A Conceptual Study on Application of Human-Centered Explainable Artificial Intelligence to Nuclear Power Plant

Young Do Koo [a,c], Sa Kil Kim [a], Yonggyun Yu [b], Man Gyun Na [c*]
[a]*SMART Reactor Technology Development Division, Korea Atomic Energy Research Institute, 989-111 Daedeok-daero, Yuseong-gu, Daejeon, Republic of Korea 34057*
[b]*Artificial Intelligence Application & Strategy Team, Korea Atomic Energy Research Institute, 989-111 Daedeok-daero, Yuseong-gu, Daejeon, Republic of Korea 34057*
[c]*Department of Nuclear Engineering, Chosun University, 309 Pilmun-daero, Dong-gu, Gwangju, Republic of Korea 61452*
[*]*Corresponding author: magyna@chosun.ac.kr*

1. Introduction

In recent years, artificial intelligence (AI), specifically machine learning including deep learning, is constantly used in various studies for operation support (e.g., detection, diagnosis, prediction/forecasting, autonomous control, etc.) in the nuclear engineering field. In most of those studies, machine learning becomes a method functioning as similar as or better than human-level. Therefore, machine learning-based AI emerges as a potential technology for nuclear power plant (NPP) operation support.

However, machine learning has its intrinsic problem, generally called 'black-box'. Here, black-box is described as opacity of inner workings of machine learning, using inputs and creating an output. Explainable artificial intelligence (XAI) was introduced to resolve the black-box characteristic of the machine learning. XAI makes black-box transparent by explaining inner workings of a machine learning model in various ways.

With introduction of various XAI techniques, human can interpret machine learning models. However, it is more helpful and understandable for a developer than an operator recently. That is, information presented by XAI is more useful for developing, diagnosing, or optimizing a machine learning model. As aforementioned above, it is feasible that machine learning can provide operators with the operation supporting information. However, the operation supporting information from machine learning may be ineffective if rationale for the AI information is not clear, namely, an operator cannot have reliability on the machine learning.

Therefore, the study is performed to suggest human-centered XAI (HCXAI), which presents understandable explanation from the viewpoint of NPP operators. The main purpose of the suggested HCXAI is to help intuitive understanding of an AI system and decision-making of an operator. In this paper, HCXAI for NPP operation support is addressed at concept level. The brief overview of the suggested HCXAI and implementation examples at the concept level are included.

2. Human-Centered Explainable Artificial Intelligence

2.1 Explainable Artificial Intelligence

The description of XAI can differ. Generally, XAI is described as a method or a technique to help human to understand or trust an AI model by resolving a black-box problem in common. In aspect of terms related to XAI, XAI is proposed to achieve trustworthiness, causality, transferability, informativeness, confidence, fairness, accessibility, interactivity, privacy awareness, and so on [1].

There are a variety of XAI techniques to explain an AI model, but not all the XAI techniques are able to be applied to any AI techniques. In the studies by A. Arrieta et al. [1] and V. Belle and I. Papantonis [2], taxonomy for XAI is presented. In summary, XAI techniques are sorted according to AI model type (e.g., transparent models vs. opaque models), application type (e.g., model-agnostic vs. model-specific), and explanation type (e.g., explanation by simplification, feature relevance explanation, local explanation, visual explanation, text explanation, or explanation by example) [1,2].

To evaluate the effectiveness of an XAI model, several dimensions, as evaluation criteria, in terms of explainability are considered: comprehensibility, fidelity, accuracy, scalability, generality, and so on [2]. In Defense Advanced Research Projects Agency (DARPA), five dimensions (i.e., user satisfaction, mental model, task performance, trust assessment, and correctability) are considered to evaluate the explanation effectiveness of XAI techniques [3].

2.2 XAI in terms of Human-Centered

In the AI field, the term 'human-centered' is emerging, and accordingly human-centered AI or human-centered machine learning are frequently used [4,5]. However, there is no unifying meaning for the term 'human-centered', so its meanings are diverse. The term 'human-centered' in the study refers to understandable and meaningful for human. Here, human is focused on end-user, namely, NPP operator.

