RESEARCH Open Access

A digital twin-based system for full-lifecycle safety management and dynamic risk assessment in nuclear power plants

Hongcheng Yu[1]*

*Correspondence:
Hongcheng Yu
tougao2025@outlook.com
[1]Hainan NPP Project Management Department, China Nuclear Power Engineering Co., Ltd, Changjiang Li Autonomous County 572735, Hainan, China

Abstract
In order to respond to the complex safety problems involved with the lifecycle of Nuclear Power Plants (NPPs) and to eliminate the bottleneck of traditional static risk assessment, this paper proposes a new safety management framework based on Digital Twin (DT) technology and deep reinforcement learning. Legacy paradigms of safety management fail in high-risk, dynamic environments, such as planned refueling outages and unplanned fire events, due to the latency of the data and rigid decision-making. To this end, the current work proposes first the "Nuclear Safety Twin Construction and Integration Framework" (NST-CIF) to build systematically a high-fidelity virtual environment time-synchronized with the real plant and coupled across multiple physics domains. With large-scale experiments performed in the Digital Twin environment on two critical scenarios—dynamic refueling outage scheduling and emergency response to fire accidents—the results indicate that Proactive Safety Policy Optimization (PSPO) completely outperforms both traditional procedure-based policies and naive deep reinforcement learning algorithms both in operational efficiency and safety. This research presents a tested and potentially effective technological pathway for the transformation of NPP safety management from passive, reactive to intelligent, proactive, and predictive.

Keywords Digital twin, Nuclear power plant, Safety management, Dynamic risk assessment, Reinforcement learning, Proactive safety policy optimization (PSPO), Graph neural network

1 Introduction

The demand for clean and reliable energy has made the position of Nuclear Power Plants (NPPs) a cornerstone of global energy policy, as a safe, low-carbon alternative to fossil fuels [1, 2]. However, ensuring the operational integrity and safety of these complex, high-consequence systems throughout their multi-decade lifecycle is a challenging engineering and management endeavor [3–5]. The increasing complexity of modern plant schematics, coupled with the extreme constraints of regulatory adherence and financial pressures for operational effectiveness, necessitates a paradigm shift from traditional management approaches to more holistic, smarter, and predictive ones [6, 7].

Monitoring, control, and maintenance systems are vital, with failures having disastrous implications, so the search for emerging management technologies is a necessity rather than an option [8, 9].

Despite a history of decades of safe operations, safety management in NPPs is generally constrained by methods founded on static, deterministic, and probabilistic analysis [10]. Traditional Probabilistic Safety Assessment (PSA), while pioneering, typically relies on time-averaged failure rates and prespecified event sequences, and struggles to capture the real-time dynamic nature of risk as the plant state evolves [11, 12]. This kind of limitation becomes particularly severe for complex operational transients, e.g., planned refueling outages or unusual events such as a fire accident [13, 14]. Optimizing the intricate network of maintenance tasks during a shutdown to minimize downtime while meeting all safety specifications is a high-dimensional scheduling problem that strains conventional planning abilities [15]. Similarly, firefighting entails rapid, context-specific decision-making that considers the dynamic propagation of the fire and its cascading impact on functionally coupled safety systems—a scenario for which pre-defined, procedure-based responses are less than optimum [16]. These circumstances refer to an intrinsic mismatch: running a dynamic, evolving system using static, fragmented data and models [17].

The issue is also compounded by the dynamic nature of the plant itself. Component wear, equipment aging, and subtle operational differences are continuously modifying the risk profile of the plant in ways that are difficult to keep track of by using periodic inspection and time-based maintenance schedules alone [18, 19]. The need for methods that can dynamically assess system health, predict failures, and optimize operation strategies in real-time is critical [20]. It calls for an integrated framework that can break down data silos between design, operation, and maintenance, creating a live, integrated view of the status and risk profile of the plant [21].

The convergence of Digital Twin (DT) technology and Cutting-Edge Artificial Intelligence (AI), in the form of Deep Reinforcement Learning (DRL), presents a compelling framework to address these complex issues [22–24]. A Digital Twin, a virtual high-fidelity representation of a physical asset coupled with real-time synchronized data, offers a unique platform for simulation, analysis, and optimization [25]. By fusing multi-physics models with real-time sensor information, a DT can provide a comprehensive picture of the plant's current and future state [26]. DRL, having already shown the ability to solve complex sequential decision-making problems, can use this DT framework to learn optimal control and management policies that balance the competing objectives of safety, reliability, and economic efficiency [27, 28]. Application of DRL on dynamic optimization problems, from power systems to robotics, has demonstrated its ability in discovering non-intuitive strategies that outperform heuristics devised by humans and traditional control policies.

However, direct application of such advanced technologies on the safety-critical nuclear power application is not trivial. The "trial-and-error" nature of typical RL is not permissible in a real NPP. Therefore, any learning must occur in a validated, high-fidelity virtual environment. To this aim, this paper presents a new full-lifecycle safety management framework based on a Proactive Safety Policy Optimization (PSPO) methodology within a specially designed NPP Digital Twin. Our framework contributes in several important ways: firstly, it outlines a structured method (the NST-CIF) for the development of a multi-fidelity DT that faithfully captures plant dynamics. Second, it uses a Graph Neural Network (GNN) to decode the complex, networked topology of the NPP, enabling richer system state understanding. Third, it introduces a safety-constrained DRL algorithm that uses a "Risk Shield" to ensure both the agent's exploration and learned policies always meet prespecified safety constraints. Through large-scale experiments simulating both outage management and fire response scenarios, we demonstrate that our framework learns policies that are empirically safer and more efficient than standard methods and popular RL baselines.

The remainder of this paper is structured as follows: Chapter 2 provides the theoretical background of Digital Twins, Dynamic Risk Assessment, and the MDP modeling of the safety management problem. Chapter 3 elaborates on the proposed PSPO method in detail, comprising its architectural components and algorithmic operation. Chapter 4 presents comprehensive experimental analyses and results. Chapter 5 concludes the paper by summarizing the key findings and suggesting promising avenues for future research.

2 Theoretical foundations and system modeling

This chapter aims to establish a robust and comprehensive theoretical and mathematical foundation for the proposed Digital Twin (DT)-based system for full-lifecycle safety management and dynamic risk assessment in Nuclear Power Plants (NPPs). The content is structured into three core sections. First, the system architecture and multidimensional modeling principles of Digital Twins are elaborated. Second, the framework of Dynamic Risk Assessment (DRA) is detailed as an evolution of traditional Probabilistic Safety Assessment (PSA), highlighting its synergistic integration with the DT paradigm.

