Note by the secretariat
(reflecting the amendments proposed
during the session)

Informal document **GRVA-18-04**
18th GRVA, 22-26 January 2024
Agenda item 3

# Proposal for amendments to ECE/TRANS/WP.29/2024/34

# Proposal for a draft guidance document on Artificial Intelligence in the context of road vehicles

The Contracting Parties to the 1958 and the 1998 Agreements, participating in the Working Party on Automated/Autonomous and Connected Vehicles,

*Having recognized* the significant penetration of some Artificial Intelligence (AI) in wheeled vehicles covered in the scope of the agreements administered by the World Forum for Harmonization of Vehicle Regulations (WP.29),

*Having noted* that industry currently could use machine-learning tools to support the development and/or testing of software before deployment,

*Having discussed* the technical fundamental aspects of some Machine Learning systems in automotive products, to which the general public refers to as Artificial Intelligence, and discussed corresponding definitions,

*Recalling* the adoption of Recommendations on uniform provisions concerning cyber security and software updates

*Having assessed* the importance of proper AI lifecycles for compatibility with existing certification regimes,

*Having acknowledged* that the use of such technology for automotive applications is still under development,

*Have agreed* on the following recommendations using AI-based algorithms within their automotive products:

**Software update**

1. This guidance document applies to certification requirements and Conformity of Production. Industry shall not issue software updates, which will significantly modify already certified functions according to the recommendations on uniform provisions concerning cyber security and software updates without resuming the relevant certification procedure.

2. It is recommended that after having trained an AI-system which is incorporated in the software it should be validated by authorised parties and or certification processes and assessed with regards to safety, security and environmental performance and other relevant requirements. Non-Certified systems containing AI, shall not influence certified systems in a way it harms the certification. Following that process, the validated software may be deployed in vehicles of a vehicle type.

**Data to be used for AI based system development**

3. It is assumed that data protection and privacy regulations, and other legal requirements are fully respected. This document is without prejudice to existing market-specific legislations and regulations concerning how personal data is collected and used. Where such regulations exist, they contribute to the overall safety of the AI system through setting personal data management safety standards.

# Annex 1

## Simplified definitions in the context of vehicles regulations - 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 and staleness, model complexity, robustness, verifiability, predictability and overfitting etc. while guaranteeing a certain level of safety and security. AI-based systems should provide possibilities for system updates.

3. Further recurrent evaluations might be necessary to check whether the provisions regarding software updates (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 behaviour 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 conventional software before and during deployment of an automotive product those processes might not be sufficient for AI based software. Software, whether it is created by machine learning or not, has to be tested prior to deployment in order to comply to all related Laws, Regulations, Recommendations 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.

6. The terms below are largely derived from the definitions at the International Standard Organization (see ISO/IEC 22989) SAE Ground Vehicle AI Committee and Organisation for Economic Co-operation and Development (OECD). The list of terms is not exhaustive, the definitions provided are simplified and may therefore not be suitable for the purpose of regulation.

7. It is customary to conduct extensive testing on White/Grey/Black box systems to ensure safe functionality of the certified system.

Note: some definitions below were taken from (or influenced by) OECD[1], SAE International[2] or ISO[3].

- **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)[3]** 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.
- **AI system[1]** is a machine-based system that is capable of influencing the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives. It uses machine learning and/or human-based data and inputs to perceive real and/or virtual environments; abstract these perceptions into models through analysis in an automated manner (e.g., with machine learning), or manually; and use model inference to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy.
- **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.
- **Conventional software** is usually created by a process called traditional programming. The programmer manually codes rules using a programming language.
- **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 analysed.
- **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 behaviour 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** is a physical, mathematical, or otherwise logical representation of a system, entity, phenomenon, process or 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 to achieve defined goals. Those parts of the system being subject to Online Learning must be decoupled from the vehicle's actuator system to ensure that the new version of the AI based system has no impact on the systems as long as it has not been tested for compliance with the relevant safety regulations.
- **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 behaviour 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. This includes the ability of a system to cope with natural and malicious perturbations within the systems input space.
- **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 labelled data and a large amount of unlabelled data, which provides the benefits of both unsupervised and supervised learning while avoiding the challenges of finding a large amount of labelled data.
- **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 connectionist 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 unlabelled data during training.
- **Validation** is done to ensure software usability and capacity to fulfil 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.

