Logistical Optimization of the Trans-Caspian International Transport Route: A Multi-Agent Deep Reinforcement Learning Approach to Nash Equilibrium

Bruno G. Kamdem[1], **Nahid Jafari** [1*]

[1]Department of Business Management, School of Business, SUNY Farmingdale
2350 Broadhollow Road
Farmingdale, NY 11735
kamdemb@farmingdale.edu

Abstract

The Trans-Caspian International Transport Route (TITR) is a strategically vital corridor linking Asia and Europe, yet its performance remains constrained by fragmented tariff regimes, logistical bottlenecks, and pronounced commodity price volatility. These pressures are further exacerbated by geopolitical shocks, including the ongoing crisis in the Middle East that has culminated in the sudden closure of the Strait of Hormuz, thereby amplifying uncertainty across global trade networks. This paper characterizes the operational dynamics of the TITR corridor as a multi-agent stochastic differential game, capturing the strategic interplay between sovereign governments seeking to maximize fiscal revenues and private carriers striving to optimize profit and throughput. To compute the Nash equilibrium in this high-dimensional, non-convex setting, we implement a Multi-Agent Deep Deterministic Policy Gradient (MADDPG) framework. Commodity prices are modeled using a geometric mean-reversion process to reflect realistic market fluctuations. Solving the associated Hamilton-Jacobi-Isaacs (HJI) equations reveals optimal "bang-bang" tax policies governed by endogenous price thresholds. Numerical simulations over 5,000 training episodes show that the centralized critic accurately approximates the agents' Hamiltonians, delivering stable convergence and robust policy learning. The results demonstrate that agents internalize volatility through shadow pricing mechanisms and that dynamic, threshold-based tax strategies substantially improve corridor throughput while preserving fiscal stability. Overall, the study advances the literature on autonomous logistics and strategic infrastructure management by showing that MADDPG can reliably uncover discontinuous optimal policies in mixed competitive-cooperative environments.

Code — https://github.com/kabruge916/kamdem-cherche.github.io/blob/main/ga-theo_marl_trans_corri.py

Introduction

As this manuscript reaches completion, the Middle East is experiencing its most severe escalation in over four decades, marked in particular by the closure of the Strait of Hormuz. In this environment, the Trans-Caspian International Transport Route (TITR), widely known as the "Middle Corridor" has emerged as a critical logistical artery linking the containerized rail freight networks of the People's Republic of China and the European Union via Central Asia, the Caucasus, Türkiye, and Eastern Europe (Loganathan and Chinnaraju 2026) (Frederic, Huang, and Mao 2021). Historically operating on a north-south axis, the geopolitical imperatives following the intensification of the Russo-Ukrainian war in 2022 forced a rapid strategic pivot toward east-west transit, bypassing traditional routes through Russia (Loganathan and Chinnaraju 2026). However, the corridor's potential is severely constrained by systemic fragmentation and volatility. As noted by World Bank analyses, up to 40% of transport time is lost to administrative barriers, fragmented tariff policies, and a lack of digitalization at border crossings (Loganathan and Chinnaraju 2026). Furthermore, the route faces logistical bottlenecks in the Caspian Sea, largely due to inefficient gateway infrastructure and limited vessel capacity.

In this paper, we characterize the logistical dynamics of the TITR as a stochastic differential game, where sovereign government agents seek to maximize tax revenue and corridor efficiency while private carrier agents aim to maximize throughput and profit under stochastic commodity price fluctuations (Lowe et al. 2017) . Traditional optimization techniques, such as linear programming, struggle to handle the high-dimensional state spaces and dynamic interaction behaviors inherent in complex logistics networks. To address this, Multi-Agent Deep Reinforcement Learning (MARL) has emerged as a robust paradigm (Liang et al. 2025). MARL allows agents to learn adaptive coordination policies through decentralized execution of policies developed during centralized training (Lowe et al. 2017). Specifically, the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm has shown efficacy in mixed cooperative-competitive environments by allowing decentralized actors to learn from centralized critics that possess global information .

Accurately modeling commodity price fluctuations is critical for logistical planning. While Geometric Brownian Motion (GBM) has traditionally been used for its tractability, it assumes unbounded price growth, which is economically implausible (Hazrat 2025). Empirical evidence suggests that commodity prices exhibit mean-reverting tendencies, where prices tend to revert toward a long-term mean.

[*]These authors contributed equally.