2.1 Digital twin system architecture and multidimensional modeling

A Digital Twin is a comprehensive technological paradigm that integrates the Internet of Things (IoT), multi-physics simulation, big data analytics, and Artificial Intelligence (AI). Its primary objective is to create a virtual, high-fidelity replica of a physical entity that remains synchronized in real-time, enabling a closed-loop interaction between the physical and digital worlds. A complete DT system comprises four fundamental components—the physical entity, the virtual model, the twinning data, and the service application layer—realized through a multi-layered architecture. The model layer, the core of the DT, integrates geometric, physical, behavioral, and rule-based models to form a holistic digital representation of the NPP [29].

The efficacy of a DT system is critically dependent on the fidelity of the virtual model, which quantifies the consistency between the virtual representation and its physical counterpart. We define the synchronization error, $E_{\text{sync}^\circ}(t)$, to measure this characteristic. Let the state vector of the physical entity be $S_P(t)$ and that of the virtual model be $S_V(t)$. The synchronization error is then defined as the Euclidean norm of their difference:

$$E_{\text{sync}^\circ}(t) = \|S_P(t) - S_V(t)\|_2 \tag{1}$$

To minimize $E_{\text{sync}^\circ}(t)$, the system must continuously calibrate and correct the virtual model using real-time data, Z_t, acquired from the physical entity. This process, known as data assimilation, can be implemented using various state estimation techniques. For instance, the Kalman Filter provides a recursive solution to this problem. If $S_V(t|t-1)$ is the a priori state estimate predicted from the previous time step, the updated a posteriori state estimate $S_V(t|t)$ is given by:

$$S_V(t|t) = S_V(t|t-1) + K_t\left(Z_t - HS_V(t|t-1)\right) \tag{2}$$

As shown in (2), H is the observation matrix that maps the true state space into the observed space, and K_t is the optimal Kalman gain, which minimizes the a posteriori error covariance. This gain term judiciously balances the uncertainty of the model's prediction with the uncertainty of the sensor measurements. Through such closed-loop feedback mechanisms, the DT model continuously converges toward the true state of the physical asset, providing a reliable foundation for decision-making.

Figure 1 illustrates the digital twin topology of a key subsystem of the system—the Essential Service Water System (SEC). In this digital twin, each critical piece of equipment (such as the pump SEC001PO, valve SEC101VE, etc.) exists as an independent intelligent node. By real-time data acquisition of the operational data of these nodes and integrating it with preset physical models and operational rules, the system is capable of dynamically assessing the overall health status and safety risks of the entire system, thereby providing decision-making support for proactive safety policy optimization. Subsequent chapters will elaborate on the theoretical modeling, implementation methods, and experimental verification of the system, all revolving around this core concept.

2.2 Digital twin-based dynamic risk assessment framework

While foundational to nuclear safety analysis, traditional Probabilistic Safety Assessment (PSA) methods are inherently static, incapable of reflecting the dynamic evolution of risk as a function of time, operational conditions, and equipment aging. Dynamic Risk Assessment (DRA) aims to overcome this limitation, and the Digital Twin provides the ideal technological platform for its implementation. Within the DRA framework, the failure probabilities of Systems, Structures, and Components (SSCs) are not treated as constant values derived from historical statistics but as dynamic functions dependent on their real-time state.

For example, the instantaneous failure rate, or hazard rate, $\lambda(t)$, of a component can be described by reliability theory. The hazard rate is defined as:

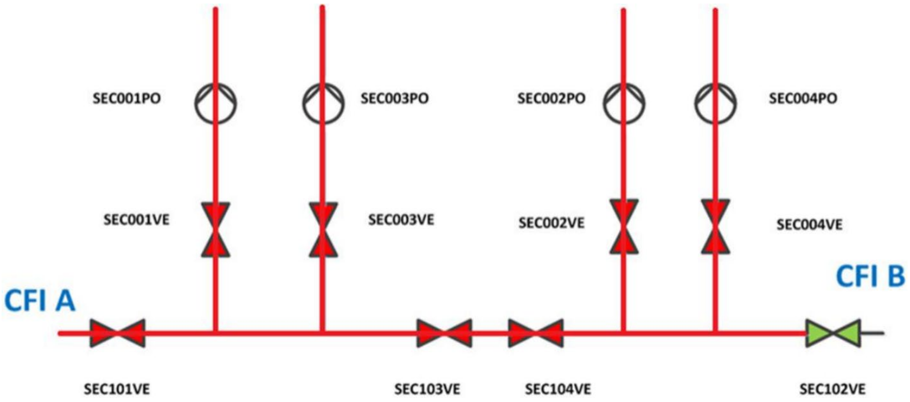

Fig. 1 DT of essential service water system (SEC)

$$\lambda(t) = \frac{f(t)}{R(t)} \tag{3}$$

where $f(t)$ is the failure time probability density function (PDF) and $R(t)$ is the reliability function, representing the probability of the component not failing before time t, such that $R(t) = 1 - F(t) = \int_t^\infty f(\tau)\,d\tau$, where $F(t)$ is the cumulative distribution function (CDF). In the DT system, we can employ models like the Weibull distribution, whose PDF is given by $f(t) = \left(\frac{\beta}{\eta}\right)\left(\frac{t}{\eta}\right)^{\beta-1} e^{-(t/\eta)^\beta}$. The shape parameter β and scale parameter η can be dynamically updated based on real-time monitoring data (e.g., vibration, temperature), yielding a dynamic hazard rate $\lambda(t \mid \beta(t), \eta(t))$.

Ultimately, the plant-level core risk metric, such as the Core Damage Frequency (CDF), can also be dynamized. The static CDF is typically calculated as $\mathrm{CDF} = \sum_i I_i \cdot P(S_i)$, where I_i is the frequency of initiating event i, and $P(S_i)$ is the conditional probability of core damage given that event. In our DRA framework, this evolves into a time-dependent function:

$$\mathrm{CDF}(t) = \sum_i I_i(s_t) \cdot P(S_i \mid s_t) \tag{4}$$

As shown in (4), both the initiating event frequency $I_i(s_t)$ and the conditional failure probability $P(S_i \mid s_t)$ are dependent on the current, real-time state of the plant, s_t. This capability for dynamic evaluation transforms risk management from a "post-mortem" reactive process to a "prognostic" proactive strategy, enabling pre-emptive actions to mitigate nascent risks.

2.3 Markov decision process modeling for safety management

To achieve automation and optimization in safety-critical decision-making, we formalize the NPP lifecycle safety management problem as a Markov Decision Process (MDP) [30]. An MDP provides a powerful mathematical framework for sequential decision-making under uncertainty and is defined by a five-tuple (S, A, P, R, γ).

