THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Separation Assurance between Heterogeneous Fleets of Small Unmanned Aerial Systems via Multi-Agent Reinforcement Learning

Iman Sharifi[1], Hyeong Tae Kim[2], Maheed Ahmed[2], Mahsa Ghasemi[2], Peng Wei[1]

[1]George Washington University

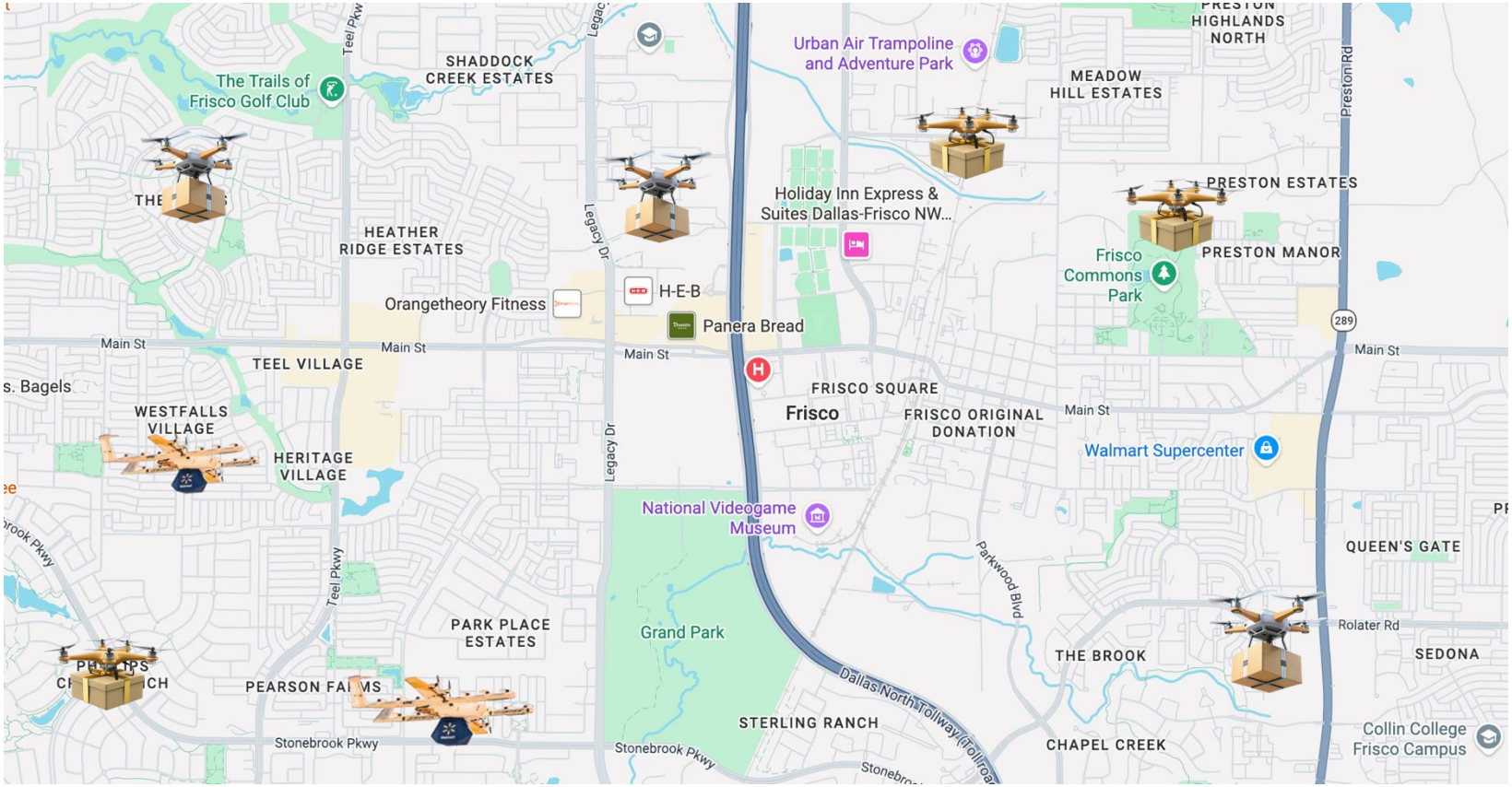
[2]Purdue University

ITSC, September 2026

Background

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Dense Urban Airspace



Proprietary Policies + Performance Configurations

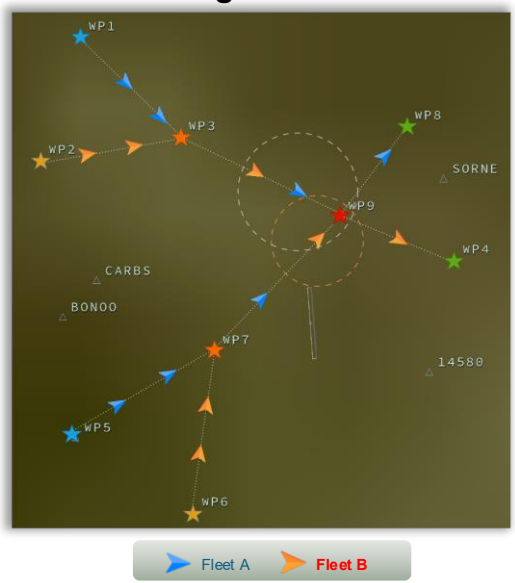
- Acceleration Range
- Speed Range
- Sensory Range

Challenging Tactical Deconfliction

Objective

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Use-case with two heterogeneous fleets



Heterogeneous fleets with independent policies and configurations

Fleet A:
Policy π_A
Config X (MK30)

Fleet B:
Policy π_B
Config Y (Wing)

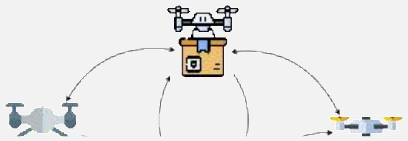
Research Questions

- Can heterogeneous fleets (A&B) reach an equilibrium with two multi-agent reinforcement learning (MARL) policies and configurations after training?
