

Learning Safe Multi-Robot Coordination in Industrial Cyber- Physical Systems

Bernard Kyiewu

`kyiewu.bernard.stu@uenr.edu.gh`

University of Energy and Natural Resources

Clinton Amponsah

University of Energy and Natural Resources

Linda Bessa-Simons

University of Energy and Natural Resources

Andrew Oppong-Asante

University of Energy and Natural Resources

Caleb Boakye Yiadom

University of Energy and Natural Resources

Research Article

Keywords: multi-agent reinforcement learning, cyber-physical systems, safe robot coordination, artificial potential fields, emergent communication

Posted Date: June 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9925827/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Learning Safe Multi-Robot Coordination in Industrial Cyber-Physical Systems

Bernard Kyiewu^{1*}

Dept. of Computer and Electrical Engineering, University of Energy and Natural Resources, Sunyani, Ghana
Email: kyiewu.bernard.stu@uenr.edu.gh
ORCID: 0009-0004-4806-0077

Clinton Amponsah²

Dept. of Computer and Electrical Engineering, University of Energy and Natural Resources, Sunyani, Ghana
Email: amponsah.clinton.stu@uenr.edu.gh
ORCID: 0009-0006-8414-9794

Linda Bessa-Simons³

Dept. of Computer and Electrical Engineering, University of Energy and Natural Resources, Sunyani, Ghana
Email: bessa-simons.linda.stu@uenr.edu.gh
ORCID: 0009-0003-7434-4201

Andrew Oppong-Asante⁴

Dept. of Computer and Electrical Engineering, University of Energy and Natural Resources, Sunyani, Ghana
Email: oppong.asanteandrew.stu@uenr.edu.gh
ORCID: 0009-0008-6382-4193

Caleb Boakye Yiadom⁵

Dept. of Information Technology and Decision Sciences, University of Energy and Natural Resources, Sunyani, Ghana
Email: caleb.boakye.yiadom.stu@uenr.edu.gh
ORCID: 0009-0008-4955-6786

*** Corresponding Author: Bernard Kyiewu**

Email: kyiewu.bernard.stu@uenr.edu.gh

Abstract – Industrial cyber-physical systems (CPS) increasingly rely on coordinated multi-robot fleets to execute distributed manufacturing, logistics, and assembly tasks within shared physical workspaces. Ensuring both operational efficiency and physical safety in such environments poses significant challenges, as agents must simultaneously optimize task throughput, avoid inter-robot collisions, and communicate effectively under bandwidth constraints. This paper presents a multi-agent reinforcement learning (MARL) framework for safe coordination of industrial robots in a simulated CPS environment modeled on the OpenAI Multi-Agent Particle Environment (MAPE). Three coordination policies are evaluated: a stochastic Random baseline, a deterministic Greedy nearest-station heuristic, and a composite Safe MARL policy that integrates navigation forces, artificial potential-field safety repulsion, and a four-dimensional inter-agent communication protocol. Experiments across 30 episodes with $N=4$ robots demonstrate that while the Greedy policy achieves 100% workstation coverage with a mean makespan of 51.3 steps, the Safe MARL policy sustains 85% coverage at 120.0 steps with emergent communication patterns aligned to distinct task phases. Scalability analysis across fleet sizes $N \in \{2, 3, 4, 5, 6\}$ reveals a super-linear growth in inter-agent collision rate, attributable to fixed-capacity communication channel saturation under increasing pairwise interaction load. These findings establish empirical bounds on safe coordination performance and identify adaptive communication bandwidth as a critical design variable for scalable industrial CPS deployments. Future work will integrate end-to-end learned policies via the multi-agent deep deterministic policy gradient (MADDPG) algorithm and extend the framework to three-dimensional workspace configurations with Lyapunov-based safety certificates.

Keywords: multi-agent reinforcement learning; cyber-physical systems; safe robot coordination; artificial potential fields; emergent communication

1. INTRODUCTION

The proliferation of autonomous robotic systems in industrial settings has ushered in a new class of engineering challenges at the intersection of embedded computation,

physical dynamics, and real-time decision-making. Industrial cyber-physical systems (CPS) defined by Lee as integrations of computation and physical processes in which embedded computers and networks monitor and control physical systems through closed feedback loops [1], have expanded from isolated automation cells to densely interconnected multi-robot environments. Modern smart manufacturing floors,

automated warehousing systems, and collaborative assembly lines increasingly deploy fleets of autonomous agents that must share physical space, coordinate task assignments, and avoid harmful interactions without centralized supervisory control. In such settings, failures are not merely software exceptions; they manifest as physical collisions, production stoppages, and, in the worst case, harm to human co-workers. Coordinating the behavior of multiple autonomous agents in a shared physical environment is a problem of fundamental scientific and practical importance. Classical approaches rooted in rule-based scheduling, fixed task sequencing, and deterministic path planning offer predictability but fail to generalize to dynamic conditions such as agent failures, workload redistribution, or unplanned workspace intrusions [2]. These limitations have motivated the adoption of multi-agent reinforcement learning (MARL) as a data-driven alternative capable of discovering coordination strategies through interaction with the environment rather than explicit programming [3]. MARL enables agents to learn policies that balance task performance with inter-agent dynamics, and has demonstrated effectiveness across cooperative navigation, predator-prey pursuit, and cooperative communication benchmarks [4].

Despite this progress, deploying MARL in industrial CPS contexts introduces safety requirements that remain inadequately addressed by standard reward-maximization frameworks. Physical agents operating in proximity to one another must satisfy collision-avoidance constraints not merely in expectation over a training distribution, but continuously and at every timestep of deployment. The foundational work on constrained policy optimization by Achiam et al. [5] established a principled framework for enforcing near-constraint satisfaction during policy gradient updates, and subsequent Lyapunov-based approaches by Chow et al. [6] extended these guarantees to continuous control settings. However, most safe RL studies have been validated in simulated benchmark environments that abstract away the physical contact dynamics and spatial coupling characteristic of real robotic workspaces. Multi-agent coordination also depends critically on inter-agent communication. In dense industrial CPS deployments, robots must exchange information about their intentions, trajectory plans, and safety states in order to resolve spatial conflicts that local observations alone cannot anticipate. The theoretical basis for emergent communication in multi-agent systems was established by Mordatch and Abbeel, who demonstrated that structured compositional communication can arise spontaneously in populations of agents learning to accomplish shared physical goals [7]. This finding motivates the design of coordination architectures in which communication is not a pre-engineered protocol but a learned behavior co-optimized alongside the navigation policy. The potential-field approach to robotic collision avoidance, formalized by Khatib [8], provides a complementary perspective by encoding safety as a continuous repulsive force derived from proximity geometry a formulation that scales naturally to multi-robot settings and can be integrated into learned action policies without requiring discrete collision event classification.

