energies

MDPI

Article

Integrating an Ensemble Reward System into an Off-Policy Reinforcement Learning Algorithm for the Economic Dispatch of Small Modular Reactor-Based Energy Systems

Athanasios Ioannis Arvanitidis * and Miltiadis Alamaniotis

Department of Electrical and Computer Engineering, University of Texas at San Antonio, San Antonio, TX 78249, USA; miltos.alamaniotis@utsa.edu

* Correspondence: athanasios.arvanitidis@my.utsa.edu

Abstract: Nuclear Integrated Energy Systems (NIES) have emerged as a comprehensive solution for navigating the changing energy landscape. They combine nuclear power plants with renewable energy sources, storage systems, and smart grid technologies to optimize energy production, distribution, and consumption across sectors, improving efficiency, reliability, and sustainability while addressing challenges associated with variability. The integration of Small Modular Reactors (SMRs) in NIES offers significant benefits over traditional nuclear facilities, although transferring involves overcoming legal and operational barriers, particularly in economic dispatch. This study proposes a novel off-policy Reinforcement Learning (RL) approach with an ensemble reward system to optimize economic dispatch for nuclear-powered generation companies equipped with an SMR, demonstrating superior accuracy and efficiency when compared to conventional methods and emphasizing RL's potential to improve NIES profitability and sustainability. Finally, the research attempts to demonstrate the viability of implementing the proposed integrated RL approach in spot energy markets to maximize profits for nuclear-driven generation companies, establishing NIES' profitability over competitors that rely on fossil fuel-based generation units to meet baseload requirements.

Keywords: nuclear integrated energy systems; small modular reactors; economic dispatch; reinforcement learning; off-policy algorithms; reward engineering; energy spot markets

Citation: Arvanitidis, A.I.; Alamaniotis, M. Integrating an Ensemble Reward System into an Off-Policy Reinforcement Learning Algorithm for the Economic Dispatch of Small Modular Reactor-Based Energy Systems. *Energies* **2024**, *17*, 2056. https://doi.org/10.3390/en17092056

Academic Editors: Grzegorz Dudek and Marcin Blachnik

Received: 22 March 2024
Revised: 10 April 2024
Accepted: 22 April 2024
Published: 26 April 2024

1. Introduction

Presently, the energy distribution sector is undergoing rapid changes marked by a growing presence of distributed renewable energy sources such as wind and solar power. While these sources contribute clean electricity to the grid, they also introduce variability, prompting the need for additional complementary energy resources. To address this evolving energy landscape, Nuclear Integrated Energy Systems (NIES) employ a holistic approach, integrating nuclear power plants with renewable energy sources, advanced energy storage systems, smart grid technologies, and other complementary components [1]. These systems aim to enhance overall energy efficiency, reliability, and sustainability by optimizing energy production, distribution, and consumption across various sectors [2]. They facilitate the seamless coordination and utilization of various energy resources, thereby maximizing the benefits of each while minimizing potential drawbacks [3]. Additionally, NIES are designed to adapt to changing energy demands and environmental considerations, ensuring a resilient and responsive energy infrastructure for the future.

The research interest has shifted towards NIES, given the significant and diverse expected advantages they offer [4]. In addition to offering the grid a reliable, carbon-free, and dispatchable source of electricity, they also contribute synchronous electromechanical inertia that improves the grid's stability and reliability. Moreover, by reducing the carbon footprint of the industrial sector and leveling energy costs, NIES contribute to environmental sustainability and economic stability. Additionally, NIES offer the opportunity

for long-term energy planning and investment, providing a stable and predictable energy supply that can support sustainable economic growth and development. However, it is crucial to recognize that as renewable energy systems gain global prominence, there may be unintended consequences for traditional baseload generators, highlighting the importance of carefully managing the transition to more integrated energy systems [5].

Transitioning from conventional integrated energy systems to those powered by nuclear energy necessitates a shift not only from typical nuclear reactors like Light Water Reactors (LWRs), but also toward newer configurations such as Small Modular Reactors (SMRs). The integration of SMRs into NIES offers clear benefits compared to traditional nuclear power plants [6]. To begin with, SMRs present lower capital costs and construction risks due to their smaller size, factory production, and standardized components, making them more attractive to investors. Their greater deployment flexibility, which includes the ability to install them inland, allows for faster construction and lower cooling capacity requirements, which may lessen the burden on transmission and distribution networks [7]. Moreover, their simplified safety features and easier decommissioning process enhance their overall operational efficiency and safety. Furthermore, SMRs' reduced refueling needs and potential for extended operation without uranium replenishment make them well-suited for integration into NIES, where reliability and sustainability are paramount [8]. Additionally, their smaller size opens up opportunities for diverse applications beyond electricity generation, such as water desalination [9], further enhancing their appeal within integrated energy systems [10]. Therefore, the unique attributes of SMRs make them indispensable components of NIES, offering a more agile, cost-effective, and versatile approach to nuclear energy generation and beyond. Figure 1 depicts the morphology, interconnection, and diverse range of energy products facilitated by NIES, further highlighting their significance in modern energy landscapes.

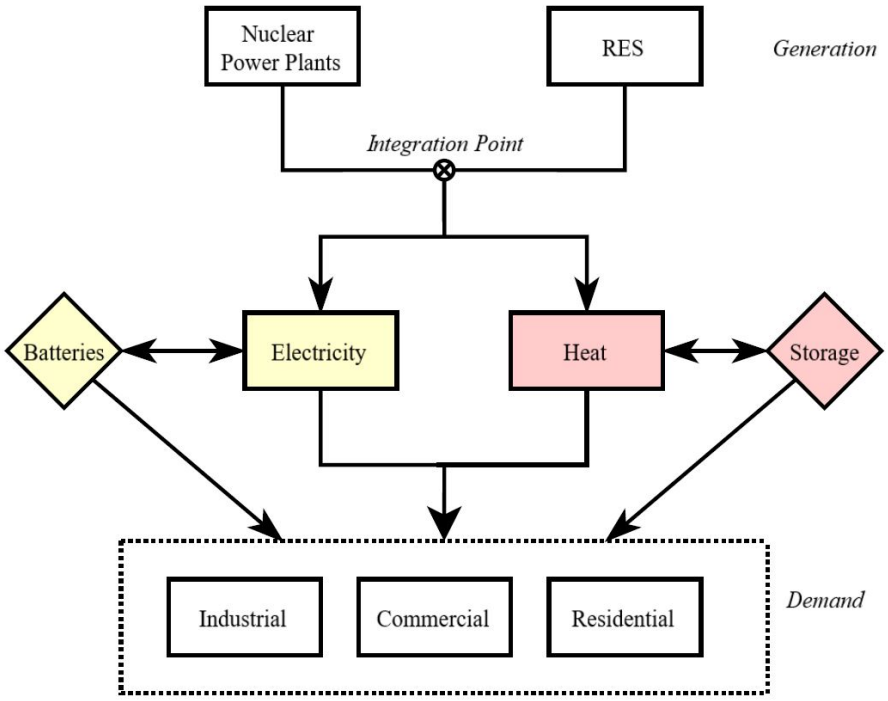

Figure 1. Integration of nuclear power with renewable energy sources for sustainability and versatility.

As SMRs are introduced into NIES, there are notable regulatory implications regarding their economic operation. In the short term, certain operational nuclear plants in the United States are adapting to evolving grid and market dynamics by adjusting their power output [11]. This strategy, known as advanced economic dispatch, involves reducing output during periods of high renewable energy generation or low net demand to avoid

selling electricity at a loss, which is a common issue for traditional baseload plants. This shift reflects the changing landscape of energy markets, which is driven by technological advancements and the increasing integration of nuclear power and renewable energy sources. This integration not only enhances the resilience and flexibility of energy systems but also facilitates the optimization of energy production and consumption in response to fluctuating demand and supply patterns.

As nuclear plants adapt to these evolving dynamics, in recent years, we have witnessed a transition from conventional power systems to smart grids, accompanied by rapid advancements in computer science. This transition has prompted the integration of sophisticated Artificial Intelligence (AI) algorithms, such as Reinforcement Learning (RL), to address research challenges within the economic dispatch issue in power systems. As RL research evolves, the focus has shifted towards addressing reward design complexities, which are crucial for developing robust algorithms capable of navigating complex environments and optimizing decision-making processes effectively [12]. The bibliography encompasses several RL techniques tailored for the economic dispatch of conventional power systems. Among these techniques, the most commonly utilized ones are referred to as Q-learning-based algorithms [13] and Deep Q-Networks (DQNs) [14]. Q-learning algorithms, a fundamental RL approach, optimize decision-making processes by iteratively learning the optimal action-value function. On the other hand, DQNs employ deep neural networks to approximate the action-value function, allowing for more complex and efficient learning in large-scale power systems.

In terms of applying Q-learning based methodologies, Ref. [15] presents a novel Q-learning-based Swarm Optimization (QSO) algorithm. This algorithm approaches optimization problems as tasks in reinforcement learning, with the objective of achieving optimal solutions through the maximization of expected rewards. Integrating Q-learning and particle swarm optimization, the QSO algorithm imitates the behavior of the global best individual in the swarm and selects individuals based on accumulated performance. Similarly, Ref. [16] tackles the necessity of adjusting economic dispatch and Unit Commitment (UC) problems to accommodate the transition to smart grids. The authors reframe these challenges into a unified framework capable of managing infinite horizon UC complexities. A centralized Q-learning-based optimization algorithm is proposed, operating online without requiring prior knowledge of cost function mathematical formulations. Additionally, a distributed version of the algorithm is developed to balance exploration and exploitation cooperatively, demonstrating effectiveness through theoretical analysis and case studies.

