Note by the secretariat
(19 May 2023: Revision includes a description of the figure in chapter 4)

Informal document **GRVA-13-04/Rev.1**
13th GRVA, 23-27 May 2022
Provisional agenda item 3

Outcome of the GRVA workshops on Artificial Intelligence and Vehicle Regulations

I. Mandate

1. Following the AC.2 decisions of November 2020 and the discussions at the last sessions of GRVA, GRVA requested the secretariat to organize a technical workshop focusing primarily on definitions for Artificial Intelligence, relevant for GRVA activities. The first workshop took place on 18 March 2022. The experts agreed to convene a second workshop on 9 May 2022 to explore the AI use cases and their relevance for GRVA with regards to safety.

2. The experts discussed whether technology neutral performance requirements are sufficient for the purpose of GRVA or if specific provisions would be necessary. The experts developed draft definitions, drafted a table with use cases and their relevance with regards to vehicle regulations and reflected on the potential activities that could be necessary in the framework of the New Assessment Test Method developed by GRVA and its IWG on Validation Method for Automated Driving (VMAD).

II. List of AI relevant definitions in the context of vehicle regulations

3. The terms below are inspired by the definitions under review at the International Standard Organization (see ISO/IEC 22989).

[4. **Artificial intelligence** is a set of methods or automated entities that together build, optimize and apply a model so that the system can, for a given set of predefined tasks, compute predictions, recommendations, or decisions.

5. **Machine learning** is a data based computational techniques to create an ability to "learn" without an explicitly programmed algorithm such that the model's behaviour reflects the data or experience.

6. **Machine learning model** is a mathematical construct that generates an inference, or prediction, based on input data.

7. **Deep learning** is an approach to creating rich hierarchical representations through the training of neural networks with many hidden layers.

8. **Supervised learning** is a type of machine learning that makes use of labelled data during training.

9. **Unsupervised learning** is a type of machine learning that makes use of unlabelled data during training.

10. **Reinforcement learning** is a type of machine learning utilizing a reward function to optimize a machine learning model by sequential interaction with an environment.

11. **Dataset** is a collection of data with a shared format and goal-relevant content.

12. **Data sampling** is a process to select a subset of data samples intended to present patterns and trends similar to that of the larger dataset being analysed.

13. **Data annotation** is the process of attaching a set of descriptive information to data without any change to that data.

14. **Training** is the process to establish or to improve the parameters of a machine learning model, based on a machine learning algorithm, by using training data.

15. **Retraining** is an approach to creating rich hierarchical representations through the training of neural networks with many hidden layers.

16. **Continuous learning** describes incremental training of an AI system throughout the lifecycle to achieve defined goals governed by pre and post operation risk acceptance criteria and human oversight.

17. **Self-learning** describes incremental training of an AI system throughout the lifecycle to achieve defined goals governed by pre and post operation risk acceptance criteria making possible a continuous activation of the new system output with or without human oversight.

18. **Online learning** describes incremental training of a new version of the AI system during operation to achieve defined goals based on post operation acceptance criteria and human oversight without activating the new system output until released.

19. **Human oversight** is AI system property guaranteeing that built-in operational constraints cannot be overridden by the system itself and is responsive to the human operator, and that the natural persons to whom human oversight is assigned.

20. **AI lifecycle** consists out of the design and development phase of the AI system, including but not limited to the collection, selection and processing of data and the choice of the model, the validation phase, the deployment phase and the monitoring phase. The life cycle ends when the AI system is no longer operational.

21. **Safe-by-design** is system property enabled by development and lifecycle activities to claim system measures bring risks to an acceptable level.

22. **Trustworthiness** is the ability to meet stakeholders' expectations in a verifiable way.

23. **Bias** is a systematic difference in treatment of certain objects, people, or groups in comparison to others.

24. **Fairness / Fairness matrix** is a way of describing bias.

25. **Predictability** is a property of an AI system that enables reliable assumptions by stakeholders about the output.

26. **Reliability** is a property of consistent intended behaviour and results.

27. **Resilience** is the ability of a system to recover operational condition quickly following an incident.

28. **Robustness** is the ability of a system to maintain its level of performance under any circumstances.

