Toward Automated Operational Assist in Satellite Fleet Management via Organizational MARL

Julien Soulé
SEDAN Research Group – SnT – University of Luxembourg
Luxembourg, Luxembourg
julien.soule@hotmail.fr

Abstract

Earth-observation satellite fleets are increasingly difficult to operate with limited mission-control staff in a highly evolving spatial environment while facing mounting challenges such as inherent uncertainty, time constraints, communication disruptions, cyber threats, resource limits, and collision risks. We propose to address this global challenge through an automated operational-assist approach with two complementary contributions: (i) a fully tunable environment that abstracts the main features of ground operations, enabling controlled, reproducible fleet-level experimentation across diverse scenarios; (ii) an explicit decision architecture combining organizational reinforcement learning to improve coordination while preserving safety, controllability, and explainability through trajectory-level analysis, yielding actionable recommendations and partial automation to support operators. Compared with handcrafted, planning-style, and unconstrained learning baselines, our approach improves long-horizon operational performance and safety compliance while providing useful diagnostics for operator audit. These results support a progressive path from simulation benchmarks to supervised operational-assist deployment.

CCS Concepts

• **Computing methodologies** → **Machine learning**.

Keywords

Satellite Fleet Management, Organizational Multi-Agent Reinforcement Learning, Multi-Agent Systems, Explainability

ACM Reference Format:
Julien Soulé. 2026. Toward Automated Operational Assist in Satellite Fleet Management via Organizational MARL. In *Appears at the International Workshop on Autonomous Agents and Multi-Agent Systems for Space Applications (MASSpace-26). Held as part of the Workshops at the 25th International Conference on Autonomous Agents and Multiagent Systems., Paphos, Cyprus, May 2026*, IFAAMAS, 9 pages.

1 Introduction

Earth-observation satellite fleets are increasingly difficult to operate with limited mission-control staff in a rapidly evolving space environment [12, 24, 26]. Operational teams must coordinate under uncertainty and time pressure while handling communication disruptions, cyber threats, resource limits, and collision risks. This creates a strong need for automated operational-assist systems that support decision-making without removing human oversight.

Throughout this paper, we use several key concepts. *Operational-assist* refers to decision-support systems that augment human operators without removing their authority. *Organizational constraints* are explicit roles and missions that guide agent behavior beyond reward optimization. *Role specialization* captures differentiation of agent behaviors to improve coordination efficiency. *Mission-phase discipline* describes structured progression between operational phases (acquire, deliver, stabilize) rather than greedy behavior.

Operationally, the target behavior is to maximize useful data delivered to the ground while preserving fleet health. A central tension is that serving observation tasks and delivering value are decoupled phases: agents can be locally productive yet globally ineffective if buffered data is not relayed within time constraints. This acquire–deliver logic motivates explicit coordination.

From a multi-agent perspective, each satellite can be modeled as an autonomous agent interacting with teammates in a partially observable setting. This naturally motivates Dec-POMDP formulations and Multi-Agent Reinforcement Learning (MARL) methods [20, 21, 23, 31]. However, raw MARL performance is insufficient for deployment, since operational practice also requires explicit controllability, traceable safety compliance for operator audit [1, 3, 11].

To address this global challenge, we propose two complementary contributions. First, we introduce *Orbital Resilient Benchmark for Interactive Task-aware Autonomous Learning* (ORBITAL), a tunable environment for controlled, reproducible fleet-level experimentation. Second, we propose a decision-making architecture combining decentralized autonomy, organizational constraints, and human oversight. The core control layer uses organizational reinforcement learning grounded in $\mathcal{M}OISE^{+}$ and MOISE+MARL [15, 16, 28] to improve coordination while preserving safety, controllability, and explainability through trajectory-level analysis.

Compared with handcrafted and unconstrained learning baselines, results indicate that our approach improves long-horizon operational performance and safety compliance, while providing actionable diagnostics for operator audit [28]. These findings support a progressive deployment path from simulation benchmarks to supervised operational-assist deployment.

Section 2 reviews related work; Section 3 presents technical background; Section 4 introduces ORBITAL and the decision-making architecture; Section 5 details the experimental protocol; Section 6 reports results; and Section 7 concludes with a deployment path.

Appears at the International Workshop on Autonomous Agents and Multi-Agent Systems for Space Applications (MASSpace-26). Held as part of the Workshops at the 25th International Conference on Autonomous Agents and Multiagent Systems., S. Chien, G. Picard, I. Zilberstein (Chairs), May 2026, Paphos, Cyprus. © 2026 Copyright held by the owner/author(s).

2 Related work

Satellite Operations and Decision Support Satellite planning and scheduling have been extensively studied in operations research, including Earth-observation mission planning, agility constraints, and integrated allocation/scheduling formulations [9, 12, 24]. These works provide strong optimization baselines and realistic constraint modeling. However, they usually focus on centrally optimized plans and are less oriented toward adaptive, decentralized, and continuously learning operational-assist loops.

In parallel, agent-based autonomy has long been investigated for spacecraft constellations [26]. This line of work supports distributed decision-making, but often without a unified benchmark that jointly emphasizes modern MARL evaluation, explicit organizational control, and operator-oriented explainability.

Digital Twin and Simulation for Space Systems Recent work on digital twins for space systems highlights their potential for system validation, software testing, and predictive decision support [6, 19]. Nevertheless, many current approaches target specific subsystems or engineering workflows. For fleet-level autonomy research, there remains a need for reproducible environments that are simple enough for controlled MARL experimentation while progressively extensible toward higher realism.

MARL, Safety, and Explainability Cooperative MARL methods have significantly progressed, from actor-critic and value factorization methods to practical training recipes [10, 20, 23, 31]. Benchmarks and software ecosystems such as SMAC, PettingZoo, Gymnasium, and MARLlib improved reproducibility and comparison [14, 25, 29, 30]. However, these environments do not directly represent satellite operational constraints and domain-specific safety priorities.

Safety-aware reinforcement learning (RL) has introduced constrained optimization and shielding methods [1, 3, 11], while explainable RL research has proposed post-hoc and introspective analysis tools [22, 27]. Yet, these strands are often developed separately from organizational modeling and from domain-specific operational-assist requirements.

