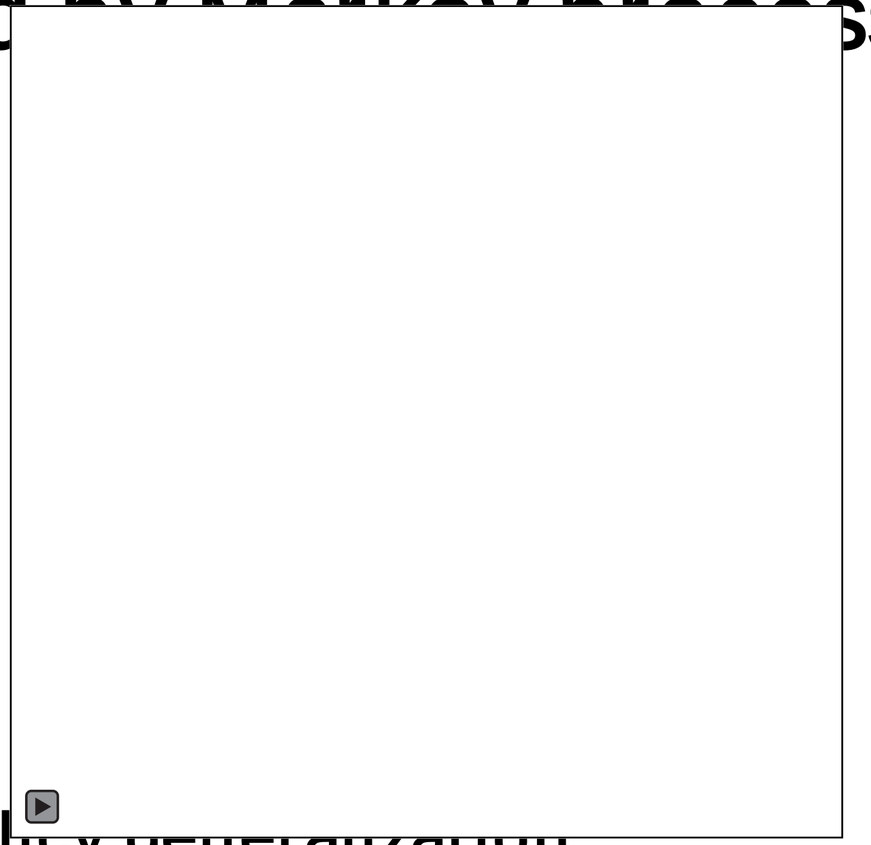
Cooperation of ad-hoc agents under partial observations

Eugeniy Vinokur
Supervisor: Dmytro Kuzmenko, Sr. lecturer, Dep. of Multimedia Systems

Fire modeled by Markov process



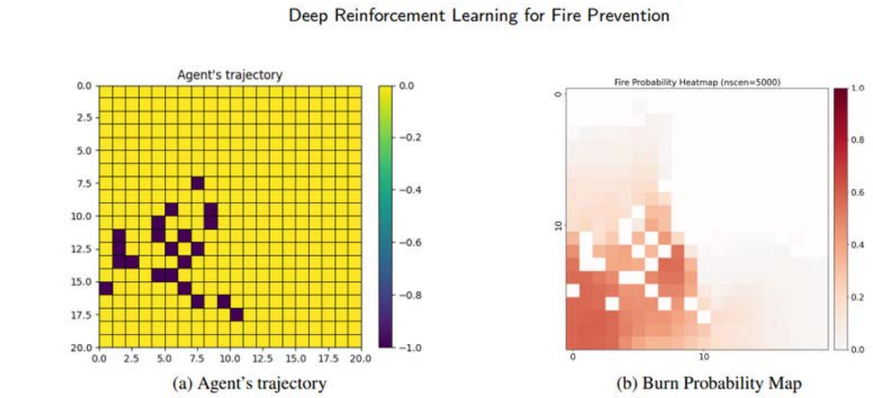
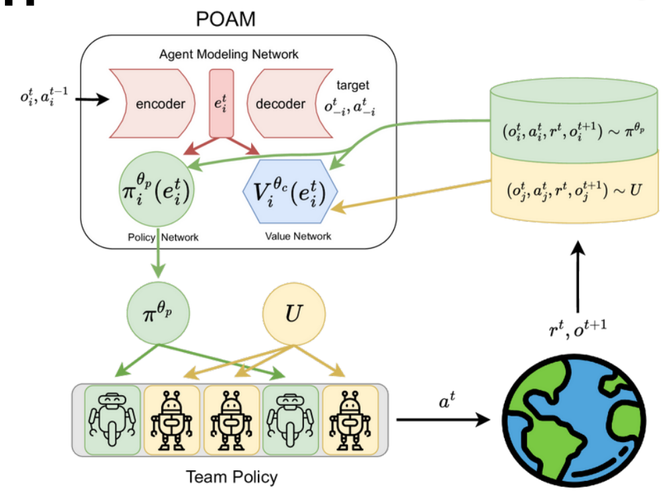


Figure 11: Best solution obtained and the corresponding fire spreading behavior for the Sub20 forest.

Figure: Murray's heuristic of single agent firebreak optimization [4]

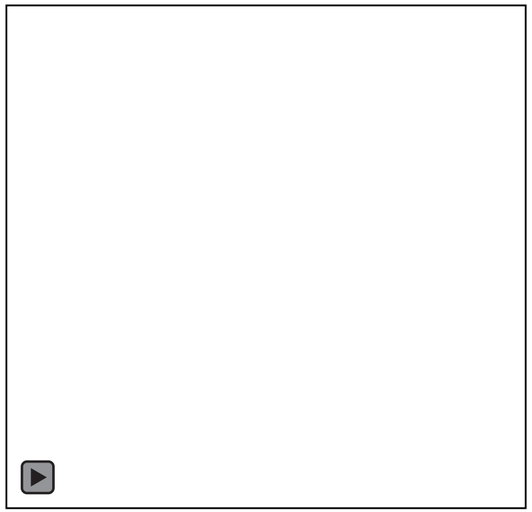
Unpredictable policy generalization



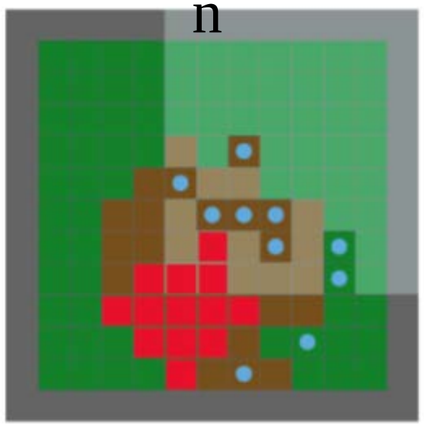
(Wang et al.) **sets** of unknown policies capable individually