29. **Transparency of an organization** is the property of an organization that appropriate activities and decisions are communicated to relevant stakeholders in a comprehensive, accessible and understandable manner.

30. **Transparency of a system** is property of a system to communicate information to stakeholders.

31. **Explainable** means a property of an AI system to express important factors influencing the AI system that results in a way that humans can understand.

32. **Black/Grey/White box [testing]** are [tests of] systems / software in which functionality are unknown / partially know / known.]

III. AI use cases in the automotive sector

Note: The following table was prepared by the experts from CLEPA and OICA

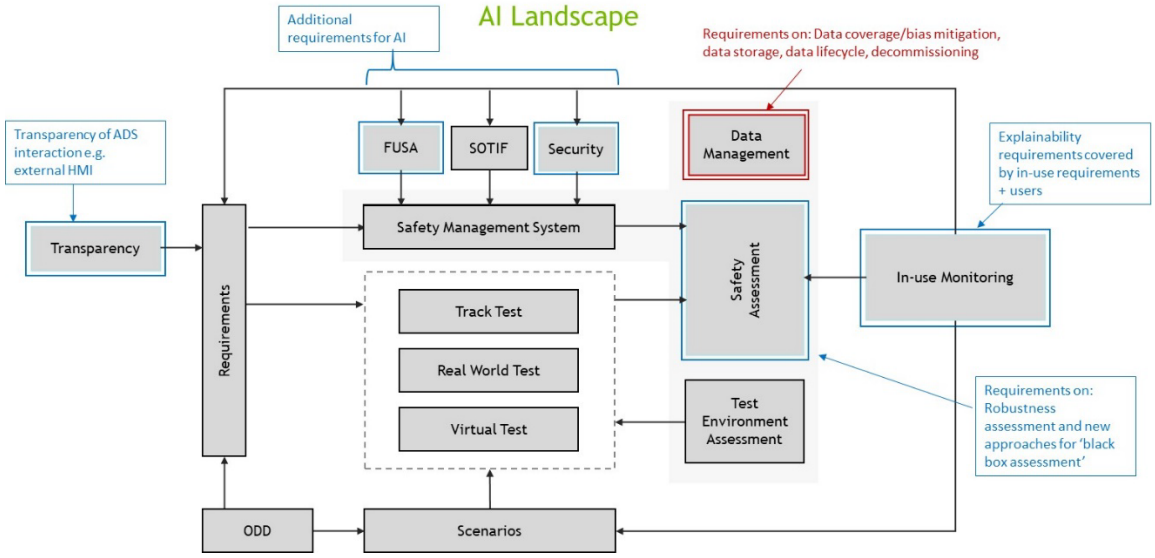
An editable version of this table is available here: https://unece.org/transport/events/grva-technical-workshop-artificial-intelligence-2nd

	AI Application		Non Safety functions e.g. Infotainment Out of Scope of type approval	Safety functions			
				Driving Function			Non Driving Functions
				Perception	Planning	Actuation	
Conventional Software	**Artifical Intelligence (AI)**	Artificial Intelligence is a set of methods or automated entities that together build, optimize and apply a model so that the system can, for a given set of predefined tasks, compute predictions, recommendations, or decisions	Natural language processing	Out of Scope [Non-AI] Detection of other road users for AEBS, ACC Detection of road infrastructure for LDW, LKAS	Out of Scope Activation of FCW and AEBS based on ego vehicle position and other road users	Not Applicable	Out of Scope Detection of driver's face for ID (under conditions ensuring privacy)
Artificial Intelligence	**Supervised Learning (SL)**	Supervised learning is a type of machine learning that makes use of labelled data during training	Gesture control Voice Recognition	Detection of other road users for AEBS, ACC Detection of passive road infrastructure for LDW, LKAS	Trajectory prediction using drivable path prediction from labelled data (e.g. HD maps)	Not Applicable	Detection of drivers eye gaze / state for DMS Fault detection, Predictive Maintenance
	Unsupervised Learning (UL)	Unsupervised learning is a type of machine learning that makes use of unlabelled data during training		Streamlining data labelling process for less safety critical systems like ISA. Extracting scenarios from real world data to suport validation Generation of synthetic data for supervised learning / distortion of real world data	Trajectory prediction using Kalman filters, KalmanNet or Gaussian Process architectures, or other architectures	Not Applicable	[?]
	Semi Supervised Learning (SSL)	Semi supervised learning is a technique that "learns" from a mix of labelled data and data that is both un-labelled and unstructured. They build on a small set of known exemplars and then use this information to guide unsupervised learning.		Streamlining data labelling process for less safety critical systems like ISA.	Shadow mode' used in development for training control algorithms	Not Applicable	[?]
	Reinforcement Learning (RL)	Reinforcement learning is a type of machine learning utilizing a reward function to optimize a machine learning model by sequential interaction with an environment		Some manufacturers are starting to use RL for perception, could potentionally be used in cooperative perception in the future.	Lane Centering or ACC systems may use RL due to the reduction in cost / data required to train the system	Not Applicable	Predictive Maintenance