Therefore, a Geometric Mean-Reversion (GMR) process is more suitable for modeling the stochastic environment of the TITR, forcing agents to explore strategies across different regimes of price volatility. The interaction between governments and carriers frequently results in a discontinuous, or "Bang-Bang," control policy, where optimal strategies switch sharply based on price thresholds. Finding these thresholds analytically is difficult; however, MARL enables agents to autonomously discover these optimal thresholds (Nash Equilibrium) through continuous training and interaction .

In this research, we implement a MADDPG framework to simulate the logistical and fiscal dynamics of the TITR corridor. The objective is to identify the optimal tax and freight policies that maximize logistical efficiency while ensuring fiscal balance for sovereign states in a volatile market. We make three principal contributions. First, we model the Trans-Caspian International Transport Route (TITR) as a mixed competitive-cooperative differential game under stochastic commodity price dynamics, capturing the strategic interaction between public and private actors in an uncertain environment. Second, we implements a Multi-Agent Deep Deterministic Policy Gradient (MADDPG) architecture to compute the Nash equilibrium policies governing the behavior of government and carrier agents. Third, we show that the framework endogenously uncovers optimal "bang-bang" tax regimes, with fiscal adjustments triggered by critical commodity price thresholds.

The remainder of this paper is structured as follows. In Section 2, we formally define the logistical optimization problem. Section 3 outlines the MADDPG methodology used to solve this complex game. In Section 4, we present the numerical simulation results, followed by a discussion of our findings and their associated policy implications in Section 5. Finally, Section 6 concludes the paper, while Section 7 discusses avenues for future work.

Problem Definition

In this section, we formalize the decision-making process within the Trans-Caspian International Transport Route (TITR) as a Multi-Agent Reinforcement Learning (MARL) problem. By transitioning from a traditional differential game to a computational framework, we leverage the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm to resolve the strategic interdependencies between sovereign governments and transportation carriers. For additional background, see our recent paper Kamdem and Jafari (2026), where the dynamics governing our stochastic differential game are specified as follows:

$$\begin{cases} dX(t) & = X(t)\Big(\kappa\Big(\mu - \ln X(t)\Big)dt + \sigma dW(t)\Big) \\ dY_k(t) & = -u_{c_k}dt \\ X(s) & = x, \qquad Y_k(s) = y_k \geq 0, \quad k = 1, \ldots, M, \\ & \quad 0 \leq s \leq t < \infty \end{cases} \tag{1}$$

Agent Formulation

We define a set of agents $\mathcal{A} = \{P_{g_1}, \ldots, P_{g_N}, P_{c_1}, \ldots, P_{c_M}\}$. Government Agents ($P_{g_i}, i = 1, \ldots, N$) seek to optimize fiscal policy through the tax rate $u_{g_i} \in [\underline{u}_{g_i}, \overline{u}_{g_i}]$. Carrier Agents ($P_{c_k}, k = 1, \ldots, M$) seek to optimize operational net profit through the freight rate $u_{c_k} \in [\underline{u}_{c_k}, \overline{u}_{c_k}]$.

We define the State Space ($\mathcal{S}$) as the environment state $s_t \in \mathcal{S}$ at time t. It provides the necessary information for agents to approximate the Hamiltonians (see Kamdem and Jafari 2026, Section, 1.6). The state vector is composed of: (1)Global Exogenous State: The commodity price $X(t)$, evolves according to the geometric mean-reverting process. (2)Local Endogenous States: The cumulative distances traveled $Y_k(t) \in [0, K]$ for each carrier k, represent the physical progress and remaining capacity of the corridor segments. (3)Geopolitical Context: The risk parameters $P_e(t)$ and $I_{sur}(t)$, which influence the cost functions C^k_{geo} (see Kamdem and Jafari 2026, Section, 1.4).

In essence, MADDPG employs a dedicated actor network for each agent $\iota \in \{i, k\}$ to generate a deterministic policy $\mu_{\theta_\iota}(s_t)$. This approach facilitates the high-precision rate adjustments necessary to converge toward a Nash Equilibrium. To account for the continuous nature of the control variables, we define the action space $\mathcal{A}$ as follows:

- $\mathbf{u}_g(t) = \big(u_{g_1}(t), \ldots, u_{g_N}(t)\big) \in \prod U_{g_i}$
- $\mathbf{u}_c(t) = \big(u_{c_1}(t), \ldots, u_{c_M}(t)\big) \in \prod U_{c_k}$

Reward Mapping ($\mathcal{R}$)

