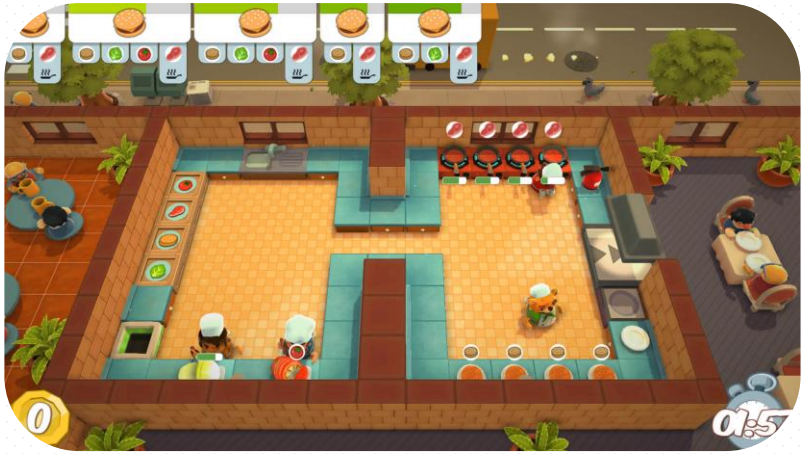
Generating Diverse Cooperative Agents by Learning Incompatible Policies (LIPO)

ICLR 2023

Rujikorn Charakorn, Poramate Manoonpong, Nat Dilokthanakul

Introduction

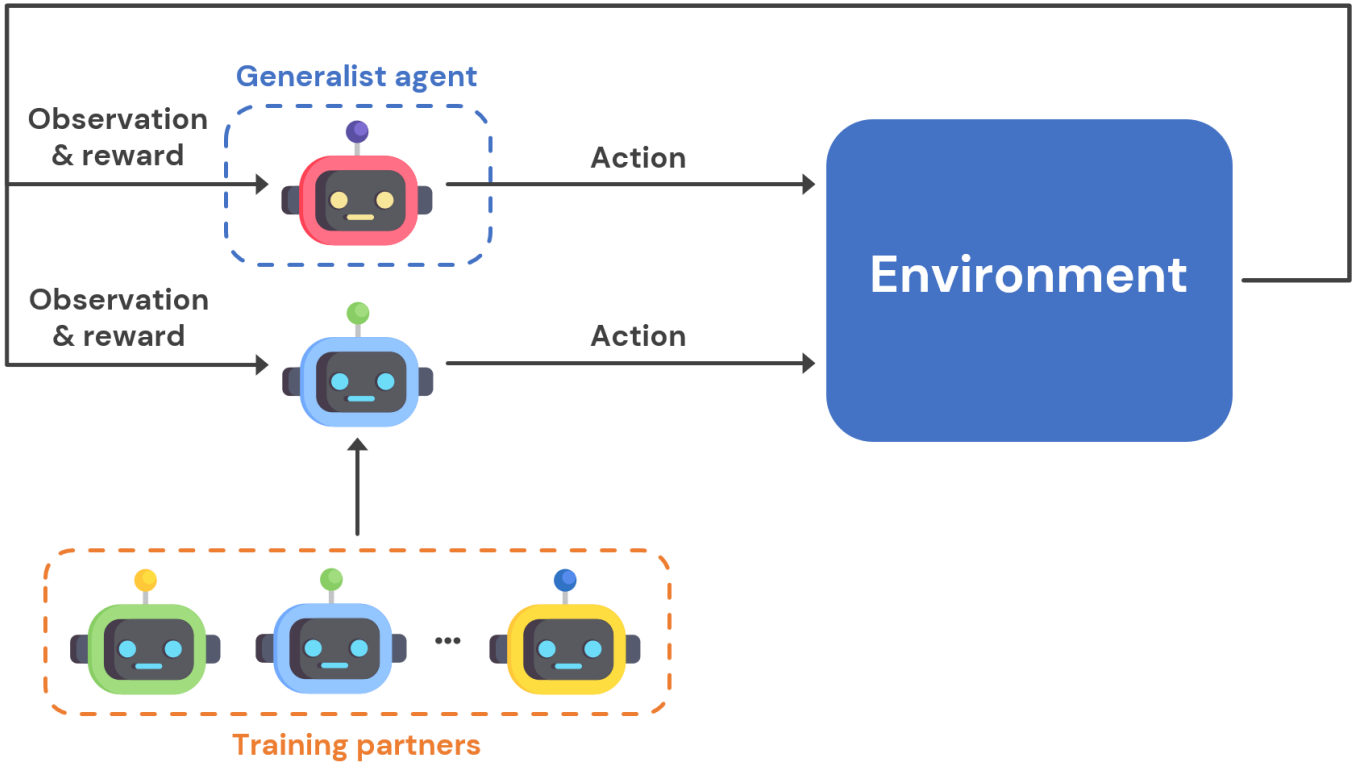
The ability to **work with unseen agents** is crucial for real-world deployment of AI systems