IV. Impact of Artificial intelligence on the New Assessment Test Method

Note: The following table was prepared by the experts from CLEPA and OICA.

An editable version of this table is available here: https://unece.org/transport/events/grva-technical-workshop-artificial-intelligence-2nd



Context and description of the figure above as presented by its author.

1. Introduction

Recent achievements and communications give the impression that the switch from conventional software to Artificial Intelligence (AI) and Machine Learning (ML) in automotive products would happen overnight and that suddenly all the software modules onboard a vehicle would be using AI ML algorithms. This isn't exactly the case.

The introduction of AI and Machine Learning in vehicles is expected to be a slow and steady journey that leads to the introduction of machine learning into an Automated Driving System (ADS) or an Advanced Driver Assistance System (ADAS), starting off with a few software modules.

To date, the use of AI and machine learning is more focused on perception algorithms. But then, as more confidence is attained in these types of algorithms, AI could be used for control algorithms and decision logic. The use of machine learning for control algorithms could be challenging, though, as there are hard sets of requirements for functional limits that ADS need to comply with. Machine learning algorithms make it hard to control compliance to those hard sets of functional boundaries like having a prescribed lateral acceleration limit for a lane keeping system not exceeding three meters per second squared or having a certain deceleration rate for Advanced Emergency Braking System (AEBS). So, this evolution from using conventional software over

to using machine leaning is going to be a slow process and something that industry will implement carefully.

The progressive introduction of AI and machine learning into vehicles requires to identify the potential elements that would be missing in the regulatory frameworks applicable to automotive systems. The policy makers defining best practices and horizontal requirements for AI based systems expect that certain aspects are duly taken into account in the various industry regulatory verticals.

The present document describes how safety is assessed for Automated Driving Systems and explores the compatibility of existing technology neutral provisions, drafted so far by the Working Party on Automated/Autonomous and Connected Vehicles (GRVA) for the assessment of ADS with the use AI and machine learning algorithms within their system itself.

2. Premarket assessment of automotive products

A. The conventional type-approval system

The most common pre-market assessment in the automotive sector is Type Approval. It evolved to assess complex electronic systems, then Advanced Driver Assistance Systems and more recently Automated Driving Systems at level three. At EU level, Type Approval can be granted for level four type systems.

Figure 1
Conventional type approval system

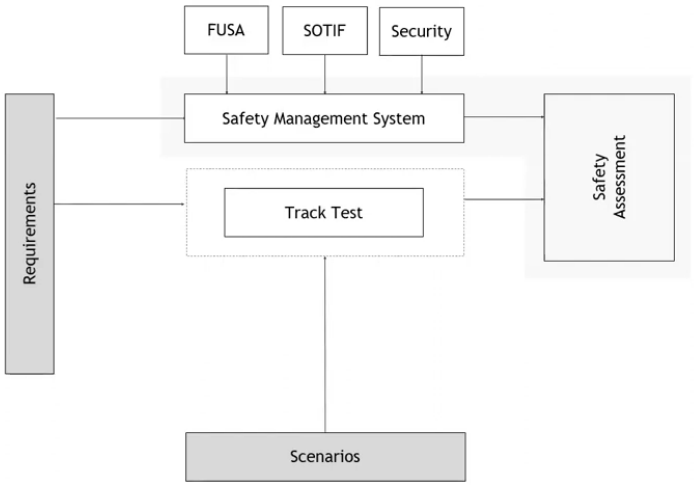


Figure 1 describes the conventional Type Approval including the Type Approval of Complex Electronic Systems such as a Lane Keeping Assist System and other ADAS.