Wildfire propagation model

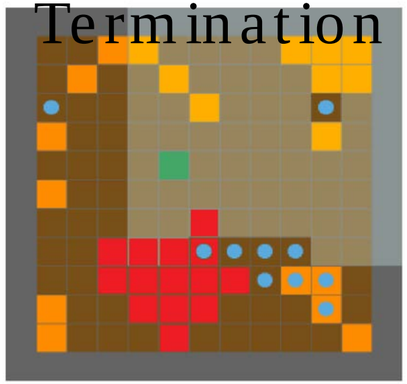
Start



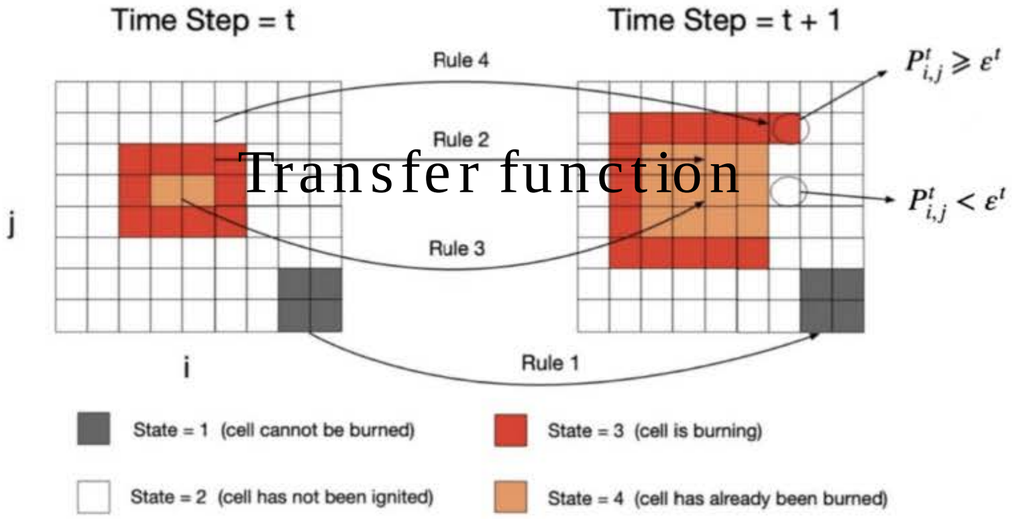
Termination



Unconstrained Termination



Transfer function



Environment fire transfer process, Figure by Yiqing Xu (2022 [1]

Affected parameter	Original	Fast fire	Inefficient
α	0.3	0.5	0.3
β	0.2	0.2	0.4
Agent impact of β	0.8	0.4	0.3

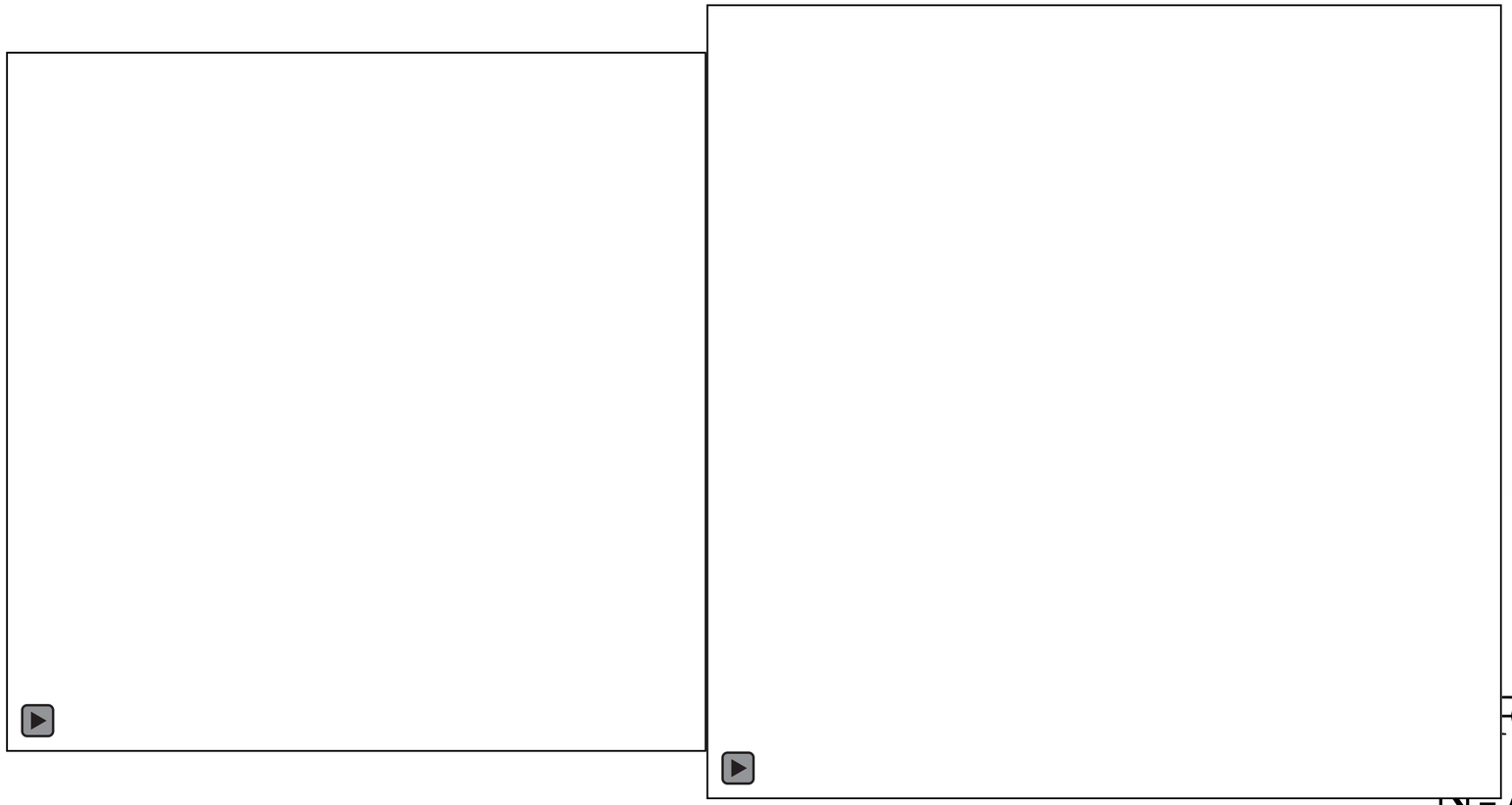
Table 3: Configuration of wildfire spread model for original and stress test setups.

Overview of challenges in environment

FO (Full observation)

Partial observability

Ad-hoc coordination



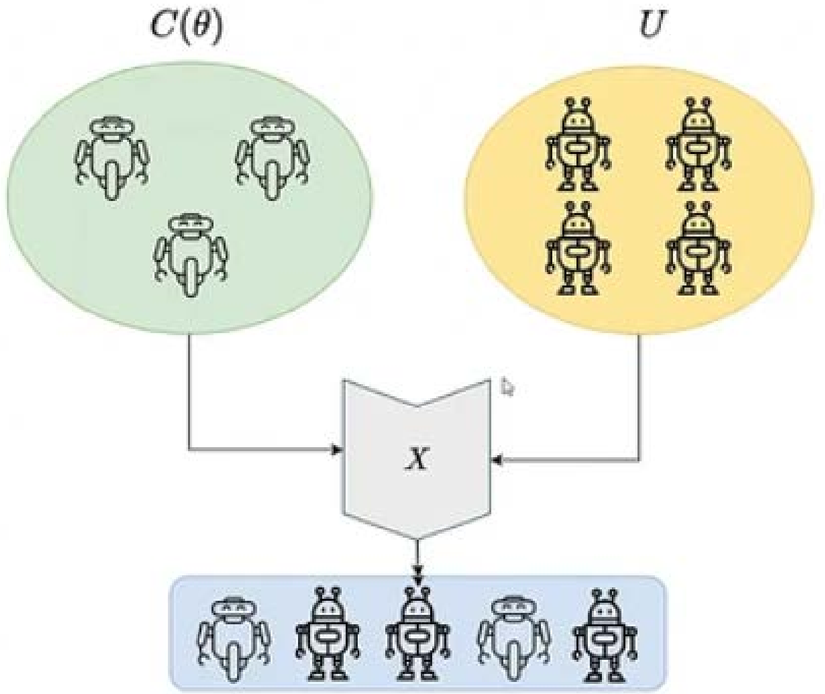


Figure: Sampling procedure in “N-Agent Adhoc teamwork” Wang et al.(2024)[4]

