Transactions of the Korean Nuclear Society Autumn Meeting
Changwon, Korea, October 30-31, 2025

Preliminary Study on the Application of Reinforcement Learning for Decision Making in Severe Accident Scenarios

Cheolwoong Kim [a], Joon Young Bae [a], Yujin Kim [a], JinHo Song [a], Joon Eon Yang [a], Taewoo Kim[b], Mi Ro Seo[b], Yoonhee Lee [c] and Sung Joong Kim [a,d*]
[a] *Department of Nuclear Engineering, Hanyang University,*
222 Wangsimni-ro, Seongdong-gu, Seoul 04763, Republic of Korea
[b] *Korea Hydro and Nuclear Power Central Research Institute,*
70, Yuseong-daero 1312beongil, Yuseong -gu, Daejeon, Republic of Korea
[c]*Department of Quantum System Engineering, Jeonbuk National University, 567, Baekje-daero, Deokjin-gu, Jeonju-si, Jeollabuk-do 54896, Republic of Korea*
[d] *Institute of Nano Science and Technology, Hanyang University,*
222 Wangsimni-ro, Seongdong-gu, Seoul 04763, Republic of Korea
[*]*Corresponding author: sungjkim@hanyang.ac.kr*

***Keywords: MAAP, Reinforcement Learning, Severe Accident, SAMG, PPO**

1. Introduction

The feasibility of applying artificial intelligence (AI) to the operation of nuclear power plants under severe accidents has been investigated with various types of models. Leveraging time-series observations of thermal-hydraulic variables obtained from the main control room, various neural-network models have been developed for state prediction[1,2]. With the development of AI models, forecasting the behavior of thermal-hydraulic variables has improved. With the precise predictions on the nuclear power plant status, valid result of the mitigation strategy enables the assessment of mitigation strategies.

Regarding the appropriate timing of the mitigation strategy, various trial must be conducted with simulation. In previous studies, establishing surrogate model with data acquired from the system code brings adequate agreements for constructing environment for the reinforcement learning (RL). With surrogate model developed, time consumption on the system codes such as MELCOR or MAAP could be saved. The prediction on the status of containment building integrity or the reactor pressure vessel integrity could be achieved with surrogate model. The surrogate model provides compatible environment for the RL training [3].

In this study, one of the stress test scenarios conducted in Hanul 3,4 was assessed [4]. The scenario is extended loss of alternate power (ELAP) and loss of ultimate heat sink (LOUHS). In this scenario, the only available safety components are, turbine driven auxiliary feedwater pump (TDAFW) and containment spray pump (CSP).

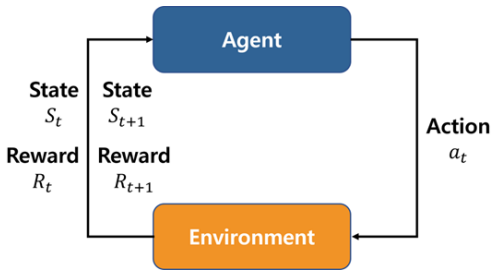

Figure 1 Concept of reinforcement learning.

2. Methodology

2.1. MAAP dataset generation

In the scenario of ELAP-LOUHS, most of the safety feature components such as high-pressure safety injection, low-pressure safety injection, motor driven auxiliary feedwater pump are unavailable. Despite the limited components available, two key mitigation strategies are for keeping the integrity of the containment building. The two mitigation strategies, TDAFW and CSP, observed in this study operates under specific circumstances such as the duration of the battery lifespan or availability of external mobility pump that could support the containment spray system (CSS). The battery duration is known to range from 4 hours to 11 hours and the external mobility pump is assumed to be ready at least 2 hours after the entrance of severe accident management guidelines (SAMG). In addition, CSS duration rely on the remaining amount of recirculation water storage tank (RWST). Therefore, the various conditions ranging from 10% available to 100% available are given to create the scenarios. The conditions assumed in this study are demonstrated in table 1.

