Deep Reinforcement Learning within Military Course of Action (COA) Analysis

Stephen Moore [1] **Nicholas Tidwell** [1]

Abstract

This paper presents a novel approach to AI-assisted military planning, leveraging reinforcement learning (RL) in a customizable simulation environment to optimize friendly Courses of Action (COAs) in diverse tactical scenarios. By integrating doctrinal knowledge into the state space and employing curriculum learning, the system generates sophisticated decision-making behaviors, including adaptive force preservation, terrain-aware navigation, and objective prioritization. Experimental results demonstrate high success rates across multiple test maps, with the agent outperforming human experts in several metrics. Future work will focus on refining doctrinal models, developing dynamic adversary strategies, and expanding the range of agent types to further enhance operational capabilities and reduce reliance on personnel experience biases in military planning. This work is based on methods and models implemented in the repository Source Code.

1. Introduction

Within the planning for military operations, course of action (COA) analysis, also known as war-gaming, is a critical step for military planners because it identifies vulnerabilities and critical decisions within each COA. This process is often executed in an analog medium in which the planners form two sides, one for the friendly forces and the other for the enemy, and then conduct a turn-based simulation of each COA on a depiction of the area of operations. Through each turn of the COA, one side states their action and then waits for the opposing side to state their counteraction. Each turn's action and counteraction are recorded and adjudicated by a designated planner using a correlation of forces and means (COFM) calculator that predicts casualty and consumption rates given the type of operation. Although it is a critical step, the effectiveness of this analysis is heavily dependent on the experience and expertise of the planners involved, particularly in representing opposing forces.

The challenges in COA analysis are further compounded by the anticipated nature of future conflicts, specifically large-scale combat operations (LSCO). Unlike recent counter-insurgency operations, LSCO is expected to result in higher casualty rates and more frequent destruction of command posts. This scenario presents a significant risk of inexperienced personnel being thrust into critical planning roles, potentially leading to poorly devised operations.

In response to these challenges, we propose a novel approach that leverages advanced machine learning techniques to generate doctrinally sound tactical plans with greater speed and consistency than traditional methods. Building upon the previously developed tools, our framework introduces two key innovations.

Firstly, we present a custom built simulation environment developed using a gymnasium framework. This tool models realistic military scenarios and allows planners to modify terrain, unit composition, and tactics to match the specific context of their operations. This capability for customization distinguishes our approach from previous work in this domain.

Secondly, by employing the Proximal Policy Optimization algorithm, our system automates tactical decision making within the simulation. This provides mathematically grounded insights that help counterbalance any lack of experience or potential bias in traditional COA analysis methods.

The remainder of this paper details the architecture of our proposed system, evaluates its performance under simulated large scale scenarios, and discusses its potential impact on future military planning.

2. Related Work

Several approaches have been developed to enhance the COA development and analysis process through artificial intelligence. However, each has limitations that our approach

[1]Department of Computer Engineering, University of Central Florida, Orlando, USA. Correspondence to: Stephen Moore <st690445@ucf.edu>, Nicholas Tidwell <ni823771@ucf.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

seeks to address.

2.1. AI-Enabled Wargaming in the Military Decision Making Process

During the COA Analysis phase of the Military Decision Making Process (MDMP), (Schwartz et al., 2020) propose an AI-enabled wargaming prototype that uses a genetic algorithm to iteratively improve a given friendly COA. Their architecture separates the AI, simulation (DAVINCI), and user interface components, allowing for flexible experimentation and future enhancement.

While this approach represents a significant step forward, it heavily relies on user input for the initial COA development, which doesn't fully address the personnel training and experience bias mentioned in our introduction. As Schwartz et al. note, "the user must manually define a COA for the enemy forces and an initial COA for the friendly forces," which means the quality of the final COA remains dependent on the expertise of the planners involved in developing the initial COA. Additionally, the genetic algorithm's search space is significantly constrained to make the search manageable, limiting the potential for discovering truly optimal solutions. All submissions must follow the specified format.

