Quality Management for Machine Learning Systems

Yutaka OIWA[1]

[1]*National Institute of Advanced Industrial Science and Technology, Tokyo, Japan*
y.oiwa@aist.go.jp

ABSTRACT

We have been developing a methodology for process-based quality management of machine learning-based AI systems. Our fruit is compiled as a guideline document named "*Machine Learning Quality Management Guideline*", published as our technical report. We will describe our background motivation, surrounding situation and our proposal for quality management.

1. BACKGROUNDS

Artificial Intelligence (AI) has become indispensable tool for developing software systems, especially for those to interact with complex real-world environments. Recently, AI software systems are used even for several safety-critical systems such as medical diagnostics, autonomous control for automotives, etc. These are also used for several humanity-critical applications such as healthcare, job recommendation, or intelligence surveillance. As those critical usage emerges, Fear for possible negative impacts of AI to humans has arisen, and demand for quality control is increasing.

There are also some government-level activities for regulating AI. In 2019, OECD has published a document named "*OECD Principles on Artificial Intelligence*", which states requirement for several aspects of AI such as fairness, robustness, safety and security. It also demands transparency, explainability and accountability to be maintained by system owners. In 2021, the European Commission has proposed a draft regulation for governing AI usages and managements, which is adopted by their Parliaments in June 2023.

Quality management for AI, especially those created with machine learning technologies (ML) is, in short, very difficult. Existing and established methods for quality management of software systems are based on divide-and-conquer approach: first they analyze all possible risks caused by misbehavior of systems, assigning a countermeasure for each enumerated risks item-wise, then implement and test each of them, independently. This approach is well accepted and reflected to many standards such as ones published by ISO and IEC. However, as ML systems are numerically and statistically derived from training data, such divide-and-conquer approach does not work well. Assigning a training data for some estimated risk condition does not guarantee effective countermeasure; furthermore, if we found some unresolved risk and put additional treatment for it, it will often invalidate existing countermeasure for other risks. Such situation implies that ML systems are not well aligned with the existing industrial standards and certification systems.

To overcome such situation, we need a new framework for ML software quality management, which may possibly amend or augment existing, established standard methods for software quality management. With discussions with Japanese industrial community partners, we compiled a set of criteria for such quality management and published as a guideline document (AIST 2021-2023).

In following sections, we will briefly introduce design and structure for our quality management approach and the outline of the published guideline.

3. BASIC APPROACH IN OUR GUIDELINES

In our guidelines, we have defined (or clarified) the *quality* into three related concepts; quality in use, external quality, and internal quality (Fig. 1). These have different viewpoints, different criteria, and some dependency between them.

The quality in use is a system-wide concept of quality which is expected or requested by system users; it is often based on technical (such as safety, security) or social (e.g. ethicalness, fairness, privacy preservation) nature. That quality is demanded externally from outside systems, and to be satisfied by external quality below.

The concept *external quality* describes an abstract guarantee to be established by the producer of AI systems. Alternatively, these can be said as a quality visible from the outside of the concerned items. In the guidelines, we have identified five specific and distinct external quality aspects for typical AI usages: Safety, Performance, Fairness, Privacy, and Security. The guideline also defines *assigned levels of*

quality for some of these aspects, to determine intensity of process management activity for assuring the external quality. For example, our definition of AI safety levels (AISL) is categorized into 7 levels, where 4 of those are roughly corresponding to already established concept of the Safety Integrity Levels (SIL) in existing international standards.

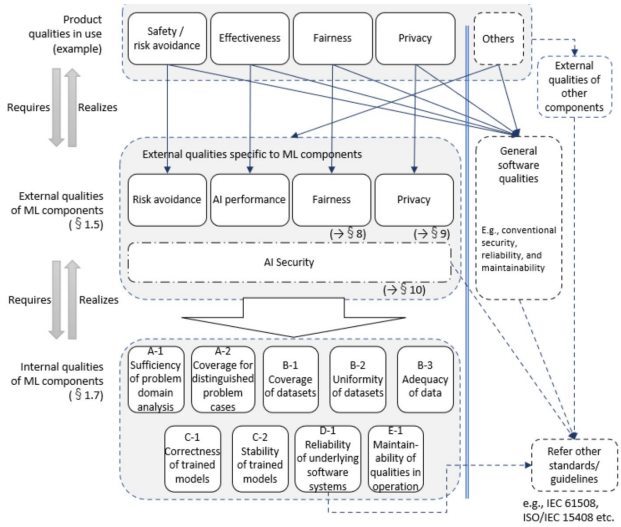

Figure 1. Structure of the quality management guidelines.