- If yes, is the interaction between fleets fair? Or biased?
- What are the factors contributing to potential discrimination?

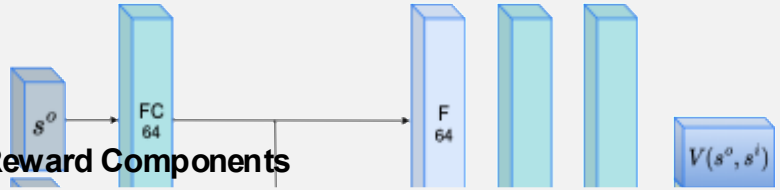
Multi-Agent Reinforcement Learning

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Multi-Agent Reinforcement Learning (MARL)



Attention-based Advantage Actor-Critic with Proximal Policy Optimization (PPO) Loss



Reward Components

Safety rewards:

Loss of Separation reward:

$$R_{\text{LoS}} = \begin{cases} -1, & \text{if } d_{\text{o}}^{(i)} < d_{\text{NMAC}}, \\ \alpha\left(-1 + \dfrac{d_{\text{o}}^{(i)} - d_{\text{NMAC}}}{d_{\text{LoWC}} - d_{\text{NMAC}}}\right), & \text{if } d_{\text{NMAC}} \le d_{\text{o}}^{(i)} \le d_{\text{LoWC}}, \\ 0, & \text{otherwise.} \end{cases}$$

Efficiency rewards:

Velocity-bounding reward:

$$R_{\text{V}} = \begin{cases} -\psi_1^{\text{v}}, & \text{if } v^{(o)} < v_{\min}^{(o)} + \eta_1^{\text{v}}, \\ -\psi_2^{\text{v}}, & \text{if } v^{(o)} > v_{\max}^{(o)} - \eta_2^{\text{v}}, \\ 0, & \text{otherwise.} \end{cases}$$

