Inferring Severe Accident Scenarios in Nuclear Power Plants with Reinforcement Learning (RL) and Supervised Learning (SL) Approaches: Part2 RL Development

Seok Ho Song[a], Yeonha Lee, Semin Joo, Jeong Ik Lee[a]*, Sung Joong Kim[b]

[a]*Department of Nuclear and Quantum Engineering N7-1 KAIST 291 Daehak-ro, Yuseong-gu*
[b]*Dept. Nuclear Eng., Hanyang University, 222, Wangsimni-ro, Seongdong-gu, Seoul, Republic of Korea*, *Daejeon, Republic of Korea 305-338, 1812wow@kaist.ac.kr*
**Corresponding author: jeongiklee@kaist.ac.kr*

1. Introduction

Severe accidents in nuclear power plants have the potential to cause catastrophic consequences, making it crucial to develop effective methods for predicting and mitigating these events. In traditional nuclear accident analysis, accident scenarios are created and analyzed based on the likelihood of failure of a particular component, the changes in thermal hydraulic variables inside a nuclear power plant that would be caused by the failure, and the opinions of experts. However, traditional methods have their limitations, and the complex nature of nuclear accidents requires a more comprehensive approach. This is where artificial intelligence (AI) comes in.

The field of AI is rapidly advancing, and its potential for revolutionizing traditional methods of analysis and prediction cannot be overstated. In the context of nuclear accident analysis, AI can be used to enhance our understanding of complex systems and to identify patterns and anomalies that would be difficult, if not impossible, for human experts to detect. By applying machine learning algorithms to vast amounts of data, new insights can be gained that were previously hidden or overlooked, allowing for more accurate and effective prediction of nuclear accidents. In this way, AI has the potential to expand our sight beyond the limitations of traditional methods and to help us develop more comprehensive and robust methods for mitigating the catastrophic consequences of nuclear accidents.

In this study, reinforcement learning (RL), one of the AI techniques, is used. It is used to generate nuclear accident scenarios and analyze the results with MAAP, an existing safety analysis code. RL is a type of AI technique in which an agent learns how to make decisions in the environment to maximize reward signals. The agent receives feedback from the environment in the form of rewards or punishments, and uses this feedback to adjust its behavior over time [1].

These characteristics of RL lend themselves well to the process of optimizing behavior to achieve a specific goal. In the nuclear industry, research has been conducted on the use of RL to operate nuclear power plants. These studies have shown that RL agents manipulate the system for a given purpose, thus performing nuclear power plant operation [2], [3].

In this study, the RL agent generates accident scenarios by selecting components that may fail in a particular accident scenario without any special constraints. In order to generate more dangerous accident scenarios, two kinds of rewards are used. The accident scenarios generated with different rewards are finally analyzed by MAAP. In this process, the surrogate model mentioned in the Part 1 companion paper is used as an environment and provides information to the agent through high-speed computation.

2. Methods

As mentioned in the Part 1 companion paper, the surrogate model used in this study describes a TLOCCW event as a function of the parameters inside a nuclear power plant, whether components fail, and whether safety measures are implemented. In the process, the RL agent will determine when the six components will fail. The six components are: heat exchanger (HX), high pressure injection pump (HPI), low pressure injection pump (LPI), containment spray system (CSS), motor driven auxiliary feedwater (MDAFW), charging pump (CHP). The RL agent takes one agent-environment interaction and decides to either fail one component or take no action. This process is optimized through the proximal policy optimization (PPO) algorithm.

PPO is one of the RL methodologies and is considered a robust RL algorithm when dealing with high-dimensional state and action spaces. It is computationally efficient compared to other RL algorithms because it uses a simple objective function and is easy to implement, making it suitable for large-scale applications. It is a flexible algorithm that can be applied to a variety of RL tasks. It is also compatible with both on-policy and off-policy learning, allowing it to learn from both current and historical data [4].

As discussed in the introduction, two different types of rewards are utilized to train the RL agent and evaluate their sensitivity. The rewards utilized in this study include pressure reward, which pertains to the primary side pressure of the nuclear power plant, and Core Exit Temperature (CET) reward, which is associated with the core exit temperature.

