MEASE: Multi-agent Episodic Action Sequence Explanation

Khaing Phyo Wai *
Singapore Management University
Singapore
khaingpw@smu.edu.sg

Minghong Geng *
Singapore Management University
Singapore
mhgeng@smu.edu.sg

Shubham Pateria
Singapore Management University
Singapore
shubhamp@smu.edu.sg

Budhitama Subagdja
Singapore Management University
Singapore
budhitamas@smu.edu.sg

Ah-Hwee Tan
Singapore Management University
Singapore
ahtan@smu.edu.sg

ABSTRACT

Multi-agent reinforcement learning (MARL) achieves remarkable performance in complex coordination tasks, yet interpreting the emergent behaviors of trained agents remains a fundamental challenge. Most current explainability methods focus on individual agent decisions, overlooking the critical interplay of joint strategies and temporal coordination patterns that define successful multi-agent policies. We present MEASE (Multi-agent Episodic Action Sequence Explanation), a novel explainable MARL (XMARL) framework that explains trained MARL policies as human-interpretable emergent cooperative joint behaviors. MEASE employs a cognition-inspired episodic memory model to learn spatio-temporal multi-agent interaction patterns, coupled with abstraction algorithms that identify significant cooperative agent behaviors. We evaluate MEASE on diverse scenarios in the VMAS and MOSMAC environments, demonstrating its generalizability across various tasks and domains. These explanations, which prescribe "when to do what" for multi-agent systems, serve as executable coordination protocols that faithfully capture the learned behaviors. Quantitative validation shows that deploying explanations as strategies achieves 93% of the original MARL policy performance. A user study with 31 participants validates the clarity and usefulness of the explanations. These results demonstrate that MEASE effectively extracts explanatory knowledge from complex multi-agent behaviors.

KEYWORDS

Explainable Multi-agent Reinforcement Learning; Multi-agent Reinforcement Learning; Sequential Decision-making

ACM Reference Format:
Khaing Phyo Wai, Minghong Geng, Shubham Pateria, Budhitama Subagdja, and Ah-Hwee Tan . 2026. MEASE: Multi-agent Episodic Action Sequence Explanation. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 10 pages. https://doi.org/10.65109/OUUG8445

*These authors contributed equally to this work.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/OUUG8445

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) has demonstrated remarkable success across diverse domains, from real-time strategy (RTS) games [34, 45] to real-world applications including robotics [15, 18, 30], security [28, 29], urban traffic control [54], and smart cities [23, 32]. However, despite agents learning emergent coordinated behaviors, their policies generally lack interpretability, functioning as *black-box* systems [13, 46, 47]. This opacity poses a critical barrier to deployment in high-stakes applications where understanding and trust in AI decision-making are essential.

While the *explainable reinforcement learning* (XRL) domain has made significant progress, existing post-hoc methods predominantly focus on explaining actions or decisions made by individual agents [1, 43], failing to capture the temporal coordination patterns and emergent behaviors that characterize successful multi-agent policies. This gap is particularly problematic because the value of MARL lies precisely in the complex coordination strategies that emerge from agent interactions — strategies that cannot be understood by examining individual decisions in isolation.

To address this research gap, we introduce *Multi-agent Episodic Action Sequence Explanation* (MEASE), a novel explainable MARL (XMARL) framework that explains MARL policies by learning *episodic memories* from multi-agent behavioral data and explaining them as interpretable symbolic sequential coordination behaviors. At its core, MEASE employs a cognitive-inspired episodic memory model to learn behavioral patterns from MARL agent trajectories. Unlike traditional black-box approaches [13], the memory models based on self-organizing neural networks [42] explicitly store behavioral patterns as interpretable *cognitive codes*, enabling systematic retrieval and analysis [11, 46, 47]. Specifically, we introduce a novel *interval-based memory retrieval* mechanism that retrieves episodic memories as generalized action patterns across arbitrary timestep windows, capturing coordination behaviors at multiple temporal scales. These patterns are then processed through our *Multi-agent Action Sequence Abstraction* (MASA) algorithm, which identifies significant joint-actions and consolidates repetitive patterns for concise symbolic explanations of agent behaviors.

We evaluate MEASE across diverse multi-agent scenarios from the VMAS [3] and MOSMAC [12] benchmarks, demonstrating its ability to explain agent behaviors across different task complexities and coordination requirements. Through an *explanation-as-strategy* validation approach, our results show that explanations not only

provide a clear understanding of agent behaviors but can also be deployed as executable strategies, achieving comparable performance to the original MARL policy (93.2% vs. 93.7% win rate), demonstrating that MEASE successfully captures the essential coordination patterns embedded in trained policies. A user study with 31 participants further reveals that MEASE generates comprehensible explanations for complex multi-agent cooperative tasks. In summary, the main contributions of this work include:

- We present MEASE, a novel XMARL framework that explains MARL policies by learning multi-agent trajectories as episodic memory and representing them in symbolic form, highlighting emergent coordination patterns.
- We develop EM-ART 2, a new episodic memory model tailored for encoding multi-agent behaviors and MASA, an abstraction algorithm for effective memory retrieval.
- Extensive experiments across VMAS and MOSMAC scenarios demonstrate MEASE accurately distills essential coordination patterns, and a user study shows MEASE provides clear, useful, and comprehensible explanations for complex multi-agent behaviors.

2 RELATED WORK

2.1 Multi-Agent Reinforcement Learning

MARL has made substantial progress in both on-policy and off-policy algorithms in recent years. Representative on-policy methods such as COMA [9], MAA2C [31], and MAPPO [53], along with off-policy methods including MADDPG [24], VDN [40], and MAVEN [27], have demonstrated strong performance on cooperative tasks. Among various MARL approaches, value decomposition methods have gained attention due to their sample efficiency. QMIX [33] introduced monotonic value function factorization to ensure consistency between centralized and decentralized policies, followed by extensions through weighted factorization in QPLEX [48], attention mechanisms in QTRAN [38]. Despite their remarkable performance in complex multi-agent tasks, these methods remain largely opaque in their decision-making processes, highlighting the critical need for explainability in MARL.

2.2 Explainability in Reinforcement Learning

The explainable reinforcement learning (XRL) community has developed various approaches for interpreting DRL models, focusing on different aspects of agent behavior, decision-making processes, and the form of interpretation [22]. Iyer et al. [20] proposed Object-sensitive Deep RL (O-DRL), which incorporates object-related information directly into the DRL network architecture, enabling object-level explanations of agent actions. Madumal et al. [26] developed an action influence model that captures causal relationships between actions and environment variables. Yau et al. [52] emphasized the importance of understanding agent *intentions*, proposing a method that projects predicted future trajectories from current observations and actions. Guo et al. [14] focused on identifying critical decision points within episodes, developing methods to highlight time steps that significantly influence final rewards. Ayala et al. [2] proposed converting internal Q-values into success probabilities for non-episodic tasks, making the decision-making process more intuitive. Recent work has explored hierarchical abstractions of action sequences [37, 44], enabling high-level behavioral explanations. However, these single-agent methods do not address the challenges of multi-agent coordination and interaction.

2.3 Explainability in Multi-Agent Systems

Despite the growing importance of MARL, explainability research in multi-agent settings remains limited. Early work by Wang et al. [50] proposed generating verbal explanations using predefined rules based on prior knowledge about agent positions and relations. Building on game-theoretic foundations, Heuillet et al. [16, 17] adapted Shapley values for MARL interpretation, providing principled methods to quantify individual agent contributions to collective outcomes. Boggess et al. [4] introduced policy explanations for MARL, focusing on temporal queries and sequential decision-making patterns. Their subsequent work [5] extended to handle complex temporal queries about multi-agent behaviors. Shi et al. [36] developed MADDPGViz, an interactive visual analytics tool specifically designed for MADDPG [24]. Itaya et al. [19] visualized actor-attention to analyze agents' cooperative decisions. Domenech i Vila et al. [8] used policy graphs to explain a trained agent's behavior in multi-agent cooperative environments. Kravaris et al. [21] explored explainability in air traffic flow management, demonstrating a practical application of interpretable MARL. Due to page limitations, we refer interested readers to recent survey papers by Dazeley et al. [7] and Wells and Bednarz [51] for more details. Despite these advances, existing methods often require extensive domain knowledge, work only with specific algorithms, and fail to capture emergent coordination strategies across temporal scales.

