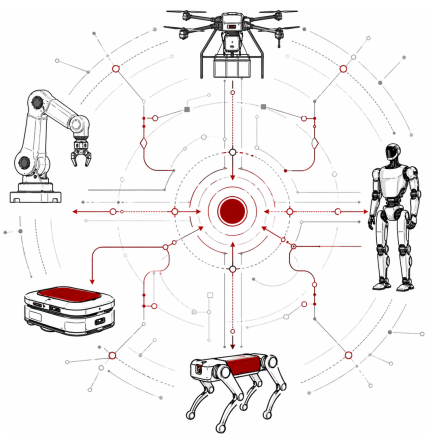
Xuezhi Niu

Enhancing Cooperation in Heterogeneous Multi-Robot Systems

Symbiosis-Inspired Representations of Directed Partner Dependence



Abstract

Heterogeneous robot teams are valuable because their members are not interchangeable, but that same non-interchangeability creates a coordination problem: whether one robot's action advances the team or merely consumes its own time and energy often depends on how the action changes a partner's resources, feasible actions, or future task value, rather than on the acting robot's local observation alone. This half-time report studies that directed, state-dependent effect, which it calls *directed partner dependence*, and asks how compact representations of it can improve decentralised coordination when introduced at three layers of the decision process: partner state in the observation, partner effect in the reward, and partner selection in the interaction topology. Ecological symbiosis contributes only a vocabulary for the sign of these effects (mutualism, commensalism, parasitism, neutrality); it is not treated as a biological explanation of robot behaviour or as a general theory of cooperation.

Four studies instantiate this representation hierarchy. At the observation layer, partner battery-state sharing in a two-AMR warehouse reduces task completion time by 10.7% and improves workload/energy balance by 13.81% relative to static recharging. At the reward layer, task-specific mutualistic reward terms in Isaac Sim/Isaac Lab have little effect on peak performance in CartPendulum, but are associated with smoother learning in the more coupled Shadow Hand and mobile-manipulation tasks, and PRISM converts event-based relationship proxies into potential-based reward shaping, reaching 40.2 deliveries per episode in TA-RWARE, 48.3% above a flat-cooperative reward and 97.1% above a task-only reward under the tested settings. At the topology layer, MORPH maintains a plastic pairwise preference graph online; in a 20-seed, 2,000-step TA-RWARE scale study its throughput is generally comparable to task-shared and full-graph baselines while using fewer active links, though a proximity baseline is significantly better at the largest tested scale.

These results support bounded claims about the tested state representations, reward functions, and topology-update rules. They do not establish biological fidelity, universal solver independence, deployment readiness, or general robustness. The second half of the PhD will therefore focus on stronger causal diagnostics, comparisons with planning and communication baselines, transfer across tasks and platforms, and validation on physical heterogeneous robot systems.

List of papers

This half time report is based on the following papers, which are referred to in the text by their Roman numerals.

I **Xuezhi Niu** and Didem Gürdür Broo. PRISM: Policy-shaping via Reward decomposition for Inter-agent Symbiosis in Multi-Robot Cooperation *under submission to Journal*

II **Xuezhi Niu** and Didem Gürdür Broo. MORPH: Self-Organising Multi-Robot Task Allocation via Neuroplasticity-Inspired Adaptive Topology. *under submission to conference*

III **Xuezhi Niu** and Didem Gürdür Broo. Investigating Symbiosis in Robotic Ecosystems: A Case Study for Multi-Robot Reinforcement Learning Reward Shaping. In *2025 9th International Conference on Robotics and Automation Sciences (ICRAS)*. IEEE, 2025.

IV **Xuezhi Niu**, Natalia Calvo Barajas and Didem Gürdür Broo. Enabling Symbiosis in Multi-Robot Systems Through Multi-Agent Reinforcement Learning. In *2025 IEEE 8th International Conference on Industrial Cyber-Physical Systems (ICPS)*. IEEE, 2025.

Reprints were made with permission from the publishers.

List of papers not included during Half Time

This section lists papers produced during the PhD that are *not* included in the thesis.

- Alexandros Rouchitsas, **Xuezhi Niu** and Didem Gürdür Broo. Dynamics of Human Adaptability to Robot Effectiveness and Efficiency Malfunctions in Collaborative Manufacturing. *under submission to Journal*

- Jing Xu, **Xuezhi Niu**, Didem Gürdür Broo, Klas Hjort. *TouchDrive: Electronics-Free Tactile Sensing Interface for Assistive Grasping*. Accepted for presentation at the RoboTac workshop, IEEE ICRA, 2026.

- Jing Xu, **Xuezhi Niu**, Jacob Andersson, Didem Gürdür Broo, Klas Hjort. Miniaturized Multifunctional Valves for Intelligent Pneumatic Systems in Soft Robotics. *under submission to Journal*

- Alexandros Rouchitsas, **Xuezhi Niu**, Ginevra Castellano and Didem Gürdür Broo. "What do I do now?": Spontaneous Human Responses to Robot Effectiveness and Efficiency Malfunctions in Collaborative Robotics. In *The ACM Conference on Human Factors in Computing Systems (CHI)*, ACM, 2026.

* Equal contribution

Contents

1. Introduction

Heterogeneous robot teams are valuable precisely because their members are not interchangeable. A mobile transporter carries what a stationary manipulator cannot reach; a picker completes a handling operation that an autonomous transporter cannot perform; and a robot with charge to spare preserves team capacity while another must recharge [1–3]. Yet that same non-interchangeability creates a coordination problem that local observations and team-level rewards routinely hide: whether one robot's action advances the team or merely consumes its own time and energy often depends less on the acting robot's own observation than on how the action changes a partner's feasible actions, resource state, or future task value. As heterogeneous teams move from laboratory demonstrations toward persistent warehouse, manufacturing, inspection, and field operations, each robot must answer three questions during execution: which partners currently matter, what information about those partners is necessary, and whether maintaining an interaction benefits or burdens the participants.

Current abstractions do not answer these questions directly. Task-allocation and planning methods optimise assignments when the relevant capabilities, costs, and constraints are specified in advance; cooperative multi-agent reinforcement learning (MARL) optimises a shared return; and communication or coordination graphs decide which agents exchange information. None of these necessarily exposes the directed, state-dependent relation that connects one robot's decision to another robot's outcome. This report studies that missing relation, which it calls *directed partner dependence*: the directed, state-dependent effect of one robot on another robot's resources, feasible actions, cost, and task value. Direction matters because the effect of robot i on robot j need not equal the effect of j on i; state dependence matters because the same pair can help each other under one resource condition and contend under another.

The central hypothesis of the thesis is that compact representations of directed partner dependence, introduced at the observation, reward, and interaction-topology layers, can preserve cross-role value that local observations and undifferentiated team signals omit. Its sharpest instance is the joint evaluation of coupled decisions before a robot is committed or removed, which recovers value that separated, locally optimal decisions discard; but the hypothesis is broader, covering any decision mechanism that makes a partner-dependent consequence explicit. Ecological symbiosis supplies a compact vocabulary (mutualism, commensalism, parasitism, neutrality) for the sign and direction of these effects, and the report keeps this vocabulary on a *framing* layer that carries no causal claim of its own, while every measurable claim

lives on a *mechanism* layer defined independently through task performance, resource use, reward signals, and graph structure. Symbiosis names the relations; coupling earns the results. The report is deliberately explicit about where its evidence stops, but it states those boundaries at the point each claim is made rather than in advance, so that the scientific object and hypothesis stand before their qualifications.

1.1 Context and Motivation

Multi-robot coordination can be organised through centralised planning [4], distributed consensus [5], graph-based control [6, 7], or multi-agent reinforcement learning (MARL) [8]. Centralised methods can optimise joint decisions from a global model but depend on reliable communication and stable team composition; decentralised methods execute locally but can miss cross-agent dependencies when observations or rewards omit information about partners. Heterogeneity sharpens this trade-off, because role, capability, and resource differences decide which joint actions are feasible and useful at any given moment. This report works within the MARL setting, not because learning is always preferable to planning, but because repeated interaction is the natural place to study how cooperative behaviour is shaped by the observed consequences of one robot's actions on another, rather than fully specified in advance.

The difficulty this report targets is that, in this setting, cooperation cannot be read off the usual learning signals. A fully shared reward encourages team performance but does not identify which relationship or event produced it; independent rewards preserve local incentives but undervalue support actions whose effect appears mainly in a partner's progress [9, 10]. Neither signal separates the *type* of interaction (mutual benefit, one-sided assistance, dependency, redundant co-presence, or interference) from its *magnitude*. For interchangeable agents this distinction is often secondary; for heterogeneous robots it is decisive, because a support action that drains a scarce battery or lengthens a route is worth taking only when its directed benefit to a partner exceeds the cost it imposes. Cooperation here is therefore a sequential, decentralised decision problem under shared constraints, in which the consequence of each local action depends on partners that are neither fixed parts of the environment nor interchangeable players.

What makes this problem tractable rather than merely hard is that the directed effect, although not exactly observable, can be approximated from quantities a decentralised robot already has. An exact measure would require counterfactual evidence, such as rerunning the same state with one partner removed or replaced, which is unavailable during online training. The interventions in this report instead approximate directed effects from observable signals: selectively exposed partner state, event-level relationship proxies inserted through potential-based shaping that leaves the optimal policy unchanged [11], and pair-

wise preferences re-estimated online from co-assignment and task-completion feedback. High-fidelity simulators for warehouse logistics and mobile manipulation make it possible to isolate each intervention through matched ablations, which is why the question can be studied rigorously now even though physical validation remains future work.

The report contributes four interventions that approximate one common relation object, the directed partner effect formalised in Chapter 2, at three successive layers of the decision process. This gives the studies a single spine, a *representation hierarchy* that moves the same partner dependence from observation, to reward, to interaction structure: (i) partner state in the observation, (ii) partner effect in the reward, and (iii) partner selection in the topology. Study IV operates at the observation layer, augmenting a decentralised learner's state with a partner's battery level and reporting faster task completion and better workload/energy balance in a two-AMR warehouse. Study III and Study I operate at the reward layer: ICRAS adds task-specific mutual-benefit terms and examines their effect on learning stability across coupled control and mobile-manipulation tasks, while PRISM converts event-level relationship proxies into potential-based shaping and tests the resulting throughput against matched reward baselines. Study II (MORPH) operates at the topology layer, maintaining a plastic pairwise preference graph re-estimated online from co-assignment, delivery, observation, and degree-regulation signals. The studies manipulate different objects and use different algorithms and simulators, so the report does not present them as one progressively refined algorithm; what makes them cumulative is that each is an approximation of the same directed-dependence quantity at a different representation layer. The present evidence supports this hierarchy only for the tested simulated tasks, and the remainder of the report states precisely where that evidence stops.

1.2 Symbiosis as a Conceptual Lens

Symbiosis supplies the vocabulary for signed, directed interaction, but only the vocabulary. In ecology, symbiosis denotes persistent interaction between dissimilar organisms whose effect on each participant may be positive, negligible, or negative, distinguished as mutualism, commensalism, parasitism, or competition [12]. These effects are context dependent, changing with resource availability, environmental conditions, and population state. The report borrows these four labels for robot pairs, with one deliberate restriction: a pair is not symbiotic by assumption. Given only a set of robots and tasks, complementarity is potential; the realised relationship exists only after an action or event changes a partner's task progress, resource use, feasible options, or cost, and it is estimated from exactly those quantities.

The lens is useful because heterogeneous cooperation is asymmetric. When one robot relays information, waits for a partner, carries an object, exposes

a resource state, or accepts a longer route, it can improve team performance while bearing an uneven share of the cost. Recording the direction and sign of such effects makes this allocation visible in a way a single team-level score cannot. It does not follow that every useful interaction is symbiotic, or that ecological population dynamics transfer to robots; the analogy motivates the questions and names the relations, and nothing more.

This restriction is decisive for the present evidence, and the report states it up front. Study III tests only task-specific mutualistic reward terms, and Study II produces operational relationship labels from event proxies rather than from partner-removal counterfactuals. The completed studies therefore do not validate commensalism or parasitism as general robot-interaction categories; those labels remain hypotheses for the asymmetric-cost and resource-contention experiments planned for the second half.

1.3 Research Gap

Existing work represents cooperation in heterogeneous teams through three dominant abstractions, each strong within its scope, but none of which provides a single directed, state-dependent relation expressing how robot i changes robot j's future task value, resource state, or feasible actions. The gap this report addresses is precisely that missing relation, and the three abstractions are reviewed in turn to locate it.

The first abstraction is *assignment and feasibility*. Gerkey and Matarić [13] formalised multi-robot task allocation as a coordination problem, Choi et al. [14] showed how decentralised auctions produce conflict-free assignments, and more recent work extends this to heterogeneous capabilities, temporal constraints, coalition requirements, and long-endurance missions [4, 6, 15, 16]. These methods decide which robots should perform which tasks, but they encode inter-agent effects indirectly through costs and constraints fixed at allocation time, so once execution begins a change in a partner's energy or availability alters how one robot affects another without changing the nominal assignment. How such a partner-state change should enter another robot's decision *during* learning is left open, and is the question Study IV examines at the observation layer.

The second abstraction is *aggregate team value*. Cooperative MARL treats cooperation as credit assignment under a shared return: Lowe et al. [9] and Foerster et al. [10] estimate agent-level contributions from joint information, while Sunehag et al. [17] and Rashid et al. [18] decompose team value into agent-wise utilities, and related work reshapes the signal through social influence, intrinsic rewards, contribution redistribution, or role structure [19–22]. These methods improve how much credit each agent receives for a team outcome, but a scalar contribution does not identify the *type* of directed interaction, support, burden, dependency, redundancy, or interference, that produced it.

Representing that directed effect in the reward is the question Studies III and I address at the reward layer.

The third abstraction is *interaction connectivity*. Nguyen et al. [3] proposes pairwise collaboration mechanisms for heterogeneous teams and Nguyen et al. [23] connects ecological mutualism to multi-robot collaboration, establishing pairwise interaction as a useful scale. However, the resulting collaboration or communication structures are often fixed, manually specified, or tied to predefined pairing rules, whereas partner relevance changes as agents are assigned, observe one another, complete deliveries, consume resources, or become redundant. Whether the interaction structure itself should be re-estimated online from such signals is the question Study II addresses at the topology layer. Read together, the three abstractions supply assignment, aggregate value, and connectivity, but not the directed relation itself; the four studies are its approximations at the observation, reward, and topology layers.

1.4 Aim and Objectives

The aim of this report is to represent directed partner dependence and use it to improve coordination in selected heterogeneous multi-robot tasks, by studying how explicit representations of partner state, relational contribution, and pairwise preference affect learning and execution; ecological symbiosis contributes only the vocabulary for the sign of each effect. Rather than treating cooperation only as task allocation, shared-reward maximisation, or fixed collaboration links, the report examines how one robot affects another through task progress, resource use, cost, and future task feasibility. Six objectives follow from this aim:

1. evaluate whether selected partner-state information improves resource-aware coordination;
2. design and compare task-specific and relationship-aware reward-shaping formulations;
3. define operational proxies for directed relational contribution and state their validity limits;
4. develop an online pairwise preference graph that adapts collaboration links during execution;
5. compare these interventions against matched reward, state, topology, and heuristic baselines; and
6. identify the additional evidence required before broader claims across tasks, simulators, and physical platforms are justified.

These objectives lead to the research questions below, which organise the studies completed so far and identify the work still needed beyond this half-time report.

1.4.1 Research Questions and Study Mapping

The thesis is guided by one technical main question and three sub-questions, one for each representation layer. The main question is stated in terms of the scientific object rather than the ecological framing, which enters only as vocabulary for the sign of the effect. Each sub-question is posed compactly through four elements: the unknown, why existing methods are insufficient, the intervention this report makes, and the evidence that would settle it; the supporting surveys are developed in Chapter 2.

Main RQ: How can directed partner dependence be represented and used to improve decentralised coordination in heterogeneous multi-robot systems under changing roles, resources, and interaction partners?

RQ1 (observation layer). *Which compact representations of partner state and partner relevance are sufficient for decentralised robots to respond to resource- and role-dependent task coupling?* The unknown is how little partner information suffices: each robot acts from a local observation, and partner relevance is state dependent, so a partner may matter for its energy, role, or task state rather than its position. Full-state exposure would approximate centralised execution and inflate communication, while fixed interaction graphs go stale as batteries, demand, and role availability change; neither answers the sufficiency question. This report intervenes with battery-state augmentation (ICPS) and an online pairwise preference graph (MORPH), treating observation and communication cost as evaluation criteria rather than as free. Settling RQ1 requires a sufficiency study over partner variables, which the completed work begins but does not complete.

RQ2 (reward layer). *When do directed pairwise event or contribution signals improve learning beyond reward-density-matched shared or local reward formulations?* The unknown is whether the *type* of interaction adds information over the *amount* of credit: the same team return can arise from support, burden, dependency, redundancy, or interference, and a scalar contribution does not distinguish them. Shared returns and independent rewards therefore leave directed, asymmetric effects unrepresented, and exact directed effects need counterfactual evidence unavailable during online training. This report intervenes with task-specific mutual-benefit terms (ICRAS) and potential-based relational shaping from event proxies (PRISM). Settling RQ2 requires reward-density-matched controls that separate relation semantics from shaped-reward density, which is a second-half priority.

RQ3 (validity axis). *Which properties of directed partner-dependence representations remain stable when task distribution, team composition, solver family, sensing quality, communication conditions, and embodiment change?* The unknown is which conclusions are invariant rather than artefacts of one simulator, reward scale, or backbone: a proxy meaningful in one layout may fail where co-occurrence does not imply cooperation, and a benefit under one solver family or in simulation need not survive sensing noise, latency, or real

energy dynamics. Existing validation practice supplies matched baselines and seeds but rarely tests cross-solver or physical transfer. This report establishes construct and internal validity within each study and identifies external validity as the principal gap. Settling RQ3 requires cross-solver, cross-task, and physical experiments, which are the core of the second half.

Table 1.1. *Mapping from representation layer and research question to the four studies. RQ3 is a cross-cutting validity axis; the MORPH scale and task-shift study is its principal current probe.*

Layer	Study	Intervention	RQ	Evidence boundary
Observation	ICPS	Partner battery-state sharing in value-decomposition MARL	RQ1	Two AMRs in one simulated warehouse setting
Reward	ICRAS	Task-specific symbiosis-inspired reward terms in MAPPO	RQ2	Three Isaac Sim/Isaac Lab tasks; mutualistic terms only
Reward	PRISM	Event-based relationship proxies in potential-based reward shaping	RQ2	TA-RWARE with IPPO/MAPPO and matched reward baselines
Topology	MORPH	Online plastic pairwise preference graph	RQ1, RQ3	TA-RWARE scale and cyclic task-shift studies

1.5 Scope and Limitations

The work focuses on interaction-level modelling and learning in heterogeneous multi-robot systems. Validation is simulation based and covers warehouse logistics, coupled control, object passing, and mobile manipulation. The report compares formulations within matched experimental substrates; it does not claim superiority over all task-allocation, communication-learning, model-predictive, or mixed-integer planning methods. Symbiosis is not treated in this report as a controller, a model-free learning algorithm, or a complete dynamics model of multi-robot interaction. It does not specify the kinematics of the robots, the transition dynamics of the environment, or the full process by which cooperation emerges from an arbitrary system. Its role is more limited and more specific.

The evidence is also uneven across studies. The ICPS study uses two AMRs, the ICRAS study uses five seeds, PRISM is evaluated with PPO-based learners in TA-RWARE, and MORPH is evaluated as an online coordination mechanism rather than an end-to-end learned policy. Cross-study comparisons are therefore interpretive rather than controlled comparisons of a common algorithm.

Potential-based policy invariance is invoked only under the standard assumptions required by the shaping construction, including a Markov-compatible potential and appropriate treatment of terminal states [11]. Event-level relationship proxies may correlate with useful co-participation without identifying causal contribution. MORPH's predictive formation signal uses nearest-picker positional information, so the current implementation should not be described as free of spatial priors.

1.6 Ethical Considerations

The reported experiments do not involve human participants, personal data, or physical robot deployment. The principal ethical obligation is accurate interpretation. Efficiency and resilience claims must remain bounded by the tested simulations, and designer-defined benefit functions should be explicit and auditable. Before physical deployment, the methods would require additional validation for communication failures, sensing uncertainty, actuator limits, safety constraints, and unequal allocation of workload or risk.

2. State of the Art

Cooperation in heterogeneous robot teams can be planned, learnt, or understood through the effects that robots have on one another. This chapter reviews coordination methods, MARL, and ecological symbiosis to identify what is already well established and where current approaches still leave inter-agent dependencies implicit. The resulting gap concerns how these dependencies can be made observable, represented in the learning process, and adapted during decentralised operation.

2.1 Heterogeneous Multi-Robot Cooperation

A heterogeneous multi-robot system is a team whose members differ in properties that affect task execution, so that the robots are not interchangeable and the value of one robot can depend on which others are present [24, 25]. Two largely orthogonal questions organise the field. The first asks *what* differs between robots, which the literature splits into physical and behavioural heterogeneity [1]. The second asks *how* decisions are organised, ranging over a control spectrum from centralised, through multi-centralised (hierarchical or cluster-based) and distributed, to fully decentralised execution [25, 26], and over an increasingly tight cooperation taxonomy of coordination (division of works), cooperation (a shared objective for a team), and collaboration (tightly coupled joint execution from heterogeneous teams). These two dimensions are largely independent: a physically heterogeneous team may be controlled centrally or learn decentralised policies, while physically identical robots may develop specialised behaviours. This section formalises heterogeneity along both dimensions and isolates the problem the report studies, namely how directed inter-agent dependencies can be made observable, represented in the learning process, and adapted during execution.

