20th IMEKO TC10 Conference
28-29 September 2026 | Lisbon | Portugal

20th IMEKO TC10 Conference
"Measurement for Diagnostics, Optimization and Control to Support Competitiveness and Innovation"
Lisbon, Portugal, September 28–29, 2026

Single-Agent Reinforcement Learning in Multi-Agent Production Environments: Effects of Observation Overlap

Bence Pálvölgyi[1], Zsolt János Viharos[1,2]

[1]*Research Laboratory on Engineering and Management Intelligence of the HUN-REN Institute for Computer Science and Control (SZTAKI), Center of Excellence of the Hungarian Academy of Sciences (MTA),*
Budapest, 1111, Hungary,
palvolgyi.bence@sztaki.hu, viharos.zsolt@sztaki.hu
[2]*Faculty of Economics and Business of the John von Neumann University,*
Kecskemét, 6000, Hungary

Abstract **– Reinforcement Learning (RL) has demonstrated strong performance in production and manufacturing systems, where decision processes can be modelled as single-agent environments with well-defined state and action spaces. However, more complex real-world production processes inherently involve multiple acting entities, making Multi-Agent Reinforcement Learning (MARL) a natural tool for optimization. While similar to RL, MARL introduces new challenges, including non-stationarity, scalability issues, and increased training complexity.**
This paper investigates an alternative approach that employs a single-agent RL framework within a multi-agent environment. Instead of learning multiple policies, a single agent is trained to sequentially control multiple entities using their combined state information. The study explores varying degrees of observation-space overlap between the controlled entities, analyzing how shared and independent components influence learning efficiency and policy quality. By comparing these configurations with a MARL formulation in a controlled test environment, the work provides insights into when a single-agent abstraction can effectively represent a multi-agent problem. The results show that MARL provides faster learning for disjoint observations, while increasing observation overlap progressively favors the single-agent formulation, with complete overlap resulting in faster convergence. The observed behavior is further related to an application-oriented production-control environment with strongly overlapping observations. These results highlight observation-space structure as an important consideration when selecting between single-agent and multi-agent RL formulations for complex production systems.

Keywords – IMEKO, TC10, Diagnostics, Optimisation, Control, Multi-Agent Reinforcement Learning, Industry, Innovation and Infrastructure.*

I. INTRODUCTION

Reinforcement Learning (RL), typically formulated as a Markov Decision Process, has been widely used for simulating and optimizing production and manufacturing systems, where sequential decision-making can be learned directly from interaction with a modelled environment. In such settings, single-agent RL provides a relatively straightforward framework, assuming a centralized controller operating over a unified state and action space.

However, more complex real-world production environments inherently consist of multiple interacting entities, such as machines, transport units, and processing stations, whose behaviors are interdependent. These systems are more naturally described within the framework of Multi-Agent Reinforcement Learning (MARL). Unlike single-agent RL, MARL problems, among others, require careful categorization along several dimensions, as stated in the paper written by Wong A. et al. [1]. Environments can be cooperative, collaborative, competitive, or mixed depending on the reward structure, while interaction patterns may follow agent-environment cycles (turn-based execution) or fully parallel action schemes, shown in Fig. 1. Additionally, agents themselves may be homogeneous, as sharing identical capabilities and policies, or heterogeneous, with distinct roles and action spaces.

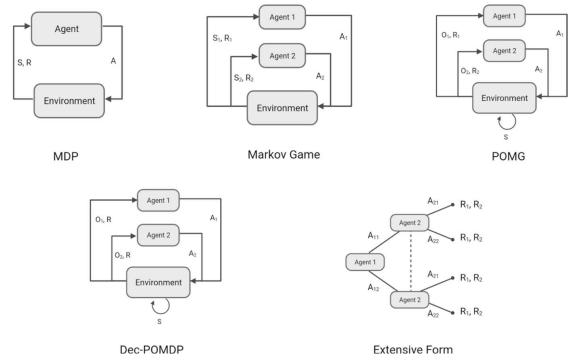

Fig. 1. Environment types (Wong et al. [1]), in order from top left to bottom right: Single Agent, Multi-Agent Environment Cycle, Parallel Environment, Parallel Environment Joint Reward, Extensive Form

From a learning perspective, MARL introduces further design choices regarding control and training structure. Agents may operate in fully decentralized settings, under centralized control, or within a hybrid paradigm such as Centralized Training with Decentralized Execution (CTDE) [5]. Techniques such as parameter sharing are often employed in homogeneous settings to improve scalability and sample efficiency [6], while heterogeneous systems typically require more specialized representations. However, the presence of multiple acting entities does not necessarily imply that separate policies are required. When entities operate sequentially and their observations share common information, their control problem may also be represented using a single policy operating on a combined state representation.

