PRISSMA Project
Plateforme de Recherche et d'Investissement pour la Sûreté
et la Sécurité de la Mobilité Autonome
04/2021 - 04/2024

[L1.3] WP1 STATE OF THE ART : FINAL-RANGE REPORT

RAPPORT FINAL : ÉTAT DES LIEUX WP1

Main authors : R. Régnier (LNE), J. Girard-Satabin (CEA), A. Kalouguine (LNE), S-S. Ieng (Univ. Eiffel), H-M. Hussain (Airbus protect), J. Fiquet (Airbus protect), J. H-A. Girard-Sabatin (CEA), D. Gruyer (Univ. Eiffel), R. de Sousa (UTAC) and A. PIPERNO (UTAC)

Keywords: Standards, Evaluation, AI systems, Validation requirements, Evaluation protocol, Regulation, Performance, Robustness, Resilience, Traceability, Interpretability, Explainability, Testability

Abstract.

his document follows on from Deliverable 1.2 of which it is an update and aims at making an inventory of the different methodologies to evaluate AI, in order to identify those applicable in the field of autonomous mobility. It will therefore list the studies under development for the evaluation of AI and the actors concerned by the subject at national, European or global level. An analysis of the reference systems in operation in related fields (railways, aeronautics, etc.) or extended to other critical systems with AI (health, defence, etc.) will be carried out, as well as a review of standards and reference systems for the autonomous vehicle (ISO 26262, ISO/PAS SOTIF 21448, etc.).

A large number of acronyms will be used in this document, the meanings of which are listed in Deliverable 8.1.

Résumé.

Ce document fait suite au livrable 1.2 dont il est une mise à jour et a pour but de faire un état des lieux et un recensement des différentes méthodologies pour évaluer l'IA, afin d'identifier celles applicables dans le domaine de la mobilité autonome. Il recensera donc les études en cours de développement pour l'évaluation de l'IA et les acteurs concernés par le sujet au niveau national, européen ou mondial. Une analyse des référentiels en exploitation dans des domaines connexes (ferroviaire, aéronautique, etc.) ou élargis à d'autres systèmes critiques avec de l'IA (santé, défense, etc.) sera effectuée, de même qu'une revue des normes et référentiels pour le véhicule autonome (ISO 26262, ISO/PAS SOTIF 21448, etc.).

Un grand nombre de sigles seront utilisés dans ce document, dont les significations sont répertoriées dans le livrable 8.1.

Contents

List of Figures

1 Introduction

The PRISSMA project aims, in response to the call launched by the Grand Défi of the Innovation Council and the Ministry of Ecological and Solidarity Transition, to generate the prototyping of a platform for the approval of autonomous mobility, addressing in particular, in the context of the safety demonstration, the impacts of the use of Artificial Intelligence techniques (for the evaluation and validation of the safety and security of autonomous mobility). It is requested that the project proposes one or more regulations, reference systems (typologies of manoeuvres, safety rules, definition and elaboration of fields of use, etc.) and adapted methodologies whose relevance will have to be demonstrated through their application to proofs of concept on experiments. Thus, the need to determine what should be tested, as well as the means to be used to carry out the tests is the subject of a specific Work Package, the WP1. The main objectives of this WP are to:

- Estimate the level of criticality of each AI-based function in relation to its function and use domain (e.g. adapt the SDF analysis of ISO 26262 for AI)
- Express demonstrable requirements (in engineering terms, either by ”auditing” during development and validation, or during testing) applicable to AI-based functions and evaluation criteria
- To deduce requirements for the means of testing and auditing to be put in place (media and methods)

This document is the third deliverable of this WP1 and is an update of Deliverable 1.2. It serves as a state of the art for methods and requirements for the evaluation of AI in general and its more specific application to the autonomous vehicle. This document will be complemented by other deliverables produced by other PRISSMA WPs which will address more technical or specific issues:

- Deliverable 2.3 of WP2: state of the art of simulation solution and methodologies for testing;
- Deliverable 3.4 of WP3: state of the art of controlled environment testing (track, test bench);
- Deliverable 4.1 of WP4: identification of field test sites and associated test procedures;
- Deliverable 6.3 of WP6: state of the art of safety risk assessments and audits;
- Deliverable 8.1 of WP8: glossary of the project;
- Deliverable 8.4 of WP8: reference report on the principles and processes as well as the actors' distribution.

2 Normative/industrial review applicable to the autonomous vehicle/shuttle (non-AI specific and AI specific)

2.1 Table of the main standards and related activities

The regulation for autonomous vehicles can be categorized as applying to different areas of autonomous vehicle functioning. We will present existing initiatives separated by areas of focus. In the table below the main standards and related activities in force are listed.

Subject	Functional need	Standardisation-related activities
Testing Approaches	• Describe how tests address functional requirements • Facilitate discussion between parties • Define test apparatus, infrastructure, procedures • Define ODD-specific OEDR tests • Define role of simulation, track testing and on-road testing	• SAE ORAD Verification and Validation Committee • SAE J3018 — Guidelines for Safe On-Road Testing of SAE Level 3, 4, and 5 Prototype Automated Driving Systems • Pegasus/AdaptIVe project • TNO Streetwise methodology • U.S. Army Tank Automotive Research, Development and Engineering Center (TARDEC) guidelines • Department of Defense Unmanned Systems Safety Guide being updated • FHWA Test and Evaluation for Vehicle Platooning • AAMVA — Jurisdictional Guidelines for the Safe Testing and Deployment of Highly Automated Vehicles • FHWA and SAE Cooperative Automation Research Modeling and Analysis (CARMA) program • US DOT V2I research program DSRC Roadside Unit (RSU) Specifications development • UNECE think-tank with the GRVA (Working party on automated/autonomous and connected vehicles), in particular the FRAV (Functional Requirements for Automated Vehicles) and NATM (New Assessment/Test Method for automated driving) groups.

Subject	**Functional need**	**Standardisation-related activities**
Functional Safety	Using verification and validation (V&V) from current standards to ensure a safe vehicle design	• **ISO 26262** — Road Vehicles — Functional Safety • IEC 62508 — Dynamic Test Procedures for Verification and Validation • SAE J3092 — Dynamic Test Procedures for Verification and Validation of Automated Driving Systems • **ISO 21448** (Publicly Available Specification, PAS)— Road vehicles — Safety of the intended functionality (SOTIF) • ISO TS5083 Safety & Cyber-security For Automated Driving (SaFAD). • UN R157: ALKS (Automated Lane Keeping Systems) • ISO/TR 4804 - Safety and cybersecurity for automated driving systems - Design, Verification and Validation • UNECE: Framework document on automated/autonomous vehicles of level 3 and higher[a]. • ISO 19237:2017 Pedestrian detection and collision mitigation systems • ISO 22078:2020 Bicyclist detection and collision mitigation systems • UL 4600: Standard for Safety for the Evaluation of Autonomous Products. A Safety case approach to ensuring autonomous product safety in general, and self-driving cars in particular. • ISO/DPAS 8926 - Road vehicles – functional safety – Use of pre-existing software architectural elements • ISO/TR 9839 - Application of predictive maintenance to hardware with ISO 26262-5 • ISO/TR 9968 - Application to generic rechargeable energy storage systems for new energy vehicle • ISO 39003 : Guidance on ethical considerations relating to safety for autonomous vehicles

[a] https://unece.org/fileadmin/DAM/trans/doc/2020/wp29grva/FDAV_Brochure.pdf

Subject	Functional need	Standardisation-related activities
Data sharing	Provide common set of parameters and interface definitions to enable sharing of scenarios	• ISO/TR 21707:2008: Integrated transport information, management, and control—Data quality in intelligent transport systems (ITS) • Pegasus Open-Simulation Interface • ITS JPO Data Program ADS • Numerous data roundtable and international work on standards harmonization
Scenarios	Provide common rules for the creation of scenarios	• ISO 34501: Scenarios Terms and definitions • ISO 34502: Engineering framework and process of scenario-based safety evaluation • ISO 34503 - Test scenarios for automated driving systems, specification for operational design domain • ISO 34504: Scenario attributes and categorization • ISO 34505 - Scenario evaluation and test case generation
ODD Definition	• Specify the boundaries of the ODD including: road type, lighting, weather, traffic volume, incidents, etc. • Boundaries may be set by vehicle capabilities and/or jurisdictional requirement or other factors.	• American Association of Motor Vehicle Administrators (AAMVA) Jurisdictional Guidelines for the Safe Testing and Deployment of Highly Automated Vehicles • States initiatives : Caltrans, Florida DOT • SAE J3016 — Definitions of ODD • ISO 34503: Taxonomy for operational design domain for automated driving systems

Subject	Functional need	Standardisation-related activities
General Atmospheric Conditions/Road Weather	• Classify various weather conditions and data formats • Identify ODD boundaries • Identify minimal risk condition and transition of control • Define approaches for testing and certification	• Reference model architecture efforts within ISO TC204 WG 1 include provision for road weather (connected vehicle focus • NHTSA Testable Cases Project : SAE J3164 — Taxonomy and Definitions for Terms Related to Automated Driving System Behaviors and Maneuvers for On-Road Motor Vehicles
Functional Architecture	Encourage interoperability and enable system-level innovation and more complex applications to emerge	• SAE On-Road Automated Driving (ORAD) • SAE J3131 — Automated Driving Reference Architecture • IEEE WG2040 — Standard for Connected, Automated and Intelligent Vehicles: Overview and Architecture • IEEE WG2040.1 — Standard for Connected, Automated and Intelligent Vehicles: Taxonomy and Definitions and IEEE WG2040.2 — Standard for Connected, Automated and Intelligent Vehicles: Testing and Verification • Automotive Functional Safety Architecture White paper[a] • Other domains: Robot Operating System (ROS), JAUS, VICTORY, AUTOSAR

[a] https://www.daimler.com/documents/innovation/other/safety-first-for-automated-driv
pdf

Subject	Functional need	Standardisation-related activities
Simulation and Software	• Using simulation to validate VAs • Use of software tools for VAs	• ISO 11010 : simulation model classification • ISO 19364 : Vehicle dynamic simulation and validation • ISO 24089 : Road vehicles, software update engineering • ISO 15497 : Development guidelines for vehicle based software

It's worth noting the emergence of a first standard on the use of AI, ISO 8800, within a brand new working group set up for the occasion (SC32/WG14). However, this draft standard is still in PAS status, which shows the lack of consensus expected and, above all, the limited scope of the first draft. The purpose of this standard is to provide specific guidance to the automotive industry on the use of safety functions based on artificial intelligence/machine learning (AI/ML) in road vehicles. This standard will include:

- A derivation of appropriate safety requirements for AI-based functions,
- Data quality and completeness considerations,
- Architectural measures for failure control and mitigation,
- Tools used to support AI,
- AI verification and validation techniques,
- As well as the evidence needed to support an assurance case for overall system safety.

As such, this standard should define a set of safety principles compatible with the existing approaches currently defined in ISO 26262 (functional safety) and ISO 21448 (safety of intended functionality). It will endeavor to fill the gaps in ISO 26262, particularly with regard to the safety of AI component implementation and supporting processes such as tool qualification. The document will also support the harmonization of concepts already described in the annexes of ISO 21448 and ISO/TR 4804, while complementing with specific guidance the definition of safety-related AI/ML properties and the creation of associated safety proofs throughout the development and deployment lifecycle.

The document should build on the guidance contained in ISO/IEC TR 5469 on safety analysis of AI-based systems (which has just been submitted to ISO for a vote), and will not be limited to specific AI/ML techniques or functions.

Generally speaking, the field of AV standardization has spawned a number of working groups that have taken up the subject, with varying degrees of involvement with AI:

- In France, most of the work underway to regulate automated vehicles is being carried out by the DGITM, through working groups such as GT GAME, GT OQA, GT Cybersécurité, GT Route, GT Scénario and GT Remote intervention.

- At the United Nations, the Working Group on Automated/Autonomous and Connected Vehicles (AGCV) has also begun to take into account the subject of AI for the automotive industry through its various sub-groups and workshops on the subject of AV. These include

  - The "Functional Requirements for Automated Vehicles" (FRAV) group, which has led to in-depth discussions on performance expectations for driver assistance systems, criteria to guide the development of requirements, and methods for determining performance specifications that will most likely have an impact on the expected performance of a vehicle equipped with AI systems.

  - The "New Assessment/Test Method" (NATM) group, an offshoot of the "Validation Method for Automated Driving group" (VMAD), was set up so that the international community could maximize the potential safety benefits of ADS, by providing access to a safety validation framework relevant to the approval of automated vehicles and consistent with the requirements set by FRAV. Most of the concepts established in the NATM framework should be included in the definition of evaluation methodologies when artificial intelligence systems are present in vehicles.

  - The organization of a technical workshop focusing primarily on definitions of artificial intelligence, relevant to the activities of the AVRG, and, schedule permitting, more detailed exploration of the potential role of vehicle regulation(s) and guidance document(s) in relation to artificial intelligence.

  - The Focus Group on AI for Automated Driving (FG-AI4AD), which supported standardization activities for services and applications enabled by AI systems in autonomous and assisted driving. The focus group aimed to create international harmonization on the definition of a minimum performance threshold for these AI systems (such as AI as driver). The FG-AI4AD concluded its activities on September 29, 2022. We will see in section 5.1 more in details on going actions about AI based-vehicle regulation.

- On the European Union side, work consists of the production of regulations, as presented later in this report, with in particular the forthcoming arrival of the AI Act based in part on the past work of the High-Level Expert Group on Artificial Intelligence (AI HLEG), which had published an "Ethical Guide for Trusted Artificial Intelligence" (https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai). The third chapter of this report contains a list of recommendations on the assessments to be made to determine whether the AI system developed, deployed, purchased or used complies with the seven requirements of trustworthy artificial intelligence (AI), as specified in the Ethical Guidelines for Trustworthy AI: human oversight, technical robustness and security, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental well-being, responsibility. A final version of this assessment list was published on July 17, 2020.
- The recent AI regulation proposal takes the same approach. The AI Act proposal defines the first-ever legal framework on AI and a new plan coordinated with Member States (updating a first plan dating from 2018), which outlines the guidelines and investments

needed to strengthen Europe's ambition to become the world leader on the subject of AI. This proposal was recently voted by the Council and has been taken to the European Parliament for a vote in 2023. Its adoption should come into force, and will be the subject of chapter 6 of this document. In parallel, the Joint Research Centre (JRC) has also begun researching the subject, organizing a first two-day exploratory workshop on the topic of AI in automobiles in March 2022 (https://publications.jrc.ec.europa.eu/repository/handle/JRC130621). The aim of this workshop was to determine what is still needed to progress towards explainable, robust and equitable AI in automated and autonomous vehicles, by listing the challenges and opportunities in terms of safety and security.

2.2 Focus on some standards

Modern autonomous vehicles are subject to many safety requirements. Two important aspects of safety are functional safety and safety of the intended functionality (SOTIF). They are covered respectively by ISO 26262 and SOTIF ISO PAS 21448. Both of these standards define safety as the absence of unreasonable risks.

The relevant standard for a given risk source can be defined as follows:

- if the risk is related to internal causes (malfunctioning behaviour), it is covered by ISO 26262
- if the risk has external causes without misuse or with foreseeable misuse, it is covered by SOTIF
- if the risk has external causes with voluntary offensive action (cyber-attack, luring, etc.), it is covered by cybersecurity standards (ISO / SAE 21434 and SAE J3061).

An approach to integration of functional safety and SOTIF requirements based on functional safety lifecycle is presented in [KG19].

2.2.1 ISO 26262 Road Vehicles - Functional Safety

ISO 26262 is an international standard for functional safety of electrical and/or electronic (E/E) systems in production automobiles (2011). It addresses possible hazards caused by malfunctioning behaviour of E/E safety-related systems, including interaction of these systems. ISO 26262-3 specifies a Hazard Analysis and Risk Assessment to determine vehicle level hazards. This evaluates the potential risks due to malfunctioning behaviour of the item and enables the definition of top-level safety requirements, i.e. the safety goals, necessary to mitigate the risks. Modifications to the standard have been underway for years to cover more precisely the autonomous functions of the system and the development of this standard should eventually cover AI applications.

ISO 26262 provides a complete safety lifecycle (covering the design, development, production and operation of the system). It covers functional safety aspects of the entire lifecycle and provides necessary safety requirements for each aspect of the system.

The current (2018) version of the standard consists of 12 parts, most of which are normative:

1. Vocabulary - this part specifies a glossary of terms used in all parts of the standard
2. Management of functional safety - this part defines overall organizational safety management measures.

3. Concept phase - this part defines the functional safety requirements pertaining to the early development of the system, including hazard analysis and risk assessment.
4. Product development at the system level - this part defines the functional safety processes in the complete ”V cycle” of development of the system
5. Product development at the hardware level - this part defines the functional safety processes in the small ”V cycle” of hardware development
6. Product development at the software level - this part defines the functional safety processes in the small ”V cycle” of software development
7. Production, operation, service and decommissioning - this part defines the functional safety processes for the final part of the system lifecycle, from production to decommissioning.
8. Supporting Processes - this part defines functional safety objectives for transverse supporting processes, such as change management, review and testing of product, or documentation.
9. Automotive Safety Integrity Level (ASIL)-oriented and safety-oriented analysis - this part considers risks adjusted for their relative likelihoods, classifying hazardous events by level of severity, likelihood and controllability. It is transverse to the complete lifecycle.
10. Guidelines on ISO 26262 - this part provides guidelines to the use of ISO 26262.
11. Guidelines on application of ISO 26262 to semiconductors
12. Adaptation of ISO 26262 for motorcycles.

In [AI20], an implementation is proposed for ISO 26262 by using:

- Functional Safety Audits
- Safety Analysis (SaAn)
- Hazard Analysis and Risk Assessment (HARA)
- Safety Concept (SaCo)
- Safety validation (SaCa)
- Audit/Assessment

Functional Safety Audits or Audits with assessment are used for all generic aspects of the standard, namely

- Part 2 ”Management of Functional Safety”
- Part 3 ”Concept Phase”
- Part 4 ”Product development at the system level”
- Part 5 ”Product development at the hardware level”
- Part 6 ”Product development at the software level”
- Part 8 ”Supporting processes”

These audits must be performed by a competent independent authority and be finalised before the corresponding phase is launched. The auditing authority must provide a report evaluating the safety plan and its implementation.

