Workshop 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, 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 (see decision 4 of the list of decisions of the 12th GRVA session).

II. Relevance for GRVA

2. 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 ad UN Regulations

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

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

5. The reproducibility of test results is an important factor for the type-approval and for the self-certification.

B. Specific features of AI systems used in automotive products

6. AI systems, used in automotive products, may provide the possibility for offline retraining combined with a thorough validation and Over-the-Air (OTA) updates. This offers a compromise that allows adaptations to model drift and model staleness processes while guaranteeing a certain level of safety and security.

7. 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 retraining and OTA updates.

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

8. The terms below are taken from the definitions under review at the International Standard Organization (see ISO/IEC 22989).

[9. **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.

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

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

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

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

14. **Unsupervised learning** is a type of machine learning that makes use of unlabeled data during training.

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

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

17. **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 analyzed.

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

19. **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.

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

21. **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.

22. **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.

23. **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.

24. **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.

25. **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.

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

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

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

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

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

31. **Reliability** is a property of consistent intended behavior and results.

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

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

34. **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.

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

36. **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.

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