2.1.1 Heterogeneity in Capabilities, Resources, and Roles

Existing taxonomies classify the *agents*. The dominant split is physical versus behavioural heterogeneity [1], alongside capability and eligibility structures from task allocation [13, 25] and broader physical, behavioural, and task-induced distinctions [24]. These answer whether two robots differ and how, but not what one robot *does to* another while a task runs. A capability difference

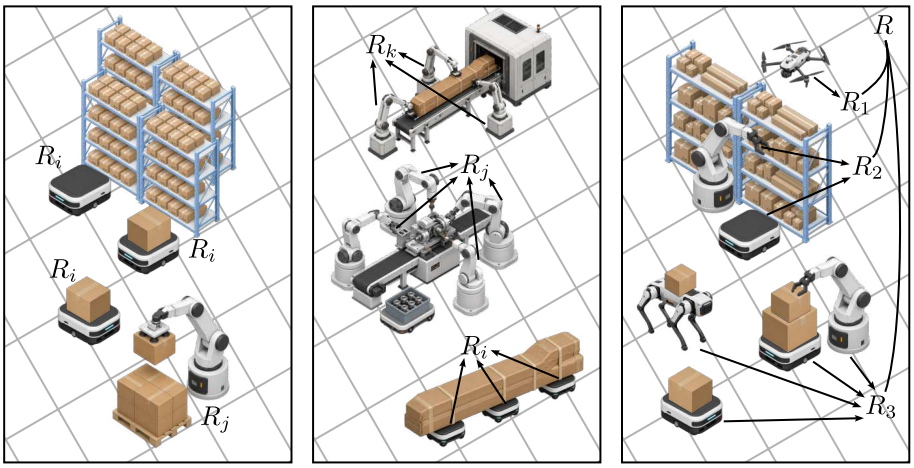


Figure 2.1. Coordination, cooperation, and collaboration in a heterogeneous warehouse robot team. Coordination distributes local objectives R_i, cooperation aligns complementary actions toward a team objective, and collaboration requires tightly coupled contributions toward a shared objective R. Robot roles may change across tasks and operating conditions as in cooperation and collaboration.

matters only when it changes the feasible actions, costs, resources, or outcomes of another robot during task execution. This report therefore treats directed inter-agent effects as the operational expression of heterogeneity. Each relation is labelled by the sign of the change it produces in the operational fitness of the two partners, following the ecological convention from Section 2.4. Table 2.1 presents this taxonomy, its operational interpretation in a robot team, and the study in which each relation is represented. This relational view might complement attribute-based classifications by showing how the same pair of robots can interact differently as tasks and resource conditions change. Even robots that are physically identical can behave asymmetrically if one has less energy, carries a different load, or takes on a supporting role. Conversely, robots with different physical characteristics may benefit each other in one task but create an imbalance in another. In this way, heterogeneity reflects not only differences in design or policy, but also the changing effects robots have on one another depending on the situation.

Heterogeneity is beneficial when it creates useful complementarity rather than difference for its own sake. Bettini et al. [1] show that heterogeneous policies can address behavioural requirements that homogeneous parameter-shared policies may fail to capture, while Bettini et al. [27] and Prorok et al. [28] associate diversity with exploration, role specialisation, sparse-reward coordination, and recovery from disturbances. These gains remain task dependent, since parameter sharing can suppress agent-specific behaviour [1],

Table 2.1. *Operational taxonomy of directed inter-agent relations in robot teams and their representation in the included studies.*

Relation	Effect (i, j)	Operational meaning in a robot team	Study
Mutualism	$(+,+)$	both robots gain feasibility, throughput, or energy; the target relation	I, II, III, IV
Commensalism	$(+,0)$	one robot is helped at no measurable cost to the other	I, III
Parasitism	$(+,-)$	one robot gains by burdening or draining the other; a failure mode to detect and suppress	–
Neutralism	$(0,0)$	no measurable directed effect; the link carries no benefit and can be pruned	I, II

indiscriminate diversity promotion can distort the learning objective [29], and heterogeneous effort allocation is useful only under suitable reward structures [30]. Physical and behavioural heterogeneity are independent but interacting, as identical robots may learn different policies while different robots may perform similar roles. The operational profile $h_i = [C_i, D_i, E_i, P_i]$ complements this distinction by separating capability, resource state, information access, and task performance. Capability C_i defines what a robot can perform, while D_i constrains whether, for how long, and at what cost those functions remain available. Information access E_i captures sensing and knowledge asymmetries, and P_i records their task consequences. This profile provides measurable variables for analysing what should be shared, with whom, when, and at what cost. Table 2.2 groups representative subclasses according to the part of the operational profile they primarily affect.

Within this profile, a role is the task- and interaction-dependent function through which a robot uses its capabilities, resources, and information. Roles may change with task conditions and need not correspond to embodiment. This parallels the ecological niche as a functional position within a community, while P_i indicates how effectively that position is realised [83, 84]. From a task-allocation perspective, Korsah et al. [25] classify how tasks are assigned to robots according to agent-task dependencies, scheduling structure, and allocation constraints. Functional responsibilities therefore follow from feasible task assignments rather than from an explicit role model. In the socially organised systems of Fagiolini et al. [60], each robot switches between discrete manoeuvres according to local events and social rules, so its current role is expressed through the behaviour expected under the active interaction context. Distributed monitors then assess whether that behaviour is consistent with the prescribed role despite partial and time-varying visibility. This formalises and verifies role compliance, but the roles and transition rules are defined in advance rather than learnt or adapted from changing resources, task outcomes, or partner-specific effects. In MARL, Wang et al. [22] introduce latent role representations that condition specialised policies and group agents according to similar behavioural responsibilities. Building on this formulation, Wang et al.

Table 2.2. *Representative subclasses of heterogeneity under the operational profile $h_i = [C_i, D_i, E_i, P_i]$. C_i denotes capability, morphology, or task eligibility; D_i denotes resource and constraint state; E_i denotes information access, communication, sensing, or interface structure; and P_i denotes task-specific performance or objective. The table uses representative citations from the local bibliography; categories may overlap.*

Heterogeneity class	Operational meaning	Representative MRS emphasis
$C \setminus (D \cup E)$	Capability, morphology, embodiment, or task eligibility differs.	Reconfiguration, allocation, morphology transfer, mapping, embodiment-aware operation, heterogeneous-policy learning, and team composition [1, 4, 24, 31–35].
$D \setminus (C \cup E)$	Energy, charging, endurance, availability, or resource state differs.	Energy-aware information gathering, monitoring, persistence, rendezvous charging, long-term operation, AGV resource management, energy optimisation, and mobile charging [7, 36–42].
$E \setminus (C \cup D)$	Sensing, interface, communication, map access, or information differs.	Sensing platforms, remote sensing, sensor fusion, multimodal interfaces, communication surveys, networked estimation, sensor-net navigation, attentional communication, and active information gathering [43–51].
$C \cap D$	Capability differences interact with resource, temporal, or uncertainty constraints.	Replenishment, charging, robust scheduling, cellular planning, conflict resolution, quadruped–wheeled teaming, rescue diversity, temporal-logic resources, resilience, and energy-aware allocation [2, 15, 16, 52–58].
$C \cap E$	Capability differences couple to communication, monitoring, autonomy, or interoperability.	Hierarchical optimisation, monitors, heterogeneous policy networks, pairwise collaboration, collaborative autonomy, mapping, interoperability, and CPS heterogeneity [3, 59–65].
$D \cap E$	Resource constraints couple to sensing, communication, remote operation, or information sharing.	Battery switching, remote operation, communication-stabilised UAV exploration, UAV offloading, energy-efficient wireless clustering, data-driven navigation, information sharing, and sparse POMDPs [66–73].
$C \cap D \cap E$	Capability, resource state, and information all affect coordination or learning.	Heterogeneous-agent RL, RL scheduling, graph MARL, subteaming, social learning, dynamic coalitions, payload-aware consensus, reaching, rearrangement, and timed-goal planning [8, 74–82].

[85] decompose the joint problem into role-specific action spaces according to action-effect similarity and separate slower role selection from faster action execution. These approaches support emergent specialisation and reduce the effective decision space, but role identity is inferred mainly from behaviour or action structure and does not directly indicate when resource depletion, information loss, changing task demands, or partner-specific effects make a previously suitable robot unable or inefficient to continue in the same role.

2.1.2 Inter-Agent Dependencies in Cooperative Tasks

A cooperative robot task is a coupled decision process: each robot acts from local observations, but its action can change another robot's state, information, resources, or feasible actions. This setting can be written as a heterogeneous partially observable Markov game, closely related to Dec-POMDP and POMG models for multi-robot decision-making under uncertainty [86–88],

$$\mathcal{G} = \langle N, S, \{A_i\}_{i \in N}, T, \{\Omega_i\}_{i \in N}, O, \{r_i\}_{i \in N}, \gamma \rangle. \tag{2.1}$$

where $N = 1, \ldots, n$ is the agent set, S is the state space, A_i and O_i are the action and observation spaces of agent i, $T(s_{t+1} \mid s_t, \mathbf{a}_t)$ is the transition kernel, $O_i(o_{i,t} \mid s_t)$ is the observation kernel, $R(s_t, \mathbf{a}_t, s_{t+1})$ is the shared team reward, and $\gamma \in [0, 1)$ is the discount factor. At time t, each agent receives a local observation $o_{i,t}$, selects an action according to $a_{i,t} \sim \pi_i(\cdot \mid o_{i,t})$,, and contributes to the joint action $\mathbf{a}_t = (a_t^i, \ldots, a_t^n)$. The environment then transitions according to

$s_{t+1} \sim T(\cdot \mid s_t, \mathbf{a}_t)$, and all agents receive the team reward $r_{t+1} = R(s_t, \mathbf{a}_t, s_{t+1})$. Exact solution is intractable in general, so practical coordination relies on structural assumptions, approximation, or learning [86, 87]. The cooperative objective is to learn a joint policy $\pi = (\pi_1, \ldots, \pi_n)$ that maximises the expected discounted team return

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \right]$$

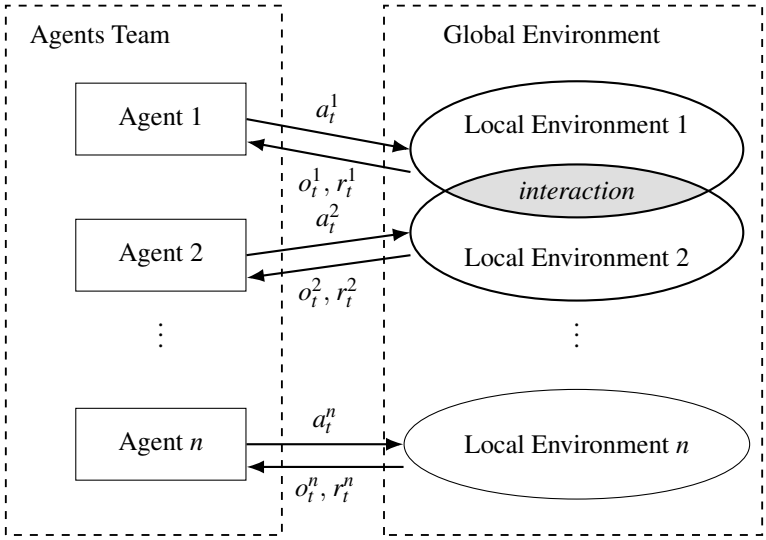


Figure 2.2. Agent interactions with overlapping local environments in a global environment.

A directed dependency is more specific than a communication edge or a fixed coupling coefficient. It is a state- and policy-dependent effect from robot j to robot i, measured against a selected operational outcome. Let $J_i^H(t;\alpha)$ denote robot i's expected outcome over horizon H when robot j takes action α at time t and subsequent transitions and actions follow T and π. Depending on the task, J_i^H may measure progress, energy-adjusted utility, delay, feasibility, safety, or return. The directed effect is $\Delta_{j\to i,t} = J_i^H(t;a_t^j) - J_i^H(t;\bar{a}_t^j)$, where $\bar{a}_t^j$ is a counterfactual reference action. Positive, negative, and near-zero values indicate support, burden, and negligible effect. This follows counterfactual reasoning used in credit assignment and causal-influence analysis [10, 19, 89], but here the quantity is kept as a signed robot-to-robot relation rather than only as a training-time credit signal.

During decentralised execution, robot i cannot observe the full state or evaluate the counterfactual directly. It therefore estimates the dependency from its local history and the operational profiles of the two robots, $w_{ij,t} = \mathbb{E}[\Delta_{j\to i,t} \mid \tau_{i,t}, h_i, h_j]$, where $\tau_{i,t}$ is its action-observation history and h_i, h_j de-

scribe the robots' operational profiles. This single object, the directed effect $\Delta_{j\to i,t}$ and its decentralised estimate $w_{ij,t}$, is what unifies the four studies: each is one approximation of the same quantity, and they differ only in how it is instantiated and where it is inserted into the decision process.

- **ICPS** exposes the partner resource state on which the effect depends, $z_j = b_j$, in the observation $o_i' = [o_i, z_j]$, and lets the decentralised learner infer $\Delta_{j\to i}$ from it.
- **ICRAS** encodes the effect directly as a designer-specified reward term, $r_i' = r_i^{\text{task}} + F_{ij}$, with F_{ij} chosen per task rather than estimated.
- **PRISM** approximates $\Delta_{j\to i}$ from event-level co-participation proxies mapped into a bounded potential Φ_i, inserting it through potential-based shaping.
- **MORPH** estimates the accumulated historical utility of selecting j, the preference $W_{ij,t}$, from co-assignment and completion signals, and adapts the interaction structure online.

Reading the studies as instances of one relation object is what makes them a cumulative programme rather than four unrelated interventions. None of them fits the expectation in the displayed equation directly; fitting $w_{ij,t}$ as a regression onto measured counterfactual effects, so that the estimator itself is validated rather than a hand-designed proxy, is the keystone study reserved for the second half of the thesis. Pairwise modelling is moreover a deliberate approximation: heavy tasks, shared chargers, and congestion can induce higher-order dependencies in which $\Delta_{j\to i|k} \neq \Delta_{j\to i}$, so $W_t = [w_{ij,t}]$ captures the pairwise projection of a structure that is not, in general, purely pairwise. Within that approximation, the matrix W_t summarises the latent, time-varying, signed, and directed dependencies in the team. Existing methods capture parts of this structure: communication graphs preserve links, role methods preserve specialisation, value decomposition preserves contribution magnitude, and causal-influence methods estimate directed effects [17, 22, 89, 90]. The missing element is an executable representation that keeps direction, sign, and state dependence available during decentralised operation.

2.2 Coordination in Heterogeneous Robot Teams

Heterogeneous multi-robot coordination converts differences in sensing, mobility, manipulation, endurance, and payload into executable assignments, schedules, and interaction structures. The need for such coordination is most visible when no single platform can cover the complete task envelope. In aerial-ground mapping and exploration, heterogeneous robots combine complementary viewpoints, mobility modes, and sensing capabilities to improve spatial coverage and information acquisition [68, 91, 92]. Bernreiter et al. [63] also develops a unified framework for coordinating heterogeneous robots in exploration and mapping, showing how platform-specific capabilities and shared environmental

information can be incorporated into a common planning structure. In search and rescue, coordination must additionally account for communication limits, risk exposure, accessibility, and the assignment of reconnaissance and intervention roles [54, 57]. Jorgensen and Pavone [93] formulates heterogeneous robot allocation under structured capability constraints, illustrating how combinatorial task requirements can be retained during coordination rather than reduced to interchangeable agent assignments. Related problems arise in persistent monitoring and informative path planning under energy, sensing, and latency limits [7, 36, 51, 94]; precision agriculture and biodiversity monitoring, where fleets combine mobility, sensing, sampling, harvesting, and environmental data collection [43, 44, 95–98]; and construction, assembly, and warehouse logistics, where task precedence, congestion, resource availability, and platform-specific manipulation capabilities constrain execution [99–103]. Tan et al. [81] addresses scalable warehouse coordination by representing interactions among robots and tasks within a structured coordination model, while Chen and Kan [104] shows how heterogeneous task requirements can be expressed and updated through formal specifications during online planning. The strongest coupling appears in physical collaboration, where robots jointly manipulate a payload or one robot directly enables the action of another. Cooperative and mobile manipulation distributes load and motion constraints across multiple arms or mobile bases [105–107], whereas Feng et al. [108] couples legged locomotion with manipulation during pushing and carrying tasks. Zhou et al. [109] and Christen et al. [110] further study assistance and hand-off behaviours in which an action by one robot changes the feasibility or progress of a task assigned to another robot. These cases make the directed dependency explicit: an action by robot i changes the progress, cost, resource state, or feasible action set of robot j. The relevant question is therefore not only how tasks are assigned, but whether the coordination method represents and preserves the sign and direction of dependencies as they change during execution. Table 2.3 summarises these application settings. Human-robot teams form a substantial parallel literature [111, 112], but this report focuses on robot-robot teams and treats humans only as exogenous task sources, leaving human-factors, safety, and interaction modelling outside the scope.

Beneath allocation lies the execution substrate that turns an assignment into collision-free motion, and this layer supplies the low-level controller and path planner that the report treats as fixed rather than as a contribution. Motion-planning methods construct reduced or sampled representations of free space. A Generalised Voronoi Graph follows the medial structure of the environment, producing paths with high obstacle clearance at the cost of potentially longer routes and sensitivity to map discretisation, while Probabilistic Roadmaps build reusable graphs by sampling collision-free configurations and connecting nearby samples, making them suitable for repeated queries in high-dimensional static environments [113]. Rapidly-exploring Random Trees instead expand incrementally toward unexplored regions and are useful when differential or

kinodynamic constraints must be considered [114], although sampling-based methods may require many samples in narrow passages and often produce irregular paths that require smoothing or optimisation. Practical motion planning must therefore account not only for connectivity between initial and goal configurations, but also for obstacle geometry, robot footprint, kinematic and dynamic feasibility, sensing and localisation uncertainty, moving obstacles, computational limits, and changes during execution, so completeness and optimality guarantees depend on assumptions about the map, discretisation, robot model, and available information. When several robots operate on the same discrete map, collision avoidance can be formulated as Multi-Agent Path Finding (MAPF), where an instance is defined by a graph $G = (V, E)$, a robot set $\mathbf{N} = 1, \ldots, n$, and start and goal vertices $s_i, g_i \in V$. Each robot follows a time-indexed path $\pi_i = (v_i^0, v_i^1, \ldots, v_i^{T_i})$, with $v_i^0 = s_i$, $v_i^{T_i} = g_i$, and transitions corresponding to graph movements or waiting actions, while feasible joint paths must avoid vertex conflicts, $v_i^t \neq v_j^t, \forall i \neq j$, and edge-swap conflicts, $(v_i^t, v_i^{t+1}) \neq (v_j^{t+1}, v_j^t)$. Common objectives minimise either the makespan, $J_{\mathrm{MS}} = \max_i T_i$, or the sum of path costs, $J_{\mathrm{SOC}} = \sum_{i=1}^{n} T_i$, but optimal MAPF is NP-hard and becomes increasingly difficult as robot density, obstacle density, and conflict coupling grow. Conflict-Based Search decomposes the problem into a high-level search over inter-agent constraints and low-level single-agent searches, providing complete and optimal solutions under standard discrete MAPF assumptions [115]; Priority-Based Search uses priority relations and replans lower-priority agents, which can reduce computation but sacrifices completeness and global optimality [116]; EECBS provides bounded-suboptimal solutions with an explicit cost guarantee [117]; and methods such as MAPF-LNS2 and LaCAM improve scalability by repairing subsets of conflicting paths or incrementally constructing joint configurations [118, 119]. These methods resolve spatial conflicts between robots treated as interchangeable path-followers; they keep the directed, resource- and role-dependent effects that this report targets outside their model, which is why the report places them in the fixed execution layer and studies dependency at the task layer above them.

Optimisation, market, and auction methods are relevant to this thesis because they provide the standard mechanism for converting heterogeneous capabilities into task assignments, routes, schedules, or coalitions. Optimisation-based methods encode robot eligibility, capability requirements, travel costs, energy limits, temporal constraints, and shared resources in a constrained allocation problem, then compute a feasible or cost-minimising solution using mathematical programming, robust optimisation, or repeated replanning [15, 32]. Calvo and Capitán [4] showed how explicit capability constraints prevent low-cost but operationally infeasible assignments, which is essential when tasks require particular robot types or combinations of capabilities. Market and auction methods distribute the same decision by allowing each robot or coalition to evaluate task

Table 2.3. *Representative application domains for coordination in heterogeneous robot teams, the platforms typically mixed, and the coordination task. Warehouse logistics is the evaluation substrate of this report; the contribution of preserving W_t is domain-general.*

Domain	Typical heterogeneous team	Coordination task	Representative work
Exploration & mapping	Aerial + ground, multi-sensor	Area coverage, frontier exploration, collaborative SLAM	[63, 68, 91]
Search & rescue	UAV + UGV	Victim search, risk-aware routing, comms-limited operation	[54, 93]
Persistent monitoring	Mobile robots/sensors, UAVs	Patrolling, informative paths, recharging, remote sensing	[7, 36, 44]
Field robotics & agriculture	Rovers, aerial robots, field sensors	Soil sampling, fleet representation, biodiversity monitoring	[43, 44, 98]
Construction & assembly	Manipulators + mobile bases	Rearrangement, multi-arm assembly, team composition	[99–101]
Mobile & loco-manipulation *(this report)*	Mobile/legged base + arm(s)	Cooperative carrying, quadrupedal pushing, hand-off, one robot assisting another	[107–109]
Warehouse logistics *(this report)*	AGVs + pick-ers/manipulators	Pickup-and-delivery, task-and-charge, routing	[81, 103]

utility, submit bids, and resolve competing claims through consensus or auction rules. Choi et al. [14] combined local bundle construction with consensus over winning bids, while Bai et al. [5] allocated tasks to robot groups that jointly satisfy heterogeneous requirements. These methods determine which robots should cooperate and which resources or tasks should be assigned to them, so they can provide the task and coalition structure within which the dependencies studied in this thesis arise. Their limitation is that inter-agent effects are represented indirectly through predefined costs, feasibility constraints, or bid values. Once execution begins, changes in battery, congestion, communication, or partner availability can alter how one robot affects another without changing the nominal assignment. Re-optimisation can update the allocation, but it does not explicitly estimate whether the current action of robot i benefits, burdens, or has no effect on robot j.

Centralised coordination computes globally optimal task allocations and plans from full joint state, enabling explicit handling of constraints and multi-robot interdependencies [25]. Non-centralised execution forgoes a single online planner by distributing decisions through consensus [26], local monitoring, and learnt communication, including heterogeneous communication protocols [61], dynamic directed structures [120], hypergraph coordination groups [121],

graph-based heterogeneous policies [75], and autonomous partner selection [122]. These methods reduce dependence on complete online state within graph $G_t = (V, E_t)$, but they primarily determine communication, grouping, or policy interaction structures. This thesis focuses on heterogeneous cooperation that arises from joint dependencies among multiple robots, tasks, and resources, rather than from isolated pairwise interactions. Decentralisation limits each robot to local observations, making team-level dependencies only partially visible and sensitive to changes in battery state, available capabilities, communication, and team composition [72, 73, 123]. The resulting challenge is to identify the partner and team information required by local policies and to represent the contribution of higher-order interactions in the learning signal.

