**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. |
| Lenovo, Motorola Mobility | 1) Agree  2) 3) 4) depends on the use case solution discussion | 1) is basically saying RAN3 does not need to dig into the AI/ML algorithm which is fine.  2) 3) 4) depends on the discussion in the use case solution, we can come back to them later. |
| Nokia | 1) Agree but  2) Agree with a clarification  3) Agree  4) Agree at least for now | 1) It is unclear to us what is the meaning of data needed from “external” functions for Inference. They are external to what? Otherwise, we agree that how Model Inference uses input to provide outputs is outside the scope of the SI.  2) Even though how model deployment/ update is done is outside the scope of RAN3, it is still in the scope to know that this is possible e.g., by checking with SA5, in cases an ML Model is trained in the OAM and deployed in the RAN.  4) It is our understanding that a RAN node depending on implementation can store data that it needs for the purpose of AI/ML. We understand this proposal in terms of having an additional memory storage that can be accessed by other nodes. Given that the current SI is based on the existing architecture, in our view such solution is more appropriate for a Rel 18 topic. |
| Verizon | 1. Agree   2), 3) Agree with clarification  4) Needs to be split into 2 parts - data storage and privacy. Agree with privacy but not with data storage | 1) Study should focus on data needed for AI/ML training/interference functions and agnostic to where there are located (internal/external etc).  2), 3) LCM aspects should be aligned with SA5  4) Data storage in RAN node likely would be needed for supporting AI/ML algorithms. However, details of data storage can be left to implementation. Mechanisms for data retrieval on demand/subscription by AI/ML functions need to be defined.  User data privacy is essential requirement from operator perspective. |
| Huawei | Agree to all | For 1), 2) and 3), in general we share similar view with DT. For 4), seems that companies all agree that security and data privacy should be respected; then for data storage, we mainly refer to data storage for offline training; for online training, we also think that it should be allowed, so we also share DT’s view. |
| Samsung | 1) Agree  2) Need clarification  3) Agree  4) Partial agree | 1) Yes, should focus on input/output of model inference and detailed model should be out of scopes.  2) For offline training, the model deployment/update is out of scope. But for online training or federated learning, the description may be needed.  3) Yes, the detailed AI/ML algorithms and models should be out of scope.  4) Same view as DT. Data privacy should be considered. For the data memory/storage, it depends on use case. |
| NEC | 1) Agree but rewording is needed  2) Agree  3) Agree  4) Need further study | 1) We agree with this principle, but we believe this needs to be re-worded as follows to include also Model Training function:  “RAN3 should focus on the analysis of data needed at the Model Training and Model Inference functions from external entities.”  2) It is our understanding that signaling between Model Training and Model Inference for the purpose of model deployment/update will not be considered in this study due to limited time of this study. Such signaling depends on particular AI/ML models, their implementation and requires a lot of study to be standardized in an interoperable form agreeable to interested companies.  This does not preclude from discussing such signaling in Release 18 study.  Definitely there are LCM and other OAM aspects for model deployment/update. This could be discussed together with SA5  3) Similar to 2), due to limited time of this study maybe it is difficult to define such signaling within this study.  However, there is some missing point in general functional framework that has to be addressed.  As pointed out in [1], [3], [6], labeled data (called “test data” in AI/ML) is required to generate model performance metrics. Basically, training data is divided into two parts and one is used for training/validation and one for testing the result of training.  There are two options to make it feasible:  Option a: If model performance metrics are generated in Model Inference, then Model Inference shall be provided with labeled test data.  Option b: Model Training could generate model performance metrics and then there is no need for model performance feedback interface from Model Inference to Model Training.  4) We definitely support that “user data privacy should be respected during AI/ML operation.”  The part related to data storage needs further study. |
| Ericsson | 1) Agree  2) Agree  3) Agree, with clarifications  4) Agree for data privacy | 1) This is a mirror image principle already captured for the for the Model Training function, namely:  “*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.*”  In reply to Nokia, to our understanding “External functions” that can provide inputs to the Inference function is “Data Collection” function. Maybe we should clarify that to make the principle clearer?  2) We have agreed that “The detailed AI/ML algorithms and models for use cases are out of RAN3 scope.”. Hence the AI Models are implementation specific and will not be standardised. With that, RAN3 cannot specify how signalling of an AI Model is performed, as this would imply specifying signalling of a proprietary information, which is in itself non-interoperable (hence outside scope – RAN3 does not define non-interoperable procedures).  We could at best take the approach SA2 took in TR 23.700-91  *”3GPP standardized sharing of models across different vendor environments is not deemed feasible in this release of the specifications. Sharing of models or model meta data is limited to single vendor environments.”*  3) Here we are open to Model Inference to provide feedback towards the Actor, e.g. with regards to the accuracy of the output generated. We agree that the Model Performance Feedback from Model Inference to Model Training function is out of scope  4) As stated by other companies above, Data Privacy is important and we should ensure that it is ensured when transferring data. We would also add that user anonymisation is a mandatory aspect of data transferring. We do not agree with the principle of banning the RAN to store data as this would preclude the possibility of the RAN hosting functions like Data Collection, Inference Function and more. |
| ZTE | P1: Agree  P2: Depends on specific solution on use case.  P3: Agree  P4: Agree for data privacy. | 1. Yes, the same principle as model training agreed last meeting. 2. How the model deployment/update from OAM and RAN should be aligned with SA5. And given that some use case perhaps needs online training, the principle can be discussed later. 3. This is related to AI/ML algorithms, which is out of scope. 4. User data privacy should be considered during the AI/ML procedure, so how to guarantee and support AI data transmission security should be studied.   Perhaps RAN should support data storage/data memory for some use cases. |
| Intel | P1: Agree  P2: Depends on use case solution  P3: Agree  P4: Disagree with data storage/memory; Agree with user data privacy | 2) In CB: #AIRAN3/4/5, model training at OAM and model inference at RAN is being discussed. For this scenario, model deployment/update is needed, at least how OAM can deploy model for RAN to perform model inference need to be studied. We prefer to conclude that aspects after we have certain level of insight on the solutions.  4) as commented by many companies, user data privacy should be respected. For the first half sentence, we think RAN is possible to act as any functionality in RAN intelligent framework, e.g. data collection, model training, model inference. Considering the latency of requesting data from external data storage node, RAN may act as data storage or data memory for some use cases.  Hence, we share the same view as DT and other companies, we suggest to split P4 into two parts:  P4-1: user data privacy should be respected during AI/ML operation.  P4-2: RAN can act as a data storage or data memory, |