3 FUSION ADAPTIVE RESONANCE THEORY

MEASE employs an *episodic memory model* as its core component for learning cooperative multi-agent behaviors. Specifically, we utilize *fusion Adaptive Resonance Theory* (fusion ART) [41, 42], a class of self-organizing neural networks that offers several advantages over deep learning approaches for this task: (1) incremental learning from streaming episodes without retraining, (2) stable category formation without catastrophic forgetting, and (3) interpretable symbolic representations rather than opaque embeddings. This section provides the necessary background on how fusion ART's properties enable the encoding and generalization of multi-agent behavioral patterns. For a comprehensive treatment of the theoretical foundations, we refer interested readers to Tan et al. [42] on fusion ART architectures, Grossberg et al. [13] on applying Adaptive Resonance Theory to explainable AI, and Wai et al. [46, 47] for recent XMARL applications with fusion ART.

As a general multi-purpose architecture for learning *cognitive codes*, fusion ART encodes multi-modal representations of memory blocks across multiple pattern channels, as a response to continuous streams of incoming patterns. Specifically, it consists of several input/output fields and a category field as defined below.

3.1 Input/Output and Category Field of Fusion ART

Let $\mathbf{I}^k = (I_i^k)_{i=1}^m$ be an input vector to be presented to the input field or channel k (F_1^k) of the fusion ART network, where $I_i^k \in [0, 1]$.

Let $\mathbf{x}^k = (x_l^k)_{l=1}^n$ be the activity vector of field k (F_1^k) receiving the input vector $\mathbf{I}^k$. *Complement coding* can be applied such that $\mathbf{x}^k = (\mathbf{I}^k, \bar{\mathbf{I}}^k)$ where $\bar{\mathbf{I}}^k = (\bar{I}_i^k)_{i=1}^m$ and $\bar{I}_i^k = 1 - I_i^k$ to prevent code proliferation (overfitting) and to allow feature generalization during learning. $\mathbf{w}_j^k$ is the weight vector associated with the j^{th} node in F_2 for learning the input pattern in F_1^k. Vector $\mathbf{y} = (y_j)_{j=1}^r$ can be defined to represent node activation in F_2.

3.2 Resonance Dynamic of Fusion ART

Given the input vector $\mathbf{x}^k$, *resonance search* is initiated by a bottom-up *choice function* to activate every node j in the *category field* F_2 as follows

$$T_j = \sum_{k=1}^{n} \gamma^k \frac{|\mathbf{x}^k \wedge \mathbf{w}_j^k|}{\alpha^k + |\mathbf{w}_j^k|} \tag{1}$$

where fuzzy AND operation $\wedge$ is defined by $(p \wedge q)_i \equiv \min(p_i, q_i)$ and the norm$|.|$ is defined by $|\mathbf{P}| \equiv \sum_i p_i$ for the vectors $\mathbf{p}$ and $\mathbf{q}$. $\alpha^k \geq 0$ is the choice parameter to avoid division by zero. $\gamma^k \in [0, 1]$ is the contribution parameter that indicates the significance of F_1^k.

Code competition selects the F_2 node J with the maximum activation value such that

$$T_J = \max\{T_j : \text{for all } F_2 \text{ node j}\} \tag{2}$$

Top-down *template matching* then applies to search for the resonance condition given by

$$m_J^k = \frac{\mathbf{x}^k \wedge \mathbf{w}_J^k}{|\mathbf{x}^k|} \geq \rho^k, \text{for all field } k \tag{3}$$

where $\rho^k \in [0, 1]$ is the vigilance parameter or resonance threshold for channel k. If for any k, $m_J^k < \rho^k$, a mismatch reset occurs on J and iteratively, another F_2 node J is selected following Equation 3 until a resonance is found or, otherwise, an uncommitted node is recruited in F_2 as a new recognition category for the novel input. If resonance is eventually found, *template learning* takes place such that

$$\mathbf{w}_j^{k(\text{new})} = (1 - \beta^k)\mathbf{w}_j^{k(\text{old})} + \beta^k(\mathbf{x}^k \wedge \mathbf{w}_j^{k(\text{old})}) \tag{4}$$

where $\beta^k \in [0, 1]$ is the learning rate parameter. Node J may also perform a readout of its weight vectors to an input field F_1^k such that $\mathbf{x}^{k(\text{new})} = \mathbf{w}_J^k$.

4 PROBLEM FORMULATION

Consider a multi-agent system with N homogeneous agents operating in a decentralized manner. Let $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$ denote the shared discrete action space of agents. An episode $\mathcal{E}$ consists of a sequence of joint actions $\mathcal{E} = \{e_1, e_2, \ldots, e_T\}$, where each event $e_t = (a_t^1, a_t^2, \ldots, a_t^N)$ represents the joint actions of all agents at timestep t and T represents the horizon of episode $\mathcal{E}$.

Our objective is to transform these primitive action sequences into human-interpretable strategies $\mathcal{S}$ that capture the essential coordination patterns while achieving significant compression. Formally, we define a strategy as: $\mathcal{S} = \{(s_1, r_1, \tau_1), \ldots, (s_K, r_K, \tau_K)\}$, where s_k represents a *significant joint action pattern* with semantic action labels, $r_k \in \mathbb{N}^+$ denotes the repetition count of pattern s_k, τ_k specifies the time interval during which pattern s_k occurs, and $K \ll T$ represents the strategy length. We seek a mapping $f : \mathcal{E} \to \mathcal{S}$ such that $|\mathcal{S}| \ll |\mathcal{E}|$ while preserving the semantic content of multi-agent coordination.

Algorithm 1 EM-ART 2 Dynamics

Require: Episode traces from MARL agent behaviors
Ensure: Set of generalized episodes
Parameters: τ (decay rate)
for each episode e in episode traces **do**
  for each event v in episode e **do**
    Present event v to layer F_1
    Select winning node J in layer F_2 via resonance search
    Set activation: $y_J \leftarrow 1$
    Update complement coding: $\bar{y}_J \leftarrow 1 - y_J^{(\text{new})}$
    Apply temporal decay to all active nodes in F_2:
      $y_i^{(\text{new})} \leftarrow (1 - \tau) \cdot y_i^{(\text{old})}$ for all i with $y_i > 0$
  end for
  Assign episode label to event label field in F_2
  Learn episode pattern and outcome status as node in F_3 via resonance search
  Reset all activations: $y \leftarrow 0$ in F_2
end for

5 METHODOLOGY

In this section, we introduce *Multi-Agent Episodic Strategy Extraction* (MEASE), a novel framework for explaining MARL policies by transferring multi-agent joint behaviors into interpretable episodic memory. Our approach extracts and abstracts multi-agent coordination strategies from execution traces, enabling post-hoc interpretation of complex joint behaviors through a two-stage process.

5.1 Episodic Memory Encoding

MEASE employs episodic memory models to memorize multi-agent behavioral patterns. Specifically, this work leverages EM-ART 2, which extends fusion ART and *Episodic Memory Adaptive Resonance Theory* (EM-ART) [49] to enable interpretable episodic memory encoding [42]. Each atomic event encapsulates three essential attributes of agent experience: agent identity (who), action selection (what), and temporal context (when). These events are subsequently organized into episodes — temporally-ordered sequences of correlated action events that capture meaningful behavioral patterns.