Where decentralised and learning-based methods address how robots coordinate under limited information, formal specification and synthesis methods define what the resulting collective behaviour must satisfy. Linear temporal logic expresses ordering, safety, recurrence, and eventual completion over execution traces, while metric and signal temporal logic add deadlines and quantitative robustness margins [124, 125]. Given a task specification and a transition model, temporal-logic planning or controller synthesis can generate behaviour with formal satisfaction guarantees. Wongpiromsarn et al. [126] synthesise policies for heterogeneous multi-agent systems from LTL specifications, Cardona and Vasile [16] integrate temporal requirements with capability and resource constraints, and Liu et al. [127] coordinate multi-robot tasks under STL specifications. Branching-time formalisms such as computation tree logic support verification over alternative system evolutions [128], although LTL and STL are used more commonly as task languages in multi-robot planning. These methods specify required goals, temporal relations, safety conditions, and capability constraints, but they usually assume that the relevant task dependencies are already represented in the model. They therefore do not directly identify how the action of one robot changes the progress, cost, or feasibility of another robot during execution. Online specification revision, including the automated task updates introduced by Fang and Kress-Gazit [129], allows requirements to change during operation, but the update still acts on an externally defined task model rather than learning inter-agent dependencies from experience. Formal synthesis is therefore complementary to the focus of this thesis. It can constrain or verify collective behaviour once task requirements are known, while the thesis investigates how local policies can observe and learn the partner and team dependencies that shape such behaviour.

2.3 Multi-Agent Reinforcement Learning for Cooperation

The partially observable Markov game of Equation (2.1) identifies the problem structure. The solution concept is a joint policy $\pi = (\pi_1, \ldots, \pi_n)$ that maximises

expected discounted team return, but computing this exactly is intractable in general, so reinforcement learning provides the practical foundation [130]. At the single-agent level, a Markov decision process (MDP) (S, A, T, r, γ) defines the state space S, action space A, transition model T, per-step reward r, and discount factor $\gamma \in [0, 1)$. The agent selects actions under a policy π, where $\pi(a \mid s) = \Pr\{a_t = a \mid s_t = s\}$, with the objective of maximising the expected discounted return

$$v_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}. \tag{2.2}$$

The state-value function $V^\pi(s) = \mathbb{E}_\pi[v_t \mid s_t = s]$ and the action-value function $Q^\pi(s, a) = \mathbb{E}_\pi[v_t \mid s_t = s, a_t = a]$ measure expected return under π, and the optimal policy π^* satisfies $V^{\pi^*}(s) \geq V^\pi(s)$ for all s and π. Value-based methods learn Q directly and derive a policy through greedy maximisation; deep Q-networks extend this to high-dimensional inputs with experience replay and a target network [131]. Policy-gradient methods parameterize π_θ and ascend the gradient of $\mathbb{E}[v_t]$; actor-critic architectures combine both by using a value estimate to reduce gradient variance. Three widely used actor-critic variants are DDPG (deterministic policy, off-policy) [132], PPO (clipped surrogate objective, on-policy) [133], and SAC (maximum-entropy regularisation, off-policy) [134].

The multi-agent extension replaces the MDP with a Markov game in which n agents act jointly under partial observability, each receiving an individual reward $r_{i,t}$ and conditioning its policy on a local observation history $\tau_{i,t}$, as formalised in Equation (2.1). The fully cooperative special case, in which all agents share one team reward, is the decentralised partially observable Markov decision process (Dec-POMDP); exact planning in a Dec-POMDP is NEXP-complete for teams of two or more agents [135], making learning from sampled interactions the only tractable approach. Multi-agent reinforcement learning (MARL) applies this strategy to the joint policy but must additionally handle non-stationarity from evolving partner policies, scalability of the joint value, and credit assignment across agents, as surveyed by Busoniu et al. [136], Hernandez-Leal et al. [137], and Wong et al. [138].

Independent learning applies a single-agent algorithm to each robot separately: each agent i maintains a local value function $Q_i(\tau_{i,t}, a_{i,t})$ and ignores partner policy changes. IQL, IDDPG, and IPPO are representative instances that permit fully decentralised execution but treat partner policy evolution as environmental non-stationarity, which can destabilise training. Centralised training with decentralised execution (CTDE) addresses this by providing a critic with access to the joint state s_t and joint actions $\mathbf{a}_t$ during training, while each actor $\pi_i(\cdot \mid \tau_{i,t})$ runs on local information at deployment [86]. MADDPG [9] extends DDPG with per-agent centralised critics that condition on all agents' observations and actions. MAPPO [139] applies PPO with a shared centralised value function across heterogeneous teams. Counterfactual multi-

agent policy gradients (COMA) [10] sharpen per-agent credit by comparing the actual action against a counterfactual baseline that marginalises the focal agent's contribution while holding partners fixed. Recent theory has placed heterogeneous-agent learning on firmer ground: Zhong et al. [74] proves convergence and monotonic-improvement guarantees for heterogeneous trust-region and actor-critic methods, and Liu et al. [140] provides a maximum-entropy formulation for exploration in heterogeneous teams. CTDE supplies the training scaffold but does not specify what each agent should observe or how directed, asymmetric inter-agent effects should enter the learning signal.

2.3.1 Value Factorisation, Parameter Sharing, and Credit Assignment

How the return is shared determines what cooperation the team can learn. A shared reward encourages cooperation but gives weak role-level credit; individual rewards localise attribution but can suppress actions whose benefit reaches another agent before returning to the actor. Pignatelli et al. [141] survey temporal credit assignment and conclude that neither formulation is uniformly superior; suitability depends on task coupling, observability, and reward density. Value-factorisation methods address the credit problem structurally by decomposing the team value into per-agent utilities. VDN uses the additive form $Q_{\text{tot}} = \sum_{i=1}^{n} Q_i$ [17], while QMIX imposes the monotonicity condition $\frac{\partial Q_{\text{tot}}}{\partial Q_i} \geq 0$ through a state-conditioned mixing network, ensuring that locally greedy actions remain consistent with global value maximisation and satisfy the individual-global-max (IGM) condition under CTDE [18]. Parameter sharing reduces sample complexity by having agents share network weights and receive an agent identifier as additional input; heterogeneous teams may instead use role-conditioned or separately parametrised policies [1].

Potential-based reward shaping provides a principled means of injecting auxiliary credit without altering the optimal policy. Andrew Y. Ng et al. [11] proves that augmenting the reward with the potential difference $F(s,s') = \gamma\Phi(s') - \Phi(s)$ preserves the optimal-policy set for any bounded real-valued Φ; Wiewiora et al. [142] extends this invariance to the state-action setting. Applied per agent over an augmented state z_t, this gives the shaped reward

$$r_i'(t) = r_i^{\text{task}}(t) + \lambda \left[\gamma\Phi_i(z_{t+1}) - \Phi_i(z_t)\right], \tag{2.3}$$

where λ is a scalar weight. This invariance requires a Markov-compatible potential and correct terminal-state treatment; the algebraic form alone is not sufficient. The semantic content of Φ_i is a design choice; Ma et al. [143], Ma and Cheng [144], and Zuo et al. [145] explore self-adaptive, assistant-guided, and preference-based variants. This report constructs Φ_i from event-level relationship proxies; relational credit can sharpen attribution but, as

Foerster et al. [10] caution, does not identify causal contribution without valid counterfactual estimates.

A complementary line of work represents the structure of cooperation explicitly through interactions and roles. Li et al. [146] couple interaction and task representations across tasks, Ye and Lu [147] regularise mutual information between agents, and Seraj et al. [61] learn heterogeneous policy networks for composite teams. A parallel strand measures and controls behavioural diversity: Bettini et al. [29] and Bettini et al. [27] quantify and tune heterogeneity, Amir et al. [30] ask when diversity actually helps, and Takai et al. [148] show that role specialisation can outperform homogeneous coordination. These methods confirm that relational structure is learnable and valuable, but they model interaction as undirected affinity or shared latent context; this report instead targets directed, asymmetric effects where the benefit one robot confers on another need not equal the benefit returned.

2.4 Ecological Symbiosis as a Conceptual Lens

Symbiosis provides an ecological vocabulary for describing how persistent interactions affect the participating organisms. Chacón et al. [12] and Wiesmann et al. [149] distinguish mutualism, commensalism, parasitism, competition, and neutralism according to whether each partner experiences a positive, neutral, or negative fitness effect. More importantly, these interaction classes are context dependent rather than fixed properties of a species pair. Chamberlain et al. [150] show that interaction outcomes vary with environmental conditions, while Wein et al. [151], Spottiswoode et al. [152], and Becks et al. [153] report changes in interaction strength or sign as resources, partner behaviour, and population composition change. At the community level, Saavedra et al. [154], Qian and Akçay [155], and Valdovinos [156] further show that persistence and stability depend on the organisation of positive and negative relations across the interaction network. These studies establish signed and context-dependent effects as the most transferable ecological ideas for heterogeneous robot systems, whereas population models such as Lotka–Volterra dynamics describe reproductive density changes that do not directly correspond to fixed robot teams [157, 158].

Engineering research has used symbiosis to describe systems in which heterogeneous components exchange information, resources, services, or capabilities. Nguyen et al. [159] examine mutualistic knowledge exchange between autonomous systems, while Nguyen et al. [23] discuss complementary capabilities and reciprocal adaptation in human-machine and machine-machine systems. Barravecchia et al. [160] extend the concept to engineered ecosystems, where interdependence is linked to adaptability, resilience, and coordinated resource use. Earlier computational studies introduced symbiosis-inspired mechanisms into evolutionary systems and reinforcement learning, using in-

teraction benefits to influence agent selection, adaptation, or policy learning [161, 162]. Related work in industrial ecology models directed exchanges of materials, energy, services, and by-products between production units, often using network or optimisation models to identify beneficial exchange structures [163–166]. Symbiosis-inspired multi-robot studies have also used shared resources, complementary roles, and interaction-based rewards to support cooperation [167, 168]. Across this literature, however, the biological terms are implemented through different operational proxies, including task performance, energy, resource availability, service exchange, and continued operation. No common formulation specifies how the direction, sign, magnitude, and temporal variation of an inter-agent effect should be estimated and incorporated into decentralised decision-making.

3. Research Methodology

This chapter sets out the methodology shared by the four studies and shows how each instantiates it. Rather than describing each study end to end, the chapter is organised by the foundations they have in common. Section 3.1 models the robots and tasks, Section 3.2 states the decision-making and learning formulation, Section 3.2.1 describes the multi-agent learning architectures and the three points at which the studies intervene, and Section 3.3 covers simulation, training, and execution.

3.1 Robot and Task Modelling

A heterogeneous multi-robot system is modelled as a set of platform-specific agents whose sensing, actuation, mobility, and control structures differ substantially. As illustrated in Fig. 3.1, each robot maintains its own state estimation, planning, and low-level control pipeline, providing the basis for decentralised multi-robot control, while task decomposition, allocation, scheduling, conflict resolution, and mission replanning are coordinated through shared mission-level information. This separation allows each robot to respond locally to disturbances and partial observations, reduces dependence on a single central controller, improves scalability, and enables continued operation when communication is delayed or individual robots fail.

3.1.1 Robot Kinematics, Dynamics, and Control to Tasks

Each robot is first treated as an autonomous dynamical system whose configuration space is determined by its embodiment. A general second-order model can be written as $\ddot{\mathbf{q}} = f(\mathbf{q}, \dot{\mathbf{q}}) + g(\mathbf{q}, \dot{\mathbf{q}})\mathbf{u}$, where q denote the configuration and may be held as a Euclidean space, a Lie group, or a product of both. For rigid-body and articulated systems, the same dynamics are commonly expressed as $M_i(\mathbf{q}_i)\ddot{\mathbf{q}}_i + C_i(\mathbf{q}_i, \dot{\mathbf{q}}_i)\dot{\mathbf{q}}_i + N_i(\mathbf{q}_i) = B_i(\mathbf{q}_i)\mathbf{u}_i + J_i(\mathbf{q}_i)^\top \lambda_i$, where M_i is the inertia matrix, C_i contains Coriolis and centrifugal effects, N_i includes gravitational and other configuration-dependent forces, and λ_i represents contact or interaction forces. The warehouse AGVs are modelled on $SE(2)$, with planar pose $\mathbf{q} = (x, y, \phi)$, while whereas aerial robots, quadruped floating bases, and end-effector frames evolve on $SE(3)$. A three-dimensional rigid-body pose is represented by $T_i = \begin{bmatrix} R & \mathbf{p} \\ \mathbf{0}^\top & 1 \end{bmatrix} \in SE(3)$, where $R \in SO(3)$

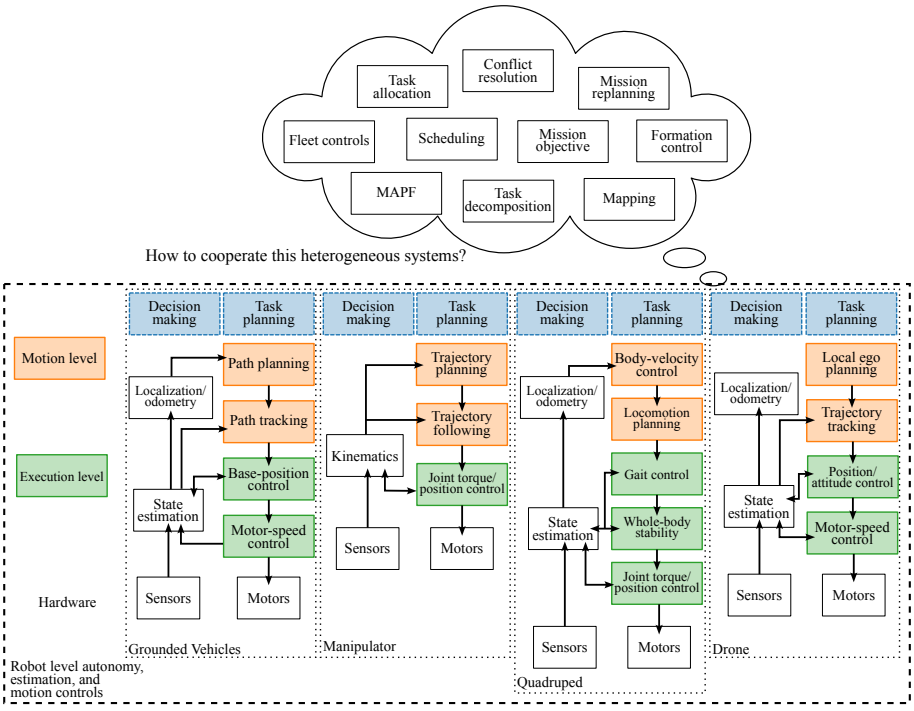


Figure 3.1. Hierarchical control architecture for heterogeneous multi-robot cooperation to date.

describes orientation and $\mathbf{p} \in \mathbb{R}^3$ describes position. Its differential kinematics may be written as $\dot{T} = T\widehat{\xi}$, where $\xi \in \mathbb{R}^6$ is the body twist and $\widehat{\xi} \in \mathfrak{se}(3)$ is its matrix representation. For a manipulator, the joint configuration remains $\mathbf{q} \in \mathbb{R}^n$, while the end-effector pose is obtained through the forward kinematics $T_e = F(\mathbf{q}) \in SE(3), \xi_e = J(\mathbf{q})\dot{\mathbf{q}}$. A mobile manipulator may therefore be described on the product space $\mathbf{Q} = SE(2) \times \mathbb{R}^n$, with the mobile base and manipulator arm governed by different kinematic and dynamic constraints.

The role of the platform-specific controller is to stabilise these dynamics and track admissible references while respecting holonomic and nonholonomic constraints, actuator limits, contact conditions, collision constraints, and platform-dependent safety requirements. Motion generation is typically formulated between an initial state $\mathbf{x}_0$ and a goal set $\mathbf{x}_g$, where the reference may specify positions, orientations, velocities, waypoints, or complete trajectories.

A general trajectory optimisation problem can be written as

$$\begin{aligned}\xi^* = \arg\min_{\xi,\mathbf{u},T} \quad & \int_0^T L(\mathbf{x}(t),\mathbf{u}(t),t)\,dt + L_f(\mathbf{x}(T)) \\ \text{subject to} \quad & \dot{\mathbf{x}} = F(\mathbf{x},\mathbf{u}), \\ & \mathbf{x}(0) = \mathbf{x}_0, \qquad \mathbf{x}(T) \in \mathbf{x}_g, \\ & h(\mathbf{x},\mathbf{u}) = 0, \qquad c(\mathbf{x},\mathbf{u}) \leq 0.\end{aligned}$$

Here, L may penalise travel time, control effort, tracking error, energy consumption, collision risk, or deviation from a desired task-space path. In Euclidean coordinates, a neighbouring candidate trajectory may be expressed as $\mathbf{x}_\alpha(t) = \mathbf{x}(t) + \alpha\eta(t), \eta(0) = \eta(T) = \mathbf{0}$. For configurations on $SE(3)$, an additive perturbation is generally not valid because rotations and rigid transformations do not form a vector space. A corresponding variation is constructed instead through the exponential map, $T_\alpha(t) = T(t)\exp(\alpha\widehat{\eta}(t)), \eta(0) = \eta(T) = \mathbf{0}$, which perturbs the trajectory while preserving its membership in $SE(3)$. The same geometric formulation applies to aerial vehicles, floating-base legged robots, and manipulator end-effector trajectories, while planar mobile robots can be treated in the subgroup $SE(2)$.

Given a geometric or dynamic model, motion planning seeks a continuous sequence of admissible configurations connecting an initial configuration $\mathbf{q}_0$ to a goal configuration or goal region $\mathbf{q}_g$. Let $\mathbf{C}_{\text{free}} = \{\mathbf{q} \in \mathbf{C}_i \mid \mathbf{B}(\mathbf{q}) \cap \mathbf{O} = \varnothing\}$ denote the collision-free configuration space of robot i, where $\mathbf{B}(\mathbf{q})$ is the region occupied by the robot and $\mathbf{O}$ is the obstacle set. The geometric path-planning problem is then to find $\sigma : [0,1] \to \mathbf{C}_{\text{free}}$ such that $\sigma(0) = \mathbf{q}_0, \sigma(1) \in \mathbf{q}_g$. Motion planning additionally requires that the path admit a feasible time parametrisation satisfying the robot dynamics, control bounds, velocity and acceleration limits, contact constraints, and nonholonomic constraints. For low-dimensional mobile navigation, the environment is commonly represented by an occupancy grid or another discrete map from which a graph $G = (V,E)$ is constructed. Search methods such as Dijkstra's algorithm and A* can then find a shortest path over the discretised free space. Their principal advantages are deterministic behaviour, direct integration with occupancy maps, and completeness and optimality with respect to the chosen graph and cost function. Their limitations arise from map resolution, increasing memory and computational cost as the state dimension grows, and the fact that a geometrically valid grid path need not directly satisfy the kinematic or dynamic constraints of the physical robot.

The MAPF methods address the routing layer after robot goals have already been assigned, whereas the central question of this thesis concerns how heterogeneous inter-robot relations should be represented and used before routing decisions are made. In the proposed relation-centred framing, symbiosis characterises how one robot changes another robot's task feasibility, resource state, execution cost, or expected contribution, particularly when tasks require complementary roles or when charging decisions alter future coalition availability.

The cooperative gains, however, are not attributed to the framing itself. They are produced by a separate coordination mechanism that uses this relational information to select tasks, form role-complete coalitions, and balance task execution against resource recovery. Local path planning and inter-robot conflict handling are retained as a common downstream execution layer across methods. This separation prevents gains arising from collision avoidance or route optimisation from being attributed to the proposed symbiosis framing. MAPF determines how already assigned robots can reach their goals without conflict, while the coordination mechanism studied in this thesis determines which robots should cooperate, which task they should perform, and under what resource conditions.

The multi-robot decision layer operates on task-level variables rather than directly on motor torques, joint commands, wheel velocities, or thrust inputs. Let $\mathcal{N} = \{1, \ldots, n\}$ denote the robot team and $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ its communication graph, where $\mathcal{N}_i$ is the neighbourhood of robot i. A first-order abstraction may represent each robot by $\dot{\mathbf{x}}_i = \mathbf{u}_i$, where $\mathbf{x}_i$ denotes a coordination-level quantity such as assigned destination, task progress, arrival time, workload, or resource condition, rather than the full mechanical state [169]. At this level, decentralised coordination is formulated as task allocation rather than agreement over a common continuous state. Following Korsah et al. [25], the problem is characterised by whether robots execute single or multiple tasks, whether tasks require individual robots or coalitions, whether allocation is instantaneous or time-extended, and whether utilities and constraints are independent or interrelated. Each robot evaluates feasible tasks from its capabilities, resources, expected execution cost, and current schedule, while communication is used to resolve competing assignments. For heterogeneous teams, the resulting allocation may include role-complete coalitions, temporal synchronisation, precedence constraints, and dependencies across robot schedules. The assigned task or task sequence is then realised by the robot-specific estimation, planning, and control stack.

The methodological problem is how to construct this coordination mechanism when robots differ in sensing, mobility, manipulation, energy, and task role. Conventional optimisation, market-based, and auction-based methods encode such differences through capabilities, costs, timing requirements, and resource constraints, from which assignments, schedules, coalitions, or winning bids are computed [4, 14, 15, 32]. The studies in this thesis instead examine how the state of one robot changes the feasible or beneficial decisions available to another during execution. Battery level, rendezvous timing, spatial access, object readiness, sensing coverage, and partner availability may all affect the task selected by a local policy. Symbiosis provides the relation-centred framing for these directed effects, while partner-state sharing, interaction-aware rewards, coalition formation, and adaptive coordination provide the mechanisms that generate cooperation. Each high-level action is translated into a local reference by a robot-specific planning and control interface, allowing

heterogeneous platforms to share a common decision layer without sharing the same dynamics or controller. In the warehouse studies, actions include moving, picking, delivering, waiting, and charging, with battery state exposing resource dependence in I and II; in the Isaac studies, actions are continuous targets above joint- or end-effector-level control in III. The ICPS study in IV modifies only the observation interface by adding the partner's battery level, so the resulting coordination effect can be attributed to partner-state observability rather than to a different optimiser, action space, or controller.

3.2 Decision-Making and Learning Formulation

The single-robot decision problem can be formulated as a Markov decision process $\langle S, A, T, r, \gamma \rangle$, where (S) and (A) denote the state and action spaces, $T(s_{t+1} \mid s_t, a_t)$ is the transition model, $r_t = R(s_t, a_t, s_{t+1})$ is the reward. When the full state is not directly available, the model becomes a partially observable MDP by introducing an observation space Ω and observation function $O(o_t \mid s_t, a_{t-1})$. This formulation extends to multiple agents as a decentralised partially observable MDP, in which agents select a joint action $\mathbf{a}_t = (a_t^1, \ldots, a_t^n)$, each agent i receives a local observation o_t^i, and the shared state evolves according to $s_{t+1} \sim T(\cdot \mid s_t, \mathbf{a}_t)$. A Dec-POMDP assumes a shared global reward or an identical optimisation objective, whereas a partially observable Markov game permits agent-specific rewards and can therefore represent cooperative, mixed-motive, or competitive interactions [73, 86, 170]. The studies in this thesis adopt the partially observable Markov game introduced in Section 2.1.2, $\mathscr{G} = \langle N, S, A_i, T, \Omega_i, O, r_i, \gamma \rangle$, where agents may differ in their action spaces, observation spaces, capabilities, and receive local rewards. The reward structure is treated as part of the coordination design, since a shared reward encourages team-level behaviour but provides weak role-specific credit, while individual rewards localise credit but may fail to represent effects realised through other agents. This sequential, partially observable, and coupled decision structure motivates the use of reinforcement learning for policies that adapt to state-dependent interactions during execution. A solution to the POMG is a decentralised policy profile $\pi = (\pi_1, \ldots, \pi_n)$, where each policy maps the agent's local observation history to an action distribution. In the fully cooperative case, the objective is to find a joint policy that maximises the expected team return, whereas distinct agent incentives lead to equilibrium concepts such as a Nash equilibrium. Computing an exact optimal policy profile is generally intractable because each agent must reason over partial observation histories, joint actions, and the policies of the other agents. In particular, finite-horizon Dec-POMDP planning with at least two agents is NEXP-complete [135]. Reinforcement learning is therefore used to approximate decentralised policies from interaction data without explicitly solving the full planning problem, although it does not in general guarantee a globally optimal policy or equilibrium.