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]. |
| Lenovo, Motorola Mobility | Maybe | We are fine either:   * Capture it as a general principle, or * Capture it as part of Model Training function. |
| Nokia | Yes but | ML Model Test/Validation should be discussed to be part of the training in order to make the framework complete. Namely, it’s existence should only be recognized and defined. |
| Verizon | Yes | High level description should be fine. Detailed description of model test/validation is not necessary as it will be dependent on use case. |
| Huawei | No, but… | In general we align with DT’s view that there is no need to describe the process in detail (part of OAM also from Huawei’s perspective), and we don’t think there is any RAN3 related standardized solution needed behind, before an offline trained model is configured to RAN node, OAM function should be involved in test/validation. |
| Samsung | Yes, but | A simple high-level description is fine. Model test/validation is a part of model training work before the inference. |
| NEC | Not sure | Not sure this should be general principle.  Rather it should be part of functional architecture (please see our comment to Q2b). |
| Ericsson | Yes, purely as high level definition | As stated by Huawei, there is no impact on RAN3 work from the process of testing and validation (all occurring within the training function). We could aim for a high level definition, such as the one below:   * ML Training: An online or offline process to train, test and validate an ML model by learning features and patterns that best present data and get the trained ML model for inference. |
| ZTE | Yes, but | ML model test/validation is one part of the ML model training. Simple high-level description is fine.  In addition, testing/validation data can be separated from the training data, which means that during model training, training datasets contains training data, validation data, and testing data. Therefore, training data in current framework is enough, or to add simple text description for training data if really needed. |
| Intel | Yes | From RAN3 general principle point of view, capturing the model validation and testing as part of the model training functionality is needed.  However, how to distinguish data for training and data for test/validation is part of LCM discussion, which may be in the scope of SA5. We suggest to send a LS to SA5 discussing LCM and CM for AI/ML. |

