processes

MDPI

Article

Hybrid AC/DC Transmission Grid Planning Based on Improved Multi-Step Backtracking Reinforcement Learning

Zhe Wang [1], Yuxin Dai [1], Wenxin Yang [1], Yunzhang Yang [1], Zhiqi Zhang [1], Yahan Hu [1], Jianquan Liao [2,*] and Tianchi Wu [2]

[1] State Grid Shaanxi Electric Power Company Limited Research Institute, Xi'an 710000, China; wangzhe87108@163.com (Z.W.); daiyuxin@jyy.sn.sgcc.com.cn (Y.D.); yangwenxin@jyy.sn.sgcc.com.cn (W.Y.); yangyunzhang@jyy.sn.sgcc.com.cn (Y.Y.); zhangzhiqi@jyy.sn.sgcc.com.cn (Z.Z.); huyahuan@jyy.sn.sgcc.com.cn (Y.H.)

[2] College of Electrical Engineering, Sichuan University, Chengdu 610065, China; 2023223035187@stu.scu.edu.cn

* Correspondence: jquanliao@scu.edu.cn

Abstract

Hybrid AC/DC transmission expansion planning must balance investment cost, supply reliability and AC/DC stability, which challenges conventional mathematical programming and heuristic methods. This paper proposes a multi-objective planning framework based on an improved multi-step backtracking α-Q(λ) reinforcement learning algorithm with eligibility traces and an adaptive learning factor. A tri-objective model minimises annual economic cost, expected power shortage and a comprehensive electrical index that combines electrical betweenness, commutation-failure margin and effective short-circuit ratio. The mixed-integer planning problem is reformulated as an interactive learning process, where the state encodes candidate line construction decisions, the action builds or cancels lines, and the eligibility-trace matrix is used to quantify line importance. Case studies on the Garver-6 system, the IEEE 24-bus reliability test system and a 500 kV regional hybrid AC/DC grid show that, compared with classical Q-learning, the proposed method yields lower annual cost, reduced expected power shortage and improved AC/DC stability; in the 500 kV system, the expected annual power shortage is reduced from 70,810 MWh to 28,320 MWh.

Keywords: transmission network planning; reinforcement learning; multi-step backtracking; multi-objective optimisation; eligibility trace; α-$Q(\lambda)$ algorithm; AC/DC stability

Check for updates

Academic Editors: Wennan Zhuang, Xia Shen, Jiayong Li and Guangsheng Pan

Received: 19 November 2025
Revised: 15 December 2025
Accepted: 16 December 2025
Published: 19 December 2025

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Traditional transmission planning often depends heavily on the experience of human planners. They develop a range of potential solutions and then choose a final plan by comparing the technical and economic pros and cons of each. As modern power systems grow more complex in both structure and operation, we have seen a shift toward mathematical optimisation. This approach uses objective functions and constraints to find the best plan. But applying these mathematical methods in real-world situations is challenging. For one, the growing number of optimisation goals makes it hard to build a single objective function that captures every real-life requirement. At the same time, as systems get larger and involve more variables, the models become highly nonlinear and sometimes non-convex. This makes finding a true global solution much more difficult. These limitations affect the reliability and effectiveness of pure mathematical optimisation in practical engineering. There is a clear need for new types of solutions. We require decision-making tools that not

only produce more credible results but also help planners better analyse grid structures and evaluate different proposals. In recent years, advanced DC devices such as modular multilevel converter–coordinated fault current limiters and DC series–parallel power flow controllers have been proposed to enhance fault current limiting capability and power flow controllability in DC and hybrid AC/DC grids [1,2], which further highlights the need for systematic planning methods that can fully exploit these flexible assets.

For solving optimisation planning models, existing literature predominantly employs two major categories of algorithms: mathematical optimisation methods [3–5] and heuristic algorithms.

Mathematical optimisation methods employ mathematical techniques such as linearisation and constraint relaxation to simplify transmission network planning models, thereby enhancing solution quality and computational speed. This facilitates algorithmic convergence towards optimal solutions. Reference [6] addresses the grid voltage issue by relaxing the branch AC power flow equations, proposing a second-order cone relaxation method for solving optimal power flow. Reference [7] employs network flow methods and the maximum-minimum cut theorem, establishing fuzzy sets based on boundary conditions, and utilises branch-and-bound techniques to derive planning solutions. Reference [8] addresses transmission congestion caused by renewable energy sources by enhancing the CCG algorithm through nested columns and constrained production, enabling it to handle large M constraints and massive variable problems. A well-built model allows mathematical methods to find the best solution quickly and reliably. However, as models grow increasingly complex, mathematical optimisation methods become progressively more challenging to formulate. When the scale of the power grid expands and the number of integer variables in the model increases significantly, it may lead to a computational complexity explosion, rendering the problem unsolvable.

Heuristic algorithms, exemplified by genetic algorithms [9] and particle swarm optimisation [10], extend the evolutionary principles or behavioural patterns observed in various organisms within the natural world to the mathematical domain. Through the process of seeking optimal solutions, they derive either optimal or suboptimal solutions to planning problems. Reference [11] proposes an optimisation model for optical cable planning in power optical transmission networks based on genetic algorithms, addressing the issues of slow computation and poor convergence in traditional planning methods. Reference [12] proposes an optimisation model for power information system network architecture based on the artificial fish swarm algorithm, addressing the issue of low identification accuracy encountered by traditional methods when optimising power communication networks. Reference [13] developed a multi-objective transmission network planning model that comprehensively considers economic efficiency, reliability, and environmental sustainability by modifying the inertia weight and learning factor of the particle swarm optimisation algorithm. This approach effectively addresses the challenges faced by traditional algorithms in solving complex planning problems, namely their susceptibility to local optima and slow convergence rates. Although heuristic algorithms offer relatively high computational efficiency, they remain fundamentally random search algorithms. When addressing large-scale optimisation problems, the quality of solutions deteriorates significantly.

The objective of these two types of algorithms is to obtain optimal solutions. The intermediate solution processes hold no practical physical significance for power grids, cannot be directly mapped to physical networks, and prove difficult to extract further knowledge for subsequent practical planning analyses, thus providing limited insights for planners.

In parallel with these algorithmic advances, recent studies have started to incorporate multiple flexibility resources and learning-based surrogates into power system decision-

making. Reference [14] propose a stochastic operating framework for systems with high wind penetration in which dynamic line rating, pumped-hydro storage, common energy storage, demand response and aggregated electric vehicles are co-optimised as a portfolio of flexibility options to enhance reliability and reduce operating costs. Reference [15] introduce input-convex neural networks to learn Optimal Power Flow value functions, enabling fast, security-aware evaluation of AC dispatch decisions under non-convex constraints. These works show that richer flexibility modelling and machine-learning-based approximations can improve economic efficiency and operational security; however, they focus mainly on short-term operation and do not explicitly address long-term hybrid AC/DC transmission expansion planning or provide an interpretable mechanism to extract line-importance information from learning algorithms. This paper therefore extends this line of research by embedding AC/DC stability indicators and eligibility-trace-based knowledge extraction into a multi-objective reinforcement learning framework for transmission grid planning.

Reinforcement learning is an algorithm based on the classical Markov Decision Process (MDP) [16], which simplifies an intelligent agent's learning task by abstracting complex phenomena into straightforward scenarios, reducing it to states, actions, and feedback rewards. Currently, reinforcement learning algorithms represented by Q-learning have been applied in power system research. However, Q-learning employs a single-step policy update, resulting in slow algorithmic convergence and suboptimal learning capabilities. Reference [17] investigated wide-area control strategies for suppressing power system oscillations following large disturbances based on the Q-learning algorithm. Reference [18] designed an optimal control system employing a Q-learning algorithm that accounts for both physical and network uncertainties. Reference [19] proposes an optimal operation model for a combined heat and power supply system, aiming to achieve both economic efficiency and environmental protection within an integrated energy system. This model employs reinforcement learning for solution, though it does not address system planning. Reference [20] employs a multi-agent collaborative Q-learning algorithm, wherein multiple agents autonomously learn within the environment and mutually transfer experience. By constructing a fuzzy logic system, this approach endows robots with the capability to autonomously analyse unfamiliar environments. Reference [21] successfully integrated eligibility traces into the Q-learning algorithm. It used a multi-step backtracking method to help the agent solve multi-objective power flow problems, achieving positive outcomes. Following these developments, This paper draws an analogy between robot path planning and transmission network expansion planning, applying the Q-learning algorithm to this new domain. This shift in perspective does more than just find an optimal solution. It allows us to learn from the algorithm's intermediate decision steps, which were often overlooked before. By analysing these steps, we can assess the importance of various candidate lines. This process enhances the reliability of the final planning results.

Despite these advances, most hybrid AC/DC planning studies still suffer from three main limitations. First, the majority of models optimise investment and reliability, but rarely incorporate explicit AC/DC stability indicators that capture commutation–failure risk and weak-grid conditions in a unified way. Second, mainstream optimisation approaches—whether mathematical programming or meta-heuristics—typically output a single expansion plan, but do not provide planners with interpretable information on which lines are structurally "important" for security and resilience. Third, although reinforcement learning has been introduced into power system decision-making, eligibility-trace-based algorithms and multi-step backtracking mechanisms have not yet been systematically exploited to solve multi-objective hybrid AC/DC.

To address these gaps, this paper pursues the following objectives:

(1) Construct a multi-objective planning model for hybrid AC/DC transmission grids that simultaneously minimises annual economic cost, expected power shortage and a comprehensive electrical index reflecting AC/DC stability.
(2) Develop an improved multi-step backtracking α-$Q(\lambda)$ reinforcement learning planner that embeds eligibility traces and an adaptive learning factor, enabling both fast convergence and explicit extraction of line importance from the agent's trajectories.
(3) Design a practical AC/DC stability evaluation framework that combines electrical betweenness, commutation-failure margin and effective short-circuit ratio, and integrate it with an intuitionistic fuzzy AHP–based weighting scheme for comprehensive plan assessment.

In short, compared with standard Q-learning, the proposed α-$Q(\lambda)$ planner accelerates convergence by leveraging multi-step backtracking and eligibility traces while adaptively tuning the learning rate. Beyond conventional mathematical-programming and heuristic planners, it also outputs an eligibility-trace matrix that highlights structurally critical lines, providing an interpretable decision aid for planners. This study was conducted to fill the above gap.

2. Evaluation Indicators for the Multi-Objective Model

This section and the following two sections constitute the methodological core of the paper. Section 2 builds an indicator system covering economic cost, supply reliability and AC/DC stability. Section 3 then organises these indicators into a tri-objective planning model with explicit constraints on power flow and generator outputs. Section 4 describes the improved α-Q(λ) reinforcement learning algorithm that solves the mixed-integer planning problem by interacting with the environment, while Section 5 focuses on reward design and eligibility-trace analysis, explaining how the agent's learning process is converted into interpretable "line importance" information for planners.

2.1. AC/DC Stability Indicators

With the continuous advancement of China's high-voltage direct current transmission technology, the grid structure of hybrid AC-DC transmission networks has been further refined, and the interconnection of large regions has progressively strengthened. While optimising resource allocation, the risk of grid disturbances and faults has also grown. This has made safety and stability critical concerns. Traditional stability metrics for AC and DC systems, like N-1, N-2, and commutation failure, mainly deal with internal problems. Therefore, we need to study how integrating DC systems affects AC grids and build a complete stability evaluation framework for hybrid AC/DC networks. This section uses AC-DC stability as a key criterion, assessing performance through three aspects: AC power flow distribution, DC commutation failure margin, and the effective short-circuit ratio.