Organizational Modeling and Organizational MARL Organizational MAS modeling (roles, groups, missions, deontic constraints) is well established through Agent-Group-Role (AGR) and $\mathcal{M}OISE^+$ [8, 15, 16]. Building on this foundation, MOISE+MARL integrates organizational constraints into MARL and uses trajectory-based post-analysis to assess organizational alignment [28]. This direction is promising for controllability and explainability, but it has not yet been fully studied in the context of satellite-fleet operational-assist settings with an explicit benchmarking perspective.

Overall, no single line of work jointly satisfies high operational relevance for satellite fleets, adaptive multi-agent learning, explicit organizational control, and operator-oriented explainability. This motivates our contributions, comprising the operationally grounded benchmark ORBITAL for satellite fleet assistance and the decision-making architecture built on organizationally constrained MARL for safety-aware controllability and trajectory-level analysis for actionable explainability.

3 Background

This section provides the technical background we built on.

3.1 Cooperative Dec-POMDP

We formalize fleet-level decision-making as a Dec-POMDP [4, 21]. This framework is suitable for ORBITAL-like settings where agents act under local observability, decentralized execution, and team-level objectives.

A Dec-POMDP instance is:

$$d = \langle S, \{A_i\}_{i=1}^n, T, R, \{\Omega_i\}_{i=1}^n, O, \gamma \rangle,$$

where S is the latent state space, A_i and Ω_i are local action and observation spaces for agent i, T is the transition kernel, O the observation kernel, R a cooperative reward function, and $\gamma \in [0, 1]$ the discount factor.

Let $\pi_i(a_i \mid \tau_i)$ denote the local policy of agent i over local action-observation history τ_i. The joint policy is $\pi = (\pi_1, \ldots, \pi_n)$ and optimizes expected return:

$$V(\pi) = \mathbb{E}_{\pi,T,O}\left[\sum_{t=0}^{H-1} \gamma^t r_t\right].$$

In ORBITAL, this objective already embeds operational trade-offs (service, energy, communication, resilience), but by itself it does not guarantee role specialization, safety-compliant behavior, or interpretability.

3.2 Organizational modeling with $\mathcal{M}OISE^+$

To encode controllable coordination semantics, we rely on $\mathcal{M}OISE^+$ [15, 16]. Its key contribution is to separate organization into complementary layers:

(1) **Structural specifications**: roles and role relations (specialization or inheritance-like structures).
(2) **Functional specifications**: goals and missions that structure collective progress.
(3) **Deontic specifications**: permissions and obligations linking roles to missions under conditions.

This layered representation is important because it separates *what* should be done (functional), *by whom* (structural), and *under which normative constraints* (deontic), instead of collapsing everything into scalar reward coefficients.

3.3 MOISE+MARL as organizational control layer

MOISE+MARL [28] injects organizational knowledge into MARL while keeping standard MARL backbones usable. The integration relies on three families of guides:

(1) **Role-action guides** (RAG): constrain or prioritize actions based on role and trajectory context.
(2) **Role-reward guides** (RRG): penalize role-inconsistent decisions.
(3) **Goal-reward guides** (GRG): reward mission-consistent progress patterns.

In the following, we use the abbreviations RAG, RRG, and GRG for readability.

For agent i at time t, the effective decision/reward mechanism can be summarized as:

$$a_{i,t} \sim \pi_i(\cdot \mid \tau_{i,t}) \text{ over } \tilde{A}_{i,t} = rag(h_{i,t}, o_{i,t}),$$

$$\tilde{r}_t = r_t + \sum_{m \in \mathcal{M}_{i,t}} grg_m(h_t) + rrg(h_{i,t}, o_{i,t}, a_{i,t}).$$

The result is a hybrid control paradigm in which learning remains data-driven while search is guided by explicit priors. This can reduce unsafe exploration regions and improve policy controllability, particularly in partially observable cooperative settings [5, 28].

3.4 Trajectory-based analysis for explainability

Even if policies are trained with organizational constraints, one still needs post-hoc evidence of the actual learned behavior. The **Trajectory-based Evaluation in MOISE+MARL (TEMM)** method [28] addresses this gap. TEMM operates on multi-episode trajectories and infers implicit organizational regularities. At a high level: i) infer structural regularities (role-like behavior clusters);; ii) infer functional regularities (goal/mission progression patterns);; iii) compare inferred and intended structures to quantify alignment.

This process yields quantitative interpretable artifacts through: $\text{OF} = \alpha \cdot \text{SF} + (1 - \alpha) \cdot \text{FF}$, where structural fit (SF) captures role consistency, functional fit (FF) captures mission/goal consistency, and organizational fit (OF) summarizes both dimensions.

In ORBITAL-like environments, pure return maximization can produce brittle policies that exploit local shortcuts while degrading long-term viability. Typical failure modes include energy collapse, communication fragmentation, or cyber-sensitive behavior such as persisting in observation or relay actions while compromise indicators are high. The combination of Dec-POMDP formalization, organizational constraints, and trajectory-level organizational analysis provides the conceptual foundation needed to evaluate policies not only by performance but also by controllability, safety compliance, and explainability.

4 Method

This section describes our two coupled contributions, namely (i) ORBITAL, an operationally grounded benchmark for satellite fleet operational-assist, and (ii) a decision-making architecture relying on an ORBITAL-oriented MOISE+MARL, which injects organizational control and interpretability into MARL policies.

4.1 The ORBITAL environment

ORBITAL [1] is intentionally designed as a benchmark for *operational-assist* use, not as a full-fidelity orbital propagator. The objective is to keep the environment simple enough for controlled MARL experimentation while preserving the interaction structure that makes fleet management difficult in practice.

ORBITAL offers two geometric modes: a default 2D orbital abstraction with states (θ, r), as shown in Figure 1, and a 3D mode that includes an inclination variable ϕ for enhanced visualization, illustrated in Figure 2. While the 3D mode increases realism, it does not alter the core decision problem of balancing task servicing, delivery, and safety under uncertainty. Thus, the 2D mode is used as the primary view for clearer comparisons across conditions.

Design requirements The benchmark was designed around five requirements derived from our research problem: i) represent **observation-task pressure** (dynamic and priority-sensitive tasks),; ii) represent **resource pressure** (energy-limited long-horizon decisions),; iii) represent **communication uncertainty** (time-varying connectivity),; iv) represent **cyber uncertainty** (degraded sensing/acting/relaying),; v) represent **conjunction-risk pressure** from orbital debris fields. This is why ORBITAL combines non-stationary observation tasks, finite energy with heterogeneous action costs, stochastic communication degradation, stochastic compromise events, and drifting debris clouds that induce local conjunction-risk signals.