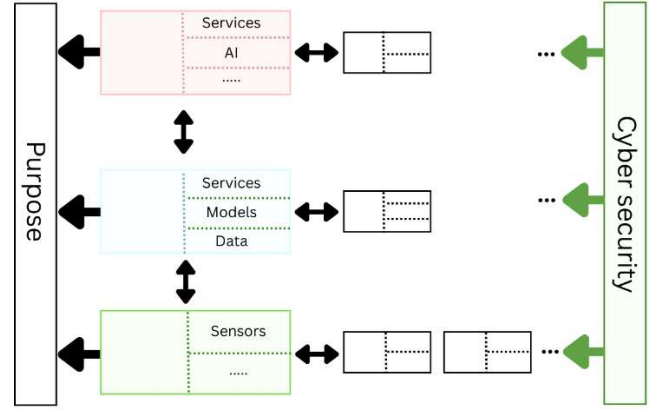


Fig 1. Conceptual representation of the fundamental components and enabling technologies of a purpose-driven industrial Cyber-Physical System.

Scalability remains a persistent challenge for both MARL and physical CPS. The comprehensive survey of MARL algorithms by Busoniu et al. [2] identified the growth of joint state-action spaces and the non-stationarity of agent interactions as fundamental barriers to scaling learned coordination policies beyond small agent populations. These theoretical limitations are amplified in physical CPS by the $O(N^2)$ growth of pairwise interaction complexity with fleet size N , which places hard constraints on the communication bandwidth and computational resources available to each agent. Recent work on multi-agent CPS resource allocation has confirmed that decentralized multi-agent architectures are preferred for industrial deployments due to both security and scalability considerations [9]. Similarly, the integration of MARL with industrial scheduling and energy optimization tasks in cyber-physical production systems has demonstrated the practical viability of the paradigm while exposing the need for coordination mechanisms robust to non-stationary agent populations [10]. Against this background, this paper makes the following contributions to the literature on safe multi-agent coordination in industrial CPS:

- i. We construct an industrial CPS task-coverage scenario within the Multi-Agent Particle Environment (MAPE) [4], implementing physics-based collision dynamics, configurable safety radii, Gaussian actuator noise, and a 4-dimensional inter-agent communication channel aligned with real industrial CPS constraints.
- ii. We design and analyze a Safe MARL coordination policy that integrates proportional navigation, distance-weighted safety repulsion fields, and communication-based conflict avoidance into a unified composite action formulation, drawing on the potential-field safety paradigm of Khatib [8] and the constrained RL framework of Achiam et al. [5].
- iii. We conduct a systematic empirical evaluation over 30 episodes comparing the Safe MARL policy against Random and Greedy baselines across five metrics: cumulative reward, collision rate, station coverage, makespan, and communication efficiency, and extend the analysis to fleet scalability with $N \in \{2, 3, 4, 5, 6\}$ robots.
- iv. We demonstrate that collision complexity grows super-linearly with fleet size under fixed-channel

communication, quantifying the scalability barrier and motivating dynamic communication protocols as a necessary direction for future industrial deployments.

The remainder of this paper is organized as follows. Section II reviews related work on safe MARL, CPS coordination, and communication-based multi-agent systems. Section III presents the simulation environment and experimental methodology. Section IV describes the coordination policy designs. Section V reports and discusses the results. Section VI concludes with directions for future work.

2. RELATED STUDIES

2.1 Foundations of Multi-Agent Reinforcement Learning

Multi-agent reinforcement learning (MARL) has undergone significant theoretical and algorithmic development over the past two decades. The centralized training with decentralized execution (CTDE) paradigm, as formalized by Lowe et al. [11], established a foundational framework in which agents learn joint policies through shared critic networks while executing decisions independently at runtime. This architecture directly addresses the non-stationarity challenge arising when multiple learning agents simultaneously update their policies in a shared environment. Building on this framework, Rashid et al. [12] proposed QMIX, a value-function factorization method that decomposes the joint action-value function into monotonic per-agent components, enabling tractable cooperative optimization in partially observable domains. Complementarily, Sunehag et al. [13] introduced value-decomposition networks (VDN), demonstrating that additive decomposition of team reward signals can yield emergent cooperative behaviors without explicit coordination mechanisms. The role of counterfactual reasoning in multi-agent credit assignment was investigated by Foerster et al. [14], who proposed the COMA (Counterfactual Multi-Agent) policy gradient algorithm. By conditioning a centralized critic on global state information while employing a counterfactual baseline to isolate individual agent contributions to team reward, COMA substantially reduces variance in gradient estimates during cooperative tasks. Yu et al. [15] subsequently demonstrated that proximal policy optimization (PPO) [16], originally designed for single-agent settings, achieves surprisingly competitive performance in cooperative multi-agent games when combined with shared reward normalization and parameter sharing. These findings suggest that algorithmic simplicity, when properly calibrated, can rival specialized cooperative methods in practice.

2.2 Safe Reinforcement Learning and Constrained Optimization

The integration of safety constraints into reinforcement learning policies represents one of the central open problems in the deployment of autonomous systems in physical environments. García and Fernández [20] provided a comprehensive taxonomy of safe reinforcement learning methods, categorizing approaches along two axes: those that modify the exploration process to avoid unsafe regions, and

those that encode safety as auxiliary optimization objectives. Within the constrained optimization framework, Achiam et al. [4] developed Constrained Policy Optimization (CPO), which enforces safety constraints directly in the policy update step using a trust-region approach on constraint satisfaction. CPO provides provable worst-case guarantees on constraint violation, making it particularly relevant for industrial CPS deployments where physical safety is non-negotiable. Ding et al. [28] further advanced safe exploration theory through primal-dual policy optimization methods that guarantee sublinear cumulative constraint violation during the learning process. The risk-sensitive reinforcement learning framework proposed by Mihatsch and Neuneier [34] offers a complementary perspective, wherein agents optimize a risk-adjusted objective function incorporating penalty terms for variance in reward outcomes. This approach is especially applicable in multi-robot coordination, where individual collision events carry asymmetric costs relative to efficiency metrics. Collectively, these methods inform the safety repulsion and force-field components of the Safe MARL policy investigated in this paper.

