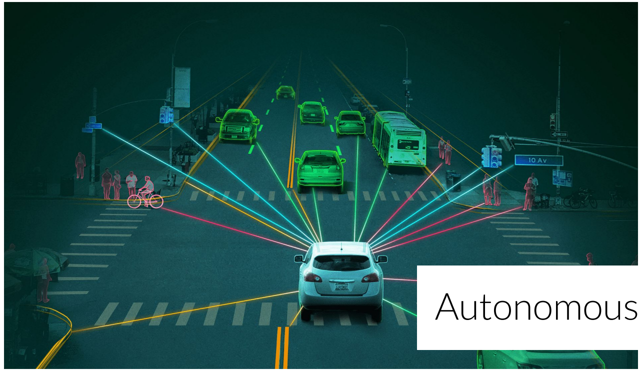
ELIGN: Expectation Alignment as a Multi-agent Intrinsic Reward

Zixian Ma, Rose Wang, Li Fei-Fei, Michael Bernstein, Ranjay Krishna
Stanford University
University of Washington
zixianma@cs.stanford.edu

Many real world applications are multi-agent systems.


Autonomous vehicles

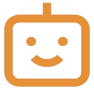

Traffic control


Resource management


Rescue robots

Many SOTA multi-agent algorithms make these assumptions.

Full observability

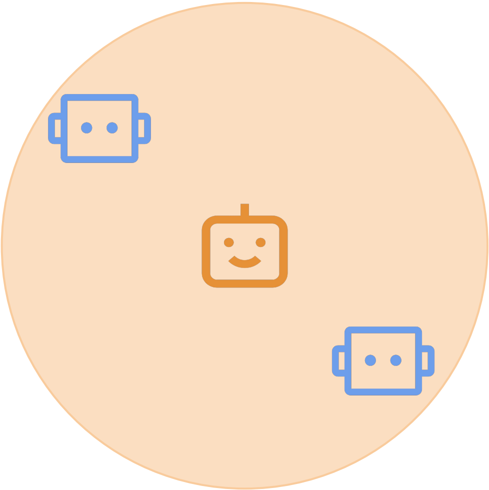


 Agent

Iqbal and Sha. "Actor-attention-critic for multi-agent reinforcement learning." International conference on machine learning. PMLR, 2019.
Liu et al. "PIC: permutation invariant critic for multi-agent deep reinforcement learning." Conference on Robot Learning. PMLR, 2020.

Many SOTA multi-agent algorithms make these assumptions.

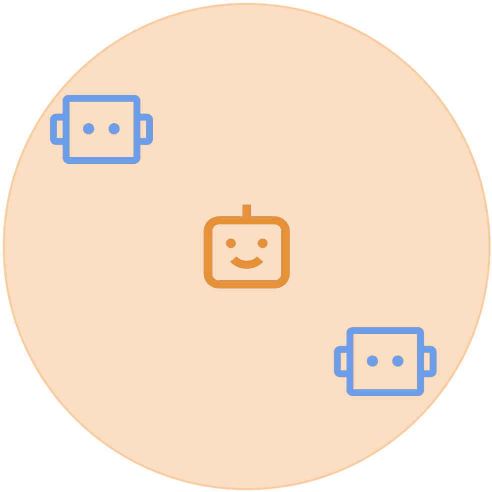
Full observability



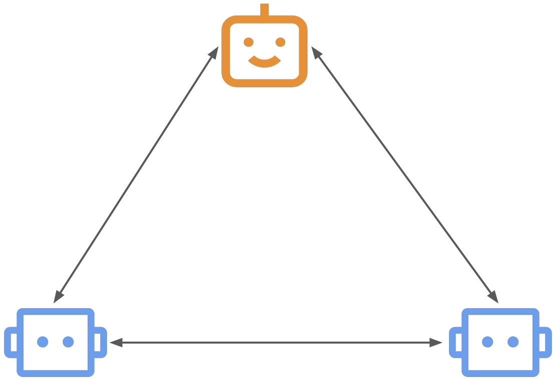
Iqbal and Sha. "Actor-attention-critic for multi-agent reinforcement learning." International conference on machine learning. PMLR, 2019.
Liu et al. "PIC: permutation invariant critic for multi-agent deep reinforcement learning." Conference on Robot Learning. PMLR, 2020.

Many SOTA multi-agent algorithms make these assumptions.