- **State Space (S)**: A state $s_t \in S$ is a comprehensive, high-dimensional representation of the NPP's safety status at time t, sourced directly from the DT. It is a vector $s_t = [D_{\text{sensor}}, H_{\text{equipment}}, C_{\text{operation}}, R_{\text{risk}}]$, containing sensor data, equipment health indices, operational modes, and the real-time risk level calculated by the DRA.
- **Action Space (A)**: An action $a_t \in A$ represents a decision that can be executed by the safety management system, such as {Activate Pump A, Adjust Valve B opening, Execute Procedure C, Schedule maintenance for Component D}.
- **Transition Probability (P)**: $P(s_{t+1} \mid s_t, a_t)$ defines the probability of transitioning to state s_{t+1} after taking action a_t in state s_t. In this context, the transition dynamics are governed by the complex physics and logic within the DT simulation engine.
- **Reward Function (R)**: The reward function $R(s_t, a_t)$ quantifies the immediate desirability of a decision, designed to balance safety, economy, and reliability, while enforcing the "safety-first" principle:

$$\begin{aligned} &R(s_t, a_t) = \\ &w_{\text{avail}\cdot}\,\mathrm{Avail}^{\cdot}(s_t) - w_{\text{cost}\cdot}\,\mathrm{Cost}(a_t) - w_{\text{risk}\cdot}\,\mathrm{Risk}(s_t) - w_{\text{penalty}\cdot}\,P_{\text{op}\cdot}(a_t, s_t) \end{aligned} \tag{5}$$

where $\mathrm{Avail}(s_t)$ is the power generation revenue, $\mathrm{Cost}(a_t)$ is the cost of the action, $\mathrm{Risk}(s_t)$ is the DRA-calculated risk value, and $P_{\mathrm{op}\cdot}(a_t, s_t)$ is a penalty term for violating operational procedures. The weighting coefficients w are tuned to reflect the prioritization of objectives.

- **Discount Factor** (γ): $\gamma \in [0, 1)$ is a hyperparameter that discounts future rewards, balancing long-term and short-term gains.

The system's objective is to find an optimal policy $\pi^* : S \to A$ that maximizes the expected cumulative discounted reward. This is achieved by solving the Bellman Optimality Equation. For the action-value function (Q-function), $Q^*(s, a)$, the Bellman equation is:

$$Q^*(s,a) = \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[R(s,a) + \gamma \max_{a' \in A} Q^*(s', a') \right] \tag{6}$$

Equation (6) asserts that the optimal value of a state-action pair is the sum of the immediate reward and the discounted optimal value of the best possible subsequent state. By employing Reinforcement Learning (RL) algorithms to solve this equation, an intelligent agent can learn the optimal safety management and risk control policy, π^*, through extensive interaction with the DT environment.

3 Proactive safety policy optimization (PSPO) in DT environment

Building upon the Markov Decision Process (MDP) model established in Chapter 2, this chapter introduces a comprehensive methodology termed "Proactive Safety Policy Optimization" (PSPO), designed to solve for the optimal safety management policy, π^*. The direct application of conventional trial-and-error reinforcement learning (RL) algorithms on a physical Nuclear Power Plant is axiomatically prohibited due to profound safety implications. Consequently, the entire process of policy learning and optimization must be conducted within a high-fidelity Digital Twin (DT) environment. The PSPO framework is shown as Fig. 2. This chapter will first detail the specific framework designed for constructing this DT environment, followed by an exposition of how the PSPO algorithm achieves safe, efficient, and robust policy learning within this virtual sandbox.

3.1 Nuclear safety twin construction and integration framework (NST-CIF)

To provide a reliable, precise, and computationally tractable environment for the PSPO agent, we have designed the "Nuclear Safety Twin Construction and Integration Framework" (NST-CIF). This framework systematically guides the construction of the NPP's

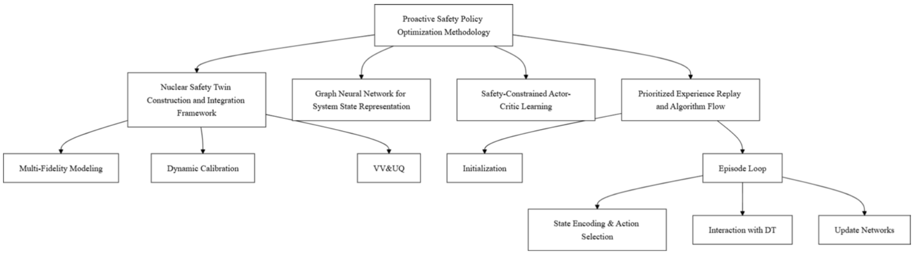
Proactive Safety Policy Optimization Methodology
Nuclear Safety Twin Construction and Integration Framework
Graph Neural Network for System State Representation
Safety-Constrained Actor-Critic Learning
Prioritized Experience Replay and Algorithm Flow
Multi-Fidelity Modeling
Dynamic Calibration
VV&UQ
Initialization
Episode Loop
State Encoding & Action Selection
Interaction with DT
Update Networks

Fig. 2 Proposed PSPO framework

digital twin, ensuring it aligns with the physical plant not only geometrically but also in its physical behavior, operational logic, and risk evolution. The NST-CIF comprises three core phases: Multi-Fidelity Modeling and Fusion, Data-Driven Dynamic Calibration, and Verification, Validation, and Uncertainty Quantification (VV&UQ).

The inherent complexity of an NPP necessitates a multi-scale, multi-physics, and multi-fidelity modeling strategy. This begins with geometric and topological modeling, where high-precision 3D models are built from CAD, BIM, and P&ID schematics, from which the system's topological graph is extracted. This is followed by multi-physics modeling, where specialized simulation codes are employed for different domains: Monte Carlo codes (e.g., MCNP) for neutronics, system-level codes (e.g., RELAP5) and CFD for thermal-hydraulics, and FEA for structural mechanics. To predict component degradation, both physics-based models (e.g., Paris's law for fatigue crack growth) and data-driven models (e.g., LSTMs for RUL prediction) are developed.

Recognizing that no single model can capture all requisite phenomena, NST-CIF employs model fusion techniques. This allows for the coupling of models with varying levels of fidelity to balance computational efficiency and accuracy. For instance, the fused state prediction, $\hat{S}_{\text{fused}^\circ}$ can be represented as a dynamic, state-dependent weighted average of predictions from N different models:

$$\hat{S}_{\text{fused}^\circ}(s_t) = \sum_{i=1}^{N} w_i(s_t) \cdot \hat{S}_i(s_t) \tag{7}$$

where $\hat{S}_i(s_t)$ is the prediction from model i, and the weight $w_i(s_t)$ reflects the confidence or applicability of that model in the current system state s_t. Bayesian Model Averaging (BMA) is a principled approach for determining these weights based on each model's posterior probability given the observed data.

