**3GPP TSG-SA5 Meeting #158 *S5-247094***

Orlando, USA, 18 - 22 November 2024

**Source: Nokia, NTT DOCOMO, Huawei, Verizon**

**Title: Rel-19 pCR TR 28.858 Add solution and evaluation for ML explainability in inference**

**Document for: Approval**

**Agenda Item: 6.19.1**

# 1 Decision/action requested

***The group is asked to discuss and agree on the proposal.***

# 2 References

[1] 3GPP TR 28.858 “Study on Artificial Intelligence / Machine Learning (AI/ML) management Phase 2”.

# 3 Rationale

This pCR is to add a solution and an evaluation for the already agreed use case on ML explainability in inference during previous meetings.

# 4 Detailed proposal

### 5.1.13 ML explainability

Void.

### 5.5.6 ML explainability

#### 5.5.6.1 Description

Explainable ML refers to a process that enables the consumers (e.g., human operator) to understand and trust the outputs provided by ML models. In essence, explainable ML is about making the decision-making of ML comprehensible to its consumers. On a broad level, explainable ML can be categorized into two types:

* Local explanation: The aim is to explain individual outputs provided by an ML model, i.e., it focuses on explaining why a specific output was generated by the ML model for a particular input data sample.
* Global explanation: The aim is to explain the whole ML model behaviour, i.e., it focuses on explaining how the ML model works in general across several or all possible input samples.

#### 5.5.6.2 Use cases

##### 5.5.6.2.1 Local explanation in AI/ML inference

Local explanation is an additional information that the MnS consumer can use to analyze the ML model’s potential impact in terms of network performance gain and make informed decisions regarding deactivation of AI/ML inference or fallback to previous version of the ML model. The ML Training Request IOC initiated by the MnS consumer does not specify requirements for local explainability support in ML model training. It is essential for the MnS consumer to indicate this need so that the MnS producer can preprocess the training data or employ suitable techniques to generate local explanations.

Once the ML model is trained to generate both outputs and corresponding explanations, it is deployed for inference. The generated local explanations by the AI/ML inference MnS producer may need to be reported to the AI/ML inference MnS consumer. These local explanations provide additional information that the MnS consumer can use to analyze the ML model’s potential impact on network performance. This analysis helps in making informed decisions regarding the deactivation of AI/ML inference or fallback to a previous version of the ML model. By considering network performance PM/KPIs, model performance, and local explanations, the MnS consumer can effectively identify the ML model contributing to network performance degradation when several ML models are active in the network.

#### 5.5.6.3 Potential requirements

**REQ-ML-EXP-1:** The 3GPP management system should provide the capability for an authorized consumer to indicate the support for local explanations in ML model training.

**REQ-ML-EXP-2:** The 3GPP management system should provide the capability for an authorized consumer to receive from the AI/ML inference MnS producer the generated local explanations for inference.

#### 5.5.6.4 Possible solutions

Introduce a new attribute, e.g., localExplanationIndicator, in the MLTrainingRequest IoC requested by an authorized MnS Consumer to the ML Training MnS Producer. When the attribute is set to True, ML Training MnS Producer should train the ML model such that it can provide local explanations for the aIMLInferenceName. The attribute is of type Boolean.

Introduce a new attribute inferenceLocalExplanation in the inferenceOutputs attribute (which is of type inferenceOutput) in the AIMLInferenceReport IoC. This information indicates the local explanation for the generated inference corresponding to a particular inferenceOutputId. The attribute is of type string.

#### 5.5.6.5 Evaluation

The solution described in clause 5.5.6.4 proposes the addition of a new attribute to the MLTrainingRequest IoC to enable the MnS consumer to indicate the need for the trained ML model to provide local explanations. Furthermore, it proposes the addition of a new attribute to the AIMLInferenceReport IoC to enable the MnS producer to provide local explanations corresponding to the generated inferences. Any advanced solutions, e.g., identifying the suitable local explanations for each ML model needs further investigation in the normative phase. Therefore, the solution described in clause 5.5.6.4 is a feasible solution to be developed further in the normative specifications.