishing, Paris, https://www.itf-oecd.org/mix-and-maas-data-architecture-mobility-service (accessed December 16, 2024).

ITF, (2023d), *Preparing Infrastructure for Automated Vehicles*, International Transport Forum, OECD Publishing, Paris, https://www.itf-oecd.org/preparing-infrastructure-automated-vehicles (accessed December 16, 2024).

Janiesch, C., Zschech, P., & Heinrich, K, (2021), "Machine learning and deep learning", *Electronic Markets*, 31, 685–695, https://doi.org/10.48550/arXiv.2104.05314.

Jones, K. H. (2014), "Engineering Antifragile Systems: A Change In Design Philosophy", *Procedia Computer Science*, 32, 870–875, https://doi.org/10.1016/J.PROCS.2014.05.504.

Khastgir, S. (2023), "Cross-Domain Safety Assurance Framework for Autonomous Transport Systems (Land, air and marine)", Presentation made to the ITF Roundtable on Artificial Intelligence, Machine Learning and Regulation January 26, 2024, https://www.itf-oecd.org/sites/default/files/repositories/session_1_-_siddartha_khastgir.pdf

Knight, J. C. (2002), "Safety critical systems: challenges and directions", *Proceedings of the 24th International Conference on Software Engineering*, 547–550, https://ieeexplore.ieee.org/document/1007998 (accessed December 16, 2024).

Koopman, P., & Widen, W. H. (2024), Breaking the Tyranny of Net Risk Metrics for Automated Vehicle Safety, https://scsc.uk/journal/index.php/scsj/article/view/31 (accessed December 16, 2024).

Krämer, J., Senellart, P., & de Streel, A. (2020), Making data portability more effective for the digital economy, https://cerre.eu/publications/report-making-data-portability-more-effective-digital-economy/ (accessed December 16, 2024).

Laplante, P., Milojicic, D., Serebryakov, S., & Bennett, D. (2020), "Artificial Intelligence and Critical Systems: From Hype to Reality", *Computer*, 53(11), 45–52, https://doi.org/10.1109/MC.2020.3006177.

Law Commission of England and Wales, & Scottish Law Commission, (2022), Automated Vehicles : joint report, https://lawcom.gov.uk/project/automated-vehicles/ (accessed December 16, 2024).

Li, X., Chen, Z., Zhang, J. M., Sarro, F., Zhang, Y., & Liu, X. (2024), "Bias Behind the Wheel: Fairness Analysis of Autonomous Driving Systems", *Transactions on Software Engineering and Methodology (TOSEM)*, 1, 1, 1, 22, https://doi.org/10.48550/arXiv.2308.02935.

Lütge, C., Poszler, F., Acosta, A. J., Danks, D., Gottehrer, G., Mihet-Popa, L., & Naseer, A. (2021), "AI4people: Ethical guidelines for the automotive sector-fundamental requirements and practical

recommendations", *International Journal of Technoethics*, 12(1), 101–125, https://doi.org/10.4018/IJT.20210101.oa2.

Matheu-García, S. N., Hernández-Ramos, J. L., Skarmeta, A. F., & Baldini, G. (2019), "Risk-based automated assessment and testing for the cybersecurity certification and labelling of IoT devices", *Computer Standards and Interfaces*, 62, 64–83, https://doi.org/10.1016/J.CSI.2018.08.003.

MOLIT Ordinace No. 684, (2019), Regulation for Performance and Safety Standards of Motor Vehicle and Vehicle Parts [in Korean: 자동차 및 자동차부품의 성능과 기준에 관한 규칙], https://bit.ly/49IAurJ (accessed December 16, 2024).

Moteff, J., & Parfomak, P. (2004), Critical Infrastructure and Key Assets: Definition and Identification, Congressional Research Service, Order Code RL32631, https://sgp.fas.org/crs/RL32631.pdf (accessed December 16, 2024).

NTC (2024), Automated Driving System Entity certification, Australian National Transport Commission, https://www.ntc.gov.au/sites/default/files/assets/files/ADSE%20certification%20April%202024.pdf (accessed December 16, 2024).

NTSB (2019), Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian Tempe, Arizona March 18, 2018, Accident Report NTSB/HAR-19/03US. National Transportation Safety Board, https://www.ntsb.gov/investigations/accidentreports/reports/har1903.pdf (accessed December 16, 2024).

New, J., & Castro, D. (2018), "How Policymakers Can Foster Algorithmic Accountability", *Information Technology and Innovation Foundation,* https://itif.org/publications/2018/05/21/how-policymakers-can-foster-algorithmic-accountability/ (accessed December 16, 2024).

NHTSA. (2023a), Early Estimate of Motor Vehicle Traffic Fatalities in 2022, US DOT National Highway Traffic Safety Administration Report No. DOT HS 813 428, https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813428 (accessed December 16, 2024).