The pressure reward was set based on data showing that a high peak in primary pressure occurs when a reactor pressure vessel (RPV) failure occurs. The pressure

reward is designed to obtain accident scenarios with more severe consequence by allowing a higher reward to occur if the pressure peak occurs at an earlier point in time, as shown in equation 1.

$$\textbf{Pressure reward} = \begin{cases} \Delta P_{1st} \times (remain\ time) & (\Delta P_{1st} > 0) \\ 0 & (\Delta P_{1st} < 0) \end{cases}$$

eq.1

In the case of CET reward, it is based on the severe accident management guide (SAMG) entry condition, so that the earlier the SAMG is entered, higher reward is obtained.

The RL agent is trained by repeating 1,000 episodes (i.e. scenario) using each reward. The configuration of the RL agent used is shown in Table 1.

Table 1. Configuration of PPO Agent

	Configuration	Activation Function	Output Layer Activation
Actor Network	7/128/256/128/7	ReLU	SoftMax
Critic Network	7/128/256/128/1	ReLU	Linear

The accident scenario is generated by the PPO method, which determines the behavior probabilistically [4]. Thus, 1,000 new scenarios were generated using the trained RL agent, and the most frequent failure time for each component was used as the final accident scenario. The severity of the accidents, such as the SAMG entry point, core uncovery, and fraction of clad reacted in the vessel, were compared using MAAP code to validate the final accident scenario. Mitigation strategies 1, 2, and 3 of SAMG will be activated during this process by following the SAMG.

3. Results & Discussions

Table. 2 shows when and how often the RL agent with pressure reward selected the time of component failure.

Table 2. The Most Frequent Component Failure Time and its Frequency with Pressure Reward RL Agent

	HX	HPI	LPI	CSS	MDAFW	CHP
Failure (hr)	2	15	3	1	3	1
Frequency	233	189	96	437	179	225

The time of failure of a given component in the table above is used for MAAP verification. As mentioned in the companion paper Part 1, the timing of the RCP seal LOCA will not be determined by the RL agent, but will be verified with MAAP as a case of failure at 1 hr. Scenarios determined by the RL agent using CET rewards under the same conditions are also validated. Table. 3 shows the most frequent component failures determined by the RL agent using CET rewards and their frequency.

Table 3. The Most Frequent Component Failure Time and its Frequency with CET Reward RL Agent

	HX	HPI	LPI	CSS	MDAFW	CHP
Failure (hr)	7	1	3	6	3	5
Frequency	101	888	195	113	335	126

Based on Tables 2 and 3, two accident scenarios were verified with MAAP, and the important parameters are shown in Table 4. The accident scenario generated from RL using CET reward was determined to have more severe consequence (earlier occurrence of core uncover and SAMG entry) than the scenario generated from RL using pressure reward.

Table 4. Accident Parameters

	Pressure Reward	CET Reward
Core Uncover Time (hr)	22.181	8.113
SAMG Entry Time (hr)	23.863	9.076
Fraction of Clad Reacted (%)	0.04	39.27
RPV Failure	None	None

The accident scenario generated with RL using CET reward shows faster occurrence of both Core uncovery and SAMG Entry. The reaction rate of the cladding material was also higher in the accident scenario generated with RL using CET reward. However, both accident scenarios did not have RPV failure as consequence. This was due to the activation of the mitigation measures under the SAMG condition, which prevented the RPV failure against all possible scenarios generated from RL agent. The RL using pressure reward produced an accident scenario with RPV failure when the mitigation measures are disabled, and this result will be discussed in a companion paper Part 3.

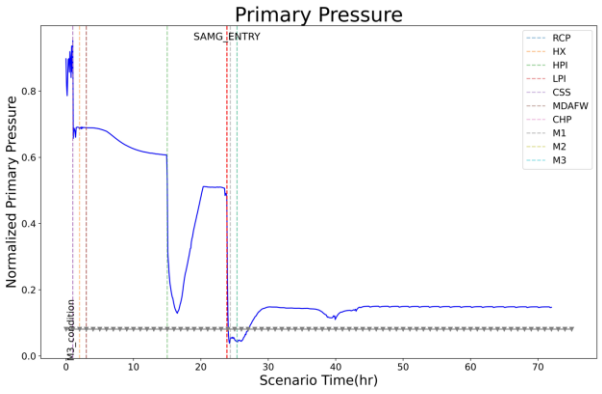

Figure 1. Primary Pressure in Pressure Reward Scenario

As shown in Figure 1, the primary pressure did not increase abruptly during the 72-hour accident, and at 15 hours into the accident, the primary pressure decreased with the HPI failure, causing the coolant inside the RPV to boil, leading to the accident scenario being determined by the reward of increasing pressure, shown in Figure 2. After the mitigation was implemented (red dotted line in Figure 2), the CET quickly decreased and stabilized.