2.2. COA-GPT: Generative Pre-trained Transformers for COA Development

Large Language Models (LLMs) have also been applied to the COA development process. (Goecks & Waytowich, 2024) introduce COA-GPT, a generative framework that incorporates military doctrine and domain knowledge through in-context learning. The system allows commanders to input mission information via text and images, generating doctrinally aligned COAs in seconds.

While COA-GPT demonstrates remarkable speed in generating initial COAs, it relies heavily on human iteration with the LLM to develop the final COA, which means the same personnel training and experience bias persists. Furthermore, COA-GPT uses a wrapper for the StarCraft II game environment, which lacks the customization capabilities needed for military planners to tailor the simulation to their specific area of operations. As the authors note, "COA-GPT relies heavily on pre-configured scenarios with limited flexibility in force composition and terrain modeling," creating significant challenges when attempting to adapt the system to different geographic theaters with unique terrain features and force capabilities.

2.3. Using AI in Wargaming Simulation as a Multi-Domain Decision Support Tool

To explore the use of reinforcement learning in adversarial planning, (Chance et al.) present Red Force Response (RFR), a decision support tool that leverages Deep Neural Networks, specifically PPO and Curiosity Learning, in a Multi-Agent Reinforcement Learning setting. This system is designed to identify both high-performing and novel Red Force Courses of Action (COAs) within a wargaming simulator.

This approach comes closest to our project, demonstrating high win rates (91% in tactical air scenarios) and employing similar RL algorithms. However, it takes on a "red team" approach by using trained adversary agents to identify weaknesses in friendly COAs, rather than generating optimized friendly COAs directly. Additionally, while their system shows promise, it doesn't address the full spectrum of customization needed for military planners across varied operational environments and doesn't fully leverage doctrinal heuristics to reduce the dimensionality of the state space.

Our approach seeks to address these gaps by creating a fully customizable environment specifically designed for military planning, incorporating doctrinal knowledge directly into the state space representation, and using reinforcement learning to generate optimized friendly COAs without requiring extensive human expertise in the initial planning stages.

3. Custom Environment

3.1. Leveraging Gymnasium for Military Planning

Our approach centers on creating a highly customizable simulation environment using OpenAI's Gymnasium, a toolkit for developing and comparing reinforcement learning algorithms. This foundation provides us with a flexible framework to model complex military scenarios while allowing military planners to tailor the environment to their specific operational requirements.

The environment is designed to faithfully represent military operations by incorporating key elements of doctrine, terrain effects, unit capabilities, and combat dynamics. Unlike commercial game engines that prioritize entertainment value over realism, our environment focuses on tactical fidelity and doctrinal accuracy while maintaining computational efficiency.

3.2. Observation Space

The observation space in our environment is carefully designed to balance tactical realism with reinforcement learning requirements. We implement a partially observable model that mirrors the information constraints faced by military commanders in real-world operations. This approach creates a hierarchical information structure that processes tactical data in layers similar to how human commanders

assess battlefield situations.

At the foundation of this structure lies agent state information, which includes position coordinates that pinpoint the unit's location on the battlefield, health status represented as a percentage to indicate combat effectiveness, ammunition levels tracked quantitatively to model resource management, and suppression status that reflects whether the unit is under effective enemy fire. This core information provides the basic situational awareness that any commander would prioritize. Building upon this foundation, tactical information enriches the observation space with formation configuration that determines the unit's battlefield geometry, orientation angles that establish primary observation and engagement sectors, and unit classification that defines available capabilities and appropriate tactics. These elements model the tactical arrangements that military units employ to maximize combat effectiveness in various situations. The situational awareness layer extends beyond the agent's immediate state to include known friendly positions shared through simulated communication networks, visible enemy positions limited by realistic line of sight calculations, and valid engagement sectors that define where weapons can be effectively employed. This information creates the tactical picture necessary for coordinated operations, while maintaining realistic constraints on what can be observed directly versus what must be communicated.