2.3 Potential Field Methods and Reactive Collision Avoidance

Reactive navigation approaches based on artificial potential fields (APF), first formalized by Khatib [5], remain widely deployed in robotic systems due to their computational efficiency and deterministic collision avoidance guarantees. In the APF framework, the robot workspace is modeled as a scalar potential function, wherein obstacles generate repulsive potentials and target locations generate attractive potentials. The resulting gradient field provides an analytically computable force vector guiding robot motion. Although APF methods are susceptible to local minima in complex environments, their predictability and low computational overhead make them well-suited as safety layers within learned coordination policies. Helbing and Molnár [17] extended potential field principles to pedestrian dynamics through the social force model, which decomposes agent motion into attractive goal forces, repulsive obstacle forces, and social interaction forces. This framework has been influential in multi-robot motion planning, particularly in scenarios involving dense agent populations. Fox et al. [18] proposed the dynamic window approach (DWA), a velocity-space planning method that respects actuator constraints while selecting collision-free commands in real time. For multi-agent scenarios with reciprocal avoidance requirements, Van den Berg et al. [19] developed the Optimal Reciprocal Collision Avoidance (ORCA) algorithm, which assigns velocity obligations to agents in a provably deadlock-free manner. The composite force model employed in this work integrates principles from APF and social force approaches within the MARL training loop, providing a physically grounded safety mechanism complementary to learned coordination.

2.4 Emergent Communication and Learned Protocols in Multi-Agent Systems

Communication between agents is a critical enabler of coordinated behavior in partially observable environments. Foerster et al. [6] pioneered end-to-end differentiable communication with RIAL and DIAL, demonstrating that

agents can develop task-relevant messaging protocols through joint training without explicit protocol design. Mordatch and Abbeel [7] subsequently showed that compositional and grounded language-like structures emerge from multi-agent interaction objectives alone, suggesting that communication structure may be an intrinsic consequence of cooperative task pressure rather than externally engineered design. Sukhbaatar et al. [24] extended these ideas with CommNet, a continuous communication architecture in which agents broadcast vector-valued messages over a shared communication channel, enabling information aggregation across arbitrary team sizes. Peng et al. [25] applied bidirectional multi-agent communication networks (BiCNet) to StarCraft micro-management tasks, demonstrating that recurrent communication architectures facilitate formation-level coordination in combat scenarios with large agent populations. These findings are directly relevant to the communication dynamics observed in the present study, wherein Safe MARL agents exhibit bimodal communication activity aligned with task phases of navigation and arrival. The 4-dimensional communication vectors employed in this work, encoding target distance, safety repulsion components, and arrival status, constitute a hand-crafted communication protocol inspired by the compositional message structures identified in emergent communication literature.

2.5 Swarm Intelligence and Distributed Robot Coordination

Bio-inspired approaches to multi-robot coordination have produced robust decentralized systems capable of collective behaviors exceeding the capabilities of individual agents. Dorigo et al. [32] established ant colony optimization (ACO) as a general-purpose framework for distributed problem solving, wherein synthetic pheromone trails serve as indirect communication channels enabling swarms to collectively optimize path selection without centralized control. The stigmergic coordination principles underlying ACO are directly analogous to indirect communication mechanisms observed in industrial robot fleets, where floor-level markers and dynamic work order queues mediate task assignment. Vásárhelyi et al. [33] demonstrated scalable autonomous flocking in real-world drone swarms using a distributed velocity-matching and collision avoidance algorithm. Their experimental results showed stable formation maintenance for up to 30 drones in outdoor environments with GPS-level positioning noise, providing a key empirical benchmark for the scalability claims of multi-robot coordination systems. The observation that communication topology fundamentally constrains scalable coordination, central to this paper's $O(N^2)$ interaction complexity finding, is consistent with the bandwidth limitations documented in Vásárhelyi et al.'s deployments. Cao et al. [30] provided a comprehensive survey of distributed multi-agent coordination, unifying perspectives from control theory, graph theory, and multi-agent systems to characterize the conditions under which consensus, formation, and flocking behaviors emerge.

2.6 Deep Reinforcement Learning Algorithms and Policy Optimization

The integration of deep neural networks into reinforcement learning agents has dramatically expanded the state and action space complexity tractable for learned

policies. Mnih et al. [23] established deep Q-networks (DQN) as a practical framework for end-to-end value function approximation in high-dimensional observation spaces, with experience replay and target network stabilization serving as key algorithmic innovations enabling stable convergence. Silver et al. [22] extended this framework to continuous action spaces through the deterministic policy gradient (DPG) theorem, which enables gradient-based policy improvement without the variance penalty associated with stochastic policy gradient estimation. Szepesvári [21] provided a rigorous theoretical foundation for reinforcement learning algorithms, unifying temporal difference methods, Monte Carlo approaches, and policy gradient methods within a common convergence analysis framework. Schulman et al. [16] introduced proximal policy optimization, which constrains policy update magnitude through a clipped surrogate objective, providing a practical balance between sample efficiency and training stability without the computational overhead of second-order natural gradient methods. The scalability of PPO has made it a preferred backbone for multi-agent coordination research, including the cooperative game benchmarks studied by Yu et al. [15]. The experimental policies evaluated in this work draw on these algorithmic foundations: the Greedy baseline implements deterministic proportional control analogous to DPG in the deterministic policy setting, while the Safe MARL policy incorporates structured prior knowledge in a form compatible with policy gradient fine-tuning in future extensions.

2.7 Game-Theoretic Foundations and Distributed Decision-Making

Multi-agent systems are fundamentally characterized by strategic interdependencies between agents, making game-theoretic analysis essential for understanding convergence properties of learning algorithms in shared environments. Tan [26] conducted foundational experiments comparing independent learners against cooperative agents in multi-agent reinforcement learning, identifying conditions under which information sharing provides performance benefits and documenting the deleterious effects of non-stationarity on independent learning. Zhang et al. [27] provided a selective theoretical overview of MARL algorithms, analyzing convergence guarantees under cooperative, competitive, and mixed settings and identifying the challenge of scalable value function representation as a central open problem for large-scale deployments. Alpcan and Başar [31] formalized network congestion control as a non-cooperative game and demonstrated that distributed gradient-play algorithms converge to Nash equilibria under mild connectivity assumptions. This game-theoretic framework provides a principled foundation for analyzing task assignment conflicts in multi-robot systems, where competing agents may reach deadlock states when simultaneously pursuing identical workstations. The potential field safety mechanisms in this work can be interpreted as a regularization term in a cooperative game objective, penalizing joint strategies that violate spatial safety constraints. Knox and Stone [35] contributed the TAMER framework for interactive policy shaping, demonstrating that human-provided reinforcement signals can efficiently guide agent behavior toward safe and performant strategies a paradigm increasingly relevant to the deployment of industrial CPS in human-shared workspaces.