Action-smoothing reward:

$$R_{\text{A}} = \begin{cases} -\psi_1^{\text{a}}, & \text{if } a_t^{(o)} \neq a_{t-1}^{(o)}, \\ -\psi_2^{\text{a}}, & \text{if } a_t^{(o)} \neq \text{hold}, \\ 0, & \text{otherwise.} \end{cases}$$

Mission-completion reward:

$$R_{\text{M}} = \begin{cases} \psi_m, & \text{if } d_f^{(o)} < \eta_m, \\ 0, & \text{otherwise.} \end{cases}$$

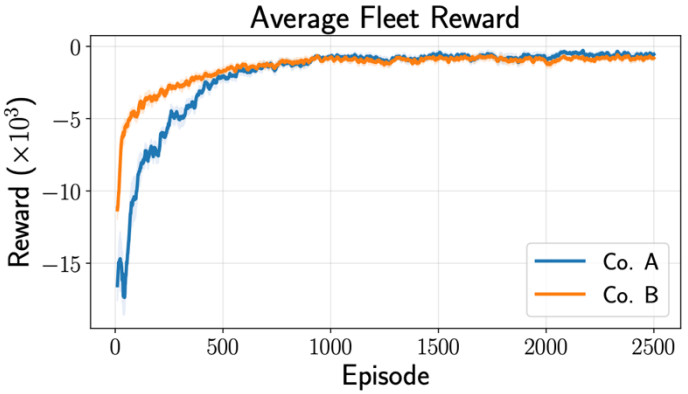
Time reward:

$$R_{\text{T}} = \begin{cases} -\psi_t, & \text{if } t < T, \\ -1, & \text{otherwise.} \end{cases}$$

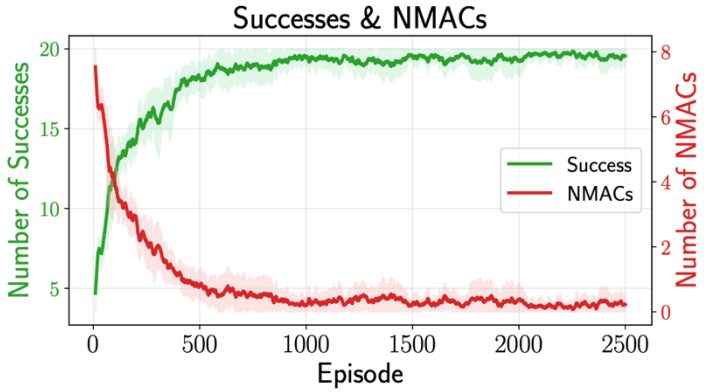
PPO Loss

$$\mathcal{L}_\pi = -E_t\left[\min\left(\zeta_t(\theta)A_t,\ \text{clip}\left(\zeta_t(\theta), 1-\epsilon, 1+\epsilon\right)A_t\right)\right] - \beta H\left(\pi(s_t)\right)$$
$$\mathcal{L}_v = A_t^2$$

Training Performance



Training rewards for fleet A & B with heterogeneous MARL policies and performance configurations



Number of successful agents and the Number of Mid-Air Collisions (NMACs) in each mission

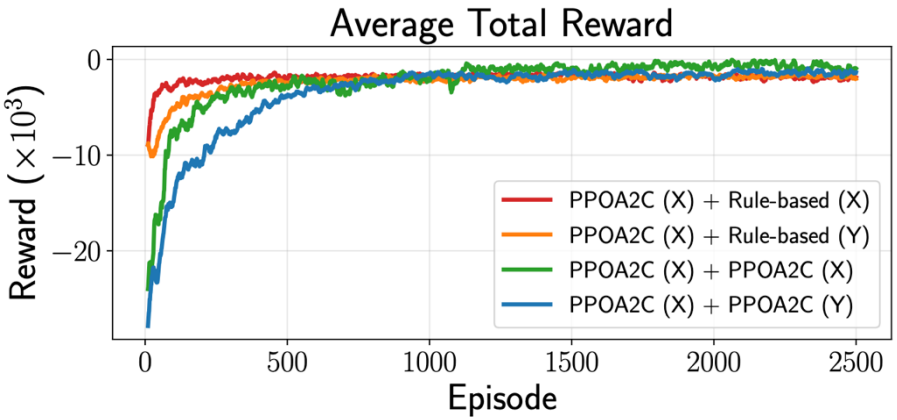
Implications:

1-Both fleets converge to the highest reward values during training.