At the highest level, mission awareness provides objective position data and directional information that guides overall movement and tactical decisions. This layer ensures that all tactical decisions remain anchored to mission objectives, reflecting the mission-centric nature of military operations. This structured approach models the actual information hierarchy military leaders process during operations: first understanding their own unit's status, then integrating awareness of friendly and enemy positions, and finally relating this information to mission objectives. The environment implements sophisticated line of sight calculations that account for terrain features, implementing a realistic fog of war that limits what units can observe. This partial observability creates learning challenges that mirror the uncertainty inherent in actual military operations, forcing agents to develop robust tactics that account for incomplete information.

3.3. Action Space

The action space in our environment addresses a critical challenge in military simulation: balancing tactical complexity with learning efficiency. Rather than simplifying military actions to abstract game mechanics, we implement a progressive constraint approach that maintains tactical richness while enabling effective reinforcement learning.

3.3.1. Action Types

Our action space encompasses five fundamental action types that mirror military operations:

- MOVE (0): Basic movement operations
- ENGAGE (1): Direct fire at enemies
- SUPPRESS (2): Area suppression fire
- BOUND (3): Tactical bounding movement
- CHANGE_FORMATION (4): Modify unit formation

Each action type is parameterized with militarily relevant attributes. Movement uses direction vectors and distance parameters. Engagement includes target position, ammunition expenditure, and fire control settings. Formation changes specify formation types and orientation angles. This design maintains tactical authenticity while allowing for computational feasibility.

3.3.2. Progressive Constraint Approach

To enable efficient learning in a complex tactical environment, we adopted a progressive constraint strategy across three implementation phases. By incrementally narrowing the agent's action space, we significantly reduced the state-action space while retaining tactical depth. This methodology draws on findings by (Tang & Agrawal, 2020), who showed that discretizing continuous action spaces—especially with factorized distributions—can improve on-policy optimization. Their results demonstrated peak performance using 7–15 discrete values per dimension, aligning closely with our own final-stage configuration.

As summarized in Table 1, our constraints led to a 99.98% reduction in the total action space (from over 2.4 million possible actions to just 406), with similar reductions in state-action pairs. This simplification preserved meaningful variation in agent behavior while accelerating convergence.

3.3.3. Doctrine Modeling

This dramatic reduction in sample complexity accelerates convergence while maintaining essential tactical decision-making capabilities. However, the most significant innovation is our implementation of sophisticated target validation that incorporates doctrinal heuristics directly into the action space. Our action space includes robust target validation logic that ensures all engagements follow sound military principles. This approach aligns with recent research by (Basak et al., 2022), who demonstrated that incorporating doctrinal knowledge into multi-agent reinforcement learning systems can significantly improve exploration efficiency and lead to more militarily relevant behaviors.

Table 1. Progressive reduction of action and state-action space.

Feature	Original	Phase 2	Phase 3	Total Reduction
Movement Distances	21 (0–20)	11 (0–10)	3 (1,5,10)	86%
Engagement Rounds	30 (1–30)	30 (1–30)	3 (1,6,12)	90%
Target Positions	10,000	$\sim$8 visible enemies	+ validation	$>$99.9%
Formation Orientations	360°	8 directions	8 directions	97.8%
Total Actions	2,403,930	2,534	406	99.98%
State-Action Pairs	$\sim$24B	$\sim$25.3M	$\sim$4.06M	99.98%

Before any engagement action is executed, the system verifies that targets must be within the unit's observation range, within the unit's engagement range, within the unit's sectors of fire, and maintain line of sight with no terrain obstruction. When an agent attempts to engage a target outside its valid sectors or beyond its engagement range, the system either reroutes the action to the closest valid target or skips the action entirely. This approach offers several critical advantages:

First, it significantly reduces training time by preventing agents from wasting exploration effort on tactically invalid actions, effectively embedding domain knowledge into the learning process. Second, it ensures that trained agents exhibit behaviors that align with military doctrine even during early training phases, making the learning process more interpretable to military experts. Finally, it creates an action space that naturally guides agents toward sound tactical principles that human operators would otherwise have to teach through trial and error or explicit programming.