Type approval regulations include requirements and fixed scenarios included in regulations for testing purposes.

To demonstrate that the requirements are met within the prescribed scenarios, track tests are performed to generate performance related data that feed into the safety assessments.

Audits are carried out as part of the assessment looking at the performance of the system and the safety management system (supported by Functional Safety (FUSA), Safety Of The Intended Function (SOTIF) and Security considerations) to verify whether the manufacturer has processes in place to ensure a safe and robust system. Audits check that vehicle manufacturers follow the principles that are defined within FUSA, SOTIF and security standards, and check if the system would just operate safely in nominal conditions or in fault conditions, too.

B. Recent evolutions of the type approval system.

1. UN Regulations Nos. 155 and 156

UNECE's WP.29 adopted UN Regulations Nos. 155 (Cyber Security and Cyber Security Management System) and 156 (Software Update and Software Update Management System).

These two regulations introduced novelties in the type approval regime.

UN Regulation No. 155 addresses security related audits that look at the manufacturer's ability to mitigate risks related to adversarial attacks that may occur onboard the vehicle but also at the manufacturer's plant or in the cloud, where some of the AI training or pure software algorithms may be performed. For this purpose, UN Regulation No. 155 introduces a novel approach related to the assessment of cyber security for automotive products. It includes extensive auditing requirements related to the manufacturer's management capability of cyber security, and the verifications. This includes that the vehicle types, for which approvals are granted, are designed and produced in line with an audited management system. In addition, the regulation introduces new notions which were not anchored in the type approval system so far, i.e. the lifecycle requirements for a vehicle type and the life time requirement for a vehicle itself.

UN Regulation No. 156 introduces requirements for software updates including type approval relevant software updates. This is an essential element that extends the type approval system beyond the pre-market approval.

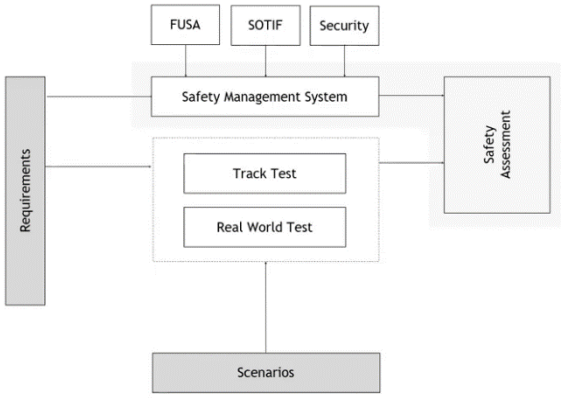
2. UN Regulation No. 157

UN Regulation No. 157, adopted in June 2020, includes the requirements for the type approval of Automated Lane Keeping System (ALKS). It is de facto the first international type approval regulation for ADS (with a limited operational domain). It allows for the deployment of level three systems in the territory of the UNECE contracting parties (signatories to UN Regulation No. 157 annexed to the 1958 Agreement). It is fundamentally similar to conventional type approval.

The regulation includes partially fixed (i.e. the some requirements and scenarios are mentioned without being detailed and linked to simple pass/fail requirements) requirements and scenarios. The safety assessment is performed through an audit and tests. Simulation may be used to support the safety assessment.

Figure 2

The type approval system as required by UN Regulation No. 157 (ALKS)



The main addition introduced in UN Regulation No. 157 to the conventional type approval is the introduction of real world testing. There were real world testing precedents, e.g. for Traffic Sign Assist in the European General Safety Regulation. This real-world testing for level three ADS differs from the said precedents as it is aimed to ensure that the system itself complies with road traffic rules across the entire Operational Domain Design (ODD), where the system subject to approval will be operating in. UN Regulation No. 157 includes fixed requirements and fixed scenarios, but there is also an expectation that manufacturers have to do an analysis of their ODD to understand in the relevant context which scenarios may be encountered in their specific use case and which road traffic rules are applicable.