Research objectives

Improve coordinated reward: investigate reward shaping and factorized value MARL techniques; Show PoC for emerging behavior

Evaluate zero-shot robustness of RL baselines: with uncertain dynamics tests in POMDP

- **Research object/subject**
  - Collaboration with unknown agents and conditions
    - Addressing credit assignment in stochastic fire simulation

Reward shaping: Overview

Name	Equation
Baseline	$-0.5\times$ (N of trees on fire)
AgentAbove	$\begin{cases} 1, & \text{if on fire tile} \\ -1, & \text{otherwise} \end{cases}$
FireAdj	$+2^N$
AgentAdj	$\begin{cases} +2^{(3-\delta)}, \\ +10 \text{ if } \delta < 2 \end{cases}$
HealthyAdj	$+5+N$

Table: Reward Features: N - outdegree of healthy trees, d - Manhattan distance

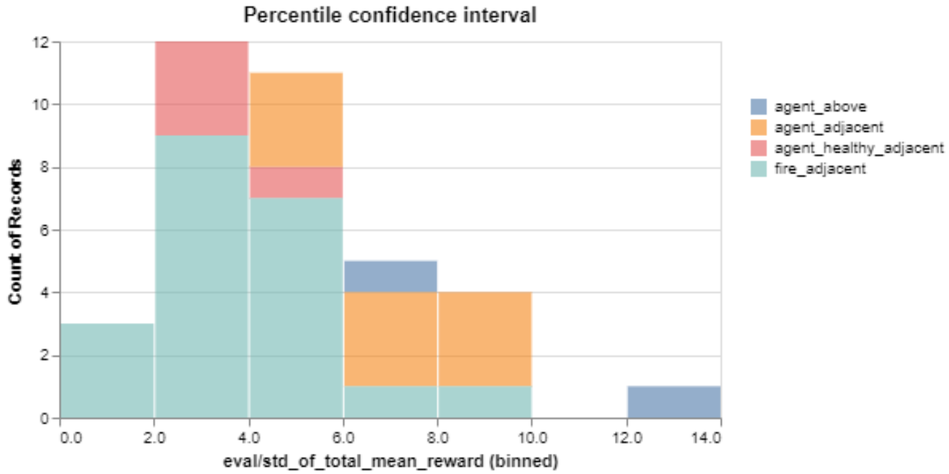
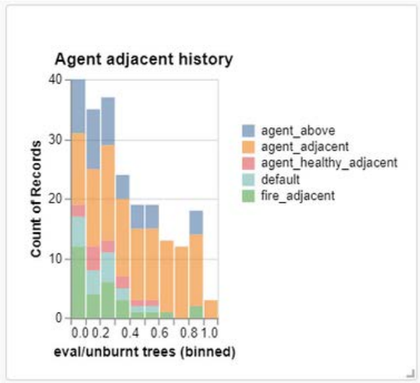
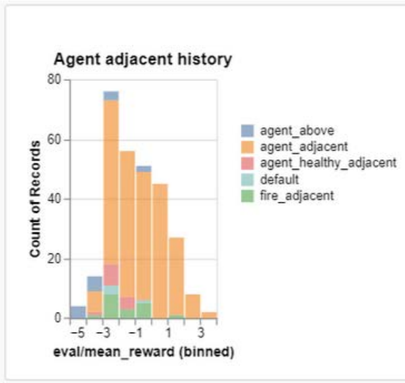


Figure: Per reward run variance distribution





Per reward run means sample distribution

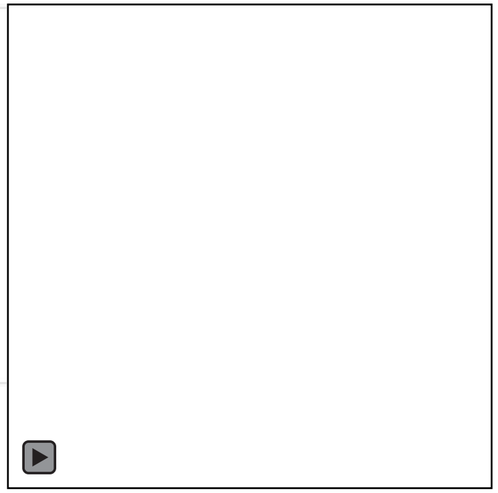


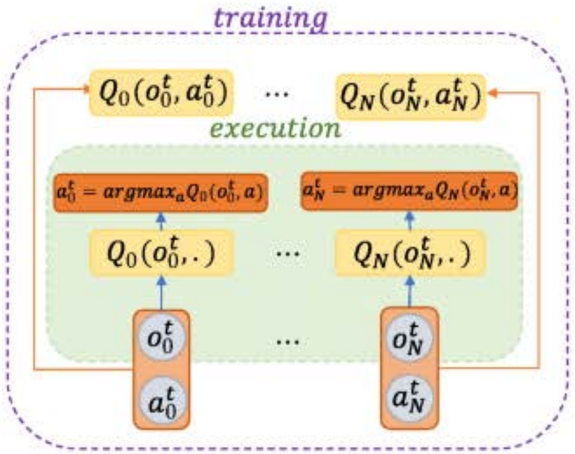
Figure: PPO trained on common reward

Multi-agent architectures: overview

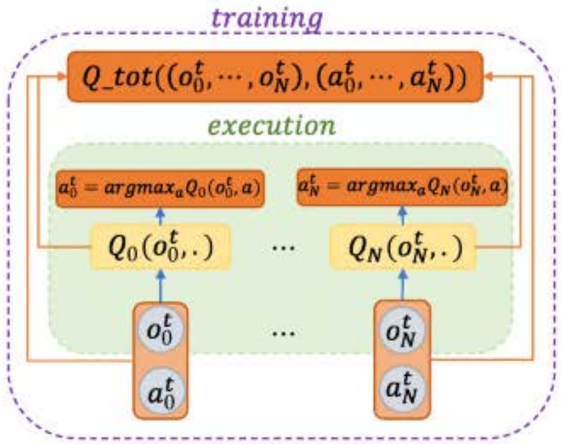
We select 5 paradigms dealing with credit assignment

Sharing **knowledge** schemes:

