Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 4, No. 11, 2025

Deep Reinforcement Learning Framework for Joint Corporate Budget-Revenue Allocation and Labor Cost Expenditure Planning

Ningyu Zhao[1*], You Lin[2], Naira Abulikemu[3]
[1]University of Southern California, Los Angeles, USA
[2]Cornell University, Ithaca, USA
[3]New York University, New York, USA
*Corresponding author: Ningyu Zhao; zzziazhao@gmail.com

Abstract: Corporate budget-revenue allocation and labor cost expenditure planning are tightly coupled decisions, yet prevailing practice treats them in isolation. This paper proposes HMARL-BW, an uncertainty-aware hierarchical multi-agent deep reinforcement learning framework that unifies budget allocation and workforce cost planning within a constrained Markov decision process. An upper-level agent apportions the cross-departmental budget pool, while lower-level agents optimize headcount and compensation under budgetary constraints; inter-departmental dependencies are encoded via a graph neural network. Distributionally robust risk constraints and a scenario simulator based on a conditional variational autoencoder (CVAE) enhance robustness against demand shocks, and attribution analysis based on Shapley additive explanations (SHAP) improves decision transparency. Experiments on a CSMAR-calibrated simulation environment and a digital-twin simulator show that the proposed method outperforms linear programming (LP), genetic algorithm (GA), MAPPO, and single-agent reinforcement learning (SA-RL) baselines, improving the revenue growth rate and labor cost efficiency by 11.8% and 9.4%, respectively, and reducing budget execution deviation by 23.3%, relative to the strongest baseline (SA-RL), while retaining 80.2% of the nominal reward under a -40% demand shock versus 64.1% for SA-RL.

Keywords: Budget-revenue allocation, distributionally robust optimization, explainable AI, graph neural network, hierarchical multi-agent reinforcement learning, labor cost expenditure planning

1. Introduction

In modern corporate management, budget allocation and workforce expenditure planning are tightly coupled: a department's budget constrains its hiring capacity, and its staffing level affects revenue generation. Despite this interdependence, the two processes are typically conducted separately.

Traditional approaches rely on static optimization such as linear programming (LP). Workforce planning has been extensively surveyed in the operations research literature [1], yet existing models seldom capture the sequential nature of budget revisions or the staffing-revenue feedback loop.

Deep reinforcement learning (DRL) can solve complex sequential decision-making problems under uncertainty [2]. Budget allocation in marketing has been addressed with offline constrained DRL [3], yet to the best of our knowledge, no prior work jointly optimizes corporate budget allocation and workforce cost planning within a unified reinforcement learning (RL) framework.

This paper makes the following contributions: (i) a constrained hierarchical multi-agent Markov decision process (MDP) formulation that couples budget allocation with workforce cost planning; (ii) a graph neural network (GNN) encoder for inter-departmental dependencies; (iii) distributionally robust optimization (DRO) constraints with a generative scenario simulator; and (iv) attribution analysis based on Shapley additive explanations (SHAP). Experiments on a CSMAR-calibrated simulator and a digital-twin environment show consistent improvements over the LP, genetic algorithm (GA), MAPPO, and single-agent RL (SA-RL) baselines.

2. Related Work

Cai et al. [3] applied offline constrained DRL to marketing budget allocation but addressed only single-level allocation. Constrained multi-agent MDPs formalize resource-sharing in energy systems and warehouse logistics [4], but applications in corporate finance remain limited.

Hierarchical RL has been explored through feudal networks [5] and HIRO [6]. MADDPG [7] trains centralized critics with decentralized actors; MAPPO [8] extends proximal policy optimization (PPO) to cooperative multi-agent games; QMIX [9] combines per-agent utilities through a monotonic mixing network. A comprehensive review of multi-agent RL (MARL) is provided in [10]. GNNs [11] encode relational structure among agents but have seen limited use in corporate settings.

The Wasserstein DRO framework [12] provides tractable robust optimization reformulations. SHAP values [13] offer

theoretically grounded attributions, with growing use for RL policy explanation [14].

3. Methodology

3.1 Problem Formulation

Consider a firm with K departments over T periods. At each period t, the firm allocates budget B_t and determines headcount n_t^k and compensation w_t^k per department. We formulate this as a constrained MDP [15] $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \mathcal{C}, \gamma)$ with $\gamma = 0.99$, where $\mathcal{C}$ comprises the budget feasibility constraints (3) and the DRO risk constraint (5).

