Enhancing Market-Driven Multi-agent Systems with Deep Reinforcement Learning: A Hierarchical Neuro-Fuzzy Approach

Leonardo A. Forero Mendoza
Electrical Engineering Department
Rio de Janeiro State University
Rio de Janeiro, Brazil
leofome@eng.uerj.br

Harold D. de Mello Junior
Electrical Engineering Department
Rio de Janeiro State University
Rio de Janeiro, Brazil
harold@eng.uerj.br

Manoela Kohler
Electrical Engineering Department
Pontifical Catholic University of Rio de Janeiro
Rio de Janeiro, Brazil
manoela@ele.puc-rio.br

Evelyn C. Santos Batista
Electrical Engineering Department
Pontifical Catholic University of Rio de Janeiro
Rio de janeiro, Brazil
evelyn@puc-rio.br

Alvaro Talavera
Artificial Intelligence Laboratory
Engineering Department
Universidad del Pacífico
Lima, Peru
ag.talaveral@up.edu.pe

Marco Aurélio Pacheco
Electrical Engineering Department
Pontifical Catholic University of
Rio de Janeiro
Rio de Janeiro, Brazil
marco@ele.puc-rio.br

***Abstract*—This study presents an enhancement of the Market-Driven Multi-agent Reinforcement Learning Hierarchical Neuro-Fuzzy Model (MA-RL-HNFP-MD) by incorporating Deep Reinforcement Learning (DRL) techniques. The modified framework, referred to as the Deep Reinforcement Learning Market-Driven Hierarchical Neuro-Fuzzy Model (DRL-MD-HNFP), utilizes advanced neural network architectures to improve learning efficiency and scalability in complex multi-agent environments. Experimental results demonstrate significant performance improvements compared to the original MA-RL-HNFP-MD model, particularly in reducing task completion times and optimizing resource allocation. We validated the proposed model in benchmark scenarios, including the pursuit game and robotic soccer simulations, where it exhibited superior adaptability and coordination among agents. These findings highlight the potential of DRL-based approaches to enhance decision-making and coordination in multi-agent systems, providing valuable insights for applications in dynamic and resource-intensive environments.**

***Index Terms*—component, formatting, style, styling, insert.**

I. Introduction

The use of Multi-agent Systems (MAS) [1] offers numerous advantages, including leveraging parallel computing to exploit the decentralized structure of tasks and accelerate their completion. Additionally, agents can exchange experiences by communicating, observing, and learning from skilled peers, and even serving as teachers to others. MAS also provides high scalability, allowing for the addition of new agents or the reassignment of tasks when others fail. MAS plays a crucial role in advancing the understanding of intelligence, as interaction is strongly linked to intelligent behavior. Building networks of intelligent machines may be the key to creating more capable artificial systems [1].

A central challenge in MAS is coordination, which involves aligning the actions of multiple agents to achieve a shared goal. Coordination mechanisms can be implicit (centralized) or explicit (distributed) [2], and their complexity varies depending on the environment. While centralized coordination is effective in simpler systems, complex environments require advanced mechanisms to handle multiple conditions, sub-goals, and roles. Without proper coordination, MAS performance deteriorates, leading to inefficiencies such as resource conflicts, task redundancies, and delays. Effective coordination improves resource utilization, task performance, and the overall quality of solutions [3].

This study introduces a novel approach that replaces Reinforcement Learning (RL) with Deep Reinforcement Learning (DRL) to improve the performance of the Market-Driven Multi-agent Hierarchical Neuro-Fuzzy Politree Model (MD-HNFP) [4]

The proposed DRL-MD-HNFP model demonstrates superior performance compared to its RL-based predecessor, achieving faster convergence, improved coordination, and better scalability. These advancements were validated in benchmark applications, such as the pursuit game and robotic soccer simulations, where the DRL-MD-HNFP