The deployment of a level three system in Germany on a highway, compared to the UK or another market, will face some nuances related to traffic signs and road traffic rules and specificities such as the evacuation corridor that is required on motorways in Germany. Using real world testing could be a very valuable tool to demonstrate that there is a sufficient level of safety and to check that the requirements are met. The regulation allows for some flexibility as to what requirements and what scenarios may be used to support the assessment of these level 3 systems. But the building blocks, i.e. the safety assessment and safety management system remain the same.

It is expected that the work done by the Informal Working Groups (IWGs) on Functional Requirements for Automated Vehicles (FRAV) and Validation Method for Automate Driving (VMAD) will be taken in to account in further updates of UN Regulation No. 157 and that these updates will support the deployment of ADS.

3. Self certification

Several countries don't rely on pre-market approvals but implement a self-certification system for the assessment of vehicle safety.

In these countries, manufacturers are required to comply with detailed safety standards and regulations (with precise testing procedure ensuring repeatability and low variability also when different stakeholders perform the same tests),

which are enforced by government agencies. Manufacturers must conduct their own testing to check that their vehicles meet the required safety standards. To ensure that manufacturers comply with these regulations, governments may also conduct their own testing and inspections of vehicles already on the market. They may also rely on consumer complaints and feedback to identify potential safety issues with specific models. They may order recalls in case of noncompliance.

Nascent technologies as well as their evaluation criteria may not be mature enough to be covered by robust standards and regulations with clear pass/fail criteria (ensuring repeatability) as needed for self-certification.

Most countries applying self-certification issue guidelines regarding ADS that guide industry. This approach, on the one hand, does not provide the safe harbour that regulations may provide, on the other handit does provide more flexibility for new technologies, as regulation may stifle innovation.

3. The new assessment method for ADS

The Working Party on Automated/Autonomous and Connected Vehicles, through its IWG on VMAD, developed a validation system for all ADS and all use cases. The outcome produced is what is described as the New Assessment / Test Method (NATM), the new assessment methods for automated driving systems.

The principles of the NATM were also adopted in the recently adopted EU Regulation for the small series of Level 4 and Automated Valet Parking.

The NATM extends the compliance assessment beyond the pre-market approval. It requires the manufacturers to report in-use data to authorities.

Figure 3

The New Assessment / Test Method description

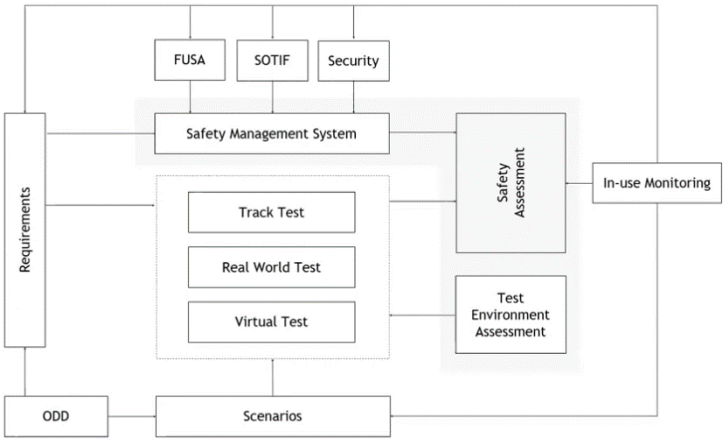


Figure 3 includes a few more components added to Figure 2 , This extended system forms the NATM designed for evaluating ADS safety. They include ODD, virtual testing, test environment assessment and in use monitoring. Due to the complexity of the ODD, the assessment is derived from requirements and scenarios relevant for their use case from the analysis of the ODD.

A. ODD

The bottom left-hand side of figure 3 shows the ODD as the starting point of the assessment. The ODD analysis leads to the derivation of requirements and scenarios.