The state is $s_t = (B_t, \mathbf{r}_{t-1}, \mathbf{n}_{t-1}, \mathbf{w}_{t-1}, \mathbf{m}_t, \mathbf{e}_t)$, where $\mathbf{r}_{t-1}$ is previous-period departmental revenue, $\mathbf{m}_t$ is the market demand indicator, and $\mathbf{e}_t$ captures organizational features (e.g., attrition rates and cost ratios). The reward combines three objectives:

$$R(s_t, a_t) = \alpha_1 \widetilde{\mathrm{RGR}}_t + \alpha_2 \mathrm{LCE}_t + \alpha_3 (1 - \mathrm{BED}_t),$$

where $\widetilde{\mathrm{RGR}}_t = \mathrm{clip}(\mathrm{RGR}_t / g_{\max}, 0, 1)$ is the revenue growth rate normalized by the simulator's maximum feasible growth $g_{\max}$, $\mathrm{LCE}_t = \mathrm{Rev}_t / (\mathrm{Rev}_t + \mathrm{LC}_t)$ is labor cost efficiency (with $\mathrm{LC}_t = \sum_k n_t^k w_t^k$; higher is better), and $\mathrm{BED}_t = \mathrm{clip}(|X_t - B_t| / B_t, 0, 1)$ is budget execution deviation (X_t = actual expenditure; lower is better). All three components lie in $[0, 1]$, yielding $R \in [0, 1]$. Weights are $\alpha_1 = 0.4, \alpha_2 = 0.3, \alpha_3 = 0.3$. Table I reports episode-level aggregate metrics in their original units (%), which are monotonically related to but not numerically identical to the per-step normalized reward plotted in Figs. 2 and 3.

3.2 Hierarchical Multi-Agent Architecture

As shown in Fig. 1, the upper-level agent outputs budget allocation $\mathbf{b}_t$ subject to $\sum_k b_t^k \le B_t$, trained via PPO [16]:

$$L^{\mathrm{upper}}(\theta) = \widehat{\mathbb{E}}_t \left[\min \left(\rho_t(\theta) \widehat{A}_t,\ \mathrm{clip}\left(\rho_t(\theta),\ 1 - \epsilon_c,\ 1 + \epsilon_c\right) \widehat{A}_t \right) \right],$$

where $\rho_t(\theta) = \pi_\theta / \pi_{\theta_{\mathrm{old}}}$, $\widehat{A}_t$ is the advantage estimate obtained via generalized advantage estimation (GAE; $\lambda_{\mathrm{GAE}} = 0.95$), and $\epsilon_c = 0.2$.

Each lower-level agent k selects $a_t^k = (n_t^k, w_t^k)$ given b_t^k and local state s_t^k. Budget feasibility is enforced via Lagrangian relaxation [17]:

$$L_k^{\mathrm{lower}}(\phi_k) = L_k^{\mathrm{actor}}(\phi_k) - \lambda_k \left(\sum_j c_{j,t}^k - b_t^k \right),$$

where $\lambda_k \ge 0$ is updated via dual gradient ascent ($\eta_\lambda = 0.01$). The upper-level policy is updated once per budgeting period, whereas lower-level agents update at every environment step with shared centralized critics.