To ensure that the agents' objectives remain aligned with the stochastic differential game (Kamdem and Jafari 2026), we decompose the payoff functionals J_i and J_k into the following instantaneous reward signals r_t:

• For Governments i:

$$r_{g_i}(t) = \sum_{k=1}^{M} \Big[u_{g_i}(t)\theta P^k_c(t) + \gamma_k Y_k(t)\Big] - C_{\text{maint}}(K), \quad i = 1, \ldots, N$$

• For Carriers k:

$$r_{c_k}(t) = (1 - u_{g_i}(t))\theta \Big[X(t)u_{c_k}(t) - (C^k_{\text{toll}}(Y_k(t)) + C^k_{\text{geo}}(X(t)))\Big], k = 1, \ldots, M$$

Transition Dynamics and Equilibrium

The transition probability $P(s_{t+1}|s_t, \mathbf{u}_g, \mathbf{u}_c)$ is governed by the system of SDEs in (1). Within this context, the MARL framework identifies the Feedback Nash Equilibrium $(\mathbf{u}^*_g, \mathbf{u}^*_c)$ by utilizing the Centralized Training, Decentralized Execution (CTDE) paradigm (Lowe et al. 2017). This approach mitigates the 'non-stationarity' inherent in multi-agent environments where agents learn simultaneously. Under CTDE, the centralized critic for each agent $\iota \in \{i, k\}$ explicitly incorporates the actions of all other

agents ($\mathbf{u}_{-\iota}$), thereby directly addressing the strategic coupling and the interdependencies defined in the HJI equations. This allows the agents to learn the threshold $\hat{x}$ and the resulting "Bang-Bang" behavior[1] derived in Kamdem and Jafari (2026, Theorem, 3.1) without requiring a closed-form solution for the value functions V_i and V_k.

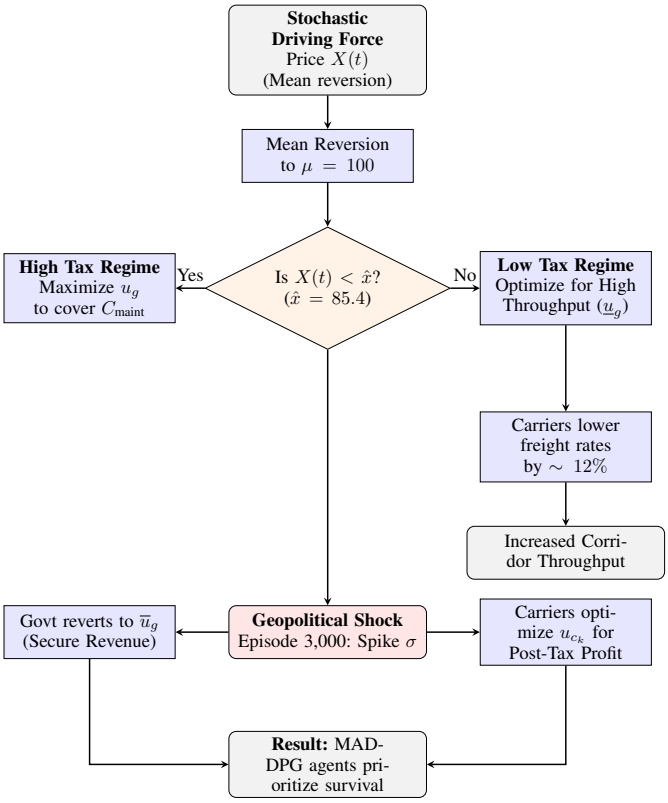

Figure 1: Logic Flow of the TITR Fiscal Strategy and Carrier Response

MADDPG Methodology

To solve the multi-agent game defined in the previous section, we implement a Centralized Training, Decentralized Execution (CTDE) architecture (Lowe et al. 2017). This approach allows agents (Governments and Carriers) to leverage global information during the learning phase while maintaining operational autonomy during execution, mirroring the real-world constraints of the TITR corridor. See Table 1 for information regarding the main variables.