All the possible combinations of ODD variations and of markets, where to deploy the different levels of functionality and different levels of automation, represent a variety of scenarios and corresponding requirements that can't be captured in an exhaustive list of requirements and scenarios. Therefore, the ODD analysis plays a central role and explains the flexibility provided by the NATM.

B. Virtual testing

The NATM provides the possibility to conduct certification relevant tests using virtual testing environments.

C. Test environment assessment

If manufacturers do use synthetic data to demonstrate that they meet the requirements relevant for the ODD, then they have an obligation (according to NATM) to demonstrate that the synthetic environment accurately represents the real world. Therefore, the NATM includes a very robust set of processes that help the manufacturer structure the evidence they have to put together to show that this synthetic environment is accurate enough for the specific purpose that it is designed for. The test environment assessments feed into the validation of the synthetic environment.

D. In-use monitoring

Another element added in NATM for the level three / level four evaluation is the concept of "in-use monitoring". It aims to demonstrate that the system remains dutiful and that there aren't any issues due to data drift or concept drift of these AI / ML algorithms (or conventional software) being used in the real world. There needs to be some way of reporting to show that manufacturers meet certain safety targets and to show that they are reacting to any previously unforeseen risks and helping mitigate those risks by taking corrective or restrictive measures.

The concept of in-use monitoring is new in vehicle regulations and in automotive products. It might become the cornerstone for the specific assessment of AI-based technologies employed in ADS.

3. How to assess AI ?

For the purpose of this chapter, the NATM, which aims to be technologically neutral, is assessed with regards to the draft EU AI act methodology defining how to assess high risk AI systems (a similar exercise could be done in comparison with similar reference documents from UNESCO or OECD - see annex), checking: (a) Accuracy, robustness and cybersecurity, (b) Transparency,

(c) Record keeping, (d) Technical documentation, (e) Data and Data governance, (f) Human oversight. The purpose of this comparison is to check whether the automotive products assessment (i.e. the automated driving systems assessment) is robust and matches the expectations for high risk AI systems.

It is noted that some of the EU AI Act requirements would be addressed by the functional requirements developed by the IWG on FRAV, e.g. in terms of technical documentation, item (d), and human oversight, item (f), or by the IWG on Event Data Recorder (EDR) / Data Storage System for Automated Driving (DSSAD) for record keeping, item (c).

Therefore, the following reviews how NATM can be used for the assessment of AI based ADS regarding items (a), (b) and (e).

Figure 4

AI relevant considerations and NATM

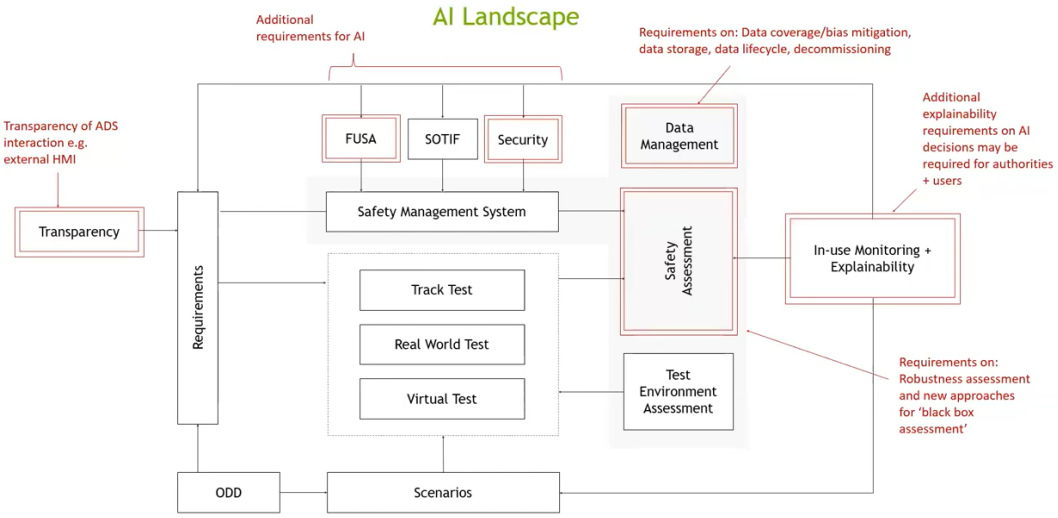


Figure 4 highlights in red the areas that may be AI specific and should address expectations to high risk AI systems.