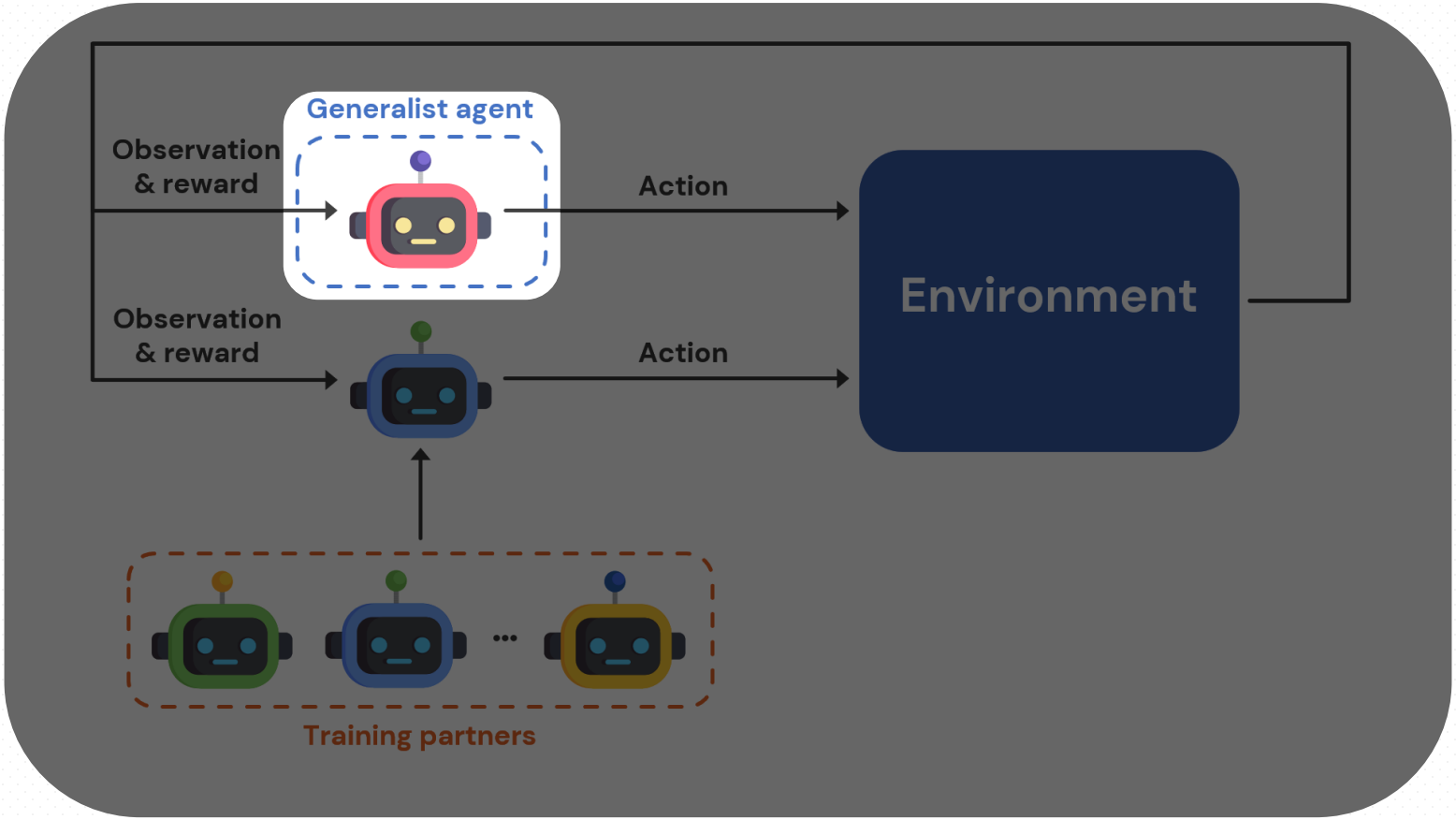
Overcooked 2 (developed by Team17)

RoboCup (photo by NSTDA)