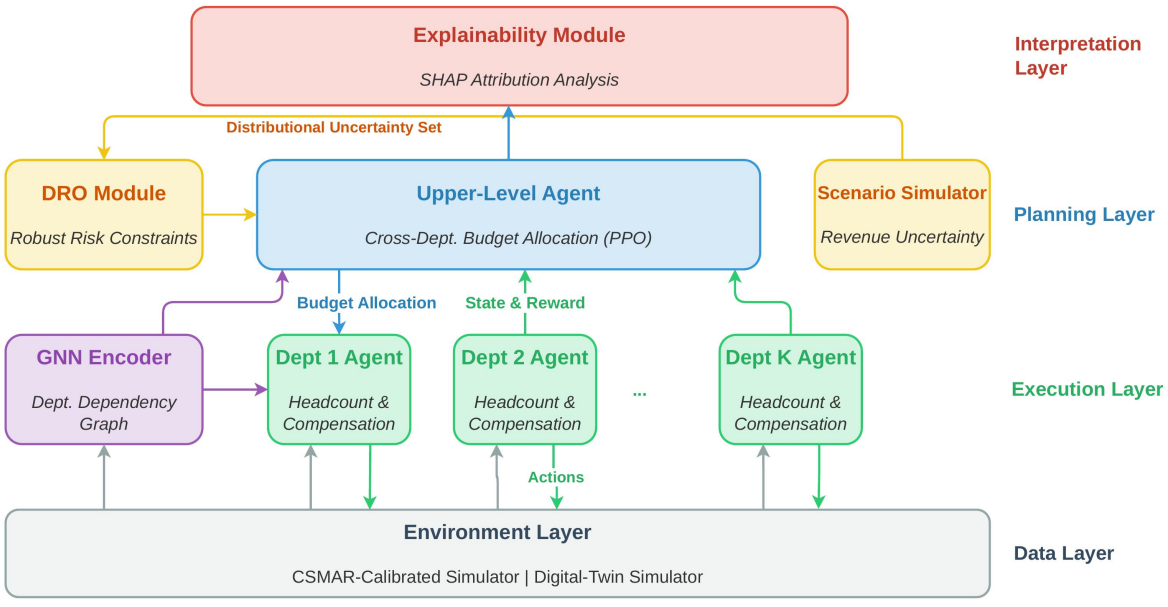

Figure 1. HMARL-BW architecture. The upper-level agent allocates budgets; lower-level agents optimize headcount and compensation and execute actions in the environment (closed loop). The GNN encoder captures inter-departmental dependencies. The scenario simulator feeds synthetic demand trajectories to the DRO module, which enforces robust risk constraints on the upper-level policy.

3.3 GNN-Based Dependency Encoder

Inter-departmental resource flows are modeled as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with adjacency derived from the production function: edge weight $\omega_{kk'}$ equals the output elasticity of department k''s contribution to k. We symmetrize ($A_{\text{sym}} = (A + A^{\top})/2$) to enable standard graph convolutional network (GCN) message passing; while this discards directional information, it ensures stable training for small graphs ($K = 5$). With self-loops ($\widetilde{A} = A_{\text{sym}} + I$, $\omega_{kk} = 1$) [11]:

$$\mathbf{h}_k^{(l+1)} = \sigma\left(\sum_{k' \in \widetilde{\mathcal{N}}(k)} \frac{\omega_{kk'}}{\sqrt{\widetilde{d}_k^{w}\widetilde{d}_{k'}^{w}}} \mathbf{W}^{(l)} \mathbf{h}_{k'}^{(l)}\right),$$

where $\widetilde{d}_k^{w} = \sum_{k'} \omega_{kk'}$ is the weighted degree.

3.4 Distributionally Robust Risk Constraints

We adopt Wasserstein DRO [12] with $N = 500$ historical observations:

$$\sup_{\mathbb{Q} \in \mathcal{B}_{\varepsilon}\left(\widehat{\mathbb{P}}_N\right)} \mathbb{E}_{\mathbb{Q}}\left[l(s, a)\right] \le \delta,$$

where $\mathcal{B}_{\varepsilon} = \{\mathbb{Q}: W_p\left(\mathbb{Q}, \widehat{\mathbb{P}}_N\right) \le \varepsilon\}$ ($\varepsilon = 0.05$, $\delta = 0.1$). A CVAE [18] with a 32-dimensional latent space generates synthetic demand trajectories.

3.5 Explainability

SHAP values [13] quantify each state variable's marginal contribution to the upper-level budget allocation policy.

4. Experiments

4.1 Setup

We use two simulation environments. The CSMAR-calibrated environment is parameterized from the CSMAR database (https://www.csmar.com; accessed 2024; institutional subscription). Panel data for 1,200 A-share listed firms (2018-2023) provide company-level revenue, costs, and headcount. Departmental proxies are constructed by mapping cost categories to departments (e.g., R&D expenses to the R&D department and selling expenses to the sales department). After filtering, 5,640 firm-year observations calibrate a simulator with $K = 5$ departments.

The digital-twin environment uses synthetic parameters with higher inter-departmental coupling and lower demand autocorrelation, providing a complementary test bed.