Full observability

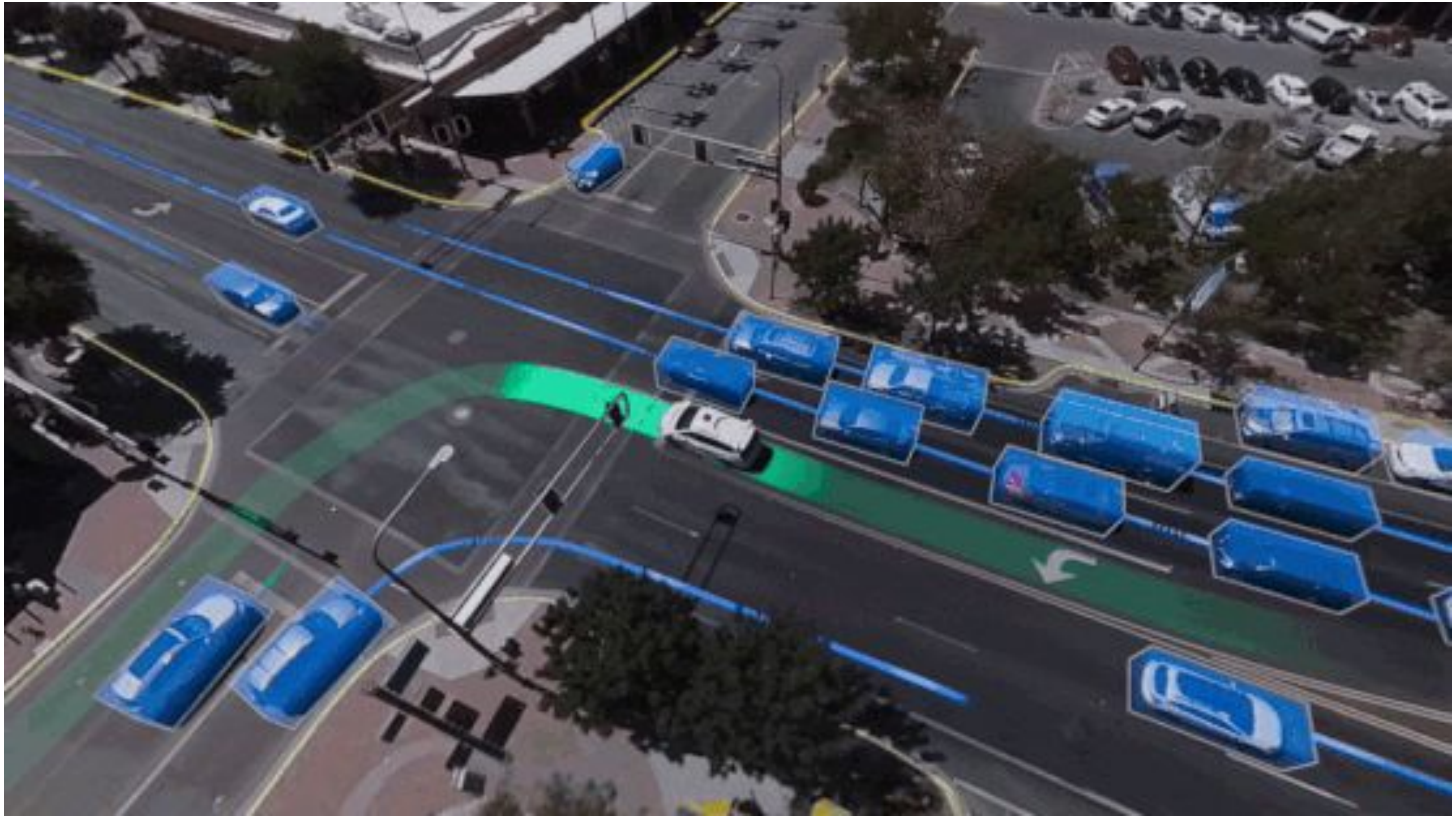


Centralized algorithm

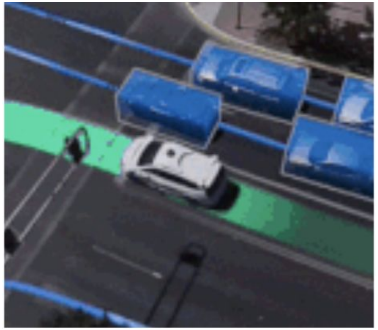


Iqbal and Sha. "Actor-attention-critic for multi-agent reinforcement learning." International conference on machine learning. PMLR, 2019.
Liu et al. "PIC: permutation invariant critic for multi-agent deep reinforcement learning." Conference on Robot Learning. PMLR, 2020.

Full observability and centralized algorithms are not ecologically valid.

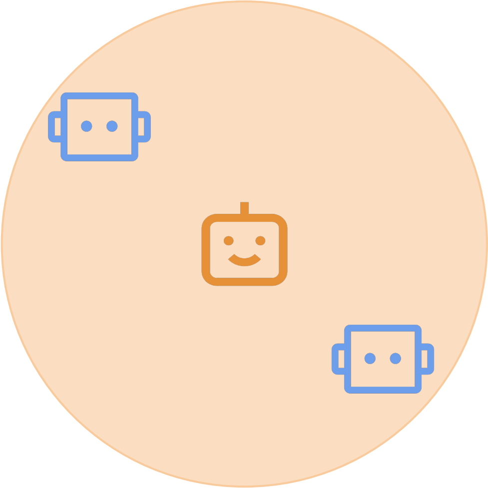


Full observability and centralized algorithms are not ecologically valid.

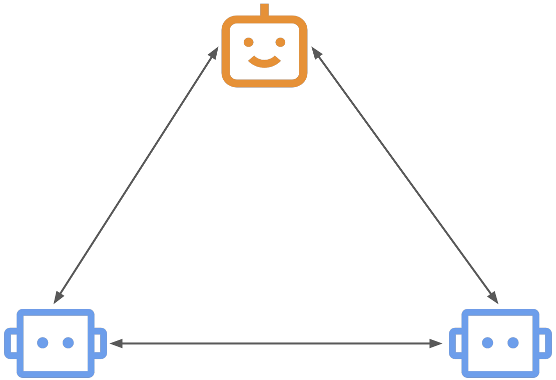


Multi-agent performance struggles without prior assumptions.

Full observability



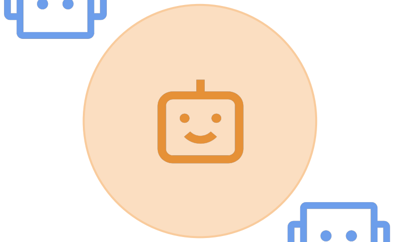
Centralized algorithm



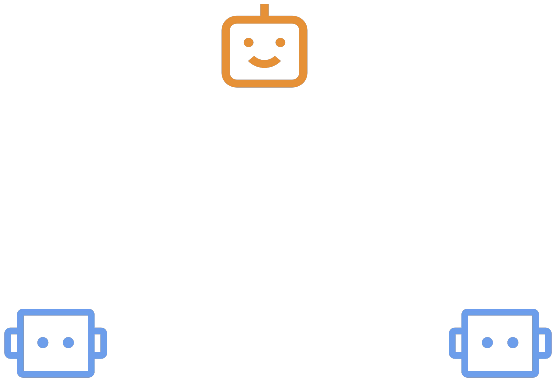
Iqbal and Sha. "Actor-attention-critic for multi-agent reinforcement learning." International conference on machine learning. PMLR, 2019.
Liu et al. "PIC: permutation invariant critic for multi-agent deep reinforcement learning." Conference on Robot Learning. PMLR, 2020.

Multi-agent performance struggles without prior assumptions.