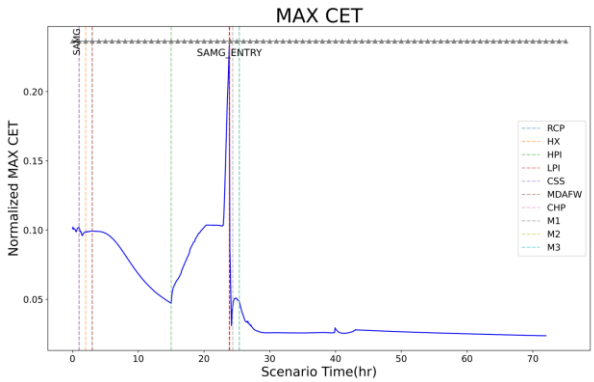

Figure 2. Maximum CET in Pressure Reward Scenario

As shown in Figure 3, the accident scenario with CET has a faster SAMG entry point (red dotted line).

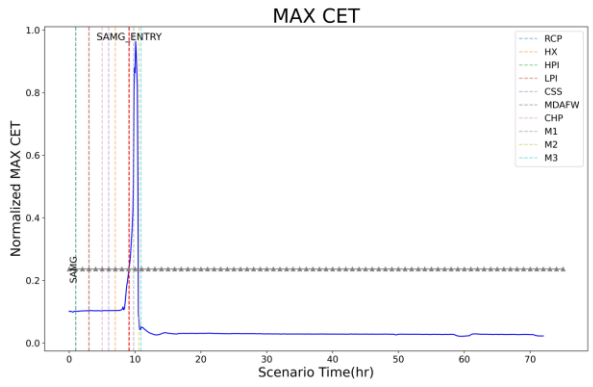

Figure 3. Maximum CET in CET Reward Scenario

In both accident scenarios, the CET decreased rapidly after the mitigation measures (red dotted lines) were activated. However, for the accident scenario generated from the RL using CET reward, the peak CET was higher compared to the accident scenario generated from the RL using pressure reward.

4. Conclusions & Further Works

In conclusion, faster core uncovery and earlier SAMG entry time, higher peak CET, and larger fraction of clad reacted values confirm that accident scenario generation using the RL with CET reward is better when severe accident mitigation measures are enabled. What is important to note from the above results is that the nature of the reward is reflected in the outcome of the nuclear accident scenarios determined by the RL agent. It is expected that accident scenario generation from RL can provide insights to the specific vulnerabilities that researchers are looking for.

However, the optimization methodology via RL is computationally intensive. The iterative nature of RL optimization is incompatible with existing thermo-hydraulic analysis tools. In this study, a surrogate model was used to address this computational load, but the process simplifies the accident description and has a low time resolution. As a result, the output of RL will inevitably be affected by the performance of the surrogate model. Improvements in the surrogate model that reduce the computational load while simplifying the accident phenomenon less are expected to improve the performance of RL.

Acknowledgment

This work was supported by KOREA HYDRO & NUCLEAR POWER CO., LTD (No. 2020-Tech-01).

REFERENCES

[1] Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. MIT press, 2018.

[2] Lee, Daeil, and Jonghyun Kim. "Autonomous emergency operation of nuclear power plant using deep reinforcement learning." *Advances in Artificial Intelligence, Software and Systems Engineering: Proceedings of the AHFE 2021 Virtual Conferences on Human Factors in Software and Systems Engineering, Artificial Intelligence and Social Computing, and Energy, July 25-29, 2021, USA*. Springer International Publishing, 2021.

[3] Park, JaeKwan, et al. "Control automation in the heat-up mode of a nuclear power plant using reinforcement learning." *Progress in Nuclear Energy* 145 (2022): 104107.

[4] Schulman, John, et al. "Proximal policy optimization algorithms." arXiv preprint ar Xiv:1707.06347 (2017).