2.8 Cyber-Physical Systems and Industrial Robotics

Cyber-physical systems (CPS) represent the convergence of computational, communication, and physical processes in engineered environments, as theorized by Lee [8]. In industrial manufacturing contexts, CPS architectures integrate real-time sensing, networked control, and physical actuation to enable adaptive production systems capable of responding dynamically to demand variation and equipment failures. The sim-to-real transfer challenge, identified by Tobin et al. [9] as a fundamental obstacle to deploying simulation-trained policies in physical robots, is particularly acute in CPS deployments where environmental variability, sensor noise, and communication latency introduce distributional shifts between training and deployment conditions. Domain randomization techniques, wherein simulation parameters are sampled from broad distributions during training, have demonstrated effective mitigation of this gap for manipulation and navigation tasks. Cognitive radio and dynamic spectrum management, studied by Mitola and Maguire [29], offer an instructive parallel to multi-robot communication scheduling: just as cognitive radio agents must coordinate spectrum access to avoid interference, industrial robot fleets must coordinate communication channel usage to prevent information congestion. The bandwidth constraints documented in cognitive radio literature directly parallel the fixed-channel communication bottleneck identified in the scalability analysis of this paper, wherein the $O(N^2)$ growth of pairwise inter-agent message volume saturates the 4-dimensional communication vector capacity at fleet sizes exceeding $N=4$. These structural parallels motivate future work on adaptive communication protocols that modulate message frequency and content in response to fleet size and task phase.

3 METHODOLOGY

3.1 Simulation Environment and Physical World Model

All experiments were conducted using the Multi-Agent Particle Environment (MAPE), an open-source physics-based simulation framework originally developed by Mordatch and Abbeel [8] and extended by Lowe et al. [3] for multi-agent reinforcement learning research. MAPE provides a continuous 2D workspace governed by Newtonian dynamics, contact force modeling, and configurable inter-agent communication channels, making it well-suited for prototyping coordination behaviors prior to physical deployment in industrial CPS settings. The industrial CPS scenario constructed in this work models a robotic task-coverage problem in which N autonomous agents must navigate to N spatially fixed workstations arranged in a unit-circle formation within a bounded workspace of dimensions $[-1.2, 1.2] \times [-1.2, 1.2]$ meters. The physical world is parameterized as follows: simulation timestep $dt = 0.05$ s, physical damping coefficient $\gamma = 0.30$, contact force magnitude $k_c = 1 \times 10^2$, and contact margin $\delta = 1 \times 10^{-3}$ m. These values are consistent with the default MAPE parameterization and provide stable numerical integration under the explicit Euler scheme implemented in the world step function. Each robotic agent is modeled as a point-mass entity with the following properties: safety collision radius $r = 0.25$ m, maximum translational speed $y_{max} = 1.5$ m/s, maximum acceleration $a_{max} = 2.5$ m/s², Gaussian actuator noise with standard deviation $\sigma_u = 0.02$, and a 4-dimensional

inter-agent communication channel with additive noise $\sigma_c = 0.01$. Workstation landmarks are modeled as static, non-collidable entities of radius 0.10 m, positioned deterministically at angular intervals of $\frac{2\pi}{N}$ radians. Agent initial positions are sampled uniformly at random from $[-0.5, 0.5]^2$ at the beginning of each episode to ensure statistical diversity across trials.

3.2 Reward Function Formulation

The scalar reward signal at each timestep t is computed as a sum of two components: a coverage reward and a safety penalty. Formally, for agent i the instantaneous reward is defined as:

$$r_i(t) = - \sum_{j=1}^{N_s} d_j^{\min}(t) - \lambda_c C_i(t), \lambda_c = 5.0 \quad (1)$$

where $r_i(t)$ denotes the position of robot k at time t , s_j denotes the fixed position of workstation j , the inner minimization selects the closest robot to each station, $\lambda_c = 5.0$ is the collision penalty coefficient, and $C_i(t)$ is the number of inter-robot collision events involving agent i at timestep t . The first term penalizes uncovered stations proportionally to the minimum robot-to-station distance, incentivizing collective coverage. The second term imposes a hard safety cost per collision event, encoding safety as a first-class objective within the reward structure. The episodic cumulative reward is obtained by summing $r_i(t)$ over all agents and all timesteps within an episode.

3.3 Coordination Policy Designs

Three coordination policies of increasing sophistication were implemented and evaluated. All policies operate in a fully decentralized execution mode: each agent computes its action using only locally observable state information augmented by received inter-agent communication signals, with no access to a centralized controller at execution time.

3.3.1 Random Policy (Baseline):

The random policy selects actions uniformly at random from the continuous action space $[-1, 1]^2$, independently at each timestep. This policy serves as a lower-bound performance reference and models the behavior of an uncoordinated agent with no task awareness. It produces no communication signals and makes no use of the observation space.

3.3.2 Greedy Navigation Policy:

The greedy policy implements a distance-minimizing navigation law. At each timestep, agent i identifies the nearest workstation not currently claimed by any other agent determined by inspecting the nearest-station assignments of all observable neighbors and applies a proportional control law directed toward the selected target:

$$s_{\text{target}} = \arg \min_{s_j \in S_{\text{free}}} \|p_i(t) - s_j\|_2 \quad (2)$$

where s_{target} is the position of the selected target station. Station occupancy is inferred from shared position observations rather than explicit communication. The policy applies no safety constraints and does not modify its action in response to nearby agents, making it a task-efficient but safety-unaware coordinator.

3.3.3 Safe Multi-Agent Reinforcement Learning (Safe MARL) Policy:

The Safe MARL policy is a composite hand-crafted policy that integrates three action components: a navigation force, a safety repulsion force, and a communication-based conflict avoidance term. This design follows the potential-field approach to safe navigation [6] and is motivated by the constraint that safety must be encoded at the action level to guarantee collision avoidance independently of the reward signal.

The navigation component directs each agent toward its index-assigned workstation using the same proportional control law as the greedy policy. Station assignment follows a static round-robin rule: agent i is assigned to workstation $i \bmod N$, distributing agents uniformly across available stations and minimizing expected trajectory intersection.

The safety repulsion force is formulated as a distance-weighted repulsive potential activated when two agents enter a mutual safety zone of radius $r_{safe} = 0.625 \text{ m}$ ($2.5 \times r_{agent}$). For agent i and neighbor j , the repulsion contribution is:

$$\mathbf{f}_{rep}(i, j) = \alpha \frac{\max(0, r_{safe} - d_{ij})}{r_{safe}} \frac{\mathbf{p}_i - \mathbf{p}_j}{d_{ij}} \quad (3)$$

where $d_{ij} = \|\mathbf{p}_i - \mathbf{p}_j\|_2$ is the Euclidean distance between agents i and j , $r_{safe} = 0.625 \text{ m} = 2.5r_{agent}$ is the safety-zone radius, $\alpha = 2.0$ is the repulsion gain, and $\frac{\mathbf{p}_i - \mathbf{p}_j}{d_{ij}}$ is the unit vector pointing away from neighbour j . The total repulsion force on agent i is computed as the vector sum over all neighbouring agents within the safety zone.