Partial observability

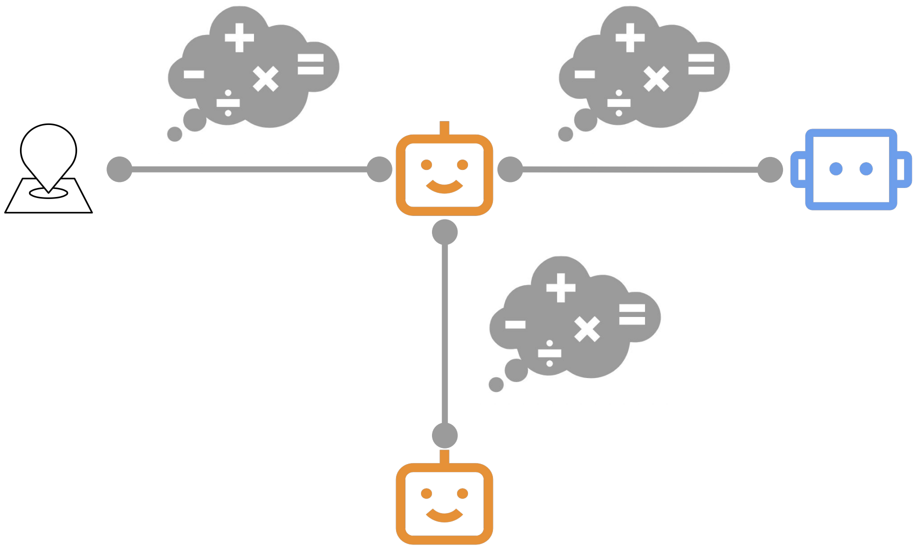


Decentralized algorithm



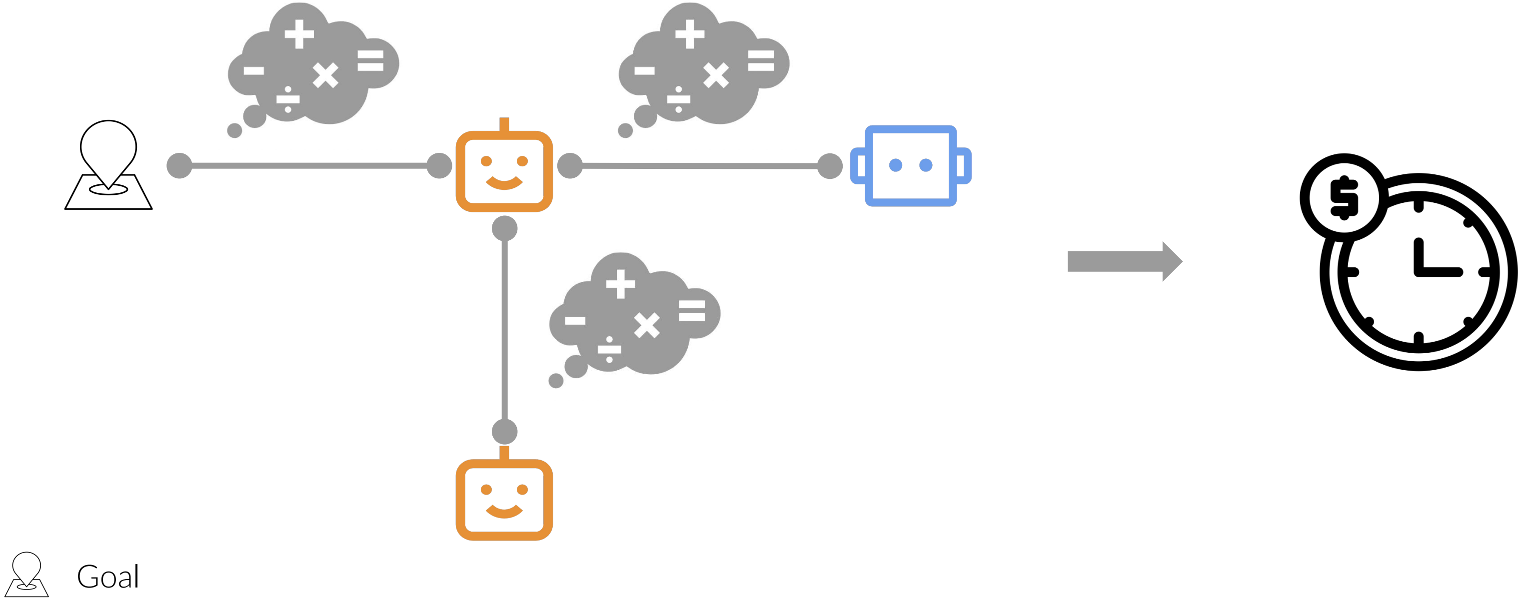
Iqbal and Sha. "Actor-attention-critic for multi-agent reinforcement learning." International conference on machine learning. PMLR, 2019.
Liu et al. "PIC: permutation invariant critic for multi-agent deep reinforcement learning." Conference on Robot Learning. PMLR, 2020.

One approach is designing task-specific dense rewards



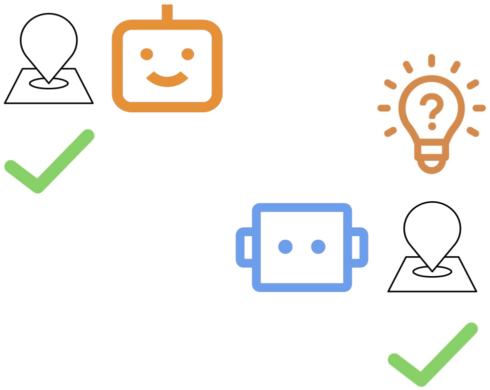
Iqbal and Sha. "Actor-attention-critic for multi-agent reinforcement learning." International conference on machine learning. PMLR, 2019.

One approach is designing task-specific dense rewards, but it's expensive.



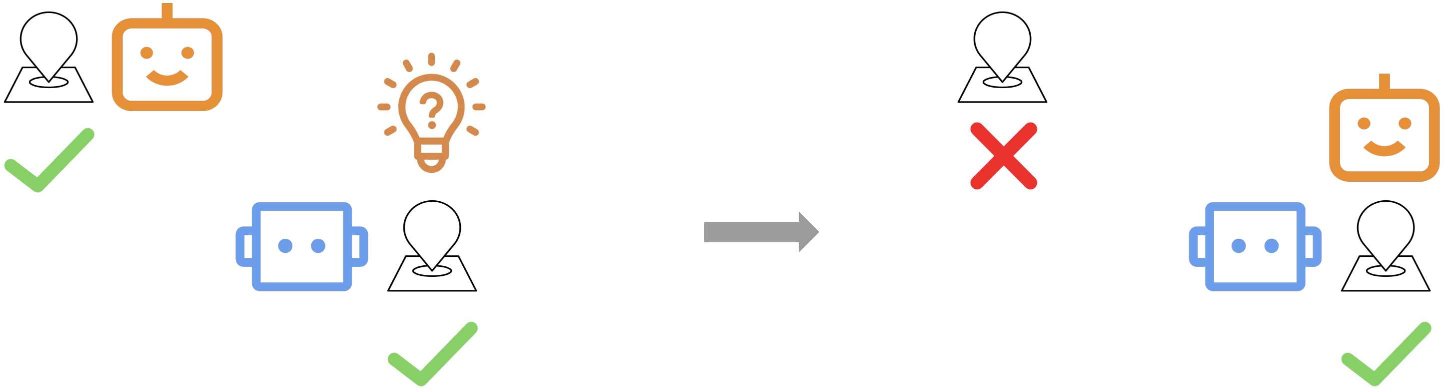
Iqbal and Sha. "Actor-attention-critic for multi-agent reinforcement learning." International conference on machine learning. PMLR, 2019.