With regard to the more intricate approach of DQNs in the context of economic dispatch, the authors of [17] present a Deep Reinforcement Learning (DRL) approach for Combined Heat and Power (CHP) system economic dispatch, offering reduced complexity and improved adaptability across various operating scenarios. Simulating CHP economic dispatch issues as Markov Decision Processes (MDP) enables the method to bypass extensive linearization efforts while preserving device characteristics. Similarly, Ref. [18] proposes a DRL algorithm for online economic dispatch in Virtual Power Plants (VPPs), addressing challenges posed by the uncertainty of distributed renewable energy generation. Utilizng DRL, the algorithm decreases computational complexity while managing the stochastic aspects of power generation. Implemented within an edge computing framework, the real-time economic dispatch strategy successfully minimizes VPP costs, demonstrating superior performance compared to deterministic policy gradient algorithms. Also, the authors of Ref. [19] propose a scenario-based robust economic dispatch approach for VPPs to address challenges posed by the complex structure and uncertain characteristics of distributed energy. By incorporating scenario-based data augmentation and deep reinforcement learning, the strategy reduces conservatism and directly solves nonlinear and non-convex problems, ultimately minimizing operational costs.

This paper aims to revolutionize economic dispatch in NIES by employing an innovative RL approach, leveraging reward engineering to significantly differentiate our work

from existing literature. In this study, a novel off-policy RL method is introduced, which integrates the core principles of the traditional Q-learning algorithm with an ensemble reward system. This hybrid approach aims to facilitate and accelerate the learning process of an agent tasked with managing the economic dispatch of a nuclear-driven generation company, equipped with a commercial SMR. Comparison with conventional economic dispatch approaches serves as a benchmark for evaluating the proposed method, showcasing its superior accuracy as the agent adeptly learns the economic dispatch portfolio. Our study enhances the area of energy systems optimization by illustrating the importance of sophisticated reinforcement learning approaches in tackling the intricate and ever-changing problems encountered by NIES. Ultimately, this paper seeks to demonstrate the viability of employing the fused RL approach within a spot energy market to maximize profits for nuclear-driven generation companies, thereby establishing the profitability of NIES over competitors relying on fossil fuel-based generation units to fulfill baseload requirements.

The remainder of the paper is organized as follows: Section 2 provides a holistic overview of the energy market environment under study, together with the underlying assumptions of this research and mathematical formulations of economic dispatch optimization challenges. Also, in Section 2, the RL environment for economic dispatch is developed, emphasizing the inclusion of the suggested ensemble reward system and the mathematical formulation of the agent's optimal policy. Section 3 presents the numerical results derived from the study, demonstrating the efficacy and accuracy of the proposed algorithm. Subsequently, Section 4 provides a discussion of the results and offers final remarks, while Section 5 concludes the paper.

2. Materials and Methods

This section presents a detailed description of the energy market environment under study, its assumptions, and mathematical formulations for economic dispatch optimization, along with the development of the RL environment, highlighting the inclusion of the proposed ensemble reward system and the formulation of the agent's optimal policy.

2.1. Structure of the Assessed Energy Market Landscape

This subsection outlines the framework of the energy spot market utilized to analyze the profitability exhibited by the examined GenCos. It is explicitly stated that the primary objective of this research is the implementation of a novel RL algorithm, enabling the developed agent to learn the economic dispatch schedule of a GenCo that utilizes a SMR, rather than the modeling or full simulation of an energy spot market. Additionally, the assumptions investigated are provided along with the mathematical aspects of the economic dispatch.

2.1.1. Energy Spot Markets

Energy spot markets are used to buy and sell physical quantities of power in a short-term timeframe ahead of delivery [20]. Essentially, in electricity markets, it is necessary to constantly balance the supply and demand. In a spot market, the goods are promptly delivered by the seller, and payment is made by the buyer "on the spot", without any conditions attached to the delivery. This operational framework ensures that neither party possesses the option to retract from the agreement once it has been initiated, fostering a swift and decisive transaction process [21]. Overall, energy spot markets play a vital role in the efficient functioning of energy systems, facilitating the reliable and cost-effective supply of electricity to consumers while providing opportunities for market participants to manage risk and optimize their operations.

In this type of market, electrical energy is produced and sold by Generation Companies (GenCos). They may also sell services such as regulation, voltage control, and reserve, which are necessary for maintaining the quality and security of the electricity supply. A single plant or a portfolio of plants of different technologies may be owned by a generating company. It is evident that every GenCo endeavoring to submit a bid for electricity

production aims to maximize profitability by efficiently meeting load demand. To achieve this objective, it is imperative for each GenCo to construct a robust energy portfolio based on its available fleet of Generation Units (GenUnits), as each unit generates a specific power output at a specific cost [22]. This cost serves as a determinant of the unit's production expense. Consequently, each GenCo should strictly follow a pre-defined generation policy, which dictates that the GenUnit with the lowest generation cost should be prioritized and dispatched first to meet subsequent load demands. This policy, known as economic dispatch, holds paramount importance for the economic viability of GenCos, as it ensures the maximization of profits [23].

The configuration of our energy market environment is shaped by the constraint that energy transactions solely take place through a spot market overseen by a central pool entity, disregarding the possibility of bilateral agreements between the demand and generation sectors. Initially, the Market Operator (MO) assumes the responsibility of coordinating bids and offers submitted by buyers and sellers of electrical energy. Upon receiving a signal from the demand side, comprising both load demand and price parameters, the MO transmits this information to the supply side, which encompasses a set of N GenCos. These GenCos are tasked with efficiently generating the required power to fulfill the load demand, with the premier objective of profit maximization. Each GenCo, in turn, submits to the market pool details regarding the quantity of energy it intends to produce and the corresponding price.

2.1.2. Mathematical Formulation of Economic Dispatch Models

Economic dispatch denotes the systematic operation of generation facilities with the objective of ensuring the reliable provision of energy at minimal expense, while acknowledging the operational constraints in generation and transmission infrastructures [24]. Typically, economic dispatch entails the utilization of the generating unit with the lowest variable cost, or in the context of a competitive energy market, the unit offering the lowest price, to increase output in response to rising loads. Conversely, it involves the reduction of output from the least cost-effective unit as demand decreases [25].

The economic dispatch problem is fundamentally contingent upon the cost function of the GenUnit [26]. Consequently, understanding the correlation between the cost and output power of a generating unit is crucial. The cost of generating a specific quantity of energy, usually measured in MWh, can exhibit significant variation depending on the technological specifications of the unit. In the short term, certain factors of production remain fixed, and the expenses associated with these factors are independent of the output quantity, constituting fixed costs. The expenses that remain constant in constructing a power plant are commonly denoted as capital costs. These encompass expenditures associated with the initial construction, installation, and development of the power plant infrastructure, including equipment and land acquisition. Such expenses are typically accumulated from the start and tend to remain consistent regardless of the amount of electricity generated by the plant. On the contrary, the volume of fuel utilized by the plant, along with operational and maintenance expenses, including the necessary manpower for its functioning, are intricately linked to the energy output. These expenditures represent the variable costs associated with the GenUnits.

In this work, the computation of the cost required to produce a specific quantity of power is conducted solely by accounting for the variable costs, encompassing fuel expenditures and Operational and Maintenance (O&M) costs associated with the generator. Additionally, the exclusion of capital costs regarding all GenUnits is presumed, and hence omitted from consideration. Consequently, the aggregated cost attributed to each GenUnit is derived by aggregating the fuel costs and the O&M costs. Therefore, the total generation cost of each GenCo is ascertained by summing the power output from all dispatched generators, with the output of each generator multiplied by its corresponding total cost, given by (1) as follows:

$$C^{\text{Total}} = \sum_{i=1}^{M} P_i^G \cdot (C_i^F + C_i^{\text{O\&M}}) \tag{1}$$

where M denotes the number of available GenUnits, with P_i^G denoting the power output generated by the ith unit. Moreover, C_i^F and $C_i^{\text{O\&M}}$ are indicative of the corresponding fuel and O&M expenses attributed to each individual GenUnit, respectively.

The economic dispatch challenge constitutes a multi-dimensional optimization task primarily aimed at minimizing the collective expenses associated with active power generation, as specified in (1). Essential to the completion of this task is the active power output of each GenUnit, denoted as P_i^G. Consequently, the economic dispatch quandary presents itself as an optimization task, established in (2), while (3) delineates the constraints inherent to this scenario, as follows:

$$\min \quad \{\sum_{i=1}^{M} P_i^G \cdot (C_i^F + C_i^{\text{O\&M}})\} \tag{2}$$

$$\text{s.t.:} \quad \sum_{i=1}^{M} P_i^G = P^L + \sum_{i=1}^{K} P_i^D \tag{3}$$

$$P_i^{G_{\min}} \leq P_i^G \leq P_i^{G_{\max}} \quad \text{for } i = 1, 2, \ldots, M$$

where M represents the number of GenUnits, reflecting the capacity of the power-generation infrastructure, while K denotes the number of demand nodes. The term P^L express the losses incurred during the transmission of electricity, highlighting the dissipation of energy. Additionally, $P_i^{G_{min}}$ and $P_i^{G_{max}}$ characterize the minimum and maximum generation capabilities, respectively, of each individual GenUnit, thus delineating the operational range within which these units can operate effectively.

The income (I_t) of a GenCo stems from the power it generates, multiplied by the prevailing electricity price in \$/MWh at each given time t, and is expressed in (4), as follows:

$$I_t = P_t^D \cdot \pi_t \tag{4}$$

where P_t^D is the power demanded and π_t is the price at a specific time t. This income is crucial for evaluating the financial performance and viability of the GenCo, reflecting its ability to leverage market conditions and efficiently allocate resources to meet demand while maximizing revenue. By effectively managing the generation of electricity in alignment with market dynamics, GenCos can optimize their income streams and ensure sustainable operations in the energy sector.

Profit represents the financial advantage gained when the revenue generated from a business activity surpasses the incurred expenses, costs, and taxes essential for sustaining said activity. Companies that demonstrate profitability often attract investors due to their promising returns. Profit, a fundamental metric, is computed by deducting total expenses from total revenue, acting as an indicator of a company's financial health and potential for growth. The applicability of this fundamental concept extends beyond traditional business domains to encompass sectors like the energy spot market. In this context, Equation (5) defines profit as follows:

$$G_t = I_t - C_t^{\text{Total}} = P_t^D \cdot \pi_t - \sum_{i=1}^{M} P_i^G \cdot (C_i^F + C_i^{\text{O\&M}}) \tag{5}$$

where G_t encapsulates the profit earned by a GenCo subsequent to dispatching the requisite GenUnits for the economic dispatch issue at time t. This term reflects the financial gain realized by the GenCo as a result of its operational decisions in response to economic dispatch requirements.