# Annex 2

## Review of use cases in vehicles provided by Industry

| 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 | |
| Conventional Software | **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,<br>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<br>(under conditions ensuring privacy), alcohol breath detection controlling immobilizer |
| Connectionist AI + Machine Learning | **Supervised Learning (SL)** | Gesture control<br>Voice Recognition | Detection of other road users for AEBS, ACC<br>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<br>Fault detection, Predictive Maintenance |
| | **Unsupervised Learning (UL)** | | Streamlining data labelling process for less safety critical systems like ISA.<br>Extracting scenarios from real world data to suport validation<br>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 |

Note: Shaded cells indicate items to be out of scope with respect to this document.]

# Annex 3

## Impact of Artificial intelligence on the New Assessment Test Methods

The figure below provides a holistic, schematic overview of the interaction of the individual pillars, scenarios, safety requirements of highly complex systems such as those used in automated driving. Reference is made to the UNECE Master Document New Assessment/Test Method for Automated Driving (NATM) in its current version. AI-based systems can also be designed in a similar way.

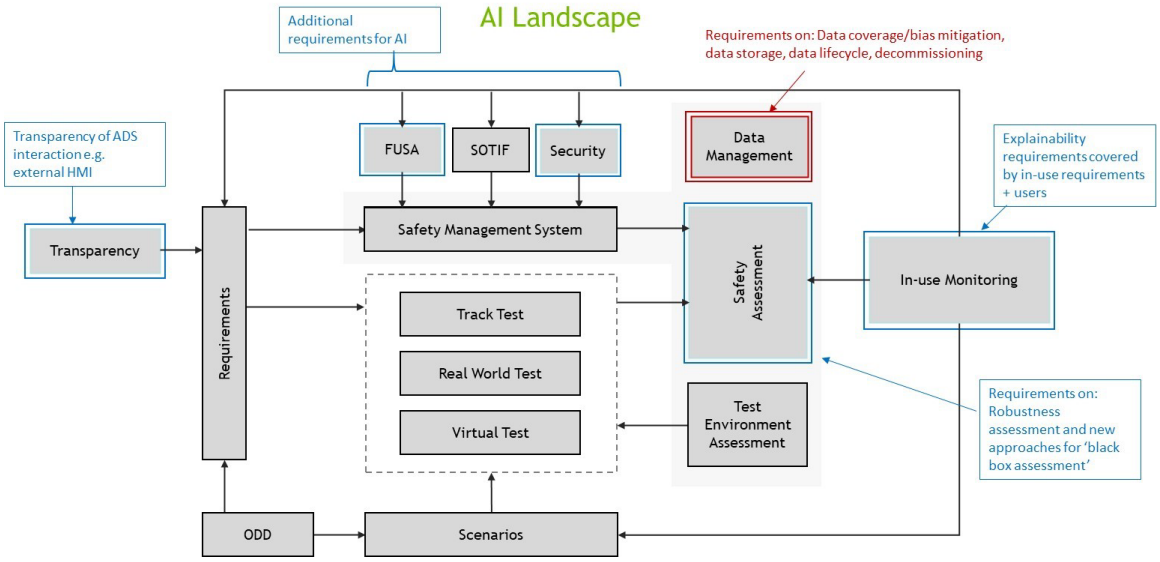


*Figure 1: Example scheme*

Note by the secretariat:

The figure above shows a representation of the NATM, in grey with two additional boxes (Transparency and Data Management). The text in color shows how the NATM can apply to AI based systems can help assessing key principles applicable to AI such as transparency, expandability and robustness.