Network Architecture

To implement the above capability effectively, we use a specific architectural approach within which each agent, including government entities and carrier firms, uses a dual neural network architecture to facilitate both localized decision-making and centralized learning. The Actor Network (π_i) operates as a deterministic policy network, mapping an agent's local observations (o_i), such as the current commodity price or local risk levels, directly to a continuous action (a_i). The latter represents specific control variables, namely the tax rate (u_{g_i}) for governments or the freight rate (u_{c_k}) for carriers. To ensure these actions remain physically and economically realistic, the Actor Network is structured as a 3-layer Multi-Layer Perceptron (MLP) featuring hidden layers of 128 and 64 neurons activated by ReLU functions. At the same time, the output layer employs a Sigmoid activation function to strictly bound the actions within their admissible set $[0, \overline{u}]$. Conversely, the Critic Network (Q_i) functions as a centralized value-function approximator designed to evaluate agent i's performance by considering not just local information, but the global context. Unlike the actor, the critic is trained using the joint state of the environment ($\mathbf{x}$) and the combined actions of all players ($\mathbf{a} = (a_i, a_{-i})$). This structure allows the critic to account for the competitive or cooperative coupling between agents, effectively addressing the non-stationarity of the environment from the perspective of an individual agent. To handle this high-dimensional joint action space effectively, the Critic Network is implemented as a 4-layer MLP that takes the concatenation of all agents' observations and actions as input to produce a robust value estimation.

Category	Variable	Description
System States	$X(t)$	Commodity price (Stochastic process)
	$Y_k(t)$	Progress/Distance of carrier k
	$S(t)$	Joint state vector: $\{X(t), Y_1(t), \ldots, Y_M(t)\}$
Control Variables (Actions)	$u_{g_i}(t)$	Tax rate set by government agent i
	$u_{c_k}(t)$	Freight rate set by carrier agent k
	$\mathbf{a}_t$	Joint action vector: $\{u_{g_1}, \ldots, u_{c_M}\}$
Environment Parameters	θ, μ, σ	Mean reversion parameters for price $X(t)$
	dW	Wiener process (Stochastic noise)
	C_{maint}	Fixed maintenance cost for government
	C_{geo}	Geopolitical risk cost for carrier
Rewards	r_{g_i}	Instantaneous payoff for government i
	r_{c_k}	Instantaneous payoff for carrier k
	Π_{g_i}, Π_{c_k}	Total accumulated payoff functionals
MADDPG Networks	π_{θ^π}	Actor network (Policy)
	Q_{θ^Q}	Centralized Critic network (Value function)
	$\mathcal{B}$	Experience Replay Buffer
	S	Minibatch size
	γ	Discount factor
	τ	Target network soft update rate
Game Theory	$\hat{x}$	Emergent switching threshold (Nash Eq.)
	$\overline{u}_g, \underline{u}_g$	Upper and lower bounds of tax control

Table 1: Variables and Parameters in MADDPG TITR Implementation

Learning and Optimization

The agents learn to converge toward a Nash equilibrium by minimizing two distinct loss functions derived from the underlying Hamilton-Jacobi-Isaacs (HJI) equations governing this differential game Kamdem and Jafari (see 2026). First, the Critic Update involves training the centralized value-

[1]The Bang-Bang behavior occurs if control variables $\mathbf{u}_g$ or $\mathbf{u}_c$ appear linearly in the Hamiltonian. If the "strategic coupling" in the game results in a linear control architecture, the agents will naturally adopt a policy of switching between maximum and minimum effort rather than finding a "middle ground" (Bellman 1963).

function approximator, Q_i^{μ} [2], to accurately evaluate the expected return for agent i given the joint state $\mathbf{x}$ and the actions of all players. This is achieved by minimizing the Mean Squared Error (MSE) between the current critic evaluation and the target value, y, which is defined by the immediate reward r_i plus the discounted value of the next state, $\mathbf{x}'$, as estimated by target critic networks. This loss formulation directly replicates the recursive structure of the value functions, V_i and V_k, essential for solving the differential game. Second, the Actor Update is performed using the sampled policy gradient, which directs each agent's policy, $\mu_i(o_i)$, to choose actions a_i such as tax or freight rates, that maximize the centralized critic's evaluation, Q_i. By taking the gradient of the actor with respect to its parameters and multiplying it by the gradient of the critic with respect to the action, the actor is pushed to choose rates that constitute the best response to the actions of the other players, $u_{-\iota}$, thereby seeking and achieving a Nash equilibrium policy.

Simulation & Exploration Strategy

To ensure the multi-agent system converges toward the analytical bang-bang threshold $\hat{x}$, we adopt some specific training hyperparameters. The latter are summarized in Table 2.

