**3GPP TSG-RAN WG3 #113e** ***Draft R3-214220***

**Online, 16th – 26th August 2021**

**Agenda Item: 18.2**

**Source: Lenovo, Motorola Mobility**

**Title: Summary of offline discussion on AI RAN general framework**

**Document for: Discussion and Approval**

# 1 Introduction

**CB: # AIRAN2\_GeneralandFramework**

**- Converge on the open issues on the AI functional framework:**

**Whether Model performance feedback/Model deployment is needed?**

**Data from Model inference?**

**- Performance Evaluation of AI/ML Models?**

**- Update/correct high level principles, definitions and terms, if needed**

**- Provide TPs if agreeable**

(Lenovo - moderator)

Summary of offline disc in [R3-214220](file:///D:\Users\Zimmermann.Gerd\Documents\Int_Gremien\3GPP\RAN3\2021\RAN3%23113-E\Inbox\Drafts\CB%20%23%20AIRAN2_GeneralandFramework\Inbox\R3-214220.zip)

The offline discussion will comprise 2 phases

* Phase 1: Try to identify easy agreements and controversial issues for Phase 2 discussion
  + **Deadline: August 20th, Friday, 4am UTC**
* Phase 2: Try to come up with TP if agreeable
  + **Deadline: August 24th, Tuesday, 4am UTC**

# 2 For the Chairman’s Notes (Phase 1)

# 3 Discussion (Phase 1)

## 3.1 High-Level Principles

In the submitted draft TR R3-214129, based on the agreed TP R3-212978 from RAN3#112e meeting, the following high-level principles are captured:

|  |
| --- |
| **R3-214129 TR 37.817 v0.2.0 (China Mobile Com. Corporation)**  4.1 High-level Principles  The following high level principles should be applied for AI-enabled RAN intelligence:   * The detailed AI/ML algorithms and models for use cases are out of RAN3 scope. * The study focuses on AI/ML functionality and corresponding types of inputs/outputs. * The input/output and the location of Model inference function should be studied case by case. * RAN3 should focus on the analysis of data needed at the Model training function from external functions, while the aspects of how the Model training function uses inputs to train a model are out of RAN3 scope. * Where AI/ML functionality resides within the current RAN architecture, depends on deployment and on the specific use cases. * The Model training and Model inference functions should be able to request, if needed, specific information to be used to train or execute the AI/ML algorithm and to avoid reception of unnecessary information. The nature of such information depends on the use case and on the algorithm. * The Model inference function should signal the outputs of the model only to nodes that have explicitly requested them (e.g. via subscription), or nodes that are subject to actions based on the output from model inference. * NG-RAN is prioritized; EN-DC is included in the scope. FFS on whether MR-DC should be down-prioritized. * A general framework and workflow for AI/ML optimization should be defined and captured in the TR. The generalized workflow should not prevent to “think beyond” the workflow if the use case requires so. |

Besides the agreed principle above, the following principles are proposed by companies to be agreed this time:

1. RAN3 should focus on the analysis of data needed at the Model Inference Function from external functions, while the aspects of how the Model Inference Function uses inputs to derive outputs are out of RAN3 scope [3][5][10]
2. Aspects of how Model Training function performs model deployment/update are out of scope of this study. [3]
3. Aspects of how Model Inference function generates model performance feedback are out of scope of this study. [3]
4. RAN should not act as a data storage or data memory, user data privacy should be respected during AI/ML operation. [12]

**Question 1: Companies are kindly asked if you agree with any/which of the above proposed high-level principles and needs to be captured in the TR in addition?**

|  |  |  |
| --- | --- | --- |
| **Company** | **1), 2), 3), or 4)** | **Comments** |
| Deutsche Telekom | 1) Agree  2) Agree w.r.t. details, but not from LCM perspective  3) Agree w.r.t. details, but not from LCM perspective  4) Principle has to be split into 2 parts; especially the use of data storage in RAN needs further explanation | 1) Focus of SI should be on inputs and outputs of Model Inference function.  2) The SI should not go into any details of model deployment/update, but the process itself has to be mentioned in the description of the functional framework. DT sees it as part of OAM (at least for offline training), therefore, RAN3 should align with SA5 on that topic to achieve a common view on lifecycle management for AI/ML model handling. For online training, that may happen in the RAN, the need for a description has to be further clarified dependent the use cases.  3) Same comment as for 2).  4) Securing user data privacy is a basic requirement from an operator’s perspective which has to be fulfilled also by the AI/ML framework in RAN.  For Model Inference and online Model Training functions data storage in the RAN is certainly required; to which amount is dependent on use cases (see also the proposal for RAN Data Repository Function (RDRF) in [4]). Therefore, we can agree to that proposed principle only for the case of offline training. |
| Futurewei | 1. Agree 2. Need clarification 3. Do not agree 4. Need clarification | 1. This proposed principal needs more clarification. If model training is performed on external node/system, e.g., OAM then the mechanism and procedures regarding how to deploy and update the model should be described. If model training and model inference reside on the same RAN node, then this is an internal implementation, thus, the mechanism and procedures can be left to vendor implementation. 2. To make the model performance feedback useful for the receiving function if requested via subscription, to the minimum, what data should be used in generating the model performance has to be clarified, otherwise, the model performance feedback generated would become no meaning. This is particularly important when there are multi-vendors in the operator’s network. If the datasets or procedures/mechanisms used in generating the model performance are different, then such information becomes useless at the operator. 3. Agreed on “user data privacy should be respected during AI/ML operation”.   For “RAN should not act as a data storage or data memory”, consistent with DT’s view. It should be clarified. |
|  |  |  |
|  |  |  |
|  |  |  |