Another is adding curiosity-based intrinsic rewards that encourage exploration



Chentanez et al. “Intrinsically motivated reinforcement learning. Advances in neural information processing systems.” 2004

Another is adding curiosity-based intrinsic rewards that encourage exploration, but exploration doesn't solve coordination.



Chentanez et al. "Intrinsically motivated reinforcement learning. Advances in neural information processing systems." 2004

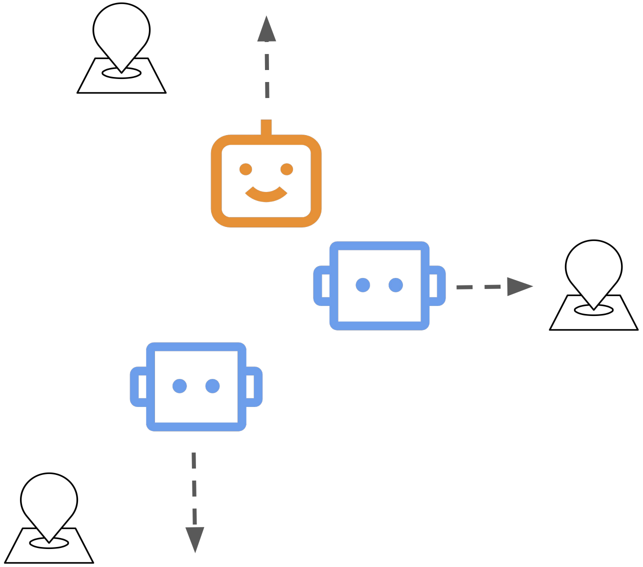
Self-organization: individual animals coordinate by ***aligning*** their behaviors within a local context.

Emergent alignment in fish

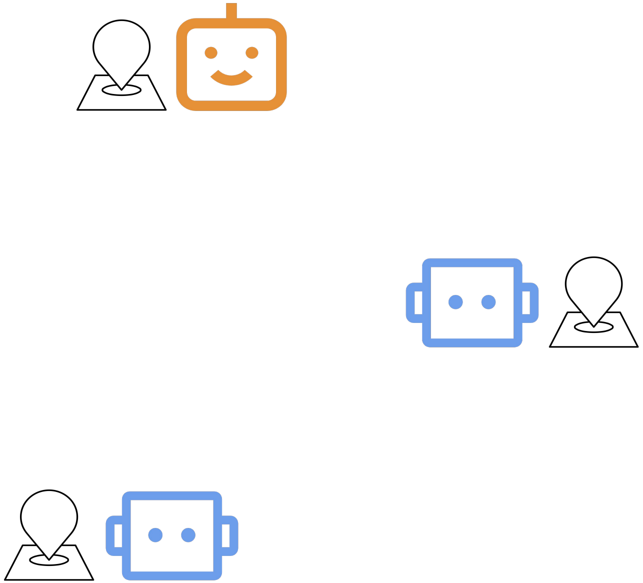
Couzin et al. Collective memory and spatial sorting in animal groups. Journal of theoretical biology. 2002.
Herbert-Read JE. Understanding how animal groups achieve coordinated movement. Journal of Experimental Biology. 2016.

We introduce ELIGN - Expectation Alignment - as a multi-agent intrinsic reward.