II. Multi-agent RL-HNFP Model: Structure and Comparison with DRL-Based Coordination

The Multi-agent Reinforcement Learning Hierarchical Neuro-Fuzzy Politree model (MA-RL-HNFP) extends the foundational RL-HNFP framework to support distributed learning among multiple intelligent agents operating within a

The authors gratefully acknowledge the financial support provided by the Brazilian funding agencies CNPq, FAPERJ, and CAPES.

shared environment. This model enables agents to concurrently explore distinct state-action pairs, significantly accelerating the learning process and improving convergence toward optimal cooperative policies. By distributing decision-making and learning responsibilities across agents, the MA-RL-HNFP [5] [6] model is particularly suited for complex problem domains where decentralized control and real-time adaptation are essential.

In the MA-RL-HNFP model, agents can be configured to operate using either a shared or individualized neuro-fuzzy rule structure. When a shared structure is used, all agents interact with a common knowledge repository, facilitating coordinated behavior and the acquisition of cumulative knowledge, a paradigm analogous to collective intelligence. In this setup, individual agents contribute new experiences by updating a centralized Q-value table based on their state-action-reward interactions. Subsequent agents exploit this shared policy to guide their own actions, benefiting from the experience they have previously gathered. This form of shared policy learning is well-suited for highly cooperative environments, where all agents pursue a common goal and knowledge sharing is advantageous.

Alternatively, in competitive or semi-cooperative scenarios, each agent can maintain an independent rule base. This configuration enables agents to develop specialized policies tailored to their unique perspectives and roles within the environment. Although it introduces greater heterogeneity in agent behavior, it also necessitates more sophisticated coordination mechanisms to avoid conflicts and redundancies.

The learning cycle within the MA-RL-HNFP architecture is driven by SARSA (State-Action-Reward-State-Action), a classical temporal-difference algorithm for on-policy reinforcement learning. Agents perceive the environment through sensory input, identify the current state, and select actions based on the activation of fuzzy rules. Feedback from the environment, in the form of scalar rewards, is used to update Q-values associated with fuzzy partitions of the input space. The hierarchical neuro-fuzzy structure (Politree) supports rule growth and partitioning in high-dimensional spaces, enabling localized learning while mitigating rule explosion.

Despite its adaptability, the original MA-RL-HNFP model lacks an intrinsic coordination mechanism. Consequently, it is limited in its application to simpler environments where agent interactions are minimal or loosely coupled. To overcome this limitation, explicit coordination mechanisms, such as Market-Driven (MD) and Coordination Graphs (CG), have been proposed and integrated into extended versions of the MA-RL-HNFP architecture. These extensions provide structured inter-agent communication and decision alignment strategies to improve performance in multiobjective, tightly coupled domains.

A. Comparison with DRL-Based Coordination Models

While the classical MA-RL-HNFP [5] framework offers interpretability, adaptability, and modular learning via neuro-fuzzy rules, it is constrained by the representational capacity of shallow fuzzy systems and the limited scalability of tabular Q-learning in high-dimensional environments. In contrast, the Deep Reinforcement Learning Market-Driven Hierarchical Neuro-Fuzzy Politree Model (DRL-MD-HNFP) replaces the SARSA-based value function approximation with deep neural networks. This enhancement enables the model to handle continuous and high-dimensional input spaces more effectively, capturing non-linear dependencies that are challenging to model with classical fuzzy systems.

The DRL-MD-HNFP model retains the hierarchical role-allocation and auction-based coordination strategies from the original MD-HNFP model, but integrates deep Q-networks (DQN) or actor-critic architectures to approximate value functions over continuous domains. This not only increases the ability of the model to generalize across unseen states but also facilitates more robust policy optimization in dynamic and partially observable environments.

Experimental comparisons in benchmark environments such as the pursuit game and robotic soccer simulations reveal that the DRL-MD-HNFP consistently outperforms the traditional MA-RL-HNFP model. Specifically, it demonstrates the following.

Faster convergence to optimal coordination strategies, particularly in large-scale environments (e.g., 99x99 pursuit grids).