To ensure real-time synchronization, NST-CIF implements a continuous, data-driven calibration loop. This involves comprehensive data acquisition via an Industrial IoT (IIoT) sensor network, followed by preprocessing and feature extraction. The core of this phase is real-time model calibration, which utilizes data assimilation techniques to integrate live data into the simulation models, dynamically correcting their states and parameters. For example, the Extended Kalman Filter (EKF) can be used to update the parameter set θ of a degradation model based on real-time observations Z_t. The update rule for the parameter estimate $\hat{\theta}_t$ is:

$$\hat{\theta}_t = \hat{\theta}_{t-1} + K_t(Z_t - h(\hat{\theta}_{t-1})) \tag{8}$$

where $h(\cdot)$ is the (potentially non-linear) observation function and K_t is the Kalman gain. This process ensures that predictions, such as Remaining Useful Life (RUL), remain accurate over time: $\text{RUL}(t) = f_{\text{model}^\cdot}(H(t) \mid \theta_{\text{calibrated}^\cdot}(t))$.

A digital twin is untrustworthy without a rigorous VV&UQ process. Verification ensures the simulation code correctly solves the mathematical equations. Validation confirms that the simulation solves the correct equations by comparing its output against experimental or historical plant data. We quantify the validation level using metrics like the Mean Absolute Percentage Error (MAPE):

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{9}$$

where A_t is the actual value and F_t is the forecasted value. A model is considered validated only if its MAPE is below a predefined threshold. For this study, a threshold of 1% was established, which was successfully achieved during the validation phase. Finally, Uncertainty Quantification (UQ) identifies and propagates all sources of uncertainty (parametric, structural, algorithmic) through the model, often using techniques like variance-based sensitivity analysis (e.g., calculating Sobol indices) to produce predictions with associated confidence intervals, which is crucial for risk-informed decision-making.

The Digital twin model test is carried through industrial PyroSim, shown as Fig. 3.

3.2 Graph neural network (GNN) for system state representation

An NPP is a complex network of thousands of interconnected components. Flattening the system state s_t into a vector, as is common in traditional RL, discards the crucial topological and relational information inherent in the plant's structure. To capture these complex systemlevel dependencies, we employ a Graph Neural Network (GNN) to encode the plant's real-time state. We abstract the system at time t as a dynamic graph $G_t = (V, E, X_V, X_E)$, where nodes V represent components and edges E represent physical or logical connections.

The GNN learns a deep representation (embedding) of the system state through an iterative message-passing mechanism. A prominent example is the Graph Convolutional Network (GCN), where the hidden representation matrix of all nodes at layer $k+1$, $I^{(k+1)}$, is computed as:

$$H^{(k+1)} = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(k)}W^{(k)}) \tag{10}$$

where $\hat{A} = A + I$ is the adjacency matrix with added self-loops, $\hat{D}$ is the diagonal degree matrix of $\hat{A}$, $W^{(k)}$ is the trainable weight matrix for layer k, and σ is a non-linear activation function. After K layers of message passing, a READOUT function (e.g., global mean pooling) aggregates the final node representations to produce a holistic graph-level embedding, $\phi(s_t)$, which serves as a feature-rich input for the RL agent.

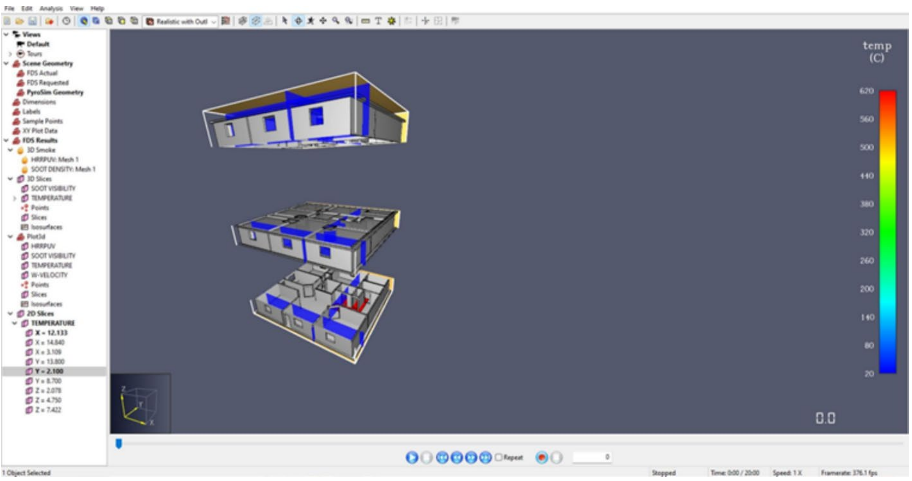

Fig. 3 Digital twin model test

3.3 Safety-constrained actor-critic learning

We employ an Actor-Critic architecture to handle the complex decision space. The key innovation is a "Risk Shield" mechanism to ensure exploration adheres to nuclear safety boundaries. The Critic network, $Q_w(s, a)$, approximates the action-value function and is updated by minimizing the TD error. The Actor network, $\pi_\theta(a \mid s)$, parameterizes the policy and is updated via policy gradients. The Risk Shield is a separately trained neural network, $C_\phi(s, a)$, that predicts the immediate risk of taking action a in state s, i.e., $C_\phi(s, a) \approx \mathbb{E}_{s' \sim P(\cdot|s,a)}[\text{Risk}(s')]$. The Risk Shield network is trained via supervised learning on data from DT simulations. We formulate the policy optimization as a constrained problem:

$$\begin{aligned} \max_{\theta} \quad & \mathbb{E}_{s\sim\rho^\pi, a\sim\pi_\theta}[R(s, a)] \\ \text{subject to} \quad & \mathbb{E}_{s\sim\rho^\pi, a\sim\pi_\theta}[C_\phi(s, a)] \leq \tau_{\text{risk}} \end{aligned} \tag{11}$$

where ρ^π is the state distribution induced by policy π and τ_{risk} is a predefined safety threshold. This constrained problem can be solved using Lagrangian methods, leading to the Actor's modified objective function, which incorporates a penalty for violating the safety constraint:

$$J_{\text{safe}}(\theta) = \mathbb{E}_{s_t\sim D}\left[Q_w(s_t, a_t)\Big|_{a_t\sim\pi_\theta} - \lambda_C \cdot \max\left(0, C_\phi(s_t, a_t) - \tau_{\text{risk}}\right)\right] \tag{12}$$

where λ_C is a large penalty coefficient (or a learned Lagrange multiplier). By maximizing this objective, the Actor learns to avoid actions that are predicted to lead to high-risk states.