With introduction of XAI, the black-box characteristics of machine learning models gets transparent in many studies. That is, the inner workings of a machine learning model were explained in ways to calculate feature's relevance, visualize important features, and so on using XAI techniques. Explanation from XAI techniques was mainly utilized to optimize a machine learning model to be developed. Hence, it seems that XAI was focused on a developer.

HCXAI aimed in the study is to give an explanation, which is useful to understand a developed machine learning model and meet criteria to evaluate effectiveness for a user such as comprehensibility [2] or user satisfaction [3] while keeping high accuracy of the AI model. Here, comprehensibility refers to the extent to which extracted representations are humanly comprehensible. User satisfaction denotes clarity or utility of the explanation, which are rated by users [2,3].

3. Implementation Examples of Human-Centered XAI Concept

Until now, there is no specific method or technique for HCXAI, but only theory-driven approach to HCXAI for AI users as well as business owner, administrator, or regulatory bodies, which is similar to HCXAI suggested in the study, has been represented [5]. There is correlation between performance and explainability (or transparency) of an AI model [3]. That is, an AI model with high-level performance has low-level of explainability. Safety is the highest level of goals in NPP; thus, high-level performance (e.g., high accuracy) of a technique is essential.

For these reasons, the concept of the suggested HCXAI will be tried to be implemented applying a deep learning method and the existing XAI technique in order to achieve understandable explainability as well as high accuracy.

4. Discussion

There are several considerations in developing and implementing a HCXAI model. Although the studies in which XAI is used increase, most studies are not user-focused, but developer-focused as aforementioned. One of the main reasons is that type of information required for a user may differ according to given condition, task, interest, and so on. Thus, a type of information understandable for an NPP operator is needed to be determined. Modelling methods are also an important factor for an HCXAI model. Optimal AI and XAI methods are needed since an HCXAI model desired in the study is aimed to sort human-centered explanation information by XAI while maintaining high accuracy of AI. Operators normally focus on information based on an operation procedure to monitor and control an NPP. Thus, information acquisition and analysis of an operator for HCXAI information are needed to be shortened for usability of operation supporting information. To do this, an approach to process and display HCXAI information is also needed to be established.

5. Conclusions

In the study, HCXAI for NPP operation support is suggested to achieve not only the high accuracy of machine learning but also its reliability and usability. The main concept of the suggested HCXAI is to present rationale intuitively understandable from the viewpoint of an operator. To develop a model to implement the suggested HCXAI concept, diversity of the type of operator-centered information for NPP operation, modelling method (i.e., AI and XAI methods), and information processing and display are needed to be necessarily taken into consideration. Consequently, an HCXAI-based model for NPP operation support will be developed reflecting these considerations in the future studies.

ACKNOWLEDGEMENT

This research was supported by a grant from the Korea Atomic Energy Research Institute (KAERI) R&D Program (No. KAERI-524450-22) and National Research Foundation (NRF) of Korea grant funded by the Korean Government (MSIT) (No. NRF-2018M2B2B1065651).

REFERENCES

[1] A. B. Arrieta et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion, Vol.58, pp.82-115, 2020.

[2] V. Belle and I. Papantonis, Principles and Practice of Explainable Machine Learning, Frontiers in Big Data, Vol.4, 688969, 2021.

[3] D. Gunning and D. W. Aha, DARPA's Explainable Artificial Intelligence (XAI) Program, AI magazine, Vol.40, pp.48-58, 2019.

[4] T. Kaluarachchi, A. Reis, and S. Nanayakkara, A Review of Recent Deep Learning Approaches in Human-Centered Machine Learning, Sensors, Vol.21, 2514, 2021.

[5] Q. Liao and K. R, Varshney, Human-Centered Explainable AI (XAI): From Algorithms to User Experiences, 2021. (https://doi.org/10.48550/arXiv.2110.10790)