**Assessment** (Between Initial Experiment and LvL 1 - Experiment 1)

Training Analysis Summary: MARL PPO for Tactical Environment

Training Session 1

Configuration

Environment: Full tactical environment with multiple enemies

Episodes: 1000

Max Steps: 1000 per episode

Hyperparameters:

Learning rate: 3e-4

Entropy coefficient: 0.01

Value coefficient: 0.5

PPO clip: 0.2

Performance

Rewards: Initially around -142.86, improved to -71.43 by episode ~600, then plateaued

Steps: Consistently reached max (1000) steps per episode

Casualties: Zero friendly and enemy casualties throughout

Actor Losses: Very small (~10^-3), indicating minimal policy updates

Critic Losses: Inconsistent across agents - only agents 2 and 6 showed significant losses

Entropy Losses: Consistently zero, suggesting deterministic policies with no exploration

Key Insights

Agents learned basic navigation/positioning (evidenced by reward improvement)

Failed to engage in combat or complete objectives

Reached a learning plateau after ~600 episodes

Zero entropy loss indicated premature convergence to deterministic policies

Limited exploration prevented discovery of effective combat tactics

Training Session 2

Configuration

Environment: Simplified curriculum with bare terrain and single enemy near objective

Episodes: 1000

Max Steps: 1000 per episode

Hyperparameters (adjusted):

Learning rate: Increased to 5e-4

Entropy coefficient: Significantly increased to 0.05-0.1

Value coefficient: Increased to 1.0

PPO clip: Increased to 0.3

Performance

Rewards: Consistently around -0.50

Steps: Still reaching max (1000) steps per episode

Casualties: Zero friendly and enemy casualties

Actor Losses: Higher (~-0.0055), indicating more meaningful policy updates

Critic Losses: More reasonable (0.0082 average)

Entropy Losses: Substantially higher (-4.51 average), indicating proper exploration

Key Insights

Successfully increased exploration through entropy coefficient adjustment

Agents actively exploring but not yet discovering successful strategies

More meaningful policy updates occurring (higher actor losses)

Better value estimation developing (more reasonable critic losses)

Learning better exploration but not yet connecting to successful outcomes

Comparative Analysis

The adjustments between sessions produced significant improvements in training dynamics:

Exploration: Dramatic improvement in exploration behavior from zero entropy loss to substantial negative entropy loss

Learning Progress: Session 1 showed plateauing with zero policy improvement, while Session 2 shows ongoing learning with meaningful updates

Policy Development: Session 1 converged prematurely to suboptimal policies, while Session 2 maintains exploration of policy space

Efficiency: Despite similar outcomes (no objectives completed, no casualties), Session 2 shows healthier learning dynamics that may lead to breakthrough performance with continued training

Recommendations for Future Sessions

Curriculum Progression:

Continue with current simplified environment until performance improves

Gradually introduce terrain features after navigation is mastered

Incrementally increase enemy count and sophistication

Reward Engineering:

Implement intermediate rewards for approaching objectives

Add small rewards for enemy detection

Consider formation-based rewards for coordinated movement

Hyperparameter Refinement:

Maintain high entropy coefficient to preserve exploration

Consider reducing learning rate if training becomes unstable

Experiment with different GAE-lambda values (0.9-0.97) for advantage estimation

Technical Improvements:

Implement early stopping when performance plateaus

Add automatic hyperparameter adjustment based on performance metrics

Develop better visualization tools to analyze agent behavior patterns

The shift from Session 1 to Session 2 represents a significant improvement in training methodology, moving from a complex environment with poor exploration to a simplified environment with robust exploration. While agents haven't yet learned successful tactics, the fundamentals for effective learning are now in place.

**Assessment** (Between LvL 1 - Experiment 1 and LvL 1 - Experiment 2)

BL – Observation Space Simplification

**Assessment** (Between LvL 1 - Experiment 2 and LvL 1 - Experiment 3)

BL – Action Space Reduction & Increased Steps   
  
Original State-Action Space Size

State Space: 100×100 grid positions × various unit states ≈ 10,000 base positions

Action Space Calculation:

MOVE: 25 directions × 21 distances = 525 combinations

ENGAGE: 10,000 positions × 30 rounds × 2 suppress options × 2 fire rate options = 1,200,000 combinations

SUPPRESS: 10,000 positions × 30 rounds × 2 suppress options × 2 fire rate options = 1,200,000 combinations