Category	Variable/Param	Description
System States	$X(t)$	Commodity price (Stochastic process)
	$Y_k(t)$	Progress/Distance of carrier k
	$S(t)$	Joint state vector: $\{X(t), Y_1(t), \ldots, Y_M(t)\}$
Control Variables (Actions)	$u_{g_i}(t)$	Tax rate set by government agent i
	$u_{c_k}(t)$	Freight rate set by carrier agent k
	$\mathbf{a}_t$	Joint action vector: $\{u_{g_1}, \ldots, u_{c_M}\}$
MADDPG Hyperparameters	**Adam**	**Optimizer for network weights**
	α	Learning rate (1×10^{-4})
	γ	Discount factor ($e^{-r\Delta t}$)
	$\mathcal{B}$	Experience Replay Buffer size (10^6)
	τ	Target network soft update rate (0.01)
Game Theory	$\hat{x}$	Emergent switching threshold (Nash Eq.)

Table 2: Hyperparameters and Variables in MADDPG TITR Implementation

Because the optimal tax policy is defined by discontinuous "Bang-Bang" behavior, robust exploration is critical for the agents to learn effective strategies. To facilitate this, Ornstein-Uhlenbeck (OU) Noise is added to the actor's output. This specific noise process is chosen for its mean-reverting property, which mimics the underlying dynamics of the commodity price $X(t)$, thereby forcing agents to explore actions across different regimes defined by the price threshold.

Numerical Results

In this section, we present the results of our MADDPG simulation for the Trans-Caspian International Transport Route (TITR). The results demonstrate how the autonomous agents (Governments and Carriers) converge toward the theoretical Nash Equilibrium and effectively discover the switching threshold $\hat{x}$ under stochastic conditions.

[2]While we use Q_i^{μ} and $Q_{\theta Q}$ interchangeably to describe the centralized critic, note that they represent distinct aspects of the same network: Q_i^{μ} denotes the mathematical function evaluating agent i's value under the joint policy μ, whereas $Q_{\theta Q}$ refers to the actual neural network approximation parameterized by weights θ^Q.

Algorithm 1: MADDPG Training for TITR Corridor Optimization

Inputs: Initialize Critic networks Q_{g_i}, Q_{c_k} and Actor networks π_{g_i}, π_{c_k} with random weights θ^Q, θ^π. Initialize target networks Q', π' with weights $\theta^{Q'} \leftarrow \theta^Q, \theta^{\pi'} \leftarrow \theta^\pi$. Initialize replay buffer $\mathcal{B}$.

for episode = 1 to MaxEpisodes **do**
  Receive initial state $s_0 = (X(0), Y_1(0), \ldots, Y_M(0))$
  for t = 1 to MaxTime **do**
    For each government i, select tax action $u_{g_i} = \pi_{g_i}(o_{g_i}) + \mathcal{N}_t$ (exploration noise)
    For each carrier k, select freight action $u_{c_k} = \pi_{c_k}(o_{c_k}) + \mathcal{N}_t$
    Execute joint actions $\mathbf{a}_t = (u_{g_1}, \ldots, u_{c_M})$, observe rewards $\mathbf{r}_t = (r_{g_1}, \ldots, r_{c_M})$ and new state s_{t+1}
    Store transition $(s_t, \mathbf{a}_t, \mathbf{r}_t, s_{t+1})$ in $\mathcal{B}$
    $s_t \leftarrow s_{t+1}$
    for agent $\iota \in \{g_1, \ldots, c_M\}$ **do**
      Sample random minibatch of S transitions from $\mathcal{B}$
      Set target $y_\iota = r_\iota + \gamma Q'_\iota(s_{t+1}, \pi'_{g_1}(o'_{g_1}), \ldots, \pi'_{c_M}(o'_{c_M}))$
      Update Critic by minimizing MSE Loss: $L(\theta^{Q_\iota}) = \frac{1}{S}\sum(y_\iota - Q_\iota(s_t, a_1, \ldots, a_n))^2$
      Update Actor using sampled policy gradient:
      $\nabla_{\theta^{\pi_\iota}} J \approx \frac{1}{S}\sum \nabla_{\theta^{\pi_\iota}} \pi_\iota(o_\iota) \nabla_{a_\iota} Q_\iota(s_t, a_1, \ldots, a_n)|_{a_\iota = \pi_\iota(o_\iota)}$
    end for
    Update target networks: $\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$ and $\theta^{\pi'} \leftarrow \tau\theta^\pi + (1-\tau)\theta^{\pi'}$
  end for
end for