This paper investigates this alternative formulation, with particular focus on the degree of overlap between the observation spaces of multiple acting entities. A controlled two-entity environment is introduced in which observation overlap can be systematically varied from completely disjoint to fully overlapping while preserving the underlying task and sequential interaction structure. A MARL formulation using separate policies is compared with a single-agent formulation in which one policy sequentially controls both entities using their combined observations and an indicator of the currently active entity. The controlled analysis is complemented by an application-oriented production environment characterized by strongly overlapping observations. The objective is to determine how observation-space structure affects the relative learning behavior of single-agent and multi-agent formulations and whether observation overlap can provide a useful consideration when selecting between them.

II. RELATED RESULTS IN THE LITERATURE

The increasing complexity of Multi-Agent Reinforcement Learning (MARL) has led to the development of structured frameworks aimed at better understanding and evaluating algorithm performance in dynamic environments. The work done by Hu, S. et al. [2] introduces adaptability as a central concept for assessing MARL systems under changing conditions. It highlights that real-world multi-agent systems are subject to variability in agent populations, task objectives, and execution environments, which limits the direct applicability of MARL methods developed for static benchmarks. The proposed framework distinguishes between learning adaptability, policy adaptability, and scenario-driven adaptability, emphasizing the importance of robustness beyond controlled experimental setups. This perspective is particularly relevant for production-like environments, where system conditions evolve continuously and require stable coordination strategies.

From an algorithmic standpoint, a survey paper by Liang J. et al. [3] provides a comprehensive categorization of MARL methods based on both learning paradigms and reward structures. Building on the transition from single-agent Markov Decision Process to multi-agent Markov games, the survey groups algorithms into value-based, policy-based, and actor-critic approaches, while further distinguishing between cooperative, competitive, and mixed settings. It also identifies core challenges inherent to MARL, including high dimensionality, non-stationarity, partial observability, and scalability. Representative methods such as Proximal Policy Optimization (PPO) and Multi-Agent Deep Deterministic Policy Gradient (MADDPG) are discussed in terms of their contributions and limitations, providing a structured overview of the design space relevant to multi-agent learning.

Complementing algorithmic analyses, Zepeng et al. [4] highlight the breadth of MARL applications across domains such as autonomous vehicles, energy systems, scheduling, and resource management. The survey emphasizes that, despite strong performance in simulated benchmarks, practical deployment remains constrained by challenges such as hyperparameter sensitivity and computational complexity. It also reviews benchmark environments and task formulations, demonstrating the adaptability of MARL to diverse real-world problems, while underlining the need for scalable and robust training methodologies.

Specific architectural approaches further refine the design of MARL systems. The concept of centralized training with decentralized execution (CTDE), as discussed by Amato et al. [5], allows agents to leverage global information during training while maintaining decentralized policies at execution time. This paradigm addresses coordination challenges without violating practical constraints of distributed systems. In homogeneous agent settings, parameter sharing has been shown to significantly improve scalability and sample efficiency. The work "Parameter Sharing Deep Deterministic Policy Gradient for Cooperative Multiagent

Reinforcement Learning" demonstrates that sharing actor and/or critic networks enables faster learning and reduced memory requirements, particularly when agents are exchangeable and operate under shared reward structures.

The distinction between homogeneous and heterogeneous agents introduces further complexity. In homogeneous settings, studies such as seen in the paper done by Chu, X. et al. [6] show that cooperative behavior can emerge from shared policies with minimal communication, and that local perception and limited interaction can sometimes outperform fully informed strategies. In contrast, heterogeneous agent systems require more expressive models to capture differing roles and capabilities.

