ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

# Research on dynamic regulation strategy optimization of supply chain greenwashing behavior based on reinforcement learning

*Qintao Peng*[1,2,*] *and Fan Chen*[3]

[1] College of Economics and Management, China Three Gorges University, Yichang 443002, Hubei, China
[2] College of Economics and Management, Jingchu University of Technology, Jingmen 448000, Hubei, China
[3] School of Artificial Intelligence, Jingchu University of Technology, Jingmen 448000, Hubei, China

**SUMMARY:** *In order to solve the problems of static identification lag, insufficient matching of regulatory actions and insufficient utilization of feedback in the regulation of supply chain greenwashing behavior, this paper constructs a dynamic regulation strategy optimization model based on reinforcement learning. The model takes the consistency of green declaration, performance deviation, certification change, text anomaly and historical feedback as the status input, sets up supervision actions such as prompt description, data review, key spot check, credit constraint and continuous tracking, and comprehensively restricts risk reduction, resource consumption and misjudgment loss through the reward function. The experiment was carried out based on 1260 supply chain subjects, 85420 structured records and 18670 text disclosure samples. The model was trained for 500 rounds, and compared with Logistic regression, SVM, random forest, XGBoost and static DQN. The results show that the Accuracy of the model in this paper reaches 93.6%, Macro-F1 reaches 91.8%, the high-risk recall rate reaches 92.4%, the invalid resource consumption rate is reduced to 13.8%, and the average response cycle is shortened to 2.4 working days. The research results show that the proposed model can improve the identification accuracy of greenwashing risk and the adaptation ability of dynamic supervision actions, and provide a computable optimization path for the intelligent supervision of supply chain greenwashing behavior.*

***KEYWORDS:*** *Reinforcement learning; Supply chain management; Greenwashing behavior; Dynamic supervision strategy*

# 1 Introduction

In the process of green transformation of supply chain, some subjects may create a green image through selective disclosure, fuzzy expression, and exaggerated emission reduction effectiveness, while the actual production, procurement, transportation or recycling links have not been improved. This kind of greenwashing behavior has the characteristics of concealment, delay and cross-link transmission. If it still relies on static rules or post-hoc verification methods to identify, it is prone to problems such as response lag, mismatched punishment intensity and uneven allocation of regulatory resources. Especially in the multi-level supply chain scenario, the sources of green declarations, transaction records,

*pengfufu228@163.com
https://doi.org/10.65102/is20261054
Published: 10 June, 2026

certification information, public opinion texts and performance data are scattered, and a single indicator is difficult to accurately reflect the real behavior state of the subject. The optimization of dynamic and intelligent regulatory strategies has become an urgent problem to be solved.

With the development of artificial intelligence and data mining technology, machine learning has been used for tasks such as greenwashing identification, text review and risk warning. There have been many discussions on greenwashing measurement, sustainable report recognition, supply chain management risk and artificial intelligence detection methods [1], but most methods still focus on static classification or rule judgment, and lack of description of the continuous relationship between regulatory actions and behavioral feedback. At the same time, reinforcement learning has shown strong adaptive ability in inventory control, ordering decision and multi-level supply chain optimization [2]. Through the cycle mechanism of "state-action-reward-policy update", it can continuously revise the decision-making scheme in the uncertain environment. The introduction of reinforcement learning into supply chain greenwashing supervision helps to transform the behavior recognition results into dynamic supervision actions, so that the model no longer stops at "judging whether there is a risk", and further answers the questions of "when to intervene, what intensity to adopt, and how to adjust the follow-up strategy" [3].

This paper constructs a dynamic regulation strategy optimization model of supply chain greenwashing behavior based on reinforcement learning. In this study, the consistency of green declaration, performance deviation, historical violation record, third-party certification change and text anomaly characteristics of the supply chain subject are used as the status input, and the prompt explanation, data review, key random check, credit constraint and continuous tracking are set as regulatory actions. The reward function is constructed by risk reduction, resource consumption, misjudgment loss and behavior correction effect. The experimental goals are set as follows: the accuracy of high risk identification of greenwashing reaches more than 90%, the ineffective consumption of regulatory resources reduces more than 20%, the missed detection rate of high-risk subjects is controlled within 8%, and the average response period after dynamic adjustment is shortened by more than 25%. Through this model, this paper hopes to form a closed-loop framework from data perception, state calculation, action selection to policy iteration, and provide a computable, traceable and updatable technical path for green behavior governance of supply chain.

*Table 1: Related research basis and improvement direction of this paper*

| Research direction | Representative literature | Main methods | Existing limitations | Implications for this study |
|---|---|---|---|---|
| Greenwashing behavior identification | [10–16] | Measuring greenwashing based on reports, statements, and entity characteristics | Mostly static identification, lacking a continuous feedback mechanism | Identification results need to be transformed into executable regulatory actions |
| Artificial intelligence detection | [11][16] | Using text mining and machine learning to identify abnormal green claims | Emphasis is placed on classification results, with insufficient strategy optimization capability | Multi-source data features can be introduced to improve the accuracy of state representation |
| Supply chain green management | [12][13][14] | Analyzing the relationships among greenwashing, supply chain collaboration, and marketing | Insufficient characterization of multi-link risk transmission | Risk diffusion and behavioral linkage among nodes need to be considered |
| Reinforcement learning optimization | [4–9][17–22] | Optimizing supply chain decisions through DQN, policy gradient, and other methods | Mostly focused on inventory, ordering, and allocation problems | Its dynamic decision-making logic can be used to construct a regulatory strategy update model |

It can be seen from Table 1 that the existing research has laid a foundation for the application of greenwashing recognition and reinforcement learning, but there is still a lack of systematic connection between the two for regulatory policy optimization. The innovation points of this paper are mainly reflected in three aspects. Firstly, the greenwashing behavior of supply chain is transformed into a computable state space, and the ability of the model to process multi-source heterogeneous information is enhanced. Secondly, the regulatory actions are designed as an upgradable set of strategies, so that the regulatory intensity can be adjusted automatically with the change of risk. Thirdly, the reward function is used to simultaneously constrain the risk control effect and the resource input level, so as to improve the stability and interpretability of the dynamic regulatory process.

# 2 Literature Review

## 2.1 Application Research of Reinforcement learning in supply chain behavior regulation

In recent years, reinforcement learning has been gradually extended from specific tasks such as inventory ordering and resource allocation to multi-agent collaborative decision-making scenarios in supply chain behavior regulation. Its core advantage is that it can continuously revise its strategy according to environmental feedback. Boute et al. introduced deep reinforcement learning into the study of inventory control and pointed out that this kind of method can form an adaptive control ability under the condition of uncertain demand and multi-period decision-making [4]. De Moor et al. further improved the inventory management strategy of perishable products from the perspective of reward shaping, so that the model could obtain more stable learning signals at the early stage of training [5]. Rolf et al. systematically reviewed the application of reinforcement learning in supply chain management and believed that it was suitable for dealing with dynamic feedback problems that were difficult to be covered by traditional static models [6].

In the multi-level supply chain scenario, reinforcement learning model not only needs to judge the optimal action of a single node, but also needs to take into account the linkage reaction between upstream and downstream agents. Hammler et al. proposed a dynamic replenishment strategy based on deep reinforcement learning and showed that the model could adjust the ordering behavior under the condition of demand fluctuation [7]. Dehaybe et al. focused on the inventory optimization problem under non-stationary uncertain demand, indicating that reinforcement learning still has certain adaptive ability when the environmental distribution changes [8]. Geevers et al. applied deep reinforcement learning to multi-level inventory optimization, which further proved its application potential in complex chain structure [9]. However, the existing research mainly focuses on inventory, ordering and allocation decisions, and rarely involves the problems of behavior deviation, statement distortion and regulatory feedback of supply chain agents. The state design and reward constraints at the level of behavior regulation still need to be further expanded.