The progressive constraint implementation, combined with doctrinal target validation, creates a learning environment that balances tactical authenticity with learning efficiency. By reducing the dimensionality while preserving essential tactical options, and by embedding doctrinal knowledge directly into the action space, our approach enables reinforcement learning that produces sophisticated tactical behaviors without requiring impractical training time or computational resources.

3.4. Reward Structure

Our implementation incorporates reward shaping, a practice of modifying the natural reward signal to provide more informative feedback that accelerates learning without distorting the optimal policy. Research by (Ng et al., 1999) established that properly implemented reward shaping can dramatically accelerate learning without distorting optimal behavior. By carefully balancing tactical realism with effective learning signals, we've created a multi-layered incentive system that guides learning while maintaining tactical authenticity.

Our hybrid reward approach combines team-based rewards with individual agent-specific rewards. This dual structure creates a learning environment that simultaneously encourages mission-oriented cooperation and role-specific tactical behaviors.

Team reward components include:

- **Enemy Elimination Reward (2.0 per enemy):** Encourages agents to reduce opposing force strength through offensive action.
- **Objective Progress Reward:** Awards agents incrementally based on proximity to objectives: +0.05 within 50 units, +0.1 within 30, +0.2 within 20, and +0.3 within 10 units, reinforcing spatial advancement toward mission goals.
- **Force Concentration Bonus:** Grants +0.5, +1.0, and +2.0 when at least half of the team is within 50, 30, and 10 units of the objective, respectively, encouraging collective positioning and coordinated maneuvers.
- **Survival Penalty (-3.0 per friendly casualty):** Penalizes loss of allied units, making force preservation a tactical priority.
- **Step Penalty (-0.01 per step):** Applies a small penalty per timestep to maintain a sense of urgency and promote efficient decision-making.

Individual agent rewards provide personalized feedback:

- **Proximity Bonus:** Rewards agents based on their distance to the objective: +0.2 within 60 units, up to +3.0 within 5 units, using a graduated scale that encourages continuous forward progress.
- **Engagement Rewards:** Structured to shape combat behavior—+0.05 for attempting valid target engagement, +0.05 when enemies are visible, up to +0.1 scaled by aiming accuracy, +1.5 for successful hits, +3.0 for eliminating an enemy, with penalties of -0.1 for invalid engagements and -0.05 for firing without line of sight.
- **Movement Rewards/Penalties:** Adjust agent rewards based on movement direction and efficiency: up to

+0.15 for moving closer to the objective and up to -0.15 for moving away, both scaled by actual versus expected movement distance.

- **Role-Specific Reward Caps:** Squad leaders receive higher reward ceilings (base 1.5, hit multiplier 3.0) compared to regular members (base 0.8, hit multiplier 2.5), reflecting their leadership influence on tactical outcomes.

4. Network - Proximal Policy Optimization (PPO)

Proximal Policy Optimization (PPO), introduced by (Schulman et al., 2017), is a state-of-the-art policy gradient method in reinforcement learning. It strikes a strong balance between performance, stability, and ease of implementation, making it especially well-suited for complex decision-making domains such as tactical military applications.

4.1. Suitability for Military Tactical Execution

PPO offers several advantages that make it particularly well-suited for military tactical simulation environments:

- **Sample Efficiency:** PPO demonstrates significantly higher sample efficiency compared to other policy gradient methods, critical for simulations with costly interactions.
- **Stability During Training:** Bounded policy updates prevent catastrophic performance drops during training.
- **Effectiveness in Partially Observable Domains:** PPO works well in "fog of war" conditions common to military simulations.
- **Multi-Agent Adaptability:** PPO handles multi-agent scenarios where units must coordinate actions toward shared objectives.
- **Scalable Parallel Implementation:** PPO supports efficient parallel sampling across many scenarios, accelerating tactical learning.