4.2 Reference Operational Scenario

We evaluate this architecture on a reference Earth-observation scenario inspired by agile LEO constellation operations. The setup considers eight satellites distributed over three orbital shells, dynamic observation demands with priorities, intermittent downlink opportunities, and coupled safety pressures (energy depletion, communication isolation, cyber degradation, and conjunction-risk proxy from debris density).

The intent is not to replicate a specific industrial mission one-to-one, but to preserve the operational doctrine that matters for assistive autonomy: *acquire–deliver–stabilize*. In this doctrine, observation throughput alone is insufficient if delivery windows are missed or if safety margins are consumed too aggressively.

Table 1: Reference operational scenario (used for all main comparisons).

Aspect	Scenario choice
Constellation geometry	8 cooperative satellites over 3 LEO shells
Mission demand	Non-stationary observation tasks with dynamic priority
Delivery model	Inter-satellite relay and intermittent ground-station windows
Safety pressures	Energy limits, link degradation, compromise events, debris-risk proxy
Success criterion	High delivered value with bounded risk and no mission collapse

Operational state and coupling At time t, the latent state includes satellite positions, energy levels, buffered data, compromise timers, active observation-task set, and communication adjacency matrix. Satellites are coupled through shared observation tasks, shared communication paths to ground, and shared team-level reward. This coupling produces the coordination tension we need to study because individual actions affect both local and fleet-level viability.

Observation and action modeling ORBITAL uses fixed-size local observations (16-dimensional vectors) and a compact discrete action space of size 7 (`Observe`, `Relay`, `OrbitDown`, `OrbitUp`, `LowPower`, `CyberScan`, `Idle`). The fixed vector format supports reproducible MARL pipelines and avoids benchmark bias toward a specific architecture. The action set was chosen to reflect the minimum operational primitives needed for operational-assist settings, including task servicing, data return, mobility, energy management, security response, and fallback behavior. These interface choices are

[1] An implementation of ORBITAL and the conducted experiments, including all details (organizational specifications, hyperparameters, and architectures), are available at https://github.com/julien6/ORBITAL.git.

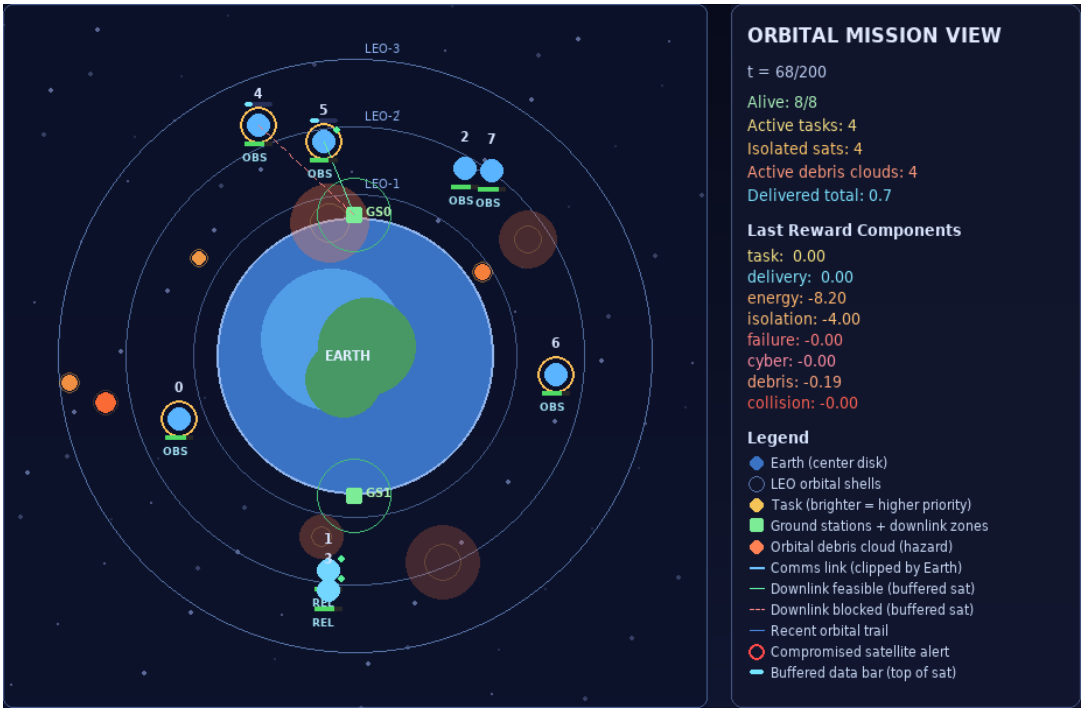

Figure 1: Screenshot of the ORBITAL environment (t = 68/200) in its orbital representation. Earth is shown at the center, surrounded by three low-Earth-orbit (LEO) shells (LEO-1 to LEO-3). Eight cooperative satellites (indexed markers) evolve on orbital states (θ, r), with role tags (`OBS`/`REL`) and per-satellite buffered-data bars. Orange markers denote active observation tasks (brightness proportional to priority). Green squares indicate ground stations (GS0, GS1) with downlink windows. Orange translucent halos denote orbital debris clouds and conjunction-risk zones. Inter-satellite communication links and downlink rays are line-of-sight constrained and clipped by Earth (no signal through Earth); feasible buffered downlinks are shown in green, blocked ones in dashed red. The right panel (*ORBITAL MISSION VIEW*) reports mission/safety indicators (alive satellites, active tasks, isolated satellites, active debris clouds, delivered total) and the per-step reward decomposition (`task`, `delivery`, `energy`, `isolation`, `failure`, `cyber`, `debris`, `collision`), highlighting the trade-off between mission productivity, connectivity, and conjunction-safe resilience.

intentionally minimal: fixed-size vectors improve reproducibility across MARL families, the 7-action set captures core operational primitives, and both Agent Environment Cycle (AEC) and Parallel application programming interfaces (APIs) are available with consistent semantics.

Task-delivery decoupling A critical modeling decision is to separate *observation-task servicing* from *value delivery*. `Observe` converts local task opportunities into buffered data, but mission value is maximized only when data is later relayed through available communication opportunities toward ground. For example, a satellite may continue observing tasks while its buffer is full and no relay opportunity is available. While this maximizes local productivity, it degrades global mission performance since collected data cannot be delivered. In parallel, a local debris-density signal yields a simplified conjunction-risk proxy (*Pc*-like indicator) that can be reduced through orbital maneuver actions (`OrbitDown`/`OrbitUp`). We use this proxy as an operational trigger variable: if risk remains high across successive steps, the architecture should shift from productivity-focused actions toward safety-preserving actions. This separation forces policies to balance sensing throughput, topology management, delivery timing, and collision-risk mitigation rather than greedily optimizing local sensing only.