In correspondence with those, we also defined *internal quality* aspects. The internal quality aspects are concepts which are visible from the inside of the development process, and closely related to the process steps of the AI developments. The guidelines define 9 particular aspects grouped into 5 clusters. In brief, our conceptual logics for realizing quality can be described as follows:

1. First, to develop ML systems with certain levels of quality requirements, we must have a *design of datasets* which represents a required quality and given problems of the systems in question. This design is critical for establishing *norm* or *measure* of quality in the following steps.
2. After establishing such norms, we can now check to ensure the quality of actual datasets, whether these satisfy all required features of the problem. Without having good datasets, we cannot believe that the final outcomes will satisfy the quality needs.
3. Even if we had a very good training dataset, ML does not, unfortunately, guarantee that the derived AI models are good. We need a separate check for the quality of the output model, in corresponding with concrete examples in the dataset.
4. As ML and AI are just software systems in some sense, all software components which are implemented in conventional software technology must be quality-assured in the conventional standards and methodology. Such components include model framework, training environments, etc.
5. Quality of AI applications often tend to deteriorate as time passes, due to changes of the nature of the usage environments, which cause some disparity between data trained and data in use. We often need to implement a monitoring and update facility to AI systems.

In the actual guideline document, we also describe some concrete technical/management aspects to be checked during development and some hints for available technologies as well.

4. STANDARDIZATION

Currently, Joint technical committee 1 (JTC1) of ISO/IEC has established SC42 (Artificial Intelligence) to discuss standardization related to AI technology. Among them, there is a draft technical report for relation between functional safety and artificial intelligence systems, currently in final discussion. We have contributed to the project with our knowledge from the guideline document. Internationally, there are strong needs for regulating AI usages (with some differences of intensity among regions), and standards will be important materials to harmonize rules, demands and activity for quality management in all the world. We will continue actively involve with standardization communities to implement a good quality management framework to the industry in whole.

ACKNOWLEDGEMENT

This work is supported by the NEDO Project P20006, funded by New Energy and Industrial Technology Development Organization (NEDO).

REFERENCES

AIST: National Institute of Advanced Industrial Science and Technology (2023). *Machine Learning Quality Management Guideline, 3rd English Edition.* Technical Report Digiarc-TR-2023-01, Digital Architecture Research Center. Tokyo, Japan. https://www.digiarc.aist.go.jp/publication/aiqm/guideline-rev3.html .

AIST: National Institute of Advanced Industrial Science and Technology (2021). *Machine Learning Quality Management Guideline, 1st English Edition.* Technical Report CPSEC-TR-202002, Cyber Physical Security Research Center, Tokyo, Japan. https://www.cpsec.aist.go.jp/achievements/aiqm/AIQM-Guideline-1.0.1-en.pdf .O

OECD: Organisation for Economic Co-operation and Development (2019), *OECD Recommendation of the Council on Artificial Intelligence.* OECD/LEGAL/0449, May 2019.

https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.
European Parliament (2023),. Artificial Intelligence Act. P9-TA(2023)0236. https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf .

Yutaka OIWA, born in Tokyo in 1976, has received Ph.D. in Computer Science from the University of Tokyo, Japan in March 2005. After joined the National Institute of Advanced Industrial Science (AIST) in April 2005, he has been serving as a researcher and research team leaders in the areas of software science and software security. From 2018, he leads a research project for machine learning quality management funded by NEDO Japan. From April 2021, he is a deputy director of the Digital Architecture Research Center in AIST.