Another issue raised by companies [1][2][3][6], is that an AI/ML model used in a Model Inference function instantiated in a logical RAN node has to be initially trained and tested before deployment (typically in an offline manner with test data). Also, a model used initially in the case of online training in a logical RAN node has to be a trained and tested one.

In moderator’s understanding, the model testing/validation before deployment is reasonable implementation of ML model to assist network optimization. The following question here is whether we need to capture the model test/validation in the high level principle, and if there is any impact on the functional framework description, e.g. if the model test/validation happens in the ML training function or ML inference function, and if we have to distinguish testing/validation data from training data and inference data.

**Question 2a: Companies are kindly asked if ML model test/validation needs to be captured as high-level principle?**

* **Yes**
* **No**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
| Deutsche Telekom | Yes | It should be simply stated that models used in the Model Inference function are trained (offline) and validated before deployment in the RAN, but there is no need to describe the process in detail (as being part of OAM from DT’s perspective).  Please note that it would be generally useful to extend some descriptions on functional framework to make issues especially on model training aspects clearer (see e.g. [2]). |
| Futurewei | Yes | It should be included in the high-level principles section. As explained in [1], AI/ML development typically involves training, validation, and testing. model training and validation happen during the training phase where training dataset is used to fit the model parameters and validation dataset is used to tune the hyperparameters of the model (still part of the training), and testing dataset, which is unseen during training and validation, is used for calculating/assessing the model performance.  While Futurewei agrees the description should not go into every detail, we believe those procedures/generated outputs that will either impact the interface or impact the interpretation when received at the other function(s) should be described. One notable example is using the agreed-upon / common dataset(s) in assessing the model performance, otherwise, the generated model performance feedback has very little meaning at the receiving function.  We believe it’s very helpful to include some background information regarding training/validation/testing and what role/responsibility model evaluation should have in the TP as provided in the text proposal and Annex section as part of [1]. |
|  |  |  |
|  |  |  |
|  |  |  |

**Question 2b: Companies are further asked if ML model test/validation shall be reflected in the functional framework?**

* **Yes, please further comment where the ML model test/validation happens, and if we need to define test/validation data in addition to training/inference data.**
* **No**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
| Deutsche Telekom | No, but … | As commented to Question 2a, there is no need to explicitly describe test/validation data for offline trained models in the functional framework, as this can be covered by a simple statement as mentioned before. |
| Futurewei | Not in the framework diagram but part of the description | We suggest keeping the diagram at the high-level, but description should be added to clarify training/validation and testing. Note that the training phase includes training + validation, thus it can be left to vendor implementation. However, testing is more critical as model performance feedback should come from the testing phase, thus, we believe it’s important to mention that performance evaluation of the AL/ML model should be measured against the testing dataset and the samples in the test dataset(s) are unseen during the training phase (including training and validation). |
|  |  |  |
|  |  |  |
|  |  |  |

It was left open in the last meeting, whether/how to capture the case that output from one model could be used as input to another model. Some companies [2][3][6][12] want to capture such model chaining in the high-level principle or capture in the functional framework description. In particular, [2][3][6] thinks to reflect such model chaining in the functional framework, Model Inference may provide output to data collection too, such that its output can be further provided to another model as input.