Reward design for operational trade-offs Default reward mode is shared team reward with positive terms for mission productivity and penalties for resilience degradation: $r_t = w_{\text{task}}\, c_{\text{task}} + w_{\text{delivery}}\, c_{\text{delivery}} - w_{\text{energy}}\, c_{\text{energy}} - w_{\text{isolation}}\, c_{\text{isolation}} - w_{\text{failure}}\, c_{\text{failure}} - w_{\text{cyber}}\, c_{\text{cyber}} - w_{\text{debris}}\, c_{\text{debris\_risk}} - w_{\text{collision}}\, c_{\text{collision}}$.

This reward structure is intentionally non-myopic, since maximizing return requires balancing service, survivability, connectivity, cyber resilience, and conjunction-risk control over the whole horizon. ORBITAL also provides a local reward mode for controlled comparisons.

Episode termination and realism scope Episodes stop at horizon or mission collapse (no alive satellites or critically low survivability). This choice makes unsafe policies self-limiting in long runs. ORBITAL remains a simplified environment, but it preserves the operational couplings required to evaluate coordination doctrine, safety control, and human-supervised assist behavior.

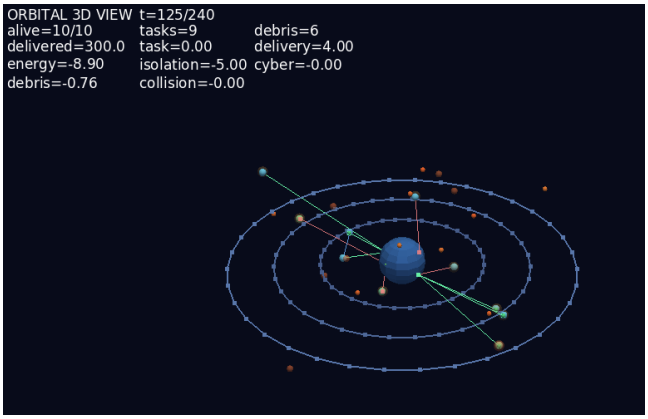

Figure 2: ORBITAL 3D mode screenshot. The same operational entities are represented in 3D (Earth, LEO shells, satellites, ground stations, debris, and communication/downlink links), with the same mission semantics as in the 2D mode.

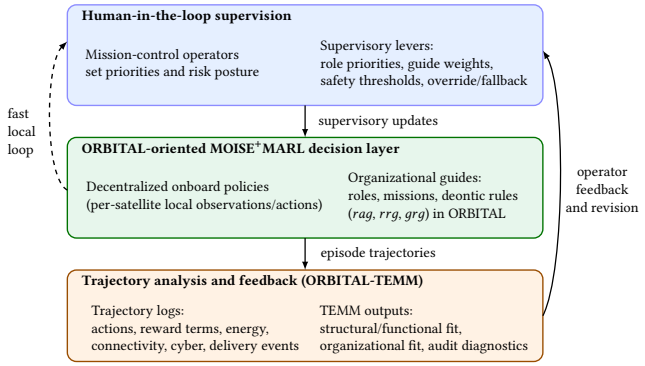

Figure 3: Operational-assist architecture based on an ORBITAL-oriented MOISE$^+$MARL. Human supervision sets priorities and safety posture; decentralized policies are constrained by organizational guides; ORBITAL-TEMM analyzes trajectories and returns audit-oriented feedback for supervisory updates and corrective override.

4.3 ORBITAL-oriented decision-making architecture

In the ORBITAL setting, static rule-based scheduling systems remain easy to inspect and certify. However, they tend to be brittle when facing evolving observation demand, fluctuating inter-satellite connectivity, and stochastic cyber or operational disturbances, as commonly observed in large-scale Earth observation constellations [32]. Distributed Constraint Optimization Problem (DCOP) formulations provide an explicit framework for decentralized coordination and property-preserving optimization in such scenarios [18], enabling safety guarantees.

However, extending DCOP-based approaches to support long-horizon adaptation under partial observability and non-stationary uncertainties typically requires non-trivial model extensions and proactive mechanisms. Adaptation is therefore less direct than in learning-based control [13]. Vanilla MARL is highly adaptive to dynamic environments because it learns directly from interaction data. However, it lacks explicit mechanisms to enforce structured behaviors such as: (i) role specialization (e.g., some satellites focusing on observation while others relay data), and (ii) mission-phase discipline (e.g., prioritizing delivery when buffers are saturated). In such cases, reward shaping alone is often insufficient to ensure consistent and safe behavior [7]. Organizational MARL is therefore adopted here as a middle ground. It enables preserving data-driven adaptation while injecting declarative role and mission constraints that improve system-level controllability, interpretability, and post-hoc auditability of agent behaviors [28].

Therefore, building on the MOISE+MARL [28] framework, we propose an ORBITAL-oriented organizational architecture [1] aligned with mission-control practice (Figure 3).

Architecture overview The architecture is organized into three coupled layers with different timescales. The top layer is human-in-the-loop supervision, where operators set mission priorities and safety posture through high-level levers (role priorities, guide weights, safety thresholds, and fallback policy). The middle layer is decentralized onboard decision-making under organizational constraints (roles, missions, and deontic rules). The bottom layer is trajectory analysis (ORBITAL-TEMM), which produces interpretable diagnostics for audit and corrective feedback.

Two control loops The design features a fast local loop and a slower supervisory loop. The fast loop allows each satellite to act based on local observations and organizational guides, while the supervisory loop updates high-level guidance without direct teleoperation. This structure ensures reactivity and maintains operator authority. For example, if communication quality declines and buffered data increases, supervisors can enhance delivery guidance (prioritizing relay roles), prompting a shift towards relay actions. Conversely, if compromise and isolation indicators rise, supervisors can tighten safety thresholds, encouraging safety-guard behavior and minimizing risky actions. In both scenarios, intervention is high-level, while action selection remains decentralized.

Feedback and auditability Trajectory logs (actions, reward components, energy, connectivity, cyber events, delivery events) are analyzed by ORBITAL-TEMM to estimate structural/functional alignment and organizational fit, then returned to supervision for the next revision cycle.