**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). |
| Lenovo, Motorola Mobility | Maybe | We are fine either:   * Capture it as a general principle, or * Capture it as part of Model Training function. |
| Nokia | No | Test/Validation Data should not be defined in the input in addition to training and inference data. Test/Validation is a process that happens as part of training to evaluate that the Model is trained properly and is specific to the ML Model/algorithm. |
| Verizon | No | Detailed description of model test/validation is not necessary as it will be dependent on use case. |
| Huawei | No. | Similar view as DT, we could have a short statement. |
| Samsung | No | Model test/validation is a part of model training work before the inference. As framework is just to provide guideline for further use case study. it seems there is no need to capture it in framework. It may be discussed upon the requirements of use case. |
| NEC |  | We believe that validation and test serves two different purposes.  Model validation is used to optimize hyper parameters during model training. Validation data is used to compare model performance metrics of several AI/ML models trained with different sets of hyper parameters in order to select the best trained model.  Model test uses test data independent from training and validation data in order to generate model performance metric for the selected trained model.  Based on the above, validation is part of training. No need to update functional architecture figure, but could be mentioned as part of model training description.  Model test and test data are currently not captured in the functional architecture. We believe functional architecture in its current form does not work in terms of generating model performance feedback.  As we have described in Q1, there are two options to fix this:  Option a: If model performance metrics are generated in Model Inference, then Model Inference shall be provided with labeled test data.  Option b: Model Training could generate model performance metrics and then there is no need for model performance feedback interface from Model Inference to Model Training.  It is proposed to select one of these two options and update functional architecture accordingly. |
| Ericsson | No | Fully agree with Nokia |
| ZTE | No | Testing/validation data can be separated from the training data, which means that during model training, training datasets contains training data, validation data, and testing data. Therefore, training data in current framework is enough, or to add simple text description for testing/validation data if really needed. |
| Intel |  | Not necessary for the figure, but we may need to update the definition of data collection and model training.  From our understanding, both validation and testing happen in model training, where the data is obtained from “data collection”, even if training is performed offline.  Following updates are proposed to be considered in the functional framework:  1. replace “Training data” from “data collection” to “model training” into “Data for Training/validation/testing”  2. update definition of data collection: add validation data and testing data as example of input data.  3. update definition of model training: a function that performs the training, validation and testing of the ML model. |

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). |
| Lenovo, Motorola Mobility | No | We tend to agree with Futurewei, it might be difficult to be a general principle for all cases, but could clarify there is no limit on whether the input of one model could be output of another. |
| Nokia | No | Model chaining highly depends on the use case. It was agreed in the last meeting that functional framework is independent of ML model types or learning models/settings so it should not be captured there. We have already captured in the definition of Data Collection that it provides input data to Model Training and Model Inference where input data can also be an AI/ML model output, as an example. |
| Verizon | Yes with clarification | As a high-level principle this (output from one model as input to another one) should be possible and the framework should allow it. There should not be a limitation in regard to using processed data. But the details (what data etc) would depend on the use case and the particular AI/ML algo and should be left to implementation. |
| Huawei | No | This is just a concrete method which might be used for some use cases, and other implementation may not rely on such method, so it should not be a high-level principle, in our understanding. |
| Samsung | No | Actually, there is no limitation for data collection to collect the raw data from RAN or the AI/ML model output to be the input of model training or inference. So there is no need to stress this case. |
| NEC | Not sure | We definitely support that output from one AI/ML model could be input to another AI/ML model.  But we also acknowledge, that this depends on use case.  Maybe it is better to capture this in functional architecture rather then as a general principle. |
| Ericsson | No | As described in [5] whether a model Inference function is implemented as one function or a chain of functions is an implementation choice. RAN3 should not shape specifications around specific implementations. Besides, Nothing prevents a Model Inference function and a Data collection function to be hosted by the same node. The latter is use case dependent, but in such case, the data collection function could take the output generated by the Model Inference function and provide it to another Model Inference function. Hence the framework already covers the case of model chaining. |
| ZTE | No | We think it is not precluded in current AI functional framework, but model chaining depends on use case (e.g., prediction information as ML model inputs). Not necessary for high-level principle, and we could clarify it in the training data definition through remove “raw” restriction and include output data from the other ML model. |
| Intel | Yes | In general, we share the same view as Verizon and DT that we can capture it in high level principle. However, whether one model output need to be used for another model highly depends on the implementation and algorithm of different use cases. It might be hard to capture the exact procedure how those output can be used by another model. Hence, the high level principle could be kept as simple and general. |