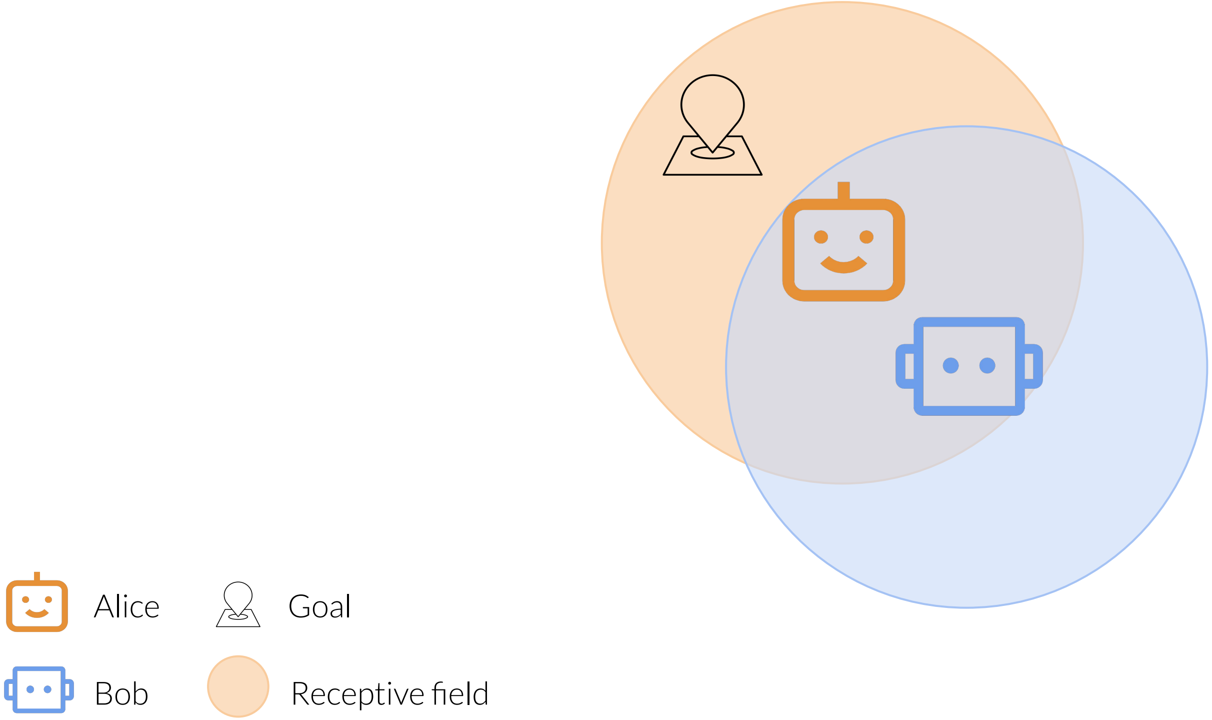
Cooperative navigation



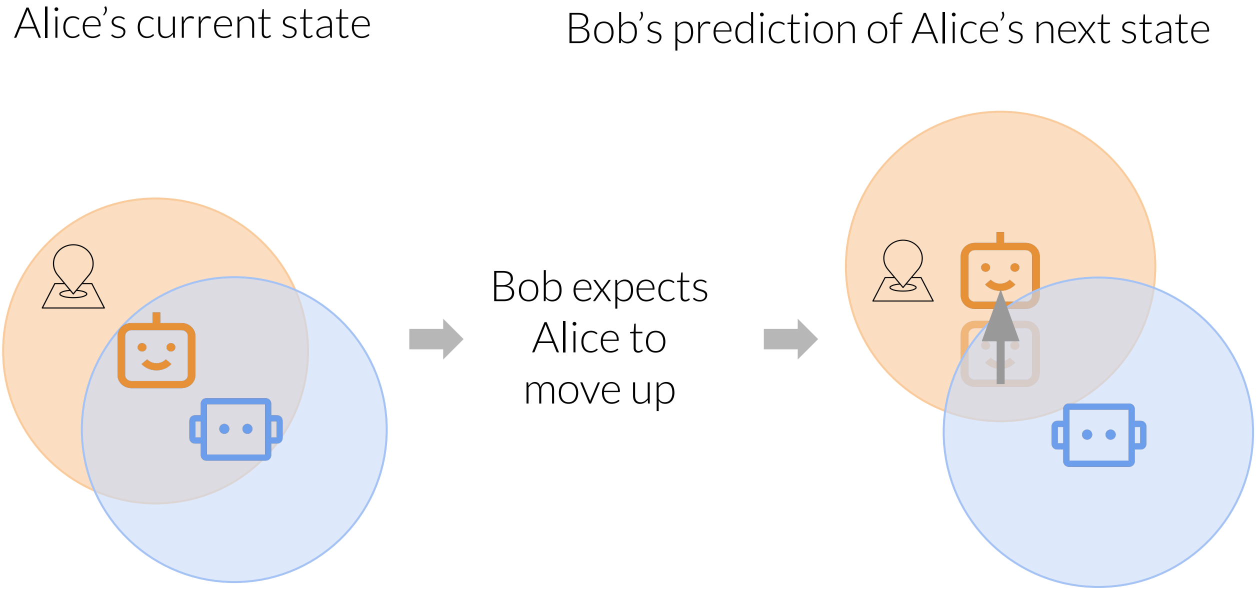
Cooperative navigation



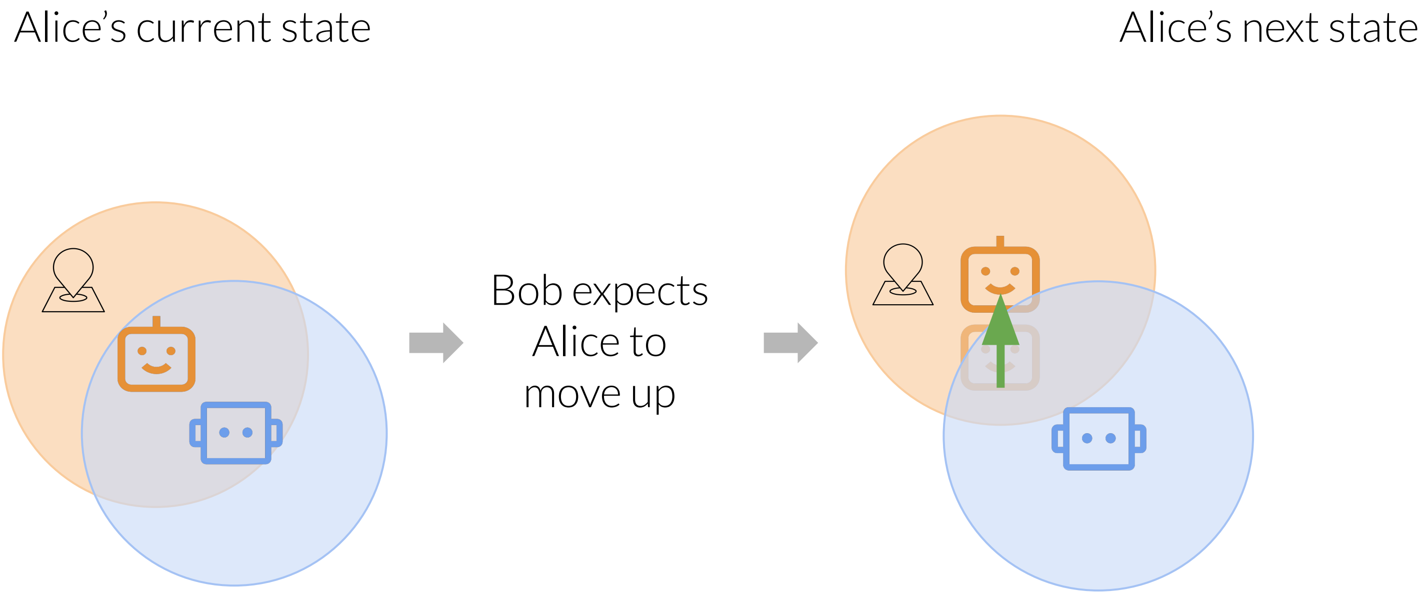
ELIGN in cooperative navigation



ELIGN in cooperative navigation

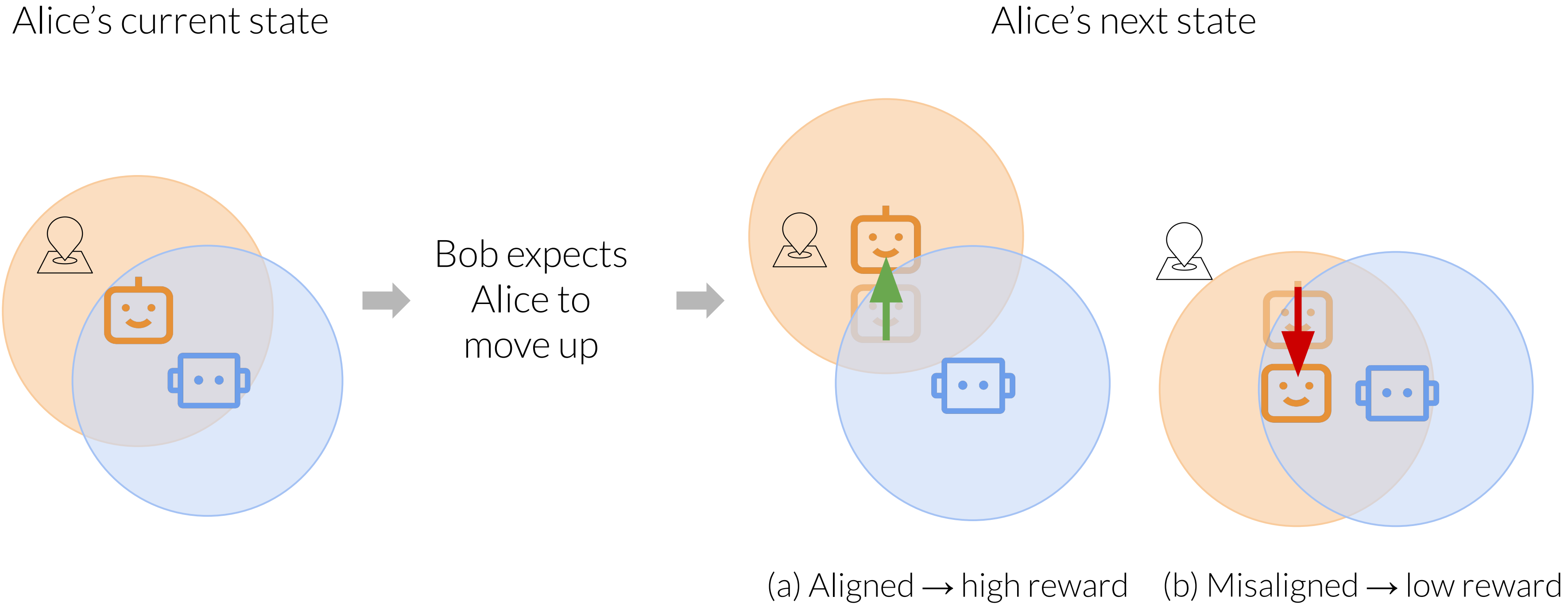


ELIGN in cooperative navigation



(a) Aligned ⟶ high reward

ELIGN in cooperative navigation



ELIGN intrinsic reward

Ideal form: $r_{\text{in}}(o_i, a_i) = -\frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \|o_i' - f_{\theta_j}(o_i, a_i)\|$

ELIGN intrinsic reward

Ideal form: $r_{\text{in}}(o_i, a_i) = -\frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \|o'_i - f_{\theta_j}(o_i, a_i)\|$

Decentralized ELIGNteam: $r_{\text{in}}(o_i, a_i) = -\frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \|o'_{i \cap j} - f_{\theta_i}(o_{i \cap j}, a_i)\|$

ELIGN intrinsic reward

Ideal form: $r_{\text{in}}(o_i, a_i) = -\frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \|o'_i - f_{\theta_j}(o_i, a_i)\|$

Decentralized ELIGNteam: $r_{\text{in}}(o_i, a_i) = -\frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \|o'_{i \cap j} - f_{\theta_i}(o_{i \cap j}, a_i)\|$

ELIGNadv: $r_{\text{in}(o_i, a_i)} = +\frac{1}{|\mathcal{N}_{\text{adv}}(i)|} \sum_{k \in \mathcal{N}_{\text{adv}}(i)} \|o'_{i \cap k} - f_{\theta_i}(o_{i \cap k}, a_i))\|$

ELIGN intrinsic reward

Ideal form: $r_{\text{in}}(o_i, a_i) = -\frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \|o'_i - f_{\theta_j}(o_i, a_i)\|$

Decentralized ELIGNteam: $r_{\text{in}}(o_i, a_i) = -\frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \|o'_{i \cap j} - f_{\theta_i}(o_{i \cap j}, a_i)\|$

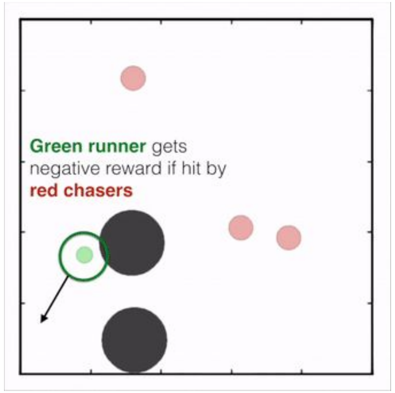
ELIGNadv: $r_{\text{in}(o_i, a_i)} = +\frac{1}{|\mathcal{N}_{\text{adv}}(i)|} \sum_{k \in \mathcal{N}_{\text{adv}}(i)} \|o'_{i \cap k} - f_{\theta_i}(o_{i \cap k}, a_i))\|$