Organizational mapping We map ORBITAL operational intents to $\mathcal{M}OISE^+$-style specifications:

(1) **roles**: behavioral specialization templates,
(2) **goals**: trajectory-level objectives with measurable progress,
(3) **missions**: coherent groups of goals in operational phases,
(4) **deontic rules**: permissions/obligations linking roles to missions.

This mapping gives a declarative control layer that can be inspected and revised independently from the MARL backbone. Table 2 summarizes the role profile used in this paper.

Constraint guides adapted to ORBITAL The ORBITAL adaptation ties guide logic to operational indicators available in trajectories and observations, including local task opportunity, buffered data

Table 2: ORBITAL-oriented organizational specifications (high-level view).

Role	Main mission focus	Typical obligations/permissions
Observer	Prioritized acquisition	Obliged to observe feasible high-priority tasks; permitted to defer low-priority opportunities under strong resource pressure
Relay	Delivery continuity	Obliged to relay when buffered data and communication opportunities are favorable; permitted to adapt relay cadence under congestion
Safety guard	Fleet survivability	Obliged to prioritize safe actions under elevated risk (energy depletion, isolation, compromise); permitted to override non-critical productivity actions

pressure, communication degree, energy margin, compromise status, and recent delivery events.

Mission semantics We structure behavior around three mission families: i) **Acquisition mission**: increase task servicing under feasibility constraints.; ii) **Delivery mission**: convert buffered data into delivered value reliably.; iii) **Resilience mission**: preserve fleet viability under resource/cyber stress. Role-specific permissions and obligations switch emphasis between these missions according to context. This produces a controllable trade-off between productivity and safety rather than relying on implicit reward.

Hard and soft control regimes The adaptation supports two control regimes. In hard regimes, RAG can enforce action masking so that non-authorized actions are unavailable. In soft regimes, actions remain selectable but violations are discouraged through RRG. This hardness axis is useful to tune exploration freedom versus organizational compliance.

ORBITAL-adapted TEMM We integrate an ORBITAL-specific TEMM process [28] as post-training validation and explainability. Trajectories include action traces, reward components, energy evolution, communication context, task-service and delivery events, and cyber-state transitions. TEMM is applied in three stages: i) infer structural regularities corresponding to implicit role behavior;; ii) infer functional progression patterns corresponding to missions/-goals;; iii) compare inferred and intended specifications to quantify alignment. We report organizational fit as $\mathrm{OF} = \frac{1}{2} \cdot (\mathrm{SF} + \mathrm{FF})$.

5 Experimental setup

This section describes the experimental protocol.

5.1 Experimental goals

We target four experimental questions: i) **Q1 (Performance):** Does ORBITAL-oriented MOISE+MARL improve long-horizon mission execution compared with unconstrained learning and handcrafted coordination?; ii) **Q2 (Control/Safety):** Does organizational guidance reduce unsafe or mission-degrading behavior under resource, communication, and cyber stress?; iii) **Q3 (Robustness):** Does the approach remain effective under intensified non-stationarity (task dynamics, link drops, compromise rate)?; iv) **Q4 (Explainability):** Do TEMM-based organizational indicators provide coherent and actionable post-hoc analysis of learned trajectories?

To answer these questions, each experiment combines: i) one environment configuration (nominal or stressed),; ii) one learning/control condition (baseline or proposed method),; iii) one MARL backbone,; iv) multiple independent random seeds. All protocol parameters and seeds are fixed before result aggregation.

5.2 Hardware and Software Configuration

Hardware profile All experiments were conducted on an academic high-performance computing (HPC) cluster using heterogeneous GPU nodes, including NVIDIA A100, NVIDIA V100, and AMD MI210 devices in a Linux-based scheduling environment. We executed 5 parallel instances per algorithm-environment combination to efficiently explore the baseline and proposed-method parameter spaces while preserving reproducibility through fixed random seeds and deterministic hyperparameter selection on validation data before final evaluation runs.

Software stack The environment is implemented through PettingZoo-compatible APIs (see Table 3). The ORBITAL-oriented decision-making architecture reuses the implementation of MOISE+MARL [2] [28] (organizational specification layer and trajectory analysis), while specializing in role/mission logic for ORBITAL.

Table 3: Software configuration and role in the pipeline.

Component	Role in experiments
Python 3.10	Training and evaluation runtime
Pygame/PyVista	2D and 3D rendering
PettingZoo 1.25.0 [29]	Multi-agent APIs (AEC and Parallel interfaces)
Gymnasium 1.2.3 [30]	Standard RL spaces/wrapping compatibility
PyTorch [17]	Neural policy/value modeling and gradient-based optimization
Optuna [2]	Hyperparameter search and trial-based configuration selection
NumPy 2.2.6	Numerical operations and logging
ORBITAL codebase	Environment dynamics, reward components, rendering/debugging
MOISE+MARL [28]	Organizational guides, role/mission constraints, TEMM pipeline

5.3 Baselines and comparison conditions

Handcrafted baselines To represent operational heuristics, we include: i) **RB-Rule:** rule-based action selection prioritizing local high-priority tasks, then relay, with simple low-energy fallback.; ii) **RB-Relay-heavy:** heuristic emphasizing delivery continuity (relay when possible, observe otherwise), with weak safety adaptation.; iii) **PB-DCOP-lite:** inspired by DCOP decomposition, with periodic assignment of observer/relay intents under communication and energy constraints, then local heuristic execution. These baselines provide interpretable references as operational scripts.

Learning baselines We compare the proposed organizational framework to unconstrained MARL and partially constrained variants: i) **LB-Unconstrained:** same MARL backbone, ORBITAL reward only, no organizational guides.; ii) **LB-RewardOnly:** unconstrained action space, additional reward shaping but no role-action masking.; iii) **LB-ActionOnly:** role-action guidance active, no mission reward guides.; iv) **Proposed (ORBITAL-$\mathcal{M}OISE^+$ MARL):** role-action + role-penalty + mission guides with deontic assignments. This decomposition isolates where gains come from, whether through action control, reward structure, or full organizational coupling.

[2] Accessible at: https://github.com/julien6/MOISE-MARL.

Backbone algorithms To reduce algorithm-specific bias, each condition is evaluated with representative cooperative MARL families: i) actor-critic / policy-gradient style methods (MAPPO [31]),; ii) value-factorization methods (QMIX [23]),; iii) multi-agent actor-critic references (MADDPG/COMA style [10, 20]).