A. Trustworthy AI

Trustworthy AI means a number of principles that are derived from ethical norms from legal frameworks, it sets requirements / obligations on manufacturers to design a system that is secure, respects privacy laws, is explainable, transparent and safe. All these elements have already been addressed by the automotive industry and their respective regulatory regimes.

B. Explainability

In NATM, the concept of explainability can take two forms:

1. Real time monitoring

For autonomous systems, the concept of having real time explanations of why a decision was made when these automated systems are making 1000s of

decisions a second may not be very useful as it could represent a lot of data that would be very uninterpretable.

2. Pre-market verifications and elements supporting investigations

The way explainable AI can be addressed in the automotive sector could be as follow: understanding of the way AI functions can be reached at the time of type approval by auditing how a system was designed, developed and validated. One could also include datasets review to the audit that were used for training, testing and validation. This can help explain how AI makes a decision or what was an issue with an AI system in the case of a critical event.

This approach can be supplemented by EDR and DSSAD, which are existing obligations for manufacturers to collect specific data in the event of a relevant event as they produce data that can leverage documentation as part of a Type Approval to explain why a AI system made a certain decision. In any case EDR/DSSAD could be used, also upon request by authorities, by manufacturers post process all collected data to come up with the explanation as to why these decisions were made.

C. Robustness and black box

Two considerations are often mentioned in the context of AI assessment: robustness and black box. AI systems based on ML and deep neural networks can be complex, they can be difficult to explain. One could argue to some extend that conventional software is also difficult to explain when software has millions of lines of code: in this context it can be difficult to explain certain functionality.

The NATM is a scenario-based approach for demonstrating that the system will remain safe in all foreseeable scenarios. This raises the question of the need to understand exactly why everything is working within a deep neural network or understand the perception algorithm, if the system has been validated and have shown it works correctly in its ODD. This is why the very robust analysis of the ODD is important such that one can support the blackbox assessment to demonstrate safety in a discrete (reasonably high) number of scenarios.

D. Transparency

Transparency can take two forms: transparency in the design and development of a system and transparency of the system.

The type approval framework covers the transparency of the design and development of the system through the audit and documentation requirements regarding the design of the system.

The transparency of the system relates to whether the system is active and how it is interacting with other road users.

GRVA and GRE are discussing external Human Machine Interface (HMI) to show the status of an automated driving system. GRVA discussed the pros and cons for doing that.

HMI inside the vehicle is a matter of transparency as well. The system can be considered as transparent if the operator or people inside the vehicle receive

information on why which manoeuvres may be occurring. Features such as the so-called confidence view, where an image of the vehicle is provided in the dashboard or head unit with a representation of the other road users around the vehicle, can provide transparency and confidence that the automated driving system is seeing the world as one expects it to.

E. Functional safety and security principles

Functional Safety concepts are not expected to fundamentally change for AI based systems. Development processes of safe AI ML systems are also expected to remain relatively similar to those of existing systems. However, there are specific AI related risks that need to be addressed. Some ISO standards are in development helping to flag what those relevant risks are and to provide appropriate tools to mitigate those risks. The same applies for security.

Very robust regulations are in place, covering the cybersecurity management systems. The cyber security management system principles remain the same also in the case of AI systems.

However new attack points / vulnerabilities may need to be considered as well as how to mitigate risks.

Data sets are used for training AI machine learning models. These data sets need to be encrypted and secured during any transfer and at rest to prevent any data poisoning attacks or adversarial attacks to the data sets that will cause the system not to operate as expected in the real world.

A specific aspect that is slightly different and may need to be addressed in the future is regarding data management.

F. Data management

Data management is the only element missing in the NATM in its current version. Data management considerations will be of importance for having appropriate data life cycles, well integrated in development processes and anchored in functional safety assessment.

Appropriate data quality requirements will support the safe operation of the system. Data decommission considerations are necessary to avoid any violation of privacy laws, including requirements about data retention or about data minimization. ISO/IEC are producing important documents around data quality and data governance.