ELIGNself: $r_{\text{in}(o_i, a_i)} = -\|o'_i - f_{\theta_i}(o_i, a_i)\|$

We include these two environments in our experiments.

Multi-agent particle environment (MAP)

- 2D
- Continuous states
- 5 actions



Google research football

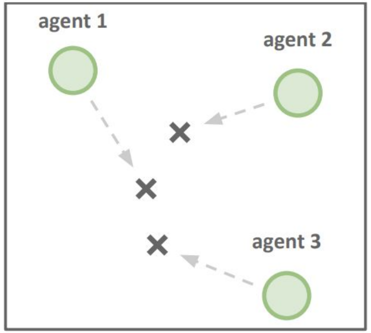
- 3D
- Continuous states
- 10 actions

Lowe et al. "Multi-agent actor-critic for mixed cooperative-competitive environments." Advances in neural information processing systems. 2017.
Kurach et a. "Google research football: A novel reinforcement learning environment." In Proceedings of the AAAI Conference on Artificial Intelligence 2020

We train and evaluate our method across cooperative and competitive tasks.

Cooperative

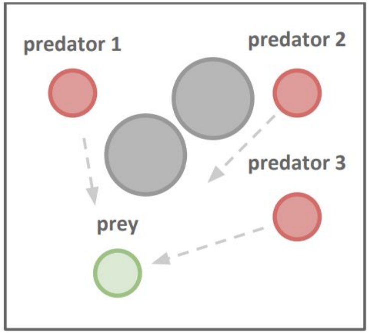
- Cooperative navigation
- Heterogeneous navigation



Cooperative Navigation

Competitive

- Keep-away
- Physical deception
- Predator-prey
- Academy 3vs1 with keeper (football)



Predator-prey

Lowe et al. “Multi-agent actor-critic for mixed cooperative-competitive environments.” Advances in neural information processing systems. 2017.
Kurach et a. “Google research football: A novel reinforcement learning environment.” In Proceedings of the AAAI Conference on Artificial Intelligence 2020

We follow these training setups.

- We use the decentralized Soft Actor-Critic for policy optimization.
- We train all algorithms across 5 random seeds
  - until convergence* in MAP;
  - for 5M timesteps in the Google football environment.

*the best test episode reward remains the same for 100 epochs (i.e. 400K episodes of 25 timesteps)

Haarnoja et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. InInternational conference on machine learning PMLR 2018.

We evaluate our method against three baselines on both test episode reward and task-specific metrics.

