Submitted by the experts
from CLEPA and OICA

Informal document **GRVA-16-05**
16th GRVA, 22 - 26 May 2023
Provisional agenda item 3

[Initial points and Terms to support discussions around the use of AI ML in automotive industry]

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).

[III. Relevance for GRVA]

1. [This short chapter provides two examples aimed at suggesting that GRVA might have to look into Artificial Intelligence in the context of vehicle regulations.]

A. [Test results reproducibility according to UN GTRs and UN Regulations]

1. [GRVA develops technical requirements and guidance that are technology neutral, unless a specific technology requires appropriate and specific provisions.]

2. [GRVA discussed (GRVA-12-06) that in the case of functions, which are based on software that is using Artificial Intelligence, the outcome associated with this AI for a given situation will not necessarily be predictable.]

3. [The predictability of test results is an important factor for the type-approval and for the self-certification.]

B. Specific features of AI-based systems used in automotive products

1. [In the following the term AI based systems refers to connectionist AI systems such as neural networks which are trained using machine learning algorithms and data. These systems exhibit qualitatively new properties leading to new opportunities as well as to new challenges.]

2. AI-based systems, used in automotive products, may allow a trade-off of various desirable model characteristics (model drift, model staleness, model complexity,[flexibility], robustness, verifiability, etc.) while guaranteeing a certain level of safety and security. AI-based systems should provide possibilities for system updates.

3. GRVA might wish to evaluate whether the provisions regarding software updates (in UN Regulation No. 156 and in the recommendations on uniform provisions concerning cyber security and software updates) adequately address updates of AI-based systems.

4. AI-based systems can contribute to improve vehicle safety, with additional beneficial consequences on road safety, e.g. by allowing AD systems to predict currently unforeseeable behavior of other road users (e.g. detection of potential collision opponents).

5. [The use of AI and machine learning algorithms in type approved functions is limited for the time being. Even though, there are already well-established processes for how to test

software before and during deployment of an automotive product. For Machine learning algorithms, that leverage unsupervised learning, those processes might not be sufficient in the mid-term future. Software has to be tested prior to deployment in order to comply to all related Laws, Regulations (e.g. UN-R) and Policies.]

[This applies also to the update process. However, it needs to be evaluated in how far current regulatory provisions can sufficiently address the specific needs for testing and updating AI based software and guarantee its safe operation.]

IV. List of AI relevant definitions for discussions in the context of vehicle regulations

The terms below are largely derived from the definitions under review at the International Standard Organization (see ISO/IEC 22989). [The intention is to create a base set of definitions for discussions. It is not an exhaustive list of definitions in a regulatory meaning.]

- **[Agent** is anything which perceives its environment, takes actions autonomously in order to achieve goals, and may improve its performance with learning or may use knowledge.]
- **AI lifecycle** consists of the design and development phase of the AI-based system, including but not limited to the collection, selection and processing of data and the choice of the model and the training process, the validation phase, the deployment phase and the monitoring phase. The life cycle ends when the AI-based system is no longer operational.
- **Artificial intelligence (AI)** 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.
- **Bias** is a systematic difference in treatment (including categorization/observation) of certain objects (e.g. natural persons, or groups) in comparison to others.
- **Black box** is a system / software in which the detailed architecture and processing is unknown
- **Black/Grey/White box testing** are tests of systems / software in which architecture and processing is unknown / partially known / known.
- **[Connectionist AI (cAI)** systems usually consist of many nodes, called neurons, which are connected with each other in specific patterns, depending on the AI model at hand. Examples of cAI systems are neural networks and support vector machines. In many applications cAI systems are more powerful when compared to sAI systems, e.g. in computer vision. In the majority of cases parameters of cAI systems may not be directly set by the developer. Instead machine learning algorithms are used together with data to train these systems. The quality of the resulting cAI system is crucially dependent on the quality and quantity of the training data. In contrast to sAI systems cAI systems are in most cases not easily interpretable and not formally verifiable.]
- **Data annotation** is the process of attaching a set of descriptive information to data without any change to that data.
- **Data sampling** is a statistical process to select a subset of data intended to present patterns and trends similar to those of the larger dataset being analyzed.
- **Dataset** is a collection of data with a shared format and goal-relevant content.
- **Deep learning** is a process whereby neural networks use multiple layers of processing intended to extract progressively higher-level features from data.
- **Explainability** means a property of an AI-based system to express important factors influencing the system's outcome in a way that humans can understand.
- **Fairness / Fairness matrix** is a way of describing bias.
- **Grey box** is a system / software in which the detailed architecture and processing is partially known.