2.1.3. Critical Assumptions Shaping Problem Formulation

This study introduces a novel greedy off-policy algorithm fused with an ensemble reward system as an RL approach for addressing the economic dispatch challenge encountered by a nuclear-driven GenCo. However, it is imperative to explicate the fundamental assumptions underpinning our investigation. These assumptions serve as foundational principles guiding the agent's learning process to ascertain the optimal economic dispatch schedule and ensure the profitability of the GenCo under consideration. The portrait of these assumptions not only shapes the formulation of the problem but also serves to unfold the underlying framework of the proposed methodology. Specifically, the following assumptions are posited:

- The structure of our energy market environment is defined by the constraint dictating that energy transactions occur exclusively through a spot energy market regulated by a central entity, thereby excluding the option for bilateral agreements between the demand side and GenCos. Consequently, the electricity price π_t remains consistent at time t, indicating that prices for the corresponding load demand remain uninfluenced by negotiations. This results in (4) remaining constant at a specific time t.
- The load demand values consistently stay below the maximum capacity of each GenCo, thereby negating the necessity for the GenCo to exhaust its entire capacity without the requirement of determining an optimal economic dispatch schedule. In the event that the load demand value exceeds the maximum capacity of the GenCos, they would be obligated to dispatch all their GenUnits without any consideration for optimization.
- GenCos function autonomously, with limited awareness of each other's portfolios, thus ensuring fairness. While GenCos may possess information regarding overall market conditions, such as wholesale electricity prices, they typically lack direct insight into the dispatch schedules of their rivals. In general, they operate independently and might lack direct access to the economic dispatch schedules of other GenCos, particularly if they are viewed as competitors in the electricity market. The economic dispatch process is often regarded as confidential to prevent manipulation or the unfair exploitation of advantages in the electricity market.
- GenCos ensure that all their GenUnits are constantly available, thereby eliminating any uncertainty regarding the generation capabilities of renewable energy resources, which are heavily reliant on weather conditions. Additionally, the possibility that certain GenUnits may be unavailable due to maintenance reasons is not taken into account.

Therefore, within the framework of the energy spot market, the adoption of RL techniques for approaching the economic dispatch method is advocated. It is emphasized that the aim of this paper is not to propose an energy spot market framework, but rather to develop an off-policy learning approach embedded with a greedy reward method. This algorithm is designed to enhance and expedite the learning capabilities of an agent simulating a nuclear-driven GenCo. Therefore, the proposed RL approach should be established, along with all the technical details of the proposed algorithm, before the case study and numerical results of our work are presented.

2.2. Reinforcement Learning Algorithm for Economic Dispatch

Reinforcement learning, a training technique within machine learning, operates on the principle of rewarding desired behaviors and penalizing undesired ones. Within the domain of RL, knowledge acquisition occurs through the interaction of agents with their environment, wherein actions are taken and subsequent rewards or penalties are received based on those actions [27]. These rewards or penalties serve as feedback mechanisms, enabling agents to adjust their action strategies, which are referred to as policies [28]. The significance of this approach lies in its capacity to enable agents to autonomously learn and adapt to the intricacies of their designated environments [29].

In the realm of RL, the fundamental concept revolves around encapsulating the essential elements of the genuine challenge encountered by a learning agent as it interacts with its environment to attain a goal. Evidently, such an agent must possess the capability to

perceive the state of the environment to a certain degree and execute actions that influence this state. Additionally, the agent must be equipped with one or more goals pertaining to the state of the environment. The formulation aims to encompass these three fundamental aspects—sensation, action, and goal—in their most basic forms.

Therefore, to comprehensively outline the proposed RL approach aimed at addressing the economic dispatch schedule, certain fundamental components for the RL environment must be established. These fundamental components consist of defining the state space, which encompasses the possible configurations or conditions of the system; specifying the action space, which outlines the available choices or decisions that the agent can make within the environment; and designing a robust reward system, which serves as the mechanism for reinforcing desired behaviors and guiding the learning process. These fundamental elements provide the groundwork for developing a precise and efficient reinforcement learning system that is adapted to handle the intricacies of the economic dispatch issue.

2.2.1. Establishing the RL Environment for Economic Dispatch Modeling

In this subsection, the RL environment utilized in our study will be elaborated upon. Within our research framework, a nuclear-driven GenCo is cast as an autonomous agent, tasked with refining its generation optimization strategy through the acquisition of expertise. This is noticeably different from antagonist GenCos, which lack RL capabilities. This research focuses on carefully determining the most effective power output for the GenCo's available GenUnits. Consequently, the agent is structured as a GenCo with a predetermined set of M GenUnits, each serving as a vital component in the overarching energy-generation process.

Secondly, within our framework, it is absolutely essential to meticulously define the representation of potential actions included in an action set $\mathbb{A}$ which are accessible to our agent in numerical terms. This entails establishing a clear link between the actions undertaken by the proposed agent and the generation process of each available GenUnit. Consequently, the agent must possess the capacity to adjust the generation levels of each of the M GenUnits. This necessitates that an equal number of potential actions be meticulously considered for each individual GenUnit, ensuring comprehensive coverage and equitable treatment across the entire system.

Finally, the definition of the state space $\mathbb{S}$ representation must be addressed. It is imperative that the agent be granted access to all pertinent data crucial for the economic dispatch procedure and for effectively minimizing the total generation cost of the nuclear-driven GenCo. Thus, within each state, the agent should be provided with information regarding both the current generation levels and the generation capacity of each of the M GenUnits. Furthermore, it is essential that the agent obtains comprehensive information concerning the cost implications of increasing the generation by 1 MW for each GenUnit. This entails understanding the costs associated with fuel and O&M expenses, which are integral components that must be factored into (2) for the purpose of minimizing the objective function. Furthermore, it is essential for the agent to have knowledge of both the power demand and the electricity price at a specific time t. This information enables the agent to accurately estimate the total income outlined in (4). Understanding the value of load demand helps the agent anticipate the amount of electricity needed by consumers, while knowledge of the electricity price allows the agent to assess the potential revenue generated from supplying electricity to the grid. Incorporating these factors into the optimization process is crucial for maximizing the overall income and effectively managing the economic dispatch procedure.

2.2.2. Attaining the Optimal Policy in Agent-Driven Economic Dispatch

In RL, a policy is defined as a strategy employed by an agent in the pursuit of goals. The actions taken by the agent are dictated by the policy, which operates as a function of both the agent's state and the environment. In the realm of RL policy training, two primary

approaches stand out: on-policy learning and off-policy learning [30]. On-policy methods involve the iterative refinement of a single policy, which in turn generates control actions within the environment, known as the behavior policy. Essentially, the behavior policy delineates the policy that is adhered to by the agent when selecting actions within the environment at each time step [31]. Additionally, there are off-policy methods in which data from the behavior policy are utilized to train a separate target policy for an optimization objective. This discrepancy in which the policy is updated with the training data exerts a profound impact on the learning behavior of the various RL algorithms [32].

In on-policy RL, actions are determined based on observed environmental states according to a specific RL policy. The results of these actions are gathered and employed to gradually enhance the parameters of the identical policy. Consequently, on-policy RL employs a shared behavior and target policy. This policy is tasked with exploring both the state and action spaces while refining the learning objective using the accumulated data. Numerous algorithms within the on-policy category incorporate a certain level of variation in actions. This variation is introduced deliberately to maintain a delicate equilibrium between the exploration of new possibilities and the exploitation of known information for optimal decision-making [33].

Conversely, off-policy RL is characterized by the utilization of two distinct strategies consisting of a behavior policy and a target policy. The behavior policy determines the actions taken in response to observed environmental states, while the target policy is continuously refined based on the outcomes of these actions. This approach allows off-policy RL to separate the process of collecting data from the training of policies. An advantageous aspect of off-policy methods is their capability to learn an optimal target policy, which may prioritize maximizing rewards, regardless of the exploratory nature of the behavior policy. It is common practice in off-policy learning to periodically update the behavior policy with the latest insights from the target policy to enhance the overall learning process [34].

In our paper, the economic dispatch procedure is intended to be addressed with an off-policy RL algorithm. On-policy algorithms involve enhancing the current behavior policy utilized for decision-making, thus acquiring knowledge of the policy's value executed by the agent. Conversely, off-policy algorithms are designed to ascertain the value of the optimal policy and possess the capability to refine policies distinct from the behavior policy. Therefore, the choice of employing an off-policy RL algorithm in our research is informed by its potential suitability, particularly in simulation scenarios. This choice is driven by its inherent tendency to more effectively learn the optimal policy, providing greater adaptability and robustness in navigating complex decision spaces, while avoiding being trapped in local minima.

Building upon the preceding discussion, understanding and defining this policy is fundamental to the success of any RL approach, ensuring effective decision-making in complex environments. Therefore, the identification of optimal actions hinges on establishing a precise mapping from each state to a probability distribution encompassing the available actions within that state. This ideal mapping is a fundamental aspect of any RL methodology, crucial for effective decision-making. In the realm of sequential decision-making, problems are commonly formulated as MDPs. The solution to such problems involves crafting an optimal mapping for each state, denoted as $s \in \mathbb{S}$, which delineates a probability distribution over the action set $\mathbb{A}$ available in that state. The mathematical representation of a policy ϕ in MDPs is articulated by (6) as follows:

$$\phi : \mathbb{S} \times \mathbb{A} \rightarrow [a_1, a_2, \ldots, a_M] \tag{6}$$

where ϕ represents the policy being defined, $\mathbb{S}$ and $\mathbb{A}$ denote the sets of all possible states and actions in the environment, respectively, and $[a_1, a_2, \ldots, a_M]$ represents the set of possible actions that the agent can choose from in response to a given state.