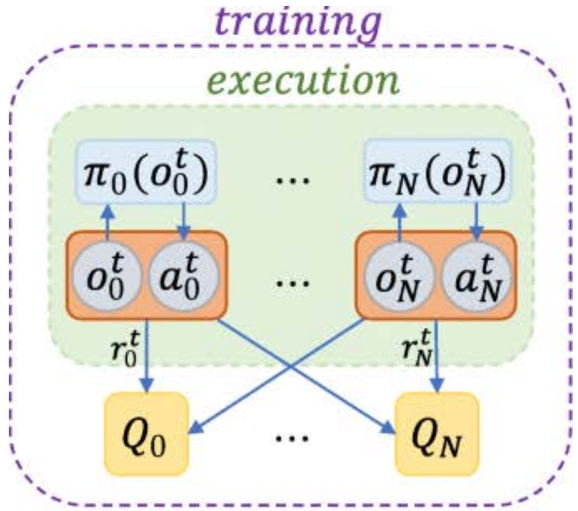
Graphs provided by “A review of cooperative multi-agent deep RL” [3]



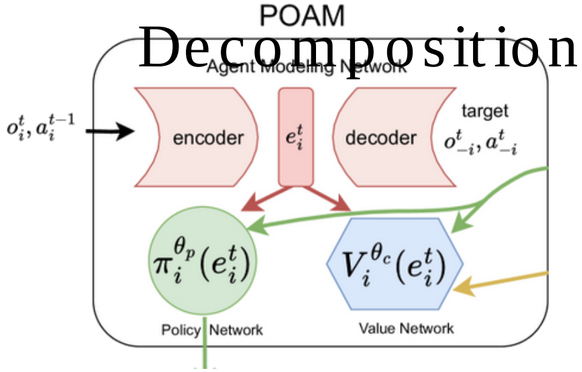
Independent



Value Decomposition



Central critic



POAM (Algorithm by Wang et al.)[4] - Recurrent encoder of teammate policies

Evaluation Approach

Metrics

- 1162 experiments, 356 models, 2.9B samples
- Training: 25 samples - aggregated means

- Evaluate in few-shot IQL vs POAM
- Confidence interval and std - from percentiles of
- normal 95% z-score
- Hardware: i7-4600U, 16GB RAM, 8 parallel runs

$$\frac{1}{N}\sum_{i=1}^{N} R_{t,i,j}, R \in \hat{X}_{[(1-\lambda);(\lambda)]}$$

$$\hat{X}_{ps} = R_i - \frac{1}{N}\sum_{i=0}^{N}(R_{-i})$$

Variance estimation method

$$\left[\sum_{j=0}^{|A|}\sum_{k=0}^{N} r_{t+k}{}^{j} \;\middle|\; o_t = o, a_t^j = a^j\right]$$

Mean reward

$$R_{Unburnt} = \frac{S_{Healthy}}{S_{Burned} + S_{Healthy}}$$

Remaining trees

Episode length-

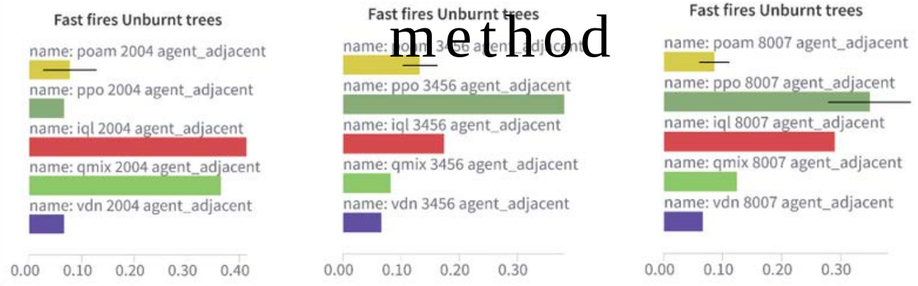


Figure: Means across 3 Seeds

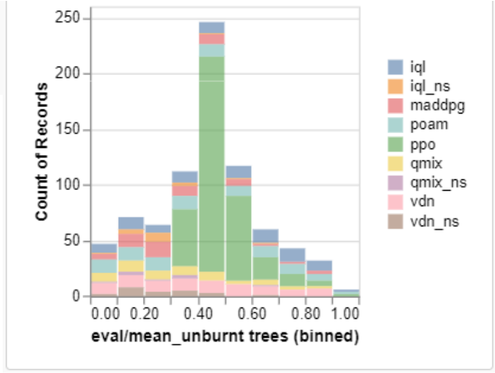
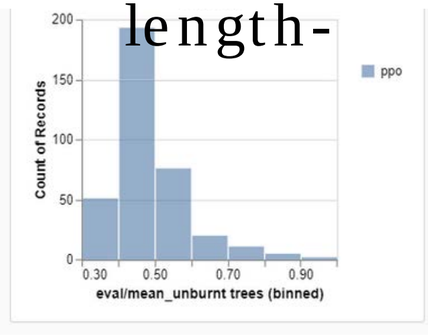


Figure: Aggregated Means from distribution for all runs



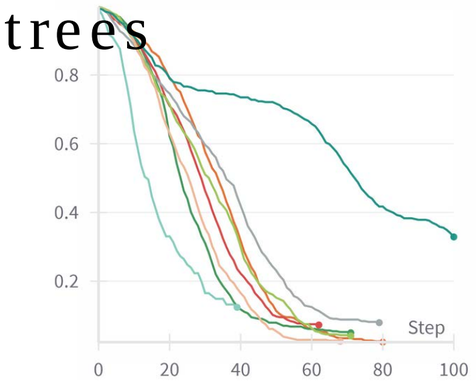
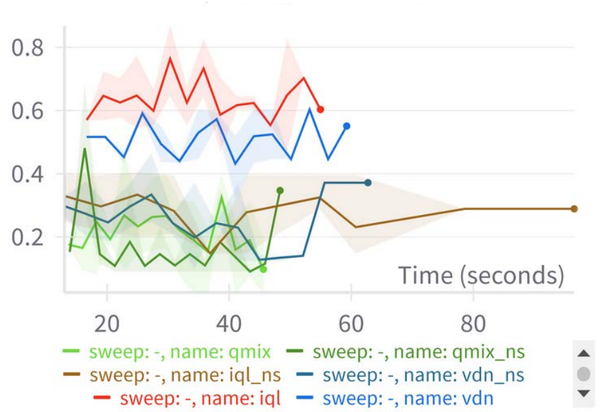