NHTSA. (2023b), Second Amended Standing General Order 2021-01: Incident Reporting for Automated Driving Systems (ADS) and Level 2 Advanced Driver Assistance Systems (ADAS), https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety#topic-road-self-driving (accessed December 16, 2024).

NIST. (2023), Artificial Intelligence Risk Management Framework (AI RMF 1.0), https://doi.org/10.6028/NIST.AI.100-1.

OECD. (2019), Scoping the OECD AI principles : Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO), Organisation for Economic Cooperation and Development, https://www.oecd-ilibrary.org/science-and-technology/scoping-the-oecd-ai-principles_d62f618a-en (accessed December 16, 2024).

OECD (2022), "OECD Framework for the Classification of Ai Systems", *OECD Digital Economy Papers*, Organisation for Economic Cooperation and Development, https://www.oecd.ai/wips (accessed December 16, 2024).

OECD (2024), OECD AI Incidents Monitor (AIM), OECD.AI Policy Observatory, Organisation for Economic Cooperation and Development, https://oecd.ai/en/incidents (accessed on 16 December, 2024).

OICA (2019), "Proposal for the Future Certification of Automated/Autonomous Driving Systems", *Informal document GRVA-02-09*, International Organization of Motor Vehicle Manufacturers, https://unece.org/DAM/trans/doc/2019/wp29grva/GRVA-02-09e.pdf (accessed on December 16, 2024).

JO (2021), Ordonnance no 2021-443 du 14 avril 2021 relative au régime de responsabilité pénale applicable en cas de circulation d'un véhicule à délégation de conduite et à ses conditions d'utilisation, Journal Officiel, https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000043370894 (accessed December 16, 2024).

Pater, G. (2018), "Challenges and Proposals for Modern Vehicles", *Presentation made to the UNECE meeting of the Working Party on Automated/Autonomous and Connected Vehicles – Introduction (GRVA)*, 25-28 September, 2018, Informal document GRVA-01-40, https://unece.org/DAM/trans/doc/2018/wp29grva/GRVA-01-40.pdf (accessed December 16, 2024).

Quinn Emanuel trial lawyers (2024), Report to the Boards of Directors of Cruise Llc, GM Cruise Holdings LLC, and General Motors Holdings LLC Regarding the October 2, 2023 Accident in San Francisco, https://assets.ctfassets.net/95kuvdv8zn1v/1mb55pLYkkXVn0nXxEXz7w/9fb0e4938a89dc5cc09bf39e86ce5b9c/2024.01.24_Quinn_Emanuel_Report_re_Cruise.pdf (accessed December 16, 2024).

EU AI Act (2024), Regulation 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence, https://eur-lex.europa.eu/eli/reg/2024/1689/oj (accessed December 16, 2024).

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018), Algorithmic Impact Assessments Report: A Practical Framework for Public Agency Accountability, https://ainowinstitute.org/publication/algorithmic-impact-assessments-report-2 (accessed December 16, 2024).

Rudin, C. (2019), "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead", *Nature Machine Intelligence* (Vol. 1, Issue 5, pp. 206–215), Nature Research, https://doi.org/10.1038/s42256-019-0048-x.

SAE International, (2021a), Surface Vehicle Recommended Practice: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, https://www.sae.org/standards/content/j3016_202104/ (accessed December 16, 2024).

Schoettle, B. (2017), Sensor Fusion: A Comparison of Sensing Capabilities of Human Drivers and Highly Automated Vehicles, Sustainable Worldwide Transportation, https://public.websites.umich.edu/~umtriswt/PDF/SWT-2017-12.pdf (accessed December 16, 2024).

Sengar, S.S., Hasan, A.B., Kumar, S., & Carroll, F. (2024), "Generative Artificial Intelligence: A Systematic Review and Applications", *Multimedia Tools and Applications,* https://doi.org/10.1007/s11042-024-20016-1.

Sheikh, H., Prins, C., & Schrijvers, E. (2023), *Artificial Intelligence: Definition and Background,* 15–41, https://doi.org/10.1007/978-3-031-21448-6_2.

Shimada, H., Yamaguchi, A., Takada, H., & Sato, K. (2015), "Implementation and Evaluation of Local Dynamic Map in Safety Driving Systems", *Journal of Transportation Technologies*, 05(02), 102–112, https://doi.org/10.4236/JTTS.2015.52010.

Song, D. (2018), "AI and Security: Lessons, Challenges & Future Directions", *Presentation made to the 1st Workshop on Deep Learning and Security at the 39th IEEE Symposium on Security and Privacy*, 24 May, 2018, https://www.ieee-security.org/TC/SPW2018/DLS/ (accessed December 16, 2024).

Srinivas Acharyulu, P. V., & Seetharamaiah, P. (2015), "A framework for safety automation of safety-critical systems operations", *Safety Science*, 77, 133–142, https://doi.org/10.1016/J.SSCI.2015.03.017.