The value function $V_\phi(s)$ characterizes the anticipated return R when commencing from the initial state s_0 and then adhering to the policy $V_\phi(s)$. This function provides an assessment of the attractiveness of being situated in a particular state. As agents move

through different states and implement decisions guided by their policies, it becomes crucial to assess the effectiveness of these policies. A fundamental aspect of evaluating policies involves comparing the performance of different action policies. This comparison process leads to the establishment of dominance relationships among the agent's policies. A relationship of dominance can be established among an agent's action policies. Specifically, policy ϕ' is deemed dominant over policy ϕ if there is no distribution of rewards that would lead policy ϕ to yield higher expected rewards than policy ϕ'. In instances where policy ϕ' holds dominance over policy ϕ, the agent consistently prioritizes ϕ' over ϕ. At the core of the concept of optimality lies the notion that a policy ϕ' possessing the highest value is considered optimal. This condition of optimality is represented by (7) as follows:

$$V^*_{\phi''}(s) = \max V_{\phi''}(s) = \max_{\phi''} E[R|s = s_0, \phi'] \tag{7}$$

This paper introduces a groundbreaking off-policy approach, which integrates the backbone of the q-learning algorithm alongside an ensemble learning accelerator. Notably, the innovation lies in the incorporation of a sophisticated reward system aimed at enhancing the learning process of the agent tasked with determining the optimal dispatch policy for the nuclear-based GenCo. The proposed reward system operates on the principle of incentivizing the agent to make decisions that align with the desired objectives. These extra rewards play a pivotal role in guiding the agent towards the most advantageous course of action, as it can learn the optimal policy and facilitate the utilization of appropriate GenUnits to efficiently address the economic dispatch issue.

2.3. Ensemble Reward Mechanisms for Enhanced and Accelerated Learning

In practical applications of RL, engineers often encounter constraints that limit their ability to directly manipulate environmental conditions. As a result, the efficacy of conventional engineering interventions is limited, necessitating further methods to effectively foster learning processes. In this context, reward-shaping techniques emerge as pivotal mechanisms for modulating the learning trajectory [35]. Reward shaping involves the strategic introduction of supplementary reward signals at intermediate stages of an RL task. These additional rewards serve as guidance mechanisms, steering the learning agent towards desired states or behaviors [36]. Moreover, through strategic design and implementation of intermediate rewards, engineers can accelerate the learning algorithm's progression towards optimal policies. This acceleration is particularly crucial in domains such as power systems, where rapid acquisition of proficient behaviors is imperative for ensuring efficient, stable, and economic operation [37].

In our setup, agents are required to follow a prescribed set of actions to understand the optimal economic dispatch schedule. The primary aim of the agent is to effectively fulfill the load demand while simultaneously developing and refining its decision-making approach. As a result, the agent should be rewarded for each potential action, indicating its effectiveness. During the decision-making process, the agent aims to select actions that maximize its returns while simultaneously boosting the production of the most cost-effective GenUnit. This entails ensuring that the selected actions align with the optimal balance between profitability and efficiency. Additionally, it is crucial for the agent to operate within the generation limits of each GenUnit, thereby preventing any instances of overgeneration. Once a GenUnit reaches its maximum capacity, it becomes necessary for the agent to prioritize dispatching the subsequent most cost-effective GenUnit. This proactive approach ensures a continuous and efficient generation capability within the autonomous system. To foster such behavior in the agent, it is crucial to establish a system of suitable rewards and/or penalties. These mechanisms serve as guiding cues, prompting the agent to consistently make decisions that are in line with the core objectives of maximizing returns and complying with generation limits [38].

Consequently, the agent should exhibit a proactive approach in its actions, aiming to satisfy both these objectives effectively. This proactive stance ensures that the agent's deci-

sions are guided by the aim of maximizing returns while also adhering to generation limits. Due to this rationale, an ensemble greedy reward scheme has been devised in our paper to ensure that the optimal strategy is acquired by the agent more rapidly, without becoming trapped in local minima, and to ensure that the GenUnits are dispatched not only at the desired generation level but also in the appropriate sequence.

The agent is awarded additional reward points for adhering to the learned policy [39]. Upon completion of the final action, the environment state is terminated, initiating a thorough examination to ascertain if the maximum capacity of the least expensive GenUnit has been reached. Following this, the embedded accelerator mechanism is promptly activated. If the least expensive GenUnit has been dispatched at the desired generation level, the agent receives a combined reward, prompting the dispatch of the next available cost-effective GenUnit. This strategy showcases its notable effectiveness, especially in the early phases of the RL algorithm when the agent's actions lack clear direction, highlighting the efficacy of the proposed approach in managing intricate environments with prolonged convergence times, thereby accelerating the acquisition of the desired behavioral policy by the agent. A comprehensive depiction of the proposed reward system is provided in Figure 2.

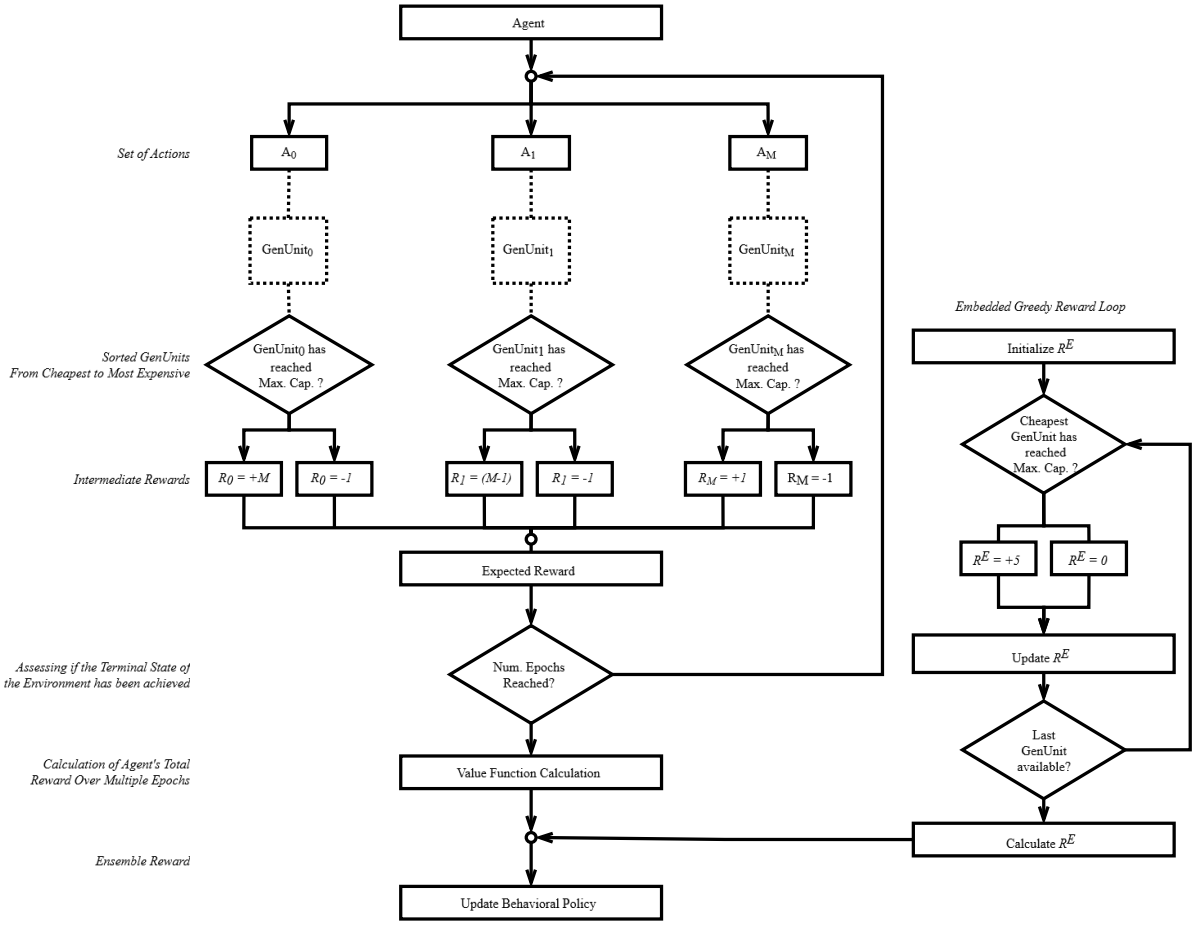

Figure 2. Reward shaping strategies and accelerated policy learning for accurate economic dispatch using an embedded greedy reward system.

The suggested off-policy technique relies on the foundation of the Q-learning process. Consequently, in order to define the optimum economic dispatch policy for the designed agent, the value function $V_\phi(s)$ of (7) should be incorporated into the Q-learning procedure with an additional reward term, with the Q-learning update rule being modified to include

this extra reward component. This update rule is given by (8), with the Q-learning update rule being modified to include this extra reward component, as follows:

$$\max_{Q}\left(Q(s,a)+\alpha\left(E[r|s=s_0,\phi]+R^{\mathrm{E}}+\gamma\max_{a'}Q(s',a')-Q(s,a)\right)\right)$$
$$\text{s.t. } \phi^*(s)=\arg\max_{a}Q^*(s,a) \tag{8}$$

where $Q(s,a)$ is the Q-value for state–action pair (s,a), α is the learning rate, r is the immediate reward, γ is the discount factor, $\max_{a'}Q(s',a')$ is the maximum Q-value for the next state s' and $E[r|s=s_0,\phi]+R^{\mathrm{E}}$ represents the total reward obtained by following policy ϕ from the initial state s_0, which is the combination of the expected return under policy ϕ and the ensemble reward value R^{E}. This modification ensures that the Q-value update incorporates both the immediate reward r and the additional reward component R^{E}, providing a more comprehensive measure of the value of taking a specific action in a state under optimal policy ϕ^*.