**Question 3a: Companies are kindly asked if “output from one model could be used as input to another model” shall be captured as one high-level principle?**

* **Yes**
* **No**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
| Deutsche Telekom | Yes | If this is not described in the functional framework, there could be limitations for solutions/implementations for some use cases. |
| Futurewei | No | Futurewei thinks this is use case/implementation dependent and believes it should not be included in the high-level principles section. However, a simple statement can be added indicating that no restriction on whether the input of an AI/ML model is raw data, processed data or output from other AI/ML model(s). |
|  |  |  |
|  |  |  |
|  |  |  |

**Question 3b: Companies are further asked if model chaining shall be reflected in the functional framework, e.g., a solid/dash line drawn from model inference to data collection providing the output?**

* **Yes**
* **No**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
| Deutsche Telekom | Yes | We would support to have it in the figure. |
| Futurewei | No | As commented on question 3a, Futurewei believes that the function framework diagram should be at high-level and clean, otherwise, a lot of interactions could be arguably added. However, a simple statement can be added as clarification as suggested in our comments for 3a. |
|  |  |  |
|  |  |  |
|  |  |  |

## 3.2 Functional Framework

|  |
| --- |
| **R3-214129 TR 37.817 v0.2.0 (China Mobile Com. Corporation)** 4.2 Functional Framework *Editor’s Note: Data Preparation aspects may be further refined*    Figure 4.2-1: Functional Framework for RAN Intelligence  This section introduces the common terminologies related to the functional framework for RAN intelligence illustrated in Figure 4.2-1.   * Data Collection is a function that provides input data to Model training and Model inference functions. AI/ML algorithm specific pre-processing of data is not carried out in the Data Collection function.  Examples of input data may include measurements from Ues or different network entities, performance feedback, AI/ML model output.   + Training Data: information needed for the AI/ML model training function.   + Inference Data: information needed as an input for the Model inference function to provide a corresponding output. * Model Training is a function that performs the training of the ML model. The Model training function is also responsible for data preparation (e.g. data pre-processing and cleaning, formatting, and transformation of raw data), if required. * Model Inference is a function that provides AI/ML model inference output (e.g. predictions or decisions). The Model inference function is also responsible for data preparation (e.g. data pre-processing and cleaning, formatting, and transformation of raw data), if required. * Actor is a function that receives the output from the Model inference function and triggers or performs corresponding actions. The Actor may trigger actions directed to other entities or to itself. * Feedback: Information that may be needed to derive training or inference data or performance feedback. |

It has been widely discussed about the metric to evaluate the performance of a ML model, and the outcome of a ML model. Those metrics could be sent together with the Inference Output or as part of the Model Performance Feedback. In particular, the following metrics were listed by different companies:

* Accuracy/Uncertainty [1][5][6][8], e.g., for classification task or for prediction task.
* Variance [1], e.g., for regression task
* Dispersion [1], e.g., for iterative optimization task
* …

Besides, [5][8] also propose to send validity time together with the inference output.

On the other hand, [9][10][11][12] believe whether/how those metrics and validity time are used and sent together with the inference output really depends on the exact AI algorithm used. Thus, [9][10][11][12] suggest to study the definition and usage of metrics and validity time case by case.

In moderator’s observation, it seems common understanding that some kind of metrics (e.g. accuracy, uncertainty, variance etc. ) will be used to evaluate the model performance and model output, however, how exactly it is defined and used in the functional framework(e.g. in which step) depends on the outcome of the use case solution discussion. Therefore, at this stage, moderator would suggest postponing introducing the metrics and validity time in the functional framework until how they are exactly used in each use case becomes clear.

**Question 4: Companies are kindly asked if it’s ok to postpone introducing the metrics (e.g., accuracy, uncertainty, variance etc.) and validity time in the functional framework until how they are exactly used in each use case becomes clear?**

* **Yes**
* **No, please further comment what shall be clearly captured in the functional framework and what is the definition/usage.**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
| Deutsche Telekom | Yes, but… | We may keep the metrics as placeholders with FFS in the description of the functional framework. |
| Futurewei | It’s ok to postpone the discussion with some text description added. | The proper model performance evaluation metrics to be used depend on the AI/ML task (e.g., classification, regression, iterative optimization while no labels type), but they do NOT depend on the AI/ML algorithms / models or implementations. As discussed in [1], we believe it is ok to discuss the exact metric definitions later, however, the type of metrics for different task types can be mentioned, e.g., accuracy type of metrics, false alarm/misdetection type of metrics, uncertainty type of metrics. As model performance feedback will be used to understand how well AI/ML models performed in unseen data or real-world scenarios and trigger proper subsequent procedures, it is very important to have common understanding and definitions and it will also have impact on the information to be exchanged over the logical functions.  If there is not sufficient time to discuss in this meeting, we think it’s ok to note this as FFS in the TP. |
|  |  |  |
|  |  |  |
|  |  |  |

As outcome of the last RAN3 meeting, the model deployment and update from ML Training to ML Inference was left FFS. Among the submitted papers, [2][6][8][9][10] think the Model Deployment and Update shall be kept, while [5] thinks it is only applicable to single vendor environment in practice.