Stolte, T., Ackermann, S., Graubohm, R., Jatzkowski, I., Klamann, B., Winner, H., & Maurer, M. (2021), "A Taxonomy to Unify Fault Tolerance Regimes for Automotive Systems: Defining Fail-Operational, Fail-Degraded, and Fail-Safe", *IEEE Transactions on Intelligent Vehicles*, 7(2), 251–262, https://doi.org/10.1109/TIV.2021.3129933.

Taeihagh, A., & Lim, H. S. M. (2019), "Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks", *Transport Reviews*, 39(1), 103–128, https://doi.org/10.1080/01441647.2018.1494640.

Taleb, N. N. (2012), *Antifragile: Things That Gain From Disorder,* Random House, ISBN: 1-400-06782-0.

Thomas, S. (2024), "Generative AI And Self-Driving Vehicles: A Potential Future", *Forbes Online – Council Post, Forbes Business Development Council*, https://www.forbes.com/councils/forbesbusinessdevelopmentcouncil/2024/11/22/generative-ai-and-self-driving-vehicles-a-potential-future/ (accessed on 16 December, 2024).

UK DfT. (2022), *Reported road casualties Great Britain, annual report: 2021,* https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-annual-report-2021/reported-road-casualties-great-britain-annual-report-2021 (accessed on 16 December, 2024).

UN Regulation No 157 – Uniform provisions concerning the approval of vehicles with regards to Automated Lane Keeping Systems [2021/389] (OJ L 82 09.03.2021, p. 75, ELI: https://eur-lex.europa.eu/eli/reg/2021/389/oj (accessed December 16, 2024).

van der Waal, S., Stikker, M., Kortlander, M., van Eeden, Q., Demeyer, T., & Bocconi, S. (2020), European Digital Public Spaces, waag technology and society- Online European Public Spaces, https://culturalfoundation.eu/wp-content/uploads/2021/05/Waag-Report-on-European-Digital-Public-Spaces.pdf (accessed December 16, 2024).

World Wide Web Foundation, (2017), Algorithmic accountability: Applying the concept to different country contexts, https://webfoundation.org/docs/2017/07/WF_Algorithms.pdf (accessed 16 December, 2024).

Annex A. List of Roundtable participants

Chair: Markus REINHARDT, German Federal Railway Authority, Germany

Special Guest: Stefan SCHNORR, Federal Ministry for Digital and Transport, Germany

Roundtable participants: affiliations current at the time of the Roundtable.

Gregorio AMEYUGO, CEA List, France

Alice ARMITAGE, University of California, USA

Gianmarco BALDINI, European Commission Joint Research Centre, Belgium

Miriam BUITEN, Centre on Regulation in Europe, Belgium

Camille COMBE, International Transport Forum

Philippe CRIST, International Transport Forum

Mary "Missy" CUMMINGS, George Mason University, USA

Louise DENNIS, University of Manchester, UK

Jagoda EGELAND, International Transport Forum

Gillian GILLETT, Caltrans, USA

Gabriele GRIMM, Federal Ministry for Digital and Transport, Germany

Aida JOAQUIN ACOSTA, Ministry of Transport, Mobility and Urban Agenda, Spain

Stig O. JOHNSEN, SINTEF Digital, Norway

Siddartha KHASTGIR, University of Warwick, UK

Jinwhan KIM, Korea Advanced Institute of Science & Technology, Korea

Young-Tae KIM, International Transport Forum

Changgi LEE, International Transport Forum

Jaehong MIN, Korean Railroad Research Institute, Korea

Mohammad Reza MOUSAVI, King's College London, England

Latifa OUKHELLOU, Université Gustave Eiffel, France

Florent PERRONIN, NAVER LABS Europe, France

Martin RUSS, AustriaTech, Austria

Margriet VAN SCHIJNDEL- DE NOOIJ, Eindhoven AI Systems Institute, Netherlands

William H. WIDEN, University of Miami, USA

Anthony WONG, AGC Legal & Advisory, Australia

International Transport Forum

AI Machine Learning and Regulation: The Case of Automated Vehicles

From road to rail to shipping, recent technology is leading to advances in automated vehicles: driverless cars, trains and boats. Known as “AVs”, they promise improved safety and accessibility. But they are also cause for concern. Potential risks are linked to data quality and representation, the development and verification of AI models, increased vehicle travel, land-use impacts, and deskilling of vehicle operators.

To safely harness the full potential of automated transport, supporting regulation must be able to demonstrate AVs trustworthiness, both in terms of safety and ability to serve the collective good.

This report examines the regulatory approaches to address these challenges – in particular focusing on road vehicles. It provides a common understanding of AI-based automated transport systems and the principles that should form the basis of institutional and regulatory actions to increase safety and social acceptability.

International Transport Forum
2 rue André Pascal
F-75775 Paris Cedex 16
+33 (0)1 73 31 25 00
contact@itf-oecd.org
www.itf-oecd.org

OECD

Published: 02/2025 | Photo credit: Aleksandra Suzi/Shutterstock