3. Numerical Results of Case Study

In this section, the case study under examination will be detailed. Firstly, the framework of the studied energy market environment and the participating GenCos will be described. Following this, the RL environment, in which the agent responsible for the economic dispatch of the nuclear-driven Genco operates, will be outlined, with its action space and state space defined. Finally, the efficiency of the proposed ensemble reward system fused with an off-learning policy will be numerically demonstrated, emphasizing the greater profitability exhibited by nuclear-driven GenCos compared to those relying on conventional GenUnits to cover their base load.

In the envisioned energy market setting, three GenCos are engaged in competition, striving to enhance their profitability while endeavoring to fulfill the entirety of demand values communicated by the MO. Table 1 accumulates comprehensive data concerning the capacity of GenUnits within each GenCo under evaluation. Electricity demand experiences fluctuations throughout the day due to various factors such as time of day, weather conditions, and industrial activities. Certain types of GenUnits, such as natural gas turbines or hydroelectric plants, possess greater flexibility and can swiftly adjust their output to align with changes in demand. Conversely, others, like nuclear plants or coal/diesel generators, offer a consistent baseload supply. The utilization of a diverse array of energy sources enables GenCos to effectively manage these demand fluctuations by employing various types of GenUnits to meet load requirements, thereby optimizing reliability, flexibility, cost-efficiency, environmental sustainability, and grid stability. Each type of GenUnit provides particular benefits and characteristics that contribute to the establishment of a balanced and resilient energy system. Thus, each GenCo under study is mandated to possess at least one GenUnit tasked with addressing the base load, alongside additional GenUnits designated for handling intermediate loads, and renewable GenUnits designated for managing peak loads.

Although there are minor differences in the O&M costs and capacities of RES, and due to the fact that each GenCo operates with its distinct combination of energy sources, it is not necessary for them to standardize their GenUnits for comparison. Initially, during the dispatch of RES, generation costs across companies generally exhibit minimal variation. However, as the demand necessitates the dispatch of baseload generation units, disparities in profitability become evident, leading to increased generation costs and subsequent profit differentials. Thus, each GenCo's unique energy portfolio necessitates flexibility in unit selection. Furthermore, the capacity of these baseload generation units does not significantly affect the economic dispatch process, since they typically operate below their full capacity. Therefore, variations in the capacities of baseload generation units among the GenCos under study are acceptable and do not hinder the comparability of results.

Table 1. Generation capacity overview of GenCos under study.

Power Plant	GenCoA	GenCoB	GenCoC
Natural Gas	–	–	1045
Coal	–	1995	–
Nuclear (SMR)	300	–	–
Solar	–	174	174
Hydropower	338	338	–
Wind	468	–	468
Biomass	–	–	105

The market conditions being analyzed are evaluated by examining hourly load and marginal price data specific to a designated region within the PJM power system for the year 2022. It is imperative for the agent overseeing the economic dispatch of GenCoA to be proficient in managing this data, as they constitute an integral component of the envisioned environment. This entails the ability to interpret and utilize the hourly load and marginal price data effectively to inform decision-making processes regarding the economic dispatch of generation resources.

In order to address the optimization problem described by (2), it is necessary for the agent to have access to data regarding fuel costs and O&M costs. These costs have been sourced from the annual reports of the U.S. Energy Information Administration (EIA) and have been converted to units of USD/MWh. The combination of all these data sets defines the majority of the RL environment in this study. Figure 3 provides a graphical representation of monthly fuel costs for the year 2022 in USD/MWh for each GenUnit under study. Enriched uranium serves as the conventional fuel choice for the SMR under study, while alternative materials, such as thorium, may be utilized. It is evident that coal exhibits the highest fuel cost, while uranium indicates the lowest cost per MWh. It is notable that the RES under examination demonstrate zero fuel costs. This emphasizes the importance of considering various energy sources and their associated costs when optimizing the system.

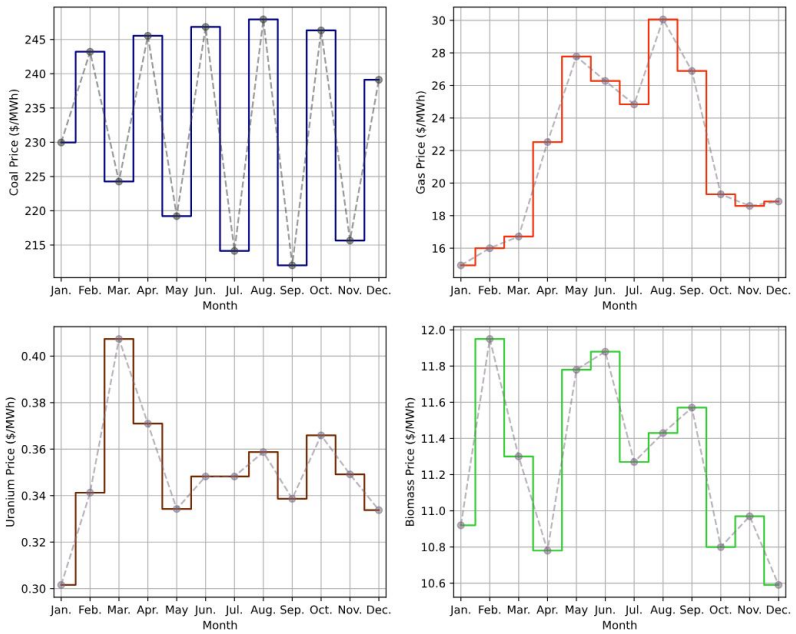

Figure 3. Fuel expenses per month (colorful lines) for the available GenUnits during periods of high fluctuations (grey dotted lines).

Furthermore, Figure 4 provides a visual representation of the O&M costs associated with each GenUnit under examination. Despite the relatively low fuel cost of uranium, it is noteworthy that GenCoA's SMR displays the highest O&M expenses. Despite their smaller

size, SMRs may still incur significant expenses due to the complexity of their designs, stringent regulatory requirements, and the need to deploy them in remote or off-grid locations. Furthermore, unlike larger reactors that have longer refueling intervals, SMRs typically require more frequent refueling due to their smaller size and power output. These refueling procedures involve shutdowns, labor, and material costs, all of which contribute to the higher O&M costs. Conversely, GenUnits reliant on fossil fuels, such as gas and coal, demonstrate comparatively lower maintenance costs. This discrepancy underscores the diverse cost structures inherent in different types of generation technologies and emphasizes the importance of considering both fuel and O&M expenses when evaluating the economic viability of energy-generation options. The agent should possess the capability to adapt to fluctuations in these data sets, as they directly impact the optimization of economic dispatch within the power system.

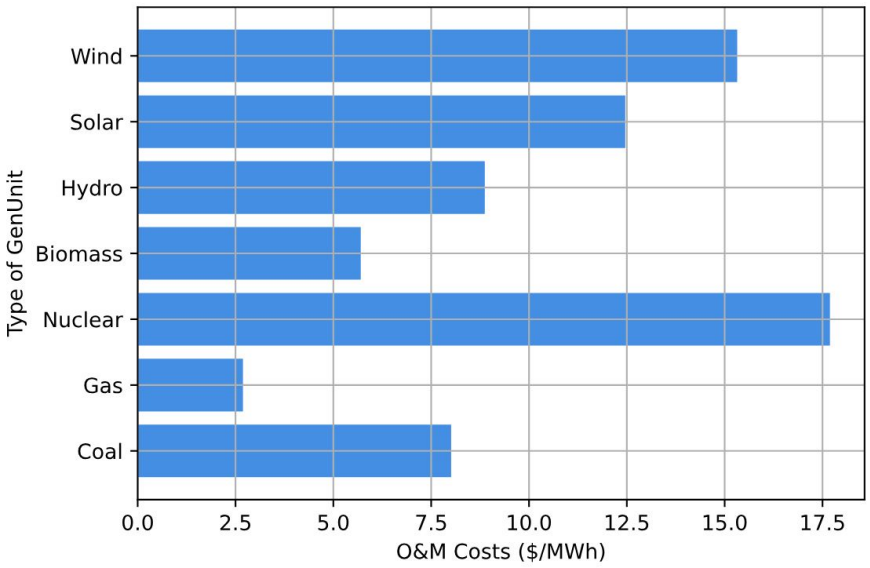

Figure 4. O&M expenses for the GenUnits under study.

Upon the establishment of the RL environment, it becomes imperative to conduct a thorough examination of the effectiveness and precision of the proposed off-policy learning approach, which encompasses the ensemble reward system. This comprehensive evaluation seeks to provide clarity regarding the overall efficacy of the developed agent's behavioural policy. Initially, serving as the benchmark for this evaluation is a traditional manual economic dispatch approach, where supply-side entities systematically organize their available generators based on ascending order of operating costs, ranging from the least to the most expensive. By contrasting the outcomes of this traditional method with those achieved by our developed agent employing the proposed approach, valuable insights into its effectiveness can be obtained. Figure 5 compiles and presents the economic dispatch results of the three GenCos for varying load values observed over a 24 h period. It is apparent from the data that GenCos adeptly manage to meet the load demand in the most economical manner adapted to their specific operational needs and constraints.