3.2.1 Reinforcement Learning and Multi-Agent Reinforcement Learning

Reinforcement learning seeks an optimal policy π^* that maximises the expected discounted return from every state. For two policies π and π', define $\pi \succeq \pi'$ when $V_\pi(s) \geq V_{\pi'}(s)$ for all $s \in S$. In a finite MDP, at least one optimal policy exists such that $\pi^* \succeq \pi$ for every admissible policy π. All optimal policies share the same optimal state-value and action-value functions,

$$V^*(s) = \max_\pi V^\pi(s), \quad \forall s \in S$$
$$Q^*(s,a) = \max_\pi Q^\pi(s,a), \quad \forall s \in S, a \in A$$

After taking action a in state s, the agent follows an optimal policy from the next state onwards. Consequently,

$$Q^*(s,a) = \sum_{s' \in S} T(s' \mid s,a) \left[R(s,a,s') + \gamma V^*(s') \right],$$
$$V^*(s) = \max_{a \in A} Q^*(s,a),$$

which gives the Bellman optimality equations

$$Q^*(s,a) = r_t + \gamma \sum_{s_{t+1} \in S} T(s_{t+1} \mid s_t, a_t) \max_{a_{t+1} \in A} Q^*(s_{t+1}, a_{t+1})$$
$$V^*(s) = \max_{a \in A} \{ r_t + \gamma \sum_{s_{t+1} \in S} T(s_{t+1} \mid s_t, a_t) V^*(s_{t+1}) \}.$$

When the transition and reward models are known, policy iteration and value iteration can apply these Bellman relations directly. When the model is unknown or the state and action spaces are too large for tabular computation, reinforcement learning estimates value functions or policies from sampled interactions. the off-policy Q-learning, for example, updates the action-value estimate with temporal difference (TD) to target $r_t + \gamma \max_{a'} Q(s_{t+1}, a')$,

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right].$$

Instead of using tabular value functions, deep reinforcement learning represents value functions and policies with neural networks. Deep Q-Networks (DQN) approximate the action-value function $Q_\phi(s,a)$ for discrete actions, mapping high-dimensional observations directly to the estimated value of each action. Training minimises the temporal-difference error using a target network $Q_{\bar{\phi}}$, whose parameters are held fixed for several updates to improve stability,

$$\mathscr{L}_Q(\phi) = \mathbb{E}_{(s_t, a_t, r_t, s*t+1) \sim \mathscr{D}} \left[\left(r_t + \gamma (1 - d_t) \max_{a'} Q_{\bar{\phi}}(s_{t+1}, a') - Q_\phi(s_t, a_t) \right)^2 \right].$$

Policies may also be represented directly by a neural network ($\pi_\theta(a \mid s)$) and optimised using policy-gradient methods. When a parametrised policy is combined with a learnt value function, the resulting method is actor-critic, where the actor selects actions and the critic estimates the value of the current policy to guide its updates. Monte Carlo methods estimate values from complete sampled returns, whereas temporal-difference methods bootstrap from current value estimates. On-policy methods learn from trajectories generated by the current policy, while off-policy methods learn a target policy from data collected by a different behaviour policy. One of the off-policy actor algorithms is the deterministic policy gradient, which is obtained by differentiating the critic through the actor instead of analytically maximising it. The critic is trained with loss

$$\mathscr{L}_Q(\psi) = \mathbb{E}_{\mathscr{D}}\left[\left(r_t + \gamma(1-d_t)Q_{\bar{\psi}}\left(s_{t+1}, \mu_{\bar{\theta}}(s_{t+1})\right) - Q_\psi(s_t, a_t)\right)^2\right],$$

while the actor maximises the critic's estimate,

$$J(\theta) = \mathbb{E}_{s\sim\mathscr{D}}\left[Q_\psi(s, \mu_\theta(s))\right].$$

Proximal Policy Optimisation directly updates a stochastic policy while limiting excessively large policy changes. Its clipped surrogate objective is

$$\mathscr{L}^{\text{CLIP}}(\theta) = \mathbb{E}_t\left[\min\left(\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}\hat{A}_t, \text{clip}\left(\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_t\right)\right],$$

By clipping the policy ratio to the interval $[1-\varepsilon,\, 1+\varepsilon]$, the incentive to move it outside this interval is removed; thus, the policy updates are constrained as well. It is possible to use an additional loss term to promote policy entropy and, thus, exploration. In practice, the policy objective is combined with a value-function loss and an entropy term.

Extending reinforcement learning to multiple decentralised agents introduces non-stationarity, scalability, and credit-assignment problems [136–138, 171]. Independent learning applies a single-agent RL algorithm to each robot using only local information. This supports decentralised execution and scales naturally with the team, but each agent treats the changing policies of its partners as part of the environment, making its learning problem non-stationary. Scalability arises as joint observation and action spaces grow rapidly with the number of agents, while credit assignment becomes difficult when a shared return does not reveal which robot, action, or interaction produced an outcome. Parameter sharing and shared experience can reduce data requirements, but they must account for heterogeneous observations, actions, and roles.

Centralised training with decentralised execution addresses some of these difficulties by allowing critics to use joint information during training while keeping each actor dependent only on locally available information at execution time. Multi-agent actor-critic methods learn centralised critics [9, 10],

MAPPO combines PPO actors with a centralised value function [139], and value-decomposition methods factor a joint action value into agent-level utilities [17, 18]. Centralised critics can reduce non-stationarity and improve value estimation, but their input size may grow with the team and may require permutation-invariant representations. Counterfactual critics further support credit assignment by comparing an agent's selected action with alternative actions while holding the actions of the remaining agents fixed.

3.2.2 Information Sharing, Reward Shaping, and Interaction Adaptation

Parameter sharing defines how learning experience is pooled across agents, but it does not answer the main research question. Sharing parameters within a robot type can reduce sample complexity and stabilise training, yet it neither represents directed inter-robot relations nor explains why one robot's state changes another robot's feasible or beneficial actions. It also does not constitute the coordination mechanism that produces cooperative gains, since agents with identical parameters may still act independently when they lack relevant partner information or interaction-specific learning signals. These studies therefore hold type-shared policies fixed where possible and evaluate separate interventions, including partner-state observability, interaction-aware rewards, coalition formation, and adaptive coordination structure. This distinction allows symbiosis to remain a relation-centred framing, while the observed gains are attributed to the mechanisms that expose and use those relations rather than to parameter sharing itself.

The core illustrations are paper-specific rather than forced into a single diagram. ICPS has an explicit algorithmic architecture for battery-state sharing (Fig. 3.2); PRISM has a relationship-aware shaping pipeline (Fig. 3.3); ICRAS is illustrated mainly through the Isaac simulation tasks in Sec. 3.3; and MORPH is represented by its online topology update equations and the recovery/scale plots in Chapter 4. This separation avoids treating simulator screenshots, controller architectures, and reward-shaping pipelines as interchangeable evidence.

Reward shaping is the mechanism through which I and III represent interaction value, but the two studies use it at different levels of formality. ICRAS adds task-specific mutual-benefit terms to MAPPO rewards in each Isaac benchmark; this is an empirical reward-design intervention, not a policy-invariance claim. PRISM instead uses potential-based shaping, which augments the reward with a difference of a potential. Andrew Y. Ng et al. [11] prove that adding $\gamma\Phi(s') - \Phi(s)$ leaves the optimal-policy set unchanged for any bounded potential Φ, and Wiewiora et al. [142] extend this to the state-action setting. Applying this per agent gives the PRISM reward

$$r_i^{\text{PRISM}}(t) = r_i^{\text{task}}(t) + \lambda \left[\gamma\Phi_i(z_{t+1}) - \Phi_i(z_t)\right], \tag{3.1}$$

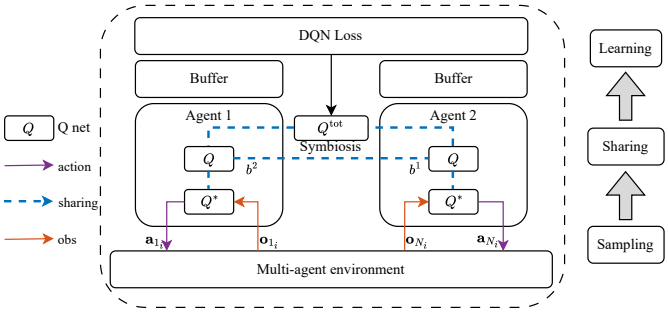


Figure 3.2. Core algorithmic illustration from IV. Battery levels are appended to the symbiotic state used by the value-decomposition learner, while decentralised execution retains local robot decision making after training.

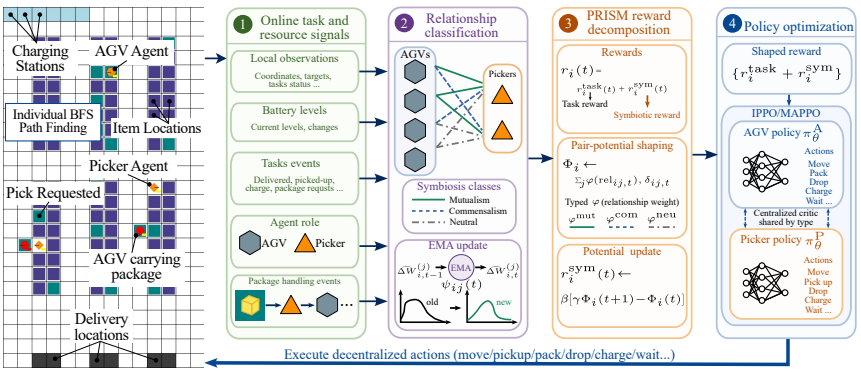


Figure 3.3. Core PRISM methodology from I. Directed relationship effects are approximated from observable interaction events, mapped to typed relation labels and bounded magnitudes, and inserted into a potential-difference shaping term for IPPO/MAPPO training and closed-loop warehouse evaluation.

where z_t is the augmented state and λ a scalar weight. The potential is built from a relational quantity. PRISM begins from the conceptual contribution

$$\Delta W_i^{(j)} = W_i(\pi) - W_i(\pi_{-j}), \tag{3.2}$$

the change in agent i's operational fitness when partner j is removed or replaced, which is the directed effect named in Chapter 2. Because exact evaluation would require counterfactual partner-removal rollouts and a specified replacement policy, I does not compute $\Delta W_i^{(j)}$ during training. It trains offline in simulation while maintaining an online event proxy from delivery, pickup, charging, and battery events observed in each rollout; this proxy is mapped to the bounded potential Φ_i. The policy-invariance guarantee transfers only when the augmented state is Markov-compatible for the potential and the terminal-state convention satisfies the standard shaping assumptions; the algebraic form alone is not

sufficient. Two further limits are stated explicitly. First, the Andrew Y. Ng et al. [11] result is a single-agent optimal-policy invariance, and applying it per agent in a Markov game does not by itself guarantee that the team optimum or the game's equilibria are preserved, so the guarantee is treated as motivation for the shaping form rather than as a convergence claim for the multi-agent setting. Second, an event-level potential risks introducing non-Markov dependence on interaction history, so the potential must be a function of a valid augmented state for even the single-agent guarantee to hold. Neither PRISM nor ICRAS claims a new convergence result for PPO.

Interaction adaptation is the mechanism through which MORPH intervenes, changing not the reward but the structure over which agents coordinate. MORPH maintains an undirected structural adjacency A_t and a pairwise preference matrix W_t, strengthening a link when co-assigned agents make task progress and decaying it otherwise [75]. A compact form of the preference update is

$$W_{ij}(t+1) = \mathrm{clip}\Big(\rho_{ij}(t)\, W_{ij}(t) + \alpha_t\, c_{ij}(t)\, q(t)\, A_{ij}(t) + \alpha_r\, d(t)\, \bar{c}_{ij}(t)\, A_{ij}(t),\, 0,\, 1\Big), \tag{3.3}$$

where $c_{ij}(t)$ is co-assignment overlap, $q(t)$ a task-progress signal, $d(t)$ a delivery event, $\bar{c}_{ij}(t)$ a recent-overlap term, and $\rho_{ij}(t)$ a base-and-threshold retention factor. Structural links are formed from observation correlation, a homeostatic degree term, and, in the evaluated TA-RWARE implementation, a predictive hint that connects a newly assigned AGV to the nearest available Picker by Manhattan distance. MORPH therefore adapts preferences online without offline graph-policy training, though the evaluated version still uses this positional hint; the plasticity terminology describes the update rule and is not a claim of neural equivalence.

3.3 Simulation, Training, and Execution

3.3.1 Simulation Environments and Architectures

The studies are evaluated in simulation because the central methodological requirement is to isolate each intervention under repeatable conditions and across enough seeds to distinguish systematic effects from training variance [172]. The simulation platform is selected according to the mechanism being tested rather than treated as a common benchmark for all studies. ICPS IV uses MATLAB/Simulink to model a two-AMR warehouse with battery depletion, charging stations, collision handling, task allocation, and a MARL decision layer. ICRAS III uses Isaac Sim and Isaac Lab [173] to test mutual-benefit rewards across physically different continuous-control tasks, including cart-pendulum balancing, Shadow Hand object passing [174], and mobile manipulation. PRISM I and MORPH II use TA-RWARE [103], derived from

Table 3.1. *Simulation illustrations, run-time architectures, and training/evaluation protocols used by the four studies.*

Study	Core illustration	Run-time architecture	Training/evaluation protocol
IV	Figs. 3.2, 3.4, and 3.6	Plant dynamics, local path/execution controllers, a global controller for collision avoidance and task allocation, and a VDN/double-DQN decision layer with partner battery-state input	Offline MARL training in the Simulink warehouse; closed-loop evaluation against non-symbiotic and static-recharging references using task time, workload, energy, and resource-use metrics
III	Fig. 3.5	Vectorised Isaac Sim/Isaac Lab tasks with MAPPO under centralised training and decentralised execution	Offline training of benchmark-reward and mutual-benefit-reward variants, followed by evaluation over repeated seeds using success, return, coordination quality, and learning-stability metrics
I	Fig. 3.3	Heterogeneous TA-RWARE with AGV–Picker macro-actions, battery/resource state, event logging, IPPO/MAPPO learners, and PRISM potential-based reward shaping	Offline policy training with online event-proxy updates inside each rollout; evaluation across task-only, flat-cooperative, and PRISM rewards with 18 learned seeds per condition/backbone
II	Eq. (3.3) and Chapter 4 figures	TA-RWARE task assignment, warehouse execution, preference matrix W_t, and structural adjacency A_t updated during execution	Closed-loop evaluation of online topology adaptation under scale and recovery settings; the mechanism adapts coordination structure rather than learning a separate graph policy offline

robotic warehouse benchmarks [175], because its discrete warehouse dynamics are lightweight enough for repeated reward, topology, scale, and perturbation comparisons.

The illustrations are therefore selective, but not arbitrary. ICPS needs both the warehouse scene and the controller architecture because its claim depends on where battery information enters the MARL layer and how robot execution is embedded in the cyber-physical loop. ICRAS needs the simulator screenshot because the contribution is a reward idea tested across three physically different Isaac tasks rather than a new controller architecture. PRISM is best illustrated by the method pipeline in Fig. 3.3, because the important object is the relationship proxy and potential-shaping path, not a rendered warehouse. MORPH is best represented by Eq. (3.3) and the scale/recovery plots in Chapter 4, because the central object is an online adaptive graph rather than a new simulator.

3.3.2 Training, Evaluation, and Execution Protocol

Training is performed offline in simulation for the learning-based studies, while evaluation is performed through closed-loop execution episodes after training.

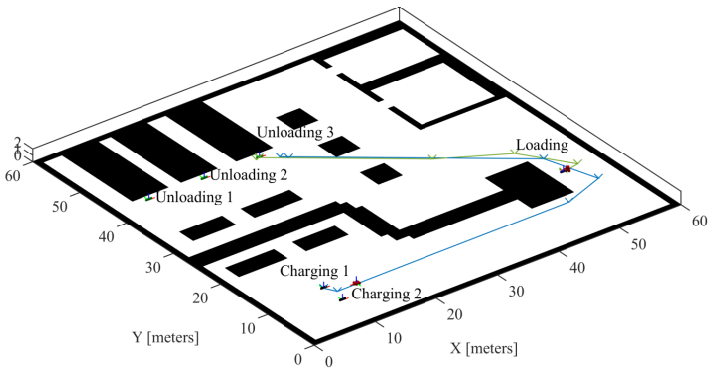


Figure 3.4. ICPS simulated warehouse environment from IV. The scene contains one loading station, three unloading stations, two charging stations, obstacles/restricted zones, and example AMR trajectories, providing the execution substrate for battery-aware task and charging decisions.

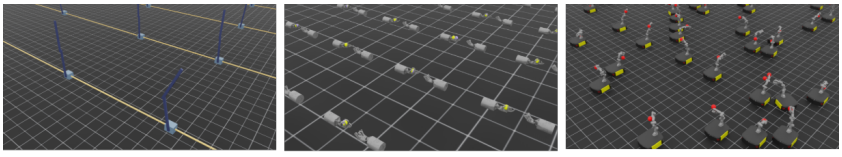


Figure 3.5. Isaac Sim/Isaac Lab tasks used in III. The tasks provide physically different substrates for the same reward-augmentation question, moving from coupled control to dexterous object passing and mobile manipulation; the source experiments use parallel simulated environments for MAPPO training.

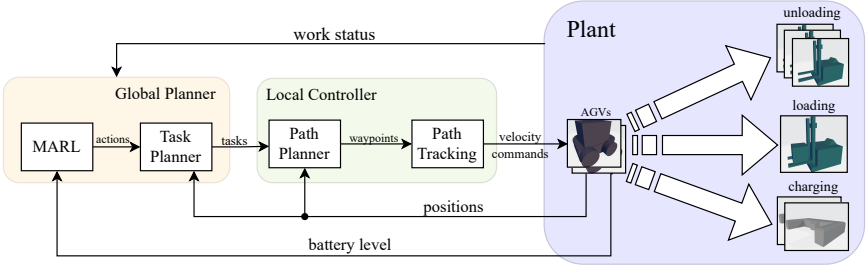


Figure 3.6. ICPS warehouse system architecture. The plant models robot dynamics, energy use, and charging; local controllers manage path following and execution; and the global controller integrates collision handling, task allocation, and the MARL decision layer.

ICPS trains a VDN/double-DQN decision layer in the Simulink warehouse. During training, each robot has local observations and an augmented symbiotic state containing other robots' battery levels; after training, agents execute

decisions through the warehouse control stack and are compared with non-augmented and static charging references. ICRAS trains MAPPO variants in Isaac Lab with the benchmark reward and with the task-specific mutual-benefit reward, then evaluates whether the added reward changes success, return, coordination balance, or learning stability. PRISM trains IPPO and MAPPO policies under matched TA-RWARE settings while changing the reward structure between task-only, flat-cooperative, and potential-shaped variants; the relationship proxy is updated from events during rollouts, but the policy-learning experiment remains an offline simulation study. After training, policies are evaluated on throughput, request type, relationship-label traces, role flexibility, and perturbation cases.

MORPH is the exception to the offline-training pattern. Its evaluated contribution is an online coordination structure, so the preference graph continues to update during execution from co-assignment, delivery, observation, degree, and performance signals. It is therefore compared as an adaptive topology mechanism rather than as a separately trained graph policy. This distinction prevents three layers from being conflated: offline policy optimisation, online inference or structure adaptation during simulated execution, and physical deployment.

Deployment is outside the evidence boundary of the completed studies. The policies and coordination rules execute closed loop in simulation, but they are not validated on physical robots. Moving from this evaluation protocol to hardware would require a separate deployment layer: calibrated sensing and localisation, communication-delay handling, safety overrides, robot-specific path tracking, battery and charging models fitted to hardware, and fail-safe behaviour when the learnt or adaptive decision layer requests an infeasible action. The present methodology therefore supports claims about simulated training and execution under controlled assumptions, while physical deployment remains future validation rather than an outcome demonstrated here.

4. Results to Date

This chapter reports the completed evidence by intervention rather than as a cumulative validation of symbiosis. The results are read against the three research questions introduced earlier: compact partner or interaction information for decentralised execution (RQ1), directed reward and credit representation (RQ2), and the validity boundary of the resulting interaction representations (RQ3). Each section therefore states the manipulated design object, the matched comparison, the observed effect, and the limit of the claim.

The studies are not directly pooled into a single aggregate score because they test different mechanism layers. ICPS changes an observation channel in a two-AMR warehouse, ICRAS changes reward terms in Isaac Sim tasks, PRISM changes reward shaping in TA-RWARE, and MORPH changes an online interaction graph. The chapter therefore treats the results as a sequence of bounded mechanism tests: first whether a selected partner state is useful, then whether reward terms can expose interaction value, then whether typed relationship proxies improve cross-role throughput, and finally whether an adaptive graph can remain competitive under scale and task-shift conditions.

4.1 Partner Battery-State Sharing in a Two-AMR Warehouse (RQ1)

The ICPS study tests whether one compact partner-resource variable changes decentralised warehouse behaviour. The intervention augments the learner's observation with the other AMR's battery level while keeping the warehouse plant, local controllers, and task setting fixed [167]. It is compared with non-augmented MARL and static recharging in the same two-AMR warehouse. This is the narrowest RQ1 test in the report: the study does not expose a full joint state or a learned communication graph, but asks whether a single resource variable is enough to affect task and charging decisions.