Kent et al. [7] address this by proposing algorithms with theoretical guarantees, such as HATRPO and HAPPO, which ensure stable learning and convergence in heterogeneous environments. These approaches demonstrate improved coordination and performance compared to standard MARL baselines, highlighting the importance of explicitly modelling agent diversity.

Overall, the literature indicates that while MARL provides a powerful framework for modelling complex multi-agent systems, its practical application is constrained by scalability, stability, and adaptability challenges. These limitations motivate alternative formulations, such as single-agent control of multi-entity systems with structured comparisons between homogeneous and heterogeneous agent configurations, as investigated in this work.

III. DESCRIPTION OF THE METHOD

The investigated problem originates from a production-control environment in which multiple acting entities operate on partially or fully overlapping state information. Such systems can be formulated either as multi-agent reinforcement learning problems, where individual policies control the acting entities, or as single-agent problems, where a common policy sequentially controls them. The suitability of these formulations may depend on the structural relationship between the information available to the individual acting entities.

To analyze this relationship independently of the complexity of the production process, a simplified discrete environment was constructed. The environment preserves the essential characteristics relevant to the comparison: multiple acting entities, sequential decision making, and some degree of overlap between their observation spaces. Other multi-agent effects, including competition, resource sharing, and direct interaction between agents, were intentionally excluded. This allows the influence of observation-space overlap on the SARL and MARL formulations to be examined in isolation.

A. ENVIRONMENT

The environment consists of two agents, denoted A_0 and A_1. Each agent operates in a discrete three-dimensional space (Fig. 2). Every dimension can take one of three possible values *{0, 1, 2}*, resulting in $3^3 = 27$ possible positions for each agent.

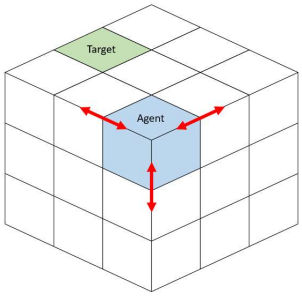

Fig. 2. Three-dimensional discrete GridWorld used for an individual agent. Each dimension contains three possible positions. The agent can move by one position in either direction or remain stationary along each dimension. A reward is received when the agent reaches the target position.

Each agent has an individual target defined in the same three-dimensional space. At initialization, the agent position and its corresponding target position are selected randomly. The objective of each agent is to navigate from its current position to its target.

The agents operate sequentially following an Agent Environment Cycle (AEC). At each environment step, only one of the two agents is active. After the active agent performs an action, control is transferred to the other agent.

The two agents are independent, as an action performed by one agent does not modify the state, available actions, target, or reward of the other or the environment.

Each episode consists of 100 environment steps, providing 50 action opportunities for each agent.

An action specifies movement independently along each of the three dimensions. For each dimension, three alternatives are available: decrease the coordinate by one, leave it unchanged, or increase it by one. The action can therefore be represented as *{-1, 0, 1}*.

Consequently, the multidimensional action space contains $3^3 = 27$ possible action combinations. Simultaneous movement along multiple dimensions is permitted. Coordinates are constrained to the boundaries of the discrete environment.

This action definition means that the minimum number of actions required to reach a target corresponds to the maximum coordinate-wise distance between the current position and the target.

A reward of *+1* is assigned when the currently active agent reaches its target position, after which both the position of that agent and its target are independently randomized, creating a new navigation task.

No positive or negative reward is obtained otherwise. This mechanism allows multiple targets to be reached within one episode and makes cumulative episode reward a direct measure of task performance.

Because the grid is small and the transition dynamics are deterministic, the expected performance of an optimal controller can also be determined analytically. For uniformly sampled starting and target positions, the expected optimal cumulative reward over 100 environment steps is approximately 65.14. This value provides a reference against which the learned policies can be evaluated.

B. OBSERVATION-SPACE OVERLAP

The central experimental variable is the number of dimensions shared between the observation spaces associated with the two agents.

Each individual agent task contains three dimensions. Four configurations are investigated, corresponding to zero, one, two, and three overlapping dimensions (Fig. 3).

