From Rules to Reasoning: Evolving Agentic AI for Strategy Synthesis in Multi-Agent Wargaming Environments

Amauri Straford[1], **Anaiya Reliford**[1], **Charles Milligan**[1]

[1]Research Institute for Tactical Autonomy
Washington, DC 20009

Abstract

Contemporary conflict environments are increasingly characterized by autonomy, decentralization, and rapid adaptation. Traditional rule-based artificial intelligence systems struggle to operate effectively in such settings due to their reliance on pre-specified logic and static assumptions. Reinforcement learning (RL) has improved adaptability, yet many implementations remain reactive and myopic, optimizing local rewards rather than synthesizing higher-level strategies. This paper presents an incremental research program exploring the evolution of agentic artificial intelligence within a custom digital wargaming environment. Beginning with deterministic rule-based agents, progressing through multi-agent reinforcement learning (MARL) with memory augmentation, and culminating in an emerging paradigm of self-directed, research-capable agents, this work examines how autonomy can be extended beyond action selection toward independent strategy discovery. Using a dynamic 3v3 drone "capture-the-flag" simulation, we demonstrate measurable gains in coordination, adaptability, and tactical effectiveness, and outline a next-phase architecture in which agents autonomously explore, evaluate, and synthesize strategies across episodes. These results contribute to ongoing efforts to move from reactive autonomy toward genuinely agentic systems suitable for complex, contested operational domains.

Motivation

Autonomous systems are rapidly reshaping the character of modern warfare. Uncrewed aerial systems (UAS), loitering munitions, and distributed sensing platforms increasingly operate in environments marked by uncertainty, deception, and rapid change. In such contexts, static decision logic is insufficient. Autonomous agents must not only act, but adapt, coordinate, and ultimately reason about strategy. AI-enabled wargaming has emerged as a critical tool for analyzing and training such systems within controlled simulation environments (Schwartz et al. 2020; Rinaudo et al. 2024).

Early AI-enabled tactical systems relied heavily on rule-based logic, encoding expert knowledge into deterministic decision trees. While predictable and interpretable, these systems are brittle when faced with novel situations. Reinforcement learning introduced adaptability through trial-and-error optimization, enabling agents to learn effective behaviors from experience. As demonstrated in combat simulation research, scaling intelligent agents in adversarial environments requires adaptive learning mechanisms capable of handling multi-agent coordination and non-stationarity (Black 2024; Black and Darken 2023). However, many RL-based systems remain limited to policy optimization within a narrow reward structure and lack the capacity to independently formulate or evaluate alternative strategies.

This research explores a staged progression toward agentic AI within a digital wargaming environment. The central question is not merely whether agents can learn what action to take, but whether they can learn why certain strategies succeed and independently generate new ones. We hypothesize that combining multi-agent reinforcement learning, memory-augmented architectures, and self-directed evaluation loops provides a viable pathway toward autonomous strategy synthesis.

Related Work

AI-enabled wargaming has received growing attention across defense and academic communities. Recent efforts have focused on integrating reinforcement learning agents into combat simulations to improve adaptability and realism (Black 2024). Scaling such agents to multi-agent adversarial environments presents nontrivial challenges, including coordination instability and reward misalignment (Black and Darken 2023). Training frameworks for AI-enabled wargaming have also emphasized structured experimentation and performance evaluation within defense research contexts (Rinaudo et al. 2024). The utility of hierarchical reinforcement learning for tactical decision-making, the scalability challenges of multi-agent environments, and the value of AI agents has has gained significant attention as a mechanism for modeling complex operational scenarios and supporting military decision-making (Schwartz et al. 2020).

Memory-augmented models, including Long Short-Term Memory (LSTM) networks, have been shown to improve performance in partially observable and non-stationary environments. In multi-agent contexts, communication and memory mechanisms further enhance coordination (Foerster et al. 2016).

However, most existing approaches focus on improving performance within a fixed task framing. Fewer efforts ad-

dress how agents might autonomously explore the strategy space itself—testing alternative doctrines, adapting coordination structures, or transferring lessons learned across scenarios. This work seeks to bridge that gap by framing reinforcement learning not as an endpoint, but as an intermediate capability in the development of agentic systems.