Figure: trees metric plateaus on longer rollouts

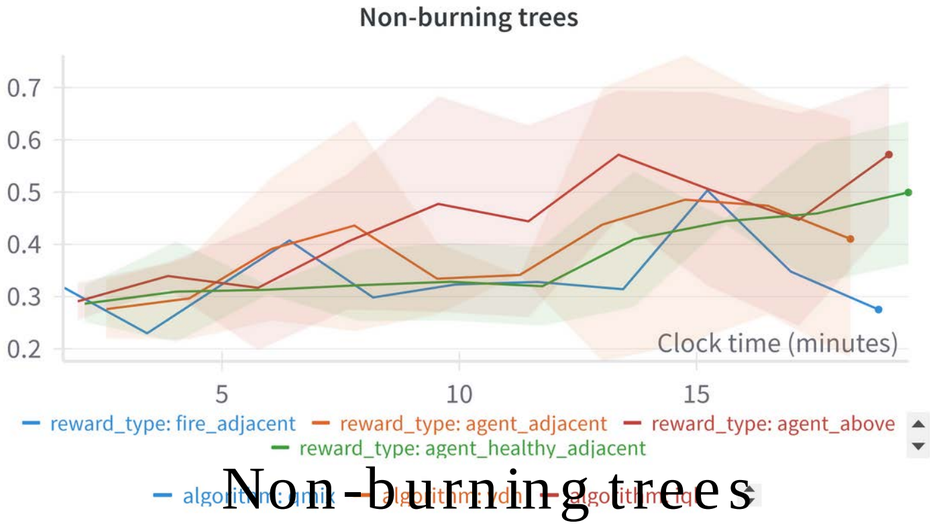
Partial-observability Finetuning

Algorithm	Obs. Limit	Reward Type	Unburnt Trees	Reward Mean
IQL	10	AgentAdj	0.666 ± 0.111	2.556 ± 0.648
IQL	6	FireAdj	0.547 ± 0.109	0.564 ± 0.647
IQL	13	FireAdj	$\mathbf{0.645 \pm 0.109}$	0.658 ± 0.749
PPO	10	AgentAdj	$\mathbf{0.725 \pm 0.106}$	0.392 ± 0.445
PPO	6	AgentAdj	$\mathbf{0.725 \pm 0.106}$	0.392 ± 0.445
PPO	6	FireAdj	0.469 ± 0.099	-2.139 ± 2.551
POAM	10	AgentAdj	0.207 ± 0.116	-2.142 ± 0.687
POAM	6	AgentAdj	0.207 ± 0.116	-2.142 ± 0.687
VDN	10	AgentAdj	0.425 ± 0.180	-0.117 ± 0.929
VDN	6	FireAdj	0.369 ± 0.168	-1.154 ± 1.066
QMIX	10	AgentAdj	0.298 ± 0.160	-1.934 ± 0.844
QMIX	6	FireAdj	0.202 ± 0.100	-2.259 ± 1.201

Table: Algorithm Summary per observability range of metrics for manually tuned (FireAdj) and bayesian (AgentAdj) reward

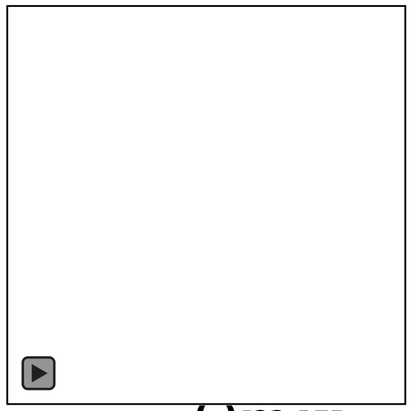


Unburnt trees for non-sharing (ns) policies evaluation



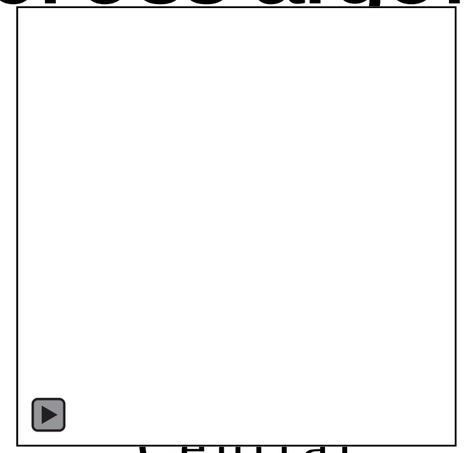
Non-burning trees Finetuning/Test evaluation