5.1.1 EM-ART 2 Architecture. EM-ART 2 extends the EM-ART model by incorporating new *Label* fields for encoding episode outcomes and L2-norm based ART 2 choice and match functions [6] to improve generalization of sequential patterns. As shown in Figure 2, EM-ART 2 employs a three-layer hierarchical architecture specifically designed to encode multi-agent behavioral sequences: Input Fields (F_1) encode multi-modal information from agent trajectories. Agent discrete actions are encoded using one-hot representations, while temporal information is normalized to values between 0 and 1. Episode outcomes, e.g., win/loss, are encoded as binary labels to provide learning targets for strategy discovery. Event Category Field (F_2) learns distinct spatio-temporal patterns representing coordinated multi-agent actions. To support strategic pattern discovery,

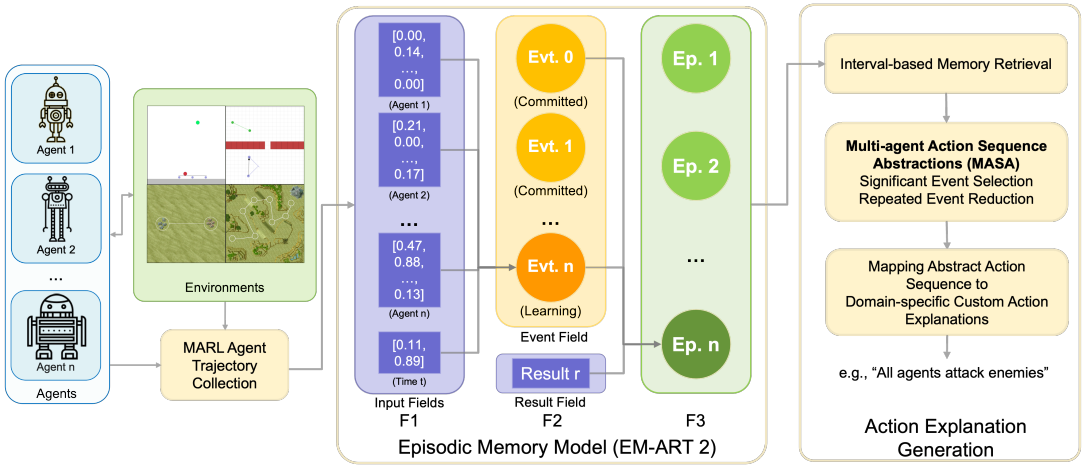


Figure 1: The MEASE framework. In the episode memory model, we exemplify a simple case on how EM-ART encodes values from the input field on F_1 as a new event (event n) in F_2 and subsequently encodes a new episode F_3 with event 0, event n, and the corresponding outcome (result). The action explanation generation utilizes the learned episodic memory in EM-ART 2.

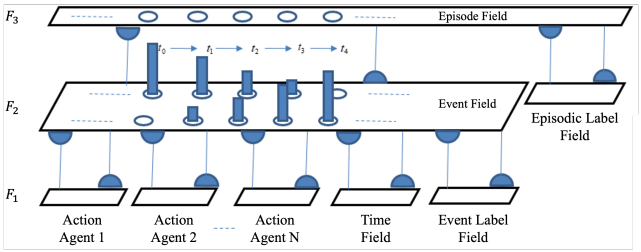


Figure 2: The EM-ART 2 architecture for encoding multi-agent action sequences.

this layer employs a dual-field structure: (1) an *event* field that captures temporal coordination patterns through gradient encoding, and (2) an *episodic* label field that encodes episode outcomes. Episode Category Field (F_3) encodes complete behavioral sequences as episodic memories, representing full multi-agent strategies from episode initiation to completion.

The learning process operates differently across network layers to capture both immediate coordination and long-term strategic patterns. $F_1 \rightarrow F_2$ transition uses standard fusion ART dynamics to encode individual coordination events, while the $F_2 \rightarrow F_3$ transition employs modified ART 2 functions specifically designed for sequence generalization. In addition, the temporal decay mechanism in F_2 implements gradient encoding, which preserves event ordering by assigning monotonically decreasing activation values to sequential elements. EM-ART 2 is considered transparent, as every learned episode/event is stored explicitly and can be read out directly as a template of the sequence or strategy.

5.1.2 EM-ART 2 Dynamics. The dynamics of EM-ART 2, as summarized in Alg. 1, are presented as follows. Let s and w be the indices referring to the event field and the label field, respectively. EM-ART 2 employs a hybrid choice function for activating node j in F_3:

$$T_j = \gamma^s \frac{\mathbf{x}^s \cdot \mathbf{w}_j^s}{\| \mathbf{x}^s \| \| \mathbf{w}_j^s \|} + \gamma^w \frac{|\mathbf{x}^w \wedge \mathbf{w}_j^w|}{\alpha^w + |\mathbf{w}_j^w|} \tag{5}$$

where the L2 norm $\| \mathbf{p} \| \equiv \sqrt{p_1^2 + \ldots + p_n^2}$. The match function for the event field s uses L2-based similarity:

$$m_j^s = \frac{\mathbf{x}^s \cdot \mathbf{w}_j^s}{\| \mathbf{x}^s \| \| \mathbf{w}_j^s \|} \geq \rho^s \tag{6}$$

while the label field w uses standard fuzzy matching (Eq. 3). Template learning differs for each field. The event field employs L2-based learning:

$$\mathbf{w}_j^{s(\text{new})} = (1 - \beta^s)\mathbf{w}_j^{s(\text{old})} + \beta^s(\mathbf{x}^s) \tag{7}$$

while the label field uses standard fuzzy learning (Eq. 4).

Using fusion ART as the building block, EM-ART 2 inherits the generalization property based on the vigilance parameters. Specifically, the vigilance parameters at the event layer determine the granularity of the learned categories, while those at the episode level influence the number of learned episodes. Once the episodes are encoded and learned, they can be recalled by reading out the weight vector for each episode node in F_3 to the F_2 layer. Each event in the readout episode, as activated in F_2, is then read out to the F_1 layer, subsequently one at a time following the order in the sequence. The sequential order of the events is based on the graded values formed in F_2 after the episode readout.

5.1.3 Interval-based Memory Retrieval. For memory retrieval, events with selected features can be recalled by firstly providing the input vector at a particular input field F_1^k to activate the event nodes at F_2. This approach can be applied to get relevant events based on a particular criterion or query.

To abstract and discern dominant or salient events within an interval of time, we apply a novel *interval-based memory retrieval* approach by averaging the features of all events that occur within the time interval. Given $\mathbf{x}^t$ as the vector for time field, every node j in F_2 can be activated such that $T_j = \gamma^t \frac{|\mathbf{x}^t \wedge \mathbf{w}_j^t|}{\alpha^t + |\mathbf{w}_j^t|}$. The abstracted

Algorithm 2 Interval-Based Memory Retrieval

Require: EM-ART 2 Neural Network; abstraction factor ϕ; maximum possible length of learned episode L
Ensure: List of abstracted events $\mathcal{E}$
1: $\mathcal{E} \leftarrow \emptyset$; set time interval $\delta \leftarrow L/\phi$
2: **for** every time interval $(t, t+\delta)$ in a sequence or episode **do**
3: set time field vector $\mathbf{x}^t$ with normalized values of timestamp $(t, 1-(t+\delta))$
4: activate nodes in F_2 such that $T_j = \gamma^t \frac{|\mathbf{x}^t \wedge \mathbf{w}_j^t|}{\alpha^t + |\mathbf{w}_j^t|}$
5: readout the abstracted event $\mathbf{x}^k \leftarrow \frac{\sum_{j=1}^{|\mathbf{y}|} \mathbf{w}_j^k \cdot T_j}{\sum_{j=1}^{|\mathbf{y}|} T_j}$
6: append $\mathbf{x}^k$ for interval $(t, t+\delta)$ to $\mathcal{E}$
7: **end for**

event vector $\mathbf{x}^k$ can then be obtained as:

$$\mathbf{x}^k \leftarrow \frac{\sum_{j=1}^{|\mathbf{y}|} \mathbf{w}_j^k \cdot T_j}{\sum_{j=1}^{|\mathbf{y}|} T_j} \tag{8}$$

where $|\mathbf{y}|$ is the length of activation vector $\mathbf{y}$ in F_2. Alg. 2 shows the process of producing the set of abstracted events given ϕ as the *abstraction factor*. The abstraction factor can be chosen according to the length of the abstracted sequences (and time interval) to generate. The time interval δ to be used for retrieval can be calculated as $\delta \leftarrow L/\phi$ where L is the maximum possible length of episodes.