BOUND: 25 directions × 21 distances = 525 combinations

CHANGE\_FORMATION: 8 formations × 360 orientations = 2,880 combinations

Total action combinations: 2,403,930 distinct actions

State-Action Pairs: 10,000 states × 2,403,930 actions ≈ 24 billion pairs

Constrained Action Space Breakdown

1. Action Types (unchanged)

Same 5 options: MOVE, ENGAGE, SUPPRESS, BOUND, CHANGE\_FORMATION

2. Movement Parameters (constrained)

Direction: Unchanged continuous 2D vector (-1 to 1)

Still approximately 25 meaningful directions

Distance: Integer between 0 and 10 cells (reduced from 0-20)

11 possible distance values (reduced from 21)

3. Engagement Parameters (heavily constrained)

Target Position: Only visible enemy positions (typically 5-10 at most)

≈ 8 possible target positions on average (reduced from 10,000)

Max Rounds: Unchanged, 1-30

30 possible values

Suppress Only: Unchanged, binary

2 possible values

Adjust for Fire Rate: Unchanged, binary

2 possible values

Additional Constraint: Can only execute these actions when enemies are visible

4. Formation Parameters (constrained)

Formation Type: Unchanged, 8 possible formations

Orientation: Constrained to 8 cardinal/ordinal directions

8 possible orientations (reduced from 360)

Constrained State-Action Space Size

State Space: Unchanged at approximately 10,000 base positions

Action Space Calculation:

MOVE: 25 directions × 11 distances = 275 combinations

ENGAGE: 8 enemy positions × 30 rounds × 2 suppress options × 2 fire rate options = 960 combinations

SUPPRESS: 8 enemy positions × 30 rounds × 2 suppress options × 2 fire rate options = 960 combinations

BOUND: 25 directions × 11 distances = 275 combinations

CHANGE\_FORMATION: 8 formations × 8 orientations = 64 combinations

Total action combinations: 2,534 distinct actions (reduced from 2,403,930)

State-Action Pairs: 10,000 states × 2,534 actions ≈ 25.3 million pairs (reduced from 24 billion)

Summary of Reductions

Movement Actions: Reduced possible distances from 0-20 to 0-10

50% reduction in movement distance options

Engagement Actions:

Limited target positions to only visible enemies

Prevented engagement actions when no enemies are visible

Combined result: ~99.9% reduction in engagement target space

Formation Actions:

Limited orientation to 8 cardinal directions instead of 360 degrees

97.8% reduction in orientation options

Overall Impact:

~99.9% reduction in total action space

State-action pairs reduced from ~24 billion to ~25.3 million

Makes the problem tractable for reinforcement learning algorithms

These constraints maintain the tactical richness of the simulation while dramatically improving learning efficiency.  
  
Optimal Steps Per Episode: 3,000

Reasoning:

Traversal Efficiency: With the maximum movement distance of 10 cells per step, an agent could theoretically cross the entire 100×100 map in about 10-20 steps in a straight line. However, tactical movement is rarely direct.

Action Exploration: With approximately 2,534 distinct actions, agents need sufficient steps to explore the action space effectively. A common heuristic suggests at least 10-100× the action space size for meaningful exploration within an episode.

Mission Complexity: Your environment simulates tactical military operations which involve:

Movement to objectives

Engagement with enemies

Formation changes

Coordination between units

Each of these phases requires numerous steps to execute properly.

Learning from Consequences: Tactical decisions have long-term consequences that may only be apparent after hundreds of steps.

Time-to-Completion Analysis:

Position-to-objective time: ~100-300 steps

Multiple engagements: ~300-600 steps per engagement

Tactical repositioning: ~100-200 steps per maneuver

Mission completion: ~1,000-3,000 steps

3,000 steps provides a good balance between:

Allowing missions to reach completion

Giving agents time to learn from their actions

Avoiding unnecessarily long episodes that waste computational resources

If you observe that agents are consistently completing missions in fewer steps, you could reduce this number. Conversely, if missions remain incomplete, consider increasing it. Start with 3,000 and adjust based on empirical results from your training runs.