Safety Analysis is used in the product development at hardware level for evaluation of the hardware architectural metrics (section 5-8 of the standard) and for evaluation of safety goal violations due to random hardware failures (section 5-9). It is also used for sections 9-8 and 9-9, ”Analysis of dependent failures” and ”Safety analysis”.

Hazard analysis and risk assessment is used for the eponymous section 3-6.

Safety Concept is used for sections

- 3-7 ”Functional safety concept”
- 4-6 ”Technical safety concept”
- 5-6 ”Specification of hardware safety requirements”
- 5-7 ”Hardware design”
- 6-6 ”Specifications of software safety requirements”
- 6-7 ”Software architectural design”
- 6-8 ”Software unit design and implementation”
- 8-6 ”Specification and management of safety requirements”
- 8-14 ”Proven in use argument”
- 9-5 ”Requirements decomposition with respect to ASIL tailoring”
- 9-6 ”Criteria for coexistence of elements”

Finally, Safety Validation is used for the following sections of the standard:

- 4-7 ”System and item validation and testing”
- 4-8 ”Safety validation”
- 5-10 ”Hardware integration and verification”
- 6-9 ”Software unit verification”
- 6-10 ”Software integration and verification”
- 6-11 ”Testing of the embedded software”
- 8-12 ”Qualification of software components
- 8-13 ”Qualification of hardware elements”

2.2.2 ISO/PAS 21148 SOTIF

SOTIF (PAS 21448 : 2019) is to be distinguished from other types of safety, as it is defined for the intended functionality in a system free from faults and in normal conditions (or with foreseeable misuse). In these conditions, the absence of unreasonable risk due to potentially hazardous behaviours related to the intended functionality or to performance limitations is defined as the safety of the intended functionality (SOTIF).

The object of this standard is to provide guidance on the applicable design, verification and validation measures needed to achieve the SOTIF.

- The intrinsic safety of the electrical / electronic components (E / E system) remains the task of Functional Safety according to ISO 26262
- SOTIF standard covers misuse as does the European Statement of Principles on human-machine interface
- Cybersecurity (external attacks) are covered by the standards ISO / SAE 21434 and SAE J3061
- Communication with road infrastructure and other vehicles (Car2x) should be considered by ISO 20077 Road Vehicles - Extended vehicle (ExVe)

2.2.3 New regulations on VA

The UN R157 ALKS (Automated Lane Keeping Systems) regulation covers the development of SAE Level 3 autonomous vehicles. It includes general requirements regarding the system safety and the failsafe response, and lays down requirements on how the driving task shall be safely handed over from the ALKS to the driver. It also includes requirements on the HMI to prevent misunderstanding and misuse by the driver.

This regulation has been applicable starting since Q3 2021[1].

The first regulations for highly automated vehicles (levels 3 and 4) are beginning to appear:

- EU-ADS (2022-1426) [2] for European regulations on automated driving systems
- NAVUR-BAUT [3] for French regulations on the use of autonomous urban shuttles.

Specific work on these two standards was carried out as part of PRISSMA deliverable 1.5.

3 Normative review and regulation in other specific AI fields

3.1 Aeronautics

3.1.1 EASA

The European Union Aviation Safety Agency published in February 2020 a report explaining its human-centric approach to AI in aviation under the name of EASA Artificial Intelligence Roadmap 1.0 [EAS20b]. It discusses the types of AI and their possible roles in aviation, associated safety and ethical concerns. It finally gives its roadmap to progress on these questions. To achieve its goals, the agency concluded several Innovation Partnership Contracts. The first project in 2020 between EASA and Deadlean produced Concepts of Design Assurance for Neural Networks (CoDANN) [EAS20a]. It aimed to identify ways to gain confidence in products embedded with Machine learning based system particularly neural networks. They could then identify enablers to support these system introductions in aviation. They could also come up with an adaptation of system engineering V cycle into a W shaped cycle of ML applications

[1] https://unece.org/transport/documents/2021/03/standards/un-regulation-no-157-automated-lane-keeping-systems-alks

[2] European Commission, "Regulation (eu) 2022/1426 - laying down rules for the application of regulation (eu) 2019/2144 of the european parliament and the council as regards uniform procedures and technical specifications for the type-approval of the automated driving system (ads) of fully automated vehicles - 5 august 2022," Brussels, 2022.

[3] MINISTERE DE LA TRANSITION ECOLOGIQUE - TRANSPORTS. Arrêté du xxx définissant les conditions d'homologation, d'exploitation et de circulation des navettes urbaines équipées d'un système de conduite automatisé.

concluding that their use was feasible. Their second report Concepts of Design Assurance for Neural Networks (CoDANN) II [EAS21a] has been published in May 2021. This report went further in discussing trustworthiness of AI building blocks and refined the concept of Learning Assurance. Its main axes are: “

- implementation and inference parts of the W-shaped process (hardware, software and system aspects), encompassing development and deployment aspects;
- definition and role of explainability;
- details on the system safety assessment process, concluding discussions on integrating neural networks into complex systems and their evaluation in safety assessments.”

EASA has already used these outputs, following its calendar to produce guidelines concerning AI in aviation, and published First usable guidance for Level 1 machine learning applications - Issue 01 [EAS21b] in April 2021.

3.1.2 EAAI HLG

European Aviation High Level Group on AI (EAAI HLG) - a high level group composed of key representatives from all aviation sectors (airlines, airports, Air Navigation Service Providers, manufacturers, EU bodies, military and staff associations) have published in March 2020 first Fly AI report [HLG20]. Experts from EUROCONTROL, the European Commission, ACI-Europe, Airbus, ASD, CANSO, Heathrow Airport, Honeywell, IATA, IFATCA, IFATSEA, the SESAR JU, Thales, as well as our military partners EDA and NATO were involved in its writing.

This report identifies domains where the use of AI could bring a significant change and sets out an action plan to accelerate the development of AI in European Union and Air Traffic Management.

3.2 Rail

Shift2Rail is a European rail initiative for focused Research and Innovation to accelerate the integration of new technologies including AI into innovative rail solutions. It contains 5 Innovation Programs (IP).

3.2.1 IP 2 Technologies for sustainable and attractive European Rail Freight [SHI19]

3.2.1.1 Automatic Train Operation (ATO)

In its IP2 working on Technologies for sustainable and attractive European Rail Freight includes Work Package 04 Automatic Train Operation (ATO) up to Grade of Automation 4 (GoA4) for mainlines. It aims to put a solution on market before 2024. It aims to take benefit of experiences in urban applications and some existing mainlines to generalize it to Mainlines: High Speed Line, Low Traffic/Regional Lines, Urban/Suburban, and Freight) and to extend these applications from fenced systems (urban rail) to open systems (single EU Railway Area). This WP has already published:

- D4.1 - ATO OVER ETCS GOA2 SPECIFICATION [XRA21a] in 2021 which focuses on GoA2 starting from inputs from Ten-T 3rd call (ATO over ETCS - Technical Interoperability Requirement for GoA2); the operation concepts updated according to the results of the European NGTC project and existing standard IEC 62290-2; (incl. IEC 62267).

- D4.3 - AOE_GOA3_4_PRELIMINARY_SPECIFICATION [XRA21b] in 2021 which performs the feasibility study and preliminary design for GoA3 and GoA4 solutions.

3.2.1.2 ASTRAIL

In the framework of ASTRAIL, WP3 Automatic driving technologies for railways has produced 3 deliverables. This WP worked to identify automatic driving technologies in other fields that could be applicable for railway sector ATO. Therefore it performed:

- D3.1 STATE OF THE ART OF AUTOMATED DRIVING TECHNOLOGIES [AST19a], 2019 in the automotive sector and other application fields considering technologies that are already on the market or in the development phase;
- D3.2 AUTOMATIC TRAIN OPERATIONS IMPLEMENTATION OPERATION CHARACTERISTICS AND TECHNOLOGIES FOR THE RA [AST19b], published in 2019 identifies the basic implementation characteristics of the automotive sector that are compliant for the implementation in the railway sector; the operation conditions that are required for the different grade of automation in ATO (e.g. driverless or unattended operations); and selects the automated driving technologies suitable for the rail sector.

3.2.2 IP5 :Technologies for sustainable and attractive European Rail Freight

3.2.2.1 ATO GoA 2 for existing fleet

This segment of IP5 is testing the interchangeability of GoA2 modules and will contribute to the standardization committees and ERA TSI CCS. WP1 Automated Train Operation within the Automated Rail Cargo Consortium published in 2018 D1.3 AUTOMATED BRAKE TEST [CON18] that provided concepts and requirements for fully and partially automated brake tests with a concern to reduce time and cost associated to the execution of brake tests on freight trains.//

3.2.2.2 Obstacle Detection System (ODS)

ODS is a cornerstone in Automatic Train Operation. This solution will allow to identify collision threats in the surrounding environment by using different sensing technologies and hence reduce the risk of collision and improve safety of the operations. SMART, Shift2Rail and Horizon2020 worked on jointly published a series of deliverables for an ODS system that could perfume up to GoA4. The WP1 focused on requirements and specification of the system, WP2 developed a prototype for obstacle detection, WP3 developed software algorithms for obstacle detection on railway tracks, and WP7 evaluated the system. Their findings are listed below:

- D1.1 Obstacle detection system requirement specification [SMA19a]
- D2.1 Report on selected sensors for multi-sensory system for obstacle detection [SMA19b]
- D2.2 Design of the passive vibration isolation system [SMA19c]
- D2.3 Report on sub-systems conformance testing [SMA19d]
- D2.4 Report on functional testing of fully integrated multi-sensor obstacle detection system [SMA19e]
- D3.1 Report on algorithms for 2D image processing [SMA19f]

- D3.2 Report on smart data fusion and distance calculation [SMA19g]
- D3.3 Report on real-time algorithm implementation and performance evaluation [SMA19h]
- D7.1 Report on evaluation of developed smart technologies [SMA19i]

3.3 Health

AI in healthcare is an emerging field of research, with several directions of application:

- Health services management
- Predictive medicine
- Patient data and diagnostics
- Clinical decision-making
- Robotics for surgical intervention
- Robotics for patient care

AI in healthcare involves many ethical and safety concerns, and current regulation is still insufficient for the full development of these systems. We will therefore provide a short summary of existing regulations for various AI applications in healthcare.

3.3.1 IEEE - ECPAIS (Ethics Certification Program for Autonomous and Intelligent Systems)

The ECPAIS[4] program is meant to create specifications for certification and scoring processes that promote transparency, accountability, and reduction in algorithmic bias for autonomous and intelligent systems. ECPAIS intends to offer a process and defines a series of labels by which organisations can seek certifications for their processes around the A/IS products, systems, and services they provide.

3.3.2 HAS Classification of digital solutions used for medical or paramedical care

HAS is the French High Health Authority. In 2019-2021, they have created a table[5] of classification for software-based systems used in healthcare. The aim of this table is to provide the Social Security and health authorities with an objective criterion to evaluate safety and criticity of these systems.

The rough categorization of these systems is presented below:

[4] https://standards.ieee.org/industry-connections/ecpais.html

[5] https://www.has-sante.fr/jcms/p_3212876/fr/un-nouvel-outil-pour-l-evaluation-des-dis

	Description	Number of detailed categories	Personalisation	Autonomy level
Level A	Patient or care professional support in care or healthcare trajectory/administrative optimisation without direct impact on patients health.	1 category	Limited	No
Level B	General information to the user, not personalised for living conditions, hygiene or any existing health condition. This level also includes information support or training tools.	1 category	No	No
Level C	Life support, assistance in prevention, screening, diagnostic, observation, surveillance or treatment of a pathology or other health condition, including handicaps. The digital solution does not have autonomy in therapeutic decisions.	8 categories	Yes	No
Level D	Autonomous decision-making after autonomous data analysis and diagnostic. The treatment is adjusted autonomously without human intervention.	1 category	Yes	Yes

3.3.3 IEEE 2801/2802 - Recommended Practice for the Quality Management of Datasets for Medical Artificial Intelligence

These two standards (in development, expected to be released May 2022)[6][7] identify the best practices for establishing a quality management system for datasets used for artificial intelligence medical devices. The recommended practice covers a full cycle of dataset management, focusing on data collection, transfer, utilization, storage, maintenance and update. The document recommends a list of critical factors that impact the quality of datasets, such as data sources, data quality, annotation, privacy protection, staff qualification/training/evaluation, tools, equipment, environment, process control and documentation.

3.3.4 ETSI standard: eHEALTH Data recording requirements for eHealth

The purpose of this technical specification is to specify the normative framework for ensuring that the events/transactions related to a patient are recorded accurately by identifiable entities (devices or health professionals) and made available with minimum delay to any other health professional. The normative framework is intended to be adopted by all groups contributing to eHealth including CYBER, smartM2M, smartBAN.

ETSI is contributing to the development of several technical standards for eHEALTH[8].

[6] https://standards.ieee.org/project/2801.html

[7] https://standards.ieee.org/project/2802.html

[8] https://www.etsi.org/committee/1396-ehealth

3.3.5 IEEE P7000 - Draft Model Process for Addressing Ethical Concerns During System Design

This set of standards[9] aims to define ethical values for the design of complex systems.

P7002 - Standard for Data Privacy Process

This standard[10] specifies how to manage privacy issues for systems or software that collect personal data. It will do so by defining requirements that cover corporate data collection policies and quality assurance. It also includes a use case and data model for organizations developing applications involving personal information. The standard will help designers, by providing ways to identify and measure privacy controls in their systems, utilizing privacy impact assessments.

P7003 - Standard for Algorithmic Bias Considerations

This standard[11] describes specific methodologies to help users certify how they worked to address and eliminate issues of negative bias in creating their algorithms–where "negative bias" infers the usage of overly subjective or uniformed data sets, or information known to be inconsistent with legislation concerning certain protected characteristics; or with instances of bias against groups not necessarily protected explicitly by legislation, but otherwise diminishing stakeholder or user well-being, and for which there are good reasons to be considered inappropriate.

P7010 - Standard for Well-being Metrics for Ethical Artificial Intelligence and Autonomous Systems

This standard[12] establishes well-being metrics related to human factors directly affected by intelligent and autonomous systems, and it establishes a baseline for the types of objective and subjective data these systems should analyze and include (in their programming and functioning) to proactively increase human well-being.

3.4 Defense

In the area of Defense, AI has many applications, ranging from active protection (for example the Iron Dome air defense system in Israel) to fully autonomous lethal weapons (such as combat drones). The issue of regulating AI for Defense is deeply political and diplomatic. Currently, there is no regulation[13] specific to autonomous weapons systems, but many actions are ongoing.

Autonomous Military Vehicles Autonomous vehicles for military use can be ground-based, on- or underwater, or airborne. Overall, they face the same challeges as ordinary unmanned vehicles, plus some challenges exclusive to the military domain. In particular, the difficulty associated with the navigation task is not comparable between a civil road and combat conditions[14].

Autonomous Weapons Lethal Autonomous Weapons (LAWs) are a major concern in AI regulation[15]. Although automatically triggered weapons have existed for centuries - land mines and naval mines, or simple traps can kill without human intervention - the introduction of AI is raising new ethical concerns.

[9] https://standards.ieee.org/standard/7000-2021.html
[10] https://standards.ieee.org/project/7002.html
[11] https://standards.ieee.org/project/7003.html
[12] https://standards.ieee.org/standard/7010-2020.html
[13] https://article36.org/updates/treaty-structure-leaflet/
[14] https://www.lawfareblog.com/challenges-us-military-designing-and-deploying-self-driv
[15] http://www.fcas-forum.eu/en/articles/responsible-use-of-artificial-intelligence-in-f

Indeed, autonomous weapon systems rise major concerns that need clarification in the following years :

- Reducing potential bias is of the utmost importance
- Accelerated decision-making may lead to escalation of combat situations
- Measures must be taken to ensure meaningful human control and supervision over unmanned weapon systems
- Automated decision-making may interfere with political and diplomatic considerations.

LAWs can be categorized into Autonomous Defensive Systems and Autonomous Offensive Systems.

3.4.1 Defining the level of autonomy of autonomous weapons

It is important to properly define the term autonomy when discussing autonomous weapon systems. The following table summarises the definition proposed by [AK16].

Level of automation	**Definition**
Human-operated	perform all of the information and decision processes only with a human input
Partially automatic	perform parts of the information and decision processes without any human input in a predictable environment
Human-supervised automatic	perform all or parts of the information and decision processes under human oversight in a predictable environment
Automatic	perform all of the information and decision processes without any human input in a predictable environment
Partially autonomous	perform parts of the information and decision processes without any human input in an unpredictable environment
Human-supervized autonomous	perform all or parts of the information and decision processes under human oversight in an unpredictable environment
Autonomous	perform all of the information and decision processes without any human input in an unpredictable environment

3.4.2 Autonomous weapon systems - European Parliament position

In this European Parliament Resolution[16] dated 12 September 2018, the EU is calling for a common position on lethal autonomous weapon systems that ensures meaningful human control over the critical functions of weapon systems.

[16] https://www.europarl.europa.eu/doceo/document/TA-8-2018-0341_EN.html

3.4.3 European Parliament Resolution

European Parliament published a resolution on 1 January 2021 on artificial intelligence: questions of interpretation and application of international law in so far as the EU is affected in the areas of civil and military uses and of state authority outside the scope of criminal justice.

In this Resolution[17], the Parliament stresses that AI used in military context must be subject to meaningful human control, and that its development must be done with respect of international law (including humanitarian law) and of fundamental rights, values and freedoms. The Parliament also insists on the need for an EU-wide strategy against LAWS and a ban on so-called 'killer robots', and considers that LAWs should only be used as last resort.