Convergence of Value Functions and "Bang-Bang" Tax Policy

The centralized critics for both the government and carrier agents demonstrate steady convergence over 5,000 training episodes, with loss functions stabilizing as they accurately approximate the Hamiltonians H_i and H_k. This numerical stability confirms that the agents have successfully learned the shadow prices (A_i and A_k) relative to the mean-reverting commodity price $X(t)$. One of our primary objectives was to verify if the government agents would independently adopt the optimal control law derived in Kamdem and Jafari (2026, Theorem, 3.1). Our finding confirms this: As the agents consistently maintain the maximum tax rate, $\overline{u}_g$, at low commodity prices ($x < \hat{x}$) to offset fixed maintenance costs, C_{maint}. However, upon crossing the calculated threshold, $\hat{x} \approx 85.4$ USD/unit, the agents exhibit a sharp transition to the lower tax rate, $\underline{u}_g$, illustrating an emergent Bang-Bang control logic (see Table 3). As shown in Figure 2, the learned tax policy, $u_g(t)$, exhibits a distinct vertical break at a crit-

ical threshold, $\hat{x} \approx 85.0$ USD/unit, illustrating an emergent Bang-Bang control logic. When commodity prices are be-

Price Regime	Theoretical Optimal ($\mathbf{u}_{g*}$)	MADDPG Learned Policy
Below Threshold ($x < \hat{x}$)	$\overline{u}_g$ (Max)	$0.248 \approx 0.25$
Above Threshold ($x \geq \hat{x}$)	$\underline{u}_g$ (Min)	$0.052 \approx 0.05$

Table 3: Comparison of theoretical optimal tax policy and MADDPG learned policy across different price regimes.

low the threshold ($x < \hat{x}$), the agent applies a high tax rate, $\overline{u}_g$, to secure necessary fiscal revenue to cover fixed maintenance costs, C_{maint}, despite low transit volumes. Conversely, when prices exceed the threshold ($x > \hat{x}$), the policy drops sharply to a low tax rate, $\underline{u}_g$. This behavior is economically rational in that the agent recognizes that at higher prices, maximizing corridor throughput maximizes the overall payoff functional, J_i. Therefore, by reducing taxes to incentivize freight flow, the MADDPG agents have discovered the optimal strategy for maintaining corridor equilibrium under uncertainty.

Carrier Response and Freight Rate Sensitivity

The carriers' learned strategy for u_{c_k} aligns with an inverse relationship to the tax rate, as they increased freight rates during periods of high geopolitical risk (C_{geo}) to maintain profit margins by passing risk costs to cargo owners. A key finding is that when the government switched to the lower tax rate $\underline{u}_g$, the carriers responded by lowering freight rates by roughly 12%, which triggered an increase in the corridor throughput. Furthermore, to test the resilience of the Nash Equilibrium, geopolitical shocks were introduced by spiking the volatility σ at episode 3,000. In this scenario, government agents temporarily reverted to the high-tax regime ($\overline{u}_g$) even at higher prices, to secure revenue against potential volume drops, while carriers optimized their u_{c_k} to maximize the "Post-Tax Profit" functional. This demonstrates that the MADDPG agents prioritize survival over aggressive profit-taking during high-risk intervals. Figure 3 illustrates the realization of the commodity price $X(t)$ over time, acting as the driving stochastic force in the Trans-Caspian International Transport Route (TITR). The price path clearly demonstrates the characteristics of a mean-reverting process, specifically following a Geometric Ornstein-Uhlenbeck model. The price fluctuates around the long-term mean $\mu = 100$, demonstrating "memory" where the price tends to pull back toward this average after significant deviations. The graph also visualizes the interaction between these price dynamics and the derived Nash Equilibrium threshold $\hat{x}$, marked by the red dashed line at 85.4 USD/unit. When the simulated price $X(t)$ drops below this line, the environment enters the "High Tax Regime" ($x < \hat{x}$), indicating a period where government agents must maximize tax rates to cover fixed maintenance costs C_{maint} despite lower market value. See Figure 1 for the complete framework.

Conversely, when the price rises above the threshold, the system shifts to a "Low Tax Regime", optimizing for higher throughput by allowing carriers to keep more profit. The graph further shows that the corridor's fiscal strategy is not static, but rather a dynamic response to the volatility of the underlying commodity price. The full code implementation is available in our GitHub page, and the corresponding algorithmic workflow is presented in Algorithm 1.