5.2 Multi-agent Action Sequence Abstraction

To bridge the gap between primitive actions and human understanding, we introduce the *Multi-agent Action Sequence Abstraction* (MASA) algorithm (Alg. 3), transforming time-segmented action sequences into the final human-interpretable strategy $\mathcal{S}$. Specifically, MASA operates on the output of interval-based memory retrieval, i.e., action sequences segmented based on time intervals τ_k. MASA transforms these segments through a two-phase process.

In Phase 1 (Significant Event Selection), MASA filters each interval t to retain only actions performed by at least M agents, where the threshold $M \in \{1, \ldots, N\}$ controls the abstraction granularity. This phase extracts coordination-critical events $\mathcal{A}_t$ where collective behavior emerges, discarding sparse individual actions. Phase 2 (Repeated Event Reduction) applies temporal compression by identifying consecutive identical events in the filtered sequence and encoding them as tuples (s, r), where s represents the event pattern and r its repetition count. This transformation typically achieves 90-95% episode compression while preserving essential coordination strategies, making complex agent behaviors interpretable.

6 THE MULTI-AGENT BEHAVIORAL DATASET

To ensure reproducibility, we collect a *Multi-Agent Behavioral* (MAB) dataset, comprising full-episode multi-agent trajectories from trained MARL agents across various scenarios. The dataset spans two established environments: the Vectorized Multi-Agent Simulator (VMAS) [3] and the Multi-Objective StarCraft Multi-Agent Challenge (MOSMAC) [10, 12]. Table 1 summarizes the MAB scenarios.

Algorithm 3 Multi-agent Action Sequence Abstraction (MASA)

Require: Time-segmented episode from memory retrieval: $\mathcal{R} = \{(a_1, \tau_1), (a_2, \tau_2), \ldots, (a_K, \tau_K)\}$ where a_k is action sequence and τ_k is time interval; abstraction threshold M
Ensure: Strategy $\mathcal{S} = \{(s_1, r_1, \tau_1), (s_2, r_2, \tau_2), \ldots, (s_L, r_L, \tau_L)\}$
Phase 1: Significant Event Selection
Initialize filtered sequence $\mathcal{F} \leftarrow \emptyset$
for $k = 1$ to K **do**
Extract joint action $a_k = (a_k^1, a_k^2, \ldots, a_k^N)$ for interval τ_k
$s_k \leftarrow \{(a, c) : c = |\{i : a_k^i = a\}| \geq M\}$
if $s_k \neq \emptyset$ **then**
$\mathcal{F} \leftarrow \mathcal{F} \cup \{(s_k, \tau_k)\}$
end if
end for

Phase 2: Repeated Event Reduction
Initialize strategy $\mathcal{S} \leftarrow \emptyset$, $j \leftarrow 1$
while $j \leq |\mathcal{F}|$ **do**
$(s_{curr}, \tau_{start}) \leftarrow \mathcal{F}[j]$
$r \leftarrow 1$, $\tau_{end} \leftarrow \tau_{start}$
while $j + r \leq |\mathcal{F}|$ **and** $\mathcal{F}[j+r].pattern = s_{curr}$ **do**
$\tau_{end} \leftarrow \mathcal{F}[j+r].interval$, $r \leftarrow r + 1$
end while
$\tau_{merged} \leftarrow [\tau_{start}.start, \tau_{end}.end]$ {Merge intervals}
$\mathcal{S} \leftarrow \mathcal{S} \cup \{(s_{curr}, r, \tau_{merged})\}$, $j \leftarrow j + r$
end while
return $\mathcal{S}$

6.1 Data Collection Protocol

We trained all agents using QMIX [33] within the EPyMARL[1] framework until convergence (> 90% success rate for MOSMAC, stable high returns for VMAS). QMIX was selected for its state-of-the-art performance across multi-agent benchmarks [12, 31]. After convergence, policies were frozen and 1,000 complete trajectories per scenario were collected using default environment configurations. The dataset provides balanced coverage across four coordination challenges: Balance and Joint Passage from VMAS; 4t1sp and 4t8sp from MOSMAC[2]. Each trajectory captures full joint action sequences throughout an episode, providing rich behavioral data for pattern extraction and explanation generation.

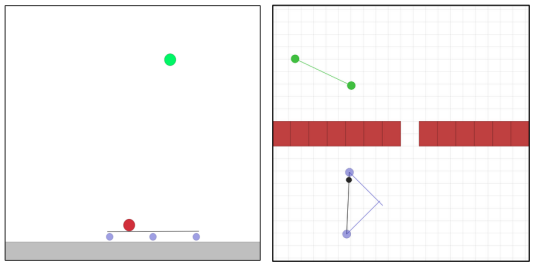
6.2 Environment Specifications

6.2.1 VMAS Environment. The Vectorized Multi-Agent Simulator (VMAS) [3] is a continuous 2D physics-based platform for multi-agent coordination tasks. We adopt the *Joint Passage* and *Balance* scenarios (see Figure 3) from VMAS, with action spaces comprising nine discrete movement actions in cardinal and diagonal directions. Object locations are fully randomized. The *Joint Passage* scenario requires two agents to synchronize movements through barriers. The *Balance* scenario involves three agents manipulating a shared platform to position a ball at target locations. Table 2 presents one example of the collected data from the VMAS balance scenario.

[1] https://github.com/uoe-agents/epymarl
[2] https://github.com/smu-ncc/MOSMAC

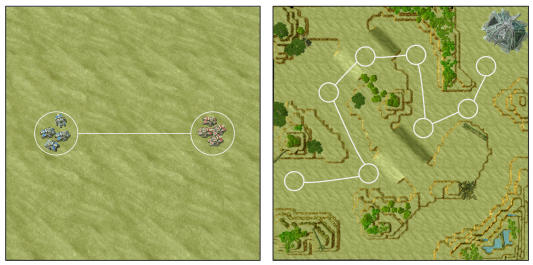
Table 1: Properties of scenarios in the Multi-Agent Behavioral (MAB) Dataset.

Env.	Scenario	Agents	Horizon	Collab.	Stochasticity	Description
VMAS	Joint Passage	2	100	High	High	Two agents navigate through a constrained environment with barriers, requiring synchronized movement to successfully traverse obstacles.
	Balance	3	100	High	High	Three agents collaboratively manipulate a shared platform to move a ball toward a target location, requiring fine-grained coordination to balance the platform stability with strategic ball positioning.
MOSMAC	4t1sp	4	50	High	Low	Four siege tank units engage in symmetric tactical combat and navigation, requiring coordinated positioning, movement, and attack strategies to defeat enemy forces.
	4t8sp	4	300	High	Low	Four siege tank units navigate through eight strategic points while engaging enemies, combining long-horizon navigation with tactical combat.



(a) Balance (b) Joint Passage

Figure 3: The VMAS scenarios implemented in this study.



(a) The 4t1sp scenario (b) The 4t8sp scenario

Figure 4: MOSMAC scenarios with four ally units navigating toward goals via traversable paths with intermediate targets.

6.2.2 MOSMAC Environment. MOSMAC [12] features long-horizon multi-objective MARL in the StarCraft II environment. Agents need to balance navigation, combat, and safety objectives to successfully complete tasks. We implement two scenarios following Wai et al. [46, 47] in MOSMAC (see Figure 4): *4t1sp* where four siege tanks navigate to a single target on a 32×32 map, and *4t8sp* where they traverse through eight strategic points on a 128×128 map of complex terrain, both while engaging enemy forces.

7 EVALUATION CONFIGURATIONS

7.1 Evaluation Protocol

We configure EM-ART 2 with vigilance parameters $\rho^s = 0.25$ for sequence learning and $\rho^e = 1.0$ for event learning, with abstraction factor $\phi = 10$ for interval-based memory retrieval. MASA requires consensus from $M = 2$ agents for action selection.