3.5 Standards on AI

Numerous IA standards other than ISO PAS 8800 have recently been released or are in production. The first standards to come out take a long, hard look at the subject. These include ISO/IEC 22989:2022, which finally establishes a terminology for AI, and in particular defines a concrete framework for the definition of AI, which was previously very nebulous and sector-specific. Other standards mainly concern data or AI risk management, such as ISO/IEC 23053:2022 or ISO/IEC 23894:2023 (risk management recommendations).

here's a small selection:

- Big data
  - ISO/IEC 20546 :2019 Bigdata :Overview and vocabulary
  - ISO/IEC 20547-1:2020 Bigdata reference architecture: methodological framework and application process
  - ISO/IEC 20547-2: 2018 Big data reference architecture: practical cases and derived requirements
  - ISO/IEC 20547-3: 2020 Bigdata reference architecture: Reference architecture
  - ISO/IEC 20547-5: 2018 Big data reference architecture: roadmap for standards
  - ISO/IEC 24668: 2022 Process management framework for bigdata analysis
- High-level concept
  - ISO/IEC 22989: 2022 AI concepts and terminology
  - ISO/IEC 23053: 2022 Framework for artificial intelligence systems using ML
  - ISO/IEC 23894: 2023 Risk management recommendations
  - ISO/IEC TR 24030: 2021 AI - Case studies. Accompanying ISO/IEC TR 24372:2021 gives the state of the art of AI in relation to case studies.
  - ISO/IEC 38507: 2022 Governance implications of organizations' use of AI

[17] https://www.europarl.europa.eu/doceo/document/TA-9-2021-0009_EN.html#ref_1_7

- Evaluation
  - ISO/IEC TS 4213: 2022 Evaluation of machine learning classification performance
  - ISO/IEC 24028-1: 2021 Robustness evaluation of neural networks - Overview
  - ISO/IEC 24028-2: 2023 Robustness evaluation of neural networks - Methodology for using formal methods
- Ethics: ISO/IEC TR 24368:2022 High-level standard on ethical and societal aspects of AI

The most important point is also the appearance of the first AI evaluation standards, ISO 4213 and ISO 24028, but these are still poorly developed and not very applicable to fields such as transport (the application of formal methods is still very difficult in this sector, and still at the research stage).

But the bulk of the standardization work is still underway at world level, with the ISO/IEC JTC 1/SC 42 working group, the CN Afnor IA for France and the CEN-CENELEC Focus group for Europe. The most important standard at this level is ISO/IEC 42001, which was voted on December 2023 and cover management systems for the use of AI. This standard shake up the various existing certifications in the field of AI, as it provide the first normative framework for the implementation of specific processes for the use of AI. Its main contribution is therefore to provide a guide for the use and integration of AI systems within a structure. It also attempt to answer questions concerning the autonomy of AI systems (notably autonomous vehicles), transparency and explicability, and trust (as a complement to the ISO/IEC TR 24028 technical report). The annexes to this standard also look important: - Annex A deal with questions on trusted AI - Annex B cover a guide to AI implementation, as well as data management issues - Appendix C deal with sources of risk and organizational objectives relating to AI - Annex D deal with the sectorial approach (which include the field of autonomous transport) and the integration of this generic AI standard with existing sectorial standards, as well as certification aspects and an approach to third-party conformity assessment of this new standard.

4 AI evaluation and validation requirements

Here are the key questions to address in the quantification and validation of AI and deep learning methods applied to the domain of autonomous systems :

- Specify DNNs and understand their dynamics w.r.t changes in the input signal.
- Design datasets and create training sets that cover the whole input ODD and the functionality specifications in a robust manner.
- Define metrics that evaluate datasets for autonomous systems.
- Quantify uncertainties associated with AI systems to stochastically characterise the ODD of a DNN.

4.1 AI Functionalities

4.1.1 Sensor/perception

The awareness of the surrounding environment is the decisive safety-critical step for a connected and automated vehicle (CAV). A CAV uses multiple on-board sensors. The most popular

are cameras and LiDARs (Light Detection and Ranging). A LiDAR emits thousands of light pulses every second and creates a 3D image of the surrounding with more accurate depth information as compared to cameras. However, cameras play a vital role in scenic understanding compared to LiDARs, such as better performance in poor weather conditions, colours detection, or interpreting traffic signs. There are many perception tasks associated with understanding the local environment such as objects detection, localisation, and lane analysis. Objects detection, in particular, plays a vital role in determining an object's location and classifying it correctly as pedestrian, car or other kind of road users and it is one of the challenging tasks in the self-driving research area. This specific detection task is also the basic function for a good semantic understanding of the road scene then allowing the machine to make adequate driving decisions. In addition, it is not enough to detect objects well, it must also be done in real time in order to be usable in an automated vehicle. With the recent advances in the Deep Neural Networks field, it is reasonable to hope that the scenes will be effectively interpreted by DNN models. Popular DNNs for the perception task includes YOLO (You Only Look Once): YOLOV3 [RF18], YOLOV4 [BWL20] and YOLOV5 (https://github.com/ultralytics/yolov5), Mask-RCNN [Abd17] for the 2D images and PointNet++ [QYSG17], VoxelNet [ZT17] or Complex-YOLO [SMAG18] for 3D objects detection on point clouds. Despite the impressive results obtained on the datasets, it is important to evaluate the performance of such a perception system and to know if it is sufficiently reliable for safety-critical decision. To ensure the reliability of such a system, two types of approaches can be found. The first is the evaluation of the DNN system before its actual use in vehicles [PNdS20]. The second is to provide a real-time mean of managing the uncertainty of the results while the system is in use. In the following paragraph, we provide a comprehensive survey of the DNN based perception system evaluation involving uncertainty management in DNN for AV.

4.1.1.1 Evaluation of DNN based perception systems

Evaluating object detection algorithms is already a widespread practice in the machine learning community. Many popular competitions bear witness to this: PASCAL VOC Challenge, COCO, ImageNet Object Detection Challenge or Waymo open dataset challenge. In 2019, the ride sharing company Lyft organised a competition: Lyft 3D Object Detection for Autonomous Vehicles. In most of these competitions the same evaluation process and metrics are used. The principle is based on the comparison between the DNN's outputs and the ground truth objects. As we can see, the evaluation material consists of the data to be analysed and the ground-truth data that can be obtained manually or using a state-of-the-art interactive, where professional human annotators iteratively correct the output of a segmentation neural network as it is explained for the object segmentation in image. The evaluation are based the component of the so-called confusion matrix. Each column of the matrix represents the instances in an actual class while each row represents the instances in a predicted class. From the confusion matrix, one derives the precision that is the measure of the accurate predictions the DNN model has made and the recall that is the measure of how well the DNN model has predicted the true positive. The figure 1 provides the scheme of the matrix and the derived measures.

Using the recall and precision that are derived from the confusion matrix, we can plot the precision vs recall curve (PRC). For a given DNN model, the precision will be high when the False Positive (FP) rate is low but on the other side, many Positives (P) can be missed. The missed Positives are False Negative (FN) that are not taken into account in precision metric. A high FN rate leads to a low Recall. A DNN model is considered accurate if it finds all the ground-truth

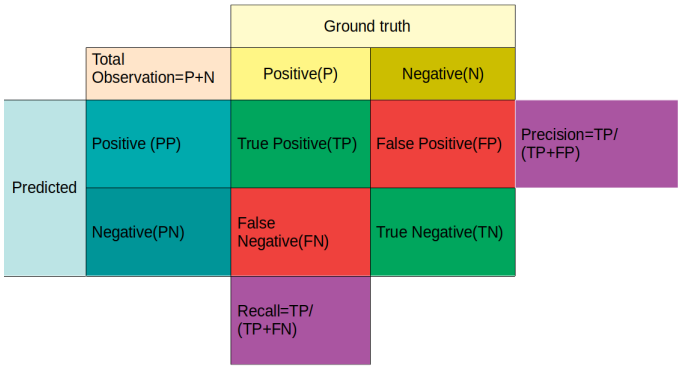

Figure 1: The confusion matrix and two derived metrics: precision and recall

objects (high Recall) while classifying all relevant object classes (high precision). Hence, an accurate object detection DNN should have high Precision and high Recall by keeping a trade-off between these metrics. The Area under the PRC (AUC) should thus be high, indicating high Precision and high Recall. However, in the practice, the PRC follows a zig-zag pattern. For this reason, the best way to sum up the PRC is to use the so-called Average Precision metric that is an approximation of AUC. Two approximations are used in the competitions. For more details about the AP approximations, one can read the presentation by J. Hui in https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173. To evaluate a classifier that is trained with N classes, the Mean Average Positive (mAP) metric is used. It is simply the average of AP over all the specified classes. For object detection DNNs, the outputs are usually bounding boxes (YOLO, RCNN, fast and faster RCNN). Thus, the ground truth are also bounding boxes and labels. The comparison between the outputs and the ground truth is the Intersection over Union (IoU) between the bounding boxes. However, evaluating AV perception system by vehicles makers is more than simply evaluating DNNs in competitions. The evaluation must take into account road environment, weather conditions, driving conditions, visibility conditions among others. There are two approaches to acquire data to train and to evaluate AV perception system:

- The first one is to acquire data in real road environment as Waymo's vehicles but it is very costly and the ground truth is not easy to measure.
- The second is to use on board sensors simulation.

The second approach is easier to carry out because the simulator can provide different scenarios and environment conditions are under control leading to very reliable perception system evaluation. Many simulators are available today: pro-Sivic (https://www.esi-group.com/) [GGV+10], Carla (https://carla.org/) [DRC+17], dSPACE (https://www.dspace.com), Ansys VRXperience (https://www.ansys.com) or Vector (https://www.vector.com) among others.

4.1.1.2 DNN's uncertainty

The evaluation of the perception systems as described in the previous paragraph is necessary but it is not sufficient to ensure the safety of such a system. This is because an automated vehicle travels in a constantly changing road environment and it is impossible to evaluate the system

in all traffic conditions. Furthermore, although these DNNs provide impressive results on the datasets, the questions arises whether they are reliable enough for safety decision. In Su et al. [SVS17], it is shown that the output of a DNN can be altered by adding a small perturbation such as an additional pixel in the input and recent research works (see for instance [GTA+21], [ASSR20], [LPB17], [CCKL19], [KG17]) show that the question is important and that DNN's uncertainty quantification is critical for AI-based perception systems to be exploited in high risk applications. As it is shown in the survey [GTA+21], there are multiple sources of uncertainties and one can classify them into epistemic uncertainty or aleatory uncertainty [HW21]. Therefore methods in estimating uncertainty in DNN is an important field. The methods for estimating the uncertainty can be split in four kind of approaches based on the nature (Stochastic or Deterministic) of the DNN and on the number (single DNN or multiple DNNs):

- Single deterministic approaches (SDA) [MG18] provide the prediction based on one single forward pass within a deterministic network. The uncertainty quantification is derived by using additional methods.
- Bayesian approaches (BA) [KG17] cover all Stochastic DNNs.
- Ensemble methods (EM) [LPB17] combine the predictions of several different deterministic networks at inference.
- Test-time augmentation methods (TTAM) [SBBG20] give the prediction based on one deterministic network but augment the input image by performing random modifications to the input image at the test time in order to generate several predictions that are used to evaluate the confidence for the prediction.

These different approaches have their advantages and their drawbacks. But depending on the application, some approaches are more suitable than others. The table 1 below has been taken from the article [GTA+21] and it shows how many networks are involved in the methods, the computational effort for the DNN training and for the inference and the memory consumption.

Methods	SDA	BA	EM	TTAM
Need to change the DNN	No	Yes	No	No
number of trained networks	1	1	several	1
Training computational effort	Low	High	High	Low
Memory consumption during training	Low	Low	High	Low
Number of input per prediction	1	1	1	several
Forward pass per prediction	1	several	several	several
Evaluated modes	single	single	multiple	single
Computational effort during inference	Low	High	High	High
Memory consumption during inference	Low	Low	High	Low

Table 1: Description of four DNN's uncertainty estimation approaches

4.1.2 Decision making

In the automotive industry, the planning and decision making can be divided into two subtasks, global route/long term planning and local path planning sometimes accompanied by a

third classification with intermediate planning. The long term planning is responsible for finding the route on the road network from origin to the final destination and is mainly based on conventional methods of graph analysis, so for the purely AI part we will be more interested in the decision making due to local path planning. While these decisions were often due to algorithms as graph-based planners, sampling-based planners, interpolating curve planners and numerical optimization approaches, deep learning-based and reinforcement learning based local planners have recently emerged [LCW17][DBP17], but they are not widely used in real-world systems yet. Some important issues have to be addressed first: lack of hard-coded safety measures and validation methods, problems of generalization, need for more labelled data.

Incorporating safety in Reinforcement Learning (DRL) for decision making: Deploying an autonomous vehicles and drones in real environments after training directly could be dangerous. Different approaches to incorporate safety into high level decision making algorithms for autonomous system. A comprehensive survey on safe reinforcement learning can be found in [GF15] for interested readers.

For imitation learning based systems, Safe DAgger [ZC17] introduces a safety policy that learns to predict the error made by a primary policy trained initially with the supervised learning approach, without querying a reference policy. An additional safe policy takes both the partial observation of a state and a primary policy as inputs, and returns a binary label indicating whether the primary policy is likely to deviate from a reference policy without querying it.

Authors of [SSSS16] addressed safety in multi-agent Reinforcement Learning for Autonomous Driving, and try to integrate the unexpected behaviour of other drivers or pedestrians, without being too defensive, so that normal traffic flow is achieved. While hard constraints are maintained to guarantee the safety of driving, the problem is decomposed into a composition of policies to enable comfort driving and trajectory planning.

The deep reinforcement learning algorithms for control such as DDPG and safety based control are combined in [XWZL16], including artificial potential field method that is widely used for robot path planning. Using TORCS environment, the Deep Deterministic Policy Gradient (DDPG) algorithm is applied first for learning a driving policy in a stable and familiar environment, then policy network and safety-based control are combined to avoid collisions. It was found that combination of DRL and safety-based control performs well in most scenarios. In order to enable DRL to escape local optima, speed up the training process and avoid danger conditions or accidents, Survival-Oriented Reinforcement Learning (SORL) model is proposed in [YMZ+17], where survival is favored over maximizing total reward through modeling the autonomous driving problem as a constrained MDP and introducing Negative-Avoidance Function to learn from previous failure. The SORL model was found to be not sensitive to reward function and can use different DRL algorithms like DDPG.

In fact validation of decision making is mainly done by expert judgement in relation to simulated test or track/road tests. Relatively few non-human dependent methods are used, while in robotic field several approaches to path planning, reinforcement learning, etc. have been developed and are commonly used for these tasks.

Other approaches are those developed in particular within the 3SA project:

- Optimization-based reference comparison: in the 3SA project [ADR20], they consider a decision making module as a black-box and try to determine a reference which represents the 'right decision', if it exists. An optimization-based reference model is created for the

control function. This model allows each scene in the environment to be mapped to the desired decision regardless of the black-box decision under test. The black-box and the reference model are run on several critical scenarios and a comparison has been made between them. In output, an assessment of decision making is performed along with systematic criticality characterization of targeted scenarios.

- Monitoring method: The objective of property monitoring is to confront the executions of a module with properties expressing expected constraints. This requires the use of a very expressive language on temporal and numerical relations as well as a monitor synthesis engine (or oracle) with the capacity to operate in line and in embedded mode for synchronous and continuous semantics. This kind of method can intervene at several levels, to capture abstract textual requirements and monitor them on low-level simulated executions (Simulink + HIL + SIL ...), on real executions (in drive), or in the simulation phase. The monitoring engine can be used to monitor on-line and in parallel a large number of behavioural requirements written in its input language. On the 3SA project, the monitoring is performed by a CEA application called ARTIMON.
- Model-based testing: the use of this type of test covers in particular processes for analysing the conformity of execution traces with respect to models that can describe security properties. This approach makes it possible to compare the execution traces of a system under test with models to analyse whether the trace reveals a non-conformity with the control flow/data flow model (including real time constraints if necessary).

4.1.3 Control

Control in an Autonomous transport generally has to do with the vehicle motions such as lane changing, lane keeping, and car following. These actions are categorised under longitudinal control (speed regulation, brake) and lateral control (i.e. automatic steering to follow track reference) [AKF10]. So the motion controller is responsible for computing the longitudinal and lateral steering commands of the vehicle. Learning algorithms are used either as part of Learning Controllers, within the motion control module or as complete End2End Control Systems which directly map sensory data to steering commands [SG20].

In complex environments, such as driving, traditional controllers (fixed parameters) cannot foresee every possible situation, so AI controllers can give more flexibility and anticipate situations that cannot be modelled before deployment. Learning techniques are commonly used to learn a dynamics model which in turn improves an a priori system model in iterative learning control (ILC, [ZYW17]) or model predictive control (MPC, [SLB16]).

End2End learning control is defined as a direct mapping from sensory data to control commands and can be formulated as a back-propagation algorithm scaled up to complex models. And since Darpa Autonomous VEhicle (DAVE, [UMC06]) managed to drive through an obstacle-filled road in 2006, they have widely spread (encouraged notably by NVIDIA®), as part of the PilotNet architecture. Another approach to design End2End driving systems is DRL. This is mainly performed in simulation, where an autonomous agent can safely explore different driving strategies.

To carry out an evaluation of the Tesla AutoPilot system, [LF17] proposed an End2End CNN framework. It is designed to determine differences between AutoPilot and its own output trained over 420 hour of real road driving, taking into consideration edge cases. The comparison between the two controllers revealed an accuracy of 90.4% in detecting differences between both systems.