2.1.1. Distribution of System Current Flows

Line load factor: DC systems often carry large amounts of power between distant AC grids. Because of this role, a DC link can act like a major power plant or a large load from the AC system's viewpoint. This significantly changes power flows in the AC network. Lines near the connection point are affected most, as this can push some lines close to their transmission limits, threatening system stability. The load factor τ_k for a transmission line is given by:

$$k_{\tau_k} = P_{\tau_k} / \overline{P}_{\tau_k} \tag{1}$$

Improved Electrical Modulus: The concept of electrical modulus originates from graph theory. It works by modelling the power system as a directed graph, where the connections between generation nodes, load nodes, and transmission lines are analysed. This model can

effectively represent the power transfer characteristics of a real-world grid. Since stability in hybrid AC/DC grids is primarily assessed through line power flows, our work focuses on refining the branch electrical modulus for this specific purpose.

Branch electrical parameters quantify the role that each transmission line plays within the directed graph model. Essentially, they describe how much a specific line contributes to moving power across the entire network. Here's how it works: imagine a case where we inject a unit of active power at node i and withdraw the same amount at node j, with all other nodes set to zero. The power flow that results on any line connecting nodes m and n is designated as P_{mn}. The electrical parameter of this line is defined as:

$$B_e(m,n) = \left| \sum_{i \in G_N, j \in D_N} \sqrt{w_i w_j} P_{mn}(i,j) \right| \tag{2}$$

In the equation, $Be(m,n)$ denotes the electrical parameters of transmission line mn, while w_i and w_j represent the weighting coefficients for generation node i and load node j, respectively, taken as the active power output and load values from the actual system.

To better assess the overall stability of hybrid AC/DC grids, we introduce the Gini coefficient. It evaluates stability by measuring how evenly power flows are distributed across transmission lines, using an improved electrical index derived from the coefficient itself. The Gini coefficient, an economic metric measuring uniformity, is combined with the electrical medium index to structurally analyse system stability by evaluating the uniformity of power flows across branches and nodes within hybrid AC/DC transmission grids. The more uniform the power flows, the smaller the Gini coefficient, indicating stronger system stability.

Within the Gini coefficient framework, the areas enclosed by the actual Lorenz curve and the lines representing absolute uniformity and absolute non-uniformity are denoted as S_A and S_B respectively. The uniformity metric G is then defined as:

$$G = \frac{S_A}{S_A + S_B} \tag{3}$$

Sort the electrical permittivities of all branches in the power grid by magnitude, denoted as $B_{e1}, B_{e2} \cdots \cdots B_{en}$. Using the Lorentz curve for calculation, the S_B area can be obtained:

$$S_B = \sum_{i=1}^{N} \frac{1}{2N} \left(\sum_{q=1}^{i-1} \left(B_q / \sum_{i=1}^{N} B_i \right) + \sum_{q=1}^{i} \left(B_q / \sum_{i=1}^{N} B_i \right) \right) \tag{4}$$

The area of segment S_A is the area of the right-angled isosceles triangle minus the area of segment S_B. Combining Equations (3) and (4), the improved expression for calculating the electrical medium constant of the branch circuit is derived as follows:

$$G_{\mathrm{B}} = \sum_{i=1}^{N-1} (N-i)(B_N - B_i) / N \sum_{i=1}^{N} B_i \tag{5}$$

2.1.2. DC System Commutation Failure Coefficient

During normal operation of the DC system, should the inverter's arc-extinction angle γ fall below the specified threshold $\gamma_{\min}$ (6–8°) at any given moment, commutation failure may occur on the inverter side of the DC system. This paper calculates the commutation failure margin of the DC system by taking the difference between the inverter's arc extinction angle and $\gamma_{\min}$, then comparing this difference with $\gamma_{\min}$. This margin reflects the current operational state of the DC system and indicates the likelihood of commutation

failure occurring. A larger value of this indicator corresponds to a lower probability of commutation failure in the DC system. Its expression is as follows:

$$K_\gamma = \begin{cases} \frac{\gamma - \gamma_{\min}}{\gamma_{\min}}, & \gamma \geq \gamma_{\min} \\ 0 \quad , & \gamma \leq \gamma_{\min} \end{cases} \tag{6}$$

2.1.3. Effective Short-Circuit Ratio

In hybrid AC-DC transmission networks, the magnitude of DC transmission capacity significantly impacts the voltage stability of the AC system. This makes it vital to assess how well the AC system can handle the power delivered by the DC link. A key measure of this strength is the effective short-circuit ratio (*ESCR*), which reflects the overall robustness of the hybrid AC-DC network. The *ESCR* can be found by calculating the combined impact of AC filters, reactive power compensation capacitors, and synchronous phase-shifting reactors at the converter stations. Its value is given by:

$$ESCR = \frac{S_{ac} - Q_c}{P_{dN}} = \frac{V_N^2}{P_{dN} \cdot |Z|} - \frac{V_N^2 \cdot B_C}{P_{dN}} \tag{7}$$

In this formulation, V_N stands for the rated AC-side voltage, while P_{dN} is the rated DC active power at the converter station. The equivalent admittance for filtering and reactive power compensation is given by B_C, and $|Z|$ represents the equivalent impedance of the AC system.

2.2. Weight Calculation Method for Evaluation Indicator Systems Based on Intuition-Fuzzy Analytic Hierarchy Process

Assessing transmission grid planning requires a balanced allocation of weights across different indicators. Most current methods for setting these weights rely on subjective approaches. Commonly used techniques include the Analytic Hierarchy Process and expert survey methods. A key drawback of these methods is their strong dependence on personal judgement. Different experts may assign very different weights to the same indicator, sometimes even giving opposite scores. This variation makes it important to reduce human bias in the calculations. Improving the objectivity and fairness of the process is a central challenge in building a good evaluation system.

The Intuition-Fuzzy Analytic Hierarchy Process offers one way to address this problem. This method combines fuzzy theory with the classic Analytic Hierarchy Process. Fuzzy theory helps handle vague or uncertain information. Combining it with AHP creates what is known as Fuzzy AHP. This version uses fuzzy numbers to make the traditional method more objective. It goes a step further by adding a non-membership function as an intuition function. Using intuitive fuzzy principles helps manage uncertainty and reduces the ambiguity that comes from human judgement. This makes the evaluation process more reliable. This renders the evaluation system more objective and effective.

In Intuition-Fuzzy AHP, to evaluate relationships between indicators, pairwise comparisons are first conducted to derive the intuition-fuzzy judgement matrix H_z, representing relative importance. To quantitatively describe the significance of attributes, definitions are provided as shown in Table 1.

Table 1. Indicator importance rating scale.

Rating Level	Intuitive Fuzzy Number
Factor *i* is of paramount importance compared to factor *j*	(0.90,0.10,0.00)
Factor *i* is significantly more important than factor *j*	(0.80,0.15,0.05)
Factor *i* is markedly more important than factor *j*	(0.70,0.20,0.10)
Factor *i* is slightly more important than factor *j*	(0.60,0.25,0.15)
Factor *i* is equally important to factor *j*	(0.50,0.30,0.20)
Factor *j* is slightly more important than factor *i*	(0.40,0.45,0.15)
Factor *j* is markedly more important than factor *i*	(0.30,0.60,0.10)
Factor *j* is significantly more important than factor *i*	(0.20,0.75,0.05)
Factor *j* is of paramount importance compared to factor *i*	(0.10,0.90,0.00)

Upon obtaining the fuzzy judgement matrix provided by Expert l, the relative weights of the criterion layers shall be calculated according to the following formula:

$$\begin{aligned}(\omega^l)^T &= [\omega_1^{(l)}\omega_2^{(l)}\cdots\omega_n^{(l)}] = [\frac{\sum_{j=1}^{n} h_{1j}^{(l)}}{\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}^{(l)}} \frac{\sum_{j=1}^{n} h_{2j}^{(l)}}{\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}^{(l)}} \cdots \frac{\sum_{j=1}^{n} h_{nj}^{(l)}}{\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}^{(l)}}] \\ &= [(\frac{\sum_{j=1}^{n} \mu_{1j}^{(l)}}{\sum_{i=1}^{n}\sum_{j=1}^{n} \mu_{ij}^{(l)}}, \frac{\sum_{j=1}^{n} \nu_{1j}^{(l)}}{\sum_{i=1}^{n}\sum_{j=1}^{n} \nu_{ij}^{(l)}})(\frac{\sum_{j=1}^{n} \mu_{2j}^{(l)}}{\sum_{i=1}^{n}\sum_{j=1}^{n} \mu_{ij}^{(l)}}, \frac{\sum_{j=1}^{n} \nu_{2j}^{(l)}}{\sum_{i=1}^{n}\sum_{j=1}^{n} \nu_{ij}^{(l)}}) \cdots (\frac{\sum_{j=1}^{n} \mu_{nj}^{(l)}}{\sum_{i=1}^{n}\sum_{j=1}^{n} \mu_{ij}^{(l)}}, \frac{\sum_{j=1}^{n} \nu_{nj}^{(l)}}{\sum_{i=1}^{n}\sum_{j=1}^{n} \nu_{ij}^{(l)}})]\end{aligned} \tag{8}$$

In the formula, ω^l denotes the criterion layer weight matrix provided by expert l; h_{ij} represents the importance of indicator i relative to indicator j within the intuitive fuzzy judgement matrix H_z; μ_{ij} and ν_{ij} correspond, respectively, to the importance and unimportance levels in the aforementioned scaling table.

According to the operation of the direct fuzzy weighted arithmetic mean operator:

$$\kappa = (\kappa_1, \kappa_2, \ldots, \kappa_n) = IF_\omega(h_1, h_2, \ldots, h_n) = (1 - \prod_{i=1}^{n}(1-u_i)^{\omega_i}, \prod_{i=1}^{n} v_i^{\omega_i}) \tag{9}$$

Combining Equations (8) and (9), the calculation weighting for the criterion layer is obtained:

$$H(\kappa_i) = \frac{1-\nu_i}{1+\mu_i}, i = 1, 2, \cdots, n \tag{10}$$

Normalised to:

$$\sigma_i = \frac{H(\lambda_i)}{\sum_{j=1}^{n} H(\lambda_i)}, i = 1, 2, \cdots, n \tag{11}$$

In the formula, σ_i denotes the final weighting coefficient of the criterion layer.

Furthermore, once the criterion-level weights are obtained, the indicator-level weights can be calculated. Assuming the experts' intuitive fuzzy judgement matrix regarding the importance of indicator-level metrics to the criterion level is denoted as H_e, $h_{eij} = (\mu_{eij}, \nu_{eij})$, the normalised indicator-level weight coefficients $\sigma^{(2)}$ can then be derived.