3.4 Prioritized experience replay and algorithm flow

To enhance learning efficiency, we utilize Prioritized Experience Replay (PER). Instead of uniform sampling, PER samples transitions from the replay buffer D based on their TD error, $|\delta_i|$, giving higher priority to more "surprising" experiences. The probability of sampling transition i is:

$$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha} \tag{13}$$

where $p_i = |\delta_i| + \epsilon$ is the priority, and α controls the degree of prioritization. To correct for the bias introduced by this non-uniform sampling, we use importance-sampling (IS) weights when updating the network parameters:

$$w_i = \left(\frac{1}{N \cdot P(i)}\right)^\beta \tag{14}$$

The Critic loss is then modified to:

$$L(w) = \mathbb{E}_{i\sim P(i)}\left[w_i \cdot (y_i - Q_w(s_i, a_i))^2\right] \tag{15}$$

where β is a hyperparameter that anneals from 0 to 1. The complete PSPO algorithm, integrating all these components, is detailed in the pseudocode below. The overall PSPO for full-lifecycle safety management and dynamic risk assessment in nuclear power plants is shown in Fig. 4.

```
Require:
  Initialize Actor network $\pi_\theta$, Critic network $Q_w$, and Risk Shield network $C_\phi$ with random parameters $\theta, w, \phi$.
  Initialize target networks with $\theta' \leftarrow \theta, w' \leftarrow w$.
  Initialize experience replay buffer $D$.
Ensure:
  Obtain the initial state graph $G_0$ from the Digital Twin environment.
  for $episode = 1$ to $M$ do
      for $t = 1$ to $T$ do
          // State Encoding & Action Selection
          Encode state graph $G_t$ via GNN to get representation $\phi(s_t)$.
          Select action $a_t = \pi_\theta(\phi(s_t)) + \mathcal{N}_t$, where $\mathcal{N}_t$ is exploration noise.
          // Interact with Digital Twin Environment
          Execute action $a_t$ in the DT, observe reward $r_t$ and next state graph $G_{t+1}$.
          Store transition $(G_t, a_t, r_t, G_{t+1})$ in replay buffer $D$.
          // Sample from Replay Buffer and Update Networks
          Sample a minibatch of $N$ transitions $(G_j, a_j, r_j, G_{j+1})$ and IS weights $w_j$ from $D$ using PER.
          for each transition $j$ in the minibatch do
              Encode $G_j \rightarrow \phi(s_j)$ and $G_{j+1} \rightarrow \phi(s_{j+1})$ using GNN.
              // Update Critic Network
              Set TD target $y_j = r_j + \gamma Q_{w'}(\phi(s_{j+1}), \pi_{\theta'}(\phi(s_{j+1})))$.
              Update Critic weights $w$ by minimizing the weighted loss:
                  $L(w) = \frac{1}{N} \sum_j w_j \left(y_j - Q_w(\phi(s_j), a_j)\right)^2$
              // Update Actor and Risk Shield Networks
              Update Actor weights $\theta$ by performing a gradient ascent step.
              Update Risk Shield weights $\phi$ by minimizing its prediction loss on simulated risk data.
              Update PER priorities $p_j$ for the sampled transitions based on their new TD errors $|\delta_j|$.
              // Soft update target networks
              $w' \leftarrow \tau w + (1 - \tau) w'$
              $\theta' \leftarrow \tau \theta + (1 - \tau) \theta'$
          end for
      end for
  end for
```

Algorithm 1 Proactive Safety Policy Optimization (PSPO)

4 Experiments and results

This chapter presents a comprehensive empirical evaluation of the Proactive Safety Policy Optimization (PSPO) framework. The primary objective is to validate its effectiveness and superiority in comparison to established baseline methods across representative safety management scenarios within a nuclear power plant. All experiments are conducted within the high-fidelity Digital Twin (DT) environment, which has been rigorously validated for its fidelity against a reference plant model. The evaluation is designed to assess the performance of the learned policies in terms of cumulative reward, safety adherence, and operational efficiency.

4.1 Digital twin fidelity validation

Before its use as a training environment, the fidelity of the Digital Twin, constructed via the NST-CIF, was rigorously validated. We selected a reference Pressurized Water

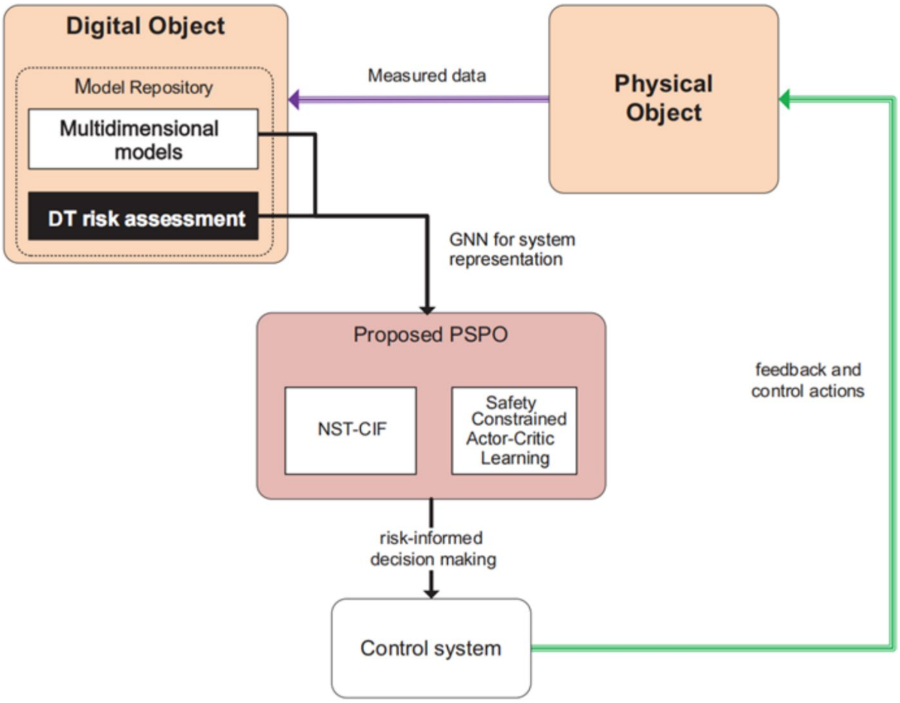

Fig. 4 PSPO for full-lifecycle safety management and dynamic risk assessment in nuclear power plants

Reactor (PWR) plant model and simulated a standard operational transient: a planned power reduction from 100% to 75% over 30 min. We monitored 20 critical plant parameters and compared the DT's simulated outputs against the "ground truth" data from the reference model. The selected parameters include key temperatures and pressures in the primary loop, as well as neutron flux and steam generator levels, which are indicative of the plant's power status. The result is shown as Table 1.