5.4 Ablation plan

To validate causal contributions of the proposed framework, we define the ablations in Table 4. We also stress-test each setting along controlled perturbation axes: i) increased task non-stationarity (spawn/priority volatility),; ii) increased communication degradation ($p_{\text{link\_drop}}$),; iii) increased compromise intensity (adversarial rate and duration),; iv) reduced energy budgets.

Table 4: Ablation settings for ORBITAL-oriented MOISE+MARL.

ID	Ablation description
A0	Full framework (role-action + role-penalty + mission guides + TEMM analysis)
A1	Remove role-action guide (rag off): no action-space organizational control
A2	Remove role-penalty guide (rrg off): no explicit role-violation penalty
A3	Remove mission guides (grg off): no mission-level shaping
A4	Soft-only control: no action masking, penalties/rewards only
A5	Hard-only control: action masking active, no additional role penalty
A6	TEMM feature reduction: remove cyber/context features in trajectory analysis

5.5 Evaluation metrics

We report control, and organizational interpretability metrics.

Mission performance metrics i) **Cumulative return**: episode return.; ii) **Task service volume**: serviced task-priority mass.; iii) **Delivery volume**: delivered data.; iv) **Mission completion rate**: episodes without mission collapse.

Safety and resilience metrics i) **Energy stress index**: low-energy occupancy over agents and time.; ii) **Isolation ratio**: alive agents with zero communication degree.; iii) **Failure count**: depleted satellites per episode.; iv) **Cyber impact score**: aggregate compromise-related penalties/events.

Control and explainability metrics i) **Constraint violation rate**: role-inconsistent action frequency.; ii) **Structural fit** and **functional fit** from ORBITAL-TEMM.; iii) **Organizational fit** using structural and functional organizational fits.; iv) **Consistency score**: intended vs inferred role/mission agreement.

Subjective operator-audit indicators Operator-facing indicators use an ordinal scale (`very low`, `low`, `high`, `very high`): i) **Mission alignment rating**: alignment between inferred motifs and expected mission logic.; ii) **Human-audit agreement rating**: agreement between TEMM diagnostics and human audit.

5.6 Training and evaluation protocol

We use disjoint seeds for training/selection/final evaluation and report aggregate metrics over at least 10 seeds per condition. Learning conditions share identical episode horizons and optimization budgets; hyperparameters are tuned on validation seeds only, then frozen for testing. We report mean, standard deviation, and results uncertainty intervals, with two-sided significance testing and effect sizes for pairwise comparisons. All runs log per-step reward components and mission/safety/organizational events, enabling targeted failure-mode analysis and reproducible reruns.

6 Results and discussion

We first compare strong, medium, and weak organizational-control regimes (Table 5). Strong constraints deliver the fastest early convergence, weak constraints preserve flexibility but slow learning and increase violations, and medium constraints provide the best global compromise. Medium control reaches the highest final return (379.8 ± 16.1), mission success (0.91 ± 0.03), and OF score (0.92 ± 0.02), while keeping convergence substantially better than weak control.

Relating these results to the gaps identified in Section 2, our contributions strongly cover adaptive multi-agent learning, explicit organizational control, and operator-oriented explainability at benchmark level: compared with handcrafted and unconstrained baselines, the proposed framework improves mission value and success, reduces role-inconsistent behavior, and increases trajectory-level interpretability/audit agreement. Coverage of the operational-realism gap is partial by design: ORBITAL captures the key coupled pressures required for operational-assist evaluation, but does not yet represent full flight-grade orbital and communication realism.

6.1 Baselines against handcrafted and learning conditions

Table 6 compares our method against handcrafted and learning baselines. As expected, handcrafted policies do not rely on organizational constraints and therefore violation metrics are marked `n/a`. Two implications stand out. First, handcrafted and planning-style policies remain interpretable and operationally plausible, but plateau at lower performance and resilience levels than the proposed framework. Second, partial organizational variants improve either control or speed, but the full framework is required to simultaneously optimize mission value, convergence, and robustness.

6.2 Convergence and robustness trade-off

The key behavioral result is a non-monotonic relation between control hardness and end-task quality. Overly hard constraints accelerate convergence but reduce robustness; overly weak constraints preserve robustness but underuse organizational priors. Medium constraints act as the best compromise, improving convergence and compliance without collapsing policy diversity.

6.3 Ablation results

Table 7 shows that each component contributes to at least one critical dimension. Removing RAG strongly reduces return and robustness, and increases violations. Removing GRG yields the largest mission-performance drop (-10.4% return). Removing RRG most strongly harms compliance, with the largest violation increase. Reducing TEMM features barely affects training performance but degrades interpretability quality, consistent with TEMM.

6.4 Explainability and trajectory analysis

Quantitative explainability indicators Table 8 shows that the proposed method achieves higher role-cluster separability (0.67 ± 0.04)

Table 5: Aggregate results by constraint regime (mean ± std over seeds). Higher is better for all metrics except convergence episode, violation rate, and energy stress.

Regime	Final return	Return AUC (early)	Convergence episode ↓	Robustness score	Constraint violation ↓	OF score	Mission success rate	Energy stress ↓
Strong constraints	352.6 ± 18.4	205.3 ± 9.8	118 ± 14	0.71 ± 0.05	1.8% ± 0.9	0.90 ± 0.03	0.86 ± 0.04	0.37 ± 0.06
Medium constraints	**379.8 ± 16.1**	192.7 ± 8.2	146 ± 17	**0.84 ± 0.04**	3.9% ± 1.2	**0.92 ± 0.02**	**0.91 ± 0.03**	**0.31 ± 0.05**
Weak constraints	334.9 ± 21.7	143.5 ± 11.6	213 ± 22	0.87 ± 0.04	8.7% ± 1.8	0.81 ± 0.05	0.83 ± 0.05	0.35 ± 0.07

Table 6: Baseline comparison. Handcrafted baselines report `n/a` for organizational violation metrics.