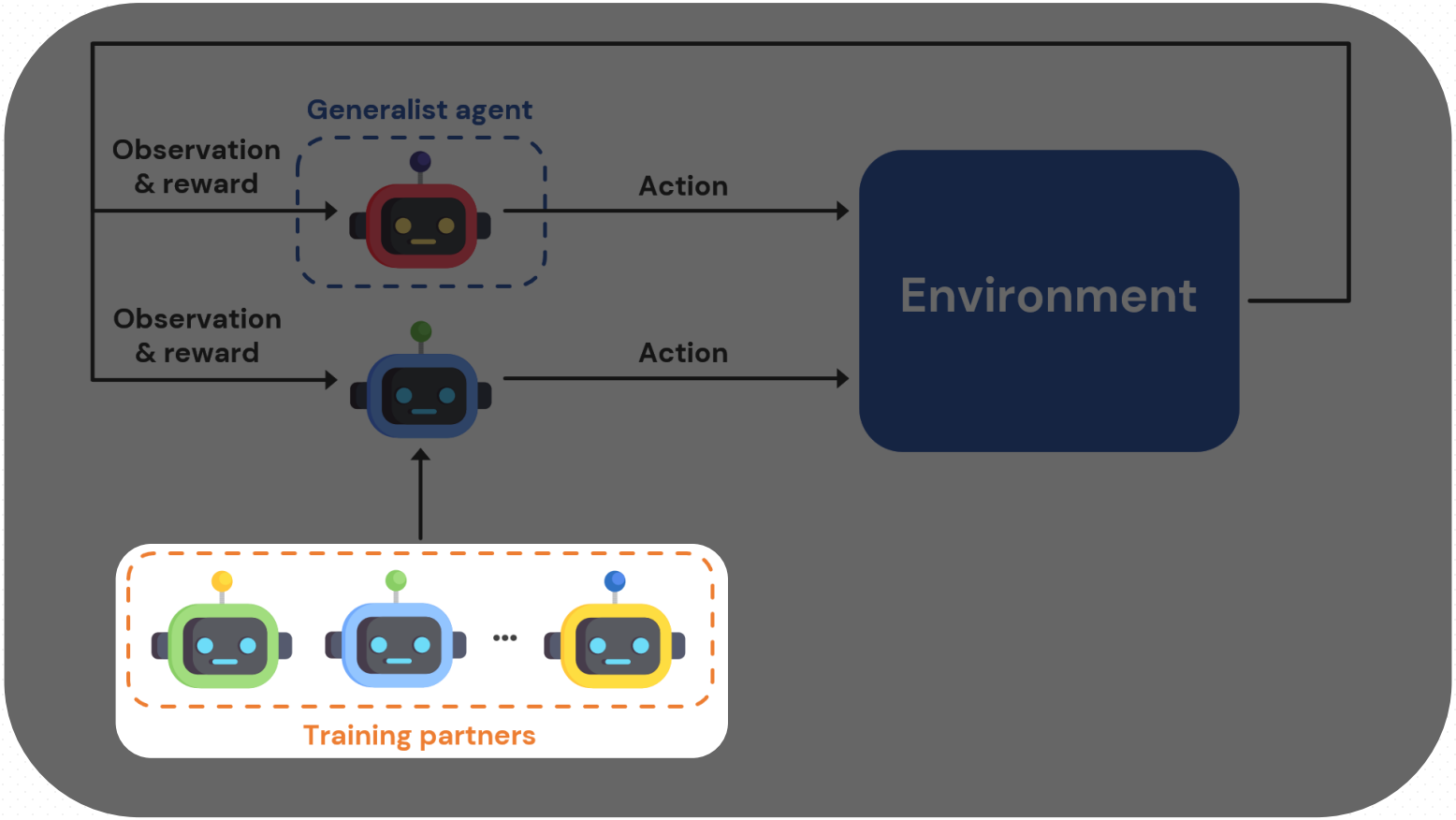
Introduction



Introduction



Introduction



Motivation

Widely used methods to generate partners are

- **Hand-crafted** policies

- Multiple runs of **self-play**

- Explicitly diversify the agents by **changing state-action distribution**

Motivation

Widely used methods to generate partners are

- **Hand-crafted** policies
  - **Not scalable**
- Multiple runs of **self-play**

- Explicitly diversify the agents by **changing state-action distribution**

Motivation

Widely used methods to generate partners are

- **Hand-crafted** policies
  - **Not scalable**
- Multiple runs of **self-play**
  - **Not guarantee to produce diverse agents**
- Explicitly diversify the agents by **changing state-action distribution**

Motivation

Widely used methods to generate partners are

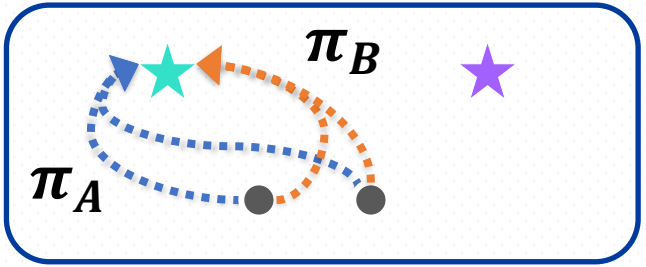
- **Hand-crafted** policies
  - **Not scalable**
- Multiple runs of **self-play**
  - **Not guarantee to produce diverse agents**
- Explicitly diversify the agents by **changing state-action distribution**
  - **Might not lead to high-level behavioral difference**

Motivation

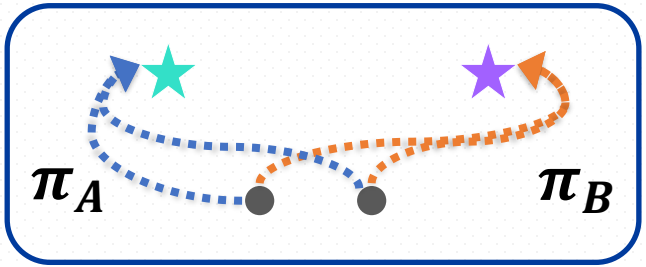
- Explicitly diversify the agents by **changing state-action distribution**
  - **Might not lead to high-level behavioral difference**

Motivation

- Explicitly diversify the agents by **changing state-action distribution**
  - **Might not lead to high-level behavioral difference**



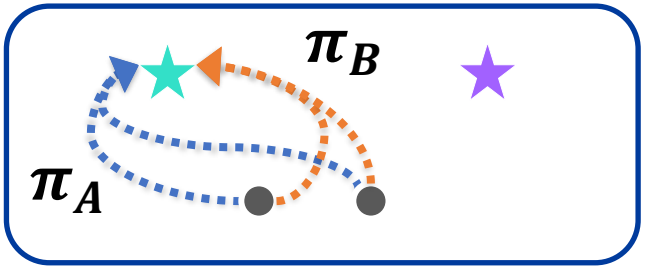
π_A and π_B have different state-action distributions **but** share the same high-level objective