The validation results, summarized in Table 1, demonstrate a high degree of fidelity, with the average relative error across all points being less than 0.5%. This high level of accuracy confirms that the DT environment is a reliable and representative surrogate of the physical plant, suitable for training and evaluating RL agents.

4.2 Scenario parameters, baselines, and evaluation metrics

To rigorously evaluate the PSPO framework, all experiments were conducted within the validated Digital Twin environment, which provides the necessary realism and complexity. The performance of the PSPO agent was systematically compared against three distinct baseline methods: a traditional Rule-Based (RB) Policy, a standard value-based RL algorithm (Deep Q-Network, DQN), and a state-of-the-art policy gradient method (Proximal Policy Optimization, PPO). The RB policy serves as a proxy for conventional, procedure-driven management, executing a static maintenance plan in the outage scenario and a fixed emergency operating procedure (EOP) in the fire scenario. Both DQN and PPO agents operate on a flattened state vector, meaning they do not explicitly model the plant's topological structure, thus providing a direct comparison to highlight the benefits of PSPO's GNN-based architecture. For each of the two experimental scenarios, the MDP was carefully defined. In the Outage Management scenario, the state vector is comprised of 52 dimensions (50 component health indices, the current outage day, and

Table 1 Digital twin fidelity validation at a key transient snapshot (T = 15 min)

Point ID	Parameter	Location	Ground truth	DT prediction	Relative error (%)
Temperature points					
T-01	T_hot_leg_1 (K)	RCS Loop 1	598.2	598.5	0.05
T-02	T_cold_leg_1 (K)	RCS Loop 1	569.5	569.2	−0.05
T-03	T_hot_leg_2 (K)	RCS Loop 2	598.3	598.1	−0.03
T-04	T_cold_leg_2 (K)	RCS Loop 2	569.6	569.6	0
T-05	T_avg (K)	Core outlet	584	583.8	−0.03
T-06	T_steam_1 (K)	SG 1 outlet	558.1	557.9	−0.04
T-07	T_steam_2 (K)	SG 2 outlet	558.2	558.3	0.02
T-08	T_containment (K)	Containment	315	315.4	0.13
Pressure points					
P-01	P_pressurizer(MPa)	Pressurizer	15.5	15.48	−0.13
P-02	P_rcs_loop1 (MPa)	RCS Loop 1	15.7	15.71	0.06
P-03	P_sg_1 (MPa)	SG 1	6.8	6.82	0.29
P-04	P_sg_2 (MPa)	SG 2	6.8	6.78	−0.29
P-05	P_containment (kPa)	Containment	101.5	101.3	−0.2
Power and level points					
N-01	Neutron flux (%)	Core	87.5	87.9	0.46
N-02	Thermal power(%)	Core	87.5	87.6	0.11
N-03	Turbine load (%)	Turbine	87.3	87.1	−0.23
L-01	Lvl_pressurizer (%)	Pressurizer	60.1	59.8	−0.5
L-02	Lvl_sg_1 (%)	SG 1	75	75.2	0.27
L-03	Lvl_sg_2 (%)	SG 2	74.9	74.8	−0.13
L-04	Elect. power (MWe)	Generator	805	802.1	−0.36

Table 2 Key hyperparameters for reinforcement learning agents

Hyperparameter	DQN	PPO	PSPO (ours)
Network architecture			
Hidden layers	3	3	3 (for actor/critic)
Neurons per layer	256	256	256
GNN Layers	N/A	N/A	2 (GraphConv)
Training parameters			
Optimizer	Adam	Adam	Adam
Learning rate (α)	1e-4	3e-4	1e-4 (actor), 3e-4 (critic)
Discount factor (γ)	0.99	0.99	0.99 (Scen. 1), 0.98 (Scen. 2)
Replay buffer size	1e6	N/A	1e6
Batch size	256	2048 (steps)	256
Algorithm-specific			
Target net update (τ)	0.005	N/A	0.005
PPO clipping (ϵ)	N/A	0.2	N/A
Risk shield penalty (λ_C)	N/A	N/A	1000
PER (α,β)	N/A	N/A	0.6, 0.4

maintenance team availability), with a discrete action space of 51 (maintain one of 50 components or wait). In the Fire Response scenario, the state vector has 35 dimensions (fire dynamics, the status of 20 safety systems, and the estimated CDF), with a discrete action space of 20 emergency actions. To ensure a fair comparison, all RL agents were built with similar network architectures and were tuned for key hyperparameters, as detailed in Table 2. Performance was assessed using a suite of metrics, including the primary Average Cumulative Reward, the Safety Risk Profile (CDF), and scenario-specific indicators such as Total Outage Duration and Time to Fire Suppression.

Table 3 Performance comparison in refueling outage scenario

Method	Avg. cumulative reward (± Std Dev)	Avg. outage duration (days)	Best case duration (days)	Avg. unaddressed critical components (H<0.4)
RB Policy	−1550.8 (± 0.0)	40.0	40.0	4.8 (± 0.0)
DQN	−1210.5 (± 85.2)	36.2	34	3.1 (± 1.2)
PPO	−1055.2 (± 68.9)	35.1	33	2.5 (± 0.9)
PSPO	−878.6 (± 45.1)	33.4	31	0.7 (± 0.4)

Table 4 Performance comparison in fire emergency scenario

Method	Avg. cumulative reward (± Std Dev)	Avg. time to suppression (min)	Avg. peak CDF	Max peak CDF (worst case)	Avg. safety systems impaired
RB Policy	−2.5e-4 (± 0.0)	28.5	8.1e-5	8.1e-5	3.0
DQN	−2.1e-4 (± 0.5e-4)	24.1	6.9e-5	9.8e-5	2.4
PPO	−1.8e-4 (± 0.3e-4)	22.7	6.2e-5	9.1e-5	2.1
PSPO	−0.9e-4 (± 0.1e-4)	17.3	3.5e-5	4.2e-5	1.1

4.3 Results and analysis

All RL agents were trained for 2,000 episodes, and their performance was averaged over 100 evaluation runs with different random seeds to ensure statistical significance.