Technical Approach

Digital Wargame Design

The experimental platform is a Python-based multi-agent simulation developed using the PettingZoo framework. The environment models a 3v3 UAS engagement in a capture-the-flag scenario, selected for its balance of competition, cooperation, and dynamic objectives. Each team must coordinate navigation, detection, and engagement behaviors while operating in a contested space. Figure 1 provides a screenshot of the digital wargame interface showing a 3v3 UAS capture-the-flag scenario with dynamic terrain obstacles. Red and Blue agents are depicted navigating the grid-based environment with flag locations and obstacles visible. This figure illustrates the spatial complexity and adversarial setting used for training and evaluation.

To prevent overfitting and scripted behavior, the environment incorporates dynamic terrain. Obstacles are randomly repositioned between episodes, forcing agents to continuously adapt their navigation and tactics. This design choice reflects real-world operational uncertainty and limits reliance on static map knowledge.

Fifteen obstacles are sampled subject to feasibility constraints that prevent placement on base locations, flag positions, or spawn regions and ensure navigable paths remain between objectives. Drone spawn positions are also randomized within predefined team-specific regions. While the grid dimensions remain fixed, this stochastic terrain initialization induces cross-episode variability, requiring agents to adapt navigation and engagement strategies rather than memorize deterministic paths.

Rules of Engagement

The environment models a discrete-time, competitive 3v3 capture-the-flag engagement between two teams (Red and Blue) on a fixed 20×15 grid. Each team consists of three drones initialized in team-specific spawn regions (Red: $x \in [1, 5]$, Blue: $x \in [14, 18]$).

Mission Objective and Termination. The primary objective is to capture the opposing team's flag and return it to one's home base. Episodes terminate according to the following ordered conditions: (1) successful capture and return of the enemy flag, (2) total elimination of the opposing team, or (3) reaching a maximum horizon of 200 turns, in which case the team with the higher cumulative score wins (ties are permitted).

Movement Model. Agents operate in discrete timesteps and may execute exactly one action per turn. Movement is restricted to one orthogonal cell (up, down, left, or right). Diagonal motion is not permitted. Agents cannot move through obstacles, other drones, or outside map boundaries. The grid contains 15 randomly placed obstacles per episode, subject to constraints preventing placement on base locations, flag locations, or spawn regions.

Health and Combat Model. Each drone is initialized with 100 health points (HP). A valid attack requires that the attacker is alive and the target lies within a Manhattan distance of two cells. Line-of-sight constraints are not modeled. A successful attack inflicts 25 HP of damage. Health is strictly non-increasing and is clamped at zero:

$$HP_{t+1} = \max(0, HP_t - 25).$$

Four successful attacks are required to eliminate a drone. Upon reaching 0 HP, a drone is permanently removed from the episode and may no longer execute actions. No healing, regeneration, or respawn mechanics are implemented. If a destroyed drone was carrying a flag, the flag immediately resets to its home position. Attack and movement actions are mutually exclusive within a timestep.

Health is strictly non-increasing and is clamped at zero according to

$$HP_{t+1} = \max(0, HP_t - 25).$$

No healing or repair mechanisms are implemented. Consequently, once damaged, a drone remains damaged until destruction. Upon reaching 0 HP, a drone is permanently removed from the episode, its `is_alive` status is set to false, and it may no longer execute actions. If the destroyed drone was carrying a flag, the flag immediately resets to its home position.

Combat Constraints. Attacks are range-limited to two Manhattan cells. Friendly fire is permitted and penalized within the reward structure. Invalid attacks result in zero damage.

Flag Mechanics. Each team has a fixed base location (Red: $(2, 7)$; Blue: $(17, 7)$) and an associated flag initially positioned at that base. A flag may be captured when an opposing drone is within Manhattan distance one of the flag and the flag is in its home state. A captured flag must be transported to the capturing agent's home base and returned (distance ≤ 1) to secure victory. Flag carriers may continue to move and attack. Only one flag may be carried per drone.

Reward Structure. The reward function is multi-objective and aligned with mission priorities. Agents receive positive reward for flag capture and return, damage inflicted (+5 per damage point), and opponent elimination (+50). Per-turn survival yields positive reward. Penalties include friendly fire (-100), flag loss, and ineffective or invalid actions. This structure promotes mission completion, survivability, and coordinated behavior rather than purely aggressive elimination.