We compare against: (i) LP (solved once at $t = 0$; its high BED reflects its inability to adapt to demand realizations); (ii) GA (population size 200, 500 generations, and approximately 1.2×10^6 fitness evaluations, i.e., more environment interactions than the RL methods); (iii) MAPPO [8]; and (iv) SA-RL. All RL baselines enforce budget feasibility through a softmax output layer that normalizes allocations to the budget pool. MAPPO's centralized critic provides coordinated credit assignment, but the architecture offers no explicit budget allocation mechanism; effective budget coordination must be learned implicitly. All methods are evaluated zero-shot under demand shocks without re-optimization. RL agents are trained for 5×10^5 environment steps (learning rate 3×10^{-4}, batch size 2,048). The upper-level PPO network uses 256 hidden units, each lower-level agent uses 128, and the GNN comprises 2 layers with 64-dimensional embeddings. Simulator code and parameters will be released upon acceptance.

4.2 Main Results

Table I summarizes performance. HMARL-BW achieves the best results on all metrics. Relative to SA-RL, it improves RGR by 11.8% and LCE by 9.4%, and reduces BED by 23.3% on the CSMAR-calibrated environment.

Table 1. Performance comparison (mean +/- std, 5 seeds). RGR, LCE: higher is better (%); BED: lower is better (%). Best results are in bold.

Method	CSMAR-Calibrated			Digital Twin		
	RGR	LCE	BED	RGR	LCE	BED
LP	6.2+/-0.3	71.5+/-0.5	14.8+/-0.6	5.8+/-0.4	70.2+/-0.7	15.3+/-0.8
GA	7.5+/-0.8	73.8+/-1.2	12.1+/-1.4	6.9+/-1.0	72.4+/-1.4	13.2+/-1.6
MAPPO	8.6+/-0.6	75.8+/-0.9	9.5+/-0.8	7.9+/-0.8	74.2+/-1.1	10.1+/-1.0
SA-RL	9.3+/-0.5	76.4+/-0.8	8.6+/-0.7	8.8+/-0.6	74.8+/-1.0	9.2+/-0.8
HMARL-BW	10.4+/-0.4	83.6+/-0.6	6.6+/-0.5	9.7+/-0.4	82.1+/-0.6	7.2+/-0.5

MAPPO underperforms SA-RL in this task because the flat multi-agent architecture offers no explicit mechanism for allocating the shared budget pool; each agent must learn budget coordination implicitly, which proves slower and less reliable than the monolithic policy's ability to learn global trade-offs. This observation directly supports the necessity of hierarchical budget control.

Fig. 2(a) shows training convergence on the digital-twin environment. HMARL-BW converges to the highest reward (≈ 0.91), followed by SA-RL (≈ 0.78) and MAPPO (≈ 0.75). GA and LP are not trained by gradient-based environment interaction; their evaluation performance is shown as horizontal reference lines. Although GA consumes more environment evaluations (1.2×10^6) than the RL methods (5×10^5), it converges to a substantially lower reward.