The training evidence suggests that the battery-state variable changes coordination before the final evaluation metrics are computed. Figure 4.1 shows that the augmented learner reaches higher cumulative reward earlier and exhibits less imbalance between the two AMRs than the non-augmented variant. The important point is not only that reward increases, but that the policy has access to the partner's future resource constraint while choosing whether to keep delivering or recharge. This makes charging conflicts and asymmetric depletion visible during learning, whereas the non-augmented learner must infer the

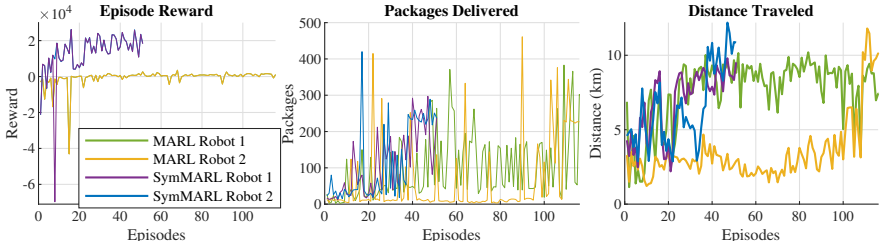


Figure 4.1. ICPS training results. Battery-state sharing makes future charging conflicts visible during learning and is associated with earlier reward improvement and more balanced activity between the two AMRs.

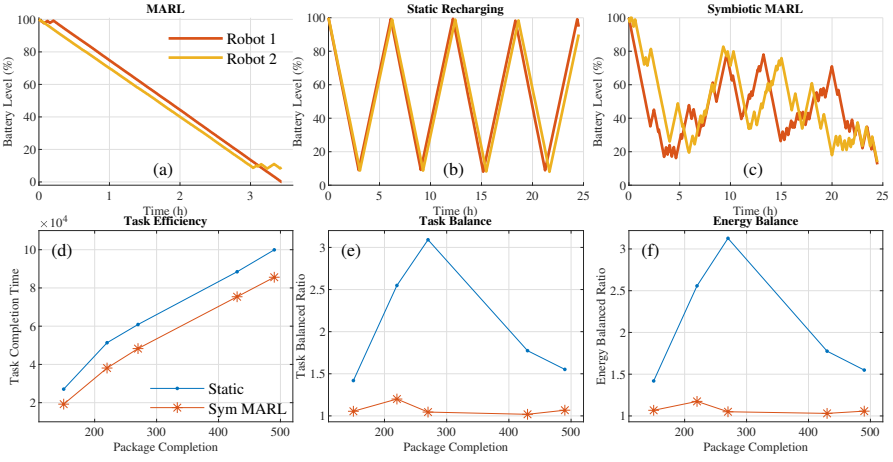


Figure 4.2. ICPS evaluation results. In the tested two-AMR warehouse, partner battery-state augmentation is associated with shorter completion time and more balanced workload and energy use.

partner's state only through delayed consequences in the shared warehouse dynamics.

The evaluation evidence in Fig. 4.2 supports the same interpretation at the operational level. Relative to static recharging, the reported evaluation shows a 10.7% reduction in task completion time and a 13.81% improvement in workload/energy balance. The non-augmented evaluation includes trajectories in which one AMR charges at a high battery level while the other depletes after approximately 140 deliveries. These observations indicate that the augmented state helps the two robots avoid unbalanced resource use, not merely that it increases the learned reward.

The result supports RQ1 only for the selected variable and environment. Partner battery state can be informative when charging and task allocation are coupled, but the study does not show that arbitrary partner information improves coordination, that full state sharing is desirable, or that the result scales beyond two AMRs. The main evidence boundary is therefore the state-sufficiency

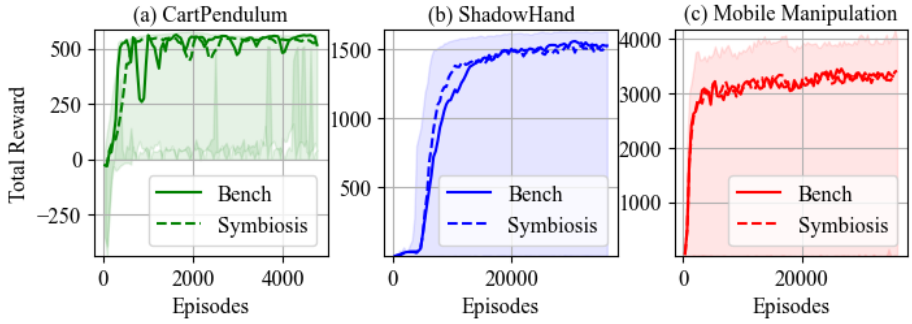


Figure 4.3. ICRAS training curves over five seeds. The reward additions have little peak effect in the simplest task and appear mainly as stability or variance changes in the more coupled settings.

question: battery level is useful in this setting, but later work must test which additional partner variables, if any, are necessary for larger heterogeneous teams.

4.2 Task-Specific Reward Augmentation in Isaac Sim (RQ2)

The ICRAS study tests the simplest reward-side version of the relation-centred idea: add task-specific mutual-benefit terms while keeping the underlying MAPPO training and simulator tasks fixed [168]. Unlike PRISM, this study does not use counterfactual fitness, event-proxy relationship labels, or potential-based invariance. Its value in the thesis is empirical: it asks whether mutual-benefit reward terms change learning behaviour across physically different simulated tasks.

The result is task dependent. In CartPendulum, both variants reach approximately 580 peak reward, with success rates of 0.993 for the benchmark and 0.992 for the augmented reward. The main visible difference is reduced fluctuation rather than improved final performance. This matters because it prevents an overgeneral claim: if the benchmark reward already captures the main coupled-control objective, adding a mutual-benefit term may smooth training without changing the achievable task outcome.

In Shadow Hand object passing, both variants reach a peak near 1631, while the augmented reward has smoother curves and lower reported variability. In mobile manipulation, the augmented variant reaches approximately 4168 compared with 4046 for the benchmark and is reported as more stable. Read together, Figs. 4.3 and 4.4 suggest that reward-side interaction terms become more useful when task success depends on higher-dimensional handover, contact, or coupled locomotion-manipulation behaviour.

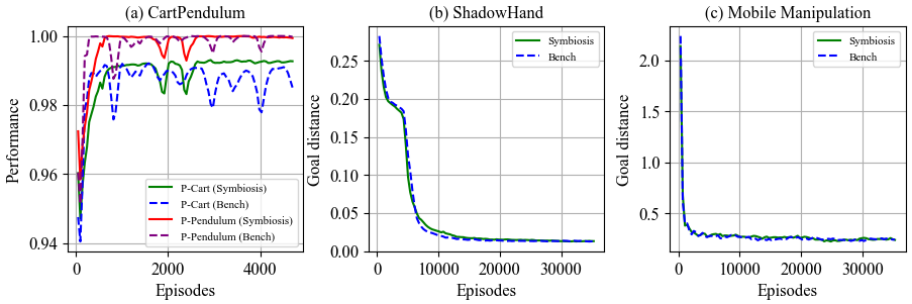


Figure 4.4. ICRAS performance summary over five seeds. The task-specific reward additions show little peak-performance change on CartPendulum and more visible stability differences in the higher-dimensional tasks.

Table 4.1. *Deliveries per episode in the PRISM study. Learned methods use 18 replicates per backbone; the heuristic is a non-learning reference.*

Condition	IPPO	MAPPO	Mean across backbones
PRISM	$\mathbf{44.2 \pm 3.8}$	$\mathbf{36.2 \pm 5.0}$	40.2
Flat-cooperative	27.0 ± 4.3	27.3 ± 3.5	27.1
Task-only	19.7 ± 6.8	21.1 ± 4.6	20.4
Heuristic	27.5 ± 15.8	27.5 ± 15.8	27.5

These results support a narrow RQ2 claim: task-specific interaction terms can change learning behaviour in some coupled-control settings. They do not demonstrate that a mutualistic term improves an arbitrary MARL task, and the five-seed design limits strong statistical interpretation. The current evidence also tests only positive mutual-benefit terms. It does not validate commensalism, parasitism, or harmful interaction labels as operational robot categories.

4.3 PRISM Reward Comparisons in TA-RWARE (RQ2)

PRISM evaluates a stronger RQ2 intervention by converting event-level relationship proxies into potential-based reward shaping. It is compared against task-only and flat-cooperative rewards under IPPO and MAPPO with the same TA-RWARE environment, macro-action controller, and training settings. The principal result is a throughput difference under matched reward-structure comparisons, but the study also reports learning curves, request decomposition, operational relationship labels, role diagnostics, and perturbation outcomes.

Across IPPO and MAPPO, PRISM reaches 40.2 deliveries per episode, 48.3% above the flat-cooperative reward and 97.1% above the task-only reward. Welch comparisons against both learned reward baselines are below $p < 10^{-3}$ under both backbones. The heuristic comparison is narrower: PRISM is

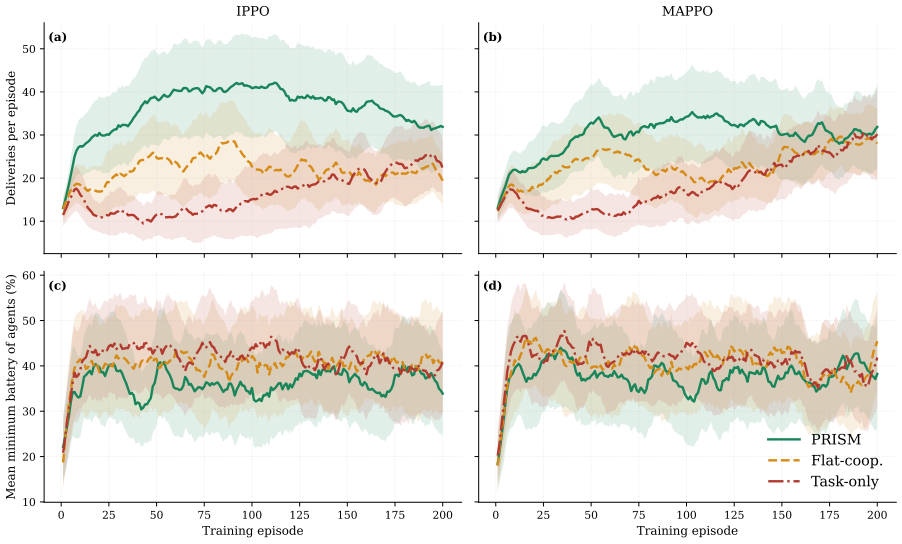


Figure 4.5. PRISM learning curves. The separation from task-only and flat-cooperative rewards occurs after agents learn viable battery management, indicating that the main difference is how operational viability is converted into coordinated deliveries.

significantly higher under IPPO and numerically higher under MAPPO. The flat-cooperative reward remains near the heuristic reference, which suggests that a generic team bonus can recover part of the cooperative structure but does not provide the same cross-role credit as the typed shaping term.

The learning curves in Fig. 4.5 show that the PRISM gain is not primarily about learning to keep batteries alive. Battery-related behaviour stabilises early across reward designs, while the delivery gap appears when agents must convert viable operation into coordinated task completion. This supports the interpretation that the reward intervention affects credit for cross-role task events rather than simply preventing energy collapse.

The request decomposition in Fig. 4.6 helps localise the difference. Simpler transport-only, Picker-only, and two-role requests can often be completed by local heuristics. The largest PRISM gain occurs on heavy requests requiring one AGV and two Pickers: PRISM completes roughly ten per episode, compared with four to six for the learned baselines and almost none for the heuristic. This is consistent with improved credit for sparse cross-role events, but does not by itself prove that the event proxy is the causal mechanism. Reward scale, coefficient selection, and interaction with the macro-action controller remain alternative explanations. A further distinction must stay explicit: the AGV–Picker complementarity of heavy requests is imposed by the TA-RWARE feasibility structure, not discovered by PRISM. The reward intervention plausibly improves how the team learns to exploit that built-in dependency, but task-imposed

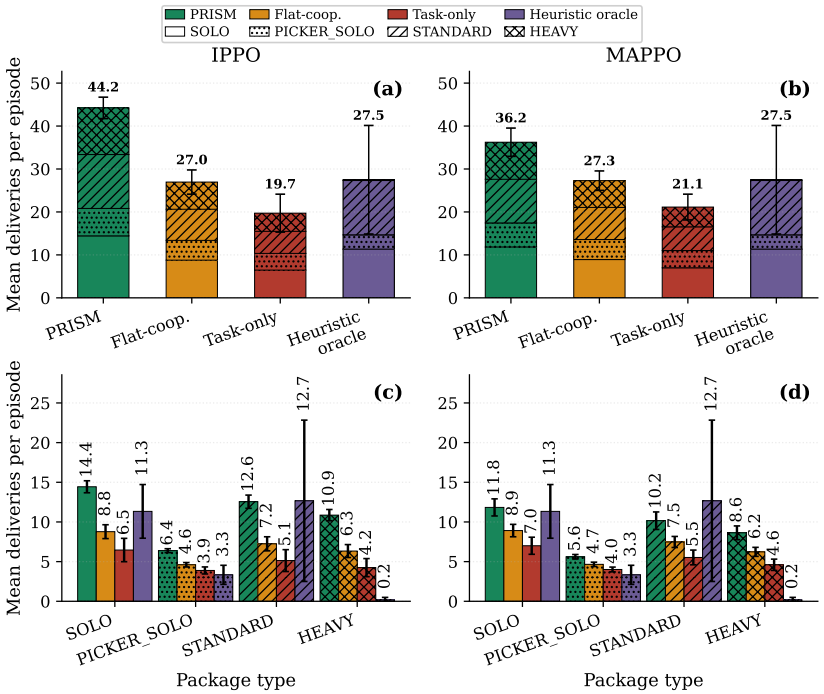


Figure 4.6. PRISM throughput by reward structure and request type. The largest separation occurs for heavy requests requiring one AGV and two Pickers.

complementarity is not the same as a learned relational representation, and the present result speaks only to the former.

4.3.1 Operational Relationship Labels

PRISM's relationship analysis reports labels produced by its event proxy. Neutral labels dominate near 99%, while the mutualistic category remains sparse. Under IPPO, the PRISM mutualism proxy peaks at 1.23% and averages 1.00% over episodes 50–150, compared with 0.62% for the flat-cooperative reward and 0.37% for task-only reward. Under MAPPO, the corresponding averages are 0.89%, 0.72%, and 0.38%.

The alignment between sparse labels and heavy-request throughput supports the proxy as a diagnostic for where cross-role interaction may matter. The mutualism percentages are small because most agent pairs are inactive, redundant, charging, or only weakly coupled at any one step; the meaningful events are concentrated around task handoff and heavy-request completion. This is consistent with the thesis view that relation-level signals can be high leverage even when rare.

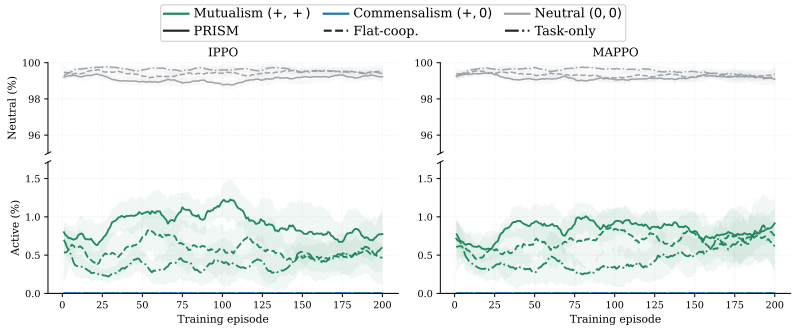


Figure 4.7. Frequency of operational mutualism labels in PRISM. The labels are generated from task-event proxies and should not be interpreted as ecological or causal classifications.

The proxy still does not validate the labels as ground-truth ecological relationships. Intervention rollouts, partner swaps, or exact counterfactual evaluations would be needed to determine whether the labelled pairs caused the performance difference. The current evidence therefore supports an operational reward-shaping mechanism, not a causal taxonomy of robot symbiosis.

4.3.2 Perturbation Results

The perturbation evidence depends on the learning backbone. Under IPPO, PRISM has delivery drops of 20.37%, 14.00%, and 23.03% in the one-AGV, sudden-AGV, and cascade AGV–Picker scenarios, compared with 63.75%, 31.11%, and 48.19% for the flat-cooperative reward. Under MAPPO, PRISM drops 60.30%, 14.89%, and 33.53%, while the flat-cooperative condition drops 37.38%, 10.18%, and 20.55%. The current study therefore does not support a backbone-independent robustness claim. It does, however, identify robustness as a follow-up question: the same typed signal that improves cross-role throughput under nominal conditions may or may not help when a role becomes scarce, and the answer depends on the learner and perturbation type.

Table 4.2. *Mean cumulative deliveries in the MORPH scale study (20 seeds, 2,000 steps).*

Condition	Tiny ($N = 8$)	Small ($N = 12$)	Medium ($N = 18$)	Large ($N = 24$)
MORPH	115.10	153.20	201.45	244.10
Proximity	113.50	165.25	207.45	260.60
Task-shared graph	116.25	157.10	198.15	246.00
Full graph	123.10	154.85	201.40	246.20

4.4 MORPH Online Preference Adaptation (RQ1/RQ3)

MORPH tests whether a compact pairwise structure can be updated during execution rather than fixed before deployment. This contributes to RQ1 as an interaction-structure design and to RQ3 as a validity test under scale and cyclic task-shift conditions. The current MORPH scale artifacts contain 20 seeds per condition and 2,000 steps per run. Table 4.2 reports final cumulative deliveries for the four principal topology conditions.

MORPH is not the highest-throughput condition at every scale. Its final deliveries are not significantly different from TSG or the full graph at any tested scale. Against proximity, differences are not significant for tiny, small, or medium teams; proximity is significantly higher for the large team (260.60 versus 244.10, Welch $p = 0.020$). This bounds the contribution: the current evidence supports online construction of a competitive task-oriented graph with fewer active links at medium and large scales, not throughput superiority over proximity. This is an important distinction because MORPH is meant to test adaptive interaction structure, not to claim that plasticity-inspired topology always maximises deliveries.

MORPH uses mean active-link counts of approximately 27.0, 38.3, 40.7, and 58.5 across the four scales. The full graphs contain 28, 66, 153, and 276 links. The comparison therefore shows a reduction in active links at medium and large scales while retaining throughput similar to the full-graph and TSG baselines. The result is most meaningful at medium and large team sizes, where the gap between a full graph and a sparse adaptive graph becomes substantial. Communication volume was not measured directly, so active-link count is a structural proxy rather than a network-cost measurement.

4.4.1 Cyclic Task-Shift Experiment

The cyclic experiment alternates demand across six 400-step spatial phases. MORPH retains its graph state, while a reset variant clears the preference and adjacency matrices at every phase boundary. Across the return phases, MORPH's mean recovery times range from approximately 30.7 to 52.2 steps. The corresponding ranges are 45.1–96.0 for proximity, 39.8–54.7 for TSG, and 51.7–66.1 for reset MORPH.

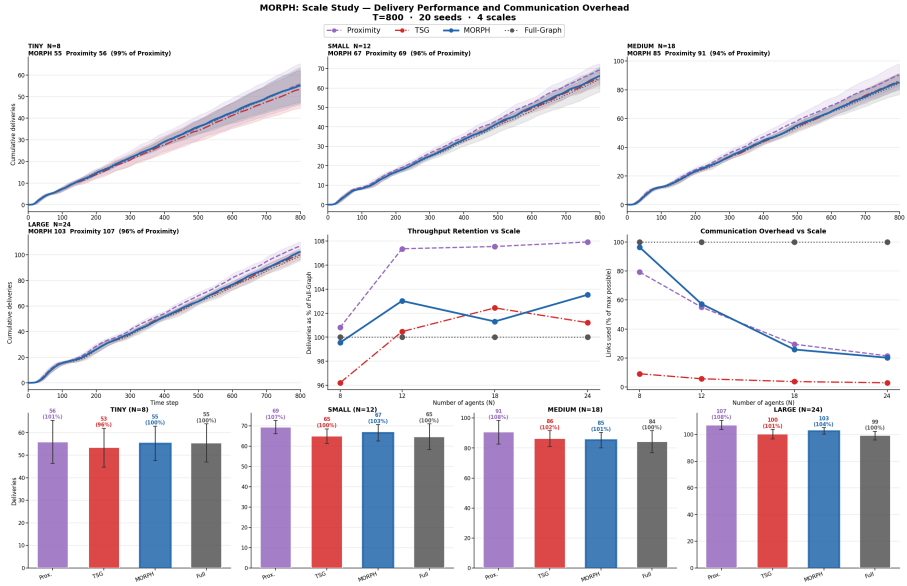


Figure 4.8. MORPH scale and communication results from 20 seeds and 2,000-step runs. Throughput must be read together with active-link count; the full graph contains all possible links, while MORPH maintains a smaller adaptive subset.

The result is consistent with useful persistence in the tested schedule, but it is not monotonic and does not establish general distribution-shift resilience. The experiment also combines preference memory with a nearest-Picker predictive hint, so adaptation cannot be attributed solely to history in W_t. MORPH should therefore be interpreted as an online execution mechanism within the TA-RWARE abstraction, not as evidence that the same graph rule transfers unchanged to physical robots or unrelated domains.

4.5 Cross-Study Summary

The strongest completed evidence is for mechanism-level effects under controlled simulation, not for a complete theory of robotic symbiosis. Read as a representation hierarchy, the four studies move the same directed partner dependence through three layers of the decision process. At the observation layer, ICPS shows that a selected partner resource variable can improve two-AMR resource coordination. At the reward layer, ICRAS shows that task-specific mutual-benefit rewards can affect learning stability, especially in more coupled tasks, but not uniformly, and PRISM shows the clearest throughput gain by inserting typed event proxies into potential-based shaping under matched reward baselines. At the topology layer, MORPH shows that an online graph

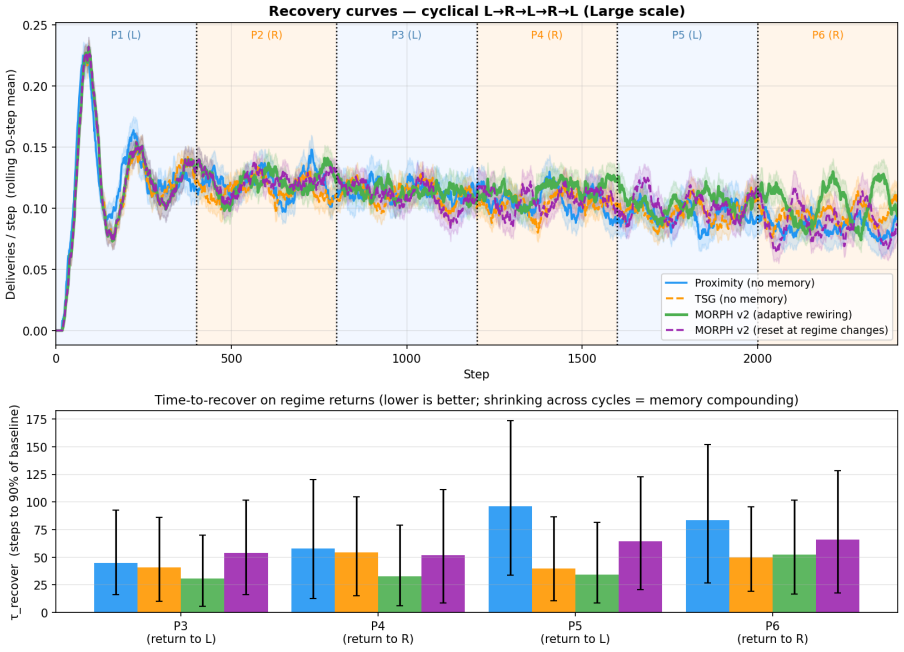


Figure 4.9. Recovery analysis for the six-phase MORPH task-shift experiment (20 seeds; 400 steps per phase). Retained preferences often reduce recovery time relative to reset MORPH and proximity, but the advantage is not uniform across phases or baselines.

can remain competitive with task-shared and full-graph structures while using fewer active links, but not that it dominates proximity in all settings.