Table 2: A joint action trajectory of three agents over the first 20 timesteps in the VMAS balance scenario. Legend: ↑ Up, ↗ Up-Right, ↖ Up-Left, → Right.

(a) $t_1 - t_5$				(b) $t_6 - t_{10}$				(c) $t_{11} - t_{15}$				(d) $t_{16} - t_{20}$			
T	Ag_1	Ag_2	Ag_3	T	Ag_1	Ag_2	Ag_3	T	Ag_1	Ag_2	Ag_3	T	Ag_1	Ag_2	Ag_3
t_1	↗	↖	↖	t_6	↖	↖	↖	t_{11}	↖	↖	↖	t_{16}	↑	↑	↖
t_2	↑	↖	↖	t_7	↗	↖	↖	t_{12}	↗	↖	↖	t_{17}	↗	↖	↖
t_3	↑	↖	↖	t_8	↑	↖	↖	t_{13}	↑	↑	↖	t_{18}	↑	↑	↑
t_4	↑	↖	↖	t_9	↑	↖	↖	t_{14}	↑	↖	↖	t_{19}	↑	↑	↑
t_5	↗	↖	↖	t_{10}	↖	↖	↖	t_{15}	↑	↑	↖	t_{20}	↑	↑	↑

Table 3: Results of memory encoding with EM-ART 2.

Env.	Scenario	Raw Events	F2 Nodes	F3 Nodes
VMAS	Joint Pass.	100,000	3,405	149
	Balance	99,964	3,883	59
MOSMAC	4t1sp	37,647	8,012	48
	4t8sp	44,500	17,608	58

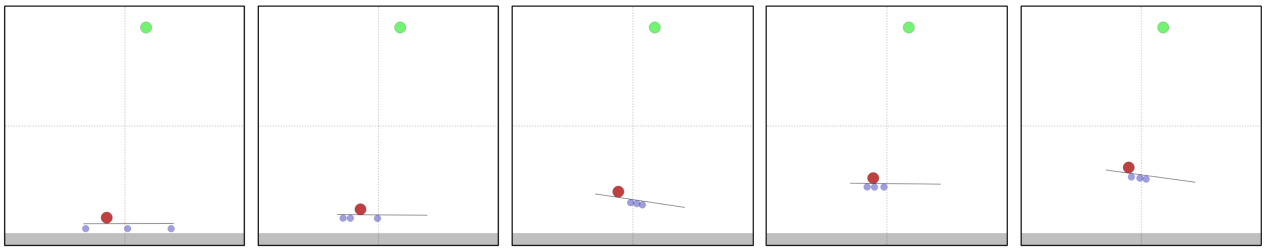
7.2 Evaluation Metrics

We evaluate MEASE along two dimensions. We first measure the compression effectiveness of EM-ART 2's episodic memory by analyzing the number of learned patterns, specifically $|F_2|$ codes representing distinct event patterns and $|F_3|$ codes representing complete episode strategies. Second, we validate explanation fidelity through an "explanation-as-strategy" approach, where extracted symbolic strategies are directly executed by agents on original tasks to assess whether compressed explanations retain essential coordination patterns. Third, following Wai et al. [46], we conducted a user study with 31 participants evaluating MEASE explanations on MOSMAC scenarios, rating explanations on clarity, usefulness, and user satisfaction, with inter-rater agreement analysis to assess consistency.

8 RESULTS

8.1 Episodic Memory Learning with EM-ART 2

Table 3 presents EM-ART 2 training results across the four scenarios. On the VMAS Balance scenario with 1,000 episodes and 99,964 raw events, EM-ART 2 performs two-stage abstraction: (i) identification of 3,883 distinct spatio-temporal patterns as F_2 cognitive codes, representing unique multi-agent interaction primitives,



(a) State 1 (b) State 2 (c) State 3 (d) State 4 (e) State 5

Figure 5: Visualization of the trajectory presented in Table 4. Agents show clear balancing behaviors: (a)→(b) upper-left positioning, (b)→(c) upper-right adjustment, (c)→(d) upper-left platform stabilization, (d)→(e) upper-right approach to goal.

Table 4: One strategy derived from the VMAS balance scenario using MEASE. Legend: See Table 2.

(a1) $t_1 - t_{20}$		(a2) $t_{21} - t_{50}$		(b1) $t_{51} - t_{70}$		(b2) $t_{71} - t_{100}$	
T	Action	T	Action	T	Action	T	Action
$t_1 - t_{10}$	↖	$t_{21} - t_{30}$	↑ ↗	$t_{51} - t_{60}$	↑ ↖	$t_{71} - t_{80}$	↑ ↗
$t_{11} - t_{20}$	↗ ↑ ↖	$t_{31} - t_{40}$	↑ ↗	$t_{61} - t_{70}$	↗ ↑ ↖	$t_{81} - t_{90}$	↑ – ↗
		$t_{41} - t_{50}$	↗ ↖			$t_{91} - t_{100}$	↑ ↖

Table 5: Symbolic strategy for 4t1sp derived through MASA ($M = 2$) with tactical interpretations. Legend - Movement: ↑ North, ↓ South, → East, ← West, × No-op, Attack[enemy$_i$]: A_1 Enemy 1, A_2 Enemy 2, A_3 Enemy 3, A_4 Enemy 4.

(a) $t_1 - t_{12}$			(b) $t_{13} - t_{30}$		
T	Action	Interpretation	T	Action	Interpretation
$t_1 - t_3$	←	Away from enemies	$t_{13} - t_{15}$	A_1	Focus fire E1
$t_4 - t_6$	↓ ↑	Spread out	$t_{16} - t_{18}$	A_4	Focus fire E4
$t_7 - t_9$	←	Away from enemies	$t_{19} - t_{25}$	A_2	Focus fire E2
$t_{10} - t_{12}$	→	Advance	$t_{25} - t_{30}$	A_3	Focus fire E3

Table 6: Symbolic strategy for 4t8sp derived through MASA ($M = 2$). Legend: See Table 5.

(a) $t_1 - t_{32}$		(b) $t_{33} - t_{68}$		(c) $t_{69} - t_{132}$		(d) $t_{133} - t_{254}$	
T	Action	T	Action	T	Action	T	Action
$t_1 - t_4$	← ↑	$t_{33} - t_{36}$	→ ↑	$t_{73} - t_{76}$	A_4	$t_{149} - t_{172}$	× ↓
$t_5 - t_8$	↓	$t_{37} - t_{48}$	↑	$t_{77} - t_{84}$	× ↑	$t_{173} - t_{184}$	× →
$t_9 - t_{16}$	→	$t_{49} - t_{52}$	↑ ←	$t_{85} - t_{88}$	× ←	$t_{185} - t_{192}$	× ↓
$t_{17} - t_{20}$	↑ →	$t_{53} - t_{56}$	↑	$t_{89} - t_{112}$	× ↑	$t_{193} - t_{212}$	× →
$t_{21} - t_{24}$	→	$t_{57} - t_{60}$	A_2	$t_{113} - t_{132}$	× →	$t_{213} - t_{254}$	× ↑
$t_{25} - t_{28}$	↑	$t_{61} - t_{68}$	A_3	$t_{133} - t_{140}$	× ↑		
$t_{29} - t_{32}$	← ↑	$t_{69} - t_{72}$	A_1	$t_{141} - t_{148}$	× →		

and (ii) compression into 59 episodic memories in F_3, each encoding a complete collaborative strategy. Similarly, MEASE abstracts 1,000 episodes from the MOSMAC 4t1sp scenario into 8,012 distinct spatio-temporal patterns and 47 distinct behavioral clusters.

8.2 Explanations from MEASE

MEASE successfully explains multi-agent policies learned from both the VMAS and MOSMAC environments. Table 4 presents a representative collaborative strategy extracted by MEASE for the VMAS Balance scenario, with visualization in Figure 5. MEASE also distills winning strategies into interpretable tactical phases (Table 5) in MOSMAC 4t1sp. The extracted strategy reveals three coordination phases in 4t1sp: (i) defensive positioning with team dispersion while maintaining distance from enemies (t_1–t_9), (ii) coordinated advance toward engagement range (t_{10}–t_{12}), and (iii) systematic focus fire eliminating enemies in prioritized order (1→4→2→3).