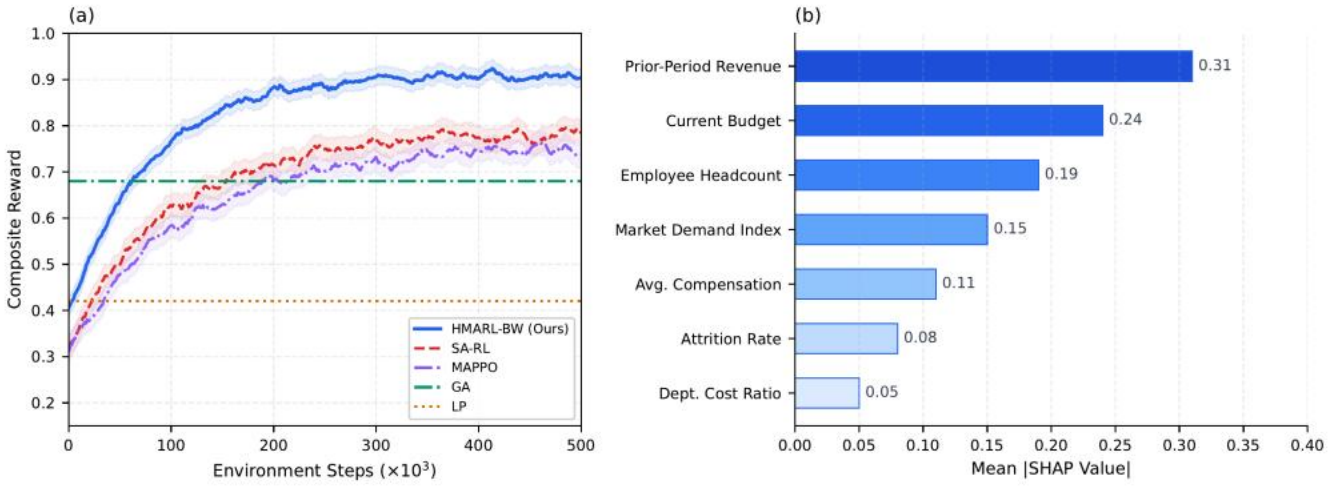

Figure 2. (a) Training convergence on the digital-twin environment. GA and LP are not trained by gradient-based environment interaction; their evaluation performance is shown as horizontal reference lines. (b) SHAP feature importance (mean |SHAP|) for the upper-level policy. Features correspond to the state definition in Section III-A.

4.3 Robustness Analysis

Fig. 3 shows composite reward under demand shocks from -40% to $+40\%$ on the digital-twin environment. HMARL-BW retains 80.2% of its nominal reward at -40% (*0.73/0.91*), versus 64.1% for SA-RL (*0.50/0.78*), 62.7% for MAPPO (*0.47/0.75*), 61.8% for GA (*0.42/0.68*), and 71.4% for LP (*0.30/0.42*). LP's high relative retention is misleading: its absolute reward (0.30) is far below HMARL-BW's (0.73). Positive shocks also degrade performance because demand surges exceed allocated capacity, causing revenue loss and elevated BED.

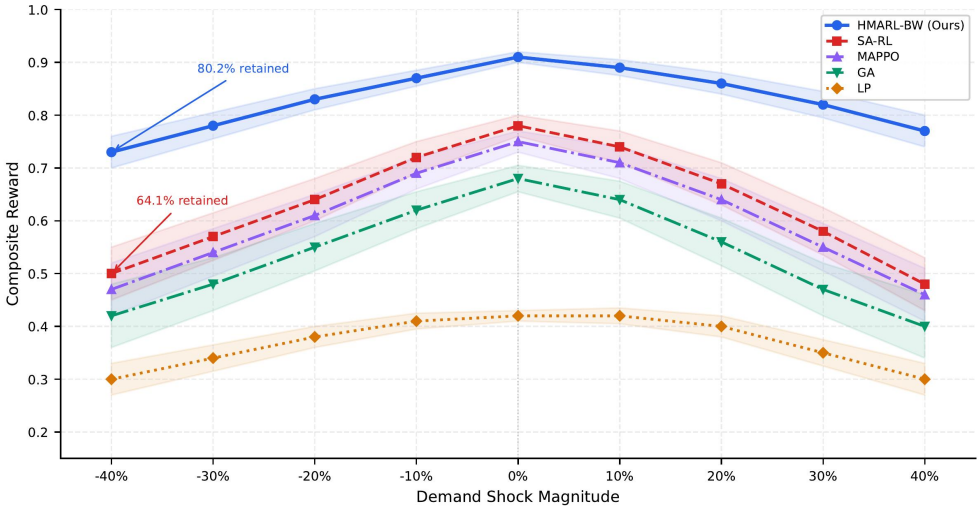

Figure 3. Composite reward under demand shocks on the digital-twin environment. Shaded bands: +/-1 std (5 seeds). LP's band reflects environment stochasticity. All methods evaluated zero-shot.

4.4 Explainability Analysis

Fig. 2(b) shows SHAP feature importance. Prior-period revenue is most influential (0.31), followed by current budget (0.24) and headcount (0.19). Market demand index ranks fourth (0.15); its moderate direct influence reflects the fact that demand uncertainty is primarily addressed through the DRO constraint and scenario simulator rather than through the raw demand feature. Average compensation (0.11) and attrition rate (0.08) play smaller roles.

4.5 Ablation Study

Table II isolates component contributions. The w/o Hierarchy variant retains GNN and DRO but removes the upper-level agent; it differs from the MAPPO baseline (Table I), which lacks all three enhancements.

Removing the GNN encoder causes the largest RGR drop (-2.1 percentage points (pp)); without it, the hierarchical architecture underperforms even SA-RL on RGR (7.6 vs. 8.8), showing that the GNN is essential for effective hierarchical

coordination. Removing the hierarchy causes the largest LCE degradation (-5.3 pp) and BED increase ($+2.2$ pp), indicating that the hierarchy primarily improves cost efficiency and budget execution. These complementary roles (GNN for revenue, hierarchy for cost discipline) underpin the full model's advantage.

