COMMENTARY TO AI ACT

INTRODUCTION

The key goal of the AI Act is to ensure that AI systems placed on the EU market "are safe and respect existing law on fundamental rights and existing values"[1]. The proposal lays down a risk methodology for "high-risk" AI systems, which "have to comply with a set of horizontal mandatory requirements for trustworthy AI and follow conformity assessment procedures"[2].

With respect to "AI systems that relate to products that are covered by relevant Old Approach legislation (e.g., aviation, cars), the AI Act does not directly apply; however, the ex-ante essential requirements for high-risk AI … will have to be taken into account when adopting relevant … legislation.[3]

For non-high-risk AI systems, only very limited transparency obligations are imposed, e.g., "flag the use of an AI system when interacting with humans"[4]. For high-risk AI systems, "the requirements of high-quality data, documentation and traceability, transparency, human oversight, accuracy and robustness, are strictly necessary"[5].

The EC "will establish a system for registering stand-alone high-risk AI applications in a public EU-wide database."[6] AI providers must "inform national … authorities about serious incidents or malfunctioning"[7].

Volkswagen AG very much welcomes the AI Act and its risk-based approach, and understands and supports the unique opportunity that the European Commission has in advancing the safe and trustworthy use of AI. We believe that this approach will give Europe a competitive advantage on the long-term development and deployment of AI systems.

Besides our below feedback we clearly indicate our support of the feedback to the AI Act given by the European Automobile Manufacturers Association (ACEA).

Out feedback includes the following.

DEFINITION OF SAFETY COMPONENT

In this regard, we ask clarification on the following aspects:

- Which AI systems can be considered high-risk safety components in motor vehicles and which ones are not?
- Which ones would fall in the scope of the AI Act (not type-approved)?
- How is a safety component characterised? A module embedded in a safety critical chain does not necessarily bears safety requirement.

[1] Explanatory memorandum 1.1, p3.
[2] Explanatory memorandum 1.1, p3.
[3] Explanatory memorandum 1.1, p3.
[4] Explanatory memorandum 2.3, p7.
[5] Explanatory memorandum 2.3, p7.
[6] Explanatory memorandum 5, p12.
[7] Explanatory memorandum 5, p12.

A precise definition and potential use-cases are required in the automotive area and will provide legal certainty.

CHAPTER 2: REQUIREMENTS FOR HIGH-RISK AI SYSTEMS

Article 10 – Data and data governance A questionable definition here is the necessity of the use of validation data sets, Article 10, 1:

> *High-risk AI systems which make use of techniques involving the training of models with data shall be developed on the basis of training, validation and testing data sets that meet the quality criteria referred to in paragraphs 2 to 5.*

While the principle of this is laudable, the concept of "validation and testing" data sets is not, for each and every parameter adaptation methodology, sensible. A "where applicable" would solve this issue.

Article 10, 3:

> *Training, validation and testing data sets shall be relevant, representative, free of errors and complete. They shall have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons on which the high-risk AI system is intended to be used. These characteristics of the data sets may be met at the level of individual data sets or a combination thereof.*

This restriction is impracticable or impossible. A data set typically cannot be guaranteed to be "free of errors", and can practically never guaranteed to be "complete"; furthermore, validation of such is not practicable. Also, having "the appropriate statistical properties" poses a chicken–egg problem, when we are dealing with statistical methods that exactly validate that. Additionally, errors or "statistical anomalies" may be beneficial to check or reduce overfitting.

We suggest a wording along the following lines: "Training, validation and testing data sets shall be relevant and representative. Errors in the data set shall be statistically negligible for the models that use these data sets. The statistical properties of the AI methods trained with the data sets have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons on which the high-risk AI system is intended to be used."

Article 10, 6:

> *Appropriate data governance and management practices shall apply for the development of high-risk AI systems*

The implication of "appropriate" is unclear.

Article 12 "Record Keeping",

> *§2. The logging capabilities shall ensure a level of traceability of the AI system's functioning throughout its lifecycle that is appropriate to the intended purpose of the system.*

The use of the word "lifecycle" is confusing, esp. for heterogeneous systems, or AI systems which are used as services. The corresponding logging costs and resource use may be very impactful.

§4:

> *… logging capabilities shall provide, at a minimum ... the input data for which the search has led to a match*

whereas this is applicable to systems for Biometric identification and categorisation of natural persons only. The idea is good, but then also requires logging of the state of the system that did the matching.

Additional argument: In the case where local storage capacity does not suffice, the data could not be saved. Bandwidth may not suffice to save all data, e.g. for image data.