2.3. Calculation of Indicator Scores Based on the Effectiveness Coefficient Method

In transmission network planning, specific numerical values obtained through data collection, statistics, and calculations cannot be directly weighted for use. They must be uniformly converted into dimensionless scoring values for comparison. Transmission network planning evaluations often incorporate both positive and negative indicators,

or deviation-type indicators. As all three types of indicators coexist, each must undergo directional standardisation. This involves first converting negative indicators into positive ones, then transforming deviation-type indicators into positive values based on their deviation distance. The conversion methodology is as follows:

$$\sigma'_{ij} = \frac{1}{p + \max|\sigma_j| + \sigma_{ij}} \tag{12}$$

In the formula, σ_{ij} and σ'_{ij} denote the indicator values for the jth indicator in the ith scheme before and after directional standardisation, respectively; $\max|\sigma_j|$ represents the maximum value of the jth indicator; p is the coordination coefficient, typically set at 0.1. Following the above processing yields the normalised evaluation indicator values. However, as the indicators differ in meaning and units, they undergo further dimensionless processing:

$$\sigma''_{ij} = \frac{\sigma'_{ij}}{\sqrt{\sum_{j=1}^{m} (x'_{ij})^2}} \tag{13}$$

Following the processes of phase alignment and dimensionless transformation, the indicator values now permit comparative analysis between metrics. However, this approach makes it difficult to intuitively ascertain the relative strengths or weaknesses of individual indicators. Consequently, the effectiveness coefficient method has been introduced to convert each indicator value into a percentage-based score, thereby enabling a more intuitive assessment of comparative performance:

$$\sigma'''_{ij} = c + \frac{\sigma''_{ij} - m_j}{M_j - m_j} \times d \tag{14}$$

In the formula, M_j and m_j denote the satisfaction value and the unacceptable limit value of indicator σ_j respectively; c and d are known constants, with c = 60 and d = 40 adopted in this paper. Following conversion via the efficacy coefficient method, the final score range obtained by the evaluation indicator system is 60–100, providing an intuitive representation of the quality of indicator assessment.

In this paper, the coordination coefficient p is set to 0.1, following common practice in multi-criteria evaluation where a small p-value avoids excessive compression of indicator differences while maintaining numerical stability. The constants c = 60 and d = 40 in the effectiveness-coefficient method define a final score range of 60–100, which corresponds to a typical pass/fail threshold used in engineering evaluation. Alternative values were tested in preliminary experiments and were found to have little impact on the relative ranking of planning schemes.

3. Multi-Objective Planning Model for Transmission Grids

3.1. Multi-Objective Planning Model

Considering the economic viability, safety and reliability, as well as AC-DC stability of hybrid AC-DC power grids, a multi-objective planning model is established to minimise the annual comprehensive economic cost, reduce the annual expected power supply shortfall

to the lowest level, and achieve the smallest system electrical coefficient. The model's objective function is as follows:

$$\begin{cases} F = \min(F_1, F_2, F_3) \\ \min F_1 = \rho_a \sum\limits_{\tau_k \in \tau} c_{\tau_k} \tau_k + (1+\rho_e)\left(\sum\limits_{i \in G_N} \rho_{bi} P_{gi} + \rho_c \sum\limits_{\tau_k \in \tau, \tau^Y} r_{\tau_k} P_{\tau_k}^2\right) \\ \min F_2 = \sum\limits_{i \in D_N} E\left(P_{di}^{clp}\right) \\ \min F_3 = \sum\limits_{\substack{i \in N \\ \tau \in \tau, \tau^Y}} E(B_i, B_\tau) \end{cases} \tag{15}$$

The constraints are as follows:

$$\begin{cases} \sum\limits_i P_{g_i} + \sum\limits_{\tau_k(in)=i} P_{\tau_k} - \sum\limits_{\tau_k(out)=i} P_{\tau_k} = P_{di} \quad i \in N \\ P_{\tau_k} - b_{\tau_k}\left(\theta_{k(out)} - \theta_{k(in)}\right) = 0 \qquad \tau_k \in \tau \\ -\overline{P}_{\tau_k} \le P_{\tau_k} \le \overline{P}_{\tau_k} \qquad \tau_k \in \tau \\ \underline{P}_{Gi} \le P_{gi} \le \overline{P}_{Gi} \quad i \in G_N \\ -\pi \le \theta_i \le \pi \qquad b \in N \\ \theta_o = 0 \end{cases} \tag{16}$$

In the equation, $\tau_k(in)$, $\tau_k(out)$, $\theta_k(in)$, and $\theta_k(out)$ denote the power flow and phase angle at the two end nodes of line τ_k, where *in* indicates the inflow node and out indicates the outflow node; N is the node set comprising all nodes in the system; P_{di} is the load value at node i; $b_{\tau k}$ is the admittance of line τk; $\overline{P}_{\tau_k}$ is the maximum transfer power of line τ_k; $\overline{P}_{Gi}$ and $\underline{P}_{Gi}$ denote the upper and lower limits of active power output for generator unit gi; θ_o is zero, representing the phase angle of the balanced node; $F1$ denotes the annual comprehensive economic cost, encompassing line construction expenditure, system operational expenses, network loss costs, and maintenance outlays; $F2$ represents the system's expected power supply deficiency, determined via Monte Carlo simulation; $F3$ signifies the system's composite electrical index, calculated by incorporating node electrical indices and branch electrical indices derived from the Gini coefficient; B_i and B_τ denote the electrical parameters of node i and line τ, respectively.

In addition, the notation in (15) and its constraints is defined as follows. Index $i \in \mathcal{N}$ denotes buses, and $g \in \mathcal{G}$ denotes generating units connected to the buses. Let $\mathcal{T}$ be the set of candidate transmission lines, and let $\tau_k \in \mathcal{T}$ denote the k-th candidate line. For each line τ_k, the binary decision variable $x_{\tau_k} \in \{0,1\}$ indicates whether this line is constructed $(x_{\tau_k} = 1)$ or not $(x_{\tau_k} = 0)$. The set $\mathcal{T}_i \subseteq \mathcal{T}$ collects all lines incident to bus i, so that the nodal power-balance constraints are written as the net power injection at bus i(generation minus load) being equal to the sum of power flows on all lines in $\mathcal{T}_i$. The active power output of generating unit g at bus i is denoted by P_{gi}, and $\underline{P}_{gi}$ and $\overline{P}_{gi}$ represent its lower and upper limits, respectively. The maximum transferable power of line τ_k is written as $\overline{P}_{\tau_k}$. In the Monte Carlo-based reliability evaluation associated with F_2, $\omega \in \Omega_{\mathrm{MC}}$ indexes the contingency and load scenarios, $\pi(\omega)$ denotes the probability of scenario ω, and $P_{\mathrm{shed}}(\omega)$ is the total load shedding in that scenario. Finally, $\rho \in \Omega_{\mathrm{sch}}$ is used to index candidate planning schemes in the multi-objective evaluation, where Ω_{sch} is the set of all feasible schemes explored by the reinforcement learning agent.

The unit investment, operation and maintenance costs adopted in this study are taken from recent planning guidelines and case studies for regional 500 kV grids in China. The network loss cost is valued at 1000 RMB/MWh, which is consistent with the average marginal cost of energy used in long-term planning studies of large-scale AC/DC systems. Although more detailed cost models (e.g., time-varying loss prices or device-specific main-

tenance contracts) could be incorporated, the adopted cost parameters provide a reasonable approximation for comparing alternative expansion schemes at the planning horizon.

3.2. DC System Equivalence

In traditional planning for AC-DC hybrid grids, the DC system is usually modelled as a two-terminal system. Power flow in the hybrid network is then solved using unified or sequential methods. This detailed representation captures the DC system's operational behaviour and offers high accuracy. However, in many practical hybrid grid projects, the DC system—which is the main focus—often connects to the AC grid at just one end. For this reason, our model treats the DC system as a single-terminal link to the AC network. It acts as an external power delivery path supplying electricity to other grids. Under this setup, the DC system can be represented as a PQ node, equivalent to a load in the planning model. This equivalent load approach simplifies computation while still capturing the effect of the DC link on AC power flow patterns.

It is acknowledged that representing the DC system as a PQ node neglects detailed DC dynamic behaviour such as commutation failures and DC voltage oscillations. However, for long-term transmission expansion planning, where the decision variables are line routes and capacities, the main influence of the DC system is its power injection pattern into the AC grid. This static PQ-node equivalence is therefore widely adopted in hybrid AC/DC planning studies, while more detailed DC dynamic models are usually reserved for operation, protection and transient-stability analyses. In this paper, the proposed framework is intended for planning-level studies; extending it to incorporate more detailed DC dynamics will be an important topic for future work.

4. Reinforcement Learning Algorithm Design: Improved α-Q(λ)

4.1. The Concept of Eligibility Traces

The concept of eligibility trace comes from cognitive science, where it describes a type of memory mechanism. In simple terms, it acts as a temporary record of something that has happened. When an agent enters a certain state or takes an action, the eligibility trace stores information linked to that event. This gives the agent useful background knowledge for future decisions—often referred to as credibility. Over time, the strength of an eligibility trace gradually fades. This reflects how we tend to forget details of events that happened longer ago. If the same event occurs again, the trace is updated with a new value, which adds to what remains of the old one. By using eligibility traces, agents can learn not only from future rewards but also by recalling past states and actions. This helps create a cognitive model that works more like human memory.

Eligibility traces employ two accumulation methods: replacement and accumulation.

The accumulation method involves incrementing the existing eligibility trace value by one when an agent accesses a particular event, whilst the eligibility traces for other events undergo decay:

$$e_t(s,a) = \begin{cases} \gamma\lambda e_{t-1}(s,a) & (s,a) \neq (s_t,a_t) \\ \gamma\lambda e_{t-1}(s,a)+1 & (s,a) = (s_t,a_t) \end{cases} \tag{17}$$

Replacement refers to the process whereby, when an agent accesses a particular event, the eligibility trace for that state is set to 1, whilst the eligibility traces for all other events undergo decay:

$$e_t(s,a) = \begin{cases} \gamma\lambda e_{t-1}(s,a) & (s,a) \neq (s_t,a_t) \\ 1 & (s,a) = (s_t,a_t) \end{cases} \tag{18}$$

In the above equation, $e_t(s,a)$ denotes the eligibility trace of event (s,a). t represents the decision time of the agent, γ is the discount factor, and λ is the decay coefficient. Following each action executed by the agent, the eligibility trace continuously decays at an exponential rate of $(\gamma\lambda)$. The accumulation methods for the two eligibility traces are illustrated in Figure 1.

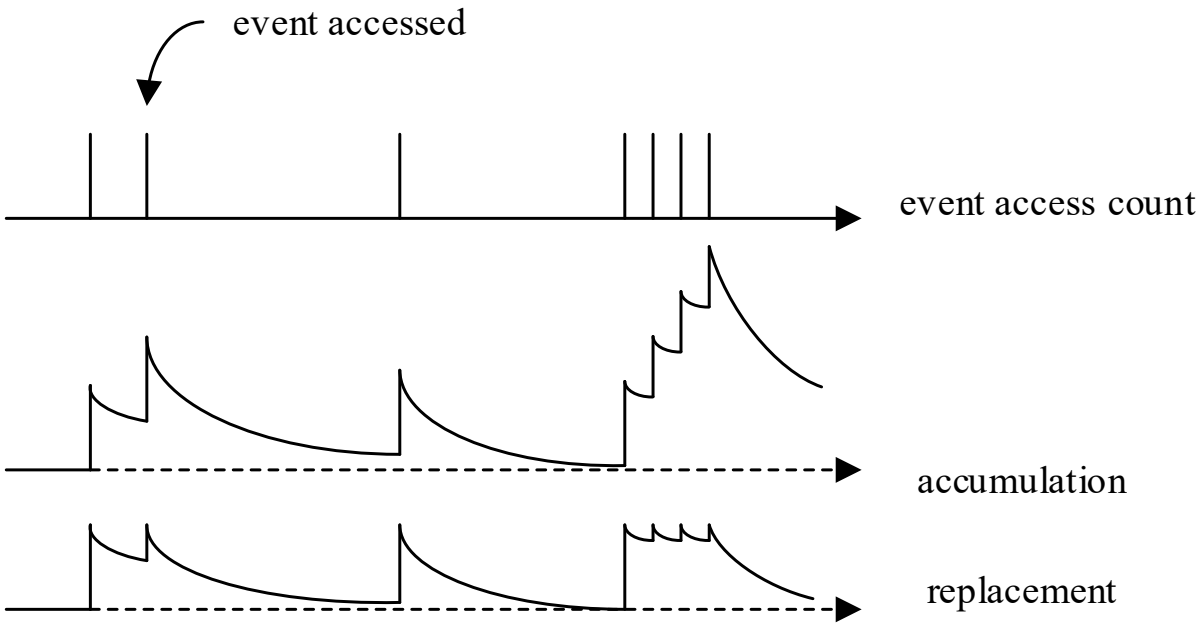


Figure 1. Two methods of accumulating eligibility traces.