**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. |
| Lenovo, Motorola Mobility | Yes, optional | Since it applies to some cases, it seems ok to draw a dash line for model inference to provide output to the data collection.  We are also fine to postpone the decision and wait for the progress in each use case solution. |
| Nokia | No | For the reasons explained above. |
| Verizon | Yes, optional | Same comments as in 3a. A dotted line and clarification text might help clarify that there is no restriction on data from one model to be used by another model. |
| Huawei | Seems not needed | We think framework should be simple and clean which should be AI/ML algorithm/learning method independent. |
| Samsung | Seems not needed | As for the issue of output from one model as input to another, the current framework does not preclude such case. The output from one model can be one of the collected data, and the functionality for each block is same as the current one. So there is no special issue for the case that output from one model as input to another, and current framework is good to cover such case. |
| NEC | Yes | Even for three selected use cases, UE trajectory prediction and traffic/load prediction are inputs to energy saving, mobility optimization, and load balancing.  If more use cases are considered later, such model chaining would have more applications.  We propose to have such line on the functional architecture figure with corresponding description. |
| Ericsson | No | As explained in question 3a, the functional framework already covers this aspect. |
| ZTE | Not necessary | Same reason explained in Q3a. |
| Intel | Depends | Whether this need to be captured in functional framework depends on whether the output of model inference will be collected by data collection in certain way, e.g. together with feedback or directly provided to data collection.  One way is to capture as a dot line, indicating that this feedback is optional. We are also ok not to update the functional framework as long as “data collection” has the visibility of output of inference data. |

## 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. |
| Lenovo, Motorola Mobility | Yes | We see them strongly dependent on the use case and specific task. Better postpone it. We are fine to leave a FFS in the description. |
| Nokia | Yes | We can delay the discussion to later. At least the current metrics, are strongly related to the ML Model and its performance. At this point it is unclear to us how these detailed metrics can be useful for our study. |
| Verizon | Yes, but | Same view as DT |
| Huawei | Yes | As mentioned in our paper, metrics are part of AI/ML modelling itself. |
| Samsung | Yes | We think it depends on the use case. We can discuss it when it is needed for the use case study. |
| NEC | Yes but | As we have mentioned in Q1 and Q2b we believe that the current functional architecture does not work to provide such model performance metrics.  We would like to propose to select between two options at earlier stage:  Option a: If model performance metrics are generated in Model Inference, then Model Inference shall be provided with labeled test data.  Option b: Model Training could generate model performance metrics and then there is no need for model performance feedback interface from Model Inference to Model Training.  After one option is selected, detailed metrics could be discussed. |
| Ericsson | Yes, but | The value we see here is in providing accuracy levels on predictions from the Model Inference function. A validity time could be useful too.  For performance feedback we already have a description that marks our intentions, i.e.  “*Feedback: Information that may be needed to derive training or inference data or performance feedback*.”  We are fine with the proposal of postponing the discussion in the framework AI, but then we should put up for agreement that the following point is taken in the use cases discussions:  *Whether and how to signal output accuracy and validity* |
| ZTE | Yes | We are not sure how to use these metrics with the output from model inference to action. Perhaps we can discuss it case by case. |
| Intel | Ok to postpone in this meeting | The accuracy level we proposed is generated from Model Training, which is calculated based on input of validation and testing data. This is a general methodology to represent how well one AI/ML model is trained, which is independent from detailed AI/ML algorithms. As majority view in Q2a is to capture model test/validation as high-level principle, we think there’s a need to provide this accuracy level from “model training” so that “model inference” could understand how well this model can be trusted. |