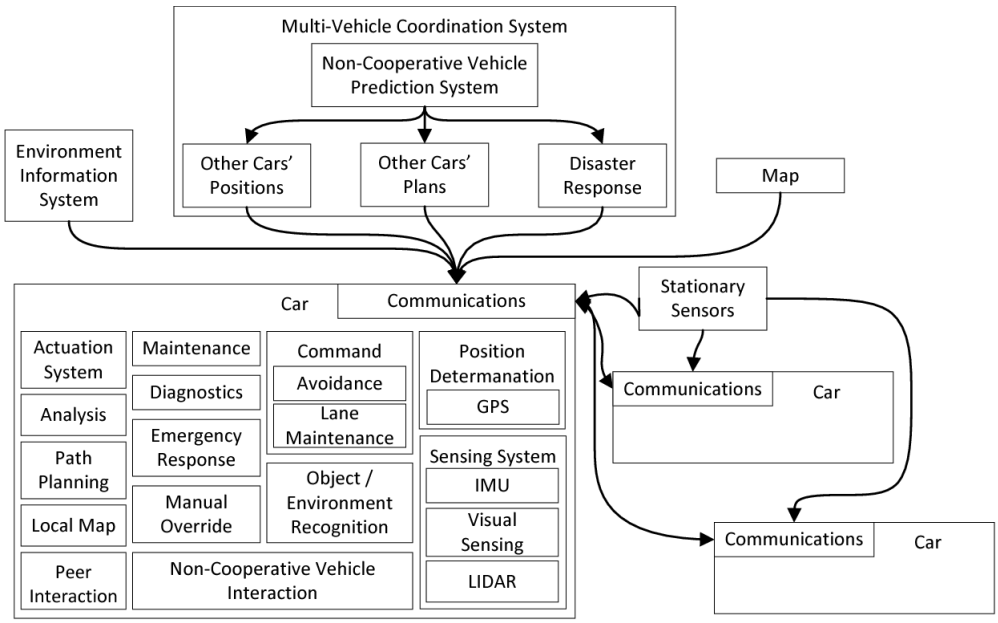

Figure 1. System Overview [39].

Figure 2: Automated Driving system with its constituent systems by [SAA+17].

4.1.4 Supervision and System of systems

We are currently witnessing a major change in mobility thanks to technological developments in telecommunications (the Internet of Things) and to the progress in the field of Artificial Intelligence, which can now process a lot of information, a bit like what humans are capable of. Advances in automation and connectivity have enabled vehicles to exchange data and improve the perception of the environment. However, these technological improvements alone will not be able to solve the problems of congestion, transport emissions and road casualties. It is important to proceed properly with a transition step between historical mobility (as we know so far) and the future transport system we imagine, which is likely to include automated and connected vehicles. And in this prospect, it is important to work on better exchanges between vehicles and other more vulnerable users and road managers.

4.1.4.1 Collaborative systems

The transport system is made up of infrastructure, vehicles, vulnerable users and road managers (among others). These components can be seen as independent agents or systems that interact with each other. With the proliferation of sensors in vehicles and infrastructures, it is even reasonable to consider these devices as agents. This system of systems (SoSs) approach is a recent approach that can accelerate the development of automated transport as it is described in [ATC19]. In [SAA+17], SoSs models have been proposed to describe the autonomous vehicle with every hardware devices and sensors but also to describe the transport system mentioned above. The figure 2 is the SoSs scheme for automated driving system including ego-vehicle, other vehicles, environment information system, map, connected infrastructure systems and every devices and driving assistance systems that are in the ego-vehicle. Many systems essential for automated vehicles have been proposed using the SoSs approach such as location and control systems [KYY+15, SG15]. There are also algorithms for sensors on the infrastructure that must consider all users and their interactions. In [SoS], the authors use data from inductive loops to

train a network based on the ensemble learning to predict short-term traffic flow. Using this approach, it is possible to improve traffic management with sensors installed in the infrastructure and processing systems located in the infrastructure to help smooth traffic and facilitate collaboration between different users. Intelligent transport management system is not recent and many published research works already proposed different algorithms using classical inductive loops or other in-pavement sensors [DSSCY14, SS15]. Today, communication between these agents is facilitated thanks to the technologies mentioned above but also to the development and standardisation of V2X communication messages[CAM19, CPM18, MAP20], and the progress made in artificial intelligence allows more efficient real-time road management [Get19] and collaborative driving [SSSPP21].

4.1.4.2 Supervision

Supervising the transport system is an approach that can exploit the SoSs to ensure the robustness of the automated driving system at two levels:

- at the level of on-board perception or control systems in VAs, in order to ensure the proper functioning of the algorithms in degraded conditions (weather forecast, electronic devices),
- at road level, the AI and V2X technologies allow a more efficient active management, as it is reported in [Get19], and smart connected perception systems in the infrastructure may enable an easier interaction and collaboration between users including the traffic and road managers.

The supervision in the road level can be an effective support to automated driving like the SAE J3016 Levels of Driving Automation (https://www.sae.org/blog/sae-j3016-update). The Infrastructure Support for Automated Driving (ISAD) can also be classified into five levels as it is shown in Figure 3 (https://www.inframix.eu/). The Levels C to E represent the current infrastructure used in the world. The infrastructure of level B can only be equipped with sensors, a computer and Road Side Unit (RSU) that broadcast information in real time so that connected vehicles are able to improve the perception of their environment. This may be particularly interesting when confronted to poor visibility intersection or roundabout, as it is presented in [SSSPP21] with data fusion algorithms and control algorithms operating inside of the connected vehicle. The infrastructure of the level A must process all the data and actively participate to the real-time traffic flow management by broadcasting orders to plan the trajectories of all vehicles.

4.1.5 Fusion

In the broad sense, data fusion can be seen as the use, through mathematical operators, of information from several sources (sensors or outputs of specific processing) to improve decision-making. This definition is an extension of that given by [Blo96]. Another very interesting definition is that proposed by Herbert Simon, in the 1970s, who proposes a scheme of decision making that is general enough to be recognized as a canonical model of decision. This definition lists, in a simple manner, the steps from data sources to decision and therefore integrates the upstream and downstream steps framing the information fusion process. This model is broken down into four stages:

	Level	Name	Description	Digital information provided to Avs			
				digital map with static road signs	VMS, warnings, incidents, weather	Microscopic traffic situation	Guidance: speed, gap, lane advice
Conventional infrastructure	E	Conventional infrastructure/ no AV support	Conventional infracstructure without digital information. Avs need to recognise road geometry, road signs, and road markings.				
	D	Static digital information/Map support	Digital map data is available with static road information (signs and markings). Map data could be complemented by physical reference points (landsmarks). Traffic lights, short term road works and VMS need to be recognized by Avs.	☑			
Digital infrastructure	C	Dynamic digital information	All dynamic and static infrastructure information is available in digital form and can be provides to Avs	☑	☑		
	B	Cooperative perception	Infrastructure is able to perceive microscopic traffic situations and to provide this data to Avs in real time.	☑	☑	☑	
	A	Cooperative driving	Based on the real-time information on vehicle movements, the infrastructure is able to guide Avs (groups of vehicles or single vehicle) in order to optimize the overall traffic flow.	☑	☑	☑	☑

Figure 3: ISAD (Infrastructure Support for Automated Driving): Levels of information services from infrastructure needed for the deployment of Automated Vehicle (H2020 INFRAMIX project)

- The first Stage deals with the diagnosis of the problem and the exploration-recognition of the conditions in which it arises: this is the ”intelligence” stage (in the military sense of intelligence). In the development of complex applications requiring one or more data fusion steps, this stage corresponds to the determination of the framework and the modelling of the knowledge and data to be manipulated. In a probabilistic framework, this will correspond to defining the triplet made up of the space of the tests, of the set of events and of a space of measurement. For instance, in belief theory, this corresponds to the construction of the framework of discernment, of the frame of definition and to the generation of the initial mass set.
- The second stage is about the designing and formulating of the possible ways offered to solve the problem. This stage consists in implementing one or more operators using the knowledge built from the information sources and using an appropriate modelling (stage 1) in order to produce more complete, more extensive, more reliable, enriched, or enhanced information (change of representation space), in order to allow judicious, reliable and relevant decision-making. We can associate this phase, either with the use of probabilistic fusion operators like the Bayes formula, or the Kalman filter, or with credible operators with the Dempster-Shafer combination or with the generalized combination proposed in [Gru99].
- The third stage is responsible for choosing a particular mode of action among the possible actions: this is the selection and decision stage. In fact, this step consists of making a choice and taking one or more decisions based on the refined knowledge of stage 2 and the objectives set. In belief theory, this step consists of using decision measures (measure of credibility, plausibility, commonality), in possibility theory, it corresponds to using measures of possibility or necessity.
- And the final stage addresses the evaluation of the solution provisionally retained as satisfactory. This stage, called the assessment stage, can lead to the reactivation of one of the three previous stages or, on the contrary, to the validation of the solution recognized

as satisfactory. This feedback is clearly a training stage (in AI-based approaches).

As we can see, many notions and mathematical concepts are used in these 2 definition tests. In order to better understand the underlying semantics and especially in order to be able to understand the complexity behind this simple word: FUSION, it is necessary to develop these concepts by following the first 3 steps of the definition of the decision proposed previously. It is obvious that the 2 definitions proposed in this part give a good working framework but are far from being exhaustive and that the reality is more complex. In addition, despite rigorous mathematical frameworks, each application is often a specific case requiring an adapted and dedicated data fusion architecture.

4.2 Evaluation requirements with associated metrics

Whatever data to be manipulated and merged, their imperfections (accuracy, uncertainty, and reliability) always exist and they will always be linked to the characteristics of the sensor(s) providing them. Despite this lack of measurement accuracy, it is often necessary to make a decision. For this decision to be rational, coherent, and exploitable, it is necessary to consider these imperfections. This therefore requires the use of tools/indicators/metrics to measure and model these notions of measurement imperfection in order to be able to deal with them as best as possible. Moreover it is essential to be able to propagate these quantities the as long as possible in order to have a final decision that is as realistic as possible and best fits the problem. This propagation of imperfections on the data will also give the possibility of constructing new information quantifying the confidence on a partial decision or the operation of a part of an algorithm. In this part, we will speak without distinction of sensors, observations and sources of information because all these designations have a very similar connotation: they provide information about an object or an object attribute belonging to a set of possibilities. A number of difficulties in data fusion arise from problems that are often independent of the theoretical framework used. These problems are often related to knowledge modelling. It is therefore important to be able to answer the following questions:

- How to model **inaccuracy** and **uncertainty** on data provided by an information source?
- How to better characterise the **reliability** of an information source?
- How to represent a **lack of information**?

In fact, more generically, the question we need to face is: which modelling should be used to better represent knowledge or a lack of knowledge? It is only by answering these questions that it will be possible to design a method of information fusion and management. Indeed, the choice of a data fusion method is strongly dependent on knowledge modelling. Either we have reliable and accurate information and in this case the data fusion mechanism to use will be very simple, or we have unreliable and imperfect data and we will need a fusion method that takes these imperfections into account. There are mainly three types of imperfections affecting the data and the information it contains, namely: uncertainty, inaccuracy and incompleteness.

Uncertainty is a notion relating to the veracity of information, and which characterises its conformity with reality. This uncertainty is mainly due to ignorance and lack of information. There are two main sources of the nature of uncertainty: bias or systematic error and randomness which is the component varying in an unpredictable way. More generally, we can define several types of errors:

- **Natural error**. The conditions of an experiment always vary a little, so they influence the measurable quantity.
- **Accidental error or statistical error**. This error arises from the fact that to obtain the set of possible values and their probabilities of occurrence, one would have to take an infinite number of measurements. The fact of being limited to a finite number of measurements introduces an additional uncertainty which is therefore the statistical error.
- **Equipment error**. Between the experimenter and the measurable object, there is an apparatus which inevitably alters the initial distribution. It disturbs it in two ways.
  - **Systematic error**. The device shifts the mean of the distribution, the most frequent case is a calibration problem. It is very difficult to find and correct these types of errors.
  - **Expansion of distribution**. This uncertainty has the same origin as the initial uncertainties (natural errors).

Uncertainty gives a qualitative representation of the imperfection in a given context.

Imprecision is described as the degree of approximation with which a desired result is achieved, in the form of a deviation between the desired value and the observed value. The imprecision therefore relates to the content of the information, it concerns a quantitative lack of knowledge about a measure. The inaccuracy is directly related to the measurements or the operating state of a source of information. Imprecision gives a quantitative representation of the imperfection in a given context.

Reliability characterises the validity of the operating ranges of an information source (physical or logical sensor). For a measurement, reliability is the probability that the measured parameter is correctly measured according to the nominal operation of the measure and is not an outlier. Depending on the reliability on the source, an attenuation mechanism can be applied to the less reliable measurements. This weakening will consist, for example, in increasing the standard deviation in the case of a Gaussian distribution modelling a data. Reliability is a qualitative representation of the imperfection on the data source or on the outcome of a data merge process.

In summary, we can say that uncertainty about a hypothesis characterises doubt about the veracity of the latter. So a probability coefficient is a degree of uncertainty and a probability distribution is a distribution of uncertainties. Imprecision is a very different concept. A source of information (sender) speaking about the world is imprecise if it leads the receiver of that information to have uncertainties about this world.

4.2.1 Performance

In the automotive field, there are two approaches to performance evaluation. The one based purely on safety and the one based on the intrinsic performance of the AI.

Several metrics are used to assume the safety of the autonomous vehicle. The standard one is the one proposed by MobilEye : the “Responsibility-Sensitive Safety (RSS)[18] proposal

[18] https://static.mobileye.com/website/corporate/rss/rss_on_nhtsa.pdf

[SSSS17] and its NHSTA implementation [SSSS18] complement the classic use of "Time-to-X" metrics (Time to Brake, Time to Collision [MMM01], [JW14], etc.) and the avoidance metrics ([Jan05]) .

For SAE 1 and 2 automation level, Virginia Tech wrote a technical report describing standardized performance testing procedures [BHDN19]. Researchers conducted tests of various driving situations on controlled and real roads. Example of test scenarios include:

1. Autonomous vehicle had to keep track of a lead car in a curve road section;
2. Autonomous vehicle had to correctly accept cut-ins and cut-outs in highway lanes;
3. Autonomous vehicle had to correctly detect and avoid an obstacle on the road using evasion manoeuvres.

Experimented drivers and on-boarded experimenters were to take note of the vehicle behaviour, observations were translated into a grade rating the overall vehicle performance over expected performance on the given scenarios. Note that evaluations were conducted at the system level: the reasons for failure to follow the evaluation scenarios were not actively investigated.

With regard to AI algorithms linked to classification, in particular those used in the perception of the environment, the performance evaluation methods were presented in the section 4.1.1.

4.2.2 Robustness

Deep neural networks can have a very brittle behaviour against perturbations on their input. For instance, a change on one pixel [SVK17] can result in a different prediction. More generally, their sensitivity to input corruption [FGCC19] such as Gaussian noise is well known. See 4.2.2 for a schematic representation of adversarial examples crafting. As they are a building block in

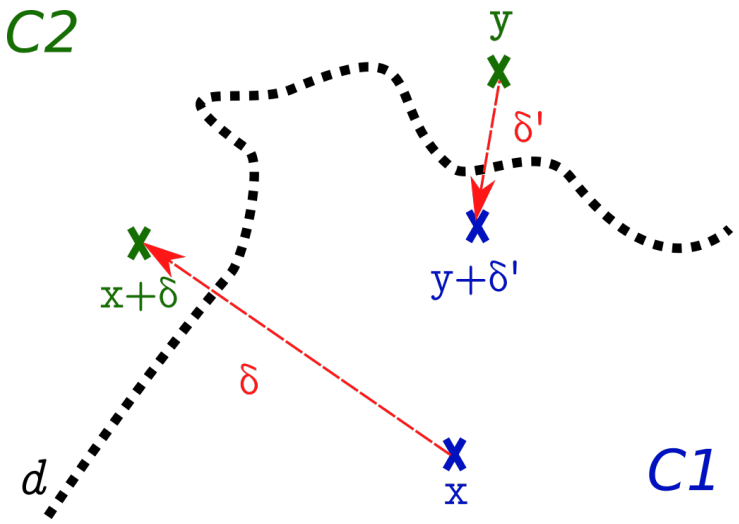

Figure 4: Adversarial example crafting for binary classification

perceptive subsystems (processing images, videos and other sensed environmental information), their inability to correctly perform their mission may lead to dire consequences: not detecting a pedestrian, not detecting a lane border...

To formally assess the robustness of machine learning algorithms, three components are necessary:

1. a formal specification of the behaviour to be verified - here, robustness against perturbations
2. a system/software to study: which part of the autonomous vehicle will be verified
3. a set of tools to formally verify

4.2.2.1 Specifying robustness

A common robustness assessment is the ability of the neural network to keep its classification around a vicinity of a given point. Vicinity is usually defined in terms of l_1, l_2 or l_∞ norm.

Robustness is both a safety and security property as it models respectively stability w.r.t. possible inaccuracy or noise specific to an application domain (e.g., sensitivity of measurements), and resistance w.r.t. adversarial attacks. Robustness properties may be seen as domain-agnostic since no particular knowledge is required to assess the stability of a neural network for a given distance metric. When available, some knowledge of the application domain is nonetheless valuable to determine an appropriate degree of perturbation to verify robustness.

If there exists a precise enough semantic of the function approximated by the neural network, then it is possible to further verify the robustness against perturbation. For instance, in the ACAS-Xu [MJ16] benchmark that implements a collision avoidance system, properties are formalized to verify the correct output.

4.2.2.2 Programs and subsystems to verify

With classical program, it is possible to embed directly into the code some part of the expected input specification. For instance, a programmer can write a conditional to prevent the use of certain values, or throw and catch exceptions. In a sense, the program control flow can be somewhat adapted to the specification.

Machine learning programs lack such mechanism. As such, particular care must be given into defining the machine learning program's expected input and output. This reasoning applies to defining the scope of the robustness assessment process as well.

It is possible to design neural networks to be more robust and/or easier to verify. Previous state-of-the-art approach revolved around adversarial training [MMS+17], but recent approaches use "robust training" [CKD+22]: a combination of formal methods and classical neural network inference to guarantee the robustness of the network around a predetermined neighborhood.

4.2.2.3 Tools for robustness verification

An important issue concerning robustness verification comes from the fact that current robustness verification methods only focus on "local" properties (around a given point). Depending on the targeted robustness properties, robustness verification will not be enough.

Indeed, there are two main cases:

1. inputs have a semantic meaning, usually in low dimensional inputs (tabular); it is possible to write an explicit formal specification to check,

2. inputs are perceptual (image, videos); in which case it is only possible to formulate properties robustness centered around a specific input.

There is a need to use specific solvers for robustness verification, since most of solvers are not scalable (for now). Examples of such specialized solvers are Marabou [KHI+19], PyRAT[19] and ERAN [RCB+21]. Properties to check are of the reachability form (eg. "for any bounded perturbation around this input, the output should not change"). In that context, ACAS-Xu provides an interesting benchmark on the specification of an Aircraft Collision Avoidance System [MJ16].