- **Human oversight** is an AI-based system property guaranteeing that built-in operational constraints cannot be overridden by the system itself and are responsive to the human operator, and that the natural persons to whom human oversight is assigned exert ultimate control.
- **Machine learning** is a collection of data-based computational techniques to create an ability to learn without following explicit instructions such that the model's behavior reflects patterns in data or experience.
- **Machine learning model** is a computer science construct that generates an inference, or prediction, based on input data.
- **Model drift** is a term from the field of machine learning. It refers to the phenomenon that the predictive accuracy of machine learning models can degrade over time. The reasons for this are, for example, that assumptions or variable dependencies that were still valid when the models were created and trained have changed over time. Measures such as retraining or tuning the models can eliminate model drift.
- **Model staleness** is defined as outdated if the trained model does not contain current data and/or does not meet current requirements. Outdated models can affect prediction quality in intelligent software.
- **Online learning** describes incremental training of a new version of the AI-based system during operation onboard production vehicles to achieve defined goals.
- **Predictability** is a property of an AI-based system that enables reliable assumptions by stakeholders about the output.
- **Reinforcement learning** is a discipline of machine learning that permits an agent to learn actions to be taken from patterns in data or experiences, optimizing a quantitative reward function gained along the time.
- **Reliability** is a property of consistent intended behavior and results.
- **Resilience** is the ability of a system to recover operational condition quickly following an incident.
- **Robustness** is the ability of a system to maintain its level of performance under a wide range of circumstances.
- **Safe-by-design** is a system property enabled by proactive development and lifecycle activities to ensure that risks are brought to an acceptable level through system measures.
- **Semi Supervised learning** is a combination of supervised and unsupervised learning. It uses a small amount of labeled data and a large amount of unlabeled data, which provides the benefits of both unsupervised and supervised learning while avoiding the challenges of finding a large amount of labeled data.
- **Software** is usually created by a process called traditional programming. The programmer manually codes rules using a programming language.**Supervised learning** is a type of machine learning that makes use of labelled data during training.**[Symbolic AI (sAI)** explicitly encodes knowledge using symbolic representations. An example of such a system is a decision tree. Interpreting and formally verifying a sAI system is generally possible and much easier to achieve when compared to connectionst AI systems.]
- **Training** is the process to tune the parameters of a machine-learning model.
- **Training data** is a subset of input data samples used to train a machine learning model
- **Transparency of an organization** is the property of an organization that appropriate activities and decisions are documented and communicated to relevant stakeholders in a comprehensive, accessible and understandable manner.
- **Transparency of a system** is property of a system to communicate information to stakeholders.
- **Trustworthiness** is the ability to meet stakeholders' expectations in a verifiable way.

- **Unsupervised learning** is a type of machine learning that makes use of unlabeled data during training.
- **Validation** is done to ensure software usability and capacity to fulfill the customer needs.
- **Validation data** is data used to assess the performance of a final machine learning model
- **Verification** is done to ensure the software is of high quality, well-engineered, robust, and error-free without getting into its usability.
- **White box** is a system / software in which the detailed architecture and processing is known.

V. AI use cases in the automotive sector

Type of AI	Type of Machine Learning	Non Safety functions	Safety functions			
			Driving Function			Non Driving Functions
		Out of Scope of type approval	Perception	Planning	Actuation	
Non -AI Soft-war e	None		Out of Scope (in the context of this document)	Out of Scope (in the context of this document)	Out of Scope (in the context of this document)	Out of Scope (in the context of this document)
Symbolic AI	None or any type of ML	e.g. Infotainment, Natural language processing	e.g. Detection of other road users for AEBS, ACC, Detection of road infrastructure for LDW, LKAS	e.g. Activation of FCW and AEBS based on ego vehicle position and other road users	Currently not applicable	e.g. Detection of driver's face for ID (under conditions ensuring privacy), alcohol breath detection controlling immobilizer
Connectionist AI + Machine Learning	Supervised Learning (SL)	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)	Currently not applicable	Detection of drivers eye gaze / state for DMS Fault detection, Predictive Maintenance
	Unsupervised Learning (UL)		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	Currently not applicable	fault detection (unsupervised anomaly detection)
	Semi Supervised Learning (SSL)		Streamlining data labelling process for less safety critical systems like ISA.	Shadow mode' used in development for training control algorithms	Currently not applicable	
	Reinforcement Learning (RL)		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	Currently not applicable	Predictive Maintenance

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

[A current block diagram that shows the extent to which test methods for future and highly complex systems can be influenced by AI can be found in the NATM document (New Assessment and Test Methods).]