4.2. α-Q(λ) Algorithm Procedure

To efficiently update and invoke the eligibility trace, eligibility trace matrix M is established with the same dimensionality as the knowledge matrix Q. This matrix is updated at each agent action. The matrix is organised with state-action as rows and columns, and is located through the events executed by the agent. To account for the number of times a line is selected, this paper employs the cumulative Formula (17) to update eligibility trace values. Consequently, the value function update formula for the α-$Q(\lambda)$ algorithm is as follows:

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha\delta_t e_t(s,a) \tag{19}$$

$$\delta_t = r_t + \gamma \underset{a}{max} Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t) \tag{20}$$

$$\alpha = \alpha_0 \cdot \frac{C_{f_{t-1}}}{C_{f_t}} \cdot \frac{C_{f_{best}}}{C_{f_t}} \tag{21}$$

In the equation, C_{f_t} and $C_{f_{t-1}}$ denote the objective function values at states s_t and $s_{t\text{-}1}$ respectively, while $C_{f_{best}}$ represents the optimal value of the objective function across multiple learning rounds. Compared to the classical Q-learning algorithm, the α-$Q(\lambda)$ algorithm introduces an eligibility trace to record past states and actions, thereby enhancing the speed of reward feedback transmission and computational performance. The learning process of the Algorithm 1 is as follows:

Algorithm 1: α-$Q(\lambda)$ algorithm.

Input
State space S, action space A, discount factor γ, decay coefficient λ
Output
Optimal planning solution τ^*
Randomly initialise $Q(s,a)$, $s \in S$, $a \in A$.
Repeat (one learning round)
  Initialise initial state s_0;
  Repeat (one step within a learning round)
    Based on state s, select action a using the given action policy;
    Execute action a, observe next state s' and reward function r;
    Update $e(s,a)$ according to (17);
    Estimate the action a' for the next state, and update $Q(s,a)$ per (19);
    $s \leftarrow s'$;
  Until state s is a terminal state;
Until the state-action value function converges;
$\forall s \in S : \tau^*(s) = \underset{a \in A}{\operatorname{argmax}} Q(s,a)$

When solving transmission network planning models using the α-$Q(\lambda)$ algorithm, the accumulation of eligibility traces enables the capture of agents' action paths, thereby identifying critical lines during the planning process. Within the eligibility trace matrix, a line corresponding to a high action eligibility trace indicates that agents have repeatedly executed that action. Consequently, it can be inferred that agents have frequently considered the construction of that line, rendering it a significant candidate line in the planning process.

5. Reward Design and Eligibility Trace Analysis

5.1. Eligibility Trace Matrix and Knowledge Extraction

During the learning process of the intelligent agent, the α-$Q(\lambda)$ algorithm progressively constructs a knowledge matrix Q. Simultaneously, the algorithm establishes an eligibility trace matrix M of identical size to Q:

$$EG = \begin{array}{c|c} & \begin{matrix} a_1^J & a_2^J & \cdots\cdots & a_n^D \end{matrix} \\ \begin{matrix} s_1 \\ s_2 \\ \vdots \\ \vdots \\ s_n \end{matrix} & \begin{bmatrix} eg\left(s_1, a_1^J\right) & eg\left(s_1, a_2^J\right) & \cdots\cdots & eg\left(s_1, a_n^D\right) \\ eg\left(s_2, a_1^J\right) & eg\left(s_2, a_2^J\right) & \cdots\cdots & eg\left(s_2, a_n^D\right) \\ \vdots & \vdots & \cdots\cdots & \vdots \\ \vdots & \vdots & \cdots\cdots & \vdots \\ eg\left(s_n, a_1^J\right) & eg\left(s_n, a_2^J\right) & \cdots\cdots & eg\left(s_n, a_n^D\right) \end{bmatrix} \end{array} \quad (22)$$

The values within the eligibility trace matrix M represent the credibility of values within the knowledge matrix Q, and participate in the updating of the knowledge matrix Q. The greater the credibility, the larger the proportion of feedback rewards derived from past events. Based on this principle, the values within the eligibility trace matrix correspond to the frequency and importance of events executed by the intelligent agent in the past. The greater the value, the more significant the event, meaning the corresponding pathway for the event action is more important.

From a planning perspective, the eligibility trace associated with a given line-building or line-cancelling action can be interpreted as a measure of how frequently and how recently the RL agent relied on that decision to obtain high-quality expansion schemes. Lines with

high eligibility traces are repeatedly selected in successful episodes and rarely removed thereafter, indicating that they play a structurally important role in maintaining economic efficiency, reliability and AC/DC stability across many scenarios. Conversely, actions with persistently low traces correspond to lines that are seldom used or frequently cancelled during learning, suggesting lower importance in the final plan. This interpretation allows the eligibility-trace matrix to be viewed as a data-driven analogue of traditional sensitivity analysis or contingency ranking, providing planners with physically meaningful insight into which corridors should be prioritised for route surveys and reinforcement.

5.2. Reward Function and Termination Criteria

In reinforcement learning, each action taken by an agent requires feedback rewards from the environment to evaluate the quality of its behaviour. Within multi-objective planning models, however, the presence of multiple objectives prevents agents from obtaining feedback rewards based on a single criterion. Therefore, this paper employs a weighting method. By applying a ratio transformation, the objective values under each goal are converted into dimensionless values within the range [0,1]. Finally, by assigning different weights, a composite feedback reward is obtained:

$$R = \beta_1 \cdot R1 + \beta_2 \cdot R2 + \beta_3 \cdot R3 \tag{23}$$

$$\begin{cases} R1 = \begin{cases} \frac{F1_{best} - F1_{now}}{F1_{best}} & N-1\text{pass} \\ -1 & N-1\text{fail} \end{cases} \\ R2 = \begin{cases} \frac{F2_{best} - F2_{now}}{F2_{best}} & N-1\text{pass} \\ -1 & N-1\text{fail} \end{cases} \\ R3 = \begin{cases} \frac{F3_{best} - F3_{now}}{F3_{best}} & N-1\text{pass} \\ -1 & N-1\text{fail} \end{cases} \end{cases} \tag{24}$$

In the formula, $F1_{best}$, $F2_{best}$, and $F3_{best}$ denote the optimal comprehensive economic cost, the minimum expected power shortage, and the minimum comprehensive electrical index, respectively; $R1$, $R2$, and $R3$ represent the feedback rewards corresponding to each of the three objectives; β_1, β_2, and β_3 denote the respective weights of $R1$, $R2$, and $R3$. In the initial state, $F1_{best}$ t ($F2_{best}/F3_{best}$) assumes a sufficiently large positive value. For planning schemes failing N-1 grid security constraint verification, a feedback reward of -1 is applied to deter the agent from selecting such schemes. When a planning scheme satisfies the N-1 grid safety constraint verification, the respective objectives for that state are computed. Taking $F1$ as an example: when $F1_{best} < F1_{now}$, a positive feedback reward $R1 > 0$ is provided. This indicates the agent receives positive reinforcement from the environment, leading it to favour exploring this scheme. Conversely, the agent tends to avoid such schemes to prevent receiving negative feedback. A larger gap between $F1_{best}$ and $F1_{now}$ leads to a stronger environmental feedback reward. This greater reward pushes the agent harder toward better solutions. This design helps the agent focus on actions that cause major improvements in $F1_{now}$, while reducing rewards when the target is nearly met. That way, the agent avoids getting stuck too early with a suboptimal solution. By tuning the weights assigned to each objective's feedback, we can also steer the agent to emphasise certain goals. For example, setting β_2 to the highest weight encourages the agent to find a more cost-effective planning solution.

The reward function in (24) adopts a piecewise design: when a newly generated plan improves the dimensionless scores of all three objectives compared with the current best plan, a positive reward is given; otherwise, a negative reward encourages the agent to search alternative expansion paths. The weights β_1 = 0.5, β_2 = 0.3 and β_3 = 0.2 reflect the planning preference that economic cost is slightly more important than reliability

and AC/DC stability. These values were selected based on expert consultation and a simple sensitivity analysis, which showed that moderate perturbations of β_1–β_3 do not change the qualitative structure of the optimal plans. A more systematic multi-stakeholder weight-setting procedure will be considered in future work.

In summary, the specific algorithmic flow details of the α-$Q(\lambda)$ algorithm are illustrated in Figure 2.