The current state-of-the-art on perception inputs focus on different directions:

- adversarial examples defence [MMS+17]
- noise robustness [RBBV20]

4.2.2.4 Tools for formal specification of neural network systems

Although the framework of robustness verification is limited, it is actually possible to use it in different contexts (for instance, in the case of explanaibility [MSI22]). The input languages of the provers tend to be limited. Furthermore, the provers usually focus on one single neural network at the time. Recent work present the perspective of neuro-symbolic verification [XKN22]. The overall approach is to propose (semi-)formal languages that include the result of neural network computations into a (semi-)formal specification. Examples of tools that [DKA+22] [GSAB+22]

4.2.3 Resilience

Authors demonstrate the case of in-production ML models that are used in the domain of API-based-service, where partial feature vectors as inputs and include confidence values with the models predictions [TZJ+16]. They demonstrate simple, efficient attacks that extract target ML models with near-perfect fidelity for popular model classes including logistic regression, neural networks, and decision trees. They demonstrate results on services such as BigML and Amazon Webservices.

The domain of privacy preserving machine learning aims at preserving the identity of objects by ensuring that the features extracted by a deep learning model does not allow generelization to other tasks besides its original domain of operation (for example counting people as primary task, back door task being identification or classification). Authors in [BS20] describe such back doors in modern ML models that could be used by potential attackers.

Additional references can be found in the surveys [UM21] and [LAL+19].

4.2.4 Traceability

Any programmable part of a high-risk system must comply at least with the recommendations of IEC 61508-3 with regard to the traceability of its design i.e to have:

- a forward traceability between the software safety requirements specification and software design,
- a forward traceability between the software design specification and the module and integration test specifications,

[19] https://pyrat-analyzer.com/

- a forward traceability between the system and software design requirements for hardware/software integration and the hardware/software integration test specifications,
- a bidirectional traceability between the software safety requirements specification and the software safety validation plan,
- a bidirectional traceability between the software safety requirements specification and the software modification plan (including rechecking and second validation) in a modification phase,
- a bidirectional traceability between the software design specification and the software verification (including data verification),
- a forward traceability between the requirements of the functional safety assessment and the plan for software functional safety assessment.

In the LNE's AI certification standard[20] called "Processes for design, development, evaluation and maintenance of artificial intelligences", a lot of recommendations for the traceability are requested, these requests are often adaptations of the classic requests formulated in most processes or management audits such as the 9001.

- The identities (or unique identifiers), roles and responsibilities of the persons involved in the key phase of design or conception (database creation, annotation,...) must be documented.
- Method of archiving (with unique identifier information) the successive versions of AI functionality, models and data must be implemented. This archiving of versions must allow a precise traceability of these versions and access to previous versions.
- The evaluation protocol must allow the detection and traceability of cases of non-repeatability of measurements and non-reproducibility of experiments.
- This mechanism must allow the detection and traceability of degradation and drift in the performance of the deployed AI functionality.

It is also required by the future European regulation for High-risk AI to keep logs of events when it is operating.

4.2.5 Interpretability and explicability

A common criticism about AI-based systems is their lack of explainability and transparency. Internal working of the algorithm results from an indirect optimization process, hence variables and control flows have no direct link with human intention. In "classical" programming, modules, interfaces, function definition and documentation are the results of human intention - one of this intention is to be understood by other humans. In machine learning programming, a program is the result of a composition of mathematical operations aiming to minimize a certain objective function. Without any penalty towards "explainability", this process has no reason to produce understandable and transparent software. Note that the wording "black-box" does not necessarily mean that there is no direct access to the program's internal. It is better understood as the opacity of the program's inner working against an external observer.

This lack of interpretability is a major obstacle to wide adoption. Indeed, trusting a system without understanding its its way of working and thinking is more akin to a religious faith

[20] https://www.lne.fr/fr/service/certification/certification-processus-ia

than an engineering process. Not being able to identify the reasons of a malfunction is making debugging and correction much harder. An AI system going against a human expert opinion will not be trusted until it provides understandable, fact-based, tamper-proof and refutable arguments along with its decision. The notion of ”interpretability” or ”explainability” is fuzzy at best (see for instance [Lip16]).

For tabular data, there exists a framework of contrasting local attribute importance [MT20]. Shapley values underline the features that contributed the most to a network's output, in contrast to a given reference. For ML purposes, the range R ”of wealth” could be any real number, namely the output of a regression model, the maximum probability estimated or the individual likelihood in a classification task, anomaly score for outliers detection, and so on. It could also be a discrete number, for instance the predicted rank in a ranking problem, a one-loss function for binary or multi-class classification which means that the reward equals 1 if two instances are in the same class and 0 otherwise.

Post-hoc explainability [RSG16] is a common approach for image inputs models. The key idea if this approach is to mask some parts of the input with black pixels and see how prediction changes. Coloring input images wrt gradient values to provide ”most activated areas”. Limitations: may not work on images that are too far from the input distribution.

A recent line of work focuses on models that are ”explainable by design” [CLT+19, Rud18]. The model architecture aims to expose parts of its reasoning, following the framework of case-based reasoning. At each inference, the program will compare the features of a new input image to relevant features (prototypical images) that were learned during training. Armed with this knowledge, it is possible to use the prototypes as nodes of a soft decision tree, which exposes how the presence (or absence) of a prototype influence the final decision [NvBS20]. Contrary to post-hoc explanations, those explanations have the potential to explain the global decision process of the model, hence providing ”global” explanations.

4.2.6 Testability

A hypothesis (output of a AI algorithm here) is testable if there is a possibility of deciding whether it is true or false based on experimentation by anyone. This allows to determine whether a theory can be supported or refuted by data. However, the interpretation of experimental data may be also inconclusive or uncertain. The problem in the context of the autonomous vehicle is therefore to be able to define whether an output of an algorithm is good or bad and therefore to define what a good output is. In the case of algorithms for which labelled data can be obtained (typically a classification algorithm), the testability issue does not arise because we have references against which to compare the predictions of the system. Obtaining such references can sometimes be hard, as labelling new data is a costly process if done by experts, and raise some ethical concerns if done on microworking platforms. It is more difficult when dealing with control or decision-making algorithms where references other than expert opinion do not always exist. Often in these cases, the reference can become quite binary in the sense that decision-making is considered good if there is no accident, which greatly limits interpretation. This is why some recent work such as [ADR20], has tried to define fine-grained references to compare decision making or control hypotheses in order to improve testability.

To assure the testability of a system, quality management systems (defined in ISO 9001:2015) usually require full documentation of the procedures used in a test.

4.3 Scenario generation

This aspect is covered by the state of the art deliverable of WP2 of the project but one of the most critical issues straddling the technical work-packages remains to extract critical scenarios from real world data as well as generating failure cases are part of testing an autonomous driving system [GHF+21].

5 Ongoing actions (French, European, international) for AI approval

5.1 Automated vehicle

Ongoing automotive actions are described in more details in deliverable 8.4 but here are some non-exhaustive references to this area:

- In France, most of the actions in progress around the regulation for automated vehicles are done within the framework led by the DGITM through several working groups as GT GAME, GT OQA, GT Cybersecurity, GT Route, GT Scenario or GT Remote Intervention. Periodically, the DGITM brings together the various actors of these groups in a meeting where they have the chance to report on discussions underway in these groups. This session, is also an opportunity to review the regulatory context under construction outside France (EU, UNECE). (see deliverable 8.4 for more detailed information);
- In Europe, the High-Level Expert Group on Artificial Intelligence (AI HLEG), set up by the European Commission, published the Ethics Guidelines for Trustworthy Artificial Intelligence. The third chapter of those Guidelines contained an Assessment List to help assess whether the AI system that is being developed, deployed, procured or used, adheres to the seven requirements of Trustworthy Artificial Intelligence (AI), as specified in the Ethics Guidelines for Trustworthy AI:
  - Human Agency and Oversight;
  - Technical Robustness and Safety;
  - Privacy and Data Governance;
  - Transparency;
  - Diversity, Non-discrimination and Fairness;
  - Societal and Environmental Well-being;
  - Accountability;

  A final version of this Assessment List was published in July 17th 2020. The recent proposal (published on April 21st) for a new framework called the AI Act quote the same approach. The AI act proposal defines the first-ever legal framework on AI and a new Coordinated Plan with Member States (updating a first plan dating from 2018), which describes the guidelines and investments needed to strengthen Europe's ambition to become the world leader on the AI topic. This proposal has recently been voted by the Council and should be brought to the Parliament for vote during the first quarter of 2023.

  In parallel, the Joint Research Centre (JRC) has also started to conduct its research on the topic with a first 2-day exploratory workshop on the subject of AI in the automobile

in March 2022. The goal of this workshop was to find out what is still needed to move towards explainable, robust and fair AI in automated and autonomous vehicles by listing the challenges and opportunities for safety and security.

- On the United Nations side, the Working Party on Automated/Autonomous and Connected Vehicles (GRVA) has also started to take into account the topic of AI for the automotive industry through its different subgroups and various workshops. We can, for example quote:

  - The Functional Requirements for Automated Vehicles (FRAV) that has held extensive discussions regarding expectations for ADS performance, criteria to guide the development of requirements, and methods for determining performance specifications that will, most likely, have consequences on the performances expected from a vehicle equipped with AI systems;

  - The New Assessment/Test Method (NATM) that has been produced by the Validation Method for Automated Driving group (VMAD), In order for the international community to maximize the potential safety benefits of ADS, by giving access to a safety validation framework relevant to the automated vehicles approval and consistent with the requirements set by the FRAV. Most of the concepts established inside the NATM are to be included in the definition of the evaluation methodologies when AI systems are present in the vehicles;

  - The organisation of a technical workshop focusing primarily on definitions for Artificial Intelligence, relevant for GRVA activities, and, if possible i.e. if time is available, exploring more in detail the potential role of vehicle regulation(s) and guidance document(s) with regard to AI.

  - The Focus Group on AI for Automated Driving (FG-AI4AD) that supported standardization activities for services and applications enabled by AI systems in autonomous and assisted driving. The focus group aimed to create international harmonisation on the definition of a minimal performance threshold for these AI systems (such as AI as a Driver). The FG-AI4AD concluded its activities on 29 September 2022.

- On the ISO side, a new project is being developed in a new WG of SC32 (SC32/WG14) which is being created. This new project has been approved by the member countries of TC22 and registered as: PAS 8800 ”Road Vehicles - Safety and Artificial Intelligence”. The purpose of this document is to provide industry-specific guidance on the use of safety-related Artificial Intelligence/Machine Learning (AI/ML)-based functions in road vehicles. As such it will define a set of safety principles compatible with existing approaches currently defined within the standards ISO 26262 (functional safety) and ISO 21448 (Safety of the intended functionality). It will seek to address any gaps in ISO 26262 specifically related to the safe implementation of AI components and support processes such as tool qualification. The document will also support the harmonization of concepts already described in Annexes of ISO 21448 and ISO/TR 4804 whilst extending these with specific guidance regarding the definition of safety-related properties of AI/ML and the creation of associated safety evidence along the development and deployment life cycle.

The document shall build upon guidance contained within ISO/IEC TR 5469 (under development) and shall not be restricted to specific AI/ML techniques or specific vehicle functions.

5.1.1 State of thinking around regulation, ODD, tests and protocols for validation & homologation in 2024

First, many AI-based functions are already implemented and so have to be type approved (which means to verify compliance with regulation requirements & type approval tests), even if some functions are currently unregulated:

- Braking (UN-ECE R13 regulations)
- Steering (UN-ECE R79 regulation)
- Automated functions like ACC (automated cruise control), (no regulation today)
- Automated Lane Keeping Warning / Alert / Centering functions (no regulation today)
- AEBS (automated emergency braking system) (UN-ECE R152 regulation)
- ESF (emergency steering function), (UN-ECE R79 regulation amendment)
- ALKS (automated lane keeping system), (UN-ECE R157 regulation)
- ADS (automated driving system), (EU ADS regulation)
- AVP (automated Valet Parking), (EU ADS regulation)
- Etc

Some of these regulations requires only one to three type approval tests (breaking, steering), but some others require 20 to 40 type approval tests (AEBS, ALKS, ADS).

Secondly, according to the requirements of all autonomous vehicle regulations (ALKS, ADS, draft of the Arreté francais autonomous urban shuttles), the OEM will have to declare to the customers and to the type approval authority its ODD (Operational Design Domain). For example, an OEM will declare that its autonomous driving functionality are safe and operational for speeds of not more than 30 km/h. In addition, the ODD limits will define the tests, the limits on which the AI based vehicle will be tested, verified, and type approved. For these two reasons, PRISSMA project duration and budget are not enough to investigate adaptation of all functions and all type approval tests for AI-based vehicles. The project will propose a global methodology going far beyond current regulations. It can be adapted to different functions (but is not function-specific) and will be implemented through POCs, to view its declination and investigate some of the most important application/function and how to adapt them for AI-based vehicle with potential safety weak points. The methodology will also present how to complement the traditional on-track approach with simulation approaches and will study the practical application of case studies on real roads.

In addition to offering a comprehensive global methodology for evaluating embedded AI, part of the task WP3 will also look at how to adapt current regulations at a minimum by developing new scenarios to evaluate anticipation and overfitting of AI-based vehicles. For example, we built a catalogue of 18 existing or new scenarios, to verify during vehicles or functions homologation that there are no weak points related repeatability, robustness, anticipation and overfitting. These tests can be chosen when they are relevant for the considered

vehicle/function/regulation. While remaining in the spirit of current regulations, these new tests and scenario are quite far ahead compared to on-going regulation or Euro NCAP discussions to evaluate AI-based vehicles.

Today GRVA regulation group discussions about AI-based vehicles homologation are not very advanced and consider that existing regulations could be sufficient to verify AI-based vehicles safety:

- EU AI act
- UNECE software regulation (UNECE 156)
- UNECE cyber regulations (UN-ECE R155)

These three regulations mainly require audits (of AI and software development, validations, production, reparations,) but not additional tests to evaluation vehicles and testing tracks.

Beginning 2024 GRVA discussions main points are:

- a proposal for a draft resolution with guidance on AI in the context of road vehicles, with
  - the status of the current situation,
  - limitations due to software updates requirements,
  - AI-based vehicles may allow a trade-off of various desirable model characteristics : model drift and staleness, model complexity, robustness, verifiability, predictability and over-fitting etc, while guaranteeing a certain level of safety and cybersecurity. AI-based systems should provides possibilities for system updates.
  - AI-based systems can contribute to improve vehicle safety, with additional benefits for road safety
- Decomposition of AI based systems and potential new requirements, see figure 5.

Therefore, PRISSMA proposals of new scenarios to tests and new metrics to verify AI-based vehicles/functions (on repeatability, robustness, anticipation and overfitting) still remain to be presented and discussed to this GRVA regulation group, and also to Euro NCAP working groups, which don't address AI-based vehicle yet.

5.1.2 State of thinking around metrics and requirements for homologation in 2024

What are the today metrics and requirements for ADAS and AD homologation?

Until this year, homologation and regulations have to be very simple to warranty safety verifications in reasonable duration and costs and which, overall, concerns only track tests.

So today tests to type approve one new function are generally very few, one to 10 most of time, and 20 to 50 tests for the most complex functions like autonomous driving functions (AEBS, ALKS, ADS,...) on many scenarios and configurations (speed, loading of the vehicle,...).

With so few tests, validation criteria are also very simple and are closer to simple KPIs than to real metrics. This is a far cry from the PRISSMA WP1 (see deliverable 1.5) recommendations

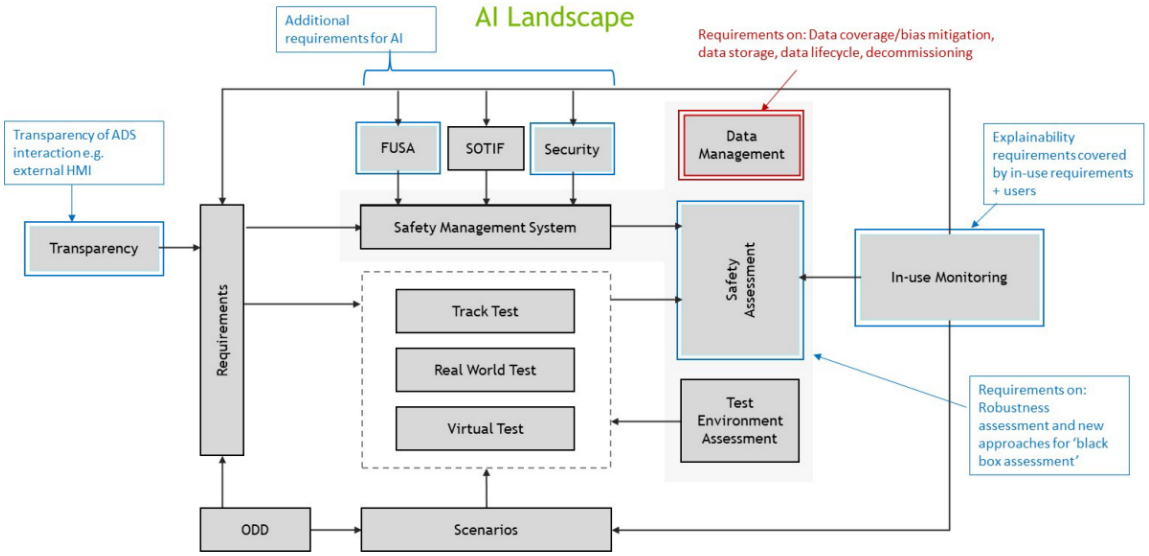

Figure 5: representation of the environment around AI for GRVA

and metrics, because they suppose to have thousands of test results, which is much easier and less costly to obtain in simulation than on your physical track tests. In current ADAS and AD regulations no virtual tests are allowed to replace physical tests on closed tracks. It's in discussion in GRVA but still not decided.