Learned behaviors across algorithms

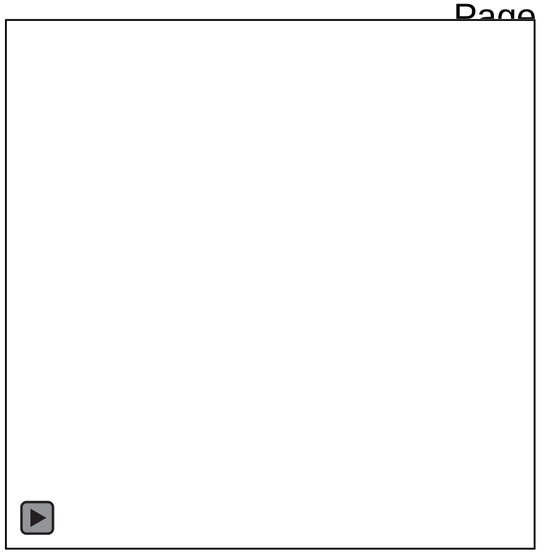


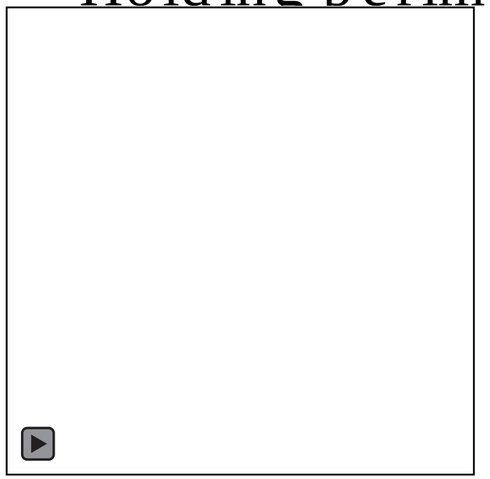
Qmix

• Holding perimeter,



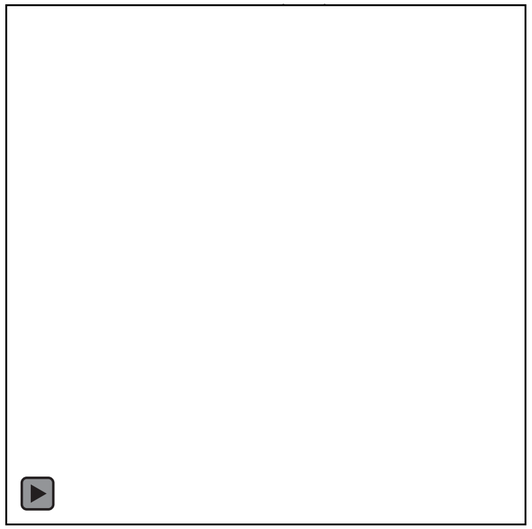
Central





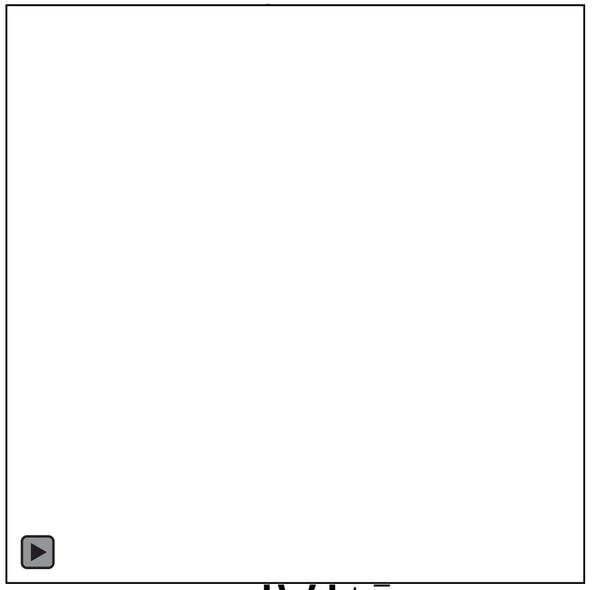
PPO with agent adjacent -

• Fire head tracking



Pareto equilibrium

• collision avoidance



IQL -

pairs formation

Robustness test: higher spread probability

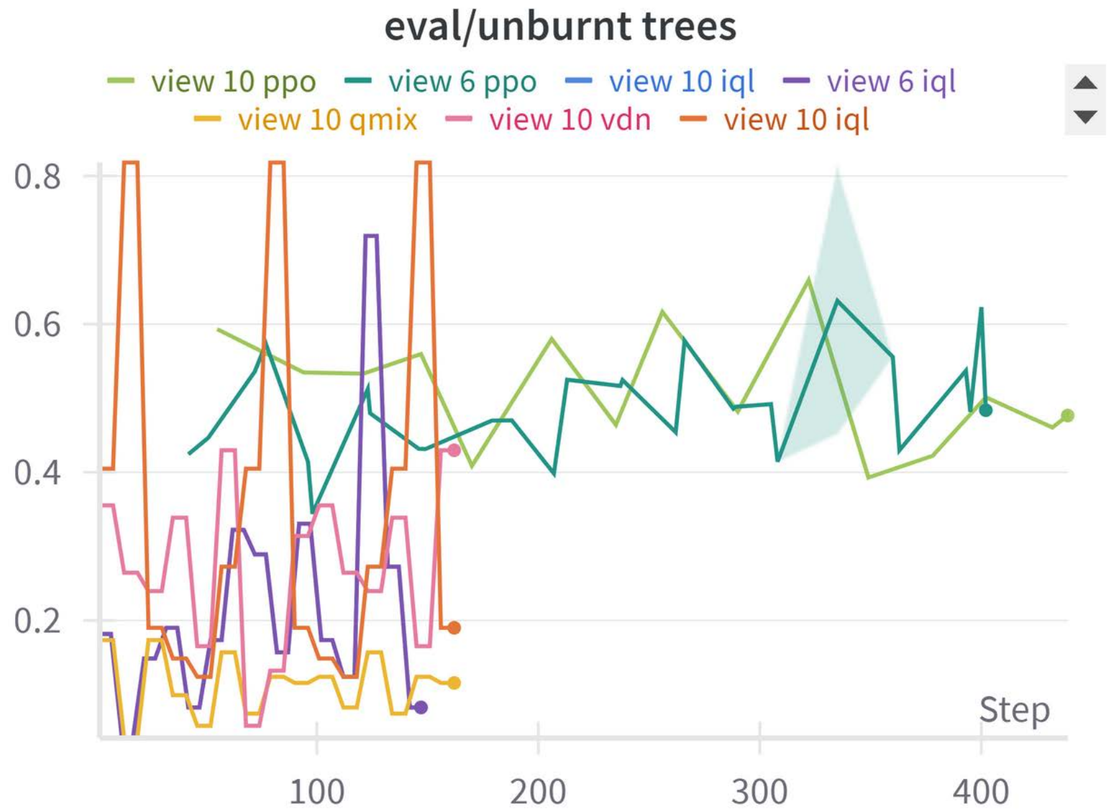
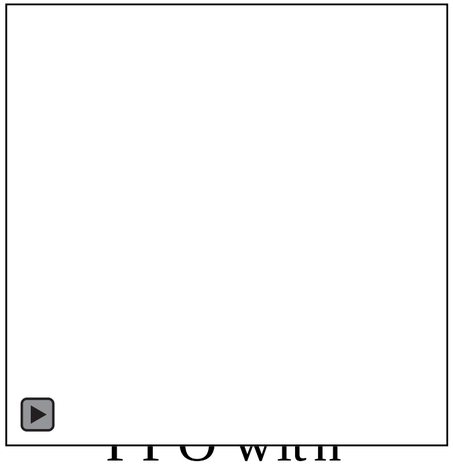
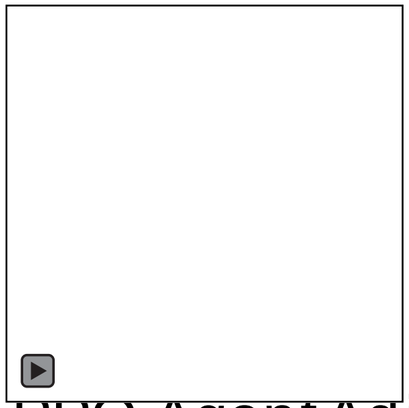


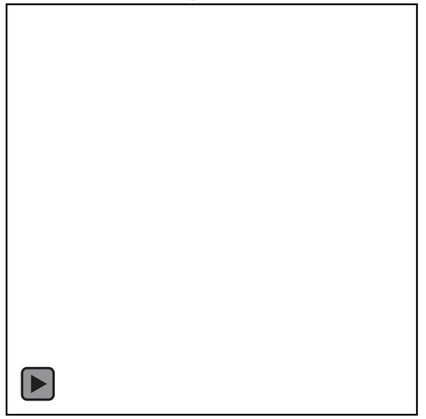
Figure. Non-burning trees on harder environment of PPO against, IQL, VDN