Article 14, "Human oversight, 1:

> *High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use.*

The implication of "*effectively overseen*" is unclear. For instance, in the control of complex machines, e.g. in manufacturing, the effectiveness of human supervision may be detrimental.

Proposal: High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tool, that they can be interrupted by a natural person and therefore either give control of the action to the natural person or put themselves in a safe state where no harm is caused by the system.

Article 17, "Quality management system", 2, p54:

> *The implementation of aspects referred to in paragraph 1 shall be proportionate to the size of the provider's organisation.*

whereas §1 starts, "Providers of high-risk AI systems shall put a quality management system in place that ensures compliance with this Regulation". As what was mentioned for Article 10, the quality management system should not overburden the development process, rather focus on proving performance. Also, it is perhaps better to make that system proportionate to the complexity cq. impact of the AI system.

Article 43, 4, p. 65:

> *For high-risk AI systems that continue to learn after being placed on the market or put into service, changes to the high-risk AI system and its performance that have been pre-determined by the provider at the moment of the initial conformity assessment ... shall* ***not constitute a substantial modification****.*

This is laudable.

Article 52 – Transparency obligations for certain AI systems

§3: the definition of "manipulation" is only intuitively clear, and that unclarity may lead to different interpretations. A list of examples may address this.

Article 54 "Further processing", 1, p. 70:

> *In the AI regulatory sandbox personal data lawfully collected for other purposes shall be processed for the purposes of developing and testing certain innovative AI systems in the sandbox under the following conditions:*

> *(a) the innovative AI systems shall be developed for safeguarding substantial public interest in one or more of the following areas: (i) ... criminal offences prevention ...; (ii) ... public safety/health; (iii) ... environment protection ...*

Why is the further use of sandboxed personal data restricted to these areas? Why not, e.g., in traffic control systems? Please note that, e.g., tracking vehicle numbers or license plates is GDPR-relevant and R&D in that area suffers from such restrictions.

Chapter 3: Enforcement

Article 64 "Access to data and documentation", 1, p. 77:

> *1. Access to data and documentation in the context of their activities, the market surveillance authorities shall be granted full access to the training, validation and testing datasets used by the provider, including through application programming interfaces ('API') or other appropriate technical means and tools enabling remote access.*

This means a full storage of all data. E.g. for moving-average filters or time-series methods in general, this may lead to prohibitively large data storage capacity, and goes against reasonable rules for efficiency and the goals of the Green Deal. Such data is also very sensitive data which is not very helpful to understand the decision-making process of an AI. The data governance and right to access should be reduce to test data only.

> *2. Where necessary to assess the conformity of the high-risk AI system with the requirements set out in Title III, Chapter 2 and upon a reasoned request, the market surveillance authorities shall be granted access to the source code of the AI system.*

We note that this could be problematic to achieve in the case where 3rd-party pretrained models are used. Of such models the data may not be available for this scrutiny. Does this requirement therefore preclude the inclusion of 3rd-party pretrained models?

ANNEX I: ARTIFICIAL INTELLIGENCE TECHNIQUES AND APPROACHES REFERRED TO IN ARTICLE 3, POINT 1

> *(a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;*
>
> *(b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;*
>
> *(c) Statistical approaches, Bayesian estimation, search and optimization methods.*

The definition of AI as laid down in Annex I is very broad, as it includes almost all software; it may cover traditional control algorithms as well as any piece of software which is based on statistical approaches.

At the same time, it is applauded that the definition does not refer to undefined concepts such as "intelligence".

Probably the key issue is that we are talking about parameterised methods where the parameters are determined based on data, using non-convex parameter optimisation – that means that, there is no guaranteed optimal solution.

It is well understood that the definition is only relevant in combination with high-risk methodologies as defined in Annex II. It may be worth pointing that out explicitly.

There are two issues. First, the exemplary character of this definition is not to be underestimated. As the AI Act will serve as a basis for many other regulatory frameworks, it is likely that this definition will be adopted by those.

Second, the proposed definition risks to capture in the regulation also traditional software systems that process data and take decisions. These systems are already rigorously tested and covered by current legislation.

Finally, there is no definition of "data". This means that an every-day interpretation of the word needs to be used.

ANNEX IV: TECHNICAL DOCUMENTATION

REFERRED TO IN ARTICLE 11(1)

> *The technical documentation referred to in Article 11(1) shall contain at least the following information, as applicable to the relevant AI system:*
>
> 1. *A general description of the AI system including:*
>   (a) *its intended purpose, the person/s developing the system the date and the version of the system*;

It is problematic to determine "the person/s developing the system", esp. in large organisations. How do we handle copy/pasted code, for instance, or other 3rd party code where the individuals cannot be named? Naming of a responsible department should suffice.