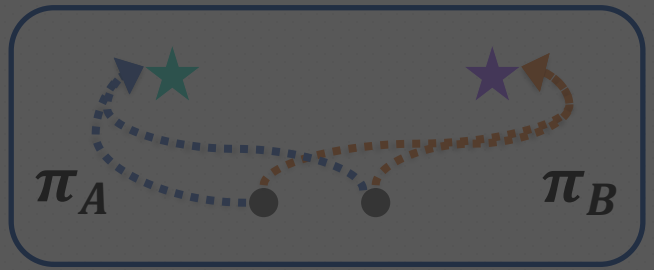
π_A and π_B have different state-action distributions **and** high-level objectives

Motivation

- Explicitly diversify the agents by **changing state-action distribution**
  - **Might not lead to high-level behavioral difference**



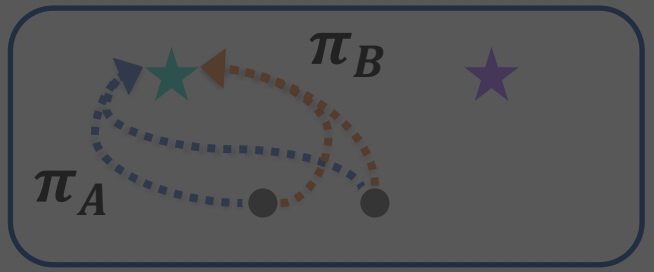
π_A and π_B have different state-action distributions **but** share the same high-level objective



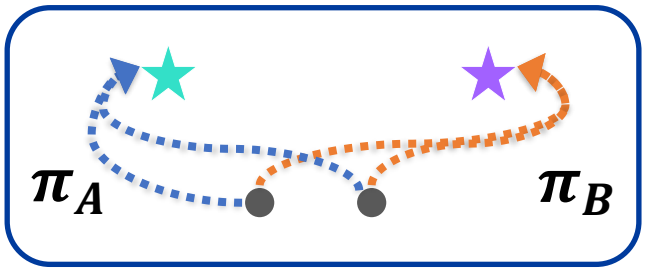
π_A and π_B have different state-action distributions **and** high-level objectives

Motivation

- Explicitly diversify the agents by **changing state-action distribution**
  - **Might not lead to high-level behavioral difference**