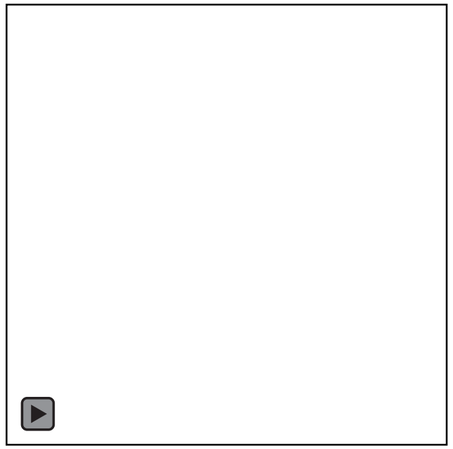
PPO with Fire Adj reward



PPO Agent Adj reward

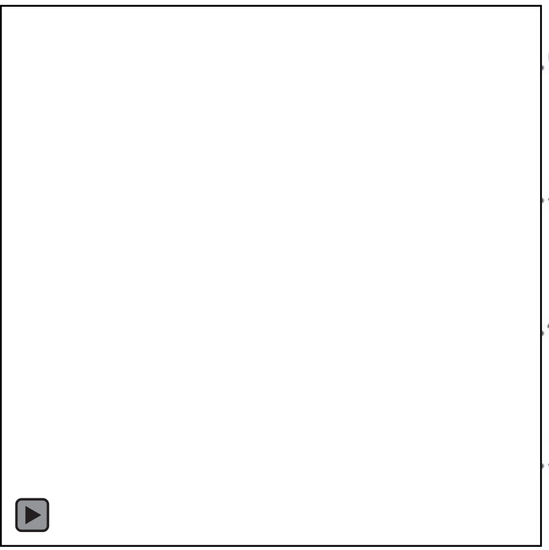


IQL Faster fires on Fire Adj

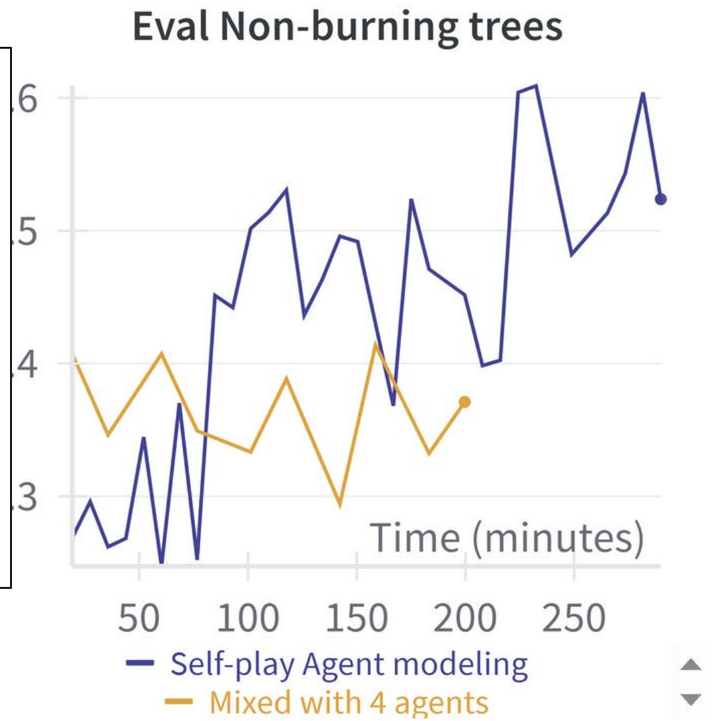


IQL with Fire Adj on **regular** fires

Agent modelling on unseen environment



POAM in few-shot training



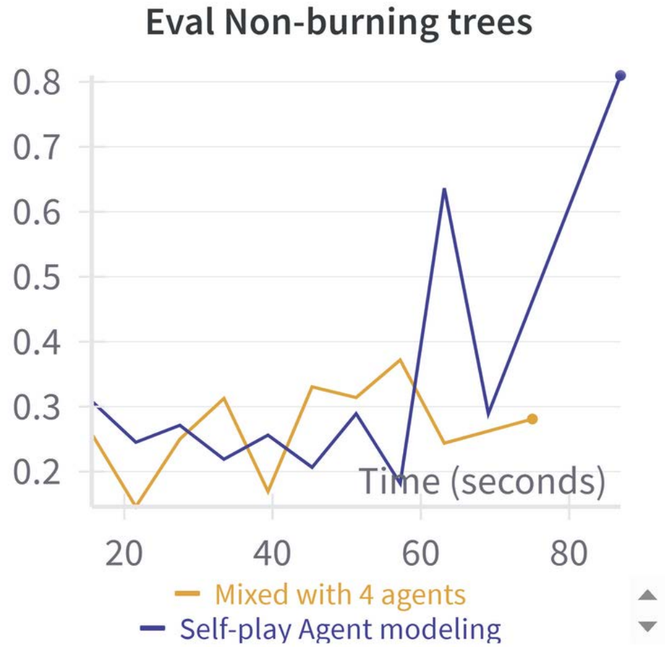
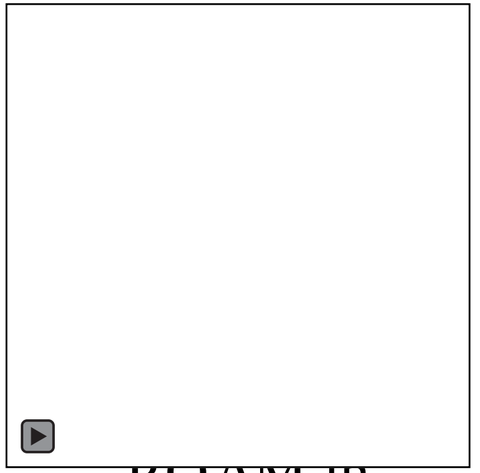
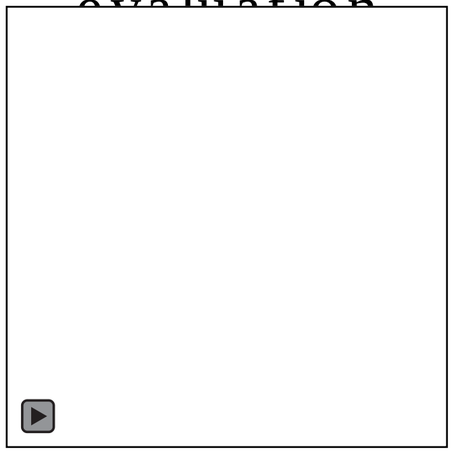


Figure: Metric in self-play/few-shot in training and evaluation

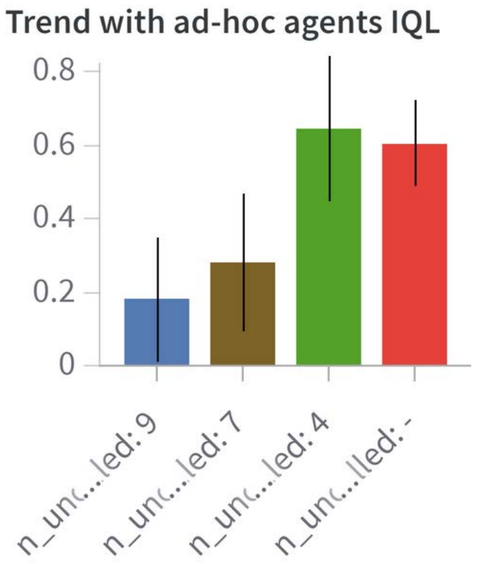


POAM in evaluation

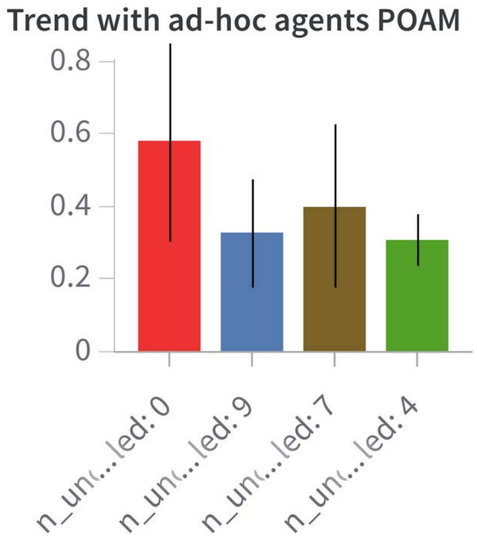


Comparison with Zero-shot IQL

Comparison of Agent Modelling trend with new teammates



IQL



POAM

Figures: metric with higher unseen teammate ratio compared to self-play(red)

Algorithm	IQL	POAM
Self-play	**0,628+0,109**	0,207+0,116
VDN+QMIX Ad-hoc	0,517+0,216	0,298+0,168
Fast fires	0.174+0,124	0.155 +0.053
Ad-hoc fast fires	0.281 0.033	0.179 + 0.086

Table: median non-burning trees with adaptation to 4 uncontrolled teammates

Key Insights

AgentAdj: Improves common baseline (0.52 0.233 vs 0.73 trees PPO) unstable with central critic.