Across the four studies, the pattern is consistent with the research framing: inter-agent information is most useful when it exposes a concrete dependency that the local policy would otherwise miss. Battery state matters when charging affects partner availability; reward shaping matters when heavy requests require non-substitutable AGV–Picker cooperation; graph adaptation matters when useful collaboration partners change with task phase and scale. The current evidence is therefore positive but bounded. It motivates stronger ablations and transfer tests rather than closing the question.

Table 4.3. *Claim-evidence assessment for the completed work.*

Claim	Evidence	Status
Partner battery state can be a useful compact observation in the tested two-AMR warehouse.	10.7% shorter completion time and 13.81% better workload/energy balance relative to static recharging.	Supported in setting
Task-specific mutualistic rewards uniformly improve MARL.	Similar CartPendulum outcome, smoother Shadow Hand learning, and a modest mobile-manipulation difference over five seeds.	Not supported
PRISM improves throughput over the tested learned reward baselines.	40.2 deliveries; 48.3% above flat-cooperative and 97.1% above task-only reward under matched TA-RWARE settings.	Supported in setting
PRISM's event proxy identifies causal directed contribution.	Proxy labels align with heavy-request performance, but no intervention or exact counterfactual test is reported.	Not demonstrated
MORPH constructs a competitive online preference graph in TA-RWARE.	Similar throughput to TSG/full graph with fewer active links at medium and large scales.	Supported in setting
The interaction representations are validated beyond controlled simulation.	Evidence remains simulator-specific; sensing, communication, safety, hardware energy dynamics, and physical deployment are untested.	Not supported

4.6 Claim-Evidence Map

5. Discussion

The completed work supports a cautious interpretation of symbiosis as a relation-centred framing, not as the mechanism that directly produces cooperation. Across the four studies, the measurable design objects are partner-state observability, reward shaping, event-level relationship proxies, and online topology adaptation. These are conventional engineering mechanisms placed at the interaction layer between individual robot actions and aggregate team return. The evidence therefore supports the claim that making selected inter-agent effects explicit can change simulated coordination behaviour, while leaving causal interaction dynamics, solver independence, and physical deployment as open questions.

5.1 Interpretation by Research Question

RQ1 asks which compact partner states or interaction structures make resource- and role-dependent cooperation visible to decentralised learners. The ICPS study gives bounded support for partner battery-state observability: in a two-AMR warehouse, exposing the other robot's battery state is associated with shorter completion time and better workload/energy balance. MORPH gives a complementary result for interaction structure: an online preference graph can remain competitive with full and task-shared topologies while using fewer active links at medium and large scales. Together, these results suggest that compact partner information can be useful, but they do not define a sufficient state representation. Battery state, co-assignment history, and delivery events are useful only under the tested task couplings, and the cost of observation or communication is still represented indirectly.

RQ2 asks how directed inter-agent effects can enter the learning signal without reducing cooperation to a shared team return. The ICRAS study shows that task-specific mutual-benefit reward terms have task-dependent effects: little peak-performance difference in CartPendulum, clearer stability differences in Shadow Hand, and a modest mobile-manipulation improvement. PRISM provides stronger evidence within TA-RWARE, where potential-based relationship-aware shaping improves throughput over task-only and flat-cooperative rewards under matched learned baselines. However, the evidence supports the reward mechanism rather than the ecological label itself. The event proxy may correlate with useful cross-role activity, but without partner-removal or intervention tests it does not identify causal directed contribution.

RQ3 asks when the proposed interaction representations remain valid beyond controlled simulation. The current answer is mostly negative or incomplete. The results are internally useful because each study compares matched variants inside a fixed simulator, but cross-study validity is limited by different substrates, seed counts, controllers, and mechanism layers. ICPS uses two AMRs in a Simulink warehouse, ICRAS uses five-seed Isaac Lab experiments, PRISM uses PPO-family learners in TA-RWARE, and MORPH evaluates an online graph heuristic rather than an offline-trained graph policy. These differences make the studies informative as bounded mechanism tests, not as evidence that one symbiosis-inspired representation transfers across embodiments, solvers, team sizes, sensing conditions, or physical robots.

5.2 Alternative Explanations

Reward scale and coefficient tuning remain important alternatives to the relational-credit explanation for PRISM. The task-only, flat-cooperative, and PRISM conditions share the same environment, controller, and training protocol, which strengthens the comparison, but PRISM may still provide a denser or better-scaled auxiliary signal. Ablations that match reward variance, expected auxiliary-return contribution, and clipping frequency would clarify whether relationship semantics add information beyond shaped reward density.

The event proxy creates a construct-validity problem. Co-participation in delivery, pickup, charging, or battery events can coincide with successful coordination without measuring the marginal contribution of a particular partner. The sparse operational mutualism labels align with heavy-request throughput, but co-occurrence can still be mistaken for assistance. Partner-removal rollouts, partner swaps, event interventions, or counterfactual traces are needed before stronger causal claims are justified.

MORPH contains several interacting mechanisms: co-assignment updates, delivery-modulated reinforcement, threshold-dependent decay, target-degree regulation, global performance modulation, observation correlation, and predictive link formation. The positional predictive hint connects newly assigned AGVs to nearby available Pickers, so the evaluated method should not be described as learning topology without spatial prior. The current evidence supports a competitive online preference heuristic, not an isolated proof that preference memory alone causes recovery.

The warehouse execution layer is another shared source of structure. PRISM and MORPH both operate above macro-actions, task eligibility, collision handling, and local routing logic that already encode part of the coordination problem. Their results concern reward and topology within that substrate, not end-to-end robot control or integrated task-and-path planning. A centralised coalition scheduler, mixed-integer optimiser, model-predictive planner, or learned communication baseline could change the comparative picture.

5.3 Validity and Generalisation

The simulation evidence is useful for internal validity but insufficient for deployment readiness. Isaac Sim and Isaac Lab provide higher-fidelity physical dynamics for the ICRAS tasks, while TA-RWARE and the ICPS warehouse provide controllable logistics abstractions. None of the completed studies fully represents communication delay, localisation error, battery ageing, hardware faults, human activity, safety overrides, or actuator-level failures. The results therefore concern algorithmic behaviour under modelled assumptions rather than real-robot autonomy.

The statistical strength of the evidence differs across studies. ICPS reports a small two-AMR scenario, ICRAS uses five seeds and mostly supports qualitative stability statements, PRISM uses 18 replicates per learned backbone and gives stronger throughput comparisons, and MORPH uses 20 seeds for the current scale and recovery artifacts. These differences should remain visible in the final thesis. A cross-study synthesis is appropriate for identifying recurring design dimensions, but not for ranking the four methods as if they shared one benchmark.

The symbiosis taxonomy remains broader than the empirical evidence. Mutualistic terminology appears in ICRAS and PRISM, but commensalism and parasitism are not systematically validated. More importantly, even mutualism is currently operational: it is inferred from task-specific reward terms or event proxies, not from independent ground truth about each robot's marginal outcome. Future work needs controlled asymmetric costs, resource contention, harmful interference, and partner-removal tests before the signed categories can be treated as validated robot-interaction classes.

The negative results in this report must also be read at the right strength. Where a manipulated term does not change an outcome, as with the mutualistic term on CartPendulum peak performance, the defensible statement is that no effect was detected under the tested seeds, not that the term is unnecessary. Distinguishing no-detected-effect, statistical equivalence, and evidence of harm requires equivalence testing and adequately powered designs, neither of which the current seed budgets supply; the second half therefore treats an equivalence claim as a separate experiment rather than a corollary of a non-significant difference. The same discipline applies to the relational metrics. Proxy mutualism frequency, typed-relationship counts, and active-link fraction are reported as diagnostics that localise where interaction may matter, and they are never used on their own to support a cooperation claim; the primary outcomes remain deliveries, completion time, energy, workload balance, idle fraction, and directly measured communication cost.

5.4 Implications for the Remaining PhD Work

The next PRISM work should separate relational semantics from generic reward shaping. This requires reward-scale-matched controls, component ablations, and intervention-based validation of the event proxy. Testing the same relational features with an alternative planner or MARL family would examine portability without presuming solver independence.

The next MORPH work should isolate each information source. In particular, experiments should compare the full method with variants that remove the positional predictive hint, remove preference memory, and match active-link budgets against learned-communication, proximity, full-graph, and task-shared graph baselines. Communication bytes, latency, and decision cost should be measured directly instead of inferred from active-link count.

The next RQ1 work should map the sufficiency frontier for partner information. Battery state is one useful variable in one warehouse task, but larger teams may require role availability, queue state, charging-station occupancy, planned arrival time, or task-eligibility information. The key experiment is not to expose more state by default, but to identify the smallest partner-state or graph representation that changes behaviour without approximating centralised execution.

Cross-platform validation should proceed in stages: additional simulated tasks, degradation tests for sensing and communication, solver-family comparisons, and then limited physical experiments with explicit safety constraints. Sim-to-real performance, fault detection, and adversarial or harmful-agent isolation remain future research questions. The present results do not justify presenting them as consequences of a negative directed-benefit signal.

5.5 Progress and Planned Work

The remaining PhD work is held together by one thesis claim, that compact representations of directed partner dependence improve coordination under state-dependent non-interchangeability, and it is sequenced so that mechanism validation precedes conceptual expansion. Fig. 5.1 gives the plan over the full doctoral period, from January 2024 to the planned defence in December 2028, at quarter resolution. The mandatory first package closes the alternative explanations identified in this chapter: a reward-information factorial that holds reward magnitude fixed while varying only the pair semantics, simple online-scoring and degree-matched-random baselines for MORPH under a fixed link budget, and direct communication-cost measurement rather than active-link count. Only then do the four study lines proceed. Two are core to the argument: joint task-and-charge allocation under shared resource state, which closes the thread opened by ICPS and adds the allocation and auction baselines the completed work lacks; and one focused physical validation of role adaptation, the largest package because it supplies the deployment evidence RQ3 currently

lacks. Two are breadth that extends the same directed-dependence question to new partner conditions rather than to new literatures: human–robot teaming, where a partner's state is observed rather than controlled, and mixed aerial–ground teams, which test whether the same partner representations transfer across embodiments. The breadth lines are the first to be narrowed or deferred if time or hardware constrains the second half. The timeline is intentionally coarse because the exact ordering will depend on review outcomes for PRISM and MORPH and on hardware availability for the physical experiments.

The plan also clarifies what should not be treated as finished. The completed ICPS and ICRAS studies provide evidence for selected observation and reward interventions; PRISM and MORPH require revision work and stronger ablations; and the mechanism-validation package must test causal directed-benefit measurement, solver-independent performance, and communication efficiency before the four study lines build on those mechanisms. The second half therefore front-loads the experiments that remove alternative explanations, and reserves the largest share of time for the physical-robot and cross-platform work that the current evidence base lacks entirely.

Finally, future reporting should retain the claim-evidence discipline used in Chapter 4. Each study should state the manipulated design object, matched baseline, seed count, primary outcome, alternative explanation, and external-validity boundary. This structure keeps the biological framing useful as a vocabulary while preventing it from carrying claims that belong to the actual observation, reward, graph, controller, and evaluation mechanisms.

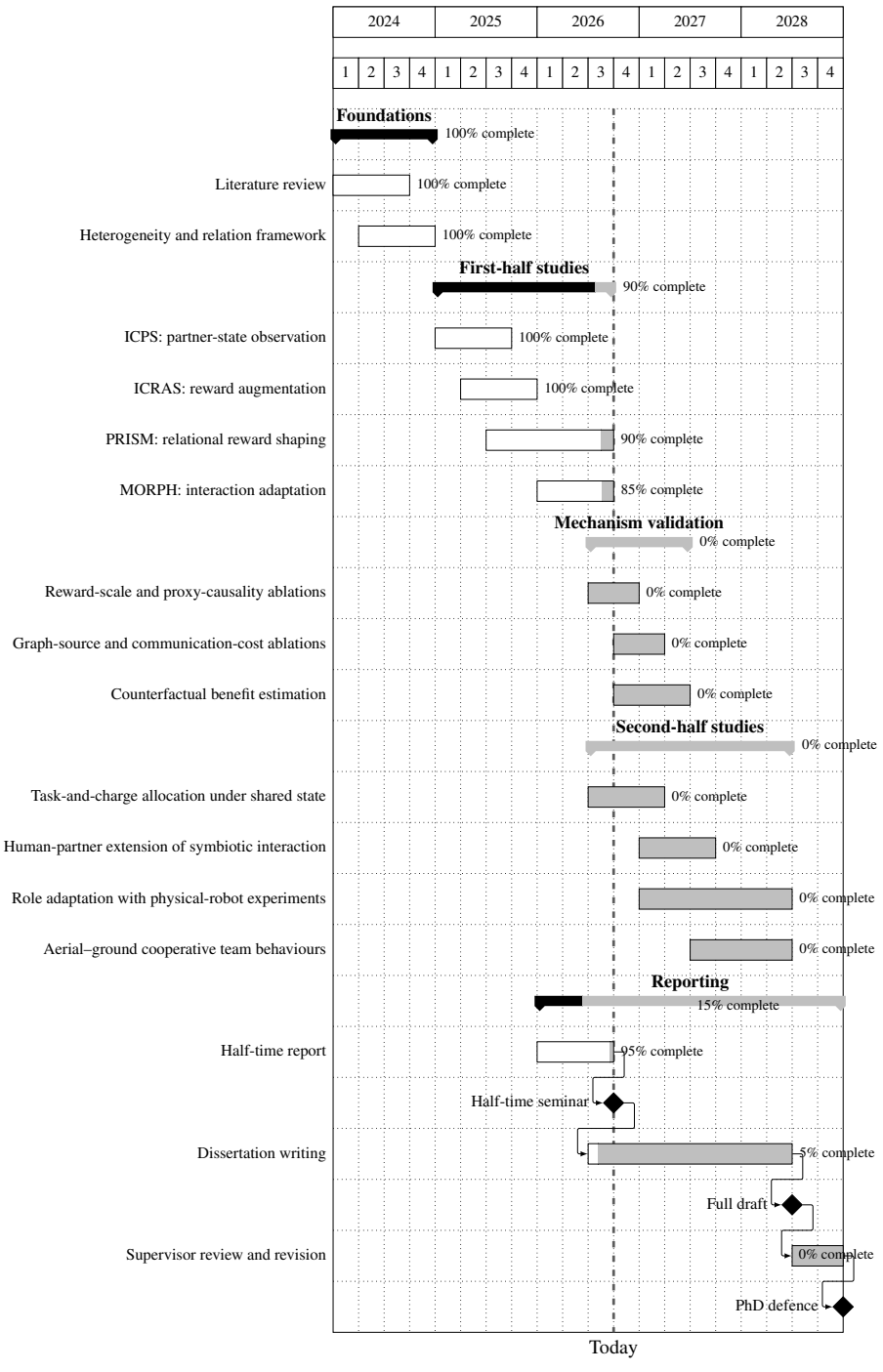


Figure 5.1. Gantt chart of the doctoral project from January 2024 to the planned defence in December 2028, at quarter resolution. Bar shading indicates progress; the dashed rule marks the half-time point.

6. Conclusion

6.1 Conclusions

This half-time report has reframed symbiosis as a limited conceptual lens for one scientific object, the directed effect of one heterogeneous robot on another, studied at three successive representation layers: partner state in the observation, partner effect in the reward, and partner selection in the topology. The four studies are approximations of this single directed-dependence object rather than four unrelated interventions, and the contribution lies in the implemented and evaluated mechanisms at each layer, not in a claim that ecological symbiosis is a fundamental theory of robot cooperation. Stated without the ecological vocabulary, the thesis studies when a robot should observe selected information about another robot, receive pair-specific auxiliary credit, or adapt which partners it coordinates with, rather than relying only on local state and a single team score.

Four study-level findings are supported within their respective settings. First, sharing a partner's battery state improves completion time and workload/energy balance in the tested two-AMR warehouse. Second, task-specific mutualistic reward additions have little effect on the simplest ICRAS task and more visible stability effects in the higher-dimensional tasks. Third, PRISM improves TA-RWARE throughput over task-only and flat-cooperative rewards, with the largest difference on heavy cross-role requests. Fourth, MORPH constructs an online pairwise preference graph that is competitive with TSG and full-graph baselines while using fewer links at larger scales, although proximity performs significantly better at the largest tested scale.

These results do not establish biological fidelity, exact causal credit, universal reward improvement, solver independence, general distribution-shift resilience, or physical-deployment readiness. PRISM's relationship labels are operational proxy outputs. MORPH is plasticity inspired and uses a positional predictive hint in its current implementation. The report therefore treats both methods as engineering designs whose usefulness must be tested against standard alternatives.

6.2 Unresolved Questions

The principal unresolved question for partner information is scalability: which compact partner variables remain useful as team size, communication constraints, and charging competition increase? The present battery result is insufficient to answer this beyond two AMRs.

For relational reward, the main unresolved question is attribution. Future work must determine whether PRISM's gain arises from relationship semantics, reward density, coefficient scale, or interaction with the warehouse controller. Intervention rollouts and reward-matched ablations are required before interpreting the proxy as causal contribution.

For adaptive topology, the main unresolved question is what information MORPH truly needs. Removing or replacing the positional predictive hint, comparing against trained communication methods, and measuring actual communication cost will clarify the contribution of the plastic pairwise preference graph.

6.3 Future Work

The remaining PhD work is governed by a single thesis claim, that compact representations of directed partner dependence improve coordination under state-dependent non-interchangeability, and every planned study is admitted only insofar as it tests that claim. This is a deliberate guard against fragmentation: the four lines below extend the *same* directed-dependence question to new partner conditions rather than opening four separate literatures. A mechanism-validation package is mandatory and gates all of them, because the four lines build on mechanisms whose alternative explanations must first be closed: a reward-information factorial that separates relation semantics from reward density (meaningful, shuffled, random-pair, and no-pair terms at matched magnitude), simple online-scoring and degree-matched-random baselines for MORPH under a fixed link budget, direct communication-cost measurement, and counterfactual benefit estimation that fits the directed-effect estimator rather than a hand-designed proxy.

Two of the four lines are core to the thesis argument and two are breadth that strengthens external validity without being required for completion:

1. **(Core)** joint task-and-charge allocation for warehouse fleets, carrying the ICPS state-sharing thread into a full resource-coupled allocation problem under shared battery state, and compared against established allocation and auction baselines on a common feasible set;
2. **(Core)** one focused physical validation of role adaptation in a heterogeneous robot team, the largest package in time and effort, supplying the deployment evidence that RQ3 currently lacks;
3. **(Breadth)** extension of the directed-dependence formulation from robot–robot to human–robot teaming, where a partner's state is observed rather than controlled; and
4. **(Breadth)** cooperative behaviours for mixed aerial–ground teams, testing whether the same partner representations transfer across embodiments.

The two breadth lines test transfer of the same claim to a human partner and to a cross-embodiment partner respectively; if time or hardware constrains the

second half, they are the first candidates to be narrowed or deferred, whereas the validation package and the two core lines are not.

6.4 Scope of Supported Claims

The report holds its own conclusions to five checks:

- **Clarity and flow:** each study is organised around one manipulated design object and one bounded question;
- **Terminology:** symbiosis and plasticity are identified as inspirations or operational labels, not biological explanations;
- **Unsupported claims:** universal solver, robustness, security, causal-credit, and deployment claims have been removed or marked as untested;
- **Evidence alignment:** quantitative conclusions are restricted to the reported environments, baselines, seeds, and metrics; and
- **Missing evidence:** causal interventions, reward-scale controls, communication-cost measurements, cross-task validation, and physical experiments are stated as future requirements.

The resulting thesis direction is deliberately narrower but more testable: determine when explicit partner information, relational reward features, and adaptive pairwise preferences improve heterogeneous coordination, and identify the conditions under which they do not.

References

[1] Matteo Bettini, Ajay Shankar, and Amanda Prorok. Heterogeneous multi-robot reinforcement learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pages 1485–1494, Richland, SC, May 2023. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-9432-1.

[2] Gennaro Notomista, Siddharth Mayya, Yousef Emam, Christopher Kroninger, Addison Bohannon, Seth Hutchinson, and Magnus Egerstedt. A Resilient and Energy-Aware Task Allocation Framework for Heterogeneous Multirobot Systems. *IEEE Transactions on Robotics*, 38(1):159–179, February 2022. ISSN 1941-0468. doi: 10.1109/TRO.2021.3102379.

[3] Alexander A. Nguyen, Luis Guerrero-Bonilla, Faryar Jabbari, and Magnus Egerstedt. Scalable, pairwise collaborations in heterogeneous multi-robot teams. *IEEE Control Systems Letters*, 8:604–609, 2024. ISSN 2475-1456. doi: 10.1109/LCSYS.2024.3400702.

[4] Álvaro Calvo and Jesús Capitán. Heterogeneous multirobot task allocation for long-endurance missions in dynamic scenarios. *IEEE Transactions on Robotics*, 41:6494–6513, 2025. ISSN 1941-0468. doi: 10.1109/TRO.2025.3626651.

[5] Xiaoshan Bai, Andres Fielbaum, Maximilian Kronmuller, Luzia Knoedler, and Javier Alonso-Mora. Group-based distributed auction algorithms for multi-robot task assignment. *IEEE Transactions on Automation Science and Engineering*, 20(2):1292–1303, April 2023. doi: 10.1109/TASE.2022.3175040.

[6] Jacopo Banfi, Andrew Messing, Christopher Kroninger, Ethan Stump, Seth Hutchinson, and Nicholas Roy. Hierarchical planning for heterogeneous multi-robot routing problems via learned subteam performance. *IEEE Robotics and Automation Letters*, 7(2):4464–4471, 2022. doi: 10.1109/lra.2022.3148489. IEEE article id: 10.1109/lra.2022.3148489.

[7] Ahmad Bilal Asghar, Shreyas Sundaram, and Stephen L. Smith. Multirobot persistent monitoring: Minimizing latency and number of robots with recharging constraints. *IEEE Transactions on Robotics*, 41:236–252, 2025. ISSN 1941-0468. doi: 10.1109/TRO.2024.3502497.