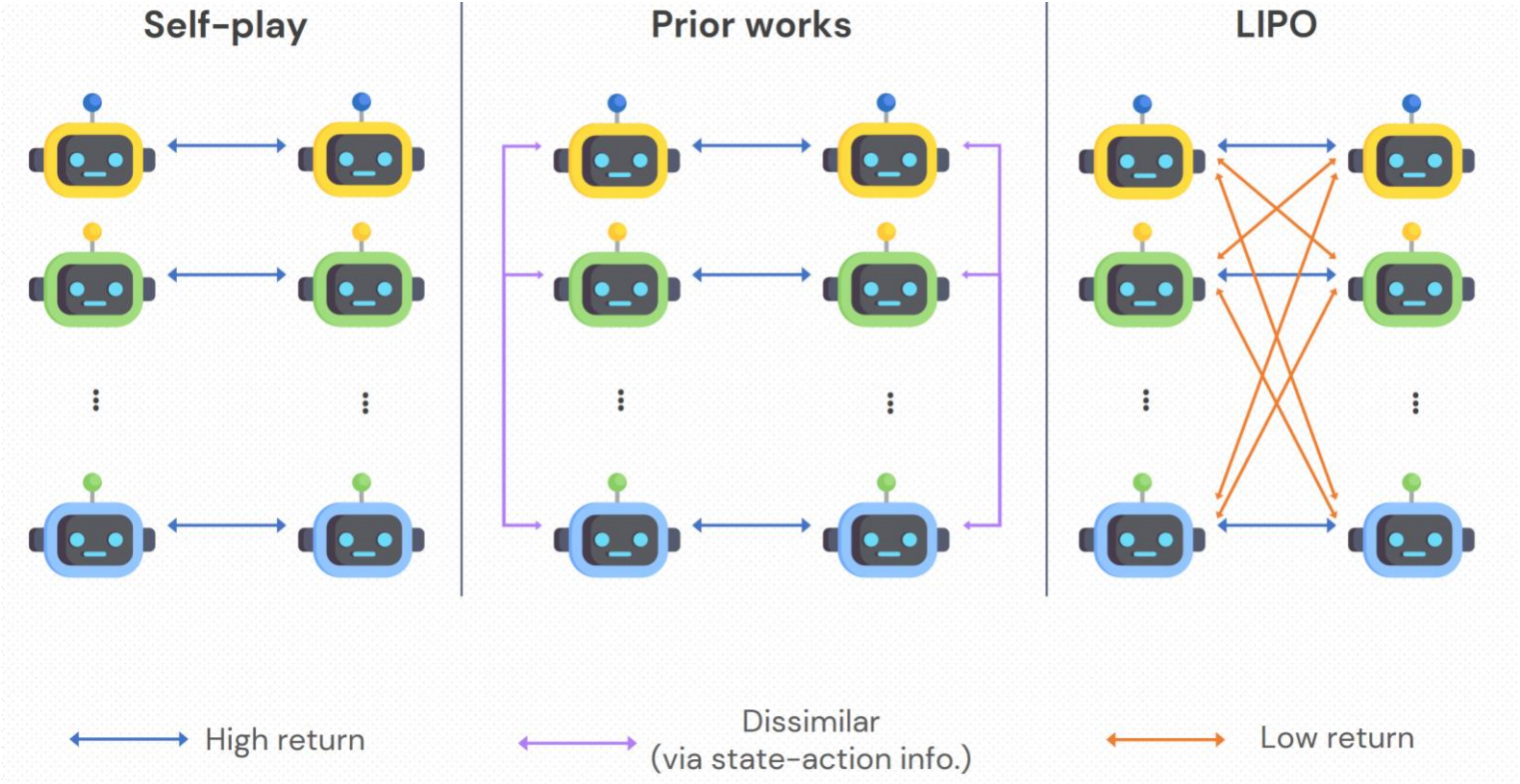


π_A and π_B have different state-action distributions **but** share the same high-level objective



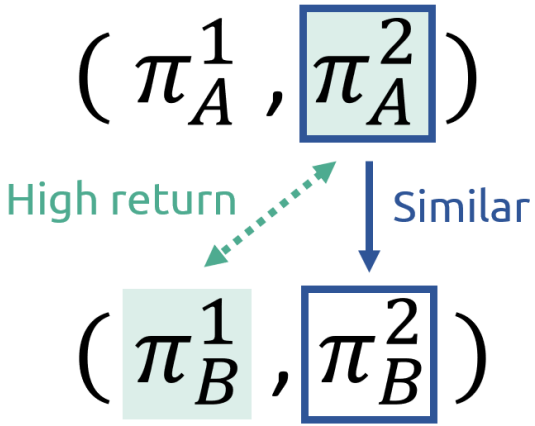
π_A and π_B have different state-action distributions **and** high-level objectives

Learning Incompatible Policies (LIPO)



Learning Incompatible Policies (LIPO)

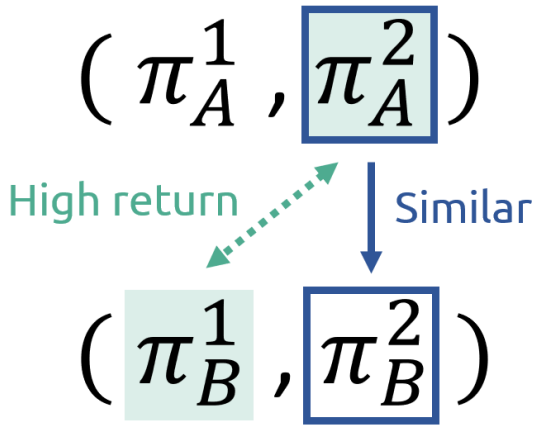
Considering $\boldsymbol{\pi_A = (\pi_A^1, \pi_A^2)}$ and $\boldsymbol{\pi_B = (\pi_B^1, \pi_B^2)}$, find $\boldsymbol{\pi_A}$ that is not similar to $\boldsymbol{\pi_B}$



If π_A^2 is **similar** to π_B^2, then π_A^2 is **compatible** with π_B

Learning Incompatible Policies (LIPO)

Considering $\boldsymbol{\pi_A} = (\boldsymbol{\pi_A^1}, \boldsymbol{\pi_A^2})$ and $\boldsymbol{\pi_B} = (\boldsymbol{\pi_B^1}, \boldsymbol{\pi_B^2})$, find $\boldsymbol{\pi_A}$ that is not similar to $\boldsymbol{\pi_B}$



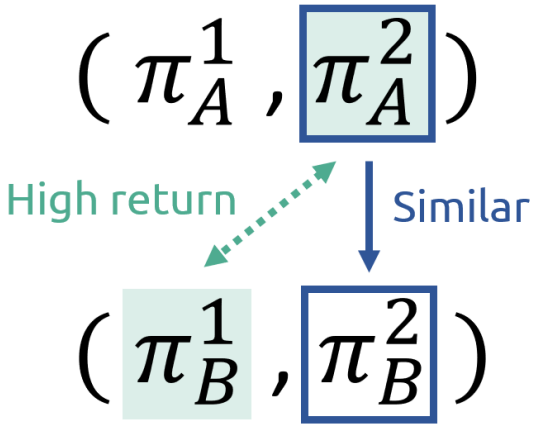
If π_A^2 is **similar** to π_B^2, then π_A^2 is **compatible** with π_B