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. |
| Lenovo, Motorola Mobility | Yes | Maybe in practice it’s most likely the model deployment is from a single vendor, but we also don’t think it’s necessary to limit in the standard. |
| Nokia | Yes | It should be kept at least for completeness of the functional framework. We don’t need to limit the scope to single vendor. |
| Verizon | Yes | Yes, it will be useful for completeness of high-level framework. Multi-vendor operation should be possible within the framework. |
| Huawei | Yes | Anyway, a trained model should be configured to the entity which will use it. |
| Samsung | Yes | It is necessary in terms of framework completeness. And we think it is better to not limit “single vendor” at current stage. |
| NEC | Yes | It is better to keep. |
| Ericsson | Yes, but | We are fine with keeping the arrow in the diagram, but we should then specify that this is out of RAN3 scope. We also insist on the aspect that Model Deployment/Update is an intra vendor procedure from a 3GPP prospective. How could it be multi-vendor if the model itself is proprietary?  In reply to DT, SA2 defined the following in TR 23.700-91  *”3GPP standardized sharing of models across different vendor environments is not deemed feasible in this release of the specifications. Sharing of models or model meta data is limited to single vendor environments.”*  Hence, SA2 already converged on this conclusion and it would be beneficial for RAN3 to align.  As a possible compromise we could propose to capture the following:  *Model/Deployment update: Out of RAN3 scope. The payload of this signalling is vendor proprietary.* |
| ZTE | Yes | Model deployment/Update is one logical procedure during AI/ML operation. If we removed the model deployment/update, it seems to break the logical functions in RAN intelligence.  Besides, no need to limit it to single vendor case. |
| Intel | Yes | As replied in Q1 and Q2a, we think this may impact the interaction between OAM and NG-RAN. According to SA5 MDA, currently, RAN as MDAS consumer can only get analytics report from OAM. It is not clear whether model deployment/update from OAM to NG-RAN is supported or not. As this is related to LCM and CM for AI/ML model between OAM and NG-RAN, we suggest to send a LS to SA5 checking whether this is feasible or not.  If both model training and model inference locate in NG-RAN (e.g. training at CU, inference at DU), we are ok to study RAN3 impact if there’s any solution of use cases is identified. |

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. |
| Lenovo, Motorola Mobility | Yes | We tend to agree the model performance feedback might not suit the decision-oriented AI task. We are fine to have it optional, with dash line. |
| Nokia | No | Model evaluation can be done through different performance metrics (e.g., confusion matrix, recall, accuracy, precision, ROC, mean square error, to name a few). But those are strongly dependent on the Learning method (Supervised learning Classification or Regression vs Reinforcement learning) and dataset properties. Since in this SI, ML Models and Algorithms are not in the scope, we think that also Model Performance feedback should not be standardized. |
| Verizon | Yes, as optional | The details of AI/ML model and its evaluation are out of scope of RAN3 and left to implementation to foster innovation. However depicting a dashed line in the framework diagram as a high level principle should be fine. |
| Huawei | No | Maybe we still need to discuss what the performance feedback here means, technically in our understanding, performance could only be obtained/evaluated after an action is taken. |
| Samsung | Yes | Model performance feedback is to feedback the effectiveness of model (such as accuracy, problem occurred during inference) from model inference to provide reference to model training. As the model performance feedback depends on the use case (AI functionality, data availability), dash line is fine for us. |
| NEC | Depends | As we have mentioned in Q1, Q2b, and Q4 we believe that the current functional architecture does not work to provide model performance metrics.  We would like to propose to select between two options:  Option a: If model performance metrics are generated in Model Inference, then Model Inference shall be provided with labeled test data.  Option b: Model Training could generate model performance metrics and then there is no need for model performance feedback interface from Model Inference to Model Training.  If option a is selected, answer to Q6 is Yes plus test data shall be additionally provided to Model Inference.  If option b is selected, answer to Q6 in No. |
| Ericsson | No | Maybe there should be a few more words on why companies in [5],[11] and [12] believe Model Performance Feedback is not needed, to make an equal representation with the opposite view.  We do not understand what Performance Feedback is. If this is feedback on the performance of e.g. a prediction, then such feedback cannot be achieved until an action is taken and measurements on the actual data (ground truth) is available. The latter is not performed at the Model Inference function. If feedback is the model accuracy derived at e.g. validation, then this information is already available at the Model Training function. We therefore do not understand what this function is supposed to provide. |
| ZTE | Yes | Model performance feedback is useful for the prediction task. We agree that model performance cannot calculated or generated at the time of inference, and should wait for the real labels or after actions, but the data could be collected from data collection to model inference. It’s model inference’s duty to calculate the model performance and send it to the model training. |
| Intel | Depends on purpose | From our understanding, performance feedback of a model could be only available from the environment after a certain action is taken. This is already reflected in “feedback” arrow from Actor to Data collection. Hence, the direct feedback for model performance from “Model inference” to “Model training” is not needed.  However, when deploying a trained model from Model training to Model inference, the network node with Model training may need to understand the capability of network node who will perform model inference. A arrow from model inference to model feedback can be used for the purpose of indicating AI/ML capability of the network node. |