✓ Reach an Equilibrium

2-NMACs converge to zero, and the number of successful agents converges to the maximum value.

✓ Safe and Efficient Separation

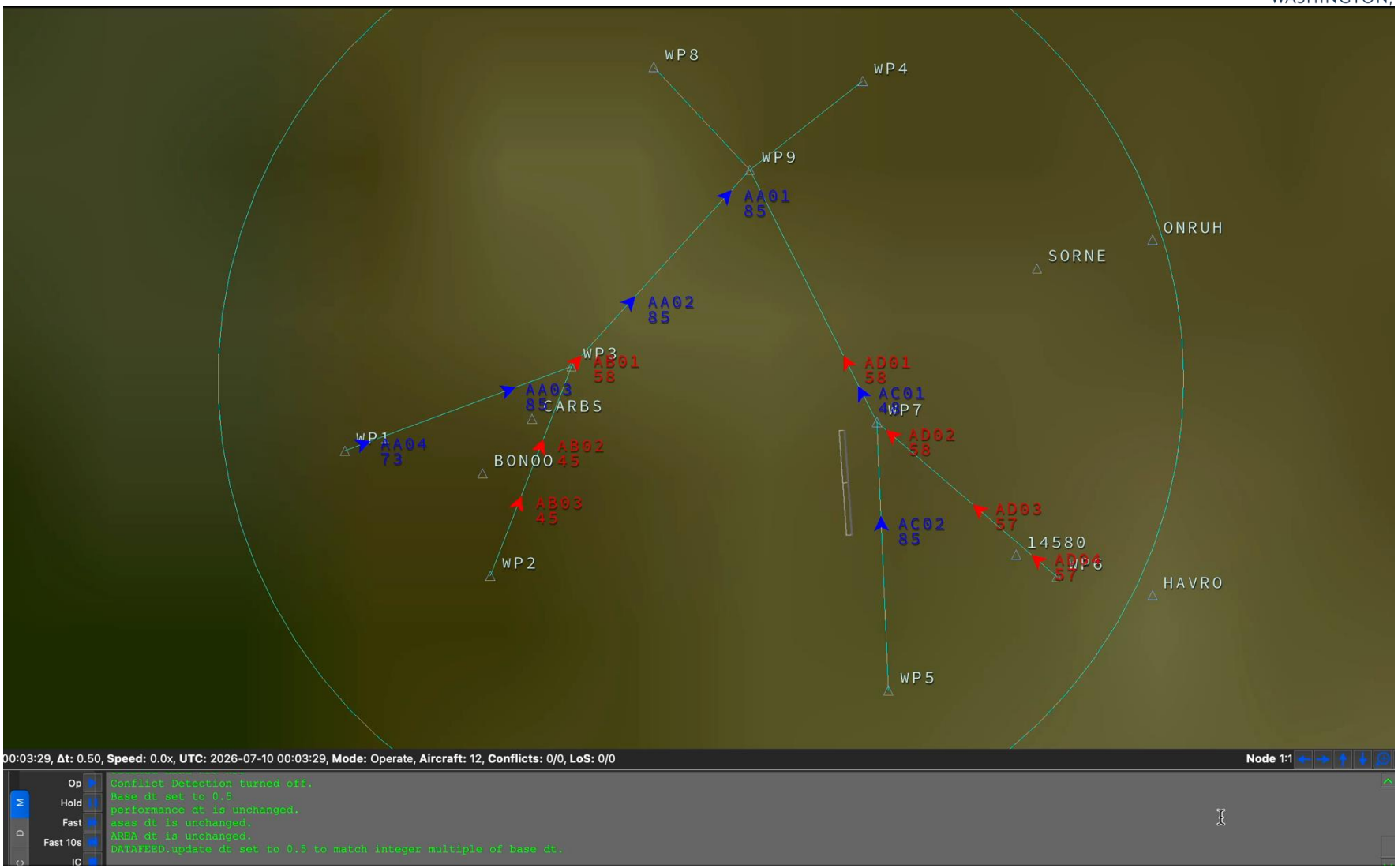


Total training rewards for both heterogeneous MARL policies with variable configurations

