A Comparative Study of Independent PPO and MAPPO in a Simplified 3D Multi-Agent Air-Combat Environment

1. Project Overview

Multi-agent reinforcement learning (MARL) has demonstrated promise in complex coordination tasks, including air-combat simulations. However, many existing studies emphasize large-scale, high-fidelity environments aimed at demonstrating capability rather than isolating and understanding learning dynamics. This project aims to study coordination and training stability in MARL through a simplified three-dimensional air-combat environment designed for controlled experimentation.

Specifically, this work compares Independent Proximal Policy Optimization (IPPO) and Multi-Agent PPO (MAPPO) under centralized training and decentralized execution (CTDE). The environment models aircraft motion using simplified 3D kinematics with yaw and pitch control only, intentionally excluding roll and detailed aerodynamics to limit complexity. This design allows the project to focus on the effects of centralized critics on learning stability, coordination behavior, and sample efficiency in continuous multi-agent control tasks.

2. Research Question

Primary Question

Does the use of a centralized critic (MAPPO) improve training stability, coordination, and sample efficiency compared to Independent PPO in a cooperative multi-agent 3D air-combat task?

Secondary Questions (Exploratory)

- How does coordination emerge over training in IPPO versus MAPPO?
- How sensitive are the two approaches to modest increases in environment complexity (e.g., additional spatial degrees of freedom or number of agents)?

3. Scope and Assumptions

In Scope

- Simplified 3D air-combat environment
- Cooperative multi-agent setting
- Continuous state space
- Discrete or low-dimensional continuous action space
- Aircraft motion modeled via kinematics with yaw and pitch control
- Comparison between IPPO and MAPPO
- Centralized training with decentralized execution
- Reproducible experiments and analysis

Out of Scope

- Realistic flight dynamics, lift/drag modeling, or aerodynamics
- Roll dynamics or energy management
- Radar, sensor fusion, or communication modeling
- Adversarial multi-agent learning (initially)
- Large-scale battles (>3 agents)
- Learning explicit communication protocols
- Performance comparisons with real-world or DARPA-scale simulators

4. Environment Definition

Agents

- Two cooperative “friendly” fighter agents
- One scripted or static enemy target (initially)

State Space (per agent)

- Own position: (x, y, z)
- Own orientation: (yaw, pitch)
- (Optional, fixed or bounded) speed
- Relative position to enemy
- Relative position to teammate

All observations are local to each agent.

Dynamics

- Aircraft motion is modeled using simple 3D kinematics
- Agents move forward along their heading direction determined by yaw and pitch
- Roll dynamics are excluded
- Speed is initially fixed to reduce complexity (variable speed may be explored later)

Action Space

The initial action space is discrete to reduce complexity:

- Yaw left
- Yaw right
- Pitch up
- Pitch down

- Accelerate / decelerate (optional, may be disabled initially)
- Fire (optional, added after baseline learning is achieved)

---

5. Algorithms

The following algorithms will be implemented and compared:

Independent PPO (IPPO)

- Each agent learns independently
- No centralized critic
- Shared reward signal

Multi-Agent PPO (MAPPO)

- Shared policy parameters
- Centralized value function during training
- Decentralized execution

Both algorithms will use:

- Identical network architectures
- Identical reward functions
- Identical environment dynamics
- Identical training budgets

---

6. Metrics and Evaluation

Core Metrics

- **Average episode reward** (mean over evaluation episodes)
- **Training stability**
  - Variance of reward across multiple random seeds
- **Sample efficiency**
  - Number of environment steps required to reach a predefined reward threshold

Coordination Metrics

At least one of the following will be used:

- Average distance between cooperative agents
- Time-to-engagement with enemy
- Rate of simultaneous engagement by both agents
- Collision or spatial overlap rate between agents

---

7. Experimental Protocol

- Number of random seeds: 3–5
- Fixed maximum training steps per experiment
- Periodic evaluation using deterministic policies
- Same hyperparameters across IPPO and MAPPO
- All experiments logged for reproducibility

---

8. Expected Outcomes

It is expected that MAPPO will demonstrate:

- More stable learning dynamics
- Reduced variance across training runs
- Improved coordination metrics compared to IPPO

Deviations from these expectations will be analyzed and reported as valid and informative results.

---

9. Deliverables

The project will be considered complete when the following are achieved:

- A functional simplified 3D multi-agent air-combat environment
- Successful training of IPPO agents
- Successful training of MAPPO agents
- Logged training metrics and plots
- Comparative analysis between IPPO and MAPPO
- Clear documentation and README

---

10. Future Work

Potential future extensions include:

- Learning explicit inter-agent communication
- Introducing information bottlenecks

- Adversarial multi-agent settings
- Increased number of agents
- Variable-speed dynamics
- Distributed training infrastructure