[8] Weiheng Dai, Utkarsh Rai, Jimmy Chiun, Yuhong Cao, and Guillaume Sartoretti. Heterogeneous multi-robot task allocation and scheduling via reinforcement learning. *IEEE Robotics and Automation Letters*, 10(3): 2654–2661, March 2025. ISSN 2377-3766. doi: 10.1109/LRA.2025.3534682.

[9] Ryan Lowe, YI WU, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30 of *NIPS'17*, pages 6382–6393, Long Beach, Califonia, USA, December 2017. Curran Associates, Inc. ISBN 978-1-5108-6096-4.

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

[11] Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pages 278–287, Bled, Slovenia, June 1999. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-612-8.

[12] Jeremy M. Chacón, Sarah P. Hammarlund, Jonathan N. V. Martinson, Leno B. Smith, and William R. Harcombe. The ecology and evolution of model microbial mutualisms. *Annual Review of Ecology, Evolution, and Systematics*, 52(Volume 52, 2021):363–384, November 2021. ISSN 1543-592X, 1545-2069. doi: 10.1146/annurev-ecolsys-012121-091753.

[13] Brian P. Gerkey and Maja J. Matarić. A formal analysis and taxonomy of task allocation in multi-robot systems. *The International Journal of Robotics Research*, 23(9):939–954, September 2004. ISSN 0278-3649. doi: 10.1177/0278364904045564.

[14] Han-Lim Choi, Luc Brunet, and Jonathan P. How. Consensus-based decentralized auctions for robust task allocation. *IEEE Transactions on Robotics*, 25(4):912–926, August 2009. ISSN 1941-0468. doi: 10.1109/TRO.2009.2022423.

[15] Bo Fu, William Smith, Denise M. Rizzo, Matthew Castanier, Maani Ghaffari, and Kira Barton. Robust task scheduling for heterogeneous robot teams under capability uncertainty. *IEEE Transactions on Robotics*, 39(2):1087–1105, April 2023. ISSN 1941-0468. doi: 10.1109/TRO.2022.3216068.

[16] Gustavo A. Cardona and Cristian-Ioan Vasile. Planning for heterogeneous teams of robots with temporal logic, capability, and resource constraints. *The International Journal of Robotics Research*, 43(13):2089–2111, November 2024. ISSN 0278-3649. doi: 10.1177/02783649241247285.

[17] 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. Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pages 2085–2087, Richland, SC, July 2018. International Foundation for Autonomous Agents and Multiagent Systems.

[18] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pages 4295–4304, 2018.

[19] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro A. Ortega, D. J. Strouse, Joel Z. Leibo, and Nando de Freitas. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *PMLR*, pages 3040–3049, 2019.

[20] Yali Du, Lei Han, Meng Fang, Ji Liu, Tianhong Dai, and Dacheng Tao. LIIR: Learning Individual Intrinsic Reward in Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 4405–4416, 2019.

[21] Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. Shapley Q-value: A Local Reward Approach to Solve Global Reward Games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7285–7292, 2020. doi: 10.1609/aaai.v34i05.6220.

[22] Tonghan Wang, Heng Dong, Victor Lesser, and Chongjie Zhang. ROMA: Multi-Agent Reinforcement Learning with Emergent Roles. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pages 9876–9886, 2020.

[23] Alexander A. Nguyen, Mauriel Rodriguez Curras, Magnus Egerstedt, and Jonathan N. Pauli. Mutualisms as a framework for multi-robot collaboration. *Frontiers in Robotics and AI*, 12:1566452, March 2025. ISSN 2296-9144. doi: 10.3389/frobt.2025.1566452.

[24] Amanda Prorok and Matteo Bettini. Heterogeneous teams. In *Encyclopedia of Robotics*, pages 1–8. Springer, Berlin, Heidelberg, 2024. ISBN 978-3-642-41610-1. doi: 10.1007/978-3-642-41610-1_230-1.

[25] G. Ayorkor Korsah, Anthony Stentz, and M. Bernardine Dias. A comprehensive taxonomy for multi-robot task allocation. *The International Journal of Robotics Research*, 32(12):1495–1512, October 2013. ISSN 0278-3649. doi: 10.1177/0278364913496484.

[26] Reza Olfati-Saber, J. Alex Fax, and Richard M. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1): 215–233, January 2007. ISSN 1558-2256. doi: 10.1109/JPROC.2006.887293.

[27] Matteo Bettini, Ajay Shankar, and Amanda Prorok. System neural diversity: Measuring behavioral heterogeneity in multi-agent learning, September 2024.

[28] Amanda Prorok, M. Ani Hsieh, and Vijay Kumar. The impact of diversity on optimal control policies for heterogeneous robot swarms. *IEEE Transactions on Robotics*, 33(2):346–358, April 2017. ISSN 1941-0468. doi: 10.1109/TRO.2016.2631593.

[29] Matteo Bettini, Ryan Kortvelesy, and Amanda Prorok. Controlling behavioral diversity in multi-agent reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 3611–3636, Vienna, Austria, July 2024. JMLR.org.

[30] Michael Amir, Matteo Bettini, and Amanda Prorok. When is diversity rewarded in cooperative multi-agent learning? In *International Conference on Learning Representations (ICLR)*, 2026.

[31] Ragesh K. Ramachandran, James A. Preiss, and Gaurav S. Sukhatme. Resilience by reconfiguration: Exploiting heterogeneity in robot teams. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6518–6525, November 2019. doi: 10.1109/IROS40897.2019.8968611.

[32] Gennaro Notomista, Siddharth Mayya, Seth Hutchinson, and Magnus Egerstedt. An optimal task allocation strategy for heterogeneous multi-robot systems. In *2019 18th European Control Conference (ECC)*, pages 2071–2076, June 2019. doi: 10.23919/ECC.2019.8795895.

[33] Yun Chang, Kamak Ebadi, Christopher E. Denniston, Muhammad Fadhil Ginting, Antoni Rosinol, Andrzej Reinke, Matteo Palieri, Jingnan Shi, Arghya Chatterjee, Benjamin Morrell, Ali-akbar Agha-mohammadi, and Luca Carlone. LAMP 2.0: A robust multi-robot SLAM system for operation in challenging large-scale underground environments. *IEEE Robotics and Automation Letters*, 7(4):9175–9182, October 2022. ISSN 2377-3766. doi: 10.1109/LRA.2022.3191204.

[34] Junting Chen, Checheng Yu, Xunzhe Zhou, Tianqi Xu, Yao Mu, Mengkang Hu, Wenqi Shao, Yikai Wang, Guohao Li, and Lin Shao. EMOS: Embodiment-aware heterogeneous multi-robot operating system with LLM agents. In *The Thirteenth International Conference on Learning Representations*, pages 56670–56690, October 2024.

[35] Hakan Aktas, Yukie Nagai, Minoru Asada, Erhan Oztop, and Emre Ugur. Correspondence learning between morphologically different robots via task demonstrations. *IEEE Robotics and Automation Letters*, 9(5):4463–4470, May 2024. ISSN 2377-3766. doi: 10.1109/LRA.2024.3382534.

[36] Xiaoyi Cai, Brent Schlotfeldt, Kasra Khosoussi, Nikolay Atanasov, George J. Pappas, and Jonathan P. How. Energy-aware, collision-free information gathering for heterogeneous robot teams. *IEEE Transactions on Robotics*, 39(4): 2585–2602, August 2023. ISSN 1941-0468. doi: 10.1109/TRO.2023.3257512.

[37] Kaleb Ben Naveed, An Dang, Rahul Kumar, and Dimitra Panagou. meSch: Multi-agent energy-aware scheduling for task persistence. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 21458–21465, October 2025. doi: 10.1109/IROS60139.2025.11247605.

[38] Neil Mathew, Stephen L. Smith, and Steven L. Waslander. Multirobot rendezvous planning for recharging in persistent tasks. *IEEE Transactions on Robotics*, 31(1):128–142, February 2015. ISSN 1941-0468. doi: 10.1109/TRO.2014.2380593.

[39] Bingxi Li, Brian R. Page, Barzin Moridian, and Nina Mahmoudian. Collaborative mission planning for long-term operation considering energy limitations. *IEEE Robotics and Automation Letters*, 5(3):4751–4758, July 2020. ISSN 2377-3766. doi: 10.1109/LRA.2020.3003881.

[40] M. De Ryck, M. Versteyhe, and K. Shariatmadar. Resource management in decentralized industrial Automated Guided Vehicle systems. *Journal of Manufacturing Systems*, 54:204–214, January 2020. ISSN 0278-6125. doi: 10.1016/j.jmsy.2019.11.003.

[41] Sarmad Riazi, Kristofer Bengtsson, and Bengt Lennartson. Energy Optimization of Large-Scale AGV Systems. *IEEE Transactions on Automation Science and Engineering*, 18(2):638–649, April 2021. ISSN 1558-3783. doi: 10.1109/TASE.2019.2963285.

[42] Nathan Goulet and Beshah Ayalew. Dynamic charging rendezvous and motion planning for a multi-AGV team including a mobile charging host. *IEEE Transactions on Robotics*, 41:5344–5359, 2025. ISSN 1941-0468. doi: 10.1109/TRO.2025.3603501.

[43] Thien Hoang Nguyen, Erik Muller, Michael R. Rubin, Xiaofei Wang, Fiorella Sibona, Alex McBratney, and Salah Sukkarieh. A semi-autonomous robotic system for In Situ soil sampling, analysis, and mapping in precision agriculture.

IEEE Transactions on Field Robotics, 3:22–39, 2026. doi: 10.1109/TFR.2025.3641874.

[44] Igor A. D. Santos, Gilvanio B. de Sousa, Kenny A. Q. Caldas, Robson R. D. Pereira, Leandro J. E. Campos, Paulo H. C. Morais, Edson H. Francelino, Roberto S. Inoue, and Marco H. Terra. Integrating robotics, remote sensing, and machine learning for scalable biodiversity monitoring in tropical rainforests. *IEEE Transactions on Field Robotics*, 3:125–140, 2026. doi: 10.1109/TFR.2025.3648380.

[45] Mary B. Alatise and Gerhard P. Hancke. A Review on Challenges of Autonomous Mobile Robot and Sensor Fusion Methods. *IEEE Access*, 8: 39830–39846, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.2975643.

[46] Ainhoa Apraiz, Ganix Lasa, Maitane Mazmela, Nestor Arana-Arexolaleiba, Íñigo Elguea, Oscar Escallada, Nagore Osa, and Amaia Etxabe. The user experience in industrial human-robot interaction: A comparative analysis of unimodal and multimodal interfaces for disassembly tasks. *Robotics and Computer-Integrated Manufacturing*, 95:103045, October 2025. ISSN 0736-5845. doi: 10.1016/j.rcim.2025.103045.

[47] Jennifer Gielis, Ajay Shankar, and Amanda Prorok. A Critical Review of Communications in Multi-robot Systems. *Current Robotics Reports*, 3(4): 213–225, December 2022. ISSN 2662-4087. doi: 10.1007/s43154-022-00090-9.

[48] Massimo Franceschetti, Mohammad Javad Khojasteh, and Moe Z. Win. The Many Facets of Information in Networked Estimation and Control. *Annual Review of Control, Robotics, and Autonomous Systems*, 6(Volume 6, 2023): 233–259, May 2023. ISSN 2573-5144. doi: 10.1146/annurev-control-042820-010811.

[49] Peter Corke, Ron Peterson, and Daniela Rus. Networked Robots: Flying Robot Navigation Using a Sensor Net. In Paolo Dario and Raja Chatila, editors, *Robotics Research. The Eleventh International Symposium*, Springer Tracts in Advanced Robotics, pages 234–243, Berlin, Heidelberg, 2005. Springer. ISBN 978-3-540-31508-7. doi: 10.1007/11008941_25.

[50] Jiechuan Jiang and Zongqing Lu. Learning Attentional Communication for Multi-Agent Cooperation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[51] Brent Schlotfeldt, Dinesh Thakur, Nikolay Atanasov, Vijay Kumar, and George J. Pappas. Anytime planning for decentralized multirobot active information gathering. *IEEE Robotics and Automation Letters*, 3(2):1025–1032, April 2018. ISSN 2377-3766. doi: 10.1109/LRA.2018.2794608.

[52] Haggi Do, Junwoo Jang, and Jinwhan Kim. Heterogeneous multi-robot system mission planning with cooperative replenishment through data-driven rendezvous point selection. *Intelligent Service Robotics*, 18(1):61–73, January 2025. ISSN 1861-2784. doi: 10.1007/s11370-024-00568-9.

[53] Rui Ding, Yuhan Zhu, Xianbin Feng, Chuanshan Zhang, and Haibin Zhu. Continuous charging assignment algorithm for heterogeneous robot clusters based on E-CARGO. *Expert Systems with Applications*, 259:125175, January 2025. ISSN 0957-4174. doi: 10.1016/j.eswa.2024.125175.

[54] Arnau Romero, Carmen Delgado, Lanfranco Zanzi, Raúl Suárez, and Xavier Costa-Pérez. Cellular-enabled collaborative robots planning and operations for search-and-rescue scenarios. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5942–5948, May 2024. doi: 10.1109/ICRA57147.2024.10611179.

[55] Michael E. Cao, Xinpei Ni, Jonas Warnke, Yunhai Han, Samuel Coogan, and Ye Zhao. Leveraging heterogeneous capabilities in multi-agent systems for environmental conflict resolution. In *2022 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 94–101, November 2022. doi: 10.1109/SSRR56537.2022.10018728.

[56] Ziyi Zhou, Dong Jae Lee, Yuki Yoshinaga, Stephen Balakirsky, Dejun Guo, and Ye Zhao. Reactive task allocation and planning for quadrupedal and wheeled robot teaming. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pages 2110–2117, August 2022. doi: 10.1109/CASE49997.2022.9926658.

[57] Sanjay O V Sarma, Ramviyas Parasuraman, and Ramana Pidaparti. Impact of heterogeneity in multi-robot systems on collective behaviors studied using a search and rescue problem. In *2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 290–297, Abu Dhabi, UAE, November 2020. IEEE Press. doi: 10.1109/SSRR50563.2020.9292588.

[58] Siddharth Mayya, Diego S. D'antonio, David Saldaña, and Vijay Kumar. Resilient Task Allocation in Heterogeneous Multi-Robot Systems. *IEEE Robotics and Automation Letters*, 6(2):1327–1334, April 2021. ISSN 2377-3766. doi: 10.1109/LRA.2021.3057559.

[59] Davide De Benedittis, Manolo Garabini, and Lucia Pallottino. Managing conflicting tasks in heterogeneous multi-robot systems through hierarchical optimization. *IEEE Robotics and Automation Letters*, 10(6):5305–5312, June 2025. ISSN 2377-3766. doi: 10.1109/LRA.2025.3559843.

[60] Adriano Fagiolini, Gianluca Dini, Federico Massa, Lucia Pallottino, and Antonio Bicchi. Distributed misbehavior monitors for socially organized autonomous systems. *The International Journal of Robotics Research*, 43(14): 2145–2182, December 2024. ISSN 0278-3649. doi: 10.1177/02783649241242812.

[61] Esmaeil Seraj, Rohan Paleja, Luis Pimentel, Kin Man Lee, Zheyuan Wang, Daniel Martin, Matthew Sklar, John Zhang, Zahi Kakish, and Matthew Gombolay. Heterogeneous Policy Networks for Composite Robot Team Communication and Coordination. *IEEE Transactions on Robotics*, 40: 3833–3849, 2024. ISSN 1941-0468. doi: 10.1109/TRO.2024.3431829.

[62] Jorge Peña Queralta. *Collaborative Autonomy in Heterogeneous Multi-Robot Systems*. PhD thesis, fi=Turun yliopistolen=University of Turkul, December 2022.

[63] Lukas Bernreiter, Shehryar Khattak, Lionel Ott, Roland Siegwart, Marco Hutter, and Cesar Cadena. A framework for collaborative multi-robot mapping using spectral graph wavelets. *The International Journal of Robotics Research*, 43 (13):2070–2088, November 2024. ISSN 0278-3649. doi: 10.1177/02783649241246847.

[64] Didem Gürdür and Fredrik Asplund. A systematic review to merge discourses: Interoperability, integration and cyber-physical systems. *Journal of Industrial Information Integration*, 9:14–23, March 2018. ISSN 2452-414X. doi: 10.1016/j.jii.2017.12.001.

[65] Akshay Rajhans, Ajinkya Bhave, Ivan Ruchkin, Bruce H. Krogh, David Garlan, André Platzer, and Bradley Schmerl. Supporting Heterogeneity in Cyber-Physical Systems Architectures. *IEEE Transactions on Automatic Control*, 59(12):3178–3193, December 2014. ISSN 1558-2523. doi: 10.1109/TAC.2014.2351672.

[66] Karan P. Jain and Mark W. Mueller. Flying batteries: In-flight battery switching to increase multirotor flight time. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3510–3516, May 2020. doi: 10.1109/ICRA40945.2020.9197580.

[67] Krzysztof Adam Szczurek, Raul Marin Prades, Eloise Matheson, Jose Rodriguez-Nogueira, and Mario Di Castro. Multimodal multi-user mixed reality human–robot interface for remote operations in hazardous environments. *IEEE Access*, 11:17305–17333, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3245833.

[68] Han Jiang, Yanchun Chang, Liying Yang, Xu Liu, and Yuqing He. Cooperative exploration of heterogeneous UAVs in mountainous environments by constructing steady communication. *IEEE Robotics and Automation Letters*, 8 (11):7249–7256, November 2023. ISSN 2377-3766. doi: 10.1109/LRA.2023.3316070.

[69] Alessio Sacco, Flavio Esposito, Guido Marchetto, and Paolo Montuschi. Sustainable Task Offloading in UAV Networks via Multi-Agent Reinforcement Learning. *IEEE Transactions on Vehicular Technology*, 70(5):5003–5015, May 2021. ISSN 1939-9359. doi: 10.1109/TVT.2021.3074304.

[70] Mohamed Elhoseny, R Sundar Rajan, Mohammad Hammoudeh, K Shankar, and Omar Aldabbas. Swarm intelligence–based energy efficient clustering with multihop routing protocol for sustainable wireless sensor networks. *International Journal of Distributed Sensor Networks*, 16(9): 1550147720949133, September 2020. ISSN 1550-1329. doi: 10.1177/1550147720949133.

[71] Alexander Schperberg, Stephanie Tsuei, Stefano Soatto, and Dennis Hong. SABER: Data-Driven Motion Planner for Autonomously Navigating Heterogeneous Robots. *IEEE Robotics and Automation Letters*, 6(4): 8086–8093, October 2021. ISSN 2377-3766. doi: 10.1109/LRA.2021.3103054.

[72] Xiangyu Liu and Kaiqing Zhang. Partially observable multi-agent RL with (quasi-)efficiency: The blessing of information sharing. In *Proceedings of the 40th International Conference on Machine Learning*, pages 22370–22419. PMLR, July 2023.

[73] Tyler Becker and Zachary Sunberg. Bridging the gap between partially observable stochastic games and sparse POMDP methods. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '25, pages 2434–2436, Richland, SC, June 2025. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 979-8-4007-1426-9.

[74] Yifan Zhong, Jakub Grudzien Kuba, Xidong Feng, Siyi Hu, Jiaming Ji, and Yaodong Yang. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research*, 25(32):1–67, 2024.

[75] Bangji Fan, Xinghua Liu, Yuanzhe Wang, Gaoxi Xiao, Yu Kang, and Danwei Wang. Graph-based heterogeneous multiagent reinforcement learning for distribution system service restoration. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 56(4):2399–2412, April 2026. ISSN 2168-2232. doi: 10.1109/TSMC.2025.3649418.

[76] Williard Joshua Jose and Hao Zhang. Learning for Dynamic Subteaming and Voluntary Waiting in Heterogeneous Multi-Robot Collaborative Scheduling. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4569–4576, May 2024. doi: 10.1109/ICRA57147.2024.10610342.

[77] Zixiang Nie, Kwang-Cheng Chen, and Kyeong Jin Kim. Social-learning coordination of collaborative multi-robot systems achieves resilient production in a smart factory. *IEEE Transactions on Automation Science and Engineering*, 22:6009–6023, 2025. ISSN 1558-3783. doi: 10.1109/TASE.2024.3435443.

[78] Jakob Bichler, Andreu Matoses Gimenez, and Javier Alonso-Mora. SADCHER: Scheduling using attention-based dynamic coalitions of heterogeneous robots in real-time, October 2025.

[79] Xuekai Qiu, Pengming Zhu, Yiming Hu, Zhiwen Zeng, and Huimin Lu. Consensus-based dynamic task allocation for multi-robot system considering payloads consumption. In *2024 China Automation Congress (CAC)*, pages 5294–5299, November 2024. doi: 10.1109/CAC63892.2024.10865375.

[80] Matthew Lai, Keegan Go, Zhibin Li, Torsten Kröger, Stefan Schaal, Kelsey Allen, and Jonathan Scholz. RoboBallet: Planning for multirobot reaching with graph neural networks and reinforcement learning. *Science Robotics*, 10(106): eads1204, September 2025. doi: 10.1126/scirobotics.ads1204.

[81] Jiaqi Tan, Yudong Luo, Sophia Huang, Yifan Yang, and Hang Ma. Crest: Constraint-release execution for multi-robot warehouse shelf rearrangement. *IEEE Robotics and Automation Letters*, 11(5):6447–6454, 2026. doi: 10.1109/lra.2026.3674027. IEEE article id: 10.1109/lra.2026.3674027.

[82] Jingkai Chen. *Hybrid Concurrent Planning with Heterogeneous Robot Teams for Timed Goals*. Thesis, Massachusetts Institute of Technology, September 2022.

[83] Charles S. Elton. *Animal Ecology*. Sidgwick and Jackson, London, 1927.

[84] G. Evelyn Hutchinson. Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, 22:415–427, 1957. doi: 10.1101/SQB.1957.022.01.039.

[85] Tonghan Wang, Tarun Gupta, Anuj Mahajan, Bei Peng, Shimon Whiteson, and Chongjie Zhang. Rode: Learning roles to decompose multi-agent tasks. In *9th International Conference on Learning Representations, ICLR 2021*, 2021.

[86] Frans A. Oliehoek and Christopher Amato. *A Concise Introduction to Decentralized POMDPs*. SpringerBriefs in Intelligent Systems. Springer International Publishing, Cham, 2016. ISBN 978-3-319-28927-4 978-3-319-28929-8. doi: 10.1007/978-3-319-28929-8.

[87] Christopher Amato, George Konidaris, Leslie P. Kaelbling, and Jonathan P. How. Modeling and Planning with Macro-Actions in Decentralized POMDPs. *The journal of artificial intelligence research*, 64:817, March 2019. doi:

10.1613/jair.1.11418.

[88] Vojtěch Kovařík, Martin Schmid, Neil Burch, Michael Bowling, and Viliam Lisý. Rethinking formal models of partially observable multiagent decision making. *Artificial Intelligence*, 303:103645, February 2022. ISSN 0004-3702. doi: 10.1016/j.artint.2021.103645.