For example, most complex and recent functions like AEBS, ALKS, and ADS have the following basic KPI, like for UNECE R152 AEBS regulation, the most deployed today:

- Do for each scenario and parameters (speed, loading of the vehicle,) 2 tests. If there is one unsuccessful test, do it a third time. If the third test is ok the scenario is successful.
- Do all tests for all categories of scenarios (scenarios with car target, pedestrian target, bicycle target,), the ratio of unsuccessful tests don't have to be higher than
  - 10% for tests of car-to-car scenarios
  - 10% for tests of car to pedestrian scenarios
  - 20% for tests of car to bicycle scenarios.

This approach to control is still very lightweight, and although it may be suitable for low levels of autonomy (level 2 for these functions) and with simple AI models, we may wonder about its suitability for more advanced levels of autonomy. And from an AI evaluation point of view, this approach can even be seen as totally obsolete (in terms of AI evaluation, you can't talk about repeatability with only 3 maximum repetitions) but has been kept for reasons in regulation mainly of cost and practicality for track testing.

5.2 Rail

5.2.1 Shift2Rail upcoming deliverables

Shift2Rail project has been described above in this document. It's an ongoing project hence we can still wait for the outcomes to be part of discussion on AI.

5.2.1.1 IP2 Advanced Traffic Management and Control Systems' solution [SHI19]

Some of previously describe WP04 Automatic Train Operation upto GoA4 deliverables are yet to come. Indeed in the framework of X2RAIL-4, which will run until February 2023, there are two separate activities; Specification of GoA2 and then Preliminary specification for GoA3/4 based on GoA2. In parallel WP3 will focus on "ATO upto GoA4" specification. It will develop an ATO on-board prototype, ATO trackside prototype, develop or integrate obstacle detection and environment sensors. The ATO prototypes will include: • Passenger Exchange, door opening/closing, safe departure • Incident & Emergency detection and management • Localisation. • Remote Control supporting degraded modes The prototypes also need to demonstrate agreed interoperability and agreed interchangeability of the prototypes provided by different suppliers. Finally, WP5 "ATO up to GoA4" Tests will develop or upgrade Reference Test Benches required to perform test in factory, factory interoperability tests involving the prototypes developed by the different suppliers, On-site tests on Pilot Line and Pilot train involving the prototypes developed by the different suppliers in compliance with the functional and vendor independent reference Specification delivered in D3.2.

In Parallel, the Europe's Rail FP2 R2DATO project is focusing on delivering scalable ATO up to GoA4 across various railway segments, including freight and urban light rail.it aims to demonstrate operational solutions for automation through specific use cases and technical enablers, based on the European Rail Traffic Management System ERTMS, targeting the maximization of train operations.

The Europe's Rail FP2 R2DATO project, launched in 2023,is significant endeavor aimed at advancing automation and digitalization of railways within the European Union.

Key Objectives include increasing capacity, reducing operational costs through advanced technologies like ETCS Hybrid level 3 and moving block systems, and developing innovative solutions such as autonomous route setting and virtually coupled train sets.

Results from FP2 R2DATO are expected by 2025, covering key topics such ATO, ETCS, digital technologies, and guidelines for the deployment of these technologies across Europe.

These Projects collectively represent a significant leap towards the automation and digitalization of the European railway system, promising increased safety, capacity, and efficiency while reducing costs and energy consumption (https://projects.rail-research.europa.eu/eurail-fp2/).

5.2.1.2 IP3 Intelligent Asset Management and High Capacity Infrastructure [SHI19]

In this IP the "Automation: robot platform" could be a solution for maintenance optimization. It is to analyze whether it will use the AI for data analytics and decision making or only automation.

5.2.1.3 IP5 Technologies for sustainable and attractive European Rail Freight

Obstacle Detection System (ODS) SMART2 took succession to SMART. It consists of 5 WP.

- WP1 - Use Cases, Requirements and Specifications
  - D1.1 FREIGHT SPECIFIC USE CASES FOR OBSTACLE DETECTION AND TRACK INTRUSION SYSTEMS

– D1.2 ANALYSIS OF REQUIREMENTS AND DEFINITION OF SPECIFICATIONS FOR OBST DETECTION & TRACK INTRUSION

- WP2 - On-board obstacle and track intrusion detection system
- WP3 - Trackside/Airborne obstacle and track intrusion detection system
- WP4 - Decision Support system
- WP5 - Prototype integration with D5.1 REPORT ON SUB-SYSTEMS CONFORMANCE TESTING
- WP6 - Evaluation

The outcomes of WP 2, 3, 4 and 6 are yet to be published.

5.2.2 Europe's Rail

Europe's Rail will be the successor of Shift2Rail on rail research and innovation and it falls under the Horizon Europe program (2020-2027). Its targets include automation. It is this field that monitoring this program can be useful.

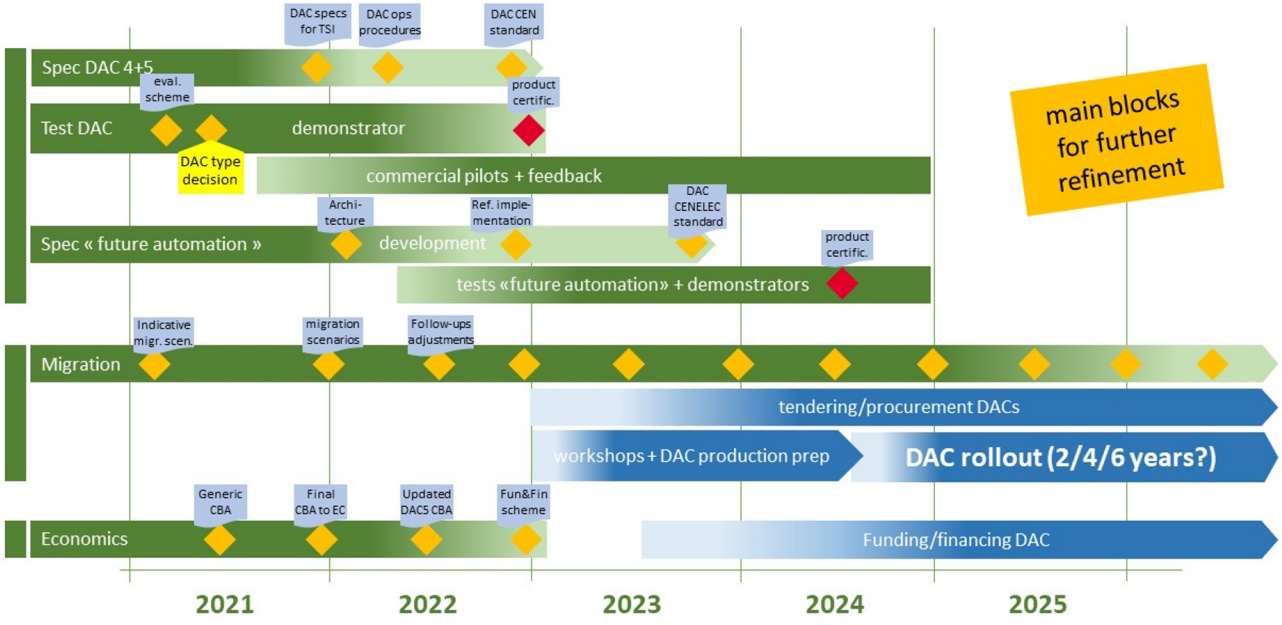

Figure 6: Europe's Rail "future automation" timeline

5.2.3 Flagship Project FP1-MOTIONAL

FP1-Motional is a flagship project focusing on advancement in rail technology and operations to enhance the efficiency, reliability and sustainability of rail services. It aims to establish and provide high-level specifications for requirements, designs and uses cases for developing technical enablers 1 to 7. The high level specifications are done in parallel with more detailed specifications from WP4, WP6 and WP8, incorporating state of the art analyses and previous Shift2Rail Results (see Fig. 7).

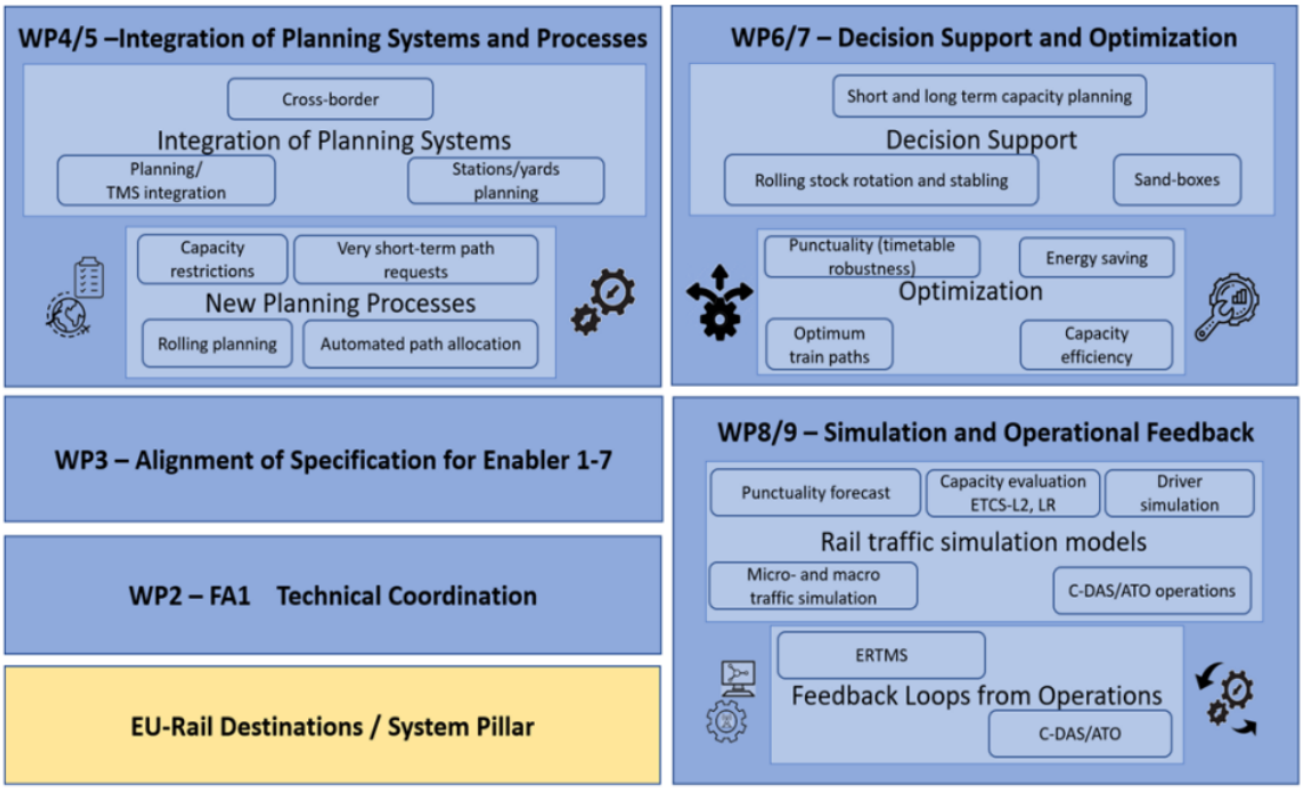

Figure 7: WPs FP1 Motional

5.3 Aeronautics

5.3.1 EASA upcoming deliverables

EASA has set European Ethical guidelines and determined Trustworthy AI building blocks illustrated in the diagram below. It set 5 Objectives and to achieve them they have built a timeline. The second guideline regarding the first level of Machine learning has been published.

The EASA AI Roadmap has been extended to encompass all techniques and approaches described in the figure 9:

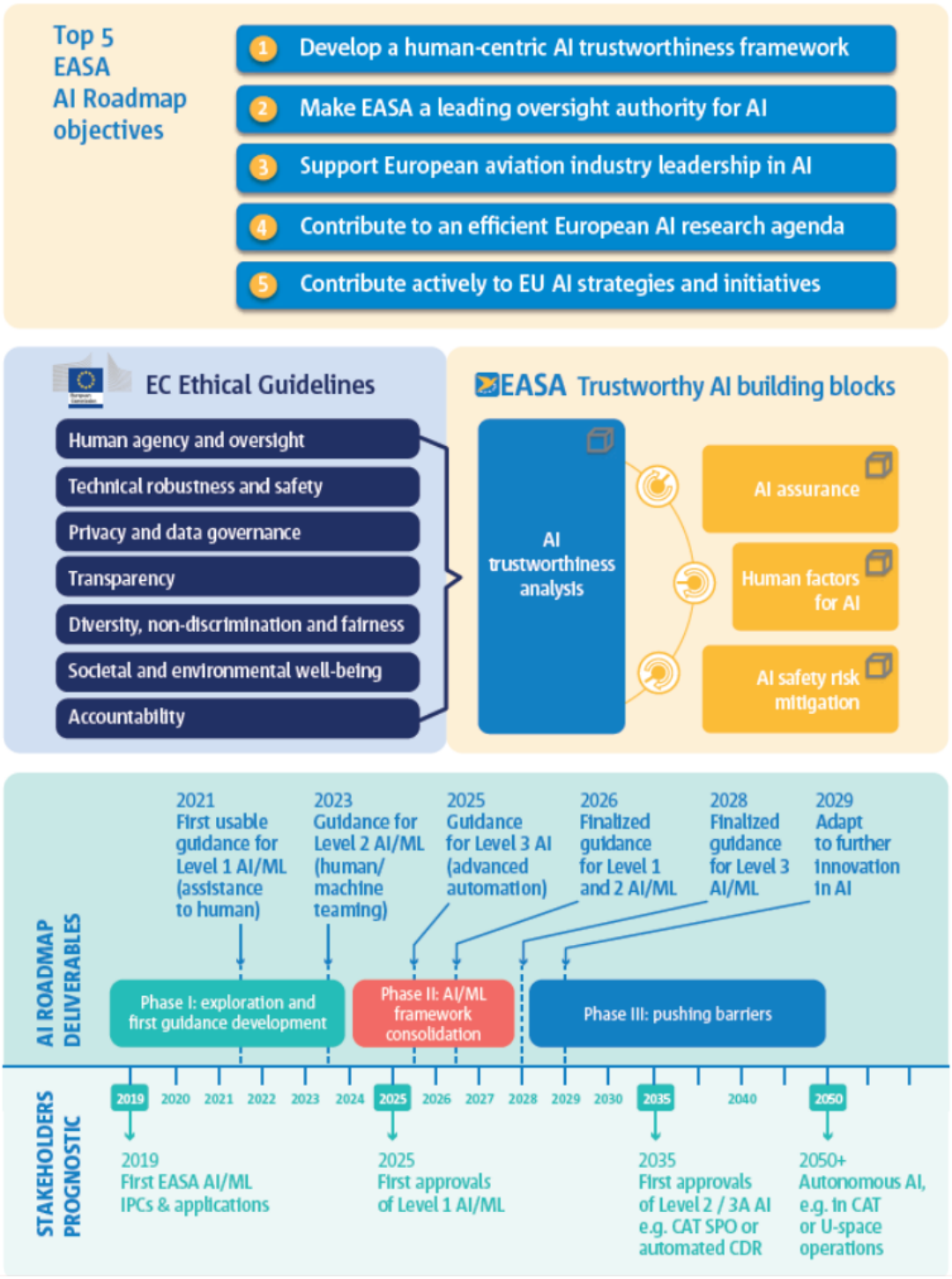

Figure 8: EASA AI Roadmap 2.0

The analysis of AI's anticipated impact across various domains,domains highlights shared issues but also identifies the need to consider domain-specific factors. This calls for a mixed rule-making approach: one involving cross-domain rules (horizontal) and another focusing on domain-specific regulations (vertical). This approach will be implemented in two steps.

First, the development of a transversal Part-AI will encompass key provisions outlined in the

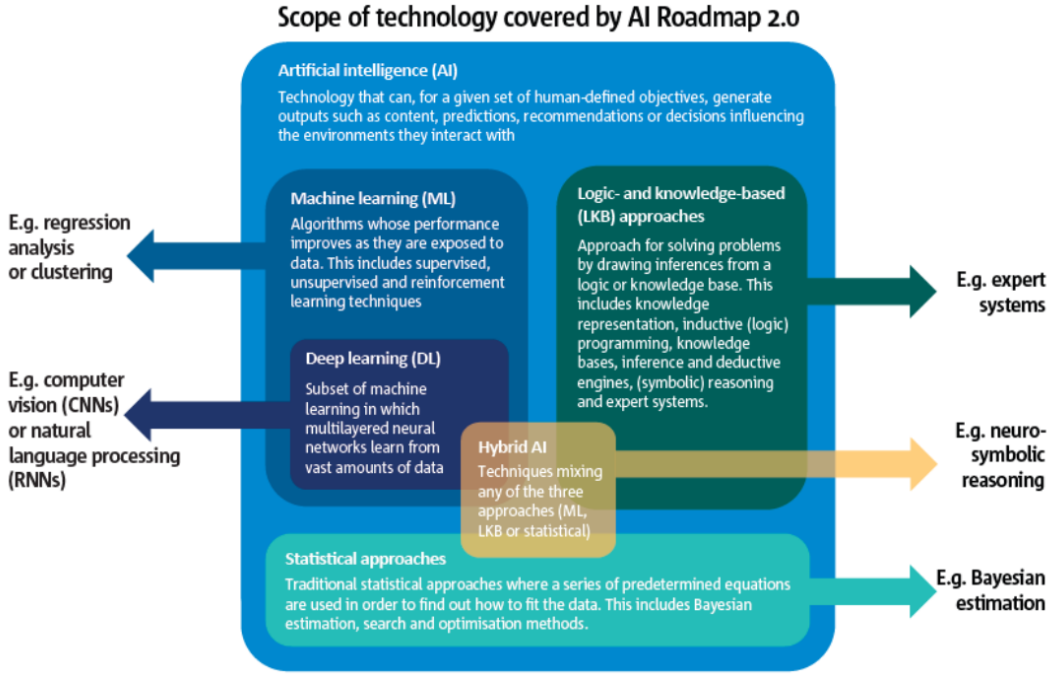

Figure 9: Scope of technology covered by AI Roadmap 2.0

Concept Papers, including requirements for authorities, organizations, and AI trustworthiness. Additionally, acceptable means of compliance and guidance material will be provided to align with industry standards where necessary.