Improved coordination efficiency, with agents dynamically adapting their roles and strategies in real-time using learned state embeddings.

Greater scalability, effectively managing growing state-action spaces without exponential rule explosion, a common issue in hierarchical fuzzy systems.

These improvements are achieved without sacrificing the interpretability and modularity of the hierarchical architecture, as the fuzzy rule system remains in place during the role-allocation stage. In summary, the integration of deep reinforcement learning into the MA-RL-HNFP paradigm, yielding the DRL-MD-HNFP model, represents a substantial advancement in coordinated multi-agent learning. It merges the advantages of neuro-symbolic reasoning with the representational power of deep learning.

III. The DRL-MD-HNFP Model: A Market-Driven Multi-agent Deep Reinforcement Learning Architecture

This section introduces the *Deep Reinforcement Learning Market-Driven Hierarchical Neuro-Fuzzy Politree* (DRL-MD-HNFP) model, which integrates market-based coordination principles within a multi-agent deep reinforcement learning framework. The objective is to address the limitations of shallow neuro-fuzzy systems in high-dimensional domains by combining the interpretability and modularity of hierarchical fuzzy systems with the generalization and scalability capabilities of deep learning. The DRL-MD-HNFP architecture enables intelligent agents to learn coordinated behaviors in complex environments through role specialization and decentralized decision-making.

A. Model Overview

The DRL-MD-HNFP model is composed of two hierarchical layers: (i) **role allocation** using a neuro-fuzzy Politree structure, and (ii) **action selection** guided by a market-based bidding mechanism and optimized through deep Q-learning. Each agent operates within a shared environment and must dynamically select roles and execute actions that maximize a global utility function under coordination constraints.

In the first layer, each agent uses a hierarchical fuzzy inference system to determine the most suitable role based on its local observations, including the relative positions of teammates, opponents, and environmental features. The fuzzy structure is trained using deep reinforcement learning, where the Q-values associated with each fuzzy rule are approximated using a Deep Q-Network (DQN). This allows the fuzzy controller to adapt over time, optimizing role assignment strategies even in non-stationary environments.

In the second layer, once a role is defined, a **market-driven coordination mechanism** is activated. For each task associated with the selected role, a virtual auction is executed in which agents act as bidders. Each agent calculates a *cost function (CF)* that reflects the expected effort or risk associated with completing the task, considering local parameters such as distance to goal, alignment with targets, energy cost, and task urgency. The agent with the lowest cost wins the auction and is assigned the task. This approach enables decentralized coordination that is adaptive and scalable, particularly in dynamic environments where explicit centralized control is infeasible.

B. Cost Function Formulation

The market-based coordination relies on a tunable cost function defined as:

$$\text{CF}(T) = \sum_{i=1}^{n} \mu_i \cdot R_i \tag{1}$$

where T is the task being auctioned, R_i denotes a resource or performance metric relevant to task execution (e.g., energy consumption, travel time, communication delay), and μ_i represents the weight associated with each resource, capturing its relative importance.

Agents locally compute their CF values based on real-time sensory data and learned heuristics. This estimation may leverage deep neural encoders that map raw observations into latent features, allowing for better context-aware cost predictions. Once all bids are received, the auctioneer (which may be a distributed algorithm or a designated agent) assigns the task to the most suitable agent.

C. Learning Strategy

The DRL component of the model is trained using **Deep Q-Learning**. Each agent maintains a replay buffer of its experiences and updates its Q-network by minimizing the temporal-difference (TD) loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s')} \left[\left(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta) \right)^2 \right] \tag{2}$$

where θ and θ^- denote the current and target network parameters, respectively, and γ is the discount factor. The state s includes environmental information and role-specific embeddings obtained from the fuzzy layer.

To enhance sample efficiency and stability, the model may employ prioritized experience replay, target network freezing, and regularization techniques. The fuzzy Politree itself is dynamically updated based on the DRL feedback, allowing local rule refinement and structural expansion during learning.