Condition	Final return	Convergence episode ↓	Robustness score	Violation rate ↓	OF score	Delivery volume	Mission success rate
RB-Rule (priority-first)	241.3 ± 12.9	n/a	0.63 ± 0.06	n/a	0.49 ± 0.07	112.4 ± 9.8	0.64 ± 0.08
RB-Relay-heavy	228.1 ± 14.7	n/a	0.67 ± 0.07	n/a	0.45 ± 0.08	121.7 ± 10.3	0.61 ± 0.07
PB-DCOP-lite	286.2 ± 15.8	n/a	0.73 ± 0.06	n/a	0.61 ± 0.06	138.9 ± 9.5	0.72 ± 0.06
LB-Unconstrained	319.7 ± 20.5	229 ± 24	0.82 ± 0.05	10.4% ± 2.1	0.74 ± 0.05	151.8 ± 11.7	0.79 ± 0.05
LB-RewardOnly	341.5 ± 19.2	188 ± 20	0.79 ± 0.05	7.1% ± 1.9	0.82 ± 0.04	162.3 ± 10.8	0.84 ± 0.04
LB-ActionOnly	348.0 ± 17.9	161 ± 18	0.76 ± 0.06	4.6% ± 1.5	0.85 ± 0.04	167.9 ± 10.1	0.86 ± 0.04
Proposed (medium constraints)	**379.8 ± 16.1**	**146 ± 17**	**0.84 ± 0.04**	3.9% ± 1.2	**0.92 ± 0.02**	**182.5 ± 9.4**	**0.91 ± 0.03**

Table 7: Ablation outcomes relative to full method.

Ablation	Δ Return	Δ Robustness	Δ Violation	Δ OF
A1 (no rag)	−8.6%	−6.2%	+2.7 pts	−0.07
A2 (no rrg)	−3.9%	−4.8%	+3.1 pts	−0.09
A3 (no grg)	−10.4%	−2.0%	+0.8 pts	−0.06
A4 (soft-only)	−4.7%	+1.1%	+1.9 pts	−0.05
A5 (hard-only)	−5.3%	−7.4%	−1.4 pts	−0.04
A6 (reduced TEMM features)	−0.2%	−0.1%	+0.0 pts	−0.11

compared to unconstrained learning (0.41 ± 0.06), indicating that organizational constraints enhance behavioral specialization. Audit-agreement ratings improve from `low` to `very high`, reflecting that TEMM patterns align closely with intended specifications.

Table 8: Explainability indicators.

Condition	Role-cluster separability	Mission alignment rating	Human-audit agreement rating
LB-Unconstrained	0.41 ± 0.06	`low`	`low`
LB-RewardOnly	0.49 ± 0.05	`high`	`high`
Proposed (medium)	**0.67 ± 0.04**	`very high`	`very high`

Qualitative and operational relevance Trajectory analysis confirms role/mission motifs in the proposed setting (observer acquisition-to-relay transitions, relay draining patterns, and safety-oriented stress responses), consistent with higher role-cluster separability in Table 8. TEMM outputs provide actionable traceability of role consistency and repeated deontic violations, with mission-alignment and audit-agreement ratings improving from `low` to `very high`.

6.5 Human-in-the-loop audit case

To evaluate auditability, we analyze one representative stress episode where communication degradation and compromise overlap. In the unconstrained condition, two satellites persist in `Observe` despite rising isolation and energy stress, which leads to buffered-data saturation and delayed delivery recovery. TEMM flags this pattern as repeated mission-role mismatch for relay-assigned agents.

Applying a supervised intervention changes the subsequent trajectory: one satellite transitions to relay behavior, one to safety-guard behavior, isolation duration is shortened, and delivery recovery occurs before mission-collapse threshold. Those observations confirmed the intended usage mode of the architecture, where the learning policy remains autonomous between updates, but operators retain structured levers for corrective control.

7 Conclusion

This paper introduced an operational-assist framework for satellite fleet management comprising: i) the ORBITAL environment as an operationally grounded abstraction in a reproducible multi-agent setting under various constraints; ii) and an ORBITAL-oriented decision-making architecture adapted from MOISE+MARL that combines organizational role/mission constraints with trajectory-based analysis to improve controllability and explainability.

The main result is a structured control trade-off: strong constraints accelerate early convergence but may reduce robustness, weak constraints preserve robustness but slow learning, and medium constraints provide the best compromise between performance, convergence speed, and resilience, suggesting that organizational specifications are most useful when guiding safety-critical behavior while preserving policy freedom for emergent cooperation.

However, ORBITAL is a simplified abstraction that lacks full communication, high-fidelity orbital mechanics, and operational complexity. The manually designed organizational specifications may introduce bias and limit scalability. TEMM analysis quality relies on trajectory features and clustering choices, necessitating careful calibration. This simplification may miss second-order effects like communication latency and orbital perturbations. Therefore, we identify three key directions: enhancing environment realism for better simulation of real-world conditions, integrating model-based reinforcement learning and multi-agent world models for improved long-horizon robustness, and reinforcing neuro-symbolic integration to align learned behaviors with safety constraints. Our goal is ultimately developing offline advisory support systems.

Acknowledgments

This DefenceTech project benefits from shared financial support by the Ministry of Foreign and European Affairs, Directorate of Defence, the Ministry of Economy and the Luxembourg National Research Fund (FNR) (DEFENCE24/IS/19272253/ALIAS).

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained Policy Optimization. In *Proceedings of the 34th International Conference on Machine Learning*. 22–31.

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. arXiv:1907.10902 [cs.LG] https://arxiv.org/abs/1907.10902

[3] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. 2018. Safe Reinforcement Learning via Shielding. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018). https://doi.org/10.1609/aaai.v32i1.11797

[4] Aurélie Beynier, François Charpillet, Daniel Szer, and Abdel-Illah Mouaddib. 2013. *DEC-MDP/POMDP*. John Wiley & Sons, Ltd, Chapter 9, 277–318. https://doi.org/10.1002/9781118557426.ch9

[5] Jiajun Chai, Zijie Zhao, Yuanheng Zhu, and Dongbin Zhao. 2025. A Survey of Cooperative Multi-Agent Reinforcement Learning for Multi-Task Scenarios. *Artificial Intelligence Science and Engineering* 1, 2 (2025), 98–121. https://doi.org/10.23919/AISE.2025.000008

[6] Andrea Colagrossi, Stefano Silvestrini, Andrea Brandonisio, and Michèle Lavagna. 2026. A Digital Twin Approach for Spacecraft On-Board Software Development and Testing. *Aerospace* 13, 1 (2026). https://doi.org/10.3390/aerospace13010055

[7] Sam Devlin and Daniel Kudenko. 2016. Plan-Based Reward Shaping for Multi-Agent Reinforcement Learning. *The Knowledge Engineering Review* 31, 1 (2016), 44–58. https://doi.org/10.1017/S0269888915000181