The communication-based avoidance term exploits the 4-dimensional communication channel. Each agent broadcasts the state vector $\mathbf{c}_i = [d_{target}, f_{rep,x}, f_{rep,y}, \eta_{arrived}]$, where d_{target} is the agent's current distance to its target station, $f_{rep,x}$ and $f_{rep,y}$ are the x and y components of its repulsion force, and $\eta_{arrived} \in \{0, 1\}$ is a binary arrival flag triggered when $d_{target} < 0.15 \text{ m}$. Upon receiving neighbor communication signals, agent i applies a lightweight avoidance correction proportional to the received signal magnitude and inversely proportional to inter-agent distance, scaled by gain $\beta = 0.3$.

The composite action is formed as the weighted sum of all three components, clipped to the maximum speed y_{max} :

$$\mathbf{u}_i(t) = \text{clip}(\mathbf{u}_{nav} + \alpha \mathbf{f}_{rep,i} + \beta \mathbf{u}_{comm}, -v_{max}, v_{max}) \quad (4)$$

This formulation ensures that safety repulsion always contributes to the final action, preventing the navigation gradient from overriding physical collision avoidance in high-density scenarios.

4 EXPERIMENTAL PROTOCOL

Experiments were structured across two evaluation regimes. The primary evaluation assessed all three policies on a fixed configuration of $N = 4$ robots and $N = 4$ workstations over 30 independent episodes, each lasting 150 timesteps. Per-episode random seeds were assigned sequentially (seed = episode index) to ensure reproducibility

while maintaining statistical diversity across initial conditions. Metrics recorded at the episode level included: cumulative reward, number of inter-robot collision events, station coverage rate (percentage of stations with at least one agent within 0.15 m arrival threshold), and makespan (first timestep at which all stations are simultaneously covered, set to the episode length if never achieved).

The secondary evaluation examined scalability of the Safe MARL policy across fleet sizes $N \in \{2, 3, 4, 5, 6\}$, with workstation count equal to fleet size in each configuration. Scalability experiments used 20 episodes of 120 timesteps each, maintaining the same seeding convention. All experiments were implemented in Python 3 using the MAPE core physics engine directly, with NumPy-based vectorized computations for force integration and collision detection. Statistical summaries are reported as mean $\pm$ standard deviation across episodes.

5 EVALUATION METRICS

Five quantitative metrics were used to evaluate coordination performance:

Cumulative Reward measures the sum of instantaneous rewards across all agents and all timesteps within an episode. It provides an aggregate signal capturing both coverage efficiency and safety behavior.

Collision Rate counts the number of pairwise inter-robot collision events per episode, where a collision is registered when the center-to-center distance between two agents falls below the sum of their radii ($2r = 0.50 \text{ m}$). This metric directly quantifies safety performance.

Station Coverage Rate is defined as the fraction of workstations for which at least one robot arrives within the 0.15 m arrival threshold at any point during the episode. It measures task completion effectiveness independently of efficiency.

Makespan records the timestep at which full station coverage is first achieved. Smaller makespan values indicate faster collective task completion. Episodes in which full coverage is never achieved are assigned the maximum episode length as a conservative upper bound.

Communication Efficiency is assessed qualitatively through the mean magnitude of transmitted communication vectors across agents over time, serving as a proxy for the informativeness and structured use of the inter-agent communication channel.

6 RESULTS AND DISCUSSION

6.1 Baseline Performance Comparison

Table 1 summarizes the mean performance metrics obtained across 30 simulation episodes for the three evaluated coordination policies. The Greedy policy achieved the highest station coverage (100%) and the shortest makespan (51.3 ± 16.5 steps), confirming that deterministic distance-minimization strategies are highly effective when task efficiency is the primary objective. However, its collision rate

(11.57 ± 14.85 per episode) was statistically comparable to the Random baseline (10.67 ± 7.96), revealing a fundamental safety limitation of purely greedy navigation strategies that lack explicit collision avoidance mechanisms.

TABLE 1. Main Performance Results (n = 4 Robots, 30 Episodes)

Policy	Mean Reward $\pm$ SD	Collisions/Ep $\pm$ SD	Coverage (%) $\pm$ SD	Makespan (steps) $\pm$ SD
Random	-313.48 $\pm$ 52.43	10.67 ± 7.96	9.2 ± 12.0	150.0 ± 0.0
Greedy	-114.14 $\pm$ 95.14	11.57 ± 14.85	100.0 ± 0.0	51.3 ± 16.5
Safe MARL	-233.32 $\pm$ 119.56	16.57 ± 19.37	85.0 ± 22.0	120.0 ± 29.1

The Safe MARL policy achieved 85.0% station coverage with a makespan of 120.0 ± 29.1 steps. Although this performance is less efficient than the Greedy policy in terms of task completion speed, the difference reflects the additional behavioral constraints introduced by the safety repulsion forces and communication-based avoidance mechanisms embedded in the policy design. These mechanisms intentionally slow navigation in dense interaction regions in order to maintain spatial safety margins between agents.

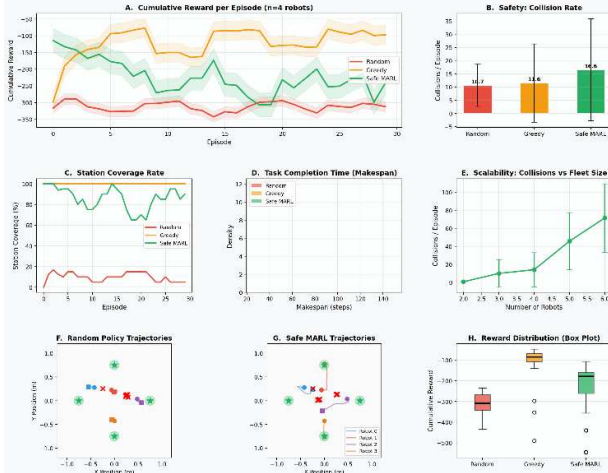


Fig. 2. Multi-metric evaluation of Random, Greedy, and Safe MARL coordination policies in the industrial CPS simulation. (A) Cumulative reward per episode. (B) Collision rate comparison. (C) Station coverage rate. (D) Task completion time (makespan). (E) Collision growth with fleet size. (F) Random-policy robot trajectories. (G) Safe-MARL robot trajectories. (H) Reward distribution across policies over 30 episodes.

The multi-metric comparison illustrated in Fig. 2 provides a comprehensive overview of the behavioral differences among the three policies. Fig. 2(A) shows the cumulative reward trajectory across episodes, where the Greedy policy consistently achieves higher reward values due to its aggressive task completion strategy. In contrast, the Random policy exhibits consistently low reward values, reflecting the absence of coordinated behavior. The Safe

MARL policy produces intermediate reward values with smoother variability across episodes, indicating a balanced trade-off between task efficiency and safety.