## 2.2 Research on identification and dynamic regulation of supply chain greenwashing behavior

Supply chain greenwashing usually manifests as the deviation between the green declaration and the actual performance. The difficulty of identifying greenwashing is that the information source is scattered, the expression way is hidden, and the consequences of the behavior are delayed. Ruiz-Blanco et al analyzed the formation factors of green bleaching from the

perspective of enterprise characteristics, and pointed out that organizational attributes, disclosure tendency and external attention degree would affect the authenticity of green declarations [10]. Through a systematic review, Moodaley and Telukdarie found that artificial intelligence methods have begun to be used in sustainability report review and greenwashing identification, but relevant studies are still inadequate in data annotation and interpretation mechanism [11]. Ines et al. discussed greenwashing from the perspective of supply chain management and believed that information asymmetry in the process of multi-link collaboration would amplify greenwashing risks [12].

The existing research on greenwashing identification has gradually shifted from concept discrimination to measurement method construction. Vangeli et al. pointed out in their research on green B2B marketing that green narratives among supply chain subjects may affect procurement, cooperation and trust relationships [13]. Santos et al. sorted out the relationship between greenwashing and stakeholder reactions and emphasized that the recognition model should pay attention to both external perception and behavioral evidence [14]. Bernini et al. summarized green bleaching measurement methods and proposed that different measurement frameworks still had differences in index selection, data sources and comparability [15]. Lagasio further proposed the idea of ESG-washing detection for sustainable reports, indicating that text anomaly, disclosure gap and index mismatch can be used as an important basis for model identification [16]. It can be seen that the research on green bleaching has a rich recognition basis, but there are still shortcomings in the dynamic supervision level, especially the lack of a computational framework to transform the recognition results into continuous supervision actions.

## 2.3 The Integration Research of reinforcement learning and the optimization of the supervision strategy of greenwashing behavior

Applying reinforcement learning to supply chain greenwashing supervision, the key is to transform the greenwashing risk identification problem into a sequential decision-making problem that can be interactive, feedback and update. Mohamadi et al. combined deep reinforcement learning with the supply chain inventory mechanism and showed that the model could form a better decision-making scheme under multi-objective constraints [17]. Stranieri et al. combined deep reinforcement learning with multi-stage stochastic programming to deal with uncertainty in complex supply chain decision-making, providing a method reference for risk state transfer in greenwashing regulation [18]. Stranieri et al. further compared the performance of different deep reinforcement learning algorithms in the two-level supply chain control system, suggesting that the choice of algorithm will directly affect the stability of strategy [19].

In the greenwashing supervision scenario, the state space can be composed of green declaration consistency, performance deviation, certification change, complaint record and text anomaly degree. Action space can include prompt rectification, data review, key spot check, credit constraint and continuous tracking; The reward function should reflect risk reduction, resource consumption and misjudgment cost simultaneously. Nahhas et al. introduced the pictorial observation space into the supply chain allocation problem, showing that reinforcement learning can deal with non-traditional structured state input [20]. Yavuz and Kaya used deep reinforcement learning in dynamic pricing and inventory management to prove that the model can adjust the strategy in time according to environmental fluctuations [21]. Kurian et al. proposed a multi-level supply chain ordering mechanism based on deep reinforcement learning, which provided reference for the design of multi-node and multi-action supervision model [22].

*Table 2: Related research topics, main methods and the entry point of this paper*

| Research topic | Representative literature | Main methods | Existing limitations | Focus of this study |
|---|---|---|---|---|
| Reinforcement learning for supply chain regulation | [4][6][9] | DQN, policy gradient, and multi-echelon inventory optimization | Emphasis is mainly placed on inventory and ordering, while behavioral regulation is insufficiently discussed | Extending reinforcement learning to the dynamic regulation of greenwashing behavior |
| Greenwashing behavior identification | [10][11][15][16] | Report review, text recognition, and indicator measurement | Mostly based on static identification, with limited strategy feedback | Constructing an updatable risk state space |
| Supply chain greenwashing analysis | [12][13][14] | Supply chain collaboration analysis and entity relationship research | Insufficient characterization of cross-link risk transmission | Incorporating multi-node behavioral deviation and linkage feedback |
| Integrated applications of reinforcement learning | [18][20][22] | Integration with stochastic programming, observation-space expansion, and multi-level policy learning | Limited integration with greenwashing regulation | Designing a linkage mechanism between regulatory actions and reward functions |

In general, reinforcement learning research emphasizes dynamic decision-making, greenwashing research emphasizes risk identification, and there is still a lack of deep integration for regulatory strategy optimization between the two types of research. In order to more clearly present the technical route, main shortcomings and the entry point of this paper, this paper summarizes and collates the existing research, as shown in Table 2. On this basis, this paper constructs a closed-loop model of "state recognition-action selection-reward feedback-strategy update", which makes the greenwashing supervision turn from static discrimination to dynamic optimization. The adaptability of the model under risk control, resource utilization and environmental changes is verified by experimental comparison.

# 3 Research Methods

## 3.1 Theoretical Foundations of supervision models for reinforcement learning

The supervision process of supply chain greenwashing is not a single judgment problem, but a dynamic decision-making problem with the characteristics of continuous state change and feedback correction. In the process of supply chain operation, green statements, performance records, certification changes, transaction vouchers, text disclosure and historical feedback

will be constantly updated, and regulatory actions will also have an impact on subsequent behavior. If only static scoring or one-time classification methods are used, the model can only judge whether there is a high risk in the current subject, but it is difficult to answer whether the risk decreases after the intervention of regulatory actions, whether the behavior is modified, and whether the subsequent strategy needs to be adjusted. Therefore, this paper abstracts the regulation process of supply chain greenwashing behavior as a Markov decision process, and introduces the reinforcement learning method to establish a dynamic regulation strategy optimization model.

In order to make the model structure clearer, this paper first divides the greenwashing supervision process into five parts: data input, risk state identification, supervision action selection, feedback reward calculation and strategy update, and its theoretical framework is shown in Figure 1. This framework emphasizes that the regulatory model is not a one-way recognition process from data to results, but re-receives feedback after each action execution, and uses the feedback results for the next round of policy adjustment.


*Figure 1: Theoretical framework of dynamic regulation of supply chain greenwashing behavior based on reinforcement learning*

Around the risk state identification layer, the supply chain greenwashing risk state at the supervision time t is denoted as $s_t$ in this paper. This state is not a single indicator, but is composed of green declaration consistency, performance deviation, certification change, text anomaly and historical feedback, which can be expressed as follows.

$$s_t = \{c_t, d_t, v_t, e_t, h_t\} \tag{1}$$

where, $c_t$ represents the consistency level between green declaration and actual performance

information; $d_t$ represents the performance deviation of supply chain nodes in procurement, production, transportation or recycling; $v_t$ represents the changes in certification, testing or third-party records; $e_t$ represents the abnormal intensity of exaggerated, vague or evading expression in text disclosure; $h_t$ represents the degree of historical regulatory feedback and risk accumulation. Through this state representation, the model can map structured records and unstructured texts into a computable space, which provides a basis for action selection. In the supervision action selection layer, the model selects the corresponding supervision action according to the current state $s_t$. In this paper, the action space is set as follows.

$$A = \{a^1, a^2, a^3, a^4, a^5\} \tag{2}$$

Among them, $a^1$ represents low intensity prompt, $a^2$ represents data review, $a^3$ represents focus spot check, $a^4$ represents credit constraint, and $a^5$ represents continuous tracking. Different actions correspond to different resource input and risk control intensity. The model does not presuppose that a certain risk level necessarily corresponds to a fixed action, but judges which action is more effective in a similar state through historical feedback, so as to improve the adaptability of the regulatory strategy.

The feedback reward calculation layer is the key to distinguish reinforcement learning models from static recognition models. After the execution of the supervision action, the system calculates the immediate reward based on the risk reduction, resource consumption and misjudgment loss. Let the reward after performing the action $a_t$ be $r_t$, which is expressed as follows.

$$r_t = \lambda_1 \Delta R_t - \lambda_2 C_t - \lambda_3 M_t \tag{3}$$

where, $\Delta R_t$ represents the decrease of greenwashing risk, $C_t$ represents the consumption of regulatory resources, $M_t$ represents the loss caused by misjudgment or excessive intervention, and $\lambda_1$、$\lambda_2$、$\lambda_3$ are the weight coefficients. The reward function makes the model reduce the risk and constrain the supervision cost at the same time, avoiding the frequent use of high-intensity actions and preventing the lack of supervision of low-intensity actions in high-risk states.