[8] Jacques Ferber, Olivier Gutknecht, and Fabien Michel. 2003. Agent/Group/Roles: Simulating with Organizations. In *ABS 2003 - 4th International Workshop on Agent-Based Simulation*, J.P. Muller (Ed.). Montpellier, France. https://hal-lirmm.ccsd.cnrs.fr/lirmm-00269714

[9] Benedetta Ferrari, Jean-François Cordeau, Maxence Delorme, Manuel Iori, and Roberto Orosei. 2025. Satellite Scheduling Problems: A survey of applications in Earth and outer space observation. *Computers & Operations Research* 173 (2025), 106875. https://doi.org/10.1016/j.cor.2024.106875

[10] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual Multi-Agent Policy Gradients. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018). https://doi.org/10.1609/aaai.v32i1.11794

[11] Javier Garcia and Fernando Fernandez. 2015. A Comprehensive Survey on Safe Reinforcement Learning. *Journal of Machine Learning Research* 16, 42 (2015), 1437–1480. http://jmlr.org/papers/v16/garcia15a.html

[12] Lorenzo Govoni, Corrado Chiatante, Bjørn Andreas Kristiansen, Tor Arne Johansen, and Andrea Cristofaro. 2026. Optimization-Based Task Allocation for Earth Observation in Multi-Satellite Systems. *Aerospace Science and Technology* 168 (2026), 111058. https://doi.org/10.1016/j.ast.2025.111058

[13] Khoi D. Hoang, Ferdinando Fioretto, Ping Hou, William Yeoh, Makoto Yokoo, and Roie Zivan. 2022. Proactive Dynamic Distributed Constraint Optimization Problems. *Journal of Artificial Intelligence Research* 74 (2022), 179–225.

[14] Siyi Hu, Yifan Zhong, Minquan Gao, Weixun Wang, Hao Dong, Xiaodan Liang, Zhihui Li, Xiaojun Chang, and Yaodong Yang. 2023. MARLlib: A Scalable and Efficient Multi-agent Reinforcement Learning Library. *Journal of Machine Learning Research* 24, 315 (2023), 1–23. http://jmlr.org/papers/v24/23-0378.html

[15] Jomi Fred Hübner, Jaime Simão Sichman, and Olivier Boissier. 2002. A Model for the Structural, Functional, and Deontic Specification of Organizations in Multiagent Systems. In *Advances in Artificial Intelligence*, Guilherme Bittencourt and Geber L. Ramalho (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 118–128.

[16] Hubner, Jomi F et. al. 2007. Developing organised multiagent systems using the MOISE+ model: programming issues at the system and agent levels. *Int. Journal of Agent-Oriented Software Engineering* (2007), 370. https://doi.org/10.1504/ijaose.2007.016266

[17] Sagar Imambi, Kolla Bhanu Prakash, and G. R. Kanagachidambaresan. 2021. *PyTorch*. Springer International Publishing, Cham, 87–104. https://doi.org/10.1007/978-3-030-57077-4_10

[18] Shai Krigman, Tal Grinshpoun, and Lihi Dery. 2024. Scheduling of Earth Observing Satellites Using Distributed Constraint Optimization. *Journal of Scheduling* 27, 5 (2024), 507–524. https://doi.org/10.1007/s10951-024-00816-x

[19] Wei Liu, Mengwei Wu, Gang Wan, and Minyi Xu. 2024. Digital Twin of Space Environment: Development, Challenges, Applications, and Future Outlook. *Remote Sensing* 16, 16 (2024). https://doi.org/10.3390/rs16163023

[20] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *Advances in Neural Information Processing Systems* 30 (2017).

[21] Frans A. Oliehoek and Christopher Amato. 2016. *A Concise Introduction to Decentralized POMDPs*. Springer. https://link.springer.com/book/10.1007/978-3-319-28929-8

[22] Erika Puiutta and Eric M. S. P. Veith. 2020. Explainable Reinforcement Learning: A Survey. Springer-Verlag, Berlin, Heidelberg, 77–95. https://doi.org/10.1007/978-3-030-57321-8_5

[23] 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 *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 4295–4304. https://proceedings.mlr.press/v80/rashid18a.html

[24] Yure Rocha, Guilherme O. Chagas, Leandro C. Coelho, and Anand Subramanian. 2025. The integrated agile Earth observation satellite scheduling problem. *Computers & Operations Research* 184 (2025), 107212. https://doi.org/10.1016/j.cor.2025.107212

[25] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philiph H. S. Torr, Jakob Foerster, and Shimon Whiteson. 2019. The StarCraft Multi-Agent Challenge. *CoRR* abs/1902.04043 (2019).

[26] Thomas Schetter, Mark Campbell, and Derek Surka. 2000. Multiple Agent-Based Autonomy for Satellite Constellations. In *Agent Systems, Mobile Agents, and Applications*, David Kotz and Friedemann Mattern (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 151–165.

[27] Pedro Sequeira and Melinda Gervasio. 2020. Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations. *Artificial Intelligence* 288 (2020), 103367. https://doi.org/10.1016/j.artint.2020.103367

[28] Julien Soulé, Jean-Paul Jamont, Michel Occello, Louis-Marie Traonouez, and Paul Théron. 2025. An Organizationally-Oriented Approach to Enhancing Explainability and Control in Multi-Agent Reinforcement Learning. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (**AAMAS** 2025)* (Detroit, MI, USA). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1968–1976.

[29] J K Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, Niall Williams, Yashas Lokesh, and Praveen Ravi. 2021. PettingZoo: Gym for Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 15032–15043. https://proceedings.neurips.cc/paper_files/paper/2021/file/803f7c4c3ff61b71be53a0c803bfb57f-Paper.pdf

[30] Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. 2025. Gymnasium: A Standard Interface for Reinforcement Learning Environments. arXiv:2407.17032 [cs.LG] https://arxiv.org/abs/2407.17032

[31] 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 *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 24611–24624. https://proceedings.neurips.cc/paper_files/paper/2022/file/9c1535a02f0ce079433344e14d910597-Paper-Datasets_and_Benchmarks.pdf

[32] Itai Zilberstein and Steve Chien. 2026. Large-Scale Continual Scheduling and Execution for Dynamic Distributed Satellite Constellation Observation Allocation. arXiv:2601.06188 [cs.AI] https://arxiv.org/abs/2601.06188