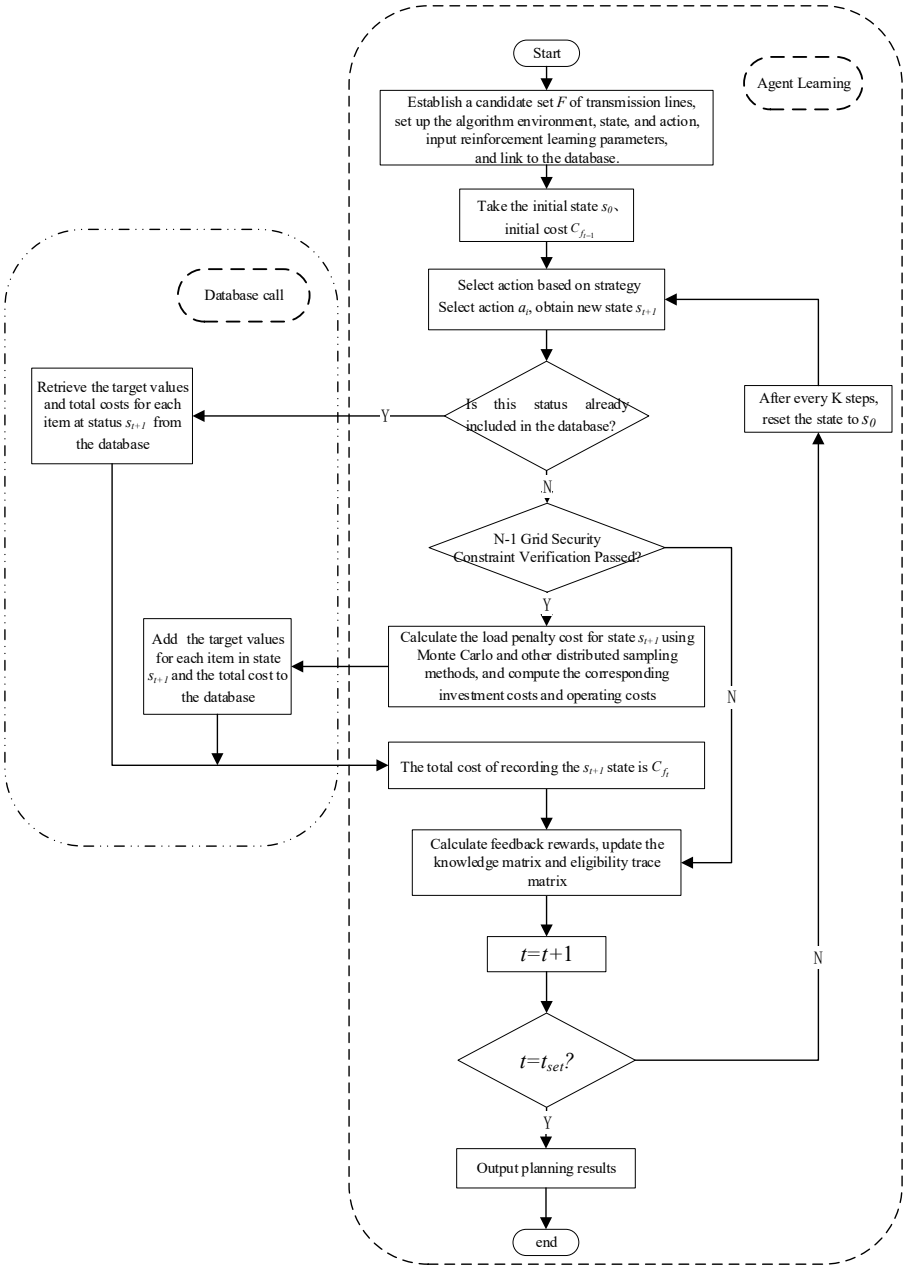


Figure 2. Flow-chats of α-$Q(\lambda)$ algorithm.

6. Results

This section presents case studies on three systems to evaluate the performance of the proposed multi-objective planning framework and the improved α-$Q(\lambda)$ planner. this section first employs the Garver-6 system and IEEE 24-RTS system as case studies to validate the α-$Q(\lambda)$ algorithm through analytical verification. Subsequently, it conducts planning scheme calculations for a DC transmission grid in a southwestern region. The system hardware configuration comprises: an i5-8750H CPU operating at 2.2 GHz, 16 GB

RAM, and a personal computer running the Windows 10 64-bit operating system. The reinforcement learning programme was developed on the Python 3.7 platform. Expected load shedding quantities were calculated by invoking CPLEX, while the contingency plan database was established using Mysql 8.0. Common parameters in the case study analysis included: Monte Carlo and other distributed sampling methods with segment count H = 4 and 1000 sampling iterations. The reinforcement learning reward function weights β_1, β_2, and β_3 were set to 0.5, 0.3, and 0.2, respectively.

6.1. Garver-6 System

6.1.1. Planning Scheme for the Garver-6 System

The Garver-6 system parameters in this paper are identical to those in Reference [3]. Based on Garver's stepwise expansion method [3], the preliminary planning results indicate that the number of lines to be constructed, $n^x = 9$. The multi-objective planning model was solved using the α-Q(λ) algorithm. The resulting planning outcomes were compared with those from the traditional Q-learning algorithm, as shown in Tables 2 and 3. The system network structure topology is depicted in Figure 3.

Table 2. Garver-6 node system planning results.

Scheme	Algorithm	Investment Costs (RMB 10,000)	Operational and Maintenance Costs (RMB 10,000)	Power Shortfall (MWh)	Comprehensive Electrical Betweenness
1	α-$Q(\lambda)$	2530.6	12,561.7	8790	0.240
2	Q-learning	2196.9	12,980.1	6420	0.265

Table 3. New lines under consideration for each planning scheme of the Garver-6 node system.

Scheme	Newly Constructed Line
1	2–6(4), 3–5(2), 4–6(3), 5–6(1)
2	1–5(1), 2–6(4), 3–5(2), 4–6(3)

According to the comparison of the above table, the α-$Q(\lambda)$ algorithm proposed in this paper can effectively solve the transmission network expansion planning problem. The difference between the Pareto planning solution set scheme 1 given by the α-$Q(\lambda)$ algorithm and the original system planning result is that more primary lines 3–5, 4–6, 5–6 are constructed, and the investment cost is 9,035,000 yuan higher. At the cost of 9,035,000 yuan, the power flow transmission structure of the power grid is optimised, the system network loss and power generation cost are effectively reduced, and the operation and maintenance cost is reduced by 9.635 million yuan. At the same time, the power supply reliability of node 3 and node 5 is improved, and the power shortage cost of the system under fault conditions is effectively reduced. The expected power shortage is reduced from 54.12 million kilowatt-hours to 8.79 million kilowatt-hours and compared with the original system. The comprehensive electrical betweenness is reduced from 0.27496 to 0.240367, and the power flow balance of the line is improved, which reduces the power loss caused by disconnecting the key lines, thus improving the reliability. The difference between the scheme 2 and the original system planning result is that more lines 1–5 and 2–6 are constructed, and the investment cost is 3.337 million yuan higher than that of the original system. Compared with the original system, the power flow is optimised to a certain extent, and the system network loss and power generation cost are effectively reduced, so that the operating cost is reduced by 4.184 million yuan. At the same time, the power supply reliability of node 3 and node 6 is improved, and the power shortage cost of the system under fault conditions is effectively reduced, so that the expected power shortage

is reduced by 47.7 million kilowatt hours. At the same time, compared with the original system, the comprehensive electrical betweenness is reduced to 0.26532, but there is still a certain gap compared with scheme 1. In the process of scheme selection, it should be based on the actual project situation, investment cost of funds, comprehensive consideration, in the scheme selection.

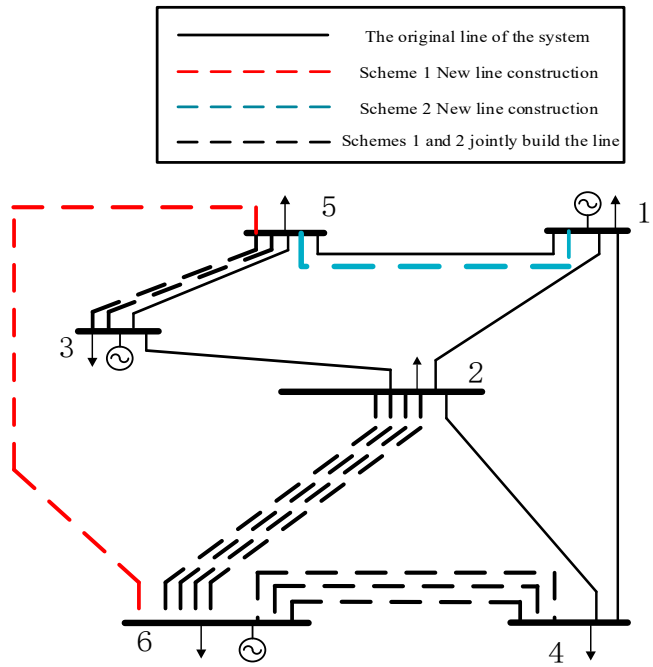


Figure 3. Garver-6 node system planning scheme.

6.1.2. Eligibility Trace Analysis for the Garver-6 System

A partial excerpt of the eligibility trace matrix learned by the intelligent agent within the Garver-6 system is presented in Table 4. In the table, values below 0.001 are represented by 0.001.

Table 4. Garver-6 node system eligibility trace matrix.

State\Action	1–2 ^{J}	2–6 ^{J}	3–5 ^{J}	5–6 ^{J}	4–6 ^{D}
[2–6(3),3–5(1),4–6(1)]	0.001	0.279	0.377	0.223	0.003
[2–6(2),3–5(2),4–6(1)]	0	0.501	0.137	0.291	0.050
[2–6(3),3–5(2),5–6(1)]	0.001	0.319	0.001	0.007	0.030
[2–6(4),3–5(2),4–6(3)]	0	0	0.031	0.671	0.184

Note: Superscripts *J* and *D* indicate, respectively, a line-construction action and a line-cancellation action applied to the corresponding candidate transmission corridor.

In the table above, action 2–6 ^{J} denotes the construction of transmission line 2–6 connecting nodes 2 and 6, while action 4–6 ^{D} signifies the cancellation of transmission line 4–6 linking nodes 4 and 6. The data reveals that actions 2–6 ^{J}, 3–5 ^{J}, and 5–6 ^{J} were repeatedly executed by the agent across multiple states, whereas action 1–2 ^{J} occurred only infrequently. Notably, in states [2–6(4), 3–5(2), 4–6(3)], the eligibility trace for action 4–6 ^{D} is significantly greater than in other states. This suggests that when three Line 4–6 connections already exist, the system tends to remove one of them. This change helps save significant investment costs.

Consider the state [2–6(3), 3–5(2), 5–6(1)] as an example. Actions 1–2 ^{J} and 3–5 ^{J} show the lowest eligibility trace values, both at 0.001. These low values mean the agent considered routes 1–2 and 3–5 less useful after trying them. It then avoided these options in later

explorations, allowing their trace values to decay over time. In contrast, actions 2–6 ^{J} and 4–6 ^{D} carry much higher trace values of 0.319 and 0.501. This suggests the agent repeatedly explored these two actions, marking lines 2–6 and 4–6 as important in its planning process. Looking across multiple states, lines 2–6, 3–5, 4–6, and 5–6 often show strong eligibility traces. This pattern points to their potential role as key pathways. Yet when comparing final plans, line 2–6 appears particularly often in scenarios with four built circuits. It costs less to build, helps lower operational expenses, and improves grid reliability. Consequently, planners need not accord Line 2–6 additional consideration during the planning process. Thus, through the intelligent agent planning process, Lines 3–5, 4–6, and 5–6 may be deemed critical candidate lines within the Garver-6 node system. Furthermore, the fact that different actions correspond to different eligibility traces under varying states indicates that the determination of important candidate lines based on eligibility traces must be predicated upon the agent's state. Different states may yield different important lines. In state [2–6(4), 3–5(2), 4–6(3)], constructing a single circuit 5–6 effectively enhances power flow uniformity, reduces network losses and operational maintenance costs, and improves grid reliability. In other states, the importance of circuit 5–6 is slightly lower. Consequently, the α-$Q(\lambda)$ algorithm effectively extracts planning experience, highlights important circuits, and assists power grid personnel in decision-making.

6.1.3. Algorithm Convergence Performance

Based on the Garver-6 system, the transmission network planning model was solved using both the α-$Q(\lambda)$ algorithm and the Q-learning algorithm. Computations revealed that the optimal planning solutions obtained by both algorithms were consistent, though their convergence characteristics differed. The convergence characteristic curves for the α-$Q(\lambda)$ algorithm and the Q-learning algorithm are shown in Figure 4.