Table 1. Various conditions for the accident

Variable	Range	Operation
TDAFW battery duration	4 ~ 11 hr	After the accident
CSP operating time	2 ~ 30 hr	After SAMG entrance
RWST water level (portion)	10 ~ 100 %	

With respect to three of the key conditions, 2320 scenarios are created. The number of scenarios were obtained by 8 possible cases in battery, 29 in CSP entrance timing, and 10 cases in RWST water level availability. With these 2320, datasets were created to train the surrogate model. The surrogate model utilized

in this study is the long-short term memory (LSTM). The concept of establishing the surrogate model is depicted in figure 2. The thermal-hydraulic variable goes through LSTM block whereas signal and state variable go through embedding block. The three variables are concatenated and passed on to LSTM to predict new thermal-hydraulic variables. Figure 3 demonstrates the key concept of this study by providing 29 cases of scenarios of 11 hours of TDAFW with RWST fully available. The example emphasizes the importance of CSP actuation timing by exhibiting both containment failure case and successfully mitigated case. The manually found optimize timing is found to be the best before 21 hours. By training RL with structured compensation system, finding the optimized timing of CSP timing is the key goal of the RL application.

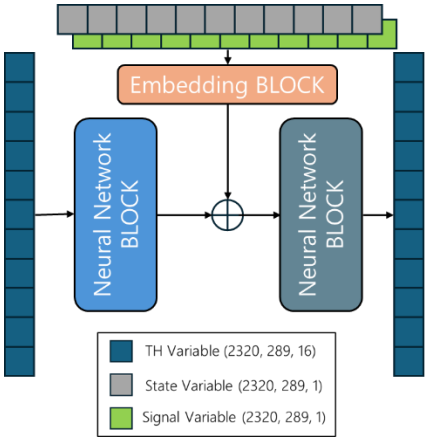

Figure 2 Establishing surrogate model

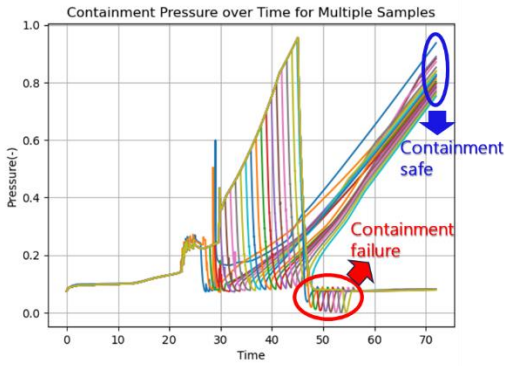

Figure 3 Pressure (normalized) behavior of 29 scenarios

2.2. Reinforcement learning

RL is an AI technique in which an agent interacts with an environment to implement a specific strategy [5]. Within the environment, the agent takes actions in order to maximize its cumulative reward. Hence, the essence of RL lies in the constructing well-designed reward system. The agent learns through Markov Decision Process (MDP), where it repeatedly selects actions and evaluates the outcomes through trial and error. The decision-making mechanism that maps states to actions is called policy. By optimizing the policy the agent improves its performance in achieving the defined objectives. The cumulative measure of success is referred as the value. Accordingly, RL approaches can be divided into value-based and policy-based method in broad sight.

In this study, for the goal is on determining the timing of mitigation strategies, policy-based RL methods are considered. Specifically, two models are compared to present the performance: Advantage Actor-critic (A2C) and Proximal Policy Optimization (PPO). In both models, the actor-critic architecture is utilized. The key difference between the two model lies in how the policy update is controlled by additional function.

2.3. A2C

The actor-critic framework combines the strengths of policy-based and value-based RL. The actor learns a policy that maps states to actions while the critic learns a value function that estimates the expected return of a state under the current policy. The critic thereby evaluates the quality of the actions and provides feedback. In A2C, the actor updates its policy directly based on the critic's evaluation (advantage function). The advantage function measures how much better (or worse) an action performs compared to the average. However, this update can be unstable due to unlimited policy changes may occur between iterations, leading to unstable calculation.

2.4. PPO

PPO is a state-of-the-art optimization method that improves upon A2C by introducing, clipping mechanism during the policy update. The clipping mechanism is retrieved by calculating probability ratio of the old and new policy. Restricting the probability ratio to the predefined threshold would stabilize the new policy calculation. As a result, PPO prevents excessively large policy updates, by constraining policy updates within this "trust region" ensuring that the actor does not deviate dramatically from the previous policy. In addition, PPO supports efficient multi-process training and balances exploration and exploitation by allowing moderate variations in updates. As a result, PPO achieves more stable convergence and higher sample efficiency compared to A2C, expecting it to be more suitable for complex environments.

2.5. Reward systems

The reward system is the key element of this study. Depending on how the reward system is structured, the RL could find the optimizing results. The key elements that need to be regarded in this study are demonstrated in Table 2. The main target of this scenario is to keep the integrity of the containment building. Therefore, when the scenario is done and the containment is not failed, the

huge reward is provided. But when the containment is failed, the same amount of disadvantage is given in negative sign. Since the goal is to find the best CSP turning timing, the reward is given to the CSP operating point as well. In order to find the optimizing time, the reward is provided as the steps are as late as possible. The maximum time step possible in this scenario is 289 (72 hours of calculation divided in 15 minutes).