4.2. Mathematical Foundation

PPO employs a clipped surrogate objective function that prevents excessive policy updates, maintaining stability during training:

$$L(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) A_t \right) \right]$$

where $r_t(\theta)$ is the ratio of probabilities under the new and old policies, A_t is the estimated advantage, and ϵ is a hyperparameter (typically 0.1 or 0.2) that constrains policy updates.

The algorithm uses Generalized Advantage Estimation (GAE) for computing advantage estimates:

$$A_t^{GAE(\gamma,\lambda)} = \sum_{i=0}^{\infty} (\gamma\lambda)^i \delta_{t+i}$$

where $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$ is the temporal difference error at time t.

4.3. Network Training - Curriculum Learning

Our training methodology employs *curriculum learning*, a technique inspired by human education that progressively introduces more complex tasks to the learning agent. This approach has been shown to significantly improve learning efficiency and performance in complex reinforcement learning environments.

Curriculum Structure

The curriculum is organized into four levels, each composed of multiple stages that systematically increase in difficulty.

Level 1: Basic Navigation (9 stages) introduces three distance categories—close (30–40 units), medium (50–60 units), and far (70–80 units)—with varying objective positions (top, middle, bottom) to promote generalization. The environment consists of bare terrain, allowing agents to focus on core navigation skills. Learning parameters are tuned to encourage exploration and facilitate rapid initial learning. This approach aligns with the "learning from easy missions" methodology first proposed by (Asada et al., 1995), where initial tasks start close to the goal state and progressively move farther away to facilitate learning.

Level 2: Terrain Navigation (9 stages) maintains the distance and objective variations of Level 1 but introduces complex terrain features that require more advanced pathfinding capabilities. Corresponding learning parameters are adjusted to account for this added complexity.

Level 3: Enemy Engagement with Limited Range (6 stages) comprises two parts: the first three stages feature far-distance navigation with enemies on bare terrain, while the final three introduce both enemies and complex terrain. This level implements a reduced engagement range of 10 cells, requiring close-quarters combat. The learning configuration places emphasis on accurate value function estimation to support tactical decision-making, similar to the approach used by Riedmiller et al. (2018) (Riedmiller et al., 2018), who employed scheduled auxiliary tasks to gradually increase the complexity of reinforcement learning agents' behavior.

Level 4: Advanced Tactical Operations (6 stages) increases the engagement range to 20 cells and introduces multiple enemies positioned around objectives. It com-

bines terrain navigation challenges with dynamic tactical engagements. Learning parameters are tuned for a balanced exploration-exploitation strategy, with a higher value coefficient to emphasize decision precision.

Across all levels, task complexity is incrementally increased through parameter tuning. Stage-to-stage consistency is maintained via parameter inheritance, allowing for smooth progression through the curriculum.

4.4. Theoretical Foundation

This curriculum structure builds upon established research in curriculum learning. (Florensa et al., 2018)

- **Progressive Distance Difficulty**: Gradually increasing the distance between agents and objectives creates an effective difficulty progression for reinforcement learning.
- **Environmental Complexity Layering**: The progression from bare terrain to complex terrain to terrain with enemies follows the principle of layered complexity introduction.
- **Tactical Skill Decomposition**: The curriculum decomposes tactical military skills into navigation, engagement, and integrated operations, aligning with the skill decomposition approach for complex task learning.
- **Parameter Adaptation by Task Complexity**: Systematic adjustment of learning parameters based on task complexity ensures that entropy coefficients, learning rates, and value coefficients are adjusted proportionally to task difficulty.

5. Experiment Design

To evaluate the generalization capabilities of the trained policy, we designed three distinct test scenarios, each introducing unique tactical challenges.

Test Map 1: Conventional Assault — The objective was to secure a position at coordinates $[350, 50]$, with two enemies located nearby at $[346, 50]$ and $[340, 48]$. Friendly forces consisted of 7 units deployed in relatively close formation to the south of the objective. This scenario emphasized a coordinated assault requiring direct engagement with enemy forces.