In the policy update layer, the model continuously modifies the subsequent action selection according to the reward feedback. Let the policy function be $\pi(a_t|s_t)$, which represents the probability of selecting action $a_t$ in state $s_t$, then the dynamic supervision objective can be expressed as follows.

$$\pi^* = \arg\max_{\pi} E\left[\sum_{t=0}^{T} \gamma^t r_t\right] \tag{4}$$

where, $\gamma$ is the discount factor used to measure the degree of influence of future rewards in the current decision, and T is the length of the supervision period. This objective function shows that the model does not pursue the local revenue of a certain round of regulatory actions, but maximizes the overall revenue in consecutive regulatory cycles. When a certain type of action can bring significant risk reduction at a lower cost, its selection probability in similar states will be improved. When a certain type of action has large consumption and limited risk improvement, the model will reduce its subsequent adoption frequency.

## 3.2 State identification and model calculation process of supply chain greenwashing behavior

In the reinforcement learning supervision model, state recognition is the pre-step of policy optimization. Supply chain greenwashing does not directly appear as a single outlier, but is hidden in the differential relationship between green claims, contract performance, detection records, certification changes, textual disclosure and historical feedback. Therefore, in the process of model calculation, a multi-source feature extraction module is constructed to convert different types of data into a unified state representation, and then input into the reinforcement learning model to complete the subsequent action selection. In order to avoid the state recognition process staying at the static scoring level, this paper considers each round of supervision cycle as a continuous calculation window, so that the model can update the risk state according to the newly incoming data. Its computational structure is shown in Figure 2.


*Figure 2: State identification and model calculation structure of supply chain greenwashing behavior*

As can be seen from Figure 2, the state recognition does not directly feed the original data into the model, but is jointly processed through three paths: structural features, change signals and semantic features. For the ith supply chain agent, the original observation vector extracted within the supervision period t can be expressed as follows.

$$x_{i,t} = [g_{i,t}, p_{i,t}, q_{i,t}, u_{i,t}, z_{i,t}] \tag{5}$$

where, $g_{i,t}$ represent the characteristics of green statements, mainly including green commitment, environmental protection description and disclosure frequency; $p_{i,t}$ represents the performance record characteristics, reflecting the implementation deviation of procurement, production, transportation and recycling links; $q_{i,t}$ denotes the authentication

and detection change signal; $u_{i,t}$ represents external feedback and historical regulatory records; $z_{i,t}$ denotes the semantic anomaly features in the text disclosure. This vector provides the base input for the subsequent risk state calculation.

Due to the differences in dimensions and value ranges of different data dimensions, continuous features are standardized in this paper. Let the original value of the KTH feature be $\tilde{x}_{i,t}^{(k)}$, and the normalized result is as follows.

$$\tilde{x}_{i,t}^{(k)} = \frac{x_{i,t}^{(k)} - \min(x^{(k)})}{\max(x^{(k)}) - \min(x^{(k)}) + \varepsilon} \tag{6}$$

where, $\varepsilon$ is the smoothing term that prevents the denominator from being zero. After normalization, features from different sources are mapped to similar numerical intervals, which reduces the interference of high-dimensional variables on model training and makes the state recognition process more stable.

For text disclosure data, this paper uses a pre-trained language model to extract semantic vectors, which are used to capture implicit features such as exaggerated expressions, ambiguous descriptions, and inconsistent statements. Let the set of texts be $T_{i,t}$. The semantic encoding result can be expressed as follows.

$$z_{i,t} = \text{Pool}(\text{Encoder}(T_{i,t})) \tag{7}$$

where Encoder(·) represents the text encoder and Pool(·) represents the pooling operation, which is used to compress the sentence-level or paragraphlevel semantic representation into a fixed-length vector. Compared with the simple word frequency statistics, this processing method can preserve the context relationship, making it easier for the model to identify the greenwashing risk signal of "high declaration strength but insufficient performance evidence".

After extracting structural features and text features, this paper uses weighted fusion method to generate greenwashing risk state vector. Let the fused state input be $S_{i,t}$, which is calculated as follows.

$$S_{i,t} = \omega_1 \tilde{g}_{i,t} + \omega_2 \tilde{p}_{i,t} + \omega_3 \tilde{q}_{i,t} + \omega_4 \tilde{u}_{i,t} + \omega_5 z_{i,t} \tag{8}$$

where, $\omega_1$ to $\omega_5$ are feature weights, which are used to control the degree of influence of different source information on state recognition. If a subject frequently appears high-intensity commitment in the green statement, but the performance record and certification change do not form the corresponding support, the model will form a high risk representation in the fusion state. If the text disclosure is consistent with the actual record and the historical feedback is stable, the risk status decreases accordingly.

During model computation, the state vector $S_{i,t}$ is fed into the reinforcement learning environment as a direct basis for action selection. During training, the system runs in the order of "observation window update - state vector generation - action execution - feedback recording". After each supervision cycle, the model re-writes the new performance data and feedback results into the observation window, and updates the state for the next round. In order to improve the computational efficiency, this paper sets a fixed-length sliding window to split the continuous regulatory records into several time slices, which not only retains the recent behavior changes, but also avoids the model hysteresis caused by too long historical information.

## 3.3 Design of supervisory action, reward function and policy update mechanism

After the state identification of greenwashing risk is completed, the model needs to transform the state results into executable regulatory actions, and continuously revise the subsequent strategy through the feedback results. Different from the simple risk identification model, the reinforcement learning supervision model emphasizes the collaborative operation between each component: the state recognition module is responsible for judging the current risk level, the action selection module is responsible for determining the supervision intervention mode, the reward feedback module is responsible for evaluating the effect of the action, and the policy update module adjusts the probability of the next round of action according to the feedback result. In order to avoid the disconnection between regulatory actions and risk status, this paper organizes the above modules into an action-feed-update linkage structure, whose synergistic relationship is shown in Figure 3.

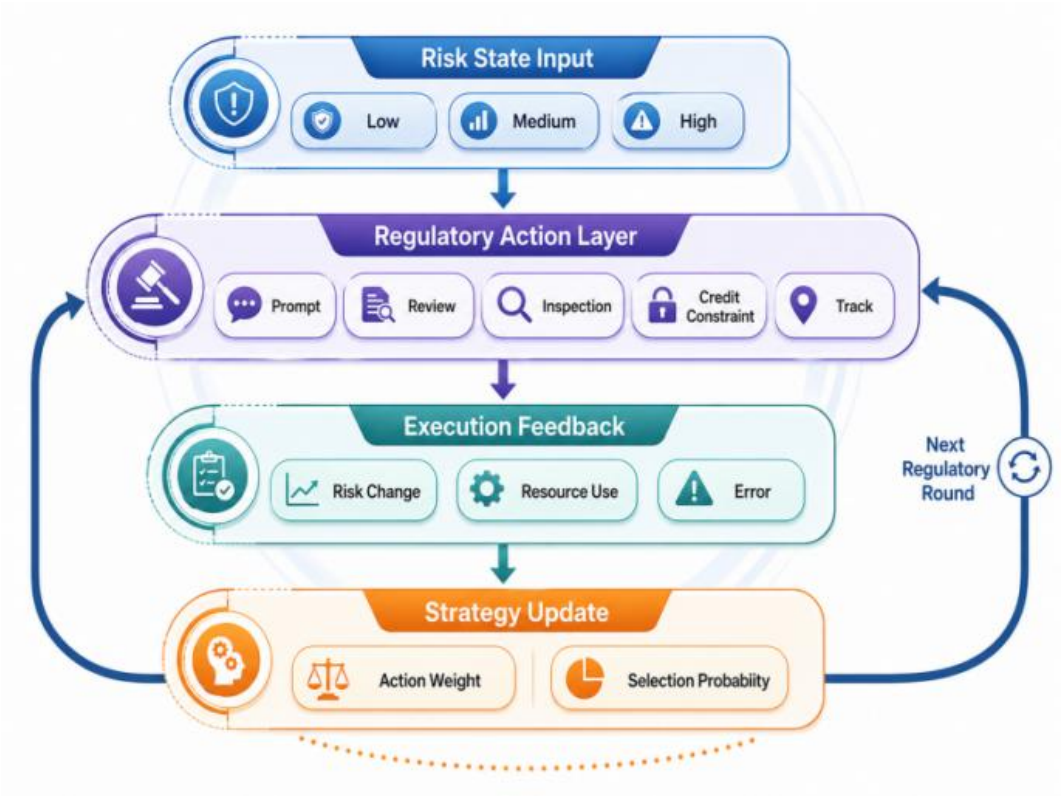


*Figure 3: Coordination structure of supervision action, reward feedback and policy update*