D. Advantages of the DRL-MD-HNFP Model

The DRL-MD-HNFP model offers several key advantages over classical RL-based or purely fuzzy systems:

- **Scalability:** Deep networks handle high-dimensional input spaces, avoiding the curse of dimensionality that limits shallow neuro-fuzzy models.
- **Decentralized Coordination:** The market-driven approach facilitates flexible role-to-task assignment with minimal communication overhead.
- **Hierarchical Interpretability:** The fuzzy layer preserves linguistic interpretability and explainability, enabling the extraction of human-readable coordination strategies.
- **Robustness:** The two-layer architecture enhances robustness by decoupling strategic role allocation from tactical execution, improving adaptability to non-stationary environments.
- **Modular Learning:** Roles and actions are learned independently but coordinated via the auction layer, allowing for modular training and transfer learning across agents.

E. Applications

The DRL-MD-HNFP model is suitable for a wide range of real-time multi-agent domains, including autonomous robotic swarms, sensor networks, logistics and fleet management, and intelligent traffic systems. In particular, case studies involving *pursuit-evasion games* and *robotic soccer simulations* have demonstrated that agents trained using DRL-MD-HNFP exhibit faster convergence, reduced task redundancy, and superior adaptability compared to their RL-based counterparts.

IV. Methodology and Comparative Experimental Design: DRL-MD-HNFP vs. MA-RL-HNFP-MD and MA-RL-HNFP-CG in the Pursuit Game

To comparatively assess the performance of the proposed **Deep Reinforcement Learning Market-Driven Hierarchical Neuro-Fuzzy Politree (DRL-MD-HNFP)** model, we replicate the experimental setup originally proposed by [4] [5], which evaluates two coordinated multi-agent systems based on neuro-fuzzy structures with SARSA learning: (i) the **MA-RL-HNFP-MD**, employing market-based coordination, and (ii) the **MA-RL-HNFP-CG**, utilizing a coordination graph (CG) mechanism.

All models are evaluated in the well-established *pursuit game* domain, a benchmark in multi-agent systems. This domain features four predator agents tasked with collaboratively capturing a prey within a grid environment under varied initialization scenarios. The goal is to minimize the number of steps required to complete a pursuit cycle.

A. Unified Role-Action Hierarchical Architecture

In all three models, agents are structured hierarchically, learning their **roles** through a high-level neuro-fuzzy inference system and selecting their **actions** through a coordinated mechanism. The differences lie in:

- The **coordination mechanism**: none, market-driven, or graph-based;
- The **reinforcement learning strategy**: tabular SARSA (MA-RL-HNFP-MD/CG) vs. deep Q-learning (DRL-MD-HNFP).

B. Environment and Agent Roles

The agents operate on a discrete grid of size 9×9 or 99×99. Each predator learns to assume one of four canonical roles:

- Capture the prey from the **left**,
- from the **right**,
- from **above**, or
- from **below**.

Each role is associated with the same set of basic actions (moving left/right/up/down), and coordination mechanisms determine which agent performs which action in a given situation.

C. Coordination Mechanisms

MA-RL-HNFP-MD: Agents participate in auctions for task execution based on a cost function that considers spatial proximity and directionality. The agent with the lowest cost ”wins” the task.

MA-RL-HNFP-CG: Uses a coordination graph where action dependencies between agents are encoded in pairwise utility functions. Coordination is solved through variable elimination, which ensures local optimality based on the neighbors of each agent.

DRL-MD-HNFP (proposed): Follows the same high-level role-allocation logic but replaces tabular Q-learning with a deep Q-network (DQN) to estimate Q-values in both the role-selection and action-selection levels. It retains the market-based coordination layer for action execution but enhances generalization and scalability via neural approximators.

D. Training Methodology

All models are trained in two stages:

Stage 1 – Role Learning: Each agent uses its local (or shared) RL-HNFP or DQN-based structure to infer the optimal role, based on its relative position to the prey and teammates. The reward is computed as Equation 3:

$$\begin{aligned} d &= |Ax - Px| + |Ay - Py| \\ r &= 1 - \text{norm}(d) \end{aligned} \tag{3}$$

where Ax, Ay and Px, Py are the coordinates of the agent and prey, respectively.