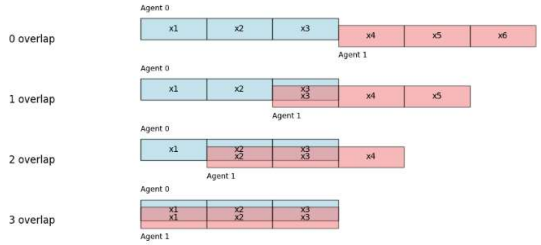

Fig. 3. Observation-space overlap configurations of the theoretical environment. Each agent has three observation dimensions. From top to bottom, the number of shared dimensions increases from zero to three. Overlapping cells represent state dimensions that are common to both agents, while non-overlapping cells represent agent-specific information.

It is important to distinguish observation-space overlap from interaction between the agents. A shared dimension indicates that the corresponding environmental quantity can be represented by the same variable; it does **not** imply that one agent can modify the state of the other. The agents remain dynamically independent in all four configurations.

The degree of overlap therefore determines the amount of redundancy present when the two individual observation spaces are combined into a centralized representation. Increasing overlap reduces the dimensionality of their union from six unique dimensions in the disjoint configuration to three unique dimensions under complete overlap.

C. MULTI-AGENT FORMULATION

In the MARL formulation, the environment is represented as an AEC multi-agent environment containing two independent agents. Each agent receives the observation associated with its own task and selects an action only when it becomes the active agent.

The policy of A_0 therefore receives observation O_0, while the policy of A_1 receives observation O_1. The observation of each agent is a subset of the environment state, that only excites on of the agents.

The two policies solve their respective navigation tasks using local observations. Importantly, increasing the overlap between the underlying dimensions does not provide either MARL policy with additional observations, as each continues to operate on its corresponding three-dimensional task representation. The overlap instead describes the structural relationship between the two observation spaces.

This formulation provides a single multi-agent reference against which the centralized single-agent representations are compared.

D. SINGLE-AGENT FORMULATION

To investigate whether the same sequential multi-agent task can be represented efficiently using a single policy, an equivalent single-agent environment was constructed.

The physical task, transition rules, reward function, episode length, action space, and alternating order of the two agents are preserved. However, instead of assigning an independent policy to each agent, one policy controls both.

At every step, the single policy receives the union of the observations corresponding to A_0 and A_1, together with an additional binary variable, agent selection, which identifies the agent currently being controlled. The selected action is then applied exclusively to that agent. After the transition, the selection variable switches and the policy controls the other agent during the following step.

The SARL observation can therefore be expressed as the union of O_0, O_1, agent-selection.

The observation dimensionality consequently depends on the degree of overlap. Ignoring the position/target duplication within each environmental dimension for the moment, the number of unique environmental dimensions available to the centralized policy is showh inFig 3.

This creates a controlled transition from a centralized policy that must distinguish between two largely separate state representations to one in which both agents operate over a joint representation. The comparison therefore evaluates whether the effectiveness of the single-policy formulation depends on the structural similarity between the tasks assigned to the logical agents.

E. TRAINING AND EVALUATION

Both formulations were trained inside a Farama Foundation environment, Gymnasium [8] for the single-agent environment and PettingZoo [9] for the multi-agent. This is a popular, standardized python package for Reinforcement Learning implementation. The algorithms used on these was the Proximal Policy Optimization (PPO) algorithm, implemented in Stable Baselines3 [10] for the single-agent and in Rllib [11] for the multi-agent solution. The same environment dynamics, reward structure, action possibilities, episode length, and comparable PPO hyperparameters were maintained across the experiments to ensure that differences in learning behavior could primarily be attributed to the policy and observation-space

formulations.

Four SARL configurations were evaluated, corresponding to zero, one, two, and three overlapping dimensions. Their learning behavior was compared with the MARL formulation.

The primary performance metric was the mean cumulative reward obtained during an episode. Since each successful target acquisition produces one reward, this metric directly represents the number of targets reached within the fixed 100-step horizon. The analytically determined optimal expected reward of approximately 65.14 provides an additional reference for evaluating convergence.