In terms of regulatory action design, this paper does not adopt a single penalty or fixed verification method, but sets a multi-level action space according to risk intensity and behavior change. Let the set of optional actions in round t of supervision be $\mathcal{A}_t$, whose expression is as follows.

$$\mathcal{A}_t = \{P_t, R_t, I_t, C_t, T_t\} \tag{9}$$

where, $P_t$ stands for prompt explanation, which is applicable to subjects with low risk but slight inconsistency in disclosure expression. $R_t$ stands for data review, applicable to the situation where there is a partial deviation between the declaration content and the performance record; $I_t$ stands for focus spot check, which is suitable for the subject of

continuous accumulation of risk signals or cross-anomaly of multi-source data; $C_t$ stands for credit constraint, which is suitable for high-risk subjects that have not been corrected after multiple rounds of feedback. $T_t$ stands for continuous tracking and is used when the risk state is not yet stable and we need to observe subsequent changes in behavior. Through action classification, the model can select the intervention mode of matching intensity under different risk states, reduce the excessive investment of low risk subjects, and avoid the risk accumulation of high risk subjects due to insufficient actions.

In the process of action selection, the model needs to maintain a balance between "following known effective actions" and "trying new actions". In this paper, the probability mechanism of action selection based on risk score is adopted, and the state vector $S_{i,t}$ is combined with the action weight matrix $\Theta$ to obtain the selection probability of the JTH type of action:

$$\Pr(a_j|S_{i,t}) = \frac{\exp(\Theta_j S_{i,t})}{\sum_{m=1}^{5} \exp(\Theta_m S_{i,t})} \tag{10}$$

where, $\Theta_j$ represents the response weights of action $a_j$ to different risk features. If a subject has a high deviation between text disclosure and performance record, the model will improve the selection probability of data review or key spot check action. If the risk of the subject is low and the historical feedback is stable, the probability of low-intensity prompting or continuous tracking will be increased accordingly. This mechanism makes the regulatory action no longer rely on manual experience fixed matching, but is determined by state characteristics and historical feedback.

The reward function design needs to reflect the actual effect of supervision actions. In this paper, the feedback results are divided into risk correction scores, resource usage scores and misjudgment penalty scores, and a comprehensive feedback value is formed:

$$F_t = \mu_1 B_t + \mu_2 D_t - \mu_3 E_t \tag{11}$$

where, $B_t$ represents the behavior modification benefit, which mainly reflects the results of improving the consistency of green statements, decreasing the performance deviation and reducing the abnormal text. $D_t$ represents the regulatory efficiency score, which reflects the degree of risk improvement obtained with limited resource input. $E_t$ stands for misjudgment penalty, which is mainly used to constrain unnecessary high-intensity supervision actions. $\mu_1$、$\mu_2$、$\mu_3$ are the feedback weights. The design enables the model to pay attention to the risk control effect and action appropriateness at the same time, avoiding the excessive occupation of resources caused by the pure pursuit of high-intensity intervention. The policy update mechanism is used to transform the feedback results into the basis for the next round of action selection. Let the update amount of the action weight after round t be $\Delta\Theta_t$, which is calculated as follows.

$$\Delta\Theta_t = \eta F_t \nabla_\Theta \log \Pr(a_t|S_{i,t}) \tag{12}$$

where $\eta$ is the update step and $\nabla_\Theta \log \Pr(a_t|S_{i,t})$ represents the gradient of the current action selection probability with respect to the weight parameter. When an action brings a higher comprehensive feedback value, its corresponding weight will be enhanced, and it is easier to be selected again in similar states. When the action feedback is poor, the model will weaken the action weight and reduce its subsequent use frequency. Through multiple iterations, the supervision model can gradually form a more stable action preference.

## 3.4 Application process of dynamic regulatory strategy optimization model

After the state recognition, supervision action design and strategy update mechanism construction, this paper further applies the reinforcement learning model to the dynamic supervision process of supply chain greenwashing behavior. The core goal of the process is not to make a one-time risk judgment for a subject, but to continuously revise the supervision strategy according to the multi-source data changes, action execution results and risk feedback in a continuous supervision cycle. Specifically, the model first receives data such as green declaration, performance record, certification change, text disclosure and historical feedback, and generates greenbleaching risk states through the state recognition module. Then, the reinforcement learning module outputs regulatory actions, and re-updates the policy according to the risk changes after the execution of the actions.

In order to make the model application process operable, this paper divides the dynamic supervision policy optimization process into five links: data access, state calculation, action generation, feedback evaluation and policy iteration, as shown in Figure 4. The process emphasizes the continuous transmission relationship between different modules, and the risk state is not the end point, but the input of supervision action selection. Supervision actions are also not static results, but will negatively affect subsequent policies through feedback results.

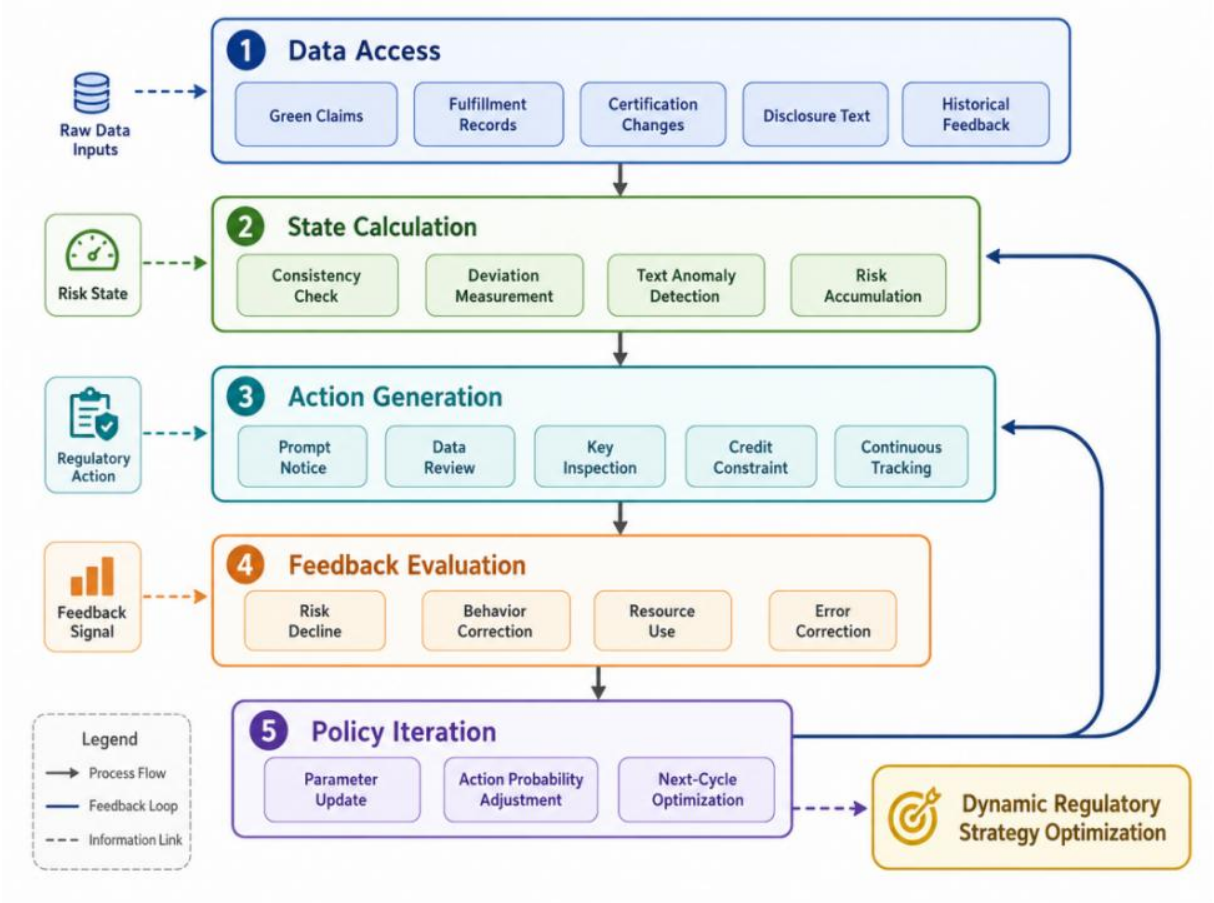


*Figure 4: Application process of dynamic regulatory strategy optimization model*