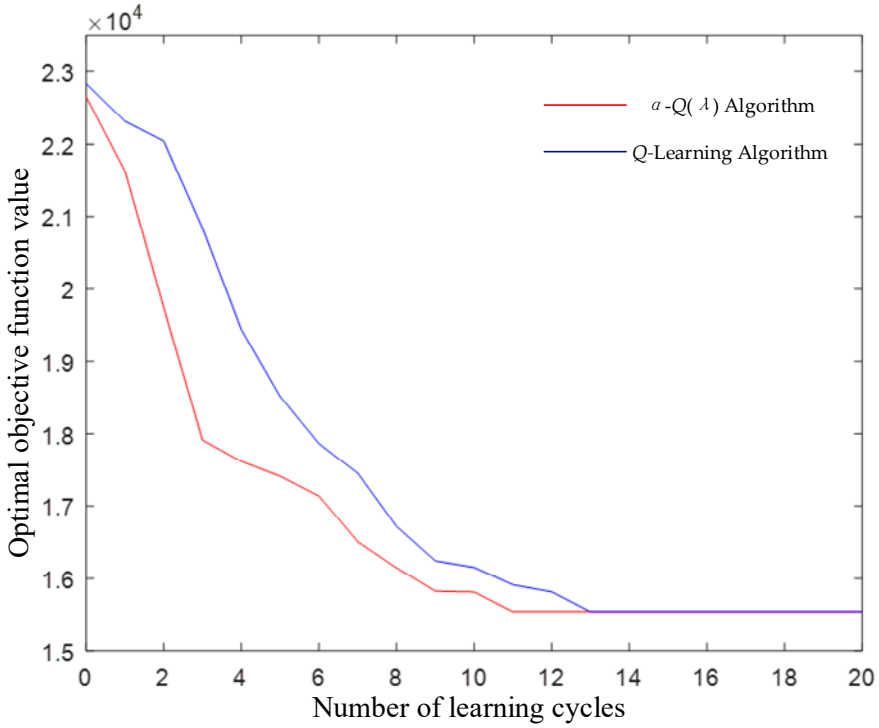


Figure 4. Convergence Curves of α-$Q(\lambda)$ Algorithm and Q-learning Algorithm.

As shown in the figure above, both the α-$Q(\lambda)$ algorithm and the Q-learning algorithm make decisions based on the optimal objective function value. As the number of learning rounds increases, the optimal objective function values achieved by both algorithms in a

single round exhibit a downward trend. The α-$Q(\lambda)$ algorithm demonstrates a faster rate of decline, indicating that after being enhanced with an adaptive learning factor, the agent possesses stronger exploratory capabilities during the early learning phase and can more quickly discover states with superior objective values. After several learning rounds, the agent in the Q-learning algorithm accumulates experience, leading to reduced exploration of unknown states and diminished learning efficiency. In contrast, α-$Q(\lambda)$ adjusts the learning rate based on the adaptive learning factor, enhancing the agent's ability to explore unknown states and enabling faster convergence to the optimal solution. By round 11, the α-$Q(\lambda)$ algorithm agent explores to the optimal state, whereas the Q-learning algorithm agent reaches the optimal state only by round 13. Thus, the adaptive learning factor α accelerates the agent's learning efficiency for superior states and enhances the algorithm's convergence speed.

6.1.4. Algorithm Computational Performance

Table 5 shows the computation time of the α-$Q(\lambda)$ algorithm before and after establishing the contingency plan database. As indicated by the data, prior to establishing the contingency plan database, the α-$Q(\lambda)$ algorithm required approximately 2 days and 4 h to converge. After establishing the contingency plan database, convergence was achieved in just 358 s, demonstrating that the contingency plan database effectively enhances the algorithm's computational performance.

Table 5. α-$Q(\lambda)$ Algorithm Computation Time Before and After Establishing the Contingency Plan Database.

Contingency Plan Database	α-$Q(\lambda)$ Algorithm Computation Time
Not established	Approximately 2 days and 4 h
After establishment	358 s

6.2. IEEE 24-RTS System

6.2.1. Planning Scheme for the IEEE 24-RTS System

The original model of the IEEE 24-RTS system comprises 24 nodes, 34 transmission lines, and 32 generators. It is assumed that in a future planning year, the load will increase to three times the current level, while the installed capacity of generators will grow to 3.3 times the present figure. In this case study, the set of candidate lines totals $n = 41$, with preliminary analysis determining the desired number of lines to be constructed as $n_x = 11$. Original network parameters, load data, generator connection locations, and other details are referenced in [22]. The upper limit for agent learning iterations is set to $i_{set} = 5000$, with a reset step size $K = 200$.

Under the aforementioned conditions, the system's optimal planning scheme was determined, with its network topology depicted in Figure 5. The planning results were compared with those generated by the traditional Q-learning algorithm, as presented in Tables 6 and 7:

The comparative results in the table above demonstrate that the α-$Q(\lambda)$ algorithm proposed herein can effectively solve the optimal expansion planning scheme. In Planning Algorithm 1, 12 new transmission lines were added at an investment cost of 92.167 million yuan; in Scheme 2, 9 new lines were added at an investment cost of 56.867 million yuan. Scheme 1 incurred an investment cost 35.3 million yuan higher than Scheme 2, while its operational cost was 32.268 million yuan lower than Scheme 2′s. Concurrently, the power shortage cost in this model is calculated via Monte Carlo sampling, accounting for system load power shortfalls under high-order faults. Consequently, the agent more effectively learns power system supply deficiencies during training. Calculations indicate that Scheme

1 exhibits a power shortage 27.12 million kWh lower than Scheme 2. Figure 3 shows that the proposed planning scheme strengthens connections between several key load and generation nodes. Under Scheme 1, part of the power flow originally on Line 1–2 is shifted to Lines 1–3 and 1–5. This change improves power flow balance around Node 1 and also greatly lowers the flow on Line 2–4. Meanwhile, Scheme 2 increases power flow along Line 5–10, which helps relieve loading near Node 10. It also causes a reversal in direction on Line 8–10, feeding 57.54 MW back toward Node 8. When comparing Lines 9–12, 15–21, and 16–17, Scheme 1 achieves more balanced flows than Scheme 2. The comprehensive electrical betweenness value is also lower in Scheme 1, reflecting a more evenly distributed power flow across the whole network. As a result, if a line fault occurs, Scheme 1 would lead to a less severe power shortfall compared to Scheme 2.

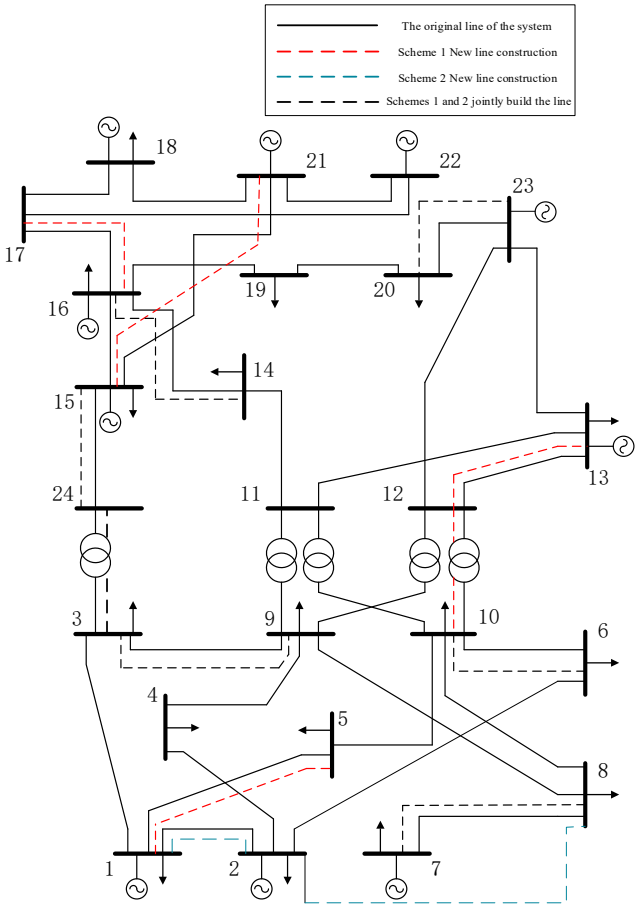


Figure 5. IEEE 24-RTS node system planning scheme.

Table 6. IEEE 24-RTS node system planning results.

Scheme	Algorithm	Investment Costs (RMB 10,000)	Operational and Maintenance Costs (RMB 10,000)	Power Shortfall (MWh)	Comprehensive Electrical Betweenness
1	α-$Q(\lambda)$	9216.7	327,830.1	26,480	0.237
2	Q-learning	5686.7	331,056.9	53,605	0.275

Table 7. New lines under consideration for each planning scheme of the IEEE 24-RTS node system.

Scheme	Newly Constructed Line
1	1–5,3–9,3–24,6–10,7–8,10–12,12–13,14–16,15–21,15–24,16–17,20–23
2	1–2,3–9,3–24,6–10,7–8,14–16,15–24,20–23,2–8

6.2.2. Eligibility Trace Analysis for the IEEE 24-RTS System

Table 8 shows the partial eligibility trace matrix derived from the model.

Table 8. IEEE-24 node system eligibility trace matrix.

State\Action	1–5 ^{J}	3–24 ^{J}	10–12 ^{J}	11–13 ^{J}	10–12 ^{D}	7–8 ^{D}
State 1	0.008	0.001	0.301	0.002	0	0.001
State 2	0.211	0.121	0.001	0.071	0.194	0.001
State 3	0.001	0.344	0.186	0.003	0.013	0

Note: Superscripts *J* and *D* indicate, respectively, a line-construction action and a line-cancellation action applied to the corresponding candidate transmission corridor.

As shown in Table 8, State 1 comprises [1–5, 3–9, 3–24, 7–8, 12–13, 14–16, 15–21, 15–24, 16–17, 20–23], with 10 new lines established; State 2 comprises [3–9, 6–10, 7–8, 10–12, 14–16, 15–21, 15–24, 16–17, 20–23], with 9 new routes established; State 3 comprises [1–5, 3–9, 6–10, 7–8, 12–13, 14–16, 15–21, 16–17, 20–23], with 9 new lines to be constructed. Based on the values in the eligibility trace matrix and a 0.1 threshold, lines 1–5, 3–24, and 10–12 stand out as critical candidate lines.

Looking at State 2 as an example, we can compare it with the final plan produced by the α-$Q(\lambda)$ algorithm. This state does not include lines 1–5, 3–24, and 12–13. Without these connections, the grid structure between nodes 1 and 5 remains relatively weak. To improve reliability, the learning agent tends to choose action 1–5 ^{J}, which adds line 1–5. When we apply the same experience-based threshold of 0.1, the key candidate lines in State 2 are 1–5, 3–24, and 10–12. For State 1, only line 10–12 stands out as critical. In State 3, lines 3–24 and 10–12 are the main candidates. It is worth noting that line 10–12 shows strong eligibility traces across all three states. This pattern suggests the agent repeatedly tried building and removing this line, marking it as one that deserves special attention during planning.

In all three states, the eligibility trace for removing line 7–8 (action 7–8 ^{D}) stays near zero. The reason lies in the system layout: node 7 connects to the main grid through just one line. This setup does not meet the N-1 safety standard [21]. Once the agent builds line 7–8, the line becomes essential to maintain security, so it is never removed again.

6.3. DC Transmission System in a Southwestern Region

6.3.1. Planning Scheme for the 500 kV Hybrid AC/DC Transmission System

Figure 6 shows the main structure of the 500 kV hybrid AC/DC transmission grid in a southwestern region. The AC network comprises 22 buses and multiple 500 kV corridors, while four point-to-point DC transmission systems inject power at nodes 6, 15, 18 and 19, forming a typical regional hybrid AC/DC system used as the planning testbed. Three of these—nodes 20, 21, and 22—are newly planned sites for the target year. In the planning year, the system incorporates four DC transmission systems. The DC systems connected to nodes 6, 18, and 19 deliver 8000 MW of power, while the system connected to node 15 delivers 7200 MW. Taking boundary power into account, the generation capacity and load at each node within the system are detailed in Table 9.