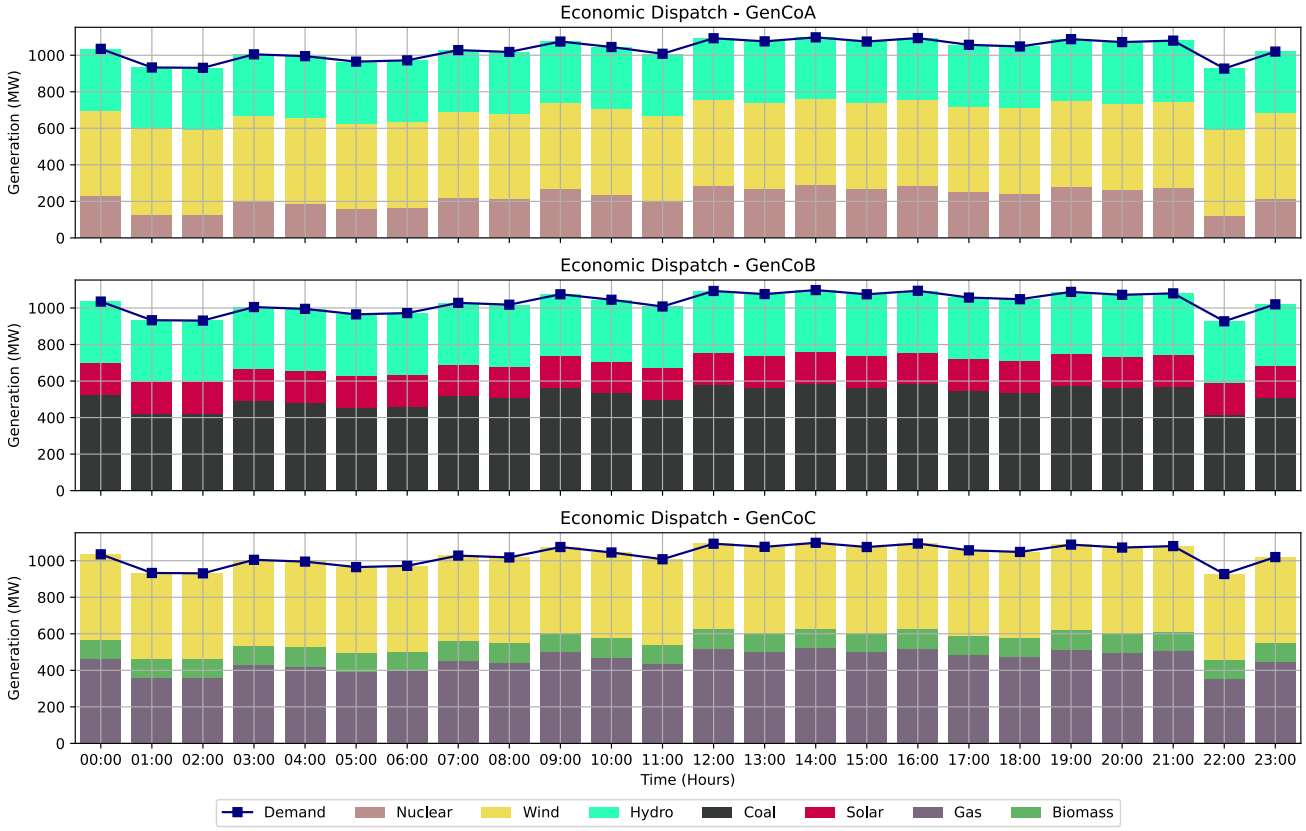

Figure 5. Economic dispatch performance of three GenCos across varying load demand values over a 24-h period.

Moreover, it is essential to conduct a comparative analysis between the outcomes generated by the traditional manual economic dispatch method and those yielded by the developed agent utilizing our proposed approach. Given the potential system fluctuations primarily attributed to RES and the imperative to validate the algorithm's effectiveness within the designated time frame, the proposed framework operates on the foundational principle of equipping the agent with an optimal strategy for efficiently allocating its generating capacity. This capability empowers the agent to adeptly address anticipated future load demands unaffected by dependence on specific time resolutions. Consequently, a single training session is necessary for the agent to acquire the capability to adhere to an economically optimized dispatch schedule aligned with its maximum generation capacity. Figure 6 provides a graphical representation facilitating the comparison of the economic dispatch results obtained from the two methodologies across varying load values received by GenCoA over a 24 h period. This comparative assessment allows for a comprehensive examination of the efficacy and performance of the developed agent in optimizing economic dispatch decisions.

Figure 6 vividly illustrates the remarkable accuracy exhibited by the proposed off-learning RL approach fused with the ensemble reward system, in addressing the economic dispatch challenges faced by the nuclear-driven GenCo. This high level of accuracy stems from both the innovative reward system proposed and the strategic approach of training the agent with GenCoA's capacity as a reference point. By adopting this training tactic, the agent becomes capable at replicating the optimal economic dispatch schedule of GenCoA without necessitating training for every individual demand signal it encounters. Consequently, this approach not only enhances accuracy but also significantly reduces the execution complexity associated with the proposed method. By minimizing the need for exhaustive training on varying demand scenarios, the agent can efficiently adapt to dynamic operating conditions, thereby improving its overall performance and applicability

in real-world settings. Furthermore, the comparative assessment of the proposed RL approach against the conventional manual method, which serves as the benchmark, reveals notable enhancements in both accuracy and efficiency. This enhancement is visible when examining the accuracy metrics presented in Table 2, showcasing the superior performance of the RL approach. The Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2 Score) are all crucial indicators for evaluating the performance of the proposed RL model.

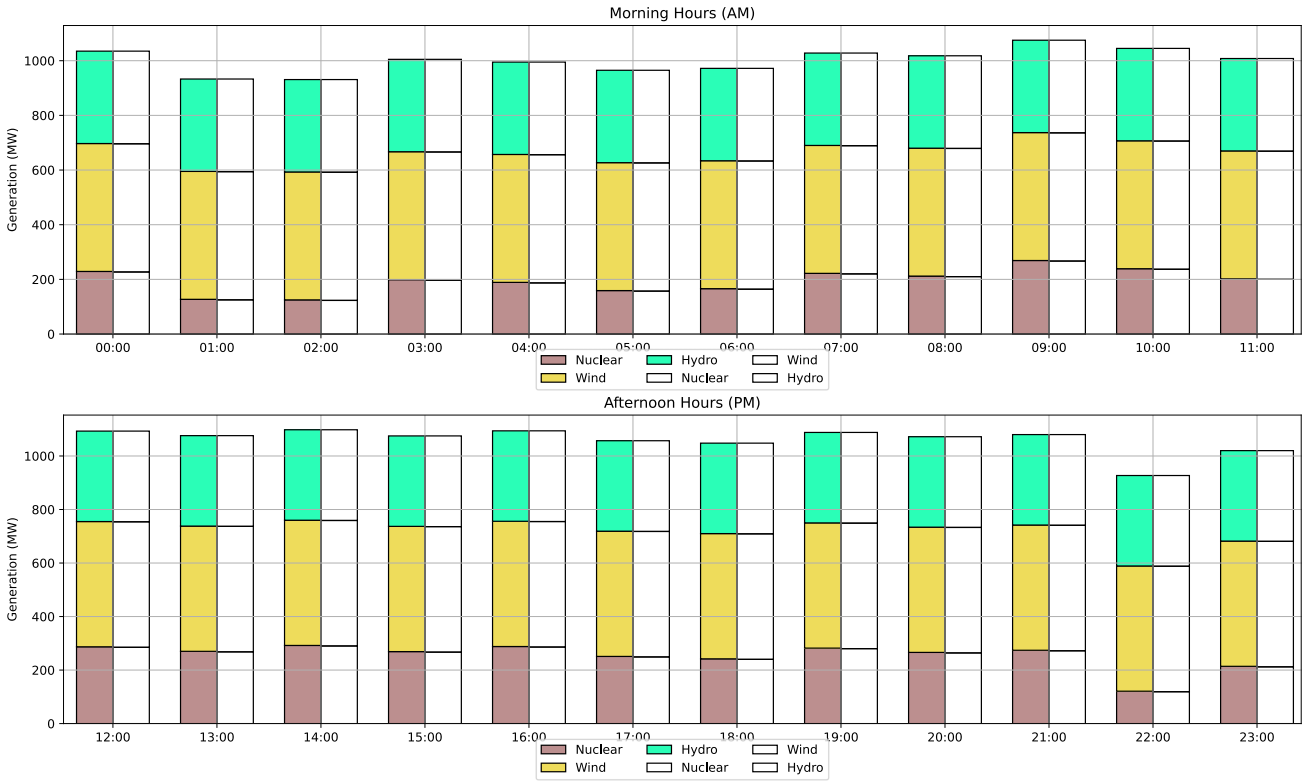

Figure 6. Comparative economic dispatch results of both approaches for GenCoA across varying load demand values over a 24 h period.

Table 2. Accuracy metrics for the proposed RL approach.

MAE (MW)	MSE (MW^2)	MAPE (%)	R^2 Score
1.333	2.00	0.52	0.333

Having demonstrated the effectiveness of the implemented agent in addressing economic dispatch challenges, it becomes essential to demonstrate the increased profitability of the nuclear-driven generation company relative to the other entities. This effort attempts to validate the second focal point of the paper concerning the profitability of GenCo utilizing a SMR to fulfill the base load requirement. Thus, it is imperative to establish that NIES not only contribute to reducing carbon dioxide emissions but also yield greater profitability. To achieve this, the total electricity production costs for each GenCo, as defined by (2), across various load values within a day need to be taken into account. It is evident that the GenCo with the lowest production costs stands to achieve the highest profitability, as supported by (5) and the underlying assumptions. Figure 7 presents a depiction of the total production costs, measured in dollars, for each GenCo undergoing examination. These costs are determined based on the economic dispatch portfolios utilized by each entity to fulfill the load demand over a 24 h period.

As depicted in Figure 7, GenCoB emerges with the highest production costs among the trio of GenCos under examination. This elevated cost is primarily attributed to the significantly higher expenses associated with coal compared to other energy sources within its energy mix. The prevailing composition of energy sources results in a notably augmented energy generation cost for GenCoB, rendering it distinctly unprofitable. Furthermore, it is important to highlight the consistent lower electricity generation costs of GenCoA compared to GenCoC throughout the observed period. This sustained pattern underscores the superior cost efficiency of GenCoA's operations relative to GenCoC. Given these noticeable disparities in production costs, our analytical focus naturally gravitates towards examining the profitability dynamics of GenCoA and GenCoC. Conversely, the notably higher costs incurred by GenCoB render it less relevant to our profitability analysis.