Table 7: Performance of the abstracted and refined strategies compared with the MARL model.

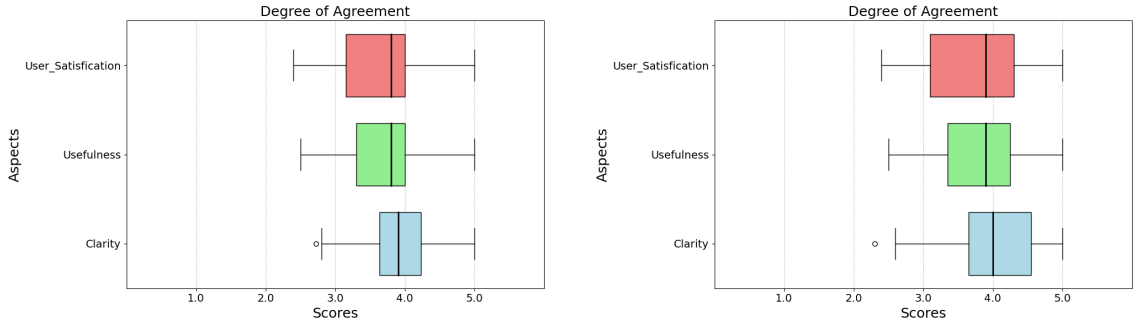
Approach	Win Rate (%)	Episode Reward
MARL (QMIX)	**93.7 ± 3.4**	19.53 ± 0.31
MEASE Explanation (AE)	93.2 ± 1.1	**19.59 ± 0.07**
Ablated Explanation (AE-4)	88.1 ± 1.2	19.26 ± 0.07
Ablated Explanation (AE-8)	69.9 ± 2.0	18.03 ± 0.13
Ablated Explanation (AE-0)	19.8 ± 3.9	14.30 ± 0.17

8.3 MEASE Empirical Evaluation

To measure the effectiveness of explanations generated by MEASE, we validate the framework through an "explanation-as-strategy" approach, where the symbolic explanations are directly executed as coordination protocols by agents in the original scenario. It tests whether our compressed explanations capture the essential decision-making patterns necessary for successful task completion.

Agents execute the MEASE abstracted action sequences across ten experimental runs, each consisting of 1,000 episodes. Table 7 demonstrates that this approach achieves performance comparable to the original MARL policy — specifically, matching 93.2% win rate versus 93.7% for QMIX while actually achieving slightly higher episode rewards (19.59 vs 19.53). This near-parity performance provides strong empirical evidence for successful knowledge distillation: the symbolic explanations faithfully represent the coordination mechanisms learned by the black-box MARL policy.

To assess the criticality of specific tactical components, we conduct systematic ablation studies examining three strategic variants. Ablated strategy AE-4 modifies spreading actions at timestep 4 from "south, north" to "west, north", reducing win rates from 93.3% to 88.1%. Ablated strategy AE-8 further disrupts coordination by extrapolating timestep 8 actions from "west" to "north, west", causing performance to drop to 69.9%. Most dramatically, AE-0 eliminates spreading actions, resulting in catastrophic performance collapse to 19.8%. This hierarchical degradation pattern reveals that MEASE



(a) 4t1sp **(b) 4t8sp**

Figure 6: Distribution of user ratings for *clarity*, *usefulness*, and *user satisfaction* across scenarios. Box plots show median (center line), interquartile range (box), and outliers (points).

identifies critical coordination primitives at multiple scales — from tactical refinements to fundamental strategic principles.

These empirical results demonstrate that MEASE explanations satisfy both fidelity and interpretability requirements for explainable MARL: they accurately distill essential coordination patterns (evidenced by comparable performance) while exposing sequential behavior structure (revealed through ablation sensitivity).

8.4 Explanation Quality Assessment

To evaluate the explanations generated by MEASE, we conducted an online user study following the methodology of Wai et al. [46], which was validated in a similar multi-agent explanation setting. The study involved 31 participants with diverse backgrounds in age, gender, and familiarity with real-time strategy games. We assessed three key dimensions of explanation quality: *clarity* (comprehensibility), *usefulness* (practical value for understanding agent behavior), and *user satisfaction* (overall contentment with explanation quality).

Each participant reviewed ten gameplay episodes from MOSMAC scenarios (five each from *4t1sp* and *4t8sp*). For each episode, participants first watched action sequences without explanations as a baseline, then viewed the same sequences with MEASE-generated explanations. Following each viewing, participants rated explanation quality on a 5-point Likert scale (1=strongly disagree, 5=strongly agree) across six evaluation questions (Table 8). The complete study required approximately 30 minutes per participant. We employed *Inter-Rater Agreement Analysis* [25, 35, 39, 46] to assess the degree of consensus among respondents. Figure 6 shows the distribution of ratings for *clarity*, *usefulness*, and *user satisfaction* across scenarios.

Both scenarios demonstrated strong positive receptions across all evaluation dimensions. For the *4t1sp* scenario, median scores reached 3.9 for clarity and 3.8 for both usefulness and user satisfaction, with relatively narrow interquartile ranges (IQR) indicating consistent agreement among participants. The *4t8sp* scenario, despite its increased complexity, achieved comparable or higher ratings: 4.0 for clarity, 3.9 for usefulness, and 3.8 for user satisfaction. These results demonstrate that MEASE effectively generates comprehensible explanations even for complex multi-agent coordination tasks. The overall consistency in scores across participants validates the robustness of our approach. The similar performance across both scenarios indicates that MEASE maintains explanation quality as task complexity increases.

Table 8: User study questions.

Evaluation Aspect	Index	Question
Explanation Clarity	Q1	"The explanation helps to better understand the action sequences of the agent team in the video."
	Q2	"The explanation given is clear and easy to follow."
Explanation Usefulness	Q3	"The explanation provided contains accurate information."
	Q4	"The explanation provided contains useful information."
User Satisfaction	Q5	"Overall, I can trust and rely on the explanations provided."
	Q6	"Overall, I am satisfied with the quality of the explanation provided."

9 CONCLUSION

This work presents MEASE, a novel framework for deriving explanations for trained multi-agent policies. Our study demonstrates the effectiveness of MEASE with EM-ART 2, interval-based retrieval, and Multi-agent Action Sequence Abstraction (MASA) algorithms in explaining multi-agent behaviors through interpretable episodic patterns. Experimental validation across VMAS and MOSMAC scenarios shows that extracted explanations achieve 93% of original MARL policy performance when deployed as strategies, verifying that MEASE captures essential coordination mechanisms. As a post-hoc XMARL method, our model-agnostic approach can explain a wide range of multi-agent models given multi-agent behavioral data. Beyond interpretability, the extracted episodic knowledge could potentially accelerate MARL training through imitation learning. Future work could involve incorporating contextual information from the state attributes to provide more fine-grained explanations of the MADRL agents' decision-making processes.

ACKNOWLEDGMENTS

This research was conducted in collaboration with the DSO National Laboratories, Singapore and supported in part by the Lee Kong Chian Professorship awarded to Ah-Hwee Tan by Singapore Management University.