Turn Execution. Each timestep proceeds through the following phases: (1) simultaneous action selection, (2) movement and action execution, (3) combat resolution and HP updates, (4) flag state updates, (5) reward computation, and (6) termination check. The environment dynamics are deterministic given selected actions; stochasticity arises only from initial obstacle placement, spawn initialization, and learning-policy exploration.

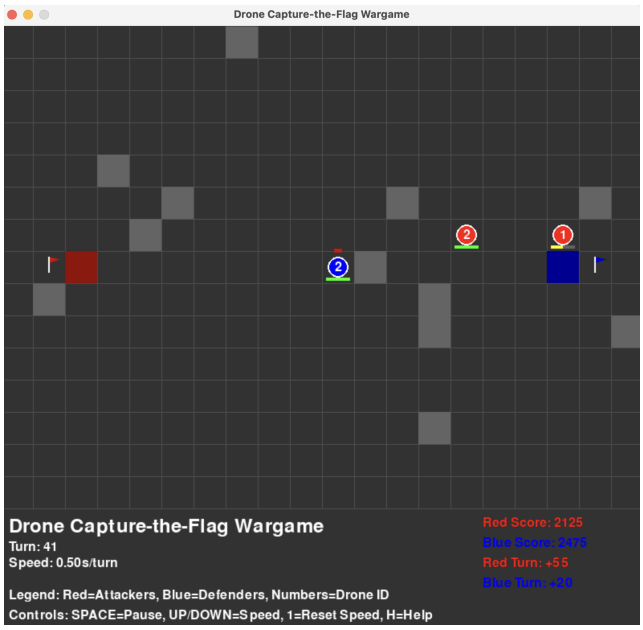

Figure 1: Snippet of 3v3 Capture-the-Flag Wargame Environment.

Agent Roles and Observations

Each agent receives partial observations of the environment, including relative positions, obstacle proximity, and limited opponent information. Rewards are structured to encourage mission success, survivability, and team coordination, rather than isolated individual performance.

Agent Architecture Evolution

Rule-Based Baseline

The initial implementation employed deterministic, rule-based agents. These agents followed predefined logic for navigation, engagement, and objective pursuit. While effective in simple or predictable situations, performance degraded significantly when terrain or opponent behavior changed. This baseline established a reference point for subsequent learning-based approaches.

Multi-Agent Reinforcement Learning

The second phase introduced reinforcement learning using Proximal Policy Optimization (PPO) and Independent PPO (IPPO) (Schulman et al. 2017). Agents learned policies through repeated interaction with the environment, gradually improving performance across episodes. MARL enabled emergent coordination behaviors, such as implicit role differentiation and cooperative movement patterns.

Memory-Augmented Learning

To address temporal dependencies and partial observability, Long Short-Term Memory (LSTM) networks were integrated into the agent architectures. LSTMs allowed agents to retain information about prior actions, opponent behavior, and environmental context within an episode.

Memory augmentation produced substantial performance gains. Episodic memory is preserved across time steps, enabling agents to leverage past observations, actions, and outcomes during decision-making. Agents demonstrated improved situational awareness, avoided repeating ineffective behaviors, and better anticipated opponent movements. These results underscore the importance of temporal reasoning in dynamic, adversarial environments.

Results and Analysis

Performance Trends

Across multiple simulation runs, LSTM-enhanced RL agents consistently outperformed rule-based baselines. Observed improvements included higher capture-the-flag success rates, more efficient navigation through dynamic terrain, and increased survivability during engagements.

Behavioral Evolution

Qualitative analysis revealed a clear evolution in agent behavior. Early training phases were characterized by exploratory and often suboptimal actions. Over time, agents adopted more coherent tactics, including coordinated advances, adaptive retreat behaviors, and implicit task specialization.

System Stability

Initial development faced challenges related to visualization, path tracing, and simulation stability. Iterative debugging and architectural refinements resulted in a robust platform capable of real-time visualization and consistent performance evaluation, enabling deeper behavioral analysis.

Toward Fully Agentic Strategy Synthesis

While memory-augmented MARL enables agents to adapt policies based on experience, it does not inherently provide agency in the sense of independent inquiry or strategic reasoning. To sharpen the agentic contribution of this work, we frame the next architectural evolution around autonomous hypothesis generation and self-evaluation loops.