π_A^1 *is compatible with* π_B *iff*

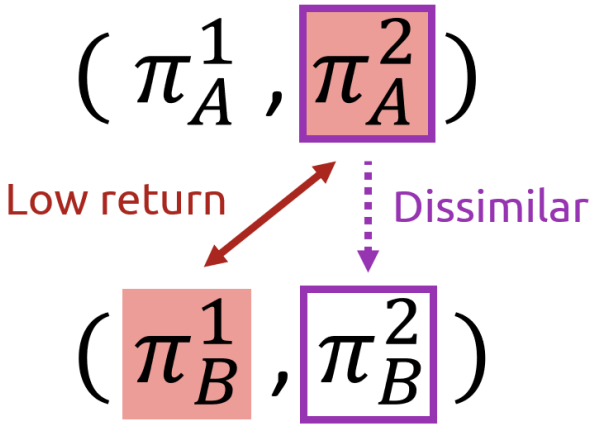
$$\mathcal{J}(\pi_B^1, \pi_A^2) \geq (1-\epsilon)\mathcal{J}_{SP}(\pi_B)$$

Learning Incompatible Policies (LIPO)

Considering $\boldsymbol{\pi_A} = (\boldsymbol{\pi_A^1}, \boldsymbol{\pi_A^2})$ and $\boldsymbol{\pi_B} = (\boldsymbol{\pi_B^1}, \boldsymbol{\pi_B^2})$, find $\boldsymbol{\pi_A}$ that is not similar to $\boldsymbol{\pi_B}$



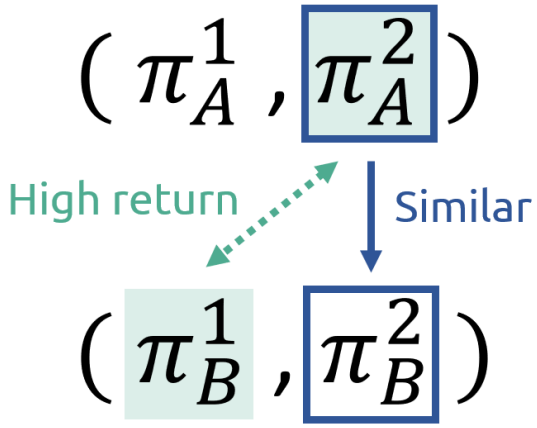
If π_A^2 is **similar** to π_B^2, then π_A^2 is **compatible** with π_B



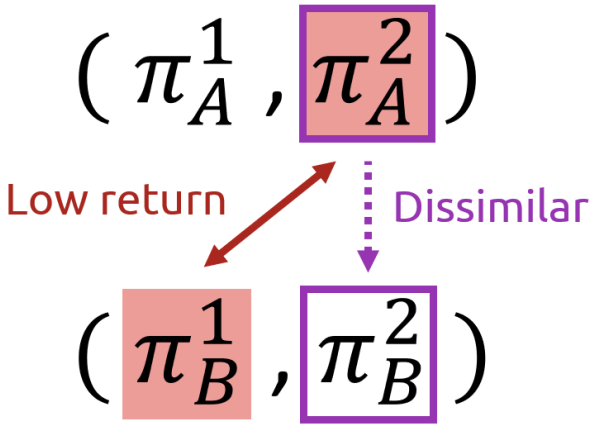
If π_A^2 is **incompatible** with π_B, then π_A^2 is **not similar** to π_B^2

Learning Incompatible Policies (LIPO)

Considering $\boldsymbol{\pi_A} = (\boldsymbol{\pi_A^1}, \boldsymbol{\pi_A^2})$ and $\boldsymbol{\pi_B} = (\boldsymbol{\pi_B^1}, \boldsymbol{\pi_B^2})$, find $\boldsymbol{\pi_A}$ that is not similar to $\boldsymbol{\pi_B}$



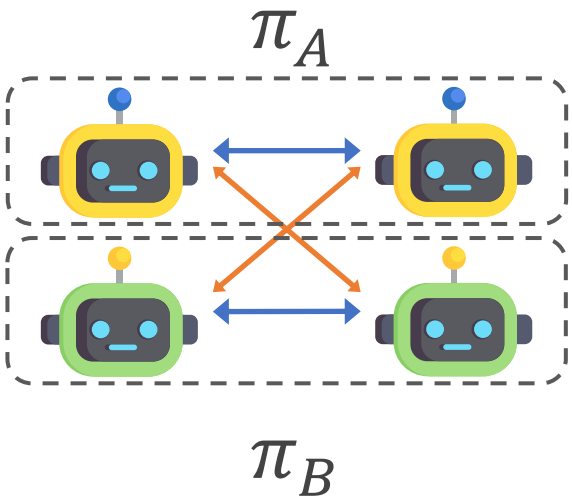
If π_A^2 is **similar** to π_B^2, then π_A^2 is **compatible** with π_B



If π_A^2 is **incompatible** with π_B, then π_A^2 is **not similar** to π_B^2

High return
(compatible)

Low return
(incompatible)