**Assessment** (Between LvL 1 - Experiment 3 and LvL 1 - Experiment 4)

BL – Reward Shaping (Risk Adverse to Engagement & Redesign of Individual and Team)  
  
**Initial Reward Structure Analysis**

**Original Values and Components**

The initial reward structure was designed with a hybrid approach combining team and individual rewards:

**Team Rewards:**

* Enemy elimination: +10.0 per casualty
* Friendly casualty penalty: -20.0 per casualty
* Objective proximity: Graduated bonuses from +1.0 to +10.0 based on proximity
* Distance improvement: Up to +10.0 for reducing average distance to objective

**Individual Rewards:**

* Movement toward objective: Up to +1.0 based on distance improvement
* Engagement with visible enemies: +0.5 basic reward
* Ammunition efficiency: Up to +0.5 additional for effective ammo usage

**Intended Learned Behaviors**

This structure was designed to encourage:

1. Force preservation (high penalty for casualties)
2. Efficient movement toward objectives
3. Cautious but effective engagement with enemies
4. Proper ammunition usage

**Observed Results from Training**

Based on the training data provided, the system achieved:

* Consistent positive rewards (476.14 in episode 59, increasing to 599.71 by episode 85)
* Perfect force preservation (0 friendly casualties)
* No enemy eliminations (0 enemy casualties)
* Use of all available time steps (3000/3000 steps)

**Learning Indicators:**

* Near-zero actor losses (-0.0011) suggesting policy convergence
* Moderate critic losses (2.8952) showing continued value function refinement
* High negative entropy losses (-4.8779) indicating deterministic policies

**Tactical Behavior:** The agents appear to have learned a highly conservative, objective-focused strategy that completely avoids combat engagement. They likely generate rewards primarily through movement toward objectives and proximity bonuses, while carefully avoiding situations that might lead to casualties.

This behavior is rational given the reward structure, where the penalty for friendly casualties (-20) is double the reward for enemy elimination (+10), creating a strong incentive for risk avoidance.

**Revised Reward Structure Analysis**

**Modified Values and Components**

**Team Rewards:**

* Enemy elimination: Increased to +20.0 per casualty (balanced with casualty penalty)
* Friendly casualty penalty: Maintained at -20.0 per casualty
* Team-based objective concentration: Scaled rewards based on number of agents at key distances:
  + Base rewards: 0.5-3.0 per agent at various distances
  + Force concentration bonuses: +5.0, +10.0, +20.0 for significant force at key distances

**Individual Rewards:**

* Individual proximity bonus: Moved from team to individual rewards (0.3-3.0 based on distance)
* Movement toward objective: Maintained (up to +1.0)
* Enemy elimination contribution: Added significant +15.0 reward for contributing to eliminations
* Engagement with visible enemies: Maintained (+0.5 basic, +0.5 for effective ammo usage)

**Intended Behavioral Changes**

This revised structure aims to:

1. Encourage more aggressive engagement by balancing casualty vs. elimination values
2. Promote force concentration at objectives through team-based bonuses
3. Reward individual initiative in approaching objectives
4. Strongly incentivize active participation in enemy elimination
5. Discourage "free-riding" where passive agents benefit from others' work

**Expected Outcomes**

We anticipate the revised reward structure will lead to:

1. More tactical engagements with enemy forces
2. Better coordination in force concentration at objectives
3. Maintaining force preservation as a priority while balancing with mission objectives
4. Greater initiative from individual agents
5. More decisive action in eliminating threats

**Research Support for MARL Reward Shaping**

Several peer-reviewed papers support the approach to reward shaping in multi-agent reinforcement learning:

1. **"Counterfactual Multi-Agent Policy Gradients" (Foerster et al., 2018)**
   * Demonstrated the importance of credit assignment in cooperative MARL
   * Showed how individual contribution rewards improve coordination
   * Published in AAAI 2018
2. **"QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning" (Rashid et al., 2018)**
   * Illustrated how combining team and individual value functions improves performance
   * Supports our hybrid reward approach
   * Published in ICML 2018
3. **"Intrinsic Motivation for Encouraging Synergistic Behavior" (Jaques et al., 2019)**
   * Showed benefits of rewarding coordinated actions between agents
   * Aligns with our force concentration rewards
   * Published in NeurIPS 2019
4. **"Learning to Coordinate with Coordination Graphs in Repeated Single-Stage Multi-Agent Decision Problems" (Bargiacchi et al., 2018)**
   * Demonstrated effectiveness of rewarding agents for group achievements
   * Supports our team-based reward for force concentration
   * Published in ICML 2018
5. **"Multi-Agent Reinforcement Learning in Sequential Social Dilemmas" (Leibo et al., 2017)**
   * Showed how reward structures influence cooperative vs. competitive behaviors
   * Relevant to balancing individual initiative with team objectives
   * Published in AAMAS 2017

These papers collectively support our approach of balancing team and individual rewards, providing appropriate credit assignment, and explicitly rewarding coordinated behaviors that align with tactical objectives.

The modifications we've made to the reward structure are well-supported by these research findings and should help create more tactically sound and coordinated behaviors among the learning agents.