(1) Scenario 1: Dynamic Refueling Outage Management

In this scenario, the PSPO agent demonstrated superior learning efficiency and converged to a significantly higher average cumulative reward. By processing the plant's state as a graph, the GNN-based agent could understand the intricate dependencies between maintenance tasks, which is critical for optimization. For example, PSPO consistently learned to prioritize maintenance on components with high "betweenness centrality" in the system dependency graph, as their timely repair unblocked numerous subsequent tasks.

The results in Table 3 show that PSPO outperforms all baselines. It achieves a 16.5% reduction in outage duration compared to the static RB policy and leaves 85% fewer components in a critical state of degradation. The lower standard deviation in its cumulative reward also indicates a more stable and reliable learned policy.

(2) Scenario 2: Fire Accident Emergency Response

The importance of the Risk Shield mechanism was paramount in this scenario. Standard RL agents, during exploration, occasionally attempted actions that, while locally optimal, were identified as high-risk by the Risk Shield network. For instance, PPO sometimes favored activating a ventilation system to clear smoke, an action which the Risk Shield correctly flagged as dangerous ($C_\phi > \tau_{\text{risk}}$) because it could direct the fire towards a critical cable tray. PSPO learned to avoid such "fatal" exploratory actions.

As detailed in Table 4, the policy of PSPO is demonstrated a significant improvement by more than 40%. It suppresses the fire over 39% faster than the fixed EOP of the RB policy. Critically, it keeps the average peak CDF 57% lower than the next best agent (PPO) and limits the worst-case CDF spike, demonstrating its robustness. The GNN assists by understanding the functional connectivity, enabling the agent to prioritize isolating systems that are not directly in the fire zone but are functionally linked to affected equipment.

The process of the fire accident emergency response in the DT is shown as Figure 5.

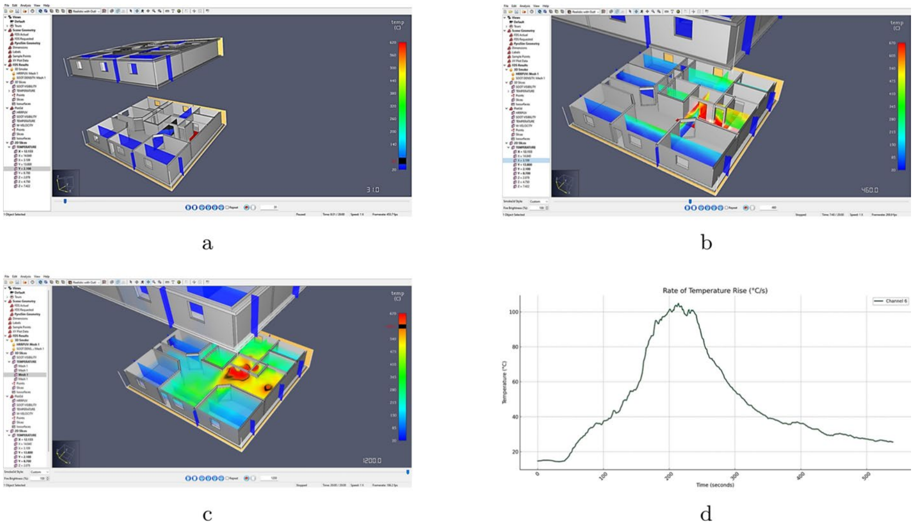

Fig. 5 Process of the fire accident emergency response in the DT. Environment in (a) 31 s (b) 460 s (c) 1200 s.(d) Rate of Temperature rise

Table 5 Ablation study results for PSPO framework

Method	Avg. cumulative reward	% Drop from full PSPO	Key observation
PSPO (Full)	−878.6	–	Optimal performance
PSPO w/o GNN	−1098.3	25.0%	Fails to identify systemic bottlenecks; sub-optimal scheduling
PSPO w/o Risk shield	−955.4	8.7%	Learns effective but occasionally unsafe policies during training
PSPO w/o PER	−980.1	11.6%	Slower convergence and less stable learning

4.4 Ablation study

The ablation study, conducted in Scenario 1, confirms the individual contributions of unique components of PSPO, as shown in Table 5.

The results clearly show that the GNN is the most critical component for achieving high performance in this complex, networked environment. Removing the GNN results in a 25% performance drop. While the Risk Shield's direct impact on the final average reward is smaller, its role is indispensable for ensuring the agent's learning process is safe, which is a non-negotiable requirement for deployment in this domain.

In summary, the experimental results, grounded in a validated high-fidelity digital twin, robustly confirm the efficacy of the PSPO framework. Its novel integration of a GNN-based state representation and a safety-constrained learning mechanism provides a significant and quantifiable advantage over both traditional management strategies and standard reinforcement learning algorithms.

5 Conclusion

We presented a novel and comprehensive framework for the full-lifecycle safety management and dynamic risk assessment of Nuclear Power Plants, grounded in the synergistic integration of Digital Twin technology and advanced Artificial Intelligence. We introduced the Nuclear Safety Twin Construction and Integration Framework (NST-CIF) as a systematic methodology for creating a high-fidelity, multi-physics, and data-driven virtual representation of an NPP. Building upon this validated DT environment, we proposed the Proactive Safety Policy Optimization (PSPO) algorithm, a sophisticated reinforcement learning approach designed specifically for safety-critical decision-making. The core innovations of PSPO—the use of a Graph Neural Network (GNN) to understand complex system interdependencies and the incorporation of a "Risk Shield" to enforce safety constraints during policy learning—were shown to be highly effective. Through extensive experiments in two challenging, representative scenarios (Dynamic Refueling Outage Management and Fire Accident Emergency Response), PSPO consistently and significantly outperformed traditional rule-based policies and standard deep reinforcement learning baselines. The results robustly demonstrated that the proposed framework can lead to more efficient operations, such as reduced outage times, and enhanced safety, evidenced by substantially lower peak risk levels during simulated accidents.

The contributions of this research mark a significant step towards transforming nuclear safety management from a conventional, reactive, and procedure-driven paradigm to a modern, proactive, and data-informed one. By providing a "what-if" sandbox for intelligent agents to learn optimal strategies, the DT-based approach can uncover complex, non-intuitive solutions that enhance both safety and efficiency, which would be impossible to discover through human analysis or real-world trial-and-error alone. Looking forward, several promising avenues for future research emerge. The NST-CIF framework can be extended to incorporate more complex physical phenomena, such as detailed neutronics-thermal-hydraulics coupling and sophisticated material aging models. The PSPO algorithm itself could be enhanced to handle multi-agent decision-making scenarios (e.g., coordinating multiple response teams) and to incorporate uncertainty quantification directly into the policy optimization process.

Despite these promising results, this study has limitations that should be acknowledged. The computational cost of training the PSPO agent is substantial, requiring significant offline simulation resources. The fidelity of the digital twin, while high, is still an approximation of the real plant and may not capture all unforeseen physical interactions. Furthermore, the scalability of the current GNN architecture to even larger and more complex plant models present a challenge for future work.