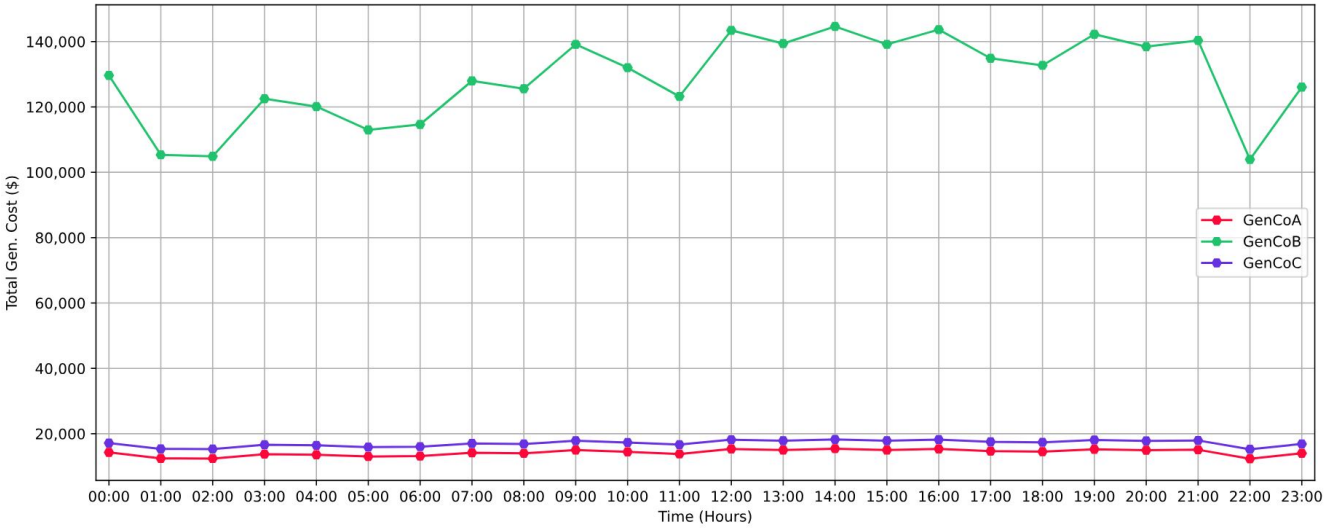

Figure 7. Total production costs of the analyzed GenCos for meeting the 24 h load demand.

Figure 8 provides a comprehensive depiction of the profits, measured in dollars, attained by the two most lucrative GenCos within the analyzed cohort. Notably, GenCoA, which utilizes an SMR, emerges as the obvious profitability leader when compared to GenCoC, which relies on fossil fuels to cover its energy needs. This disparity in profitability necessitates an investigation of the underlying causes influencing their economic performance. Traditionally, one may predict that the significant O&M expenses associated with SMRs will serve as a limiting factor for the profitability of nuclear-driven GenCo over its fossil fuel-dependent equivalent, GenCoC. However, this research shows that GenCoA's competitive advantage is mostly due to the favorable cost dynamics concerning uranium. Furthermore, despite possessing a more diverse energy mix and the flexibility to deploy a greater variety of generators, GenCoC confronts a significant rise in the cost of electricity generation, resulting in lower profits. The fluctuations in GenCoC's generating costs are mostly due to the extremely variable nature of natural gas prices throughout 2022. The high cost of gas and higher O&M expenditures for GenCoC's biomass generator and RES contribute to reduced profitability. These combined parameters highlight the delicate relationship between energy source selection, operating costs, and overall profitability in the electricity-producing environment.

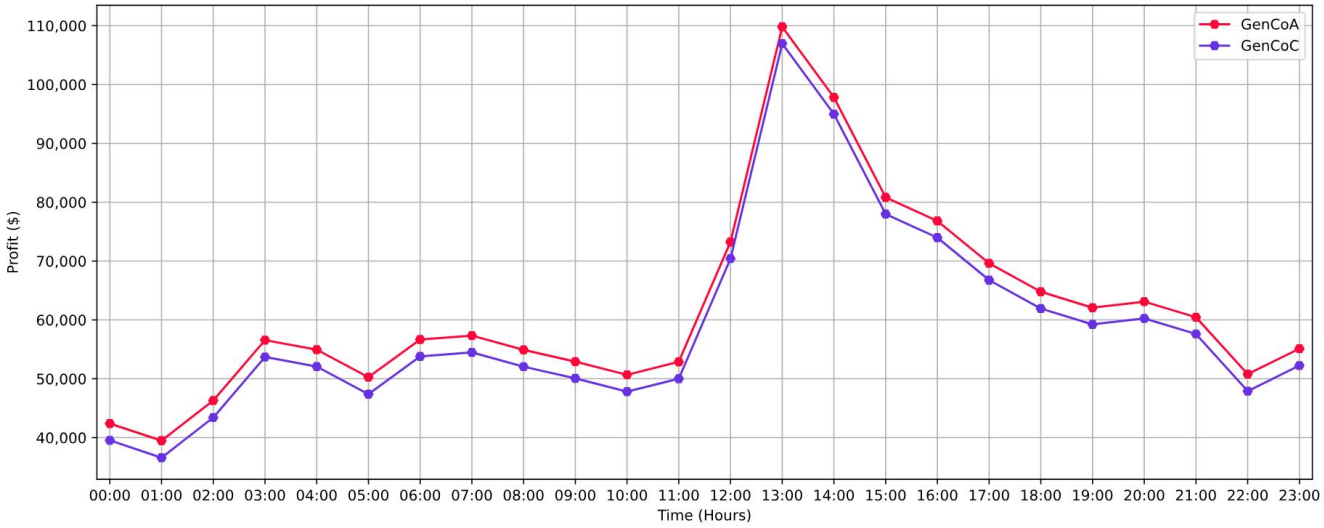

Figure 8. Profit comparison of top-performing GenCos: SMR vs. fossil fuel-based operations.

4. Discussion

This research presents a novel off-learning RL approach combined with an ensemble reward system, offering a promising solution to the economic dispatch challenges encountered by a nuclear-driven GenCo. This study underscores the critical interplay between energy-source selection, operational costs, and profitability within the electricity production sector, providing valuable insights for decision-makers aiming to optimize economic dispatch strategies and enhance the overall power-generation performance. The proposed approach showcases remarkable accuracy, which is attributed to both the innovative reward system and the strategic training tactic that leverages a reference point of GenCoA's capacity. By minimizing the need for exhaustive training on varying demand scenarios, the proposed method not only enhances accuracy but also reduces execution complexity, thereby improving overall performance and applicability in real-world settings.

Comparative assessments against conventional manual methods reveal significant enhancements in accuracy and efficiency. The analysis of production costs among GenCos highlights distinct disparities, with GenCoB exhibiting notably higher costs, primarily due to its reliance on coal. Conversely, GenCoA consistently demonstrates superior cost efficiency compared to GenCoC, which is attributable to its energy mix and operational strategies. In terms of profitability, GenCoA emerges as the leader, leveraging its nuclear-based energy generation with favorable cost dynamics concerning uranium. Despite possessing a more diverse energy mix, GenCoC faces challenges, particularly due to fluctuations in natural gas prices and higher operating expenses for renewable energy sources.

5. Conclusions

In conclusion, this study introduces a novel off-policy learning RL approach coupled with an ensemble reward system to tackle economic dispatch challenges within a nuclear-driven GenCo. Our research sheds light on the intricate interplay among energy source selection, operational costs, and profitability in the electricity production sector, providing valuable insights for dispatch strategy optimization. Comparative assessments against conventional methods underscore significant improvements in accuracy and efficiency. Analysis of production costs reveals notable disparities among GenCos, with a NIES emerging as a leader in cost efficiency owing to its SMR-based energy generation.

In future research, the uncertainties associated with various RES will be thoroughly considered to enhance the real-world applicability of the proposed approach. Incorporating these uncertainties will enable a more comprehensive understanding of the challenges faced by GenCos in managing their energy resources efficiently. This approach endeavors to establish a more generalized energy spot market. Consequently, further refinement of

the proposed RL algorithm is essential, as is facilitating its adoption by further developed agents. Consequently, this will lead to the establishment of a robust and efficient multi-agent system. Additionally, a broader spectrum of economic dispatch strategies will be developed, with a primary focus on minimizing the generation costs of nuclear-driven GenCos while simultaneously addressing objectives related to CO_2 emissions reduction. This expanded scope will facilitate the exploration of more sophisticated optimization techniques and trade-offs within the energy generation sector. Moreover, there will be an exploration into formulating a hierarchical multi-agent system to control and monitor the energy-management system of nuclear-driven GenCos. This system will offer a robust framework for integrating advanced control mechanisms into the proposed environment, enabling more sophisticated decision-making processes and enhancing overall operational efficiency. By integrating these advancements, future research endeavors aim to provide comprehensive solutions for addressing the evolving challenges in economic dispatch and energy management within the NIES.

Author Contributions: Conceptualization, A.I.A.; methodology, A.I.A.; software, A.I.A.; validation, A.I.A. and M.A.; formal analysis, A.I.A.; investigation, A.I.A.; resources, A.I.A. and M.A.; data curation, A.I.A. and M.A.; writing—original draft preparation, A.I.A.; writing—review and editing, A.I.A. and M.A.; visualization, A.I.A.; supervision, M.A.; project administration, M.A.; funding acquisition, M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Department of Energy under grant number DE-NE0009278.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data regarding power generation for various generation units and the load demand values utilized in this research are openly accessible through Data Miner, PJM's enhanced data management tool, at https://dataminer2.pjm.com/list (accessed on 9 November 2023). Also, the data pertaining to operational and maintenance costs, as well as fuel costs, utilized in this study are openly accessible through the EIA's annual reports, available at https://www.eia.gov/ (accessed on 11 November 2023). This data set was utilized as input for the development and evaluation of the reinforcement learning approach proposed in this article.