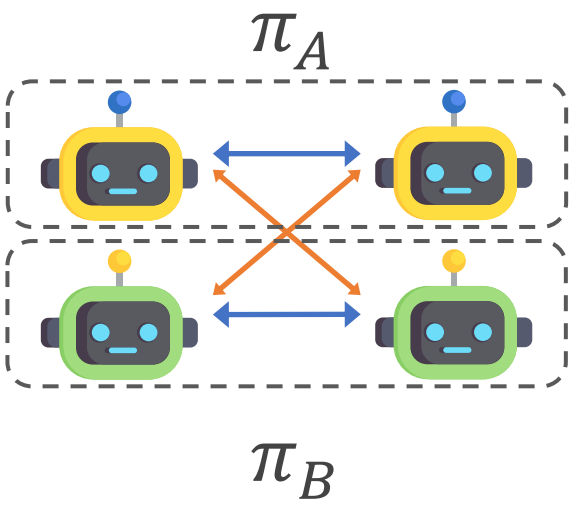


$$\max_{\pi_A} \mathcal{J}_{SP}(\pi_A) - \lambda_{XP}\mathcal{J}_{XP}(\pi_A, \pi_B)$$

$$\max_{\pi_B} \mathcal{J}_{SP}(\pi_B) - \lambda_{XP}\mathcal{J}_{XP}(\pi_B, \pi_A)$$

High return
(compatible)

Low return
(incompatible)



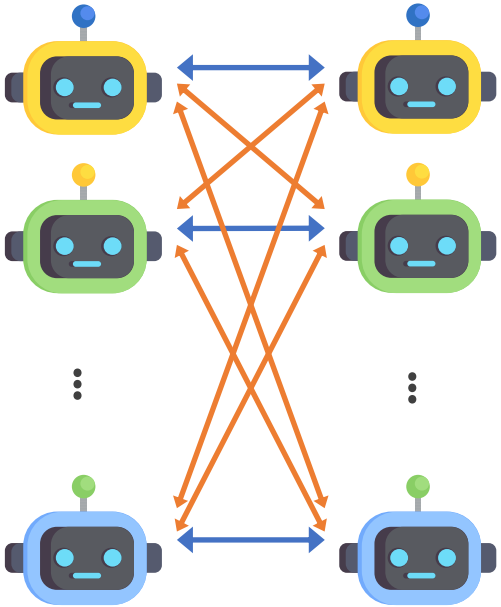
Self-play return Cross-play return

$$\max_{\pi_A} \mathcal{J}_{SP}(\pi_A) - \lambda_{XP}\mathcal{J}_{XP}(\pi_A, \pi_B)$$

$$\max_{\pi_B} \mathcal{J}_{SP}(\pi_B) - \lambda_{XP}\mathcal{J}_{XP}(\pi_B, \pi_A)$$

High return
(compatible)

Low return
(incompatible)



$$\mathcal{P} = \{\pi_i | 1 \leq i \leq N\}$$

Self-play return

Aggregated cross-play return

$$\max_{\pi_A} J_{\text{LIPO}}(\pi_A, \mathcal{P}) = J_{\text{SP}}(\pi_A) - \lambda_{\text{XP}} \tilde{J}_{\text{XP}}(\pi_A, \mathcal{P})$$

Capturing diverse behaviors that are compatible

- There may exist different behaviors that are fully **compatible**

- We propose to capture such behavioral variations by using a **mutual information** objective

Utilizing a mutual information (MI) objective

$$\pi_A(a|\tau) = \mathbb{E}_{z^1 \sim p(z^1), z^2 \sim p(z^2)} \pi_A^1(a^1|\tau^1, z^1) \pi_A^2(a^2|\tau^2, z^2)$$

Utilizing a mutual information (MI) objective

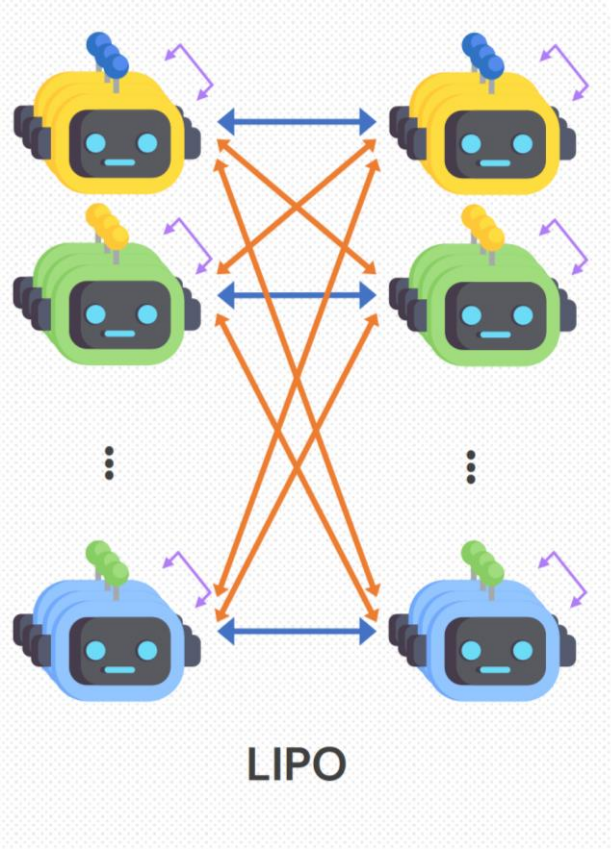
$$\pi_A(a|\tau) = \mathbb{E}_{z^1 \sim p(z^1), z^2 \sim p(z^2)} \pi_A^1(a^1|\tau^1, z^1) \pi_A^2(a^2|\tau^2, z^2)$$

$$L_{\mathrm{MI}}(\pi_A, \phi_A) = -\frac{1}{2} \sum_{i=1}^{2} \mathbb{E}_{z^i, (o^i, a^i)} \log q_{\phi_A}(z^i | o^i, \pi_A^i(\cdot | o^i, z^i)))$$