**Question 5: Companies are kindly asked if Model Deployment and Update from ML Training to ML Inference is needed in the functional framework?**

* **Yes, please further comment if we shall limit it to single vendor case.**
* **No**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
| Deutsche Telekom | Yes | We see the need to have this link in the functional framework, even if RAN3 is not required to explain details in case of addressing the connection from offline model training to inference function (responsibility of SA5 as it should be part of OAM). Therefore, there is currently also no need to limit it to a single vendor case. SA2 explicitly addressed the model exchange in their framework without intervention of OAM which may be different to our RAN framework (further discussion needed if RAN intends to go a similar way on AI/ML approaches as CN/5GC). |
| Futurewei | Yes | We believe it makes logical sense to let the link stay in the RAN intelligence framework diagram as the diagram presents a logical/functional view. We share the same view as DT not to add strict limitation regarding whether model training and model inference should belong to the same vendor or reside on the same RAN node. |
|  |  |  |
|  |  |  |
|  |  |  |

Regarding whether ML Inference shall provide model performance feedback to ML training, [2][3][6][7][8][9][10] believe it is needed in one way or another to trigger the ML model retrain/update at the ML training in case the model performance degrades. [2] believe the model performance feedback is applicable to both offline training and online training:

* In case of offline model training the Model Performance Feedback interconnection is only an optional interface and needed only if certain information from Model Inference function is suitable for improvement of the initially trained model. This information could be prediction accuracy or similar statistical data achieved with the model during run time, resulting response time, or processing and memory size/load available/required in Model Inference function.
* For online model training, e.g. using a reinforcement learning approach, information from Model Inference function is fed back to the Model Training function to further improve the model according to adaptation of model-related parameter settings. Such information could be e.g. on prediction accuracy or similar statistical data achieved with the model during run time, on output data drifts, output data quality (granularity/pattern), or output data mismatch.

In [9], traffic load prediction was taken as an example, the model performance could be evaluated by the accuracy or confidence interval by comparing the predicted traffic load with the actual measured traffic load in a certain time period. And in a reasonable implementation, the traffic load prediction will make use of the historical measured traffic load, which means the historical traffic load measurement will be provided to the model inference function as “Inference Data” and by nature model inference function can determine the model performance in a certain time period.

[7] thinks the model performance feedback arrow from Model Inference to Model Training does not really apply to reinforcement learning, thus suggests using a dash line arrow and keeping it optional.

On the other hand, [5][11][12] propose to remove the model performance feedback from ML Inference to ML training due to lack of motivation or justification.

**Question 6: Companies are kindly asked whether to keep the model performance feedback from ML inference to ML training in the functional framework?**

* **Yes, please further comment whether it should be optional, e.g., dash line.**
* **No, please further comment on the reason considering the explanation above.**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
| Deutsche Telekom | Yes | As DT sees it as optional link, it is ok for us to use a dashed line. |
| Futurewei | This should be optional as model performance feedback can be generated at other functional module, like model training function. | Technically speaking, model performance feedback can be generated or calculated at either the model training or model inference function if the needed data is available, i.e., the prediction results (which can be obtained via subscription and provided by the model inference) and the labels (also can be obtained via subscription and can include the feedback from actor and/or other attributes). However, in some cases model inference module may have limited capacity, thus acquiring and storing labels may increase its burden and may be more suitable for the model training function to calculate such information.  Please note that the model performance feedback is not known at the time of inference and has to be calculated till the labels are received or calculated after receiving the feedback from the actor at the next time interval.  Given that model performance feedback is important, we suggest adding descriptive text to clarify what data is needed in generating the model performance feedback and this performance attributes can be available to the functions that subscribe to such information as suggested in [1]. Regarding which function should generate the model performance feedback, it can be left to use case discussion or implementation. |
|  |  |  |
|  |  |  |
|  |  |  |

# 3 Conclusion

# Reference

1. R3-213196 AI/ML Model Performance and Reliability Evaluation (Futurewei) discussion
2. R3-213211 Discussion on AI/ML-based functional framework for RAN intelligence (Deutsche Telekom AG) other
3. R3-213294 Proposed TP to TR 37.817 on high-level principles and functional framework (NEC) other
4. R3-213373 AI/ML Architecture (Qualcomm Incorporated, Deutsche Telekom) discussion
5. R3-213418 (TP for SON BL CR for TS 37.817): Framework for RAN intelligence (Ericsson) other
6. R3-213468 High level principle and Functional Framework of AI/ML enabled NG-RAN Network (Intel Corporation) discussion
7. R3-213540 Discussion on framework of AI (CATT) discussion
8. R3-213712 Discussion on Functional Framework and High-Level Principles (Samsung) discussion
9. R3-213723 Remaining issues on AI functional framework (Lenovo, Motorola Mobility) discussion
10. R3-213756 Left issue on AI Functional Framework for RAN Intelligence (ZTE Corporation, China Unicom) other
11. R3-213892 (TP for TR 37.817): Further discussions on the AI/ML Framework (Nokia, Nokia Shanghai Bell) other
12. R3-214078 Further discussions on general principles and frame work (Huawei) other

# Annex