Second, a domain-specific analysis will be conducted to identify additional requirements needed for a comprehensive regulatory framework. This approach also considers the EU AI Act and aligns with the identified rule-making activities in the European Plan for Aviation Safety (EPAS) 2023-2025. The regulatory structure is anticipated as follows (Figure 10):

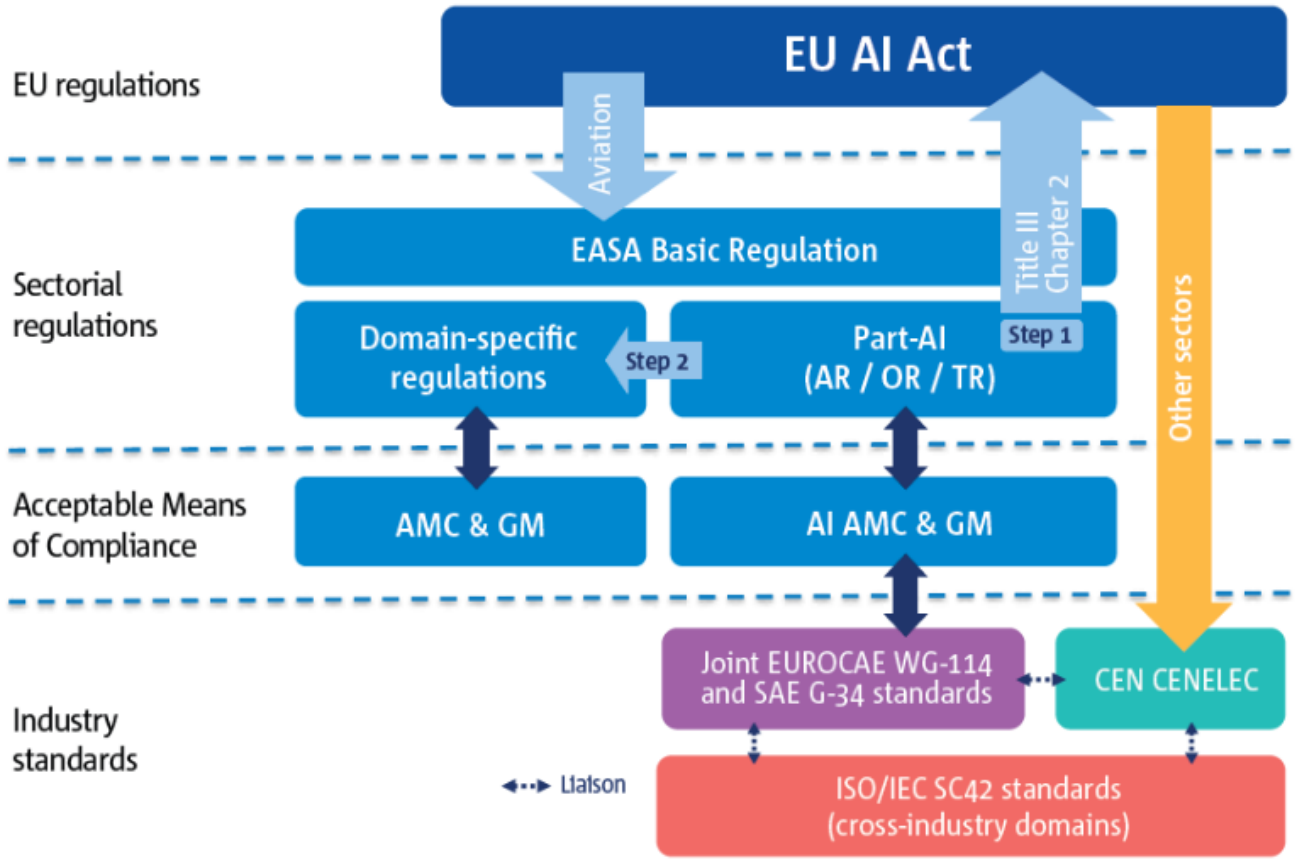

Figure 10: Anticipated regulatory structure for AI

5.3.2 Machine Learning Application Approval MLEAP Project

EASA has been actively exploring the integration of artificial intelligence and machine learning technologies into various aspects of aviation safety, regulation and operations, including the MLEAP Projects.

The MLEAP (Machine Learning Environment for Aircraft Performance) project is an initiative by the EASA. It Aims to explore the application of machine learning techniques to enhance aircraft performance evaluation and safety within the aviation industry.

The Objectives is Streamline certification and approval processes by identifying concrete means of compliance with the learning assurance objectives of the EASA guidance for ML applications.

The project focuses on capitalizing advanced data analytics and machine learning algorithms to improve the understanding of aircraft behavior, performance, monitoring, and predictive maintenance.

The MLEAP project is a two-year work initiated by the EASA to collate and evaluate the state of the art on three main topics:

- Task1: Data completeness and representativeness:
  Provide a list of factors influencing the choice of tools and approaches in order to assess the completeness and representativeness of databases, with corresponding justifications and bibliographical references.
- Task2: Model development: Generalization properties:
  Identification or development of efficient methods and tools for the quantification of generalization assurance level in the generic case of data-driven ML/DL development
  - Test available methods and tools to evaluate generalization bounds;
  - Barriers in generalization guarantees for a given model: ML and DL;
  - Identification/proposal of means to promote models generalization.
- Task3: Evaluation: robustness and stability
  - Review of methods and tools
  - Review of methods to identify corner cases and abnormal inputs
  - Identification of sources of instabilities during the design phase
  - Identification of sources of instabilities during the operational phase
  - Demonstration on a use-case for the intended application

The methodology of the MLEAP project involves the following several key steps:

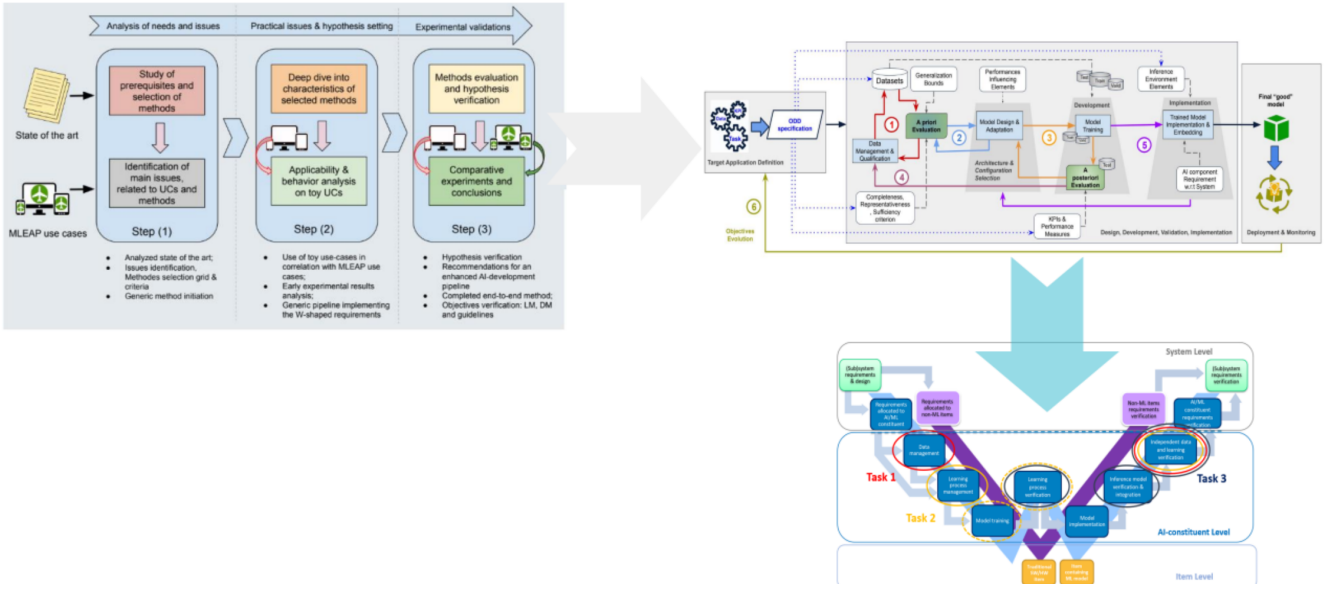

Figure 11: Key steps of Mleap

5.4 Health

FDA Artificial Intelligence and Machine Learning in Software as a Medical Device Action Plan, January 2021

In April 2019, the FDA published a Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device[21], a discussion paper requesting feedback. This paper has inspired significant discussion in the area of AI for Medical Devices, generating hundreds of comments and numerous peer-reviewed articles. Public workshops have also been held to gather feedback from various stakeholders.

Based on the received feedback, FDA has established an Action Plan for AI/ML software as a medical device. The following actions have been identified:

1. Tailored Regulatory Framework for AI/ML-based SaMD. The proposed framework relies on a "Predetermined Change Control Plan", considering two aspects of AI/ML systems separately: first, what parts of the software are intended to change with machine learning, and second how these changes will be implemented while preserving safety and efficacy of the software.
2. Good Machine Learning Practice. This is a major direction of action, as the FDA is contributing to harmonization of standards and best practices in AI/ML. Several groups in which the FDA is taking part are: IEEE P2801 "Artificial Intelligence Medical Device Working Group", ISO/IEC JTC 1/SC 42 "Artificial intelligence" Technical committee, AAMI/BSI Initiative on AI in medical technology.
3. Patient-Centered Approach Incorporating Transparency to Users. AI/ML based devices necessitate a proactive patient-centered approach. To that end, the FDA has held a Patient Engagement Advisory Committee that has allowed extracting recommendations on information that should be included in device labelling. The FDA will also hold a public workshop on device labelling for transparency and trust in AI/ML based devices.
4. Regulatory Science Methods Related to Algorithm Bias and Robustness. It is necessary to have improved methods to evaluate and address algorithmic bias and to promote algorithm robustness. The FDA will support regulatory efforts in this domain and help develop methodology for the evaluation and improvement of machine learning algorithms, including for the identification and elimination of bias, and for the evaluation and promotion of algorithm robustness.
5. Real-World Performance. There is a need for clarity on Real-World Performance monitoring for AI/ML software. The FFDA will support the piloting of real-world performance monitoring by working with stakeholders on a voluntary basis.

As the latest update in October 2023, a comprehensive analysis of 691 FDA-approved AI/ML-enabled medical devices was performed. This study provided insights into clearance pathways, approval timelines, regulations tapes, medical specialities, decision types, and recall history, highlighting a significant increase in approvals since 2018, particularly in radiology due to the abundance of clinical data.

[21] https://www.fda.gov/media/122535/download

5.5 Defence

Group of Governmental Experts related to emerging technologies in the area of lethal autonomous weapons systems - 19/04/2021

The mandate of this Working Group[22] is to establish guiding principles for the development of legal, technological and military aspects of Lethal Autonomous Weapon Systems (LAWS) ahead of the Conference of the Convention on Certain Conventional Weapons, organised by the UN and to be held in September 2021.

11 guiding principles have been decided upon. The Group may further develop and elaborate these principles.

1. Humanitarian law continues to apply fully to all weapons systems, including the potential development and use of lethal autonomous weapons systems.
2. Human responsibility for decisions on the use of weapons systems must be retained, since accountability cannot be transferred to machines. This should be considered across the entire life cycle of the weapons system.
3. Human-machine interaction, which may take various forms and be implemented at various stages of the life cycle of a weapon, should ensure that the potential use of weapons systems based on emerging technologies in the area of lethal autonomous weapons systems complies with applicable international law, in particular IHL. In determining the quality and extent of human-machine interaction, a range of factors should be considered including the operational context, and the characteristics and capabilities of the weapons system as a whole.
4. Accountability for developing, deploying and using any emerging weapons system in the framework of the CCW must be ensured in accordance with applicable international law, including through the operation of such systems within a responsible chain of human command and control.
5. In accordance with States' obligations under international law, in the study, development, acquisition, or adoption of a new weapon, means or method of warfare, determination must be made whether its employment would, in some or all circumstances, be prohibited by international law.
6. When developing or acquiring new weapons systems based on emerging technologies in the area of lethal autonomous weapons systems, physical security, appropriate non-physical safeguards (including cyber-security against hacking or data spoofing), the risk of acquisition by terrorist groups and the risk of proliferation should be considered.
7. Risk assessments and mitigation measures should be part of the design, development, testing and deployment cycle of emerging technologies in any weapons systems.
8. Consideration should be given to the use of emerging technologies in the area of lethal autonomous weapons systems in upholding compliance with IHL and other applicable international legal obligations.
9. In crafting potential policy measures, emerging technologies in the area of lethal autonomous weapons systems should not be anthropomorphized.

[22] https://meetings.unoda.org/meeting/ccw-gge-2021/

10. Discussions and any potential policy measures taken within the context of the CCW should not hamper progress in or access to peaceful uses of intelligent autonomous technologies.
11. The CCW offers an appropriate framework for dealing with the issue of emerging technologies in the area of lethal autonomous weapons systems within the context of the objectives and purposes of the Convention, which seeks to strike a balance between military necessity and humanitarian considerations.

In conclusion, the Group has underlined the importance of comprehensive, context-based human judgement in ensuring that the potential use of weapons systems based on emerging technologies is in compliance with international law, and in particular IHL. This human judgement must be unclouded and not affected by the system interface. The system must reliably and predictably perform its functions in accordance with the intention of the human operator, and the responsibility of the operator is retained in actions of the system.

However, the Group did not come up with concrete policy recommendations or elements of a legally binding instrument on LAWS.

5.6 Future European regulation on high-risk AI systems

This proposal for a regulation, which has just been definitively adopted and should come into force in 2025, also called "Artificial Intelligence Act", has the purpose of laying down harmonised rules on artificial intelligence. It aims to preserve the European technological leadership while ensuring the respect of EU values, fundamental rights and principles. This proposal should implement the principles of "The White Paper on AI" published in 2020.

The regulatory approach of the proposal is limited to the minimum necessary requirements ; it does not create unnecessary restrictions to trade, but imposes a set of mandatory requirements for "high-risk" AI systems (AI in VAs to be classified as high-risk).

The AI Act will eventually be accompanied by a Data Governance Act, the construction of which was launched this year after initial reflections in 2022. We can therefore expect the AI Act to come into force in 2025, with the Data ACt coming later. As a preamble, there are one thing to remember about the AI ACT's scope of application for the automotive industry, it will be mainly through the introduction of harmonized standards, which should be published in 2024 or 2025 at the latest. This means that the standards aspect of the automotive industry will undergo major changes in the near future.

This new regulation will require careful monitoring of certain aspects of AI development and use, which we will describe in the next parts of this section.

5.6.1 Risk management system

This section describes the constraints on the risk management system for high-risk AI systems.

Firstly, the risk management system shall be a continuous iterative process run throughout the entire life-cycle of the high-risk AI system. Firstly, it should proceed to identification and analysis of the known and foreseeable risks, as well as risks that may emerge during use. Further risk evaluation shall be carried out based on data gathered from the post-market monitoring system. Finally, the risk management system shall adopt suitable risk management measures, such that any residual risk is judged acceptable. The residual risks must be communicated to the user.

The appropriate risk management measures include elimination or maximal reduction of risks through design and development, adequate mitigation and control measures for risks that cannot be eliminated, and provision of adequate information and training to users.

Finally, high-risk AI systems shall be tested to verify the AI system performs consistently and to identify the most appropriate risk management measures. This testing will be performed at any appropriate point in time in the development process. It will be based on preliminary defined metrics and will not need to go beyond the intended purpose of the AI system.

5.6.2 Data and data governance

This section applies to any system that uses a model trained with data, and also to all high-risk AI system. It defines quality criteria for the training, validation and testing data.

Appropriate data governance and management practices must be implemented, concerning design choices, data collection and preparation (annotation, labelling, cleaning, enrichment and aggregation), assessment of the availability, quantity and suitability of the needed data sets, and identification of possible biases, data gaps and shortcomings.

The training, validation and testing data sets shall be relevant, representative, free of errors and complete. They should take into account the characteristics specific to the environment and function of the AI system.

The providers of high-risk AI systems may process some categories of personal data in order to ensure bias monitoring (following strict security and privacy rules such as pseudonimisation).

5.6.3 Technical documentation

This section defines requirements on technical documentation that will apply to any high-risk AI system.

The technical documentation must demonstrate that the AI system complies with the requirements. It must include at least the following information:

- A general description of the AI system (intended purpose, developer, versions, interaction with other systems and instructions for the user)
- A detailed description of the elements of the AI system and the development process (methods and steps performed for the development, use of pre-trained and third-party systems, design specifications i.e description of the general logic of the system and algorithms, optimisation parameters and trade-offs)
- A description of the system architecture (integration of software components into the overall processing)
- If relevant, the data requirements (with a description of training methodologies, of training datasets and of data preparation methodologies)
- An assessment of the necessary human oversight measures
- If applicable, a detailed description of pre-determined future changes and of the technical solutions to ensure continuous compliance of the AI system with this regulation.
- A description of the validation and testing procedures, including information about the data and metrics used, test logs and all test reports.

- Detailed information about the monitoring, functioning and control of the AI system (capabilities and limitations in performance, foreseeable unintended outcomes and sources of risks). Detail the necessary human oversight.
- A detailed description of the risk management system and of any change made to it
- A list of harmonised standards applied in full or in part, or relevant to the AI system.
- A copy of the EU declaration of conformity.
- A detailed description of the performance monitoring system.

5.6.4 Record-keeping

High-risk AI systems must keep logs of events when it is operating. The level of logging must be appropriate to the intended purpose of the system.

These logging capabilities must enable the monitoring of the AI system, especially in situations where it may present a risk. In the case of AI systems for biometric identification and categorisation of persons, the logs must include at least a recording of the period of each use of the system, the reference database against which input data is checked, the input data that has lead to a match, and the identity of natural persons involved in the verification of the results.

5.6.5 Transparency and provision of information to users

High-risk AI systems must be designed in a way that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately. Firstly, the system must provide instructions for use that are concise, complete, correct and clear. This documentation must specify the following information:

- the identity and contact details of the provider of the AI system
- the characteristics of the high-risk AI system: its intended purpose, its level of accuracy, robustness and cybersecurity, circumstances that may lead to risks, the level of performance, and specifications for the input data of the system
- the pre-determined future changes to the system
- the human oversight measures (see next section)
- the expected lifetime of the system and the necessary maintenance and care measures to ensure its proper functioning

5.6.6 Human oversight

High-risk AI systems must be built in such a way that their use can be effectively overseen by natural persons. This oversight will aim at preventing or minimizing the risks that remain despite other risk mitigation procedures.