✓ Robust and adaptive to the counterpart policy

✓ Robust to configuration settings

Visualization



Fleet A Fleet B

Evaluation on BlueSky Simulator

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

First Row: Average total, inter-fleet, and bottleneck NMACs during 300 evaluation episodes for different combinations of policies and configurations. Second Row: success rates and evaluations rewards. Third Row: Mission times for both fleets and the Fairness Metric.

		XY Configurations				XX Configurations		
Policy A (Configuration A):		**Random(X)**	**Rule-based(X)**	**PPOA2C(X)**	**PPOA2C(X)**	**Rule-based(X)**	**PPOA2C(X)**	**PPOA2C(X)**
Policy B (Configuration B):		**Random(Y)**	**Rule-based(Y)**	**PPOA2C(Y)**	**Rule-based(Y)**	**Rule-based(X)**	**PPOA2C(X)**	**Rule-based(X)**
Average NMAC (↓)	**AA**	3.30	0.24	0.03	0.000	0.19	0.02	0.01
	AB	1.23	0.25	0.15	0.005	0.30	0.26	0.13
	BB	3.44	0.26	0.02	0.001	0.04	0.02	0.00
	M1	3.92	0.00	0.00	0.005	0.16	0.13	0.06
	M2	3.96	0.00	0.00	0.000	0.13	0.17	0.08
	IN	0.10	0.77	0.22	0.001	0.24	0.00	0.00
	Total	7.99	0.77	**0.22**	**0.006**	0.53	**0.30**	**0.14**
Success (↑)	**N_s / 20**	3.96±1.77	18.53±1.65	**19.55**±0.98	**19.98**±0.10	19.42±0.7	**19.79**±0.60	**19.90**±0.34
Reward (↑)	**R** ($\times 10^3$)	–	-3.15	**-1.41**	-1.76	-3.27	**-0.80**	-1.87
Mission Time	$\bar{T}_A$ (min)	–	5.88	7.23	6.58	5.05	7.42	5.46
	$\bar{T}_B$ (min)	–	7.13	8.94	7.60	5.05	7.54	5.02
Fairness (↑)	F_t (%)	–	82.4	80.8	86.5	**100**	**98.4**	91.9

Safety & Efficiency Implications:
1-Two MARL policies offer safer fleet interactions compared to two strong Rule-based and Random Policies.
2-One MARL policy adapts even more safely with one Rule-based policy.
3-Interaction of two MARL policies is more efficient than one MARL policy and one Rule-based policy.

Time-based Fairness Implications:
1-Fairness increases for similar policy types and performance configurations for both fleets of agents.
2-Fairness decreases when policy types and configurations are not the same, requiring fairness-aware algorithms.

$$F_t = \left(1 - \frac{|T_A - T_B|}{\max(T_A, T_B)}\right) * 100$$

Conclusion

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Key Contributions

- ✓ Attention-enhanced PPOA2C policy per fleet
- ✓ Shared within fleet, private across fleets
- ✓ Dense Dallas use case: 20 aircraft, 3 bottlenecks
- ✓ Time-based fairness metric across fleets

Key Findings

- ✓ Equilibrium reached: about 98% mission success
- ✓ NMACs cut by about 71% vs strong rule-based
- ✓ Versus rule-based: safer, yet less efficient
- ✓ Fairness: about 99% matched, 82% mismatched

Takeaway

- ✓ Safe separation achievable, fairness not guaranteed: need fairness-aware conflict management

THANK YOU!

Email:
i.sharifi@gwu.edu