In the process shown in Figure 4, the data access link is responsible for integrating the records formed by the supply chain principals in different business nodes. Since greenwashing often manifests as an inconsistent state of "strong statement, weak evidence" or "more

disclosure, less performance", the model needs to deal with both structured data and text data in the state calculation stage. Let the comprehensive greenwashing risk score of the ith subject in the TTH supervision period be $G_{i,t}$, which can be expressed as follows.

$$G_{i,t} = \rho_1 L_{i,t} + \rho_2 D_{i,t} + \rho_3 N_{i,t} + \rho_4 H_{i,t} \tag{13}$$

where, $L_{i,t}$ represents the degree of inconsistency between the green statement and the actual evidence, $D_{i,t}$ represents the level of performance deviation, $N_{i,t}$ represents the intensity of text anomaly, $H_{i,t}$ represents the cumulative risk of historical feedback, and $\rho_1$ to $\rho_4$ are weight parameters. The higher this score is, the more prominent the greenwashing risk of the subject is in the current cycle, and the model will tend to choose the supervision action with higher intensity in the action generation stage.

In the action generation step, the reinforcement learning module selects the corresponding strategy according to the current risk score and state vector. If $G_{i,t}$ is in the low risk interval, the model can output prompt instructions or continuous tracking. If the risk is at a moderate level, the data review will be preferred to verify whether the green statement is consistent with the actual performance evidence; If the risk continues to increase, the model will increase the selection probability of the key spot check or credit constraint action. Unlike the fixed rule, this procedure allows the model to dynamically adjust the action intensity according to the feedback results, so that the action selection in similar risk states gradually tends to be stable.

The feedback evaluation link is used to judge whether the supervision action is effective. Let the risk change rate after action execution be $\Delta G_{i,t}$, which is calculated as follows.

$$\Delta G_{i,t} = \frac{G_{i,t} - G_{i,t+1}}{G_{i,t} + \delta} \tag{14}$$

where $G_{i,t+1}$ represents the risk score in the next regulatory period and δ is the smoothing term used to avoid the denominator being zero. When $\Delta G_{i,t}$ is large, it means that the current regulatory action has a good effect on the risk pressure drop. When the value is small or even negative, it indicates that the action fails to effectively promote behavior modification, and the model needs to adjust the action combination or improve the supervision intensity in subsequent cycles.

In the specific training process, this paper sets the supervision cycle as 12 rounds, and each round corresponds to a data update and strategy adjustment. The number of training rounds of the reinforcement learning model is set to 500, the learning rate is set to 0.001, the discount factor is set to 0.90, and the initial exploration rate is set to 0.20, which is gradually attenuated to 0.02 during the training process to give consideration to both strategy exploration and stable utilization. The running environment of the model uses Python 3.11 and PyTorch 2.2, and the hardware configuration is Intel Core i7-12700 processor, 32 GB memory and NVIDIA RTX 4080 GPU. The structural features are standardized and input into the state calculation module, and the text disclosure data is encoded by the pre-trained language model to participate in risk fusion, so as to ensure that the model can identify numerical deviations and semantic anomalies at the same time.

# 4 Experimental Evaluation

## 4.1 Experimental Design

In order to verify the effectiveness of the dynamic regulation strategy optimization model of supply chain greenwashing behavior based on reinforcement learning, this paper designs an experimental evaluation link to comprehensively test the identification effect of greenwashing risk, the matching degree of regulatory actions, the efficiency of resource input and the adaptability of dynamic environment. Different from the static classification model, the proposed model not only needs to determine whether the subject has greenwashing risk, but also needs to generate appropriate regulatory actions under different risk states, and continuously adjust the strategy according to the feedback results. Therefore, the experiment focuses on the model's ability to identify high-risk subjects, the stability of regulatory action selection, and the effect of strategy optimization after multiple rounds of feedback.

The experimental data comes from the dataset constructed by simulating the green disclosure and performance supervision of the supply chain, which includes green declaration text, contract performance records, certification change information, detection records, complaint feedback and historical supervision results. The dataset contains a total of 1260 supply chain subjects, 85420 structured records and 18670 text disclosure samples over 12 regulatory periods. In order to ensure the continuity of the experiment, this paper divides the data in chronological order, where the first 8 cycles are used for model training, the 9 to 10 cycles are used for parameter validation, and the 11 to 12 cycles are used for test evaluation. The data division and experimental index Settings are shown in Table 3.

*Table 3: Experimental data division and evaluation index Settings*

| Item | Setting content | Data or parameter |
|---|---|---|
| Number of supply chain entities | Node samples participating in the experiment | 1,260 |
| Structured records | Records of fulfillment, certification, testing, feedback, etc. | 85,420 |
| Text disclosure samples | Green claims, explanatory texts, and feedback texts | 18,670 |
| Regulation cycles | Continuous dynamic evaluation cycles | 12 rounds |
| Training set | Regulation data from rounds 1 to 8 | 66.7% |
| Validation set | Regulation data from rounds 9 to 10 | 16.7% |
| Test set | Regulation data from rounds 11 to 12 | 16.6% |
| Main evaluation metrics | Accuracy, Macro-F1, and high-risk recall | Evaluation of identification performance |
| Strategy evaluation metrics | Ineffective resource consumption rate, average response cycle, and misjudged intervention rate | Evaluation of regulation optimization |

In the data preprocessing stage, missing values, duplicate records and abnormal timestamps are cleaned, and then numerical features are normalized, and categorical variables are one-hot encoded. After word segmentation, denoising and semantic coding, the text disclosure data is transformed into a fixed-length vector, which is combined with features such as performance deviation, authentication change, and historical feedback to form the state input. In order to simulate the dynamic evolution process of greenwashing behavior, the experiment uses a sliding window to organize data, and each window corresponds to a

supervision cycle. The state in the window is used to generate the current action, and the feedback results after the window are used to update the policy parameters.

The model is trained in three stages. The training phase is used to learn the initial mapping relationship between the state and the supervision action. The validation phase is used to adjust the hyperparameters such as learning rate, discount factor and exploration rate. The testing phase is used to compare the recognition and decision-making effect of the model in the unknown supervision cycle. The number of training rounds of the reinforcement learning model is set as 500, the learning rate is 0.001, the discount factor is 0.90, and the initial exploration rate is 0.20, which gradually decays to 0.02 with the training process. The experimental comparison models include Logistic regression, support vector machine, random forest, XGBoost and static DQN models to test the advantages of the dynamic strategy update mechanism in the greenwashing supervision task. Through the above design, this paper can evaluate the practical application effect of the proposed model from three levels: recognition accuracy, action rationality and feedback optimization ability.

## 4.2 Analysis of experimental results

In the testing phase, the performance of the model was evaluated from two aspects: the identification effect of greenwashing risk and the optimization effect of regulatory strategy. The recognition results of different models on the test set are shown in Figure 5. The proposed model achieves better performance in the three indicators of Accuracy, Macro-F1 and high-risk recall, where Accuracy reaches 93.6%, Macro-F1 reaches 91.8%, and high-risk recall reaches 92.4%. This shows that the reinforcement learning model can continuously correct the state-action mapping relationship in multiple rounds of feedback, and the identification of high-risk subjects is more stable, especially it can reduce the missed judgment caused by fuzzy text disclosure and lagging performance records.