REFERENCES

[1] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence. *Information Fusion* 99 (Nov. 2023), 101805. https://doi.org/10.1016/j.inffus.2023.101805

[2] Angel Ayala, Francisco Cruz, Bruno Fernandes, and Richard Dazeley. 2021. Explainable Deep Reinforcement Learning Using Introspection in a Non-episodic Task. https://doi.org/10.48550/arXiv.2108.08911 arXiv:2108.08911 [cs]

[3] Matteo Bettini, Ryan Kortvelesy, Jan Blumenkamp, and Amanda Prorok. 2022. VMAS: A Vectorized Multi-agent Simulator for Collective Robot Learning. In *DARS 2022*. Springer, 42–56. https://link.springer.com/10.1007/978-3-031-51497-5_4

[4] Kayla Boggess, Sarit Kraus, and Lu Feng. 2022. Toward Policy Explanations for Multi-Agent Reinforcement Learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 109–115. https://doi.org/10.24963/ijcai.2022/16

[5] Kayla Boggess, Sarit Kraus, and Lu Feng. 2023. Explainable Multi-Agent Reinforcement Learning for Temporal Queries. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI '23)*. Macao, P.R.China, 55–63. https://doi.org/10.24963/ijcai.2023/7

[6] Gail A. Carpenter and Stephen Grossberg. 1987. ART 2: Self-organization of Stable Category Recognition Codes for Analog Input Patterns. *Applied optics* 26, 23 (1987), 4919–4930.

[7] Richard Dazeley, Peter Vamplew, and Francisco Cruz. 2023. Explainable Reinforcement Learning for Broad-XAI: A Conceptual Framework and Survey. *Neural Computing and Applications* 35, 23 (March 2023), 16893–16916. https://doi.org/10.1007/s00521-023-08423-1

[8] Marc Domenech i Vila, Dmitry Gnatyshak, Adrian Tormos, Victor Gimenez-Abalos, and Sergio Alvarez-Napagao. 2024. Explaining the Behaviour of Reinforcement Learning Agents in a Multi-Agent Cooperative Environment Using Policy Graphs. *Electronics* 13, 3 (Jan. 2024), 573. https://doi.org/10.3390/electronics13030573

[9] Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual Multi-Agent Policy Gradients. In *AAAI 2018*. AAAI Press, 2974–2982.

[10] Minghong Geng, Shubham Pateria, Budhitama Subagdja, and Ah-Hwee Tan. 2024. Benchmarking MARL on Long Horizon Sequential Multi-Objective Tasks. In *AAMAS 2024*. IFAAMAS, 2279–2281.

[11] Minghong Geng, Shubham Pateria, Budhitama Subagdja, and Ah-Hwee Tan. 2024. HiSOMA: A Hierarchical Multi-Agent Model Integrating Self-Organizing Neural Networks with Multi-Agent Deep Reinforcement Learning. *Expert Syst. Appl.* 252 (Oct. 2024), 124117. https://doi.org/10.1016/j.eswa.2024.124117

[12] Minghong Geng, Shubham Pateria, Budhitama Subagdja, and Ah-Hwee Tan. 2025. MOSMAC: A Multi-agent Reinforcement Learning Benchmark on Sequential Multi-objective Tasks. In *AAMAS 2025*. IFAAMAS, 867–876.

[13] Stephen Grossberg. 2020. A Path Toward Explainable AI and Autonomous Adaptive Intelligence: Deep Learning, Adaptive Resonance, and Models of Perception, Emotion, and Action. *Frontiers Neurorobotics* 14, 36 (June 2020). https://doi.org/10.3389/fnbot.2020.00036

[14] Wenbo Guo, Xian Wu, Usmann Khan, and Xinyu Xing. 2021. EDGE: Explaining Deep Reinforcement Learning Policies. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., 12222–12236. https://proceedings.neurips.cc/paper/2021/hash/65c89f5a9501a04c073b354f03791b1f-Abstract.html

[15] Yazied A. Hasan, Arpit Garg, Satomi Sugaya, and Lydia Tapia. 2020. Defensive Escort Teams for Navigation in Crowds via Multi-Agent Deep Reinforcement Learning. *IEEE Robotics and Automation Letters* 5, 4 (Oct. 2020), 5645–5652. https://doi.org/10.1109/LRA.2020.3010203

[16] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. 2021. Explainability in Deep Reinforcement Learning. *Knowledge-Based Systems* 214 (Feb. 2021), 106685. https://doi.org/10.1016/j.knosys.2020.106685

[17] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. 2022. Collective eXplainable AI: Explaining Cooperative Strategies and Agent Contribution in Multiagent Reinforcement Learning With Shapley Values. *IEEE Computational Intelligence Magazine* 17, 1 (Feb. 2022), 59–71. https://doi.org/10.1109/MCI.2021.3129959

[18] Yu Huang, Daxin Liu, Zhenyu Liu, Ke Wang, Qide Wang, and Jianrong Tan. 2024. A Novel Robotic Grasping Method for Moving Objects Based on Multi-Agent Deep Reinforcement Learning. *Robotics and Computer-Integrated Manufacturing* 86 (April 2024), 102644. https://doi.org/10.1016/j.rcim.2023.102644

[19] Hidenori Itaya, Tom Sagawa, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. 2023. Visual Explanation for Cooperative Behavior in Multi-Agent Reinforcement Learning. In *2023 International Joint Conference on Neural Networks (IJCNN)*. 1–7. https://doi.org/10.1109/IJCNN54540.2023.10191563

[20] Rahul Iyer, Yuezhang Li, Huao Li, Michael Lewis, Ramitha Sundar, and Katia Sycara. 2018. Transparency and Explanation in Deep Reinforcement Learning Neural Networks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. Association for Computing Machinery, New York, NY, USA, 144–150. https://doi.org/10.1145/3278721.3278776

[21] Theocharis Kravaris, Konstantinos Lentzos, Georgios Santipantakis, George A. Vouros, Gennady Andrienko, Natalia Andrienko, Ian Crook, Jose Manuel Cordero Garcia, and Enrique Iglesias Martinez. 2022. Explaining Deep Reinforcement Learning Decisions in Complex Multiagent Settings: Towards Enabling Automation in Air Traffic Flow Management. *Applied Intelligence* 53, 4 (June 2022), 4063–4098. https://doi.org/10.1007/s10489-022-03605-1

[22] Q. Vera Liao and Kush R. Varshney. 2022. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. https://doi.org/10.48550/arXiv.2110.10790 arXiv:2110.10790v5 [cs]

[23] Ali Louati, Hassen Louati, Elham Kariri, Wafa Neifar, Mohamed K. Hassan, Mutaz H. H. Khairi, Mohammed A. Farahat, and Heba M. El-Hoseny. 2024. Sustainable Smart Cities through Multi-Agent Reinforcement Learning-Based Cooperative Autonomous Vehicles. *Sustainability* 16, 5 (Feb. 2024), 1779. https://doi.org/10.3390/su16051779

[24] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *Adv. Neural Inf. Process. Syst.*, Vol. 30. Curran Associates Inc., 6379–6390. https://proceedings.neurips.cc/paper_files/paper/2017/file/68a9750337a418a86fe06c1991a1d64c-Paper.pdf

[25] Chu Fei Luo, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. 2022. Evaluating Explanation Correctness in Legal Decision Making. *Proceedings of the Canadian Conference on Artificial Intelligence* (May 2022). https://doi.org/10.21428/594757db.8718dc8b

[26] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2020. Explainable Reinforcement Learning through a Causal Lens. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 03 (June 2020), 2493–2500. https://doi.org/10.1609/aaai.v34i03.5631

[27] Anuj Mahajan, Tabish Rashid, Mikayel Samvelyan, and Shimon Whiteson. 2019. MAVEN: Multi-Agent Variational Exploration. In *Advances in Neural Information Processing Systems* (Vancouver, Canada), Vol. 32. Curran Associates, Inc., Red Hook, NY, USA. https://proceedings.neurips.cc/paper/2019/hash/f816dc0acface7498e10496222e9db10-Abstract.html

[28] Federico Mason, Federico Chiariotti, Andrea Zanella, and Petar Popovski. 2024. Multi-Agent Reinforcement Learning for Coordinating Communication and Control. *IEEE Transactions on Cognitive Communications and Networking* 10, 4 (Aug. 2024), 1566–1581. https://doi.org/10.1109/TCCN.2024.3384492

[29] Thanh Thi Nguyen and Vijay Janapa Reddi. 2023. Deep Reinforcement Learning for Cyber Security. *IEEE Transactions on Neural Networks and Learning Systems* 34, 8 (Aug. 2023), 3779–3795. https://doi.org/10.1109/TNNLS.2021.3121870

[30] Afshin Oroojlooy and Davood Hajinezhad. 2023. A Review of Cooperative Multi-Agent Deep Reinforcement Learning. *Appl. Intell.* 53, 11 (June 2023), 13677–13722. https://doi.org/10.1007/s10489-022-04105-y

[31] Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. 2021. Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms in Cooperative Tasks. In *Proc. Neural Inf. Process. Syst. Track on Datasets and Benchmarks*, Vol. 1. Curran Associates Inc.