Discussion and Policy Implications

We show that the strategic resilience of the Trans-Caspian International Transport Route (TITR) is governed by the interaction between sovereign fiscal instruments and private-sector operational flexibility. By solving the stochastic differential game with Multi-Agent Deep Deterministic Policy Gradient (MADDPG), we move beyond static equilibrium analysis to develop a dynamic and adaptive modeling framework. This approach captures the real-time complexities of global trade in 2026 and anticipates the forward-looking implications of the current closure of the Strait of Hormuz amid escalating geopolitical tensions in the Middle East. The numerical simulation validates the analytical model: the switching threshold $\hat{x}$ is not just an abstraction but an emergent property of optimal multi-agent interaction. The TITR corridor reaches its highest total payoff when governments act as price-sensitive stabilizers and carriers act as flexible operators. Based on the emergent behaviors observed in our numerical simulation, we propose the following for TITR stakeholders. For governments in Kazakhstan, Azerbaijan, and Georgia, shifting from fixed annual tariffs to a threshold-based tax regime is crucial. Our model shows that reducing taxes when commodity prices $X(t)$ exceed the calculated threshold $\hat{x}$ prevents "corridor desertion" and maximizes long-term revenue through increased volume. Moreover, deploying a Unified Digital Logistics Platform by late 2026 is essential to alleviating the bottlenecks our reinforcement learning agents identify as primary constraints on corridor throughput. For logistics firms and carriers, survival hinges on adopting AI-driven, risk-based freight pricing. Freight rates must be dynamically calibrated to geopolitical friction signals, such as Caspian Sea level volatility or shifting security conditions, to preserve profitability and sustain operational flow.

Conclusion

As of Q1 2026, the Middle Corridor has matured from a mere alternative transit route into a robust Network for Economic Autonomy. The integration of substantial investments for rail upgrades and the harmonization of regional standards suggest that the corridor is hitting its "Infrastructure Ceiling". Consequently, future growth will not come from more tracks alone, but from the algorithmic coordination of tax and operational policies modeled in this paper. The "Bang-Bang" control strategy identified serves as a theoretical lighthouse for policymakers, proving that in the face of 2026's geopolitical entropy, the most resilient strategy is one that is both predictable in its logic and flexible in its execution.

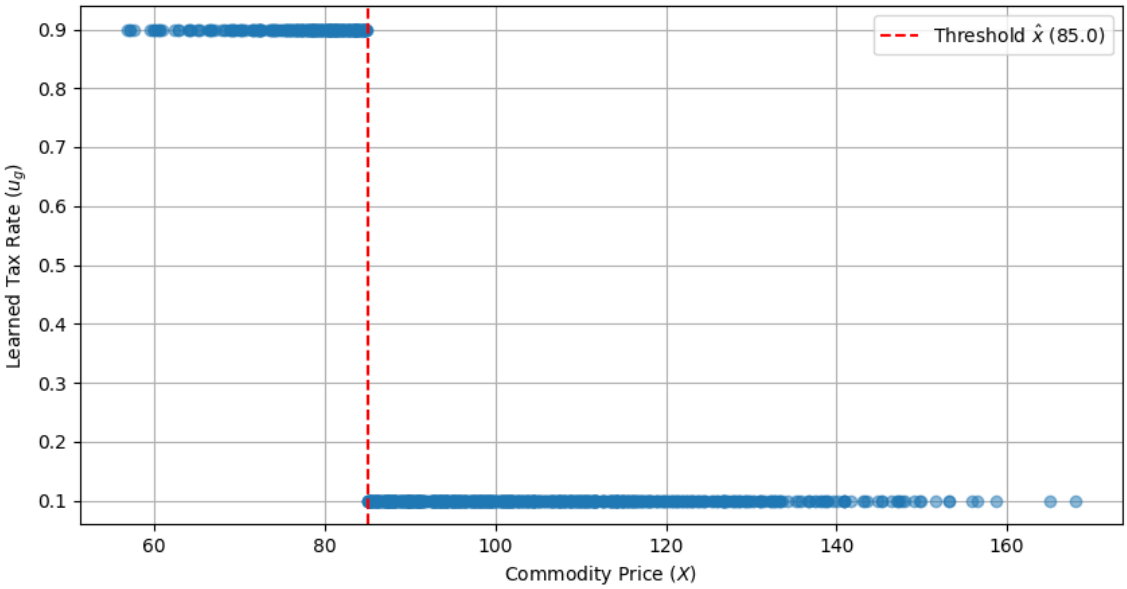

Figure 2: Emergence of Bang Bang Control Policy (Learned)