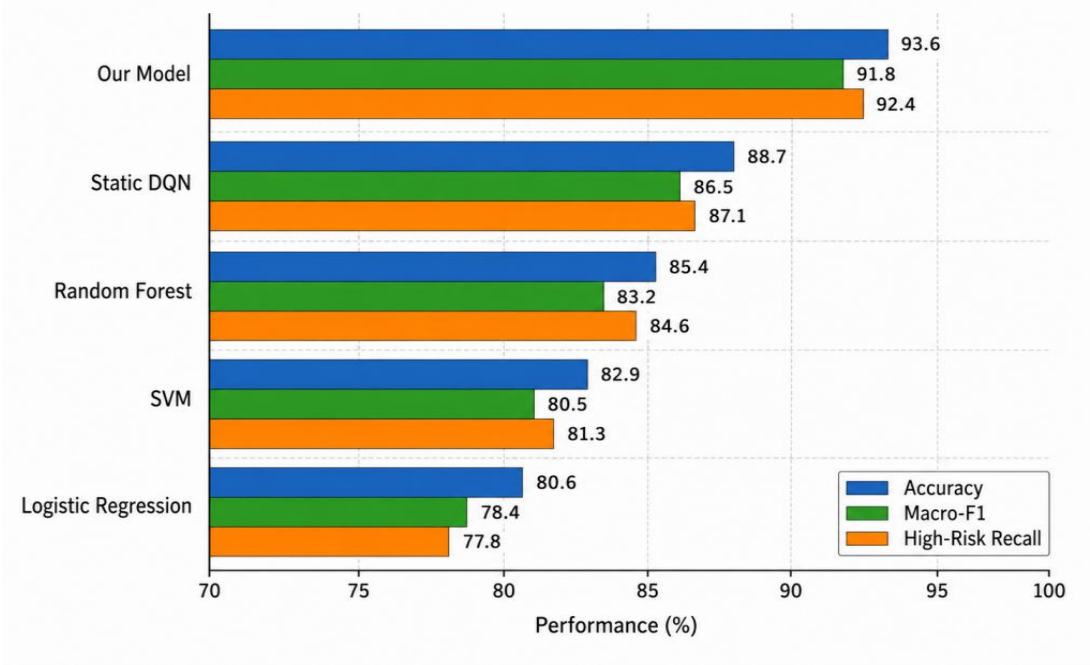


*Figure 5: Comparison of identification effects of different models for greenwashing risk*

It can be seen from Figure 5 that although the traditional classification model can complete the basic risk discrimination, it has insufficient ability to capture high-risk samples. Logistic regression has a strong dependence on linear relationships, and its recognition ability

is limited when multi-source features cross changes. SVM is stable in small and medium-sized feature Spaces, but it does not make sufficient use of continuous regulatory feedback. Random forest can deal with nonlinear features, but its output is still biased towards static judgment. Static DQN is better than the traditional model, but it lacks the continuous correction of feedback benefits, so it is still weaker than the proposed model in the optimization of subsequent actions.

In order to further compare the application effect of the model at the regulatory strategy level, this paper counts the ineffective resource consumption rate, average response period, misjudgment intervention rate and risk reduction rate of different models, and the results are shown in Table 4. The invalid resource consumption rate of the proposed model is 13.8%, which is lower than that of static DQN (18.6%) and random forest (23.4%). The average response period is shortened to 2.4 working days, indicating that the model can complete risk identification and action generation faster. The misjudgment intervention rate was controlled at 6.2%, indicating that multiple rounds of reward feedback have a certain inhibitory effect on excessive intervention.

*Table 4: Comparison of optimization effects of regulatory strategies of different models*

| Model method | Ineffective resource consumption rate (%) | Average response cycle (working days) | Misjudged intervention rate (%) | Risk reduction rate (%) |
|---|---|---|---|---|
| Proposed model | 13.8 | 2.4 | 6.2 | 31.7 |
| Static DQN | 18.6 | 3.1 | 8.5 | 25.4 |
| Random Forest | 23.4 | 4.0 | 10.8 | 21.6 |
| SVM | 26.1 | 4.5 | 12.3 | 18.9 |
| Logistic Regression | 29.7 | 5.2 | 14.6 | 15.8 |

Table 4 shows that the advantage of the proposed model is not only reflected in the recognition accuracy, but also in the matching ability of the supervision actions. Since risk reduction, resource investment and misjudgment loss are incorporated into the feedback evaluation of the model, the repeated execution of low-value actions can be reduced and the matching degree of actions such as review, spot check and continuous tracking can be improved after policy update. Compared with the static model, the proposed model shows stronger adaptive ability in the continuous supervision cycle, and can adjust the supervision intensity in time according to the change of the subject's behavior, so that the greenwashing risk identification and dynamic strategy optimization form a relatively stable closed loop.

## 4.3 Hyperparameter sensitivity analysis

In order to test the stability of the reinforcement learning supervision model under different parameter Settings, this paper carries out sensitivity experiments around the learning rate, discount factor and the decay speed of exploration rate. The learning rate affects the update range of model parameters, the discount factor determines the model's attention to long-term regulatory benefits, and the decay speed of the exploration rate relates to the balance between action attempts and strategy utilization of the model. In the experiment, Accuracy, Macro-F1, high risk recall, risk decline rate and convergence rounds are used as evaluation indicators, and the specific results are shown in Table 5.

*Table 5: Experimental results of hyperparameter sensitivity*

| Parameter type | Parameter value | Accuracy (%) | Macro-F1 (%) | High-risk recall (%) | Risk reduction rate (%) | Convergence epochs |
|---|---|---|---|---|---|---|
| Learning rate | 0.0100 | 90.8 | 88.1 | 89.3 | 27.4 | 260 |
| Learning rate | 0.0010 | 93.6 | 91.8 | 92.4 | 31.7 | 335 |
| Learning rate | 0.0001 | 92.5 | 90.6 | 91.2 | 30.5 | 470 |
| Discount factor | 0.80 | 91.2 | 88.9 | 89.7 | 28.1 | 310 |
| Discount factor | 0.90 | 93.6 | 91.8 | 92.4 | 31.7 | 335 |
| Discount factor | 0.95 | 92.9 | 91.0 | 91.8 | 30.9 | 390 |
| Exploration rate decay | 0.0010 | 91.7 | 89.4 | 90.1 | 28.8 | 300 |
| Exploration rate decay | 0.0005 | 93.6 | 91.8 | 92.4 | 31.7 | 335 |
| Exploration rate decay | 0.0001 | 92.1 | 90.2 | 91.0 | 29.6 | 430 |

It can be seen from Table 5 that when the learning rate is 0.0100, the model converges faster in the early stage of training, but the parameter update range is large, which easily leads to the fluctuation of the supervision action selection between adjacent risk states, the Accuracy is 90.8%, and the risk reduction rate is 27.4%. When the learning rate is reduced to 0.0010, the model can better balance the convergence speed and strategy stability, and the Accuracy is improved to 93.6%, Macro-F1 reaches 91.8%, and the high-risk recall rate reaches 92.4%. When the learning rate is further reduced to 0.0001, the model update process is relatively slow. Although the results remain stable, the number of convergence rounds increases to 470, and the training efficiency decreases.

The discount factor has an obvious impact on the long-term regulatory revenue of the model. When the discount factor is 0.80, the model pays more attention to the risk pressure drop in the current cycle and responds quickly in the short term, but the continuous tracking and subsequent correction are underutilized, and the risk reduction rate is 28.1%. When the discount factor is set to 0.90, the model can take into account the current action effect and the subsequent feedback benefit, and the overall performance is the best. After the discount factor was increased to 0.95, the model paid too much attention to long-term returns, and the action adjustment of some medium and low risk samples became conservative, and the number of convergence rounds increased to 390.

The exploration rate decay speed also affects the policy learning effect of the model. When the decay speed is 0.0010, the model enters the strategy utilization stage earlier, and the exploration of some implicit greenwashing samples is insufficient. When the decay speed is 0.0001, the model maintains a high exploration level for a long time, resulting in unstable action selection. Considering various indicators, learning rate 0.0010, discount factor 0.90 and exploration rate decay rate 0.0005 are more suitable as experimental parameters of the model in this paper.

## 4.4 Comparison with alternative machine learning models

In order to determine whether the proposed model has practical advantages over alternative machine learning methods, Logistic regression, SVM, random forest, XGBoost and static DQN are selected as comparison models. The comparison focuses on the recognition accuracy of greenwashing risk, Macro-F1, high risk recall and average response period. The comprehensive performance of different models is shown in Figure 6.