[32] Chanyoung Park, Gyu Seon Kim, Soohyun Park, Soyi Jung, and Joongheon Kim. 2023. Multi-Agent Reinforcement Learning for Cooperative Air Transportation Services in City-Wide Autonomous Urban Air Mobility. *IEEE Transactions on Intelligent Vehicles* 8, 8 (Aug. 2023), 4016–4030. https://doi.org/10.1109/TIV.2023.3283235

[33] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *ICML 2018* (Stockholm, Sweden) *(Proc. Mach. Learn. Res., Vol. 80)*. PMLR, 4295–4304. https://proceedings.mlr.press/v80/rashid18a.html

[34] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob Foerster, and Shimon Whiteson. 2019. The StarCraft Multi-Agent Challenge. In *AAMAS 2019*. IFAAMAS, 2186–2188. https://dl.acm.org/doi/10.5555/3306127.3332052

[35] Yael Septon, Tobias Huber, Elisabeth André, and Ofra Amir. 2023. Integrating Policy Summaries with Reward Decomposition for Explaining Reinforcement Learning Agents. In *Advances in Practical Applications of Agents, Multi-Agent Systems, and Cognitive Mimetics. The PAAMS Collection*. Springer Nature Switzerland, Cham, 320–332. https://doi.org/10.1007/978-3-031-37616-0_27

[36] Xiaoying Shi, Jiaming Zhang, Ziyi Liang, and Dewen Seng. 2023. MADDPGViz: A Visual Analytics Approach to Understand Multi-Agent Deep Reinforcement Learning. *Journal of Visualization* 26, 5 (May 2023), 1189–1205. https://doi.org/10.1007/s12650-023-00928-0

[37] Tom Silver, Ashay Athalye, Joshua B. Tenenbaum, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. 2023. Learning Neuro-Symbolic Skills for Bilevel Planning. In *Proceedings of The 6th Conference on Robot Learning*. https://proceedings.mlr.

press/v205/silver23a.html
[38] Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Hostallero, and Yung Yi. 2019. QTRAN: Learning to Factorize with Transformation for Cooperative Multi-Agent Reinforcement Learning. In *ICML 2019* (Long Beach, California, USA) *(Proceedings of Machine Learning Research, Vol. 97)*. PMLR, 5887–5896. http://proceedings.mlr.press/v97/son19a.html
[39] Budhitama Subagdja, Ah-Hwee Tan, and Yilin Kang. 2019. A Coordination Framework for Multi-Agent Persuasion and Adviser Systems. *Expert Systems with Applications* 116 (Feb. 2019), 31–51. https://doi.org/10.1016/j.eswa.2018.08.030
[40] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. 2018. Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward. In *AAMAS 2018*. IFAAMAS, 2085–2087. https://dl.acm.org/doi/10.5555/3237383.3238080
[41] Ah-Hwee Tan, Gail A. Carpenter, and Stephen Grossberg. 2007. Intelligence Through Interaction: Towards a Unified Theory for Learning. In *Advances in Neural Networks*, Derong Liu, Shumin Fei, Zeng-Guang Hou, Huaguang Zhang, and Changyin Sun (Eds.). Vol. 4491. Springer Berlin Heidelberg, Berlin, Heidelberg, 1094–1103. https://doi.org/10.1007/978-3-540-72383-7_128
[42] Ah-Hwee Tan, Budhitama Subagdja, Di Wang, and Lei Meng. 2019. Self-Organizing Neural Networks for Universal Learning and Multimodal Memory Encoding. *Neural Networks* 120 (Dec. 2019), 58–73. https://doi.org/10.1016/j.neunet.2019.08.020
[43] Philipp Theumer, Florian Edenhofner, Roland Zimmermann, and Alexander Zipfel. 2022. Explainable Deep Reinforcement Learning for Production Control. In *Proceedings of the Conference on Production Systems and Logistics: CPSL 2022*. https://repo.uni-hannover.de/items/01da9c0b-47e0-4a88-90e7-98b48c97df0b
[44] Pulkit Verma, Shashank Rao Marpally, and Siddharth Srivastava. 2022. Discovering User-Interpretable Capabilities of Black-Box Planning Agents. In *Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning* (Haifa, Israel). https://proceedings.kr.org/2022/36/
[45] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, John Quan, Stephen Gaffney, Stig Petersen, Karen Simonyan, Tom Schaul, Hado van Hasselt, David Silver, Timothy Lillicrap, Kevin Calderone, Paul Keet, Anthony Brunasso, David Lawrence, Anders Ekermo, Jacob Repp, and Rodney Tsing. 2017. StarCraft II: A New Challenge for Reinforcement Learning. *arXiv:1708.04782v1 [cs.LG]* (Aug. 2017). arXiv:1708.04782v1 [cs.LG]
[46] Khaing Phyo Wai, Minghong Geng, Shubham Pateria, Budhitama Subagdja, and Ah-Hwee Tan. 2024. Explaining Sequences of Actions in Multi-agent Deep Reinforcement Learning Models. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS '24)*. IFAAMAS, Richland, SC, 2537–2539.
[47] Khaing Phyo Wai, Minghong Geng, Budhitama Subagdja, Shubham Pateria, and Ah-Hwee Tan. 2023. Towards Explaining Sequences of Actions in Multi-Agent Deep Reinforcement Learning Models. In *Proc. of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '23)*. IFAAMAS, Richland, SC, 2325–2327.
[48] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. 2021. QPLEX: Duplex Dueling Multi-Agent Q-Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Rcmk0xxIQV
[49] Wenwen Wang, Budhitama Subagdja, Ah-Hwee Tan, and Janusz A. Starzyk. 2012. Neural Modeling of Episodic Memory: Encoding, Retrieval, and Forgetting. *IEEE Transactions on Neural Networks and Learning Systems* 23, 10 (Oct. 2012), 1574–1586. https://doi.org/10.1109/TNNLS.2012.2208477
[50] Xinzhi Wang, Huao Li, Hui Zhang, Michael Lewis, and Katia Sycara. 2020. Explanation of Reinforcement Learning Model in Dynamic Multi-Agent System. https://doi.org/10.48550/arXiv.2008.01508 arXiv:2008.01508v2 [cs]
[51] Lindsay Wells and Tomasz Bednarz. 2021. Explainable AI and Reinforcement Learning—A Systematic Review of Current Approaches and Trends. *Frontiers Artif. Intell.* 4 (2021), 550030. https://doi.org/10.3389/frai.2021.550030
[52] Herman Yau, Chris Russell, and Simon Hadfield. 2020. What Did You Think Would Happen? Explaining Agent Behaviour through Intended Outcomes. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 18375–18386. https://papers.nips.cc/paper/2020/hash/d5ab8dc7ef67ca92e41d730982c5c602-Abstract.html
[53] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and YI WU. 2022. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In *Adv. Neural Inf. Process. Syst.*, Vol. 35. Curran Associates, Inc., 24611–24624. https://proceedings.neurips.cc/paper_files/paper/2022/hash/9c1535a02f0ce079433344e14d910597-Abstract-Datasets_and_Benchmarks.html
[54] Huichu Zhang, Siyuan Feng, Chang Liu, Yaoyao Ding, Yichen Zhu, Zihan Zhou, Weinan Zhang, Yong Yu, Haiming Jin, and Zhenhui Li. 2019. CityFlow: A Multi-Agent Reinforcement Learning Environment for Large Scale City Traffic Scenario. In *WWW 2019*. ACM, 3620–3624. https://doi.org/10.1145/3308558.3314139