[Additional Question]

Another issue worth checking is that although it seems common understanding that the data collection will only provide the demanded data as requested by Model training/inference as captured as one high-level principle:

* 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.

However, the step for the Model training/inference to request specific information from the Data Collection is not drawn in the framework. [1] proposes to draw a new arrow requesting data from Model training/inference to Data Collection to better reflect the high-level principle.

**Question 7: Companies are kindly asked if it’s ok to draw an “data request” arrow from Model training/inference to Data Collection in the framework figure to better reflect the high-level principle?**

|  |  |  |
| --- | --- | --- |
| **Company** | **Yes/No** | **Comments** |
| Lenovo, Motorola Mobility |  | We agree with the understanding, while we suppose it’s sort of common understanding from last meeting it’s enough to capture it as one general principle. No strong view. |
| Deutsche Telekom | Not yet | Similar view as Lenovo. We have already captured the general statement in the bullet point listed above as a high-level principle for the framework.  W.r.t. changing the figure, unfortunately, I couldn’t find the mentioned proposal in [1]. We have nothing against it, but we have first to think about how to present the principle for “data requests” as this may go beyond the relation between training/interference function and data collection as also an actor may request dedicated output data from an inference function. This is currently also not covered in the framework. In principle, this could be covered by an SBA-related approach where functions can subscribe to services offered by other functions. |
| Futurewei | No | While the concept is ok to avoid reception of unnecessary information, we suggest keeping the diagram at high-level and clean and there is already a bullet item in the section to cover this aspect. We share the similar view as DT as this data request type of interface is applicable between other functional blocks, e.g., inference result can be delivered to other functions via data request or subscription, thus we should avoid adding many arrows to the diagram for situations like this. |
| Nokia | Yes | It would be good to align the text of the TR in the high-level principles to what is illustrated in the figure to avoid confusion and to be consistent. This arrow appears in many of the described solutions e.g., for scenarios where training functionality is located at the gNB and gNB requests data for training from a UE or from a neighbouring gNB, but this interaction is not shown in the framework. |
| Verizon | Yes | Agree with Nokia. |
| Huawei | Not sure | We are talking about framework but not procedure or flow chart, data request seems to imply a request/response/reject procedure… |
| Samsung | No | We got the intention. The high-level principles about this request is agreed. The framework is to provide the guideline for use case study, so prefer to keep it clean and brief. The detailed “data request” can be discussed on use case basis. |
| NEC | No | Functional framework figure shows general framework in a rather conceptual way underlining main data flows.  Definitely, if we consider exact signaling implementation, all interfaces in the figure are bi-directional with various aspects like interface setup, service subscription, request/response/update, etc.  We do not understand why out of all these interface and various signaling messages we have to pick up one (data request) and put it in the Figure. |
| Ericsson | Yes | We already captured a high level principle on this aspect. We could add an arrow, but we see it as a non essential addition, although it would increase clarity |
| ZTE | Not recommended. | We agree that data request/response message are needed for data collection. That’s why we propose to enhance the interface to support AI measurement request/response. But we don’t recommend to add “data request” in the current framework, which is just a AI/ML processing framework. It’s better to keep it clean. |
| Intel | Ok to capture | Agree with Ericsson and Nokia. |

# 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