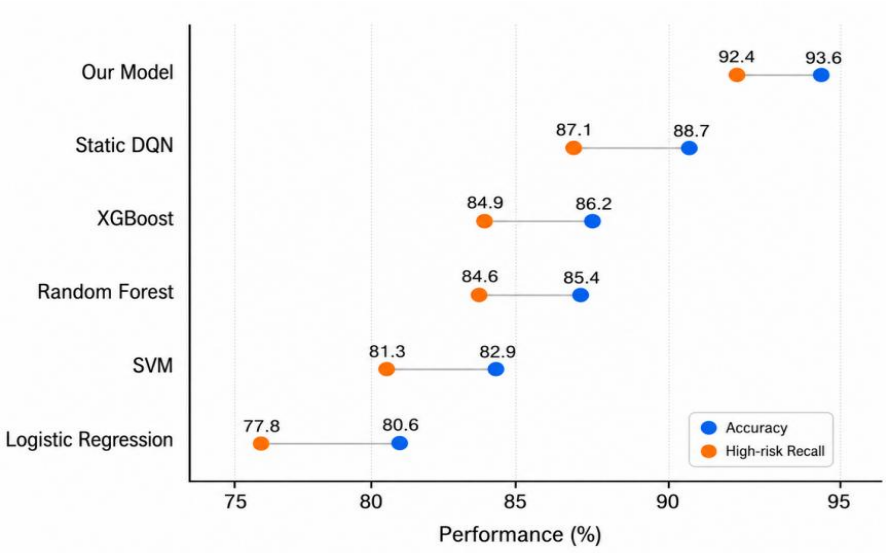


*Figure 6: Comparison of greenwashing supervision effects of different machine learning models*

It can be seen from Figure 6 that Logistic regression has certain advantages in calculation speed, but it is mainly suitable for classification scenarios with obvious linear relationship, and it is difficult to fully deal with the cross influence between green declaration, performance deviation and text anomaly, so the high-risk recall rate is only 77.8%. SVM has a good ability to distinguish samples with clear boundaries, but the model cannot use feedback information to modify the strategy when facing state changes in continuous supervision cycles, and the Accuracy is 82.9%. Random forest can depict part of the nonlinear feature relationship, and the Accuracy is improved to 85.4%, but its output still stays at the static risk judgment level, and can not directly generate subsequent regulatory actions. XGBoost is superior to traditional models in feature combination and classification performance, with an Accuracy of 86.2%. However, the model is still mainly supervised learning, and lacks dynamic response ability to risk changes after the execution of supervision actions. Static DQN can introduce an action selection mechanism to further improve the recognition effect, with an Accuracy of 88.7% and a high risk recall rate of 87.1%. However, its strategy update does not fully incorporate the resource consumption and misjudgment loss in multiple rounds of feedback. In contrast, the proposed model connects state recognition, action selection and reward feedback as a continuous optimization process, with an Accuracy of 93.6%, a high risk recall of 92.4%, and an average response period of 2.4 working days. The results show that the proposed model not only improves the identification effect of greenwashing risk, but also enhances the matching ability between regulatory actions and risk states.

## 4.5 Robustness testing in dynamic environments

In order to test the stability of the model in dynamic environment, the experiment further set up three kinds of disturbance scenarios: the sudden increase of green declaration text, the delayed update of performance data, and the centralized occurrence of abnormal feedback. The three types of scenarios correspond to the common problems of increasing information noise, temporary loss of evidence chain and short-term aggregation of risk signals in greenwashing supervision, respectively. During the testing process, the training parameters were kept unchanged, and only the environmental disturbance intensity in the test set was changed, which was used to observe the recognition ability of the model and the adjustment

effect of regulatory actions under non-stationary conditions. The model performance under different perturbation scenarios is shown in Figure 7.

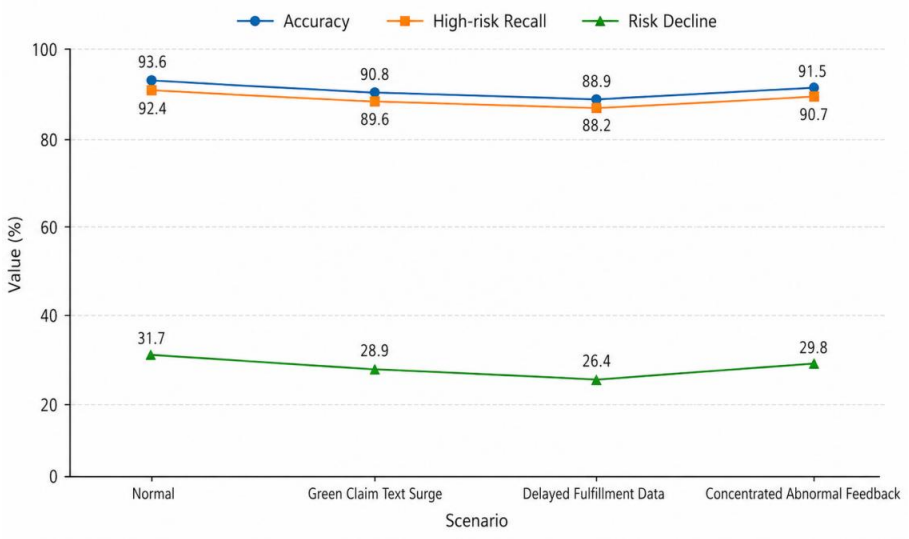


*Figure 7: Model robustness test results under dynamic disturbance scenario*

As can be seen from Figure 7, in the scene of sudden increase of green declaration text, some subjects increase the frequency of green expression in a short period of time, resulting in dense abnormal text signals. The Accuracy of the model in this paper decreases from 93.6% to 90.8%, and the high-risk recall rate remains at 89.6%, indicating that the joint input of semantic coding and performance bias can weaken the interference caused by single text noise. In contrast, the static classification model has a decrease in high risk recall of more than 6 percentage points in this scenario, which is prone to misjudge high-frequency disclosure as low risk performance. In the performance data delayed update scenario, part of the procurement, transportation and recovery records are written later, and the model can use historical feedback and certification change signals to maintain the basic judgment, with an Accuracy of 88.9% and a risk reduction rate of 26.4%. Although this scenario has a great impact on the state integrity, the reinforcement learning module will improve the selection probability of continuous tracking and data review actions, and avoid directly adopting high-intensity actions when there is insufficient evidence. In the scenario of abnormal feedback concentration, short-term complaints, abnormal detection and historical risk accumulation are integrated into the state calculation, the Accuracy reaches 91.5%, and the recall rate of high risk is 90.7%. The results show that the proposed model can still maintain good recognition stability and action adaptation ability under dynamic disturbance conditions, and has certain robustness.

## 4.6 Discussion

According to the experimental results, the dynamic regulatory strategy optimization model of supply chain greenwashing behavior based on reinforcement learning constructed in this paper shows good application effects in greenwashing risk identification, regulatory action matching and dynamic environment adaptation. The test results show that the Accuracy of the proposed model reaches 93.6%, Macro-F1 reaches 91.8%, and the high-risk recall rate reaches 92.4%, which are all higher than those of the comparison models such as Logistic regression, SVM, Random forest, XGBoost, and static DQN. This shows that the multi-source state recognition and reinforcement learning strategy update mechanism can effectively

improve the shortcomings of traditional static classification methods in greenwashing supervision tasks, so that the model can not only identify risks, but also adjust subsequent supervision actions according to the feedback results.

From the perspective of model action mechanism, the advantages of the proposed method mainly come from two aspects. On the one hand, the status recognition module integrates green declaration consistency, performance deviation, certification change, text anomaly and historical feedback into the calculation process, which reduces the deviation caused by single indicator judgment. Greenwashing behaviors are often not directly manifested as obvious violations, but are hidden in the inconsistent relationship between declaration strength, evidence support and actual performance. Therefore, multi-source feature fusion can improve the ability of the model to capture hidden risks. On the other hand, the reinforcement learning module continuously modified the action selection through reward feedback, so that the actions such as prompt, data review, key spot check, credit constraint and continuous tracking could form a more reasonable matching relationship with different risk states. In the experiment, the invalid resource consumption rate of the proposed model is 13.8%, the misjudgment intervention rate is 6.2%, and the average response period is shortened to 2.4 working days, which indicates that the model can reduce the cost of unnecessary intervention while controlling the risk. Compared with static DQN, the model further introduces resource consumption and misjudgment loss constraints, so that the policy update not only seeks the short-term risk reduction caused by high-intensity regulatory actions, but also pays more attention to the comprehensive income in consecutive cycles. In the dynamic disturbance test, even in the face of scenes such as sudden increase of green declaration text, delayed update of performance data and centralized occurrence of abnormal feedback, the model can still maintain an Accuracy of more than 88.9% and a high-risk recall rate of more than 88.2%, indicating that it has certain robustness. This result shows that reinforcement learning is suitable for dealing with the problem of state change and feedback delay in greenwashing regulation, and can make the regulation strategy change from fixed rules to dynamic optimization.