Table 2. Reward system for RL

Condition	Reward
Step number	+1 for each step
CSP turn on	+ 120
Containment failure	-200
Step > 289 and Containment integrity kept	+200

3. Result and Analysis

The RL results are compared in this section. Under the same reward system, the comparison between A2C and PPO highlights the role of the clipping mechanism. Both models were trained for 20,000 episodes, and their cumulative rewards were evaluated to assess the performance differences in Fig 4.

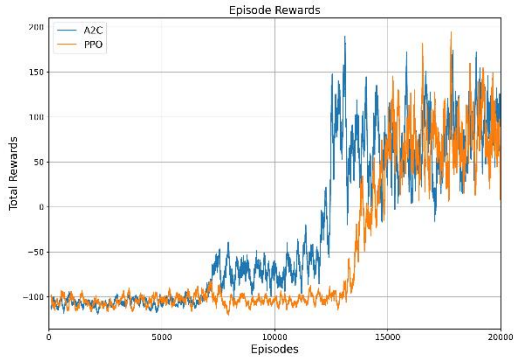

Figure 4 Comparison of rewards between A2C and PPO

Table 3. Key performance indices for A2C and PPO

Key performance index	A2C	PPO
Final Converged Reward	87.09	75.04
Learning Stability (Standard Deviation)	252.62	224.30

From figure 4, A2C exhibit faster exploration of the reward space and reaches higher peak rewards compared to PPO. The learning stability metric in Table 3 indicates that A2C explores a wider range of possibilities, whereas PPO progresses more conservatively within a stable region. For this study, where the primary goal is to identify the optimal timing for CSP operation, the faster convergence of A2C provides an advantage in seeking the solution efficiently. PPO, on the other hand, demonstrates slower but steadier improvements, reflecting its conservative update strategy due to the clipping mechanism.

4. Conclusion

In this study, two policy-based reinforcement learning methods, A2C and PPO, were applied to the problem of determining the appropriate timing for CSP operation under severe accident conditions. The results showed that A2C achieved faster exploration and higher converged rewards, making it more suitable for scenarios requiring rapid decision-making. PPO, although slower, demonstrated more stable learning behavior, reflecting its strength in long-term reliability.

The key findings from this study suggests that:

- A2C is advantageous when quick adaptation and solution retrieval are important.
- PPO may be more appropriate in common when application require stability and robustness against overfitting.

Overall, this study demonstrates the feasibility of applying RL to nuclear power plants severe accident mitigation strategies. Comparing two widely used algorithms highlights how difference policy update mechanisms can influence training outcomes. For future work, the representative value-based RL method, Deep Q-Network (DQN), will be investigated. A comparison between DQN and the policy-based methods (either PPO or A2C) is expected to provide further insights into the relative suitability of value based and policy-based approaches for predicting mitigation strategies under severe accident conditions.

Acknowledgement

This work was supported by KOREA HYDRO & NUCLEAR POWER CO., LTD (No. 2024-Tech-08). This work was supported by the Innovative Small Modular Reactor Development Agency grant funded by the Korea Government (MSIT) (No. RS-2023-00259516).

REFERENCES

[1] Bae, Joon Young, and Sung Joong Kim. "Applicability Study of Deep Learning-Based Surrogate Model to Severe Accident Simulation.", 2023

[2] Semin Joo, Yeonha Lee, Seok Ho Song, Kyusang Song, Mi Ro Seo, Sung Joong Kim, Jeong Ik Lee, Leveraging explainable AI for reliable prediction of nuclear power plant severe accident progression, Reliability Engineering & System Safety, Volume 264, Part A, 2025.

[3] Seok Ho Song, Yeonha Lee, Jun Yong Bae, Kyu Sang Song, Mi Ro Seo, SungJoong Kim, Jeong Ik Lee, Application of reinforcement learning to deduce nuclear power plant severe accident scenario, Annals of Nuclear Energy, Volume 205, 2024

[4] 한울 3,4 호기 스트레스 테스트 수행보고서 (공개본), 한수원, 2018

[5] Junyong Bae, Jae Min Kim, Seung Jun Lee, Deep reinforcement learning for a multi-objective operation in a nuclear power plant, Nuclear Engineering and Technology, Volume 55, Issue 9, 2023