- We evaluate ELIGN against three baseline rewards:
  - SPARSE
  - SPARSE + CURIOsity intrinsic rewards
    - CURIOself and CURIOteam
- Our evaluation metrics include:
  - average episode reward across 1K test episodes of 25 timesteps
  - task-specific metrics
    - agent-goal occupancy
    - agent-adversary collision count in the Predator-prey task
- We report the mean value of each metric and its standard error across 5 random seeds.

Results in partially observable environments with decentralized training

In cooperative tasks, ELIGN$_{self,team}$ outperform SPARSE and both curiosity-based intrinsic rewards on test episode rewards.

Task	Cooperative nav. 3v0	Heterogenous nav. 4v0
SPARSE[1]	139.07 ± 13.63	284.42 ± 12.83
CURIOself[2]	133.93 ± 7.66	286.22 ± 9.97
CURIOteam[3]	125.42 ± 11.95	262.28 ± 22.59
ELIGNself	**155.88 ± 5.11**	292.34 ± 9.24
ELIGNteam	141.04 ± 8.04	**311.67 ± 10.88**

[1] Lowe et al. "Multi-agent actor-critic for mixed cooperative-competitive environments." Advances in neural information processing systems. 2017.
[2] Stadie et al. Incentivizing Exploration In Reinforcement Learning With Deep Predictive Models. CoRR 2015.
[3] Iqbal and Sha. Coordinated exploration via intrinsic rewards for multi-agent reinforcement learning. 2019.

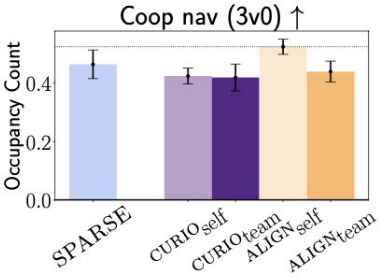
In competitive tasks, ELIGNadv achieves the best performance except for Physical deception, where ELIGNteam is the best.

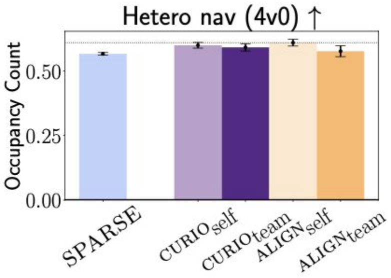
Task	Phy decep. 2v1	Predator-prey 2v2	Keep-away 2v2	Football 3v1 w/ keeper
SPARSE[1]	93.60 ± 8.61	−4.72 ± 2.4	4.58 ± 3.27	0.020 ± 0.001
CURIOself[2]	68.80 ± 7.93	−6.50 ± 2.18	11.88 ± 2.88	0.024 ± 0.004
CURIOteam[3]	85.31 ± 11.93	−3.57 ± 1.75	9.54 ± 5.04	0.021 ± 0.002
ELIGNself	69.91 ± 4.51	−7.58 ± 2.55	12.84 ± 4.29	0.003 ± 0.018
ELIGNteam	**101.72 ± 6.31**	−7.69 ± 2.69	2.96 ± 4.03	0.022 ± 0.001
ELIGNadv	92.20 ± 4.23	**−2.51 ± 1.70**	**19.46 ± 5.05**	**0.025 ± 0.001**

[1] Lowe et al. "Multi-agent actor-critic for mixed cooperative-competitive environments." Advances in neural information processing systems. 2017.
[2] Stadie et al. Incentivizing Exploration In Reinforcement Learning With Deep Predictive Models. CoRR 2015.
[3] Iqbal and Sha. Coordinated exploration via intrinsic rewards for multi-agent reinforcement learning. 2019.

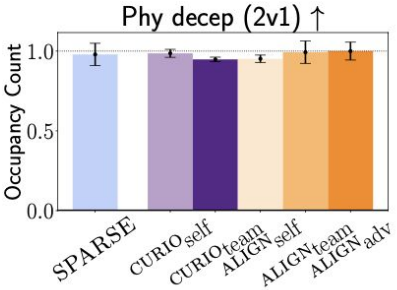
On task-specific metrics, ELIGNself and ELIGNadv perform the best in cooperative and competitive tasks respectively.