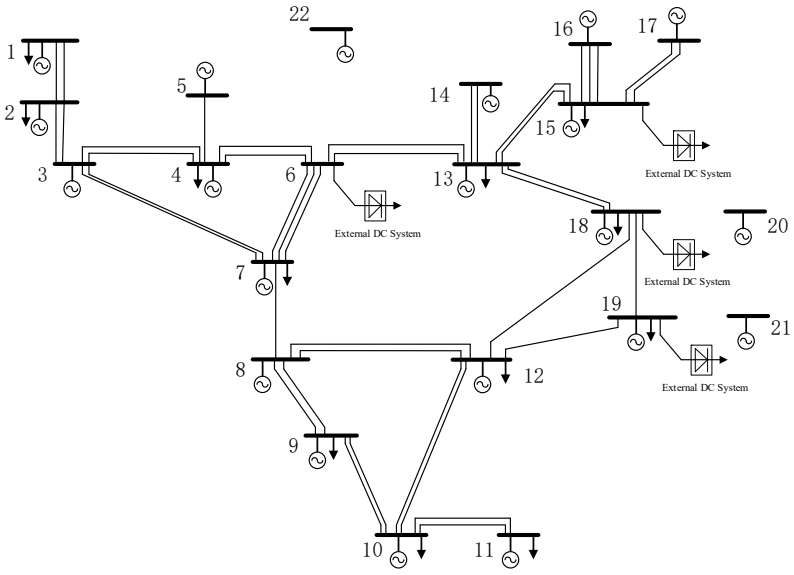


Figure 6. 500 kV grid structure in a southwestern region.

Table 9. Generation capacity and load at various nodes in a southwestern region.

Node	Generating Capacity (MW)	Load (MW)	Node	Generating Capacity (MW)	Load (MW)
1	589.2	18.2	12	1878.9	620.0
2	1301.3	40.5	13	3803.6	2680.7
3	1112.0	\	14	2400.0	\
4	1378.2	40.6	15	\	915.3 + 7200 (DC)
5	1000.0	\	16	4800.0	\
6	\	8000 (DC)	17	3600.0	\
7	139.0	218.9	18	4332.6	2609.1
8	2800.0	\	19	504.6	78.3
9	1738.3	1395.0	20	8000	8000 (DC)
10	39.1	275.1	21	8000	8000 (DC)
11	1257.9	1266.3	22	1500	\

To accommodate the planned annual line π-type connection, lines 13–18 and 18–19 were removed to reduce corresponding construction costs. Network loss costs were set at ¥1000/MWh, with an annual average utilisation of 3786 h and a 6% forced outage rate for all system components. Solving using the proposed α-$Q(\lambda)$ algorithm yielded the planning scheme shown in Tables 10 and 11:

Table 10. Planning outcomes for a DC transmission system in a southwestern region.

Scheme	Investment Costs (RMB 10,000)	Operational and Maintenance Costs (RMB 10,000)	Power Shortfall (MWh)	Comprehensive Electrical Betweenness
1	522,510	4,192,600	28,320	0.319
2	440,890	4,183,770	70,810	0.334

The grid structure of the two planning schemes is shown in Figure 7. From the above table, it can be seen that in the planning scheme 1, there are 16 new lines, with a total investment cost of 5225.1 million yuan; planning scheme 2 includes 14 new lines, and the total investment cost is 440.89 million yuan. The investment cost of planning scheme 1 is 816.2 million yuan higher than that of scheme 2, mainly because the line 7–18 and

12–19 in planning scheme 1 are longer and the construction cost is higher. The operation and maintenance costs of the two planning schemes are close. Planning scheme 1 and planning scheme 2 take into account the operation and maintenance costs. The economic cost of planning scheme 1 is 904.5 million yuan higher than that of planning scheme 2, and planning scheme 2 has better economy. However, from the perspective of power shortage, the power shortage of Plan 1 is 28,320 MWh, which is only one third of Plan 2, and the Comprehensive electrical betweenness of Plan 1 is less than that of Plan 2, indicating that the power flow distribution of Plan 1 is more balanced and the safety and reliability of the grid are higher. This is because the construction of line 7–18 can stably transmit the power generation near node 7 to 18 nodes, reducing the load of line 6–7 and 6–13. Similarly, the construction of line 11–12 can effectively prevent node 11 from becoming an island and improve the security and stability of the system. The above results verify the correctness and effectiveness of the multi-objective planning model for hybrid AC-DC transmission grids proposed in this paper. The large difference in expected power shortage between schemes 1 and 2 (28,320 MWh vs. 70,810 MWh) is mainly driven by contingencies involving heavily loaded AC/DC corridors near nodes 6, 13 and 18. In scheme 2, the absence of line 7–18 and 12–19 forces more power to flow through lines 6–7 and 6–13, making the system more vulnerable to N-1 outages of these corridors. Scheme 1 mitigates this risk by providing alternative transmission paths, thereby substantially reducing the probability and duration of load curtailment in Monte Carlo simulations.

Table 11. New transmission lines under various planning schemes for the DC transmission system in a southwestern region.

Scheme	Newly Constructed Line
1	7–18, 4–5, 6–22(2), 11–12, 12–19, 13–19(2), 18–20(4), 19–21(4)
2	4–5, 6–22(2), 13–19(2), 13–22, 18–20(4), 19–21(4)

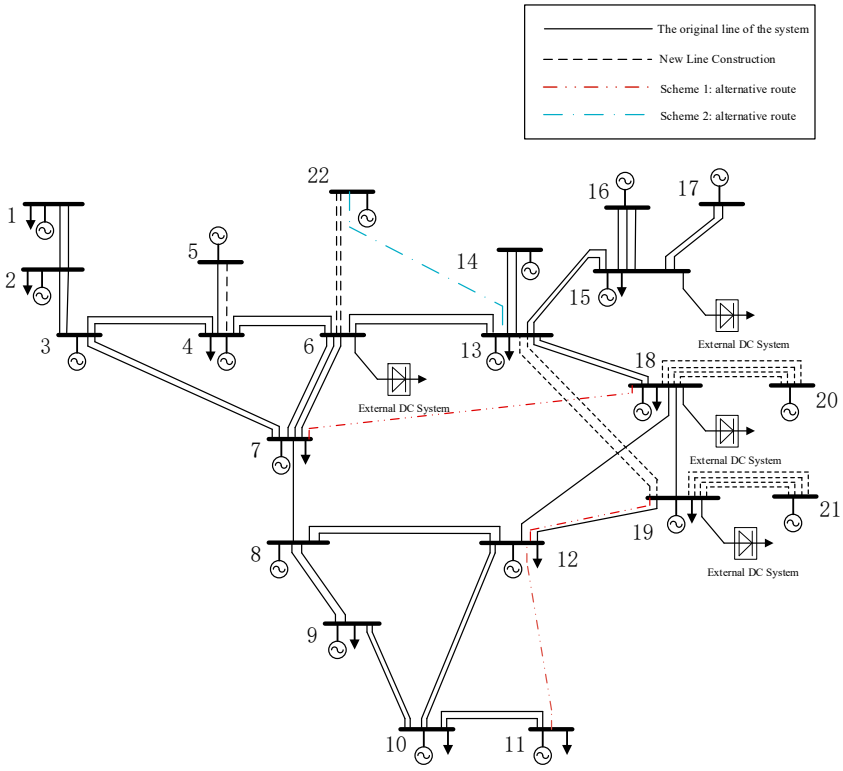


Figure 7. 500 kV network structure for the planned schemes in a southwestern region.

6.3.2. Evaluation of Planning Schemes

To conduct a rational evaluation analysis of the proposed planning scheme, this paper selects the economic indicators, safety and reliability indicators, and AC/DC stability indicators from the established metric set for assessment. These three metrics were chosen because the planning model focuses solely on decision-making for power grid line construction, excluding adaptability metrics relevant to grid development. Consequently, adaptability metrics are not employed in evaluating the planning scheme. This paper has compiled critical evaluation tables from two experts, based on pairwise attribute comparisons across each criterion layer for all indicators. As the evaluation system comprises three criterion layers, each expert provided evaluation tables for all three layers, a total of six tables. Table 12 presents a criticality assessment table for the safety and reliability criteria layer provided by an expert.

Table 12. Importance evaluation table based on safety and reliability.

Indicator	Investment and Construction Costs	Operating Costs	Maintenance Costs	Probability of Power Shortage Duration	Expected Power Shortage	N-1 Pass Rate	Distribution of System Current Flows	Commutation Failure Coefficient	Short-Circuit Ratio
Investment and construction costs	(0.5,0.3)	(0.4,0.45)	(0.5,0.3)	(0.3,0.6)	(0.3,0.6)	(0.2,0.75)	(0.5,0.3)	(0.6,0.25)	(0.4,0.45)
Operating costs	(0.6,0.25)	(0.5,0.3)	(0.6,0.25)	(0.4,0.45)	(0.4,0.45)	(0.3,0.6)	(0.7,0.2)	(0.8,0.15)	(0.8,0.15)
Maintenance costs	(0.5,0.3)	(0.4,0.45)	(0.5,0.3)	(0.3,0.6)	(0.4,0.45)	(0.2,0.75)	(0.6,0.25)	(0.8,0.15)	(0.8,0.15)
Probability of power shortage duration	(0.7,0.2)	(0.6,0.25)	(0.7,0.2)	(0.5,0.3)	(0.6,0.25)	(0.4,0.45)	(0.7,0.2)	(0.6,0.25)	(0.7,0.2)
Expected power shortage	(0.7,0.2)	(0.6,0.25)	(0.6,0.25)	(0.4,0.45)	(0.5,0.3)	(0.4,0.45)	(0.6,0.25)	(0.6,0.25)	(0.7,0.2)
N-1 pass rate	(0.8,0.15)	(0.7,0.2)	(0.8,0.15)	(0.6,0.25)	(0.6,0.25)	(0.5,0.3)	(0.7,0.2)	(0.6,0.25)	(0.7,0.2)
Distribution of system current flows	(0.5,0.3)	(0.3,0.6)	(0.4,0.45)	(0.3,0.6)	(0.4,0.45)	(0.3,0.6)	(0.5,0.3)	(0.5,0.3)	(0.6,0.25)
Commutation failure coefficient	(0.4,0.45)	(0.2,0.75)	(0.2,0.75)	(0.4,0.45)	(0.4,0.45)	(0.4,0.45)	(0.5,0.3)	(0.5,0.3)	(0.5,0.3)
Short-circuit ratio	(0.6,0.25)	(0.2,0.75)	(0.2,0.75)	(0.3,0.6)	(0.3,0.6)	(0.3,0.6)	(0.4,0.45)	(0.5,0.3)	(0.5,0.3)

Upon obtaining the importance evaluation table, the criterion layer weights were calculated according to the Intuitionistic Fuzzy Analytic Hierarchy Process (FAHP), yielding the criterion layer weights as shown in Table 13.

Table 13. Criterion layer weights.

Criterion Layer	Indicator	Weight
Economy	Investment and construction costs	0.1207
	Operating costs	0.1175
	Maintenance costs	0.1095
Safety and reliability	Probability of power shortage duration	0.1031
	Expected power shortage	0.1156
	N-1 pass rate	0.1264
AC/DC Stability	Distribution of system current flows	0.1073
	Commutation failure coefficient	0.1012
	Short-circuit ratio	0.0987

Finally, the two aforementioned schemes were scored, yielding the results presented in Table 14 below.