Test Map 2: Distributed Approach — The goal was to secure the position at $[336, 50]$, defended by enemies at $[330, 50]$ and $[336, 45]$. Friendly forces were dispersed widely, necessitating coordination across greater distances. The key tactical challenge here was synchronizing actions from distributed starting positions.

Test Map 3: Complex Terrain Navigation — The mission involved capturing a position at $[308, 78]$, with enemies in defensive positions at $[302, 75]$ and $[308, 82]$. Friendly units, again numbering 7, were arranged in a semi-circular formation. This test focused on navigating complex terrain that potentially restricted lines of sight and movement.

5.1. Testing Protocol

For each scenario, the agent was evaluated over 50 episodes with consistent parameter. The experimental setup involves a maximum of 500 steps per episode, utilizes consistent model weights loaded from the best training model, does not specify a random seed to allow for natural tactical variation, and includes the collection of performance metrics across all episodes.

6. Results

6.1. Performance Metrics

We evaluated the agent's tactical proficiency across three primary categories: mission effectiveness, force preservation, and operational efficiency.

Mission effectiveness was assessed using the following indicators: the *success rate*, which measures the percentage of episodes ending in mission success; the *objective reach rate*, indicating how often agents reached mission objectives; and the *average reward*, representing the mean cumulative reward per episode.

Force preservation metrics included the *team survival rate* (average proportion of surviving friendly units), *friendly casualties* (mean number of friendly units lost), and *enemy casualties* (mean number of enemy units neutralized).

Operational efficiency was evaluated by the *duration* (average number of steps to complete the mission), *rounds fired* (average ammunition expenditure), and the *efficiency ratio*, calculated as the total reward per unit of time (`reward/duration`).

6.2. Experiment Results

Our experimental evaluation demonstrates that the proposed curriculum-based training strategy yielded robust performance across all test scenarios.

Overall Performance Metrics

The agent exhibited consistently strong tactical behavior across diverse scenarios, with performance varying according to scenario complexity. Full quantitative results are provided in Table 2, showing the trade-offs between effectiveness, preservation, and operational efficiency across the test maps.

Metric	Map 1	Map 2	Map 3
Success Rate	94%	100%	86%
Avg. Reward	24.96	25.56	10.79
Objective Reach Rate	94%	100%	92%
Team Survival Rate	96.3%	96.6%	97.2%
Enemy Casualties	1.92	2.00	1.66
Friendly Casualties	0.26	0.24	0.20
Avg. Duration (s)	72.76	80.39	70.61
Efficiency	0.36	0.33	0.16

Table 2. Performance metrics across different test maps.

7. Tactical Behavior Analysis

The experimental results demonstrate that the PPO-trained agent developed sophisticated tactical behaviors that generalized across different scenarios. The consistently high team survival rates ($> 96\%$) across all scenarios indicate robust force preservation tactics. The agent consistently achieved high objective reach rates (94%, 100%, 92%) while balancing priorities like force preservation and enemy engagement, demonstrating effective mission focus. Performance differences between Map 3 and the other scenarios suggest terrain-aware navigation capabilities, although efficiency dropped in complex environments. The correlation between reward and team survival, along with varying enemy engagement patterns, indicates the agent learned to make contextually appropriate tactical decisions rather than relying on fixed behaviors.

Despite positive results, challenges were identified, including performance variability in later quartiles of Map 1 and fluctuating performance in Map 3, suggesting that the policy may not maintain consistent performance in extended operations. The significantly lower reward and efficiency metrics in Map 3 indicate that complex terrain navigation remains challenging. Time-limit failures in Maps 1 and 3 highlight opportunities to improve path planning and mission execution efficiency. Finally, the varying reward-metric correlations across scenarios suggest that the learned policy's behavior is sensitive to the reward structure, requiring scenario-specific tuning for optimal performance.