In addition to the controlled environment, the comparison was related to the production-control problem motivating the study. The production environment contains two acting entities that sequentially perform material-handling decisions within the same production system. In contrast to the synthetic environment, its dynamics include processing operations, product flow, feasibility constraints, and substantially larger state and action representations. The acting entities operate on nearly 100% overlapping information, making the environment structurally comparable to the high-overlap case of the proposed theoretical model.

The production experiment is therefore not treated as an additional point in the synthetic overlap series. Instead, it is used as an application-oriented case to examine whether the learning behavior identified in the controlled analysis is also observable in a substantially more complex environment.

IV. RESULTS AND DISCUSSIONS

The learning curves of the investigated configurations are presented in Fig. 4. For the theoretical environment, an optimal policy can achieve an expected mean reward of approximately 65.14 within the 100-step episode. All investigated SARL and MARL configurations eventually approach this value, indicating that the different observation structures primarily influence learning efficiency rather than the final achievable performance.

The MARL formulation's convergence is considerably faster than the SARL configurations with zero, one, and initially two overlapping dimensions. However, this advantage decreases as the observation spaces become increasingly similar. With complete overlap, the SARL formulation reaches high performance earlier than MARL. The resulting progression suggests that the relative advantage of decomposing the problem into multiple policies depends on the amount of independent information associated with the acting entities.

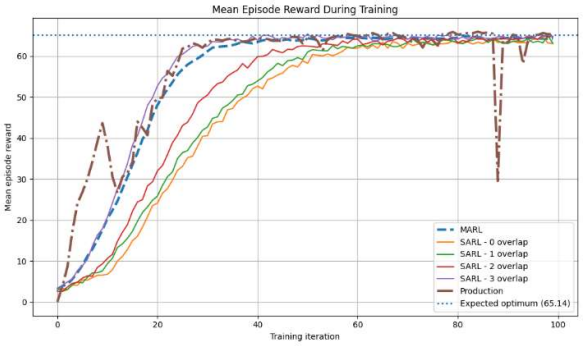

Fig. 4. Mean episode reward during training for the MARL and SARL formulations with zero to three overlapping observation dimensions. The dotted horizontal line indicates the theoretical expected optimum of 65.14 for the GridWorld environment. The production-control environment is included as an application-oriented example corresponding to a fully overlapping observation structure.

A clear relationship can be observed between the degree of observation-space overlap and the convergence rate of the SARL formulation. With completely disjoint observation spaces (0 overlap), the single-agent solution exhibits the slowest learning, requiring approximately 60 training iterations to approach its stable performance. Introducing one overlapping dimension results in a moderate improvement, whereas two overlapping dimensions substantially accelerate convergence. With complete observation overlap, the SARL formulation converges fastest, reaching a mean reward close to 60 after approximately 30 iterations and subsequently approaching the theoretical optimum.

This behavior can be explained by the different representations of the same underlying decision problem. With disjoint observations, the centralized SARL policy must process the union of two distinct observation spaces and use the agent-selection variable to determine which information is relevant to the current decision. In contrast, the MARL formulation separates the problem into two smaller agent-specific policies. Increasing observation overlap reduces the number of unique variables in the centralized representation and increases the structural similarity of the decisions made for the two entities. Under complete overlap, a single policy can therefore reuse the same representation for both logical agents, reducing the disadvantage associated with centralized control. The results consequently indicate a transition from a setting favoring separate policies towards one favoring a shared centralized representation as observation overlap increases.

The production-control result shown provides an application-oriented example of the latter case. The production environment contains two acting entities with strongly overlapping observations and is controlled sequentially by a single policy. Despite the considerably more complex production dynamics, its normalized

learning curve shows a convergence trend similar to the fully overlapping theoretical SARL configuration, reaching its high-performance region within approximately the same stage of training. The production curve exhibits substantially greater variability, including isolated performance decreases after convergence, reflecting the increased complexity and variability of the application environment compared with the deterministic theoretical model.

The production experiment is not intended as a direct quantitative validation of the theoretical GridWorld, since the environments differ substantially in their dynamics, reward definitions, and complexity. Instead, it demonstrates that the structural condition examined by the theoretical model, multiple acting entities with highly overlapping observations, also occurs in an application-oriented production problem. The combined results suggest that observation-space overlap is a relevant structural property when deciding whether a sequential control problem should be decomposed into multiple RL policies or represented using a single centralized policy.