Table 14. Planning scheme assessment results.

Scheme	Economy Score	Safety and Reliability Score	AC/DC Stability Score	Total Score
1	89.24	88.51	83.86	87.33
2	91.13	86.15	83.01	86.93

From the perspective of AC/DC stability, the proposed indicators confirm that the lines repeatedly selected by the α-Q(λ) planner are indeed critical corridors. In the

500 kV hybrid AC/DC grid, line 7–18 and line 11–12 carry a large proportion of the power exchanged between weak receiving areas and the bulk generation centres. When these lines are built, the commutation-failure margin of the DC systems connected at buses 6 and 18 increases, the effective short-circuit ratio of the key converter stations improves. The eligibility-trace matrix assigns consistently high values to the actions "build 7–18" and "build 11–12", indicating that the RL agent relies on these corridors in many high-quality expansion schemes. This alignment between high eligibility traces and improved AC/DC stability measures supports the physical interpretability of the learned planning policy. Regarding criterion-level scores, scheme 1 demonstrated poorer economic viability than scheme 2, yet superior safety reliability and AC/DC stability, resulting in a higher overall score than scheme 2 This conclusion aligns with the quantitative analysis in the preceding section concerning economic viability, safety reliability, and AC/DC stability. Crucially, it provides a final evaluation score unavailable through quantitative analysis alone, enabling an intuitive comparison of the overall merits of both schemes.

To further demonstrate the accuracy and practicality of the proposed framework, the obtained planning schemes are compared with representative methods in the literature. For the Garver-6 and IEEE 24-node systems, the α-Q(λ)-based planning scheme demonstrates comparable or marginally superior performance to benchmark mathematical programming and heuristic planning results reported in existing literature across investment levels, projected power shortages, and N-1 pass rates. In particular, the comprehensive electrical index obtained in this paper falls within the same range as those studies that use electrical centrality or vulnerability indices to assess grid robustness, confirming that the proposed Gini-based electrical betweenness index provides a physically meaningful measure of flow dispersion and stability risk. For the hybrid AC/DC regional system, the total investment and power shortage values are close to the results of traditional two-stage AC/DC planning models, while the additional AC/DC stability gains highlight the benefit of explicitly embedding commutation-failure margin and ESCR into the multi-objective formulation. Overall, these comparisons indicate that the proposed method yields planning schemes that are compatible with state-of-the-art approaches, while offering enhanced interpretability and stability assurance.

7. Conclusions

This paper has presented a multi-objective hybrid AC/DC transmission expansion planning framework based on an improved multi-step backtracking α-$Q(\lambda)$ reinforcement learning algorithm. A tri-objective model was formulated to minimise annual comprehensive economic cost, expected power shortage and a comprehensive electrical index that reflects AC/DC stability by combining electrical betweenness, commutation-failure margin and ESCR. An intuitionistic fuzzy AHP method and an effectiveness-coefficient-based scoring procedure were adopted to obtain dimensionless evaluation scores and overall plan rankings.

Numerical studies on the Garver-6 and IEEE 24-bus systems showed that the proposed α-$Q(\lambda)$ planner achieves comparable or lower investment and operating costs than classical Q-learning, while improving the N-1 pass rate and reducing the expected power shortage. For the 500 kV regional hybrid AC/DC grid, the preferred expansion scheme reduces the expected annual power shortage from 70,810 MWh to 28,320 MWh, indicating a more uniform flow distribution and enhanced grid robustness. At the same time, the eligibility-trace matrix provides interpretable "line importance" information that highlights key AC/DC corridors such as 7–18 and 11–12, which align well with engineering intuition.

Overall, the results demonstrate that the proposed multi-step backtracking α-$Q(\lambda)$ reinforcement learning framework can effectively solve multi-objective hybrid

AC/DC TNEP and offers planners both high-quality expansion schemes and interpretable diagnostic information.

Despite the above advantages, several issues deserve further investigation. First, the present study considers a static planning horizon and uses a PQ-node equivalence for DC systems, whereas multi-period investment decisions and more detailed AC/DC dynamic interactions may alter the optimal expansion scheme. Second, only one form of eligibility-trace-based reinforcement learning is explored; extending the framework to actor–critic architectures or deep Q-networks could further improve scalability to very large-scale grids. Third, the current Monte Carlo reliability assessment assumes independent outages and simplified renewable profiles. Incorporating correlated failures and high-resolution renewable generation data would allow a more realistic quantification of planning risk. These directions will be pursued in future research to enhance the robustness and applicability of the proposed hybrid AC/DC planning framework.

Author Contributions: Conceptualization, Z.W., W.Y. and Y.Y.; methodology, Z.W. and Z.Z.; software, Y.D.,Y.Y. and Y.H.; validation, Z.W.,Y.D.,Y.Y., Z.Z. and Y.H.; formal analysis, Y.D. and Y.H.; investigation, W.Y., Z.Z. and T.W.; resources, Z.W., Y.D., Y.Y. and Z.Z.; data curation, W.Y., Y.H. and T.W.; writing—original draft preparation, J.L. and T.W.; writing—review and editing, J.L.; visualisation, Y.D., W.Y. and T.W.; supervision, J.L.; project administration, J.L.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the State Grid Shaanxi Electric Power Company Limited Research Institute (SGSNJY00XGJS2500046).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: Authors Zhe Wang, Yuxin Dai, Wenxin Yang, Yunzhang Yang, Zhiqi Zhang and Yahan Hu were employed by the State Grid Shaanxi Electric Power Company Limited Research Institute. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Liu, Y.; Liao, J.; Guo, C.; Tan, Z.; Wang, Y.; Wei, N.; Zhou, N.; Song, Y. Fault severity classification based coordination control strategy of fault current limiter and modular multilevel converter for adaptive fault current limiting. *J. Mod. Power Syst. Clean Energy* **2025**, *13*, 1432–1443. [CrossRef]
2. Liu, Y.; Liao, J.; Guo, C.; Tan, Z.; Wang, Q.; Wang, Y.; Zhou, N. Self-adaptive action and parameter optimization of DC series-parallel power flow controller for fault current limiting in bipolar DC distribution systems. *J. Mod. Power Syst. Clean Energy* **2025**, *13*, 732–746. [CrossRef]
3. Garver, L.L. Transmission network estimation using linear programming. *IEEE Trans. Power Appar. Syst.* **1970**, *PAS-89*, 1688–1697. [CrossRef]
4. Majidi-Qadikolai, M.; Baldick, R. Stochastic transmission capacity expansion planning with special scenario selection for integrating N-1 contingency analysis. *IEEE Trans. Power Syst.* **2016**, *31*, 4901–4912. [CrossRef]
5. Majidi-Qadikolai, M.; Baldick, R. Integration of N-1 contingency analysis with systematic transmission capacity expansion planning: ERCOT case study. *IEEE Trans. Power Syst.* **2016**, *31*, 2234–2245. [CrossRef]
6. Hong, S.; Cheng, H.; Zeng, P.; Zhang, J.; Lu, J. Composite generation and transmission expansion planning with second order conic relaxation of AC power flow. In Proceedings of the IEEE PES Asia-Pacific Power and Energy Engineering Conference, Xi'an, China, 5–28 October 2016; pp. 1688–1693.
7. Choi, J.; El-Keib, A.A.; Tran, T. A fuzzy branch and bound-based transmission system expansion planning for the highest satisfaction level of the decision maker. *IEEE Trans. Power Syst.* **2005**, *20*, 476–484. [CrossRef]
8. Wang, S.; Geng, G.; Jiang, Q. Robust Co-Planning of Energy Storage and Transmission Line with Mixed Integer Recourse. *IEEE Trans. Power Syst.* **2019**, *34*, 4728–4738. [CrossRef]
9. Yichen, L.; Bo, L.; Chenqian, Z.; Teng, M. Intelligent Frequency Assignment Algorithm Based on Hybrid Genetic Algorithm. In Proceedings of the 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL), Chongqing, China, 10–12 July 2020; pp. 461–467.

10. Karimi, E.; Ebrahimi, A. Inclusion of blackouts risk in probabilistic transmission expansion planning by a multi-objective framework. *IEEE Trans. Power Syst.* **2015**, *30*, 2810–2817. [CrossRef]
11. Zhao, B. Optical Cable Planning of Power Optical Transmission Network based on Genetic Algorithm. In Proceedings of the 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES), Chickballapur, India, 28–29 December 2022; pp. 1–5.
12. Wang, H.; Li, P. Optimization Model of Information System Network Structure Based on Artificial Fish Swarm Algorithm. In Proceedings of the 2023 International Conference on Telecommunications, Electronics and Informatics (ICTEI), Lisbon, Portugal, 11–13 September 2023; pp. 591–595.
13. Chen, H.; Zhuang, K.; Wang, Z.; Xue, J.; Li, Y. A Transmission Network Planning Method Based on Particle Swarm Optimization. In Proceedings of the 2024 3rd International Conference on Energy, Power and Electrical Technology (ICEPET), Chengdu, China, 17–19 May 2024; pp. 768–771.
14. Gülşen Erdinç, F.; Çiçek, A.; Erdinç, O.; Yumurtacı, R.; Oskouei, M.Z.; Mohammadi-Ivatloo, B. Decision-making framework for power system with RES including responsive demand, ESSs, EV aggregator and dynamic line rating as multiple flexibility resources. *Electr. Power Syst. Res.* **2022**, *204*, 107702. [CrossRef]
15. Rosemberg, A.; Tanneau, M.; Fanzeres, B.; Garcia, J.; Van Hentenryck, P. Learning Optimal Power Flow value functions with input-convex neural networks. *Electr. Power Syst. Res.* **2024**, *235*, 110643. [CrossRef]
16. Rust, J. Structural estimation of markov decision processes. *Handb. Econom.* **1994**, *4*, 3081–3143.
17. Hadidi, R.; Jeyasurya, B. Reinforcement learning based real-time wide-area stabilizing control agents to enhance power system stability. *IEEE Trans. Smart Grid* **2013**, *4*, 489–497. [CrossRef]
18. Duan, J.; Xu, H.; Liu, W. Q-learning-based damping control of wide-area power systems under cyber uncertainties. *IEEE Trans. Smart Grid* **2018**, *9*, 6408–6418. [CrossRef]
19. Sun, Q.; Wang, D.; Ma, D.; Huang, B. Multiobjective energy management for we-energy in Energy Internet using reinforcement learning. In Proceedings of the IEEE Symposium Series on Computational Intelligence, Honolulu, HI, USA, 27 November–1 December 2017; pp. 1630–1635.
20. Lamini, C.; Fathi, Y.; Benhlima, S. H-MAS architecture and reinforcement learning method for autonomous robot path planning. In Proceedings of the 2017 Intelligent Systems and Computer Vision, Fez, Morocco, 17–19 April 2017; pp. 1–7.
21. Yu, T.; Wang, H.Z.; Zhou, B.; Chan, K.W.; Tang, J. Multi-Agent Correlated Equilibrium Q(λ) Learning for Coordinated Smart Generation Control of Interconnected Power Grids. *IEEE Trans. Power Syst.* **2015**, *30*, 1669–1679. [CrossRef]
22. Tomasson, E.; Soder, L. Improved importance sampling for reliability evaluation of composite power systems. *IEEE Trans. Power Syst.* **2017**, *32*, 2426–2433. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.