The oversight measures must be identified before it is put into service, and ether built into the system or implemented by the user. These measures must allow the human to understand the capacities and limitations of the AI system and to monitor its operation while remaining aware of the ”automation bias”. The human must be able to correctly interpret the system's output and, if necessary, to disregard, override or reverse this output. The oversight tools must allow the human to intervene or interrupt the system through a ”stop” button or similar.

In the case of AI systems for biometric identification and categorisation of persons, any identification resulting from the system must be confirmed by at least two persons before it leads to any action or decision.

5.6.7 Accuracy, robustness and cybersecurity

The system instructions of use must include a declaration of the system levels of accuracy, as well as the relevant accuracy metrics. The system must be resilient to errors, faults or inconsistencies (especially related to interaction with natural persons or other systems), as well as to third-party attacks and exploits, such as data poisoning or adversarial examples.

Systems that continue to learn after being placed on the market must provide mitigation measures for the possible biased outputs caused by ”feedback loops” (outputs of the system reused as inputs).

REFERENCES

[Abd17] Waleed Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017.

[ADR20] Loic Coquelin Adel Djoudi and Rémi Régnier. A simulation-based framework for functional testing of automated driving controllers. IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), 2020.

[AI20] Automative-IQ. ISO 26262: Cost-optimized FUSA, ASPICE assessment and part II challenges. 2020.

[AK16] Christian Alwardt and Martin Krüger. Autonomy of weapon systems. Institut für Friedensforschung und Sicherheitspolitik an der Universität Hamburg, 2016.

[AKF10] S. Ameli A. Khodayari, A. Ghaffari and J. Flahatgar. A historical review on lateral and longitudinal control of autonomous vehicle motions. International Conference on Mechanical and Electrical Technology, page 421–429, 2010.

[ASSR20] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression, 2020.

[AST19a] D3.1 state of the art of automated driving technologies. Technical report, ASTRAIL, 2019.

[AST19b] D3.2 automatic train operations implementation operation characteristics and technologies for the rail. Technical report, ASTRAIL, 2019.

[ATC19] Mohamad Ali Assaad, Reine Talj, and Ali Charara. Autonomous driving as system of systems: roadmap for accelerating development. In 2019 14th Annual Conference System of Systems Engineering (SoSE), pages 102–107, 2019.

[BHDN19] Alexis Basantis, Leslie Harwood, Zachary Doerzaph, and Luke Neurauter. Standardized Performance Evaluation of Vehicles with Automated Capabilities. Technical Report VTTI-00-020, December 2019.

[Blo96] Isabelle Bloch. Information combination operators for data fusion: A comparative review with classification. IEEE Transaction on Systems, Man, and Cybernetics-part A, pages 52–67, 1996.

[BS20] Eugene Bagdasaryan and Vitaly Shmatikov. Blind backdoors in deep learning models. arXiv preprint arXiv:2005.03823, 2020.

[BWL20] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection, 2020.

[CAM19] Intelligent transport systems (its) vehicular communications basic set of applications part 2: Specification of cooperative awareness basic service. https://www.etsi.org, 04 2019.

[CCKL19] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving, 2019.

[CKD+22] Marco Casadio, Ekaterina Komendantskaya, Matthew L Daggitt, Wen Kokke, Guy Katz, Guy Amir, and Idan Refaeli. Neural network robustness as a verification property: a principled case study. In International Conference on Computer Aided Verification, pages 219–231. Springer, 2022.

[CLT+19] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.

[CON18] D1.3 automated brake test. Technical report, A. R. C. CONSORTIUM, 2018.

[CPM18] Intelligent transport systems (its) vehicular communications basic set of applications part 2: Informative report for the collective perception service. https://www.etsi.org, 05 2018.

[DBP17] W. Maddern D. Barnes and I. Posner. Find your own way: Weakly-supervised segmentation of path proposals for urban autonomy. IEEE International Conference on Robotics and Automation (ICRA), pages 203–210, 2017.

[DKA+22] Matthew L. Daggitt, Wen Kokke, Robert Atkey, Luca Arnaboldi, and Ekaterina Komendantskya. Vehicle: Interfacing neural network verifiers with interactive theorem provers, 2022.

[DRC+17] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In Proceedings of the 1st Annual Conference on Robot Learning, pages 1–16, 2017.

[DSSCY14] Guilbert D., Ieng S.-S., Le Bastard C., and Wang Y. Robust blind deconvolution process for vehicle re-identification by an inductive loop detector. IEEE Sensors Journal, 14(12):4315–4322, 12 2014.

[EAS20a] Concepts of design assurance for neural networks (codann). Technical report, EASA, 2020.

[EAS20b] Easa artificial intelligence roadmap 1.0. Technical report, EASA, 2020.

[EAS21a] Concepts of design assurance for neural networks (codann) ii (europa.eu). Technical report, EASA, 2021.

[EAS21b] First usable guidance for level 1 machine learning applications - issue 01. Technical report, EASA, 2021.

[FGCC19] Nic Ford, Justin Gilmer, Nicolas Carlini, and Dogus Cubuk. Adversarial Examples Are a Natural Consequence of Test Error in Noise. arXiv:1901.10513 [cs, stat], January 2019.

[Get19] Douglas Gettman. Raising awareness of artificial intelligence for transportation systems management and operations. Technical Report FHWA-HOP-19-052, U.S. Federal Highway Administration, december 2019.

[GF15] Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research, 16(1):1437–1480, 2015.

[GGV+10] Dominique Gruyer, Sebastien Glaser, Benoit Vanholme, Nicolas Hiblot, and Bertrand Monnier. Sivic, a virtual platform for adas and padas prototyping, test and evaluation. In FISITA World Automotive Congress, 2010.

[GHF+21] Zahra Ghodsi, Siva Kumar Sastry Hari, Iuri Frosio, Timothy Tsai, Alejandro Troccoli, Stephen W Keckler, Siddharth Garg, and Anima Anandkumar. Generating and characterizing scenarios for safety testing of autonomous vehicles. arXiv preprint arXiv:2103.07403, 2021.

[Gru99] Dominique Gruyer. Étude du traitement de données imparfaites pour le suivi multi-objets : application aux situations routières. Thèse, Université Technologique de Compiègne (UTC), Compiègne, France, 1999.

[GSAB+22] Julien Girard-Satabin, Michele Alberti, François Bobot, Zakaria Chihani, and Augustin Lemesle. CAISAR: A platform for Characterizing Artificial Intelligence Safety and Robustness. In AISafety, CEUR-Workshop Proceedings, Vienne, Austria, July 2022.

[GTA+21] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseo Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, P. Jung, R. Roscher, M. Shahzad, Wen Yang, R. Bamler, and Xiaoxiang Zhu. A survey of uncertainty in deep neural networks. ArXiv, abs/2107.03342, 2021.

[HLG20] Fly ai. Technical report, E. HLG, 2020.

[HW21] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Machine Learning, 110(3):457–506, Mar 2021.

[Jan05] J. Janson. Collision Avoidance Theory with Application to Automotive Collision Mitigation. PhD thesis, Linköping Universitet, 2005.

[JW14] Stewart Worrall Eduardo Nebot James Ward, Gabriel Agamennoni. Vehicle collision probability calculation for general traffic scenarios under uncertainty. 2014 IEEE Intelligent Vehicles Symposium Proceedings, pages 986–992, 2014.

[KG17] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision?, 2017.

[KG19] O.M. Kirovskii and V.A. Gorelov. Driver assistance systems: analysis, tests and the safety case. iso 26262 and iso pas 21448. IOP Conf. Series: Materials Science and Engineering, 2019.

[KHI+19] Guy Katz, Derek A. Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljić, David L. Dill, Mykel J. Kochenderfer, and Clark Barrett. The Marabou Framework for Verification and Analysis of Deep Neural Networks. In Isil Dillig and Serdar Tasiran, editors, Computer Aided Verification, Lecture Notes in Computer Science, pages 443–452, Cham, 2019. Springer International Publishing.

[KYY+15] İslam Kılıç, Ahmet Yazıcı, Ömür Yıldız, Mustafa Özçelikors, and Atakan Ondoğan. Intelligent adaptive cruise control system design and implementation. In 2015 10th System of Systems Engineering Conference (SoSE), pages 232–237, 2015.

[LAL+19] Changliu Liu, Tomer Arnon, Christopher Lazarus, Clark Barrett, and Mykel J. Kochenderfer. Algorithms for Verifying Deep Neural Networks. arXiv:1903.06758 [cs, stat], March 2019.

[LCW17] L. Svensson L. Caltagirone, M. Bellone and M. Wahde. Lidar-baseddriving path generation using fully convolutional neural networks. IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), pages 1–6, 2017.

[LF17] M. Glazer W. Angell S. Dodd B. Jenik B. Reimer et al L. Fridman, D.E. Brown. Mit autonomous vehicle technology study: Large-scale deep learning based analysis of driver behavior and interaction with automation. IEEE Access, 2017.

[Lip16] Zachary C. Lipton. The Mythos of Model Interpretability. arXiv:1606.03490 [cs, stat], June 2016.

[LPB17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017.

[MAP20] Intelligent transport systems (its) vehicular communications basic set of applications part 2: Facilities layer protocols and communication requirements for infrastructure services. https://www.etsi.org, 02 2020.

[MG18] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks, 2018.

[MJ16] Guido Manfredi and Yannick Jestin. An introduction to ACAS Xu and the challenges ahead. In 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), pages 1–9, Sacramento, CA, USA, September 2016. IEEE.

[MMM01] P. H Bovy M. M Minderhoud. Extended time-to-collision measures for road traffic safety assessment. Accident; Analysis and Prevention, pages 89–97, 2001.

[MMS+17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv:1706.06083 [cs, stat], June 2017.

[MSI22] Joao Marques-Silva and Alexey Ignatiev. Delivering trustworthy ai through formal xai. Proceedings of the AAAI Conference on Artificial Intelligence, 36(11):12342–12350, Jun. 2022.

[MT20] Luke Merrick and Ankur Taly. The Explanation Game: Explaining Machine Learning Models Using Shapley Values. arXiv:1909.08128 [cs, stat], June 2020.

[NvBS20] Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14928–14938, 2020.

[PNdS20] Rafael Padilla, Sergio L. Netto, and Eduardo A. B. da Silva. A survey on performance metrics for object-detection algorithms. In 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), pages 237–242, 2020.

[QYSG17] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. CoRR, 2017.

[RBBV20] Anian Ruoss, Maximilian Baader, Mislav Balunović, and Martin Vechev. Efficient Certification of Spatial Robustness. arXiv:2009.09318 [cs, stat], September 2020.

[RCB+21] Wonryong Ryou, Jiayu Chen, Mislav Balunovic, Gagandeep Singh, Andrei Dan, and Martin Vechev. Scalable Polyhedral Verification of Recurrent Neural Networks. In Alexandra Silva and K. Rustan M. Leino, editors, Computer Aided Verification, volume 12759, pages 225–248. Springer International Publishing, Cham, 2021.

[RF18] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv, 2018.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs, stat], August 2016.

[Rud18] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1:206 – 215, 2018.

[SAA+17] Jeremy Straub, Wafaa Amer, Christian Ames, Karanam Ravichandran Dayananda, Andrew Jones, Goutham Miryala, Nathan Olson, Noah Rockenback, Franklin Slaby, Santipab Tipparach, Samuel Fehringer, David Jedynak, Haiming Lou, Dakota Martin, Marc Olberding, Austin Oltmanns, Brady Goenner, Jessie Lee, and Dylan Shipman. An internetworked self-driving car system-of-systems. In 2017 12th System of Systems Engineering Conference (SoSE), pages 1–6, 2017.

[SBBG20] Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. When and why test-time augmentation works, 2020.

[SG15] Junnan Song and Shalabh Gupta. Slam based shape adaptive coverage control using autonomous vehicles. In 2015 10th System of Systems Engineering Conference (SoSE), pages 268–273, 2015.

[SG20] T. Cocias G. Macesanu S. Grigorescu, B. Trasnea. A survey of deep learning techniques for autonomous driving. J. Field Robotics, page 362–386, 2020.

[SHI19] Shift2rail catalogue of solutions. Technical report, Shift2Rail, 08 2019.

[SLB16] A. Carvalho S. Lefèvre and F. Borrelli. A learning-based frameworkfor velocity control in autonomous driving. IEEE Transactions on Automation Science and Engineering, pages 32–42, 2016.

[SMA19a] D1.1 obstacle detection system requirements specification. Technical report, S. H. SMART, 2019.

[SMA19b] D2.1 report on selected sensors for multi-sensory system for obstacle detection. Technical report, S. H. Smart, 2019.

[SMA19c] D2.2 design of the passive vibration isolation system. Technical report, S. H. SMART, 2019.

[SMA19d] D2.3 report on sub-systems conformance testing. Technical report, S. H. SMART, 2019.

[SMA19e] D2.4 report on functional testing of fully integrated multi-sensor obstacle detection system. Technical report, S. H. SMART, 2019.

[SMA19f] D3.1 report on algorithms for 2d image processing. Technical report, S. H. SMART, 2019.

[SMA19g] D3.2 report on smart data fusion and distance calculations. Technical report, S. H. SMART, 2019.

[SMA19h] D3.3 report on real-time algorithm implementation and performance evaluation. Technical report, 2019.

[SMA19i] D7.1 report on evaluation of developed smart technologies. Technical report, S. H. SMART, 2019.

[SMAG18] Martin Simon, Stefan Milz, Karl Amende, and Horst-Michael Gross. Complex-yolo: Real-time 3d object detection on point clouds. CoRR, abs/1803.06199, 2018.

[SoS]

[SS15] Ieng S.-S. Bridge influence line estimation for bridge weigh-in-motion system. ASCE's Journal of Computing in Civil Engineering, 29(1), 02 2015.

[SSSPP21] Masi Stefano, Ieng Sio-Song, Xu Philippe, and Bonnifait Philippe. Augmented perception with cooperative roadside vision systems for autonomous driving in complex scenarios. In 2021 24th IEEE International Intelligent Transportation Systems Conference (ITSC), 2021.

[SSSS16] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. arXiv preprint arXiv:1610.03295, 2016.

[SSSS17] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. On a formal model of safe and scalable self-driving cars. arXiv preprint arXiv:1708.06374, 2017.

[SSSS18] A Shashua, S Shalev-Shwartz, and S Shammah. Implementing the rss model on nhtsa pre-crash scenarios. tech. rep, 2018.

[SVK17] Jiawei Su, Danilo Vasconcellos Vargas, and Sakurai Kouichi. One pixel attack for fooling deep neural networks. arXiv:1710.08864 [cs, stat], October 2017.

[SVS17] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. CoRR, abs/1710.08864, 2017.

[TZJ+16] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In 25th USENIX Security Symposium (USENIX Security 16), pages 601–618, Austin, TX, 2016. USENIX Association.

[UM21] Caterina Urban and Antoine Miné. A Review of Formal Methods applied to Machine Learning. arXiv:2104.02466 [cs], April 2021.

[UMC06] E. Cosatto B. Flepp U. Muller, J. Ben and Y.L. Cun. Off-roadobstacle avoidance through end-to-end learning.advances in neuralinformation processing system. NIPS, pages 739–746, 2006.

[XKN22] Xuan Xie, Kristian Kersting, and Daniel Neider. Neuro-symbolic verification of deep neural networks. In Lud De Raedt, editor, Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, pages 3622–3628. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[XRA21a] D4.1 - ato over etcs goa2 specification. Technical report, X2RAIL1, 2021.

[XRA21b] D4.3_-_aoe_goa3_4_preliminary_specification. Technical report, X2RAIL1, 2021.

[XWZL16] Xi Xiong, Jianqiang Wang, Fang Zhang, and Keqiang Li. Combining deep reinforcement learning and safety based control for autonomous driving. arXiv preprint arXiv:1612.00147, 2016.

[YMZ+17] Changkun Ye, Huimin Ma, Xiaoqin Zhang, Kai Zhang, and Shaodi You. Survival-oriented reinforcement learning model: An effcient and robust deep reinforcement learning algorithm for autonomous driving problem. In International Conference on Image and Graphics, pages 417–429. Springer, 2017.

[ZC17] Jiakai Zhang and Kyunghyun Cho. Query-efficient imitation learning for end-to-end simulated driving. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California, USA., pages 2891–2897, 2017.

[ZT17] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. CoRR, abs/1711.06396, 2017.

[ZYW17] Y. Li Z. Yang, F. Zhou and Y. Wang. A novel iterative learning path-tracking control for non holonomic mobile robots against initial shifts. International Journal of Advanced Robotic Systems, 2017.

A annex A

Basic functions of train operation		GoA0	GoA1	GoA2	GoA3	GoA4
		On-sight train operation	*Non-automated train operation*	*Semi-automated train operation*	*Driverless train operation*	*Unattended train operation*
Ensuring safe movement of trains	*Ensure safe route*	X (points command/control in system)	system	system	system	system
	Ensure safe separation of trains	X	system	system	system	system
Driving	*Ensure safe speed*	X	X (partly supervised by system)	system	system	system
	Control acceleration and braking	X	X	system	system	system
Supervising guideway	*Prevent collision with obstacles*	X	X	X	system	system
	Prevent collision with persons on track	X	X	X	system	system
Supervising passenger transfer	*Control passengers' doors*	X	X	X	X	system
	Prevent person injuries between cars or between platform and train	X	X	X	X	system
	Ensure safe starting conditions	X	X	X	X	system
Operating a train	*Set in/ set off operation*	X	X	X	X	system
	Supervise the status of the train	X	X	X	X	system
Ensuring detection and management of emergency situations	*Perform train diagnostic*	X	X	X	X	System and/or staff in OCC

Figure 12: Train operation for the different grades of Automation. D3.2 Automatic Train Operations: implementation, operation characteristics and technologies for the Railway field 1.2 - 28/01/2019