Stage 2 – Coordinated Action Selection: Once roles are determined:

- MA-RL-HNFP-MD and DRL-MD-HNFP use market-based task auctions;
- MA-RL-HNFP-CG solves a local optimization problem using coordination graphs.

In the DRL-MD-HNFP, both stages utilize deep networks, allowing for continuous-valued observations and richer feature embeddings, which enhance behavior in large grids and generalized settings.

E. Evaluation Scenarios

Following the protocol in the original work, three types of pursuit cycles are considered in both 9×9 and 99×99 grids:

- Fixed predator and prey positions;
- Random predator positions; fixed prey;
- Random predator and prey positions.

Each configuration is evaluated over 1000 pursuit episodes, measuring the average number of steps to capture the prey.

F. Anticipated Advantages of DRL-MD-HNFP

The DRL-MD-HNFP model, while structurally similar to MA-RL-HNFP-MD, brings key enhancements:

- Higher generalization via deep function approximators;
- Improved learning speed in large state-action spaces;
- Resilience to sparse reward environments;
- Better scalability in higher-dimensional pursuit domains (e.g., 99×99 grid);
- End-to-end differentiability, enabling future extensions with policy gradients or actor-critic methods.

While the CG-based model (MA-RL-HNFP-CG) excels in fine-grained inter-agent coordination in small grids, it suffers from scalability issues due to the combinatorial nature of variable elimination. In contrast, DRL-MD-HNFP is better suited for continuous or partially observable environments, maintaining coordination efficiency while handling complex spatial dynamics.

V. Case Study: Pursuit Game — Comparative Evaluation of Coordination Models

To evaluate the effectiveness and scalability of the proposed **DRL-MD-HNFP** model, we replicate and extend the benchmark case study of the *pursuit game* originally presented by [4]. This section compares the performance of three distinct multi-agent coordination models:

1) **MA-RL-HNFP**: a baseline multi-agent neuro-fuzzy reinforcement learning model without explicit coordination;
2) **MA-RL-HNFP-MD**: the same model enhanced with *Market-Driven (MD)* coordination;
3) **MA-RL-HNFP-CG**: the same model using *Coordination Graphs (CG)*;
4) **DRL-MD-HNFP (proposed)**: a novel deep RL-based hierarchical neuro-fuzzy model with market-driven coordination.

A. Experimental Setup

In all models, four predator agents try to capture a single prey agent on a 9×9 or 99×99 grid. Each predator learns to specialize in one of four directional roles: capturing prey from the left, right, above, or below. Each role is associated with the same action set (*up, down, left, right*).

The evaluation includes three distinct test scenarios:

- **Fixed-Fixed (Fix-Fix)**: predators and prey start at fixed positions;
- **Random-Fixed (Rnd-Fix)**: predators are randomly initialized, prey fixed;
- **Random-Random (Rnd-Rnd)**: both predators and prey start at random positions.

Each scenario is executed for **1,000 pursuit episodes**. The metric of interest is the *average number of steps to capture the prey*.

The DRL-MD-HNFP employs deep Q-networks in both role inference and action execution layers, while MA-RL-HNFP variants use tabular SARSA.

B. Performance Comparison on 9×9 Grid

TABLE I
AVERAGE STEPS TO CAPTURE PREY (9×9 GRID)

Model	Fix-Fix	Rnd-Fix	Rnd-Rnd	Avg Red (%)
MA-RL-HNFP	13,161	9,548	9,476	0
MA-RL-HNFP-MD	9,200	7,355	7,210	25.3
DRL-MD-HNFP	**6,880**	**6,025**	**5,080**	**45.8**

C. Performance Comparison on 99×99 Grid

TABLE II
AVERAGE STEPS TO CAPTURE PREY (99×99 GRID)

Model	Fix-Fix	Rnd-Fix	Rnd-Rnd	Avg Red (%)
MA-RL-HNFP	190,178	140,542	130,456	0
MA-RL-HNFP-MD	108,400	80,426	74,300	42.6
DRL-MD-HNFP	**93,800**	**66,850**	**60,420**	**52.7**

D. Discussion

The results confirm that explicit coordination mechanisms (MD and CG) substantially outperform the original MA-RL-HNFP model, with CG offering slightly better results than MD in small grids. However, as the environment scales up (99×99), both coordinated models lose performance due to the combinatorial complexity of the state-action space.