For robustness, w/o DRO has the lowest R@−40% (0.60), retaining 68.2% ($0.60/0.88$) versus 80.2% for the full model, a 12.0 pp gap. The w/o GNN and w/o Hierarchy variants retain 75.9% ($0.63/0.83$) and 73.8% ($0.62/0.84$), respectively. DRO thus has the largest marginal impact on robustness.

Table 2. Ablation on the digital-twin environment (mean +/- std, 5 seeds).

Variant	RGR	LCE	BED	R@0%	R@-40%
Full	9.7+/-0.3	82.1+/-0.5	7.2+/-0.4	0.91+/-.01	0.73+/-.03
w/o GNN	7.6+/-0.6	78.0+/-0.9	9.1+/-0.7	0.83+/-.02	0.63+/-.04
w/o DRO	9.3+/-0.4	80.5+/-0.7	7.6+/-0.5	0.88+/-.01	0.60+/-.04
w/o Hier.	8.1+/-0.6	76.8+/-1.0	9.4+/-0.8	0.84+/-.02	0.62+/-.04

5. Conclusion

HMARL-BW integrates GNN-based dependency modeling, Wasserstein DRO, and SHAP explainability into a hierarchical multi-agent constrained MDP for joint budget-revenue allocation and labor cost planning. Experiments show consistent improvements over LP, GA, MAPPO, and SA-RL, with a 12.0 pp robustness advantage from DRO under a -40% demand shock. Ablation reveals complementary roles: the GNN drives revenue growth while the hierarchy enforces cost discipline. Limitations include simplified simulation dynamics, restriction to Chinese A-share firm parameters, and a fixed organizational structure. Model outputs should be treated as decision-support recommendations; deployment requires labor law compliance and fairness considerations.

References

[1] P. De Bruecker, J. Van den Bergh, J. Beliën, and E. Demeulemeester, "Workforce planning incorporating skills: State of the art," European Journal of Operational Research, vol. 243, no. 1, pp. 1-16, 2015.

[2] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. MIT Press, 2018.

[3] T. Cai, J. Jiang, W. Zhang, S. Zhou, X. Song et al., "Marketing budget allocation with offline constrained deep reinforcement learning," in Proceedings of the 16th ACM International Conference on Web Search and Data Mining. ACM, 2023, pp. 186-194.

[4] F. de Nijs, E. Walraven, M. de Weerdt, and M. Spaan, "Constrained multiagent Markov decision processes: a taxonomy of problems and algorithms," Journal of Artificial Intelligence Research, vol. 70, pp. 955-1001, 2021.

[5] A. S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, and K. Kavukcuoglu, "Feudal networks for hierarchical reinforcement learning," in Proceedings of the 34th International Conference on Machine Learning. PMLR, 2017, pp. 3540-3549.

[6] O. Nachum, S. Gu, H. Lee, and S. Levine, "Data-efficient hierarchical reinforcement learning," in Advances in Neural Information Processing Systems, vol. 31, 2018.

[7] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[8] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of PPO in cooperative multi-agent games," in Advances in Neural Information Processing Systems, vol. 35, 2022.

[9] T. Rashid, M. Samvelyan, C. S. de Witt, G. Farquhar, J. Foerster, and S. Whiteson, "Monotonic value function factorisation for deep multi-agent reinforcement learning," Journal of Machine Learning Research, vol. 21, no. 178, pp. 1-51, 2020.

[10] A. Oroojlooy and D. Hajinezhad, "A review of cooperative multi-agent deep reinforcement learning," Applied Intelligence, vol. 53, no. 11, pp. 13677-13722, 2023.

[11] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in International Conference on Learning Representations, 2017.

[12] P. Mohajerin Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," Mathematical Programming, vol. 171, no. 1-2, pp. 115-166, 2018.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[14] T. Hickling, A. Zenati, N. Aouf, and P. Spencer, "Explainability in deep reinforcement learning: A review into current methods and applications," ACM Computing Surveys, vol. 56, no. 5, pp. 1-35, 2023.

[15] E. Altman, Constrained Markov Decision Processes. Routledge, 2021.

[16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.

[17] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," in International Conference on Learning Representations, 2019.

[18] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in Advances in Neural Information Processing Systems, vol. 28, 2015.