However, there are still some limitations in this paper. The experimental data is constructed based on the simulated green disclosure and performance supervision scenario of supply chain. Although it covers multiple types of information such as text, authentication, performance and feedback, compared with the real complex scenario, data noise, differences in subject behavior and cross-node risk transmission may still be more complex. Subsequent research can further introduce real operation data and combine graph neural network or multi-agent reinforcement learning methods to characterize the correlation influence between supply chain nodes. At the same time, the interpretability analysis of the model can be strengthened, and the contribution of different features to the selection of regulatory actions can be clarified, so as to improve the understandability and applicability of the model in actual regulatory scenarios.

# 5 Conclusion

Focusing on the problems of static identification lag, insufficient action matching and insufficient feedback utilization in the regulation of supply chain greenwashing behavior, this paper constructs a dynamic regulation strategy optimization model based on reinforcement learning. The model takes the consistency of green declaration, performance deviation, certification change, text anomaly and historical feedback as the state input, and incorporates prompt description, data review, key spot check, credit constraint and continuous tracking into the action space. The risk reduction, resource consumption and misjudgment loss are

comprehensively measured through the reward function. Thus, a closed-loop calculation process of "state recognition, action selection, feedback evaluation, policy update" is formed. Experimental results show that the proposed model has better performance in identifying greenwashing risk and optimizing regulatory strategies. In the test set, the Accuracy reaches 93.6%, Macro-F1 reaches 91.8%, and the high risk recall reaches 92.4%, which are better than the comparison models such as Logistic regression, SVM, random forest, XGBoost and static DQN. At the same time, the invalid resource consumption rate was controlled at 13.8%, the misjudgment intervention rate was reduced to 6.2%, and the average response period was shortened to 2.4 working days, which showed that the reinforcement learning mechanism could dynamically adjust the supervision intensity according to the behavior feedback, and improve the adaptability and stability of action selection. There are still some limitations in this paper. The experimental data are mainly constructed based on simulated green disclosure and performance scenarios of supply chain. Data noise, node association and behavior changes in real applications may be more complex. The subsequent research can further introduce real business data, combine graph neural network and multi-agent reinforcement learning methods to characterize the risk transmission relationship between nodes, and enhance the interpretation ability of the model, so as to provide more applicable technical support for the dynamic identification and strategy optimization of supply chain greenwashing behavior.

# About the Author

**Qintao Peng** was born in Jingmen, Hubei, China, in 1980. He received the M.S. degree in computer science and technology from Wuhan University of Technology, Wuhan, China, in 2010. He is currently an Associate Professor with the School of Economics and Management, Jingchu University of Technology, and is pursuing the Ph.D. degree in management science and engineering with China Three Gorges University. His research interests include decision analysis and management science.

**Fan Chen** was born in April 1981 in Jingmen City, Hubei Province, China. She holds a doctoral degree and is currently a teacher at Jingchu University of Technology in Jingmen, Hubei Province. Her research interests include Artificial Intelligence (AI), Information Security, and Big Data Analysis.

# References

[1] Sundarasen S, Zyznarska-Dworczak B, Goel S. Sustainability reporting and greenwashing: A bibliometrics assessment in G7 and non-G7 nations[J]. Cogent Business & Management, 2024, 11(1): 2320812.

[2] Ibrahim Nnindini S, Dankwah J B. Describing brown as green: an examination of the relationship between greenwashing and consumer negative emotive outcomes[J]. Cogent Business & Management, 2024, 11(1): 2367781.

[3] Carreño I. To address “greenwashing” and misleading environmental claims, the European Commission publishes a proposal on “green claims” and their substantiation[J]. European Journal of Risk Regulation, 2023, 14(3): 607-611.

[4] Boute R N, Gijsbrechts J, Van Jaarsveld W, et al. Deep reinforcement learning for inventory control: A roadmap[J]. European journal of operational research, 2022, 298(2):

401-412.

[5] De Moor B J, Gijsbrechts J, Boute R N. Reward shaping to improve the performance of deep reinforcement learning in perishable inventory management[J]. European Journal of Operational Research, 2022, 301(2): 535-545.

[6] Rolf B, Jackson I, Müller M, et al. A review on reinforcement learning algorithms and applications in supply chain management[J]. International Journal of Production Research, 2023, 61(20): 7151-7179.

[7] Hammler P, Riesterer N, Braun T. Fully dynamic reorder policies with deep reinforcement learning for multi-echelon inventory management[J]. Informatik Spektrum, 2023, 46(5): 240-251.

[8] Dehaybe H, Catanzaro D, Chevalier P. Deep reinforcement learning for inventory optimization with non-stationary uncertain demand[J]. European Journal of Operational Research, 2024, 314(2): 433-445.

[9] Geevers K, Van Hezewijk L, Mes M R K. Multi-echelon inventory optimization using deep reinforcement learning[J]. Central European Journal of Operations Research, 2024, 32(3): 653-683.

[10] Ruiz-Blanco S, Romero S, Fernandez-Feijoo B. Green, blue or black, but washing–What company characteristics determine greenwashing?[J]. Environment, Development and Sustainability, 2022, 24(3): 4024-4045.

[11] Moodaley W, Telukdarie A. Greenwashing, sustainability reporting, and artificial intelligence: A systematic literature review[J]. Sustainability, 2023, 15(2): 1481.

[12] Inês A, Diniz A, Moreira A C. A review of greenwashing and supply chain management: Challenges ahead[J]. Cleaner Environmental Systems, 2023, 11: 100136.

[13] Vangeli A, Małecka A, Mitręga M, et al. From greenwashing to green B2B marketing: A systematic literature review[J]. Industrial Marketing Management, 2023, 115: 281-299.

[14] Santos C, Coelho A, Marques A. A systematic literature review on greenwashing and its relationship to stakeholders: state of art and future research agenda[J]. Management Review Quarterly, 2024, 74(3): 1397-1421.

[15] Bernini F, Giuliani M, La Rosa F. Measuring greenwashing: A systematic methodological literature review[J]. Business Ethics, the Environment & Responsibility, 2024, 33(4): 649-667.

[16] Lagasio V. ESG-washing detection in corporate sustainability reports[J]. International Review of Financial Analysis, 2024, 96: 103742.

[17] Mohamadi N, Niaki S T A, Taher M, et al. An application of deep reinforcement learning and vendor-managed inventory in perishable supply chain management[J]. Engineering Applications of Artificial Intelligence, 2024, 127: 107403.

[18] Stranieri F, Fadda E, Stella F. Combining deep reinforcement learning and multi-stage

stochastic programming to address the supply chain inventory management problem[J]. International Journal of Production Economics, 2024, 268: 109099.

[19] Stranieri F, Stella F, Kouki C. Performance of deep reinforcement learning algorithms in two-echelon inventory control systems[J]. International Journal of Production Research, 2024, 62(17): 6211-6226.

[20] Nahhas A, Kharitonov A, Turowski K. Deep reinforcement learning for solving allocation problems in supply chain: An image-based observation space[J]. Procedia computer science, 2024, 232: 2570-2579.

[21] Yavuz T, Kaya O. Deep reinforcement learning algorithms for dynamic pricing and inventory management of perishable products[J]. Applied Soft Computing, 2024, 163: 111864.

[22] Kurian D S, Pillai V M, Raut A, et al. Deep reinforcement learning-based ordering mechanism for performance optimization in multi-echelon supply chains[J]. Applied Stochastic Models in Business and Industry, 2024, 40(5): 1433-1454.