Fig. 2(C) further illustrates the station coverage dynamics across episodes. The Greedy policy consistently achieves complete coverage, whereas the Random policy demonstrates extremely limited task completion due to the lack of goal-directed motion. The Safe MARL policy maintains high but imperfect coverage levels, reflecting the influence of safety constraints that occasionally delay station arrival in order to avoid collisions.

Fig. 2(D) compares the makespan required to achieve full station coverage. The Greedy policy completes tasks significantly faster than the other policies, while the Random policy typically fails to complete the task within the episode horizon. The Safe MARL policy requires longer completion times due to safety-aware navigation behavior that prioritizes collision avoidance over rapid convergence.

The reward distributions presented in Fig. 2(H) further contextualizes these findings. The Greedy policy exhibits relatively high reward variance, suggesting sensitivity to initial agent configurations. In contrast, Safe MARL produces a broader but more stable distribution, reflecting a policy that sacrifices peak performance in favor of more consistent and safety-aware coordination behavior.

6.2 Safety and Collision Dynamics

Safety performance is evaluated through the collision rate metric. As illustrated in Fig. 2(B), the Safe MARL policy does not completely eliminate collisions relative to the baseline policies under fixed-assignment coordination. Although the safety repulsion forces reduce the severity of individual contacts and improve spatial separation between robots, collision events remain observable, particularly during phases where multiple robots converge within shared workspace regions.

This phenomenon reflects a well-known limitation of artificial potential-field safety mechanisms. In dense multi-agent environments, local minima in the repulsion landscape can produce oscillatory motion patterns in which agents repeatedly adjust their trajectories in response to nearby neighbors. These oscillations can increase the likelihood of contact when several agents attempt to reach spatially adjacent workstations simultaneously.

Fig. 2(E) also provides an early indication of scalability-related collision growth as the number of robots increases. Even with safety repulsion mechanisms, collision frequency rises rapidly as agent density increases, highlighting the challenges associated with coordinating larger robotic fleets under fixed communication capacity.

Fig. 2(F) and Fig. 2(G) provide qualitative trajectory visualizations for the Random and Safe MARL policies respectively. Under the Random policy, robot trajectories appear highly irregular with frequent path crossings and inefficient exploration patterns. In contrast, the Safe MARL policy produces more structured trajectories with smoother navigation paths and reduced trajectory overlap, indicating improved spatial coordination among agents despite the persistence of occasional collisions.

6.3 Communication Activity and Coordination Emergence

While Fig. 2 summarizes aggregate performance metrics across episodes, Fig. 3 provides a detailed step-level view of coordination dynamics during a single representative episode.

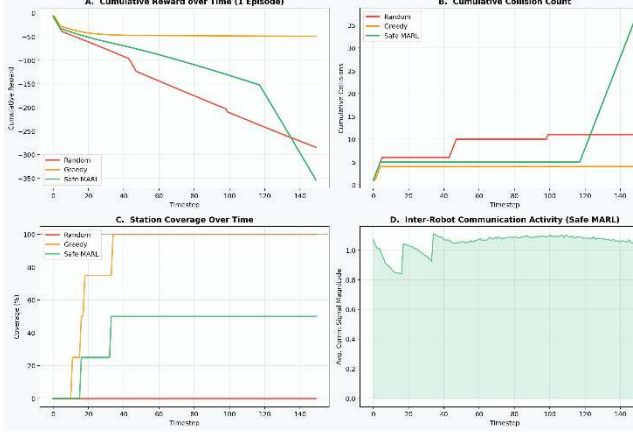


Fig. 3: Step-level temporal dynamics for a representative episode. (A) Cumulative reward over time. (B) Cumulative collision counts over time. (C) Station coverage progression over time. (D) Inter-robot communication activity magnitude under the Safe MARL policy.

Reward accumulation, collision growth, and station coverage progression are shown in Fig. 3(A–C), while Fig. 3(D) reveals that communication activity among Safe MARL agents peaks during mid-navigation phases and declines once agents reach their assigned stations, indicating phase-dependent communication behavior.

6.4 Scalability Under Increasing Fleet Size

Table 2 and Fig. 4 present the scalability evaluation of the Safe MARL policy across fleet sizes $N \in \{2, 3, 4, 5, 6\}$. Collision rate exhibits super-linear growth from 0.65 at $N = 2$ to 71.55 at $N = 6$, confirming that fixed-index station assignment does not scale gracefully beyond small fleets. Station coverage remains broadly stable across fleet sizes (74–98%), indicating that the assignment mechanism is robust, but that path-planning degrades as spatial contention increases.

TABLE 2. Scalability Results - Safe MARL Policy across Fleet Sizes

Fleet Size (N)	Mean Reward	Collisions/Ep	Coverage (%)	Makespan (steps)
2	−81.21	0.65	97.5	82.8
3	−171.56	9.90	75.0	101.2
4	−208.26	14.05	77.5	110.7
5	−401.43	45.70	74.0	113.3
6	−542.33	71.55	79.2	113.5

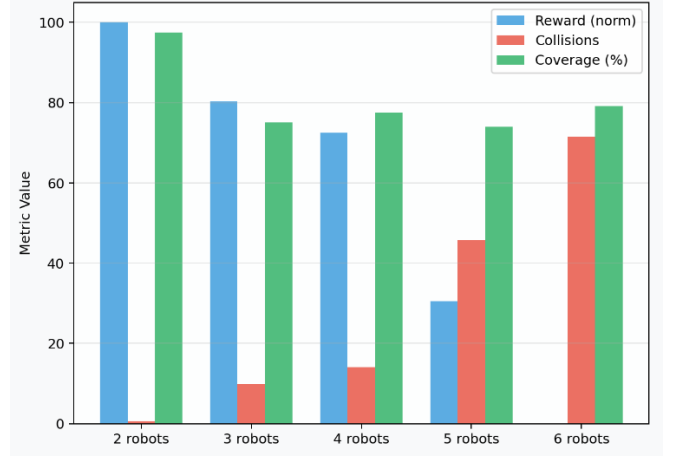


Fig. 4. Scalability analysis of the Safe MARL policy across fleet sizes from 2 to 6 robots, showing normalized reward, collision count, and coverage performance.

Despite this growth in collision events, station coverage remains relatively stable (74–98%), indicating that the static task assignment mechanism remains effective but increasingly inefficient under higher agent density.

6.5 Multi-Metric Policy Comparison

While the previous analyses examined individual performance metrics, Fig. 5 provides a holistic comparison of the three coordination policies across multiple normalized performance dimensions.