FireAdj - Best reward alignment (0.512 vs 0.328 saved trees on IQL)
Fire head tracking - most general tactic (visual analysis)
PPO- robust to OOD shocks - (0.82 vs 0.51) -
parameter sharing crucial for speed

IQL - adaptive in zero shot (0.36 vs 0.28 POAM),
high variance with seed patterns (0.73 vs 0.28).

POAM - overfitting from agent bootsstrapping (0.49 vs 0.202),

Limitations of our work

1. **Computational constraints** leading to inconsistent logging
2. **Inefficient replay buffer design** limiting experiments with larger environments and longer horizons
3. **Static ad-hoc agent policies** preventing evaluation of concurrent adaptation and specialized roles
4. **Simplified fire model** and **Heuristic agent initialization**: limiting real-world applicability

Future work

- **Implement curriculum** learning for progressive reward complexity to address credit assignment challenges
- **Diversify agent pretraining** with action switching and human-proxy adaptation to prevent few-shot overfitting
- Integrate **higher-fidelity** environment models with image-based observations aligned with deployment requirements into PettingZoo catalogue (EPyMARL, BenchMARL etc.)

References

[1]Murray, Lucas, et al. "Advancing Forest Fire Prevention: Deep Reinforcement Learning for Effective Firebreak Placement." arXiv preprint arXiv:2404.08523 (2024).

[2]Xu Y, Li D, Ma H, Lin R, Zhang F. Modeling Forest Fire Spread Using Machine Learning-Based Cellular Automata in a GIS Environment. Forests. 2022; 13(12):1974. https://doi.org/10.3390/f13121974

[3]Oroojlooy, A., Hajinezhad, D. A review of cooperative multi-agent deep reinforcement learning. Appl Intell 53, 13677–13722 (2023). https://doi.org/10.1007/s10489-022-04105-y4.

[4]Wang, C., Rahman, M. A., Durugkar, I., Liebman, E., & Stone, P. (2024). N-agent ad hoc teamwork. Advances in Neural Information Processing Systems, 37, 111832-111862.

Thank you for attention!
AMA/Q&A

Full-observability Training metrics: Reward Selection

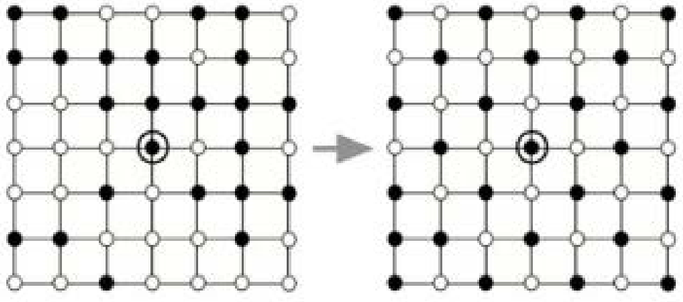
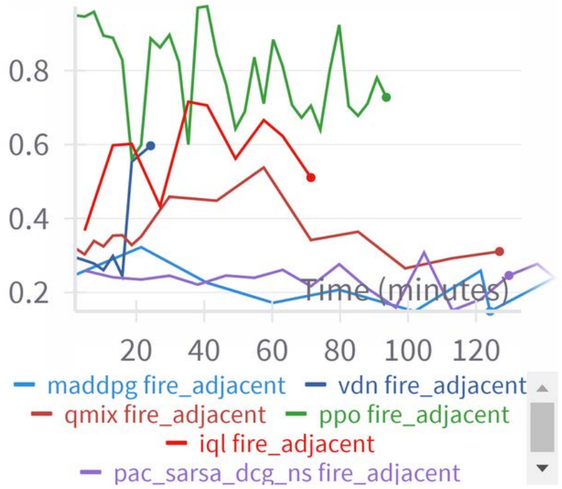


Figure. diagonal pattern for AgentAdj [5]

Per-reward Non-burning trees for Q-learning



Non-burning trees per-algorithm on manual training

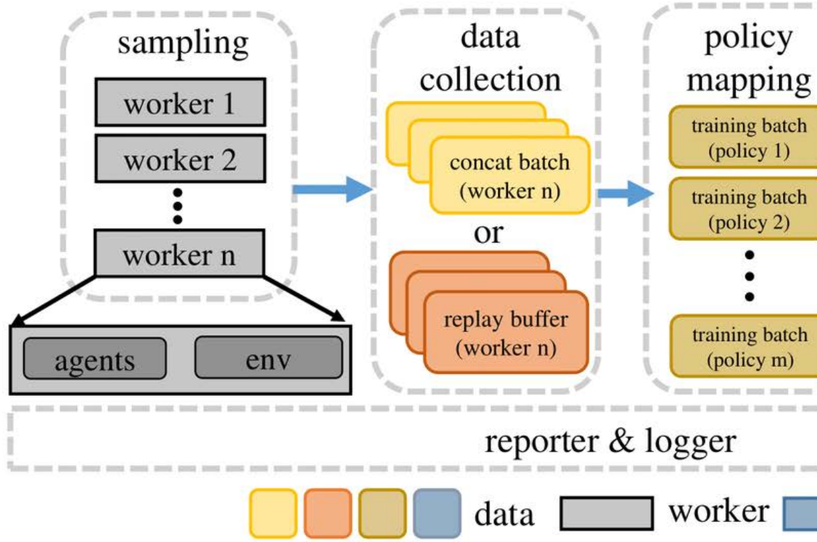
Algorithm	Reward type	Non-burning mean	std	Reward mean	std
ppo (team 10)	AgentAdjacent	0,725	0,106	0,381	0,413
ppo (team 15)	AgentAbove	0.600	0.248	0,311	0,352
ppo (team 10)	Default	0,524	0,283	-1,145	1,233
ppo (team 10)	FireAdj	0,530	0,281	0,591	0,196
iql (team 10)	AgentAdjacent	0,666	0,111	2,556	0,648
iql (team 10)	FireAdj	0,624	0,116	0,539	0,743
Maddpg (team 10)	Default	0,252	0,154	-2,494	0,822
Maddpg (team 15)	FireAdj	0,273	0,050	-1,939	0,308
Maddpg (team 10)	FireAdj	0,182	0,043	-2,922	0,308
Maddpg (team 5)	FireAdj	0,529	0,163	-0,247	0,885

Table: Aggregate FO metrics for PPO, IQL, Comparing baseline, scalability of DDPG.

Robustness: hard environment aggregated results

Algorithm	Reward Type	Trees Mean ± STD	Reward Mean ± STD	Episode Length
IQL	AgentAdj	0.174 ± 0.124	-0.208 ± 1.153	19.000
	FireAdj	0.301 ± 0.125	-0.506 ± 0.274	17.234
PPO	AgentAdj	0.319 ± 0.110	**0.334 ± 0.174**	18.200
	FireAdj	**0.512 ± 0.119**	-0.503 ± 0.146	-
VDN	AgentAdj	0.425 ± 0.180	-2.131 ± 0.642	17.400
	FireAdj	0.369 ± 0.168	-0.464 ± 0.293	18.267
QMIX	AgentAdj	0.298 ± 0.160	-1.456 ± 0.552	20.000
	FireAdj	0.202 ± 0.100	-1.197 ± 0.080	19.533
POAM	AgentAdj	0.207 ± 0.116	-2.011 ± 0.340	18.571

Table: Summarized evaluation for each algorithm, reward type



MARLlib/EpyMARL p
configuration [citat