Acknowledgments: This material is based upon work supported by the Department of Energy under Award Number(s) DE-NE0009278. This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CHP	Combined Heat and Power
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
EIA	Energy Information Administration
GenCo	Generation Company
GenUnit	Generation Unit
LWR	Light Water Reactor

MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MDP	Markov Decision Processes
MO	Market Operator
MSE	Mean Squared Error
NIES	Nuclear Integrated Energy Systems
O&M	Operational and Maintenance
QSO	Q-learning-based Swarm Optimization
RL	Reinforcement Learning
SMR	Small Modular Reactor
UC	Unit Commitment
VPP	Virtual Power Plant

References

1. Bragg-Sitton, S.M.; Boardman, R.; Rabiti, C.; O'Brien, J. Reimagining future energy systems: Overview of the US program to maximize energy utilization via integrated nuclear-renewable energy systems. *Int. J. Energy Res.* **2020**, *44*, 8156–8169. [CrossRef]
2. Arvanitidis, A.I.; Agarwal, V.; Alamaniotis, M. Nuclear-Driven Integrated Energy Systems: A State-of-the-Art Review. *Energies* **2023**, *16*, 4293. [CrossRef]
3. Arent, D.J.; Bragg-Sitton, S.M.; Miller, D.C.; Tarka, T.J.; Engel-Cox, J.A.; Boardman, R.D.; Balash, P.C.; Ruth, M.F.; Cox, J.; Garfield, D.J. Multi-input, multi-output hybrid energy systems. *Joule* **2021**, *5*, 47–58. [CrossRef]
4. Frick, K.; Wendt, D.; Talbot, P.; Rabiti, C.; Boardman, R. Technoeconomic assessment of hydrogen cogeneration via high temperature steam electrolysis with a light-water reactor. *Appl. Energy* **2022**, *306*, 118044. [CrossRef]
5. Ruth, M.F.; Zinaman, O.R.; Antkowiak, M.; Boardman, R.D.; Cherry, R.S.; Bazilian, M.D. Nuclear-renewable hybrid energy systems: Opportunities, interconnections, and needs. *Energy Convers. Manag.* **2014**, *78*, 684–694. [CrossRef]
6. Rowinski, M.K.; White, T.J.; Zhao, J. Small and Medium sized Reactors (SMR): A review of technology. *Renew. Sustain. Energy Rev.* **2015**, *44*, 643–656. [CrossRef]
7. Lloyd, C.A.; Roulstone, T.; Lyons, R.E. Transport, constructability, and economic advantages of SMR modularization. *Prog. Nucl. Energy* **2021**, *134*, 103672. [CrossRef]
8. Tian, L.; Zhu, C.; Deng, T. Day-ahead scheduling of SMR integrated energy system considering heat-electric-cold demand coupling response characteristics. *Energy Rep.* **2022**, *8*, 13302–13319. [CrossRef]
9. Hills, S.; Dana, S.; Wang, H. Dynamic modeling and simulation of nuclear hybrid energy systems using freeze desalination and reverse osmosis for clean water production. *Energy Convers. Manag.* **2021**, *247*, 114724. [CrossRef]
10. Poudel, B.; Gokaraju, R. Small modular reactor (SMR) based hybrid energy system for electricity & district heating. *IEEE Trans. Energy Convers.* **2021**, *36*, 2794–2802.
11. Epiney, A.; Rabiti, C.; Talbot, P.; Alfonsi, A. Economic analysis of a nuclear hybrid energy system in a stochastic environment including wind turbines in an electricity grid. *Appl. Energy* **2020**, *260*, 114227. [CrossRef]
12. Kaelbling, L.P.; Littman, M.L.; Moore, A.W. Reinforcement learning: A survey. *J. Artif. Intell. Res.* **1996**, *4*, 237–285. [CrossRef]
13. Watkins, C.J.; Dayan, P. Q-learning. *Mach. Learn.* **1992**, *8*, 279–292. [CrossRef]
14. Arulkumaran, K.; Deisenroth, M.P.; Brundage, M.; Bharath, A.A. Deep reinforcement learning: A brief survey. *IEEE Signal Process. Mag.* **2017**, *34*, 26–38. [CrossRef]
15. Hsieh, Y.Z.; Su, M.C. A Q-learning-based swarm optimization algorithm for economic dispatch problem. *Neural Comput. Appl.* **2016**, *27*, 2333–2350. [CrossRef]
16. Li, F.; Qin, J.; Zheng, W.X. Distributed Q-Learning-Based Online Optimization Algorithm for Unit Commitment and Dispatch in Smart Grid. *IEEE Trans. Cybern.* **2019**, *50*, 4146–4156. [CrossRef] [PubMed]
17. Zhou, S.; Hu, Z.; Gu, W.; Jiang, M.; Chen, M.; Hong, Q.; Booth, C. Combined heat and power system intelligent economic dispatch: A deep reinforcement learning approach. *Int. J. Electr. Power Energy Syst.* **2020**, *120*, 106016. [CrossRef]
18. Lin, L.; Guan, X.; Peng, Y.; Wang, N.; Maharjan, S.; Ohtsuki, T. Deep reinforcement learning for economic dispatch of virtual power plant in internet of energy. *IEEE Internet Things J.* **2020**, *7*, 6288–6301. [CrossRef]
19. Fang, D.; Guan, X.; Hu, B.; Peng, Y.; Chen, M.; Hwang, K. Deep reinforcement learning for scenario-based robust economic dispatch strategy in internet of energy. *IEEE Internet Things J.* **2020**, *8*, 9654–9663. [CrossRef]
20. Schweppe, F.C.; Caramanis, M.C.; Tabors, R.D.; Bohn, R.E. *Spot Pricing of Electricity*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
21. Wang, X.; Peng, P.; Chen, N. Review and reflection on new energy participating in electricity spot market mechanism. In Proceedings of the 2021 IEEE Sustainable Power and Energy Conference (iSPEC), Nanjing, China, 23–25 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 757–762.
22. Li, G.; Shi, J.; Qu, X. Modeling methods for GenCo bidding strategy optimization in the liberalized electricity spot market—A state-of-the-art review. *Energy* **2011**, *36*, 4686–4700. [CrossRef]
23. Wen, G.; Yu, X.; Liu, Z. Recent progress on the study of distributed economic dispatch in smart grid: An overview. *Front. Inf. Technol. Electron. Eng.* **2021**, *22*, 25–39. [CrossRef]

24. Kunya, A.B.; Abubakar, A.S.; Yusuf, S.S. Review of economic dispatch in multi-area power system: State-of-the-art and future prospective. *Electr. Power Syst. Res.* **2023**, *217*, 109089. [CrossRef]
25. Marzbani, F.; Abdelfatah, A. Economic Dispatch Optimization Strategies and Problem Formulation: A Comprehensive Review. *Energies* **2024**, *17*, 550. [CrossRef]
26. Xu, Y.; Song, Y.; Deng, Y.; Liu, Z.; Guo, X.; Zhao, D. Low-carbon economic dispatch of integrated energy system considering the uncertainty of energy efficiency. *Energy Rep.* **2023**, *9*, 1003–1010. [CrossRef]
27. Matsuo, Y.; LeCun, Y.; Sahani, M.; Precup, D.; Silver, D.; Sugiyama, M.; Uchibe, E.; Morimoto, J. Deep learning, reinforcement learning, and world models. *Neural Netw.* **2022**, *152*, 267–275. [CrossRef]
28. Bennett, D.; Niv, Y.; Langdon, A.J. Value-free reinforcement learning: Policy optimization as a minimal model of operant behavior. *Curr. Opin. Behav. Sci.* **2021**, *41*, 114–121. [CrossRef]
29. Gu, S.; Yang, L.; Du, Y.; Chen, G.; Walter, F.; Wang, J.; Yang, Y.; Knoll, A. A review of safe reinforcement learning: Methods, theory and applications. *arXiv* **2022**, arXiv:2205.10330.
30. Hausknecht, M.; Stone, P.; Mc, O.P. On-policy vs. off-policy updates for deep reinforcement learning. In *Deep Reinforcement Learning: Frontiers and Challenges, Proceedings of the IJCAI 2016 Workshop, New York, NY, USA, 9–11 July 2016*; AAAI Press: New York, NY, USA, 2016.
31. Singh, S.; Jaakkola, T.; Littman, M.L.; Szepesvári, C. Convergence results for single-step on-policy reinforcement-learning algorithms. *Mach. Learn.* **2000**, *38*, 287–308. [CrossRef]
32. Munos, R.; Stepleton, T.; Harutyunyan, A.; Bellemare, M. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems, Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16; Barcelona, Spain, 5–10 December 2016*; Curran Associates Inc.: Red Hook, NY, USA, 2016; pp. 1054–1062.
33. Andrychowicz, M.; Raichuk, A.; Stańczyk, P.; Orsini, M.; Girgin, S.; Marinier, R.; Hussenot, L.; Geist, M.; Pietquin, O.; Michalski, M.; et al. What matters in on-policy reinforcement learning? a large-scale empirical study. *arXiv* **2020**, arXiv:2006.05990.
34. Thomas, P.; Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; PMLR: London, UK, 2016; pp. 2139–2148.
35. Dewey, D. Reinforcement Learning and the Reward Engineering Principle. 2014 AAAI Spring Symposium Series. 2014. Available online: https://www.danieldewey.net/reward-engineering-principle.pdf (accessed on 23 September 2023).
36. Gupta, A.; Pacchiano, A.; Zhai, Y.; Kakade, S.; Levine, S. Unpacking reward shaping: Understanding the benefits of reward engineering on sample complexity. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 15281–15295.
37. Icarte, R.T.; Klassen, T.Q.; Valenzano, R.; McIlraith, S.A. Reward machines: Exploiting reward function structure in reinforcement learning. *J. Artif. Intell. Res.* **2022**, *73*, 173–208. [CrossRef]
38. Zhai, Y.; Baek, C.; Zhou, Z.; Jiao, J.; Ma, Y. Computational benefits of intermediate rewards for goal-reaching policy learning. *J. Artif. Intell. Res.* **2022**, *73*, 847–896. [CrossRef]
39. Van Seijen, H.; Fatemi, M.; Romoff, J.; Laroche, R.; Barnes, T.; Tsang, J. Hybrid Reward Architecture for Reinforcement Learning. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2017; Volume 30.