Response to NIST RFI on Artificial Intelligence Risk Management Framework

First of all, we thank you for all your time in checking this proposal and would like to introduce what we are doing for securing trustworthy Artificial Intelligence (AI).

In recent years, the Ministry of Science and ICT of Korea have released the national strategy on "Realization of Trustworthy AI ('14.5.21)". As a next step for realizing that strategy, TTA has been assigned as the main performer in the part of developing the technical methodology for evaluating trustworthiness. Therefore, we have been developing the technical methodology including requirements for implementation of trustworthy AI, detailed techniques for the requirement, and testing and evaluation methods for Verification and Validation (V&V) of requirements.

As an identical objective, we found NIST is working on trustworthy AI. Thus, we have concluded that not only responding to the NIST RFI but also sharing our work and getting the feedback from professional research institutes have to be performed. In this regard, we would like to introduce our systematic approach on how we try to assure trustworthiness of AI technically and hope have meaningful discussion about this

Please have a look at it and If you have questions feel free to contact us.

Your sincerely,

Junho Kwak, TTA

Our approach toward assuring trustworthiness

We define assuring trustworthiness of AI as the matter of controlling risk. Basic concept of controlling risk of AI is the same as risk management used in software/system engineering(ISO 31000) and functional safety(IEC 61508). However, we need to define the risk of AI and how we measure the degree of risk. ISO/IEC JTC1/SC42 is developing concepts of risk management of AI(ISO/IEC 23894 CD). But specific methodology for risk management is not clear. Therefore, we suggest following process, which defines the risk of AI and identify, assess(measure), response to the risk.

▶ **Process 1: Define risk**

In fact, this is a premise for risk management. Trustworthiness consists of several attributes such as transparency, Robustness, Fairness, etc. In addition, risk should be defined separately in aspect of each attributes. Based on ISO/IEC 24028:2020 and EC Ethics Guidelines for Trustworthy AI, we defined 10 attributes of trustworthiness of AI. In this process, we also conducted survey and Delphi analysis to attain consensus about the technical concept of trustworthiness among stakeholders and researchers in Korea.

▶ **Process 2: Identify and assess risk**

Not all risk may exist in a product or service. For example, deep learning model for predictive maintenance service in smart manufacturing may have issue in aspect of safety and transparency, but it does not care about fairness. We are in the process of development of a framework, which derives necessary attributes by analyzing a target product. For analysis, following factors are being considered.

: To Which function or service AI is used

: How decision of system is made by AI

: If AI goes wrong, how severe the impact is (in aspect of environment, human, society, etc.)

Through considering these factors, necessary attributes of trustworthiness and related hazards(causal factors of risk) should be clarified. The result is utilized as risk profile. Our goal is visualizing the degree of each attributes. Visualization will be the form of spider web graph, which shows the degree of level that means importance of each attributes(i.e. required level of trustworthiness)

▶ **Process 3: response to risk**

In advance, 32 general requirements were elicited. Those are provided as requirement pool or

category. Each requirement may cover one or more attributes. Covering means mitigating, soothing, getting rid of the risk in aspect of the attribute. According to the level of trustworthiness, requirements are tailored more strictly or applied loosely. We are working on those criteria of each requirements, but tailoring for various domains and types of service is not our scope.

We also should think about verification and validation(V&V) of requirements. We proposed 132 core evaluation method for V&V according to 32 requirements. Evaluation method may contain test technique, test scenario, audit process and checkpoints to make sure identified risk and hazards are properly mitigated or removed. It is also tailored by the level of trustworthiness.

When implementing and evaluating requirements are properly done, we are able to declare that target product achieves required level of trustworthiness. Throughout the lifecycle of AI product or service(may contain so-called AI pipeline), these processes may be iteratively done until accomplishment of required level.