Acknowledgements
Not applicable.

Author contributions
Hongcheng Yu independently completed the experiments and writing of this paper and is responsible for all aspects of the paper.

Funding
This research did not receive any specific funding.

Data availability
Our data are presented in the manuscript. Partial support for the datasets is available from the authors.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 23 June 2025 / Accepted: 27 October 2025

Published online: 26 November 2025

References

1. Marja Y,Kim B (2023) Integrated Management of Safety and Security (IMSS) in the Nuclear Industry – Organizational Culture Perspective. Safety Science, 166, 106236.
2. Miqueles L, Ahmed I, Maio F, Zio E (2022) A Grey-Box Digital Twin-Based Approach for Risk Monitoring of Nuclear Power Plants. Book of Extended Abstracts for the 32nd European Safety and Reliability Conference.
3. Grieves M (2014) Digital Twin: Manufacturing Excellence Through Virtual Factory Replication, White Paper.
4. Kritzinger W, Karner M, Traar G, Henjes J, Sihn W (2018) Digital Twin in Manufacturing: A Categorical Literature Review and Classification, Ifac-Papersonline, 51(11), 1016–1022.
5. Wang S, Wang M, Zhao R X, Liu L, Wang Y (2025) An Interpretable Quantum Adjoint Convolutional Layer for Image Class Ification. IEEE Transactions on Cybernetics, 55(8), 3948–3959.
6. Tao F, Zhang M (2017) Digital Twin Shop-Floor: A New Shop-Floor Paradigm for Smart Manufacturing, Ieee Access, 5, 20418–20427.
7. Uhlemann T H-J, Lehmann C, Steinhilper R (2017) The Digital Twin: Realizing the Cyber-Physical Production System for Industry 4.0, Procedia CIRP, 61, 335–340.
8. Zio E, Miqueles L (2024) Digital Twins in Safety Analysis, Risk Assessment and Emergency Management. Reliability Engineering & System Sa Fety, 246, 110040.
9. Zio, E (2018) The Future of Risk Assessment, Reliability Engineering & System Safety, 177, 1–13.
10. Zheng X, Chen X, Gao Z, Jin Q, Wei H (2024) Application Exploration of Digital Twin Technology in Second-Loop System of Nuclear Power Plants. 2024 8th International Symposium on Computer Science and Intelligent Control (ISCSIC), 330–334.
11. Siu N (1994) Risk Assessment for Dynamic Systems: An Overview, Reliability Engineering & System Safety, 43(2), 157–173.
12. Aldemir T (2013) A Survey of Dynamic Methodologies for Probabilistic Safety Assessment of Nuclear Power Plants, Annals of Nuclear Energy, 52, 113–124.
13. Ayo-Imoru, R., Ali, A., & Bokoro, P. (2021). An Enhanced Fault Diagnosis in Nuclear Power Plants for a Digital Twin Framework. 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET), 1–6.
14. Bevilacqua M, Bottani E, Ciarapica F, Costantino F, Di Donato L, Ferraro, A., Mazzuto G, Monteriù A, Nardini G, Ortenzi M, Paroncini M, Pirozzi M, Prist M, Quatrini E, Tronci M, Vignali G. (2020). Digital Twin Reference Model Development to Prevent Operators' Risk in Process Plants. Sustainability, 12(3), 1088.
15. Kochunas B, Huan X (2021) Digital Twin Concepts with Uncertainty for Nuclear Power Applications. Energies, 14(14), 4235.
16. Song H, Song M, Liu X (2022) Online Autonomous Calibration of Digital Twins Using Machine Learning with Application to Nuclear Power Plants. Applied Energy, 325, 119995.
17. Ren Z, Wan J, Deng P (2022) Machine-Learning-Driven Digital Twin for Lifecycle Management of Complex Equipment. IEEE Transactions on Emerging Topics in Computing, 10(1), 9–22.
18. Hu M, Zhang X, Peng C, Zhang Y, Jun Y (2024) Current status of digital twin architecture and application in nuclear energy field. Annals of Nuclear Energy, 199, 110491.
19. Abo-Khalil A (2023) Digital Twin Real-Time Hybrid Simulation Platform for Power System Stability. Case Studies in Thermal Engineering, 49, 103237.
20. Zhang S, Lu R, Zhou H, Link S, Yang Y, Li Z, Gong S (2020) Surface Acoustic Wave Devices Using Lithium Niobate on Silicon Carbide. IEEE Transactions on Microwave Theory and Techniques, 68(9), 3653–3666.
21. Wang J, Huang Y, Li J, Zhai W, Ouyang S, Gao H, Liu, Y, Wang G (2023) Research on Coal Mine Safety Management Based on Digital Twin. Heliyon, 9(2), E13608.
22. Sutton R S, Barto A G (2018) Reinforcement Learning: An Introduction. MIT Press.
23. Mnih V, Kavukcuoglu K, Silver D, Rusu, A A, Veness J, Bellemare M G, Hassabis, D. (2015). Human-Level Control Through Deep Reinforcement Learning. Nature, 518(7540), 529-533.
24. Zhao R X, Shi J, Li X (2024) Qksan: A Quantum Kernel Self-Attention Network. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 10184–10195.
25. Zhang Z, Liu J, Zeng W, Huang Q, Liu X (2025) Digital Twin Technology Architecture and Application for Nuclear Reactor Intelligent Operation and Maintenance. IEEE Access, 13, 91494–91504.
26. Li Z, Wang H, Peng M, Xu R, Yu Y, Zhou G (2022) Digital Twin Based Operation Support System of Nuclear Power Plant. 2022 IEEE 2nd International Conference on Digital Twins and Parallel Intelligence (DTPI), 1–6.
27. Lillicrap, T. P. et al. (2015). "Continuous control with deep reinforcement learning," arXiv preprint arXiv:1509.02971.
28. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017) Proximal Policy Optimization Algorithms," arXiv preprint arXiv:1707.06347.
29. Zheng X, Tamaki H, Sugiyama T, Maruyama Y (2022) Dynamic Probabilistic Risk Assessment of Nuclear Power Plants Using Multi-fidelity Simulations. Reliability Engineering & System Safety, 223, 108503.
30. Bao H, Zhang H, Shorthill T, Chen E, Lawrence S (2023) Quantitative Evaluation of Common Cause Failures in High Safety-Significant Safety-Related Digital Instrumentation and Control Systems in Nuclear Power Plants. Reliability Engineering & System Safety, 230, 108973.

Publisher's Note