[89] Xiao Du, Yutong Ye, Pengyu Zhang, Yaning Yang, Mingsong Chen, and Ting Wang. Situation-dependent causal influence-based cooperative multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17362–17370, 2024. doi: 10.1609/aaai.v38i16.29684. URL https://doi.org/10.1609/aaai.v38i16.29684.

[90] Woojun Kim, Jongeui Park, and Youngchul Sung. Communication in multi-agent reinforcement learning: Intention sharing. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qpsl2dR9twy.

[91] Graeme Best, Rohit Garg, John Keller, Geoffrey A. Hollinger, and Sebastian Scherer. Multi-robot, multi-sensor exploration of multifarious environments with full mission aerial autonomy. *The International Journal of Robotics Research*, 43(4):485–512, 2024. doi: 10.1177/02783649231203342.

[92] Andrew Howard, Lynne E. Parker, and Gaurav S. Sukhatme. Experiments with a large heterogeneous mobile robot team: Exploration, mapping, deployment and detection. *The International Journal of Robotics Research*, 25(5-6): 431–447, 2006. doi: 10.1177/0278364906065378.

[93] Stefan Jorgensen and Marco Pavone. The matroid team surviving orienteers problem and its variants: Constrained routing of heterogeneous teams with risky traversal. *The International Journal of Robotics Research*, 43(1):34–52, 2024. doi: 10.1177/02783649231210326.

[94] J. Cortes, S. Martinez, T. Karatas, and F. Bullo. Coverage control for mobile sensing networks. *IEEE Transactions on Robotics and Automation*, 20(2): 243–255, April 2004. ISSN 2374-958X. doi: 10.1109/TRA.2004.824698.

[95] David A. Johnson, David J. Naffin, Jeffrey S. Puhalla, Julian Sanchez, and Carl K. Wellington. Development and implementation of a team of robotic tractors for autonomous peat moss harvesting. *Journal of Field Robotics*, 26 (6-7):549–571, 2009. doi: 10.1002/rob.20297.

[96] Antonio Barrientos, Julian Colorado, Jaime del Cerro, Alexander Martinez, Claudio Rossi, David Sanz, and João Valente. Aerial remote sensing in agriculture: A practical approach to area coverage and path planning for fleets of mini aerial robots. *Journal of Field Robotics*, 28(5):667–689, 2011. doi: 10.1002/rob.20403.

[97] Audrey Guillet, Roland Lenain, Benoit Thuilot, and Vincent Rousseau. Formation control of agricultural mobile robots: A bidirectional weighted constraints approach. *Journal of Field Robotics*, 34(7):1260–1274, 2017. doi: 10.1002/rob.21704.

[98] Gautham Das, Grzegorz Cielniak, James Heselden, Simon Pearson, Francesco Del Duchetto, Zuyuan Zhu, Johann Dichtl, Marc Hanheide, Jaime Pulido Fentanes, Adam Binch, Michael Hutchinson, and Pal From. A unified topological representation for robotic fleets in agricultural applications.

Journal of Field Robotics, 42(3):760–786, 2025. doi: 10.1002/rob.22494.

[99] Valentin N. Hartmann, Andreas Orthey, Danny Driess, Ozgur S. Oguz, and Marc Toussaint. Long-Horizon Multi-Robot Rearrangement Planning for Construction Assembly. *IEEE Transactions on Robotics*, 39(1):239–252, February 2023. ISSN 1941-0468. doi: 10.1109/TRO.2022.3198020.

[100] Zaolin Pan and Yantao Yu. Optimizing heterogeneous multi-robot team composition for long-horizon construction tasks: Time- and utilization-guided simulation. *Automation in Construction*, 165:105520, September 2024. ISSN 0926-5805. doi: 10.1016/j.autcon.2024.105520.

[101] Jingkai Chen, Jiaoyang Li, Yijiang Huang, Caelan Garrett, Dawei Sun, Chuchu Fan, Andreas Hofmann, Caitlin Mueller, Sven Koenig, and Brian C. Williams. Cooperative Task and Motion Planning for Multi-Arm Assembly Systems, March 2022.

[102] Zhe Chen, Javier Alonso-Mora, Xiaoshan Bai, Daniel D. Harabor, and Peter J. Stuckey. Integrated task assignment and path planning for capacitated multi-agent pickup and delivery. *IEEE Robotics and Automation Letters*, 6(3): 5816–5823, 2021. doi: 10.1109/LRA.2021.3074883.

[103] Aleksandar Krnjaic, Jonathan D. Thomas, Georgios Papoudakis, Filippos Christianos, and Stefano V. Albrecht. Scalable multi-agent reinforcement learning for warehouse logistics with robotic and human co-workers, 2024. URL https://arxiv.org/abs/2212.11498.

[104] Ziyang Chen and Zhen Kan. Real-time reactive task allocation and planning of large heterogeneous multi-robot systems with temporal logic specifications. *The International Journal of Robotics Research*, 44(4):640–664, 2025. doi: 10.1177/02783649241278372.

[105] Anastasios Tsiamis, Christos K. Verginis, Charalampos P. Bechlioulis, and Kostas J. Kyriakopoulos. Cooperative manipulation exploiting only implicit communication. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 864–869, September 2015. doi: 10.1109/IROS.2015.7353473.

[106] Yi Ren, Stefan Sosnowski, and Sandra Hirche. Fully Distributed Cooperation for Networked Uncertain Mobile Manipulators. *IEEE Transactions on Robotics*, 36(4):984–1003, August 2020. ISSN 1941-0468. doi: 10.1109/TRO.2020.2971416.

[107] Yi Ren, Zhehua Zhou, Ziwei Xu, Yang Yang, Guangyao Zhai, Marion Leibold, Fenglei Ni, Zhengyou Zhang, Martin Buss, and Yu Zheng. Enabling Versatility and Dexterity of the Dual-Arm Manipulators: A General Framework Toward Universal Cooperative Manipulation. *IEEE Transactions on Robotics*, 40: 2024–2045, 2024. ISSN 1941-0468. doi: 10.1109/TRO.2024.3370048.

[108] Yuming Feng, Chuye Hong, Yaru Niu, Shiqi Liu, Yuxiang Yang, and Ding Zhao. Learning multi-agent loco-manipulation for long-horizon quadrupedal pushing. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14441–14448, May 2025. doi: 10.1109/ICRA55743.2025.11128478.

[109] Peng Zhou, Pai Zheng, Jiaming Qi, Chengxi Li, Hoi-Yin Lee, Anqing Duan, Liang Lu, Zhongxuan Li, Luyin Hu, and David Navarro-Alarcon. Reactive human–robot collaborative manipulation of deformable linear objects using a new topological latent control model. *Robotics and Computer-Integrated*

Manufacturing, 88:102727, August 2024. ISSN 0736-5845. doi: 10.1016/j.rcim.2024.102727.

[110] Sammy Christen, Wei Yang, Claudia Pérez-D'Arpino, Otmar Hilliges, Dieter Fox, and Yu-Wei Chao. Learning Human-to-Robot Handovers From Point Clouds. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9654–9664, 2023.

[111] Mandeep Dhanda, Benedict Alexander Rogers, Stephanie Hall, Elies Dekoninck, and Vimal Dhokia. Reviewing human-robot collaboration in manufacturing: Opportunities and challenges in the context of industry 5.0. *Robotics and Computer-Integrated Manufacturing*, 93:102937, June 2025. ISSN 0736-5845. doi: 10.1016/j.rcim.2024.102937.

[112] Roohollah Jahanmahin, Sara Masoud, Jeremy Rickli, and Ana Djuric. Human-robot interactions in manufacturing: A survey of human behavior modeling. *Robotics and Computer-Integrated Manufacturing*, 78:102404, December 2022. ISSN 0736-5845. doi: 10.1016/j.rcim.2022.102404.

[113] Lydia E Kavraki, Mihail N Kolountzakis, and J-C Latombe. Analysis of probabilistic roadmaps for path planning. *IEEE Transactions on Robotics and automation*, 14(1):166–171, 1998.

[114] Steven LaValle. Rapidly-exploring random trees: A new tool for path planning. *Research Report 9811*, 1998.

[115] Guni Sharon, Roni Stern, Ariel Felner, and Nathan R Sturtevant. Conflict-based search for optimal multi-agent pathfinding. *Artificial intelligence*, 219:40–66, 2015.

[116] Hang Ma, Daniel Harabor, Peter J. Stuckey, Jiaoyang Li, and Sven Koenig. Searching with consistent prioritization for multi-agent path finding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7643–7650, 2019.

[117] Jiaoyang Li, Wheeler Ruml, and Sven Koenig. EECBS: A bounded-suboptimal search for multi-agent path finding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12353–12362, 2021.

[118] Jiaoyang Li, Zhe Chen, Daniel Harabor, Peter J. Stuckey, and Sven Koenig. MAPF-LNS2: Fast repairing for multi-agent path finding via large neighborhood search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10256–10265, 2022.

[119] Keisuke Okumura. LaCAM: Search-based algorithm for quick multi-agent pathfinding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11655–11662, 2023.

[120] Zhuohui Zhang, Bin He, Bin Cheng, and Gang Li. Bridging training and execution via dynamic directed graph-based communication in cooperative multi-agent systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23395–23403, 2025. doi: 10.1609/aaai.v39i22.34507. URL https://doi.org/10.1609/aaai.v39i22.34507.

[121] Chiqiang Liu and Dazi Li. HYGMA: Hypergraph coordination networks with dynamic grouping for multi-agent reinforcement learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd*

International Conference on Machine Learning, volume 267 of *Proceedings of Machine Learning Research*, pages 38767–38788. PMLR, 2025. URL https://proceedings.mlr.press/v267/liu25af.html.

[122] Rui Tang, Biao Luo, and Yongzheng Cui. Autonomous partner selection for cooperative multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 25840–25848, 2026. doi: 10.1609/aaai.v40i30.39783. URL https://doi.org/10.1609/aaai.v40i30.39783.

[123] Yuchen Xiao, Weihao Tan, Joshua Hoffman, Tian Xia, and Christopher Amato. Asynchronous multi-agent deep reinforcement learning under partial observability. *The International Journal of Robotics Research*, 44(8): 1257–1286, July 2025. ISSN 0278-3649. doi: 10.1177/02783649241306124.

[124] Lars Lindemann and Dimos V. Dimarogonas. Feedback control strategies for multi-agent systems under a fragment of signal temporal logic tasks. *Automatica*, 106:284–293, August 2019. ISSN 0005-1098. doi: 10.1016/j.automatica.2019.05.013.

[125] Dawei Sun, Jingkai Chen, Sayan Mitra, and Chuchu Fan. Multi-Agent Motion Planning From Signal Temporal Logic Specifications. *IEEE Robotics and Automation Letters*, 7(2):3451–3458, April 2022. ISSN 2377-3766. doi: 10.1109/LRA.2022.3146951.

[126] Tichakorn Wongpiromsarn, Alphan Ulusoy, Calin Belta, Emilio Frazzoli, and Daniela Rus. Incremental synthesis of control policies for heterogeneous multi-agent systems with linear temporal logic specifications. In *2013 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5011–5018, 2013. doi: 10.1109/ICRA.2013.6631293.

[127] Wenliang Liu, Nathalie Majcherczyk, and Federico Pecora. Scalable multi-robot task allocation and coordination under signal temporal logic specifications. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3933–3939, 2025. doi: 10.1109/ICRA55743.2025.11128245.

[128] E. M. Clarke, E. A. Emerson, and A. P. Sistla. Automatic verification of finite-state concurrent systems using temporal logic specifications. *ACM Transactions on Programming Languages and Systems*, 8(2):244–263, 1986. doi: 10.1145/5397.5399.

[129] Amy Fang and Hadas Kress-Gazit. Automated task updates of temporal logic specifications for heterogeneous robots. In *2022 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4363–4369, 2022. doi: 10.1109/ICRA46639.2022.9812045.

[130] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2 edition, 2018.

[131] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

[132] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with

deep reinforcement learning. In *International Conference on Learning Representations*, 2016.

[133] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[134] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870, 2018.

[135] Daniel S. Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4):819–840, 2002.

[136] Lucian Busoniu, Robert Babuska, and Bart De Schutter. A Comprehensive Survey of Multiagent Reinforcement Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 38(2):156–172, March 2008. ISSN 1558-2442. doi: 10.1109/TSMCC.2007.913919.

[137] Pablo Hernandez-Leal, Bilal Kartal, and Matthew E. Taylor. A Survey and Critique of Multiagent Deep Reinforcement Learning. *Autonomous Agents and Multi-Agent Systems*, 33(6):750–797, November 2019. ISSN 1387-2532, 1573-7454. doi: 10.1007/s10458-019-09421-1.

[138] Annie Wong, Thomas Bäck, Anna V. Kononova, and Aske Plaat. Deep multiagent reinforcement learning: Challenges and directions. *Artificial Intelligence Review*, 56(6):5023–5056, June 2023. ISSN 1573-7462. doi: 10.1007/s10462-022-10299-x.

[139] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35: 24611–24624, December 2022.

[140] Jiarong Liu, Yifan Zhong, Siyi Hu, Haobo Fu, QIANG FU, Xiaojun Chang, and Yaodong Yang. Maximum entropy heterogeneous-agent reinforcement learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 57333–57374, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/fe066022bab2a6c6a3c57032a1623c70-Paper-Conference.pdf.

[141] Eduardo Pignatelli, Johan Ferret, Matthieu Geist, Thomas Mesnard, Hado van Hasselt, and Laura Toni. A survey of temporal credit assignment in deep reinforcement learning. *Transactions on Machine Learning Research*, December 2023. ISSN 2835-8856.

[142] Eric Wiewiora, Garrison Cottrell, and Charles Elkan. Principled Methods for Advising Reinforcement Learning Agents. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 792–799, Washington, DC, USA, August 2003. AAAI Press. ISBN 978-1-57735-189-4.

[143] Haozhe Ma, Zhengding Luo, Thanh Vinh Vo, Kuankuan Sima, and Tze-Yun Leong. Highly efficient self-adaptive reward shaping for reinforcement learning.

In *The Thirteenth International Conference on Learning Representations*, October 2024.

[144] Muyuan Ma and Long Cheng. A Human–Robot Collaboration Controller Utilizing Confidence for Disagreement Adjustment. *IEEE Transactions on Robotics*, 40:2081–2097, 2024. ISSN 1941-0468. doi: 10.1109/TRO.2024.3370025.

[145] Xuan Zuo, Pu Zhang, Hui-Yan Li, and Zhun-Ga Liu. Preference-based experience sharing scheme for multi-agent reinforcement learning in multi-target environments. *Evolving Systems*, May 2024. ISSN 1868-6486. doi: 10.1007/s12530-024-09587-4.

[146] Chao Li, Shaokang Dong, Shangdong Yang, Yujing Hu, Tianyu Ding, Wenbin Li, and Yang Gao. Multi-task multi-agent reinforcement learning with interaction and task representations. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2024. ISSN 2162-2388. doi: 10.1109/TNNLS.2024.3475216.

[147] Deheng Ye and Zongqing Lu. Mutual-Information Regularized Multi-Agent Policy Iteration. *Annual Conference on Neural Information Processing Systems*, 36, December 2023.

[148] Asuka Takai, Qiushi Fu, Yuzuru Doibata, Giuseppe Lisi, Toshiki Tsuchiya, Keivan Mojtahedi, Toshinori Yoshioka, Mitsuo Kawato, Jun Morimoto, and Marco Santello. Role specialization enables superior task performance by human dyads than individuals. *The International Journal of Robotics Research*, page 02783649251363274, August 2025. ISSN 0278-3649. doi: 10.1177/02783649251363274.

[149] Christina L Wiesmann, Nicole R Wang, Yue Zhang, Zhexian Liu, and Cara H Haney. Origins of symbiosis: Shared mechanisms underlying microbial pathogenesis, commensalism and mutualism of plants and animals. *FEMS Microbiology Reviews*, 47(6):fuac048, November 2023. ISSN 0168-6445. doi: 10.1093/femsre/fuac048.

[150] Scott A Chamberlain, Judith L Bronstein, and Jennifer A Rudgers. How context dependent are species interactions? *Ecology letters*, 17(7):881–890, 2014.

[151] Tanita Wein, Devani Romero Picazo, Frances Blow, Christian Woehle, Elie Jami, Thorsten B. H. Reusch, William F. Martin, and Tal Dagan. Currency, exchange, and inheritance in the evolution of symbiosis. *Trends in Microbiology*, 27(10): 836–849, October 2019. ISSN 0966-842X. doi: 10.1016/j.tim.2019.05.010.

[152] Claire N. Spottiswoode, Keith S. Begg, and Colleen M. Begg. Reciprocal signaling in honeyguide-human mutualism. *Science*, 353(6297):387–389, July 2016. doi: 10.1126/science.aaf4885.

[153] Lutz Becks, Ursula Gaedke, and Toni Klauschies. Emergent feedback between symbiosis form and population dynamics. *Trends in Ecology & Evolution*, 40 (5):449–459, May 2025. ISSN 0169-5347. doi: 10.1016/j.tree.2025.02.006.

[154] Serguei Saavedra, Rudolf P Rohr, Jordi Bascompte, Oscar Godoy, Nathan JB Kraft, and Jonathan M Levine. A structural approach for understanding multispecies coexistence. *Ecological Monographs*, 87(3):470–486, 2017.

[155] Jimmy J Qian and Erol Akçay. The balance of interaction types determines the assembly and stability of ecological communities. *Nature ecology & evolution*, 4(3):356–365, 2020.

[156] Fernanda S Valdovinos. Mutualistic networks: moving closer to a predictive theory. *Ecology letters*, 22(9):1517–1534, 2019.
[157] Alfred J. Lotka. *Elements of Physical Biology*. Williams & Wilkins, Baltimore, 1925.
[158] Vito Volterra. Fluctuations in the abundance of a species considered mathematically. *Nature*, 118:558–560, 1926. doi: 10.1038/118558a0.
[159] Alexander A. Nguyen, Faryar Jabbari, and Magnus Egerstedt. Mutualistic Interactions in Heterogeneous Multi-Agent Systems. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 411–418, December 2023. doi: 10.1109/CDC49753.2023.10384078.
[160] Federico Barravecchia, Mirco Bartolomei, Luca Mastrogiacomo, and Fiorenzo Franceschini. Redefining human–robot symbiosis: A bio-inspired approach to collaborative assembly. *The International Journal of Advanced Manufacturing Technology*, 128(5):2043–2058, September 2023. ISSN 1433-3015. doi: 10.1007/s00170-023-11920-1.
[161] T. Eguchi, K. Hirasawa, J. Hu, and N. Ota. A study of evolutionary multiagent models based on symbiosis. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(1):179–193, February 2006. ISSN 1941-0492. doi: 10.1109/TSMCB.2005.856720.
[162] Shingo Mabu, Masanao Obayashi, and Takashi Kuremoto. Reinforcement Learning with Symbiotic Relationships for Multiagent Environments. *J. Robotics Netw. Artif. Life*, 2(1):40–45, 2015.
[163] Melissa Demartini, Flavio Tonelli, and Kannan Govindan. An investigation into modelling approaches for industrial symbiosis: A literature review and research agenda. *Cleaner Logistics and Supply Chain*, 3:100020, March 2022. ISSN 2772-3909. doi: 10.1016/j.clscn.2021.100020.
[164] Liang Dong, Gideon Nkam Taka, Daye Lee, Yujin Park, and Hung Suck Park. Tracking industrial symbiosis performance with ecological network approach integrating economic and environmental benefits analysis. *Resources, Conservation and Recycling*, 185:106454, October 2022. ISSN 0921-3449. doi: 10.1016/j.resconrec.2022.106454.
[165] Xin Cao, Mingxuan Wu, Zechen Zhang, Yumeng Li, and Sai Liang. Designing cross-city symbiosis strategy in urban agglomeration by trading off decarbonization, health and economic benefits. *npj Urban Sustainability*, 6(1): 18, December 2025. ISSN 2661-8001. doi: 10.1038/s42949-025-00323-8.
[166] Yingxin Wang, Jian Sun, Hua Shang, Le Sun, Xiangyu Lan, Qiezhuo Lamao, and Jiayue Lu. An ecological network approach to assessing the site suitability of photovoltaic power stations. *Journal of Applied Ecology*, 63(1):e70222, 2026. ISSN 1365-2664. doi: 10.1111/1365-2664.70222.
[167] Xuezhi Niu, Natalia Calvo Barajas, and Didem Gürdür Broo. Enabling symbiosis in multi-robot systems through multi-agent reinforcement learning. In *2025 IEEE 8th International Conference on Industrial Cyber-Physical Systems (ICPS)*, pages 1–7, May 2025. doi: 10.1109/ICPS65515.2025.11087893.
[168] Xuezhi Niu and Didem Gürdür Broo. Investigating symbiosis in robotic ecosystems: A case study for multi-robot reinforcement learning reward shaping. In *2025 9th International Conference on Robotics and Automation Sciences (ICRAS)*, pages 112–117, June 2025. doi:

10.1109/ICRAS65818.2025.11108729.

[169] Dimos V. Dimarogonas, Emilio Frazzoli, and Karl H. Johansson. Distributed Event-Triggered Control for Multi-Agent Systems. *IEEE Transactions on Automatic Control*, 57(5):1291–1297, May 2012. ISSN 1558-2523. doi: 10.1109/TAC.2011.2174666.

[170] Shayegan Omidshafiei, Ali-Akbar Agha–Mohammadi, Christopher Amato, Shih-Yuan Liu, Jonathan P How, and John Vian. Decentralized control of multi-robot partially observable markov decision processes using belief space macro-actions. *The International Journal of Robotics Research*, 36(2):231–258, February 2017. ISSN 0278-3649. doi: 10.1177/0278364917692864.

[171] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In Kyriakos G. Vamvoudakis, Yan Wan, Frank L. Lewis, and Derya Cansever, editors, *Handbook of Reinforcement Learning and Control*, Studies in Systems, Decision and Control, pages 321–384. Springer International Publishing, Cham, 2021. ISBN 978-3-030-60990-0. doi: 10.1007/978-3-030-60990-0_12.

[172] Marian Körber, Johann Lange, Stephan Rediske, Simon Steinmann, and Roland Glück. Comparing Popular Simulation Environments in the Scope of Robotics and Reinforcement Learning, March 2021.

[173] Mayank Mittal et al. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning, 2025. URL https://arxiv.org/abs/2511.04831.

[174] Yuanpei Chen, Yiran Geng, Fangwei Zhong, Jiaming Ji, Jiechuang Jiang, Zongqing Lu, Hao Dong, and Yaodong Yang. Bi-DexHands: Towards Human-Level Bimanual Dexterous Manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15, 2023. ISSN 1939-3539. doi: 10.1109/TPAMI.2023.3339515.

[175] Georgios Papoudakis, Filippos Christianos, Lukas Schafer, and Stefano V. Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2006.07869.