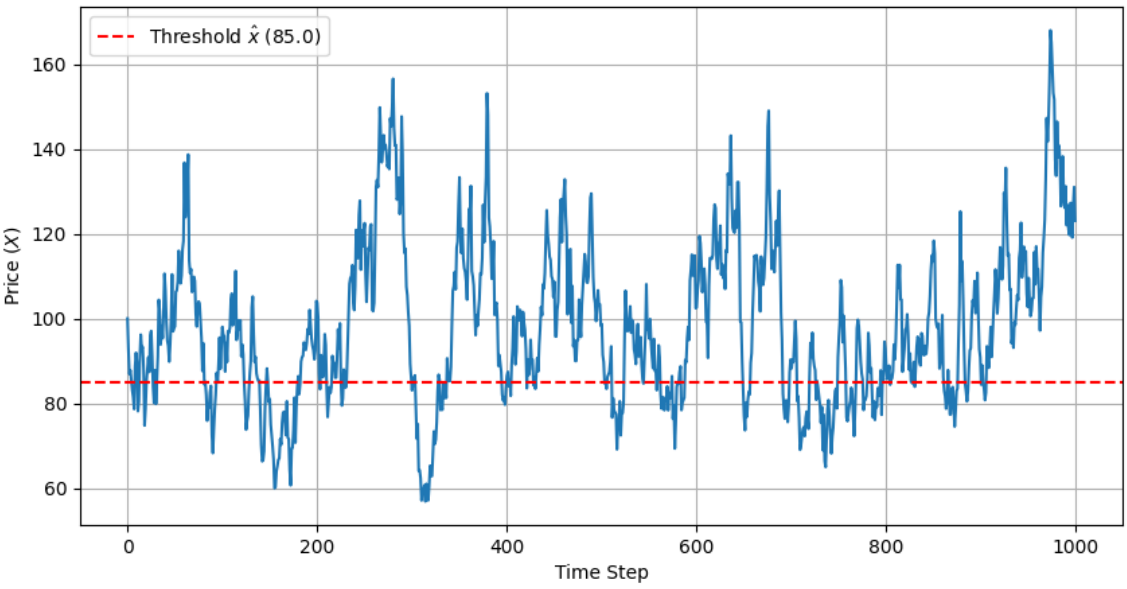

Figure 3: TITR Commodity Price Dynamics (Mean Reversion)

Future Work

While the current MADDPG framework successfully optimizes for profit and logistical speed, the future of the Trans-Caspian International Transport Route (TITR) depends on integrating sustainability metrics into the agent reward structures. To align with the European Union's Global Gateway initiative and the commitment to harmonize customs procedures and sustainable approaches in logistics (as of early 2026), future work will focus on expanding the differential game to include environmental variables. The integration of carbon pricing will necessitate augmenting the cost function C^k_{geo} to include a "Carbon Cost" term, which represents a function of the freight rate, representing speed and fuel efficiency, and the distance covered.

References

Bellman, R. E. 1963. A Bang-Bang Control Problem. *American Mathematical Society, Quarterly of Applied Mathematics*, 21(2): 159–161.

Frederic, D.; Huang, H.; and Mao, C. 2021. The Challenges Faced on Transit Transport Corridors by Landlocked Countries in Central Africa: Literature Review. *Open Journal of Applied Sciences*, 11(1).

Hazrat, A. 2025. Key Problems in the Development of Transport Corridors in Central Asia and Ways to Solve Them. *American Journal of Interdisciplinary Research and Development*, 44.

Kamdem, B. G.; and Jafari, N. 2026. A Differential Game of Profit-Sharing in Multinational Transport Corridors: Strategic Interactions between Governments and Transportation Firms. *Working Paper*.

Liang, J.; Miao, H.; Li, K.; Tan, J.; Wang, X.; Luo, R.; and Jiang, Y. 2025. A Review of Multi-Agent Reinforcement Learning Algorithms. *Electronics*, 14(4): 820.

Loganathan, K. A.; and Chinnaraju, A. 2026. Intelligent Multi-Agent Reinforcement Learning Architectures for Coordinated Autonomous Logistics and Real-Time Network Optimization. *International Journal of Research and Innovation in Social Science*, 10(1): 2666–2742.

Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; and Mordatch, I. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *Advances in Neural Information Processing Systems*, 30: 6379–6390.