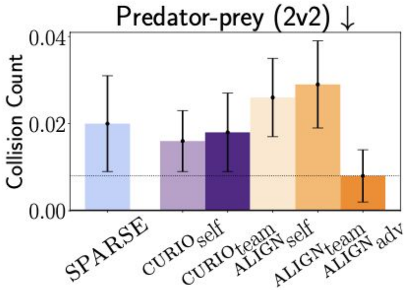
Cooperative

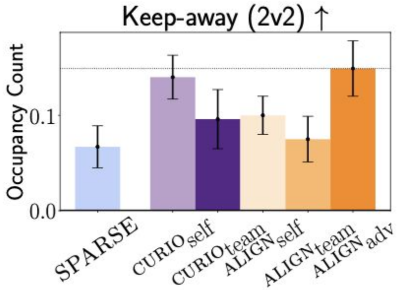




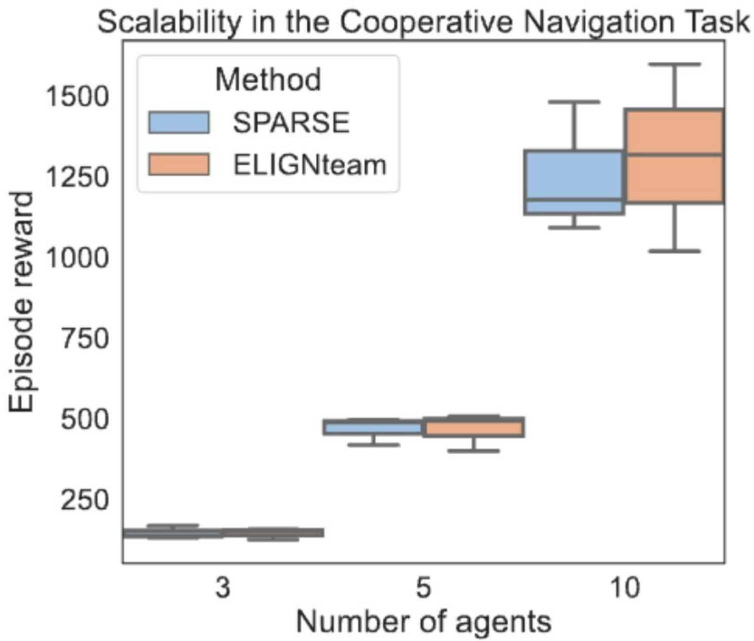
Competitive





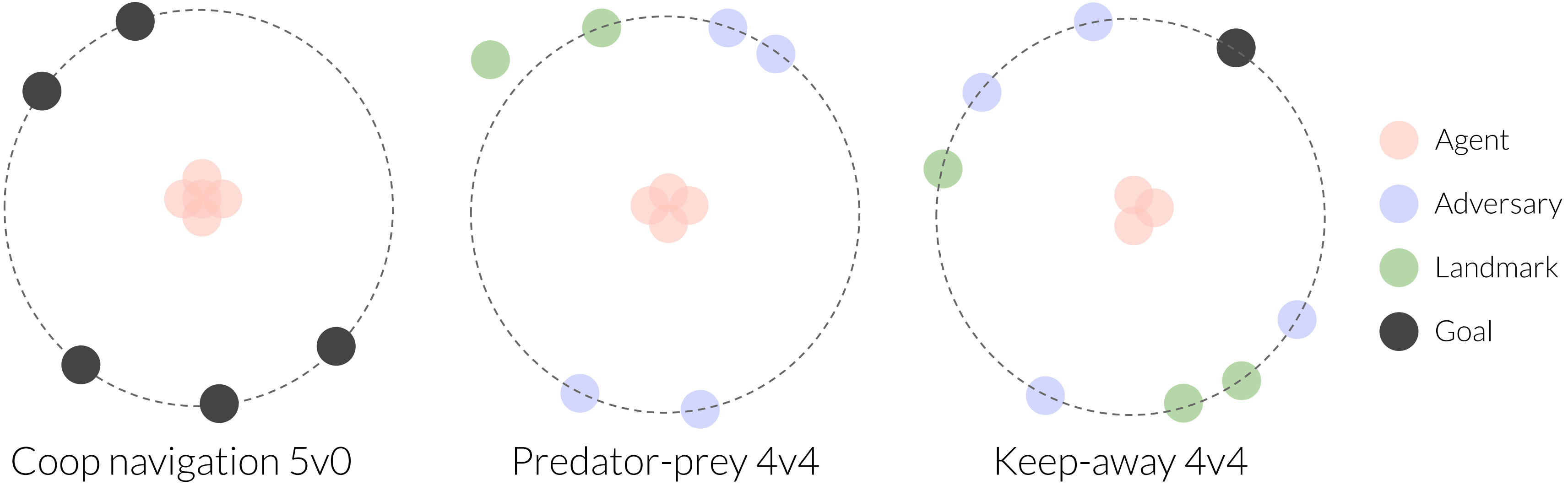


ELIGNteam scales well even when the number of agents increases to 10 in Cooperative navigation.

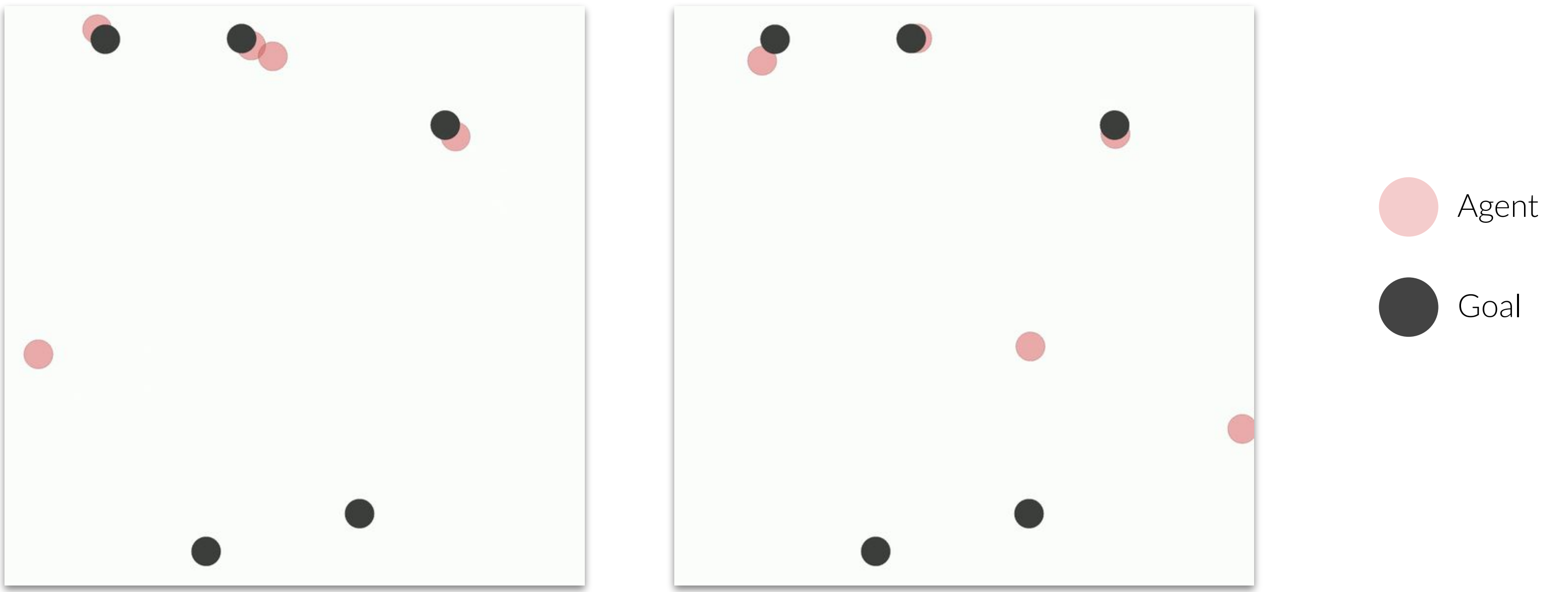


Investigating how Expectation Alignment helps

We initialize agents in states without an optimal sub-task allocation, necessitating symmetry-breaking.



Expectation alignment helps agents divide tasks.



With SPARSE only, agents cluster and cover few goals.

With ELIGN, agents **spread out to cover more goals.**

Conclusions

- Inspired by the self-organizing principle in Zoology, we formulate Expectation Alignment - ELIGN - as an multi-agent intrinsic reward.
- ELIGN rewards agents when they act predictably to their teammates and unpredictably to their adversaries.
- ELIGN improves multi-agent performance across cooperative and competitive tasks in the MAP and Google football environments.
- It also scales well, and helps agents break symmetries.

ELIGN: Expectation Alignment as a Multi-agent Intrinsic Reward

ELIGN is a simple, task-agnostic, and self-supervised multi-agent intrinsic reward, and it can be added to any multi-agent algorithm.

For more details, please refer to our paper from the QR code.

Code: https://github.com/StanfordVL/alignment

Contact: zixianma@cs.stanford.edu.