In this paradigm, agents are no longer limited to optimizing a fixed reward function within a single training regime. Instead, they operate within a higher-level control loop that enables them to:

- Formulate hypotheses about effective strategies (e.g., aggressive early capture, decentralized scouting, defensive flag protection)
- Instantiate and test these hypotheses across multiple simulation episodes
- Evaluate outcomes using mission-level metrics such as win rate, survivability, coordination efficiency, and time-to-objective
- Update internal beliefs about which strategic patterns generalize across environments and opponents

This process treats strategies as first-class objects of reasoning rather than implicit byproducts of gradient descent.

Hypotheses may be represented explicitly (e.g., parameterized behavior templates or coordination schemas) or implicitly through meta-policies that select among learned tactical modes.

Crucially, self-evaluation is performed out-of-episode, enabling agents to reason across temporal horizons longer than those accessible to standard RL. Episodic logs, memory embeddings, and outcome summaries provide the substrate for reflective analysis, allowing agents to distinguish between situational success and robust strategic advantage.

By embedding this hypothesis–test–evaluate cycle within the agent architecture, the system begins to approximate strategy synthesis rather than mere policy adaptation. Such agents are capable of exploratory reasoning, comparative assessment, and strategic transfer—hallmarks of genuinely agentic behavior. This capability is particularly relevant in adversarial and non-stationary environments, where static doctrines rapidly lose effectiveness.

The proposed framework positions reinforcement learning as an enabling substrate rather than an endpoint, supporting a shift toward autonomous agents that can independently research, critique, and refine their own tactical approaches.

While memory-augmented MARL represents a significant advance, it remains largely reactive. The next phase of this research explores agents that autonomously research their own performance. This includes:

- Maintaining episodic memory and performance logs
- Generating and testing alternative strategies across runs
- Evaluating outcomes relative to mission objectives
- Synthesizing higher-level tactical insights

Rather than optimizing a single policy, such agents would operate within a meta-learning loop, reasoning about strategy selection itself. This approach moves beyond policy learning toward autonomous doctrine exploration, a key requirement for future contested and rapidly evolving operational environments.

Limitations and Future Work

The current system remains a beta-level research platform. Limitations include computational cost, limited action space, and the absence of long-horizon cross-episode learning. Planned future work includes:

- Scalable team sizes and heterogeneous agent roles
- Expanded mission types (e.g., search-and-destroy, area denial)
- Advanced logging and causal analysis tools
- Meta-learning and agent self-reflection mechanisms

Conclusion

This work demonstrates a structured pathway from rule-based autonomy to reinforcement learning and toward fully agentic AI within a digital wargaming context. Memory-augmented MARL significantly improves adaptability and coordination, while emerging agentic concepts promise a further leap toward autonomous strategy synthesis. These findings support the viability of agentic AI as a foundational capability for future autonomous systems operating in complex, adversarial domains.

Acknowledgments

The authors acknowledge the contributions of the Research Institute for Tactical Autonomy (RITA) at Howard University. Special thanks are extended to Dr. Sonya Smith for project sponsorship and to the student research team for their collaborative efforts in system development and experimentation.

References

Black, S. 2024. Mastering the digital art of war: Developing intelligent combat simulation agents for wargaming using hierarchical reinforcement learning. Ph.D. dissertation, Naval Postgraduate School, Monterey, CA. arXiv:2408.13333.

Black, S., and Darken, C. 2023. Scaling intelligent agents in combat simulations for wargaming. In Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC). arXiv:2402.06694.

Foerster, J.; Assael, I. A.; de Freitas, N.; and Whiteson, S. 2016. Learning to communicate with deep multi-agent reinforcement learning. In Advances in Neural Information Processing Systems 29, 2137–2145.

Rinaudo, C.; Leonard, W. B.; Morey, C.; Coumbe, T.; Hopson, J.; and Hillborn, R. 2024. Artificial intelligence–enabled wargaming agent training. U.S. Army Engineer Research and Development Center (ERDC) Technical Report.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. arXiv: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. 2020. AI-enabled wargaming in the military decision-making process. In Proceedings of SPIE 11413, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II, 114130H.