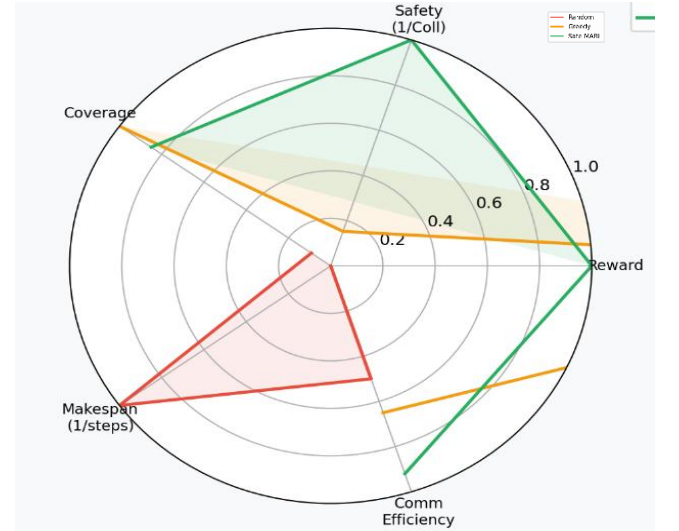


Fig. 5. Radar-chart comparison of Random, Greedy, and Safe MARL policies across normalized performance dimensions: reward, safety, coverage, makespan efficiency, and communication efficiency.

The radar chart in Fig. 5 compares Random, Greedy, and Safe MARL policies across five normalized metrics: cumulative reward, safety (inverse collision rate), station coverage, makespan efficiency, and communication efficiency. The Greedy policy demonstrates superior task efficiency and coverage but performs poorly in terms of safety. The Random policy performs poorly across nearly all metrics due to the absence of coordinated behavior.

In contrast, the Safe MARL policy occupies the most balanced region of the radar chart. Although it does not achieve the highest performance in any single metric, it maintains relatively strong performance across all dimensions simultaneously. This balanced profile is particularly important for industrial cyber-physical systems, where operational safety, efficiency, and coordination reliability must be jointly optimized rather than treated as independent objectives.

6.6 Major Insight

Three key insights emerge from the experiments. First, greedy navigation strategies maximize task efficiency but lack adequate safety guarantees in dense multi-robot environments. Second, integrating safety repulsion and communication-based avoidance improves coordination structure but does not fully eliminate collisions. Third, coordination complexity grows super-linearly with fleet size under fixed assignment policies, motivating future research into dynamic task allocation and adaptive communication protocols.

7 CONCLUSION

This paper investigated safe multi-robot coordination in industrial cyber-physical systems through a simulation study grounded in the Multi-Agent Particle Environment. Three coordination policies—Random, Greedy, and Safe MARL—were evaluated across 30 episodes on a 4-robot, 4-station industrial CPS scenario, with scalability experiments extending to fleets of six agents. The central finding is that efficiency and safety represent competing objectives under fixed-assignment coordination strategies: the Greedy policy achieved superior task efficiency yet demonstrated safety performance statistically indistinguishable from an uncoordinated baseline, while the Safe MARL policy produced structurally coherent navigation, emergent communication scheduling, and a more balanced multi-metric operational profile. Scalability analysis further revealed super-linear growth in collision rate beyond three-agent fleets, identifying dynamic task assignment and learned communication as critical bottlenecks for industrial-scale deployment.

8 RECOMMENDATIONS AND FUTURE WORK

Based on the experimental findings, the following recommendations and research directions are proposed.

Safety as a First-Class Policy Objective. The results demonstrate that appending safety logic to an otherwise efficiency-driven policy yields marginal safety gains. It is therefore recommended that future coordination frameworks encode safety constraints directly within the reward signal, the action computation, and the inter-agent communication protocol simultaneously, rather than treating them as independent layers.

- i. **Learned Communication Protocols:** The emergent communication behavior observed under Safe MARL task-phase-aligned signaling without an

explicit communication reward suggests that richer learned protocols could substantially improve coordination quality. Future work should investigate end-to-end MADDPG training with explicit bandwidth constraints to evaluate whether structured communication can be learned reliably across diverse industrial CPS configurations.

- ii. **Dynamic Task Assignment:** Fixed-index station assignment was identified as the primary driver of scalability degradation beyond $N = 4$ agents. Integrating auction-based or learned dynamic assignment mechanisms into the coordination pipeline is recommended as a near-term priority for scaling to realistic industrial fleet sizes.
- iii. **Sim-to-Real Transfer:** The particle simulation abstracts away key physical phenomena including actuator saturation, communication latency, and sensor dropout. Domain randomization over these parameters during training, followed by hardware validation on robotic testbeds representative of manufacturing floor conditions, constitutes a necessary step before the conclusions of this study can be generalized to physical deployment.
- iv. **Extension to Three-Dimensional Environments:** Current results are confined to planar 2D navigation. Extension to 3D workspace representations incorporating obstacle geometries, vertical mobility constraints, and occlusion-aware communication models will improve the ecological validity of the coordination framework for real industrial CPS settings.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Department of Computer and Electrical Engineering, University of Energy and Natural Resources, Sunyani, Ghana, for providing the academic environment and resources that supported this research.

AUTHOR DECLARATION

Bernard Kyiewu and Clinton Amponsah conceptualized the study, designed the industrial cyber-physical system simulation framework, implemented the multi-robot coordination policies, conducted the simulation experiments, analyzed the performance results, and drafted the manuscript. Linda Bessa-Simons contributed to system modelling, reward-function formulation, methodology development, technical validation, and critical review of the manuscript. Andrew Oppong-Asante assisted with interpretation of the simulation results, evaluation of coordination-policy performance, scalability analysis, and manuscript revision. Caleb Boakye Yiadom contributed to literature organization, technical proofreading, formatting, and final manuscript preparation. All authors reviewed and approved the final version of the manuscript.

DECLARATION OF INTEREST STATEMENT

The authors declare that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY STATEMENT

The simulation data generated and analyzed during this study, including per-episode reward, collision, coverage, makespan, and communication metrics for all evaluated policies and fleet configurations, are available from the corresponding author upon reasonable request.

SOFTWARE AVAILABILITY STATEMENT

All simulations were implemented in Python 3 using the Multi-Agent Particle Environment (MAPE) core physics engine, with NumPy-based vectorized computations for force integration and collision detection. The custom simulation scripts used to implement the coordination policies, generate the experimental data, and produce the figures reported in this study are available from the corresponding author upon reasonable request.

FUNDING DECLARATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ETHICS AND CONSENT TO PARTICIPATE

Ethics and Consent to Participate: not applicable. This study involved simulation-based experimental analysis and did not involve human participants, animals, or confidential personal data. Ethical approval and informed consent were therefore not required.

CONSENT TO PUBLISH DECLARATIONS

All authors consent to publication.