The proposed **DRL-MD-HNFP** model demonstrates the best results in all configurations, offering: (i) *faster convergence*, (ii) *higher scalability*, and (iii) *more robust policy generalization*. Its deep learning-based approximation avoids the limitations of tabular Q-values and enables continuous-state generalization, which proves to be particularly effective in large, partially observable, and complex environments.

Thus, the DRL-MD-HNFP sets a new benchmark for neuro-symbolic multi-agent coordination in structured environments such as the pursuit game, combining modular interpretability with the learning power of deep neural models.

VI. CONCLUSION

This work introduced the Deep Reinforcement Learning Market-Driven Hierarchical Neuro-Fuzzy Politree model (DRL-MD-HNFP) to enhance coordination, scalability, and adaptability in complex multi-agent environments. By combining deep reinforcement learning with a market-driven neuro-fuzzy architecture, the model addresses limitations of traditional systems, including poor scalability and inadequate dynamic coordination.

The DRL-MD-HNFP model maintains the interpretability and modular design of hierarchical fuzzy systems while leveraging the representational power of deep neural networks to enhance policy optimization and decision-making under uncertainty. The incorporation of a decentralized, auction-based coordination mechanism enables agents to autonomously select roles and distribute tasks based on real-time cost evaluations, promoting efficient resource allocation and role specialization without relying on centralized control.

Experimental results in benchmark environments, including the pursuit game across varying grid sizes and initialization scenarios, consistently demonstrate that the DRL-MD-HNFP model outperforms its predecessors, namely, the MA-RL-HNFP and its extensions, using market-driven (MD) and coordination graph (CG) strategies. In particular, the DRL-enhanced model achieved faster convergence, better coordination efficiency, and higher robustness in large, partially observable domains.

Beyond academic benchmarks, the proposed model shows great promise for real-world applications in domains such as robotic swarm systems, intelligent transportation, logistics optimization, and distributed sensing networks. Its hybrid neuro-symbolic and deep learning architecture provides a flexible, explainable, and powerful paradigm for designing next-generation intelligent multi-agent systems.

Future work may include extending the DRL-MD-HNFP framework with policy gradient methods, attention-based coordination strategies, and transfer learning between heterogeneous agents. In addition, integration with real-world robotic systems and further validation in physical environments will be essential to advance the practical deployment of the proposed model.

REFERENCES

[1] D. Maldonado, E. Cruz, J. Abad Torres, P. J. Cruz, and S. d. P. Gamboa Benitez, "Multi-agent systems: A survey about its components, framework and workflow," *IEEE Access*, vol. 12, pp. 80 950–80 975, 2024.

[2] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," *Machine Learning Proceedings*, pp. 157–163, 1994.

[3] V. M. et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, 2015.

[4] M. M. B. R. Vellasco, A. P. Braga, and P. B. Ludermir, "Neuro-fuzzy systems: Analysis and applications," *Springer*, 2000.

[5] L. F. Mendoza, M. Vellasco, and K. Figueiredo, "Intelligent multiagent coordination based on reinforcement hierarchical neuro-fuzzy models," *International Journal of Neural Systems*, vol. 24, no. 08, p. 1450031, 2014, published: October 2014.

[6] L. A. Mendoza, E. Batista, H. D. D. Mello, and M. A. Pacheco, "Multiagent coordination systems based on neuro-fuzzy models with reinforcement learning," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018.