These results show that our approach successfully learns generalizable tactical behaviors that perform well across diverse scenarios, with sophisticated decision-making prioritizing mission success and force preservation.

8. Conclusion and Future Work

This paper presents an advancement in AI-assisted military planning through the development of a custom simulation environment, hierarchical route planning, and reinforcement learning-driven decision-making. Our approach addresses critical gaps in existing systems by providing a fully customizable environment, incorporating doctrinal knowledge directly into the state space representation, and using reinforcement learning to generate optimized friendly COAs without requiring extensive human expertise.

The experimental results demonstrate that our system can achieve high success rates across diverse military scenarios, often outperforming human experts and other AI baselines. The learned behaviors exhibit sophisticated tactical understanding, including adaptive force preservation, objective prioritization, terrain-aware navigation, and contextually appropriate tactical decisions.

Future Work

Future research should focus on several key areas:

- **Doctrinal Model Validation:** Collaborating with a more robust team of domain experts to validate and refine the doctrinal models used in the simulation environment.
- **Non-Stationary / Dynamic Adversary:** Developing more sophisticated adversary models based on differing threat doctrines (Russia, Iran, China, North Korea, etc.) to better prepare friendly forces for diverse enemy tactics.
- **Expanded Agent Types:** Extending the range of available agent types (artillery, armor, aviation, etc.) to better represent the aspects of combined arms maneuver.

By addressing these areas, we believe our approach can significantly enhance military planning capabilities, reduce the impact of personnel training and experience biases, and ultimately improve operational outcomes in complex and dynamic battlefields.

References

Asada, M., Noda, S., Tawaratsumida, S., and Hosoda, K. Vision-based reinforcement learning for purposive behavior acquisition. In *Proceedings of 1995 IEEE International Conference on Robotics and Automation*, volume 1, pp. 146–153 vol.1, 1995. doi: 10.1109/ROBOT.1995.525277.

Basak, A., Zaroukian, E. G., Corder, K., Fernandez, R., Hsu, C. D., Sharma, P. K., Waytowich, N. R., and Asher, D. E. Utility of doctrine with multi-agent RL for military engagements. In Pham, T. and Solomon, L. (eds.), *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV*, volume 12113, pp. 1211323. International Society for Optics and Photonics, SPIE, 2022. doi: 10.1117/12.2621242. URL https://doi.org/10.1117/12.2621242.

Chance, G., Pender, C., Holland, R., and Halliday, C. Using ai in wargaming simulation as a multi-domain decision support tool.

Florensa, C., Held, D., Wulfmeier, M., Zhang, M., and Abbeel, P. Reverse curriculum generation for reinforcement learning, 2018. URL https://arxiv.org/abs/1707.05300.

Goecks, V. G. and Waytowich, N. Coa-gpt: Generative pre-trained transformers for accelerated course of action development in military operations, 2024. URL https://arxiv.org/abs/2402.01786.

Ng, A. Y., Harada, D., and Russell, S. J. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pp. 278–287, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606122.

Riedmiller, M., Hafner, R., Lampe, T., Neunert, M., Degrave, J., de Wiele, T. V., Mnih, V., Heess, N., and Springenberg, J. T. Learning by playing - solving sparse reward tasks from scratch, 2018. URL https://arxiv.org/abs/1802.10567.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Schwartz, P. J., O'Neill, D. V., Bentz, M. E., Brown, A., Doyle, B. S., Liepa, O. C., Lawrence, R., and Hull, R. D. AI-enabled wargaming in the military decision making process. In Pham, T., Solomon, L., and Rainey, K. (eds.), *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II*, volume 11413, pp. 114130H. International Society for Optics and Photonics, SPIE, 2020. doi: 10.1117/12.2560494. URL https://doi.org/10.1117/12.2560494.

Tang, Y. and Agrawal, S. Discretizing continuous action space for on-policy optimization, 2020. URL https://arxiv.org/abs/1901.10500.