V. CONCLUSIONS AND OUTLOOK

This work investigated the effect of observation-space overlap on the suitability of single-agent and multi-agent reinforcement learning formulations. A controlled environment was introduced in which two independent acting entities operate sequentially, allowing the degree of overlap between their observation spaces to be varied while keeping the underlying task and dynamics unchanged. The multi-agent formulation, using separate policies and local observations, was compared with a centralized single-agent formulation in which one policy sequentially controls both entities using their combined observation and an agent-selection variable.

The results show that both approaches are capable of reaching near-optimal performance, while their learning rates depend considerably on the structure of the observation spaces. MARL provides an advantage when the observations are disjoint or weakly overlapping, whereas increasing overlap progressively improves the convergence of the centralized SARL formulation. With complete overlap, SARL reaches near-optimal performance slightly faster than MARL. These findings indicate that the degree of observation overlap can be an important consideration when deciding whether multiple acting entities should be represented by separate policies or incorporated into a common policy.

The investigated structure is also present in the application-oriented production-control environment, where multiple acting entities operate using strongly overlapping information. Despite its substantially greater complexity and variability, the production experiment exhibits learning behavior consistent with the fully overlapping synthetic case. This provides an example of how the structural properties investigated in the controlled environment can occur in practical control problems.

Future work will extend the analysis to larger state and action spaces, different numbers of agents, and intermediate degrees and structures of observation overlap. Of particular interest is the introduction of dependencies between agents, including shared resources, constraints, and mutually influenced state transitions. Such extensions could determine whether observation overlap remains a useful criteria for selecting between SARL and MARL formulations when moving from independent acting entities toward genuinely cooperative or competitive multi-agent systems.

VI. ACKNOWLEDGMENTS

The research was supported by the European Commission (EC) through the DiGreeS project under grant No. 101178079 and by AI EDIH Hungary project under EC grant No. 101120929.

Special thank has to be expressed for Jenő Csanaki, Krisztián Meskó, Zsolt Nagy at the Opel Szentgotthárd Kft. for their continuous collaboration.

On behalf of Project "Comprehensive testing of machine-learning algorithms" we thank for the usage of HUN-REN Cloud (https://science-cloud.hu/) that significantly helped us achieving the results published in this paper.

REFERENCES

[1] **Wong, A., Bäck, T., Kononova, A.V. et al.**: Deep multiagent reinforcement learning: challenges and directions. Artif Intell Rev 56, 5023–5056, 2023. https://doi.org/10.1007/s10462-022-10299-x

[2] **Hu, S., Hady, M. A., Qiao, J., Cao, J., Pratama, M., & Kowalczyk, R.** Adaptability in Multi-Agent Reinforcement Learning: A Framework and Unified Review. (2025). arXiv preprint arXiv:2507.10142.

[3] **Liang J, Miao H, Li K, Tan J, Wang X, Luo R, Jiang Y.** A Review of Multi-Agent Reinforcement Learning Algorithms. *Electronics*. 2025; 14(4):820. https://doi.org/10.3390/electronics14040820

[4] **Zepeng, N., Lihua, X.,**: A survey on multi-agent reinforcement learning and its application, 2024, *Journal of Automation and Intelligence*, Volume 3, Issue 2.

[5] **Amato, C.:** An introduction to centralized training for decentralized execution in cooperative multi-agent reinforcement learning. 2024. arXiv preprint arXiv:2409.03052.

[6] **Chu, X., & Ye, H.**: Parameter sharing deep deterministic policy gradient for cooperative multi-agent reinforcement learning. 2017. arXiv preprint arXiv:1710.00336.

[7] **Kent T, Richards A, Johnson A.**: Homogeneous Agent Behaviours for the Multi-Agent Simultaneous Searching and Routing Problem. Drones. 2022; 6(2):51. https://doi.org/10.3390/drones6020051.

[8] https://gymnasium.farama.org/

[9] https://pettingzoo.farama.org/index.html

[10] https://stable-baselines3.readthedocs.io/en/master/

[11] https://docs.ray.io/en/latest/rllib/index.html