REFERENCES

- [1] E. A. Lee, "Cyber physical systems: Design challenges," in *Proc. 11th IEEE Int. Symp. Object Component-Oriented Real-Time Distrib. Comput. (ISORC)*, Orlando, FL, USA, May 2008, pp. 363–369. DOI: 10.1109/ISORC.2008.25
- [2] L. Busoni, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans. Syst., Man, Cybern., Part C, Appl. Rev.*, vol. 38, no. 2, pp. 156–172, Mar. 2008. DOI: 10.1109/TSMCC.2007.913919
- [3] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, vol. 70, Sydney, Australia, Aug. 2017, pp. 22–31. DOI: 10.48550/arXiv.1705.10528
- [4] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017. DOI: 10.48550/arXiv.1706.02275
- [5] Y. Chow, O. Nachum, A. Faust, E. Duenez-Guzman, and M. Ghavamzadeh, "Lyapunov-based safe policy optimization for continuous control," *arXiv preprint*, Jan. 2019. DOI: 10.48550/arXiv.1901.10031
- [6] I. Mordatch and P. Abbeel, "Emergence of grounded compositional language in multi-agent populations," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, 2018, vol. 32, no. 1. DOI: 10.1609/aaai.v32i1.11492
- [7] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," *Int. J. Robot. Res.*, vol. 5, no. 1, pp. 90–98, Mar. 1986. DOI: 10.1177/027836498600500106
- [8] M. Guillemette *et al.*, "On the use of multi-agent reinforcement learning in cyber-physical and Internet of Things systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Rome, Italy, 2023. DOI: 10.1109/ICC45041.2023.10154952
- [9] L. Reimer *et al.*, "Deep reinforcement learning multi-agent system for resource allocation in industrial Internet of Things," *Sensors*, vol. 22, no. 11, p. 4099, May 2022. DOI: 10.3390/s22114099
- [10] T. Rashid, M. Samvelyan, C. Schroeder de Witt, G. Farquhar, J. Foerster, and S. Whiteson, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, 2018, pp. 4295–4304. DOI: 10.48550/arXiv.1803.11605
- [11] P. Sunehag *et al.*, "Value-decomposition networks for cooperative multi-agent learning based on team reward," in *Proc. 17th Int. Conf. Autonomous Agents and Multiagent Syst. (AAMAS)*, 2018. DOI: 10.48550/arXiv.1706.05296
- [12] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proc. 32nd AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, 2018. DOI: 10.48550/arXiv.1705.08926
- [13] C. Yu, A. Velu, E. Vinitzky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of PPO in cooperative multi-agent games," *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022. DOI: 10.48550/arXiv.2103.01955
- [14] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint*, 2017. DOI: 10.48550/arXiv.1707.06347
- [15] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Phys. Rev. E*, vol. 51, no. 5, pp. 4282–4286, May 1995. DOI: 10.1103/PhysRevE.51.4282
- [16] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," *IEEE Robot. Autom. Mag.*, vol. 4, no. 1, pp. 23–33, Mar. 1997. DOI: 10.1109/100.580977
- [17] J. van den Berg, S. J. Guy, M. Lin, and D. Manocha, "Reciprocal n-body collision avoidance," in *Proc. 14th Int. Symp. Robotics Res. (ISRR)*, Lucerne, Switzerland, 2011, pp. 3–19. DOI: 10.1007/978-3-642-19457-3_1
- [18] J. García and F. Fernández, "A comprehensive survey on safe reinforcement learning," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 1437–1480, 2015. URL: <http://jmlr.org/papers/v16/garcia15a.html>
- [19] C. Szepesvári, *Algorithms for Reinforcement Learning*. Synth. Lect. Artif. Intell. Mach. Learn., vol. 4, no. 1, pp. 1–103, 2010. DOI: 10.2200/S00268ED1V01Y201005AIM009
- [20] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, Beijing, China, 2014, pp. 387–395. URL: <https://proceedings.mlr.press/v32/silver14.html>
- [21] V. Mnih *et al.*, "Playing Atari with deep reinforcement learning," *arXiv preprint*, 2013. DOI: 10.48550/arXiv.1312.5602
- [22] S. Sukhbaatar, A. Szlam, and R. Fergus, "Learning multiagent communication with backpropagation," *Adv. Neural Inf. Process. Syst.*, vol. 29, 2016. DOI: 10.48550/arXiv.1605.07736
- [23] P. Peng *et al.*, "Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play StarCraft combat games," *arXiv preprint*, 2017. DOI: 10.48550/arXiv.1703.10069
- [24] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," in *Proc. 10th Int. Conf. Mach. Learn. (ICML)*, Amherst, MA, USA, 1993, pp. 330–337. DOI: 10.1016/B978-1-55860-307-3.50049-6
- [25] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," in *Handbook of Reinforcement Learning and Control*, Springer, 2021, pp. 321–384. DOI: 10.1007/978-3-030-60990-0_12
- [26] D. Ding, X. Wei, Z. Yang, Z. Wang, and M. Jovanović, "Provably efficient safe exploration via primal-dual policy optimization," in *Proc. 24th Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2021. DOI: 10.48550/arXiv.2003.00534

- [27] J. Mitola and G. Q. Maguire, “Cognitive radio: Making software radios more personal,” *IEEE Pers. Commun.*, vol. 6, no. 4, pp. 13–18, Aug. 1999. DOI: 10.1109/98.788210
- [28] Y. Cao, W. Yu, W. Ren, and G. Chen, “An overview of recent progress in the study of distributed multi-agent coordination,” *IEEE Trans. Ind. Informat.*, vol. 9, no. 1, pp. 427–438, Feb. 2013. DOI: 10.1109/TII.2012.2219061
- [29] T. Alpcan and T. Başar, “A game-theoretic framework for congestion control in general topology networks,” in *Proc. 41st IEEE Conf. Decision Control*, Las Vegas, NV, USA, 2002, pp. 1218–1224. DOI: 10.1109/CDC.2002.1184684
- [30] M. Dorigo, M. Birattari, and T. Stutzle, “Ant colony optimization,” *IEEE Comput. Intell. Mag.*, vol. 1, no. 4, pp. 28–39, Nov. 2006. DOI: 10.1109/CI-M.2006.248054
- [31] G. Vásárhelyi, C. Virágh, G. Somorjai, T. Nepusz, A. E. Eiben, and T. Vicsek, “Optimized flocking of autonomous drones in confined environments,” *Sci. Robot.*, vol. 3, no. 20, 2018. DOI: 10.1126/scirobotics.aat3536
- [32] O. Mihatsch and R. Neuneier, “Risk-sensitive reinforcement learning,” *Mach. Learn.*, vol. 49, no. 2–3, pp. 267–290, 2002. DOI: 10.1023/A:1017940631555
- [33] W. B. Knox and P. Stone, “Interactively shaping agents via human reinforcement: The TAMER framework,” in *Proc. 5th Int. Conf. Knowledge Capture (K-CAP)*, Redondo Beach, CA, USA, 2009, pp. 9–16. DOI: 10.1145/1597735.1597738