Overall training objective for each joint policy

$$\max_{\pi_A, \phi_A} J_{\mathrm{SP}}(\pi_A) - \lambda_{\mathrm{XP}} \tilde{J}_{\mathrm{XP}}(\pi_A, \mathcal{P}) - \lambda_{\mathrm{MI}} L_{\mathrm{MI}}(\pi_A, \phi_A)$$

Overall training objective for each joint policy



Self-play return

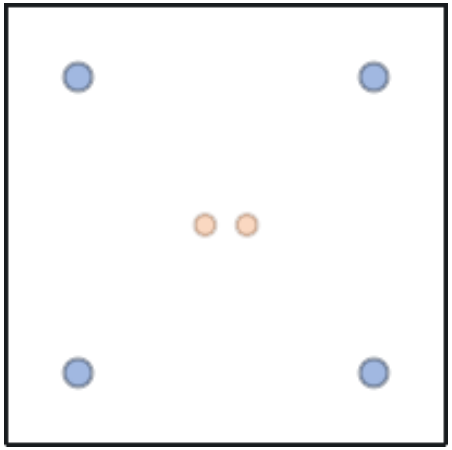
Aggregated cross-play return

MI ELBO loss

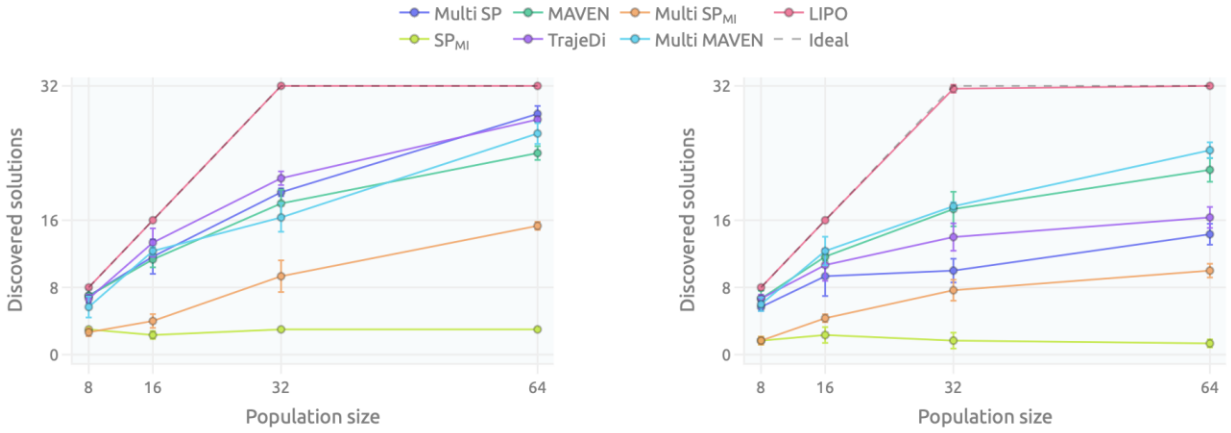
$$\max_{\pi_A, \phi_A} J_{\mathrm{SP}}(\pi_A) - \lambda_{\mathrm{XP}} \tilde{J}_{\mathrm{XP}}(\pi_A, \mathcal{P}) - \lambda_{\mathrm{MI}} L_{\mathrm{MI}}(\pi_A, \phi_A)$$

Experiments

1	0	0	0	0	0
0	2	2	0	0	0
0	2	2	0	0	0
0	0	0	3	3	3
0	0	0	3	3	3
0	0	0	3	3	3

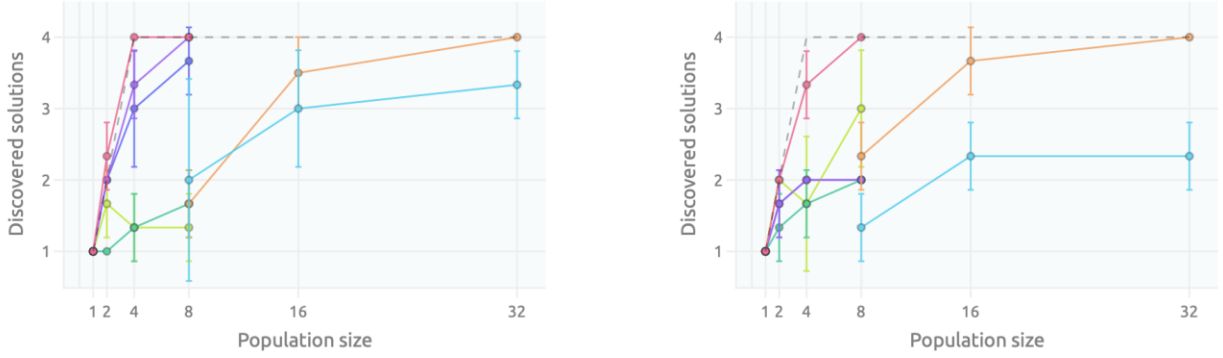


LIPO finds more solutions than the baselines



(a) Number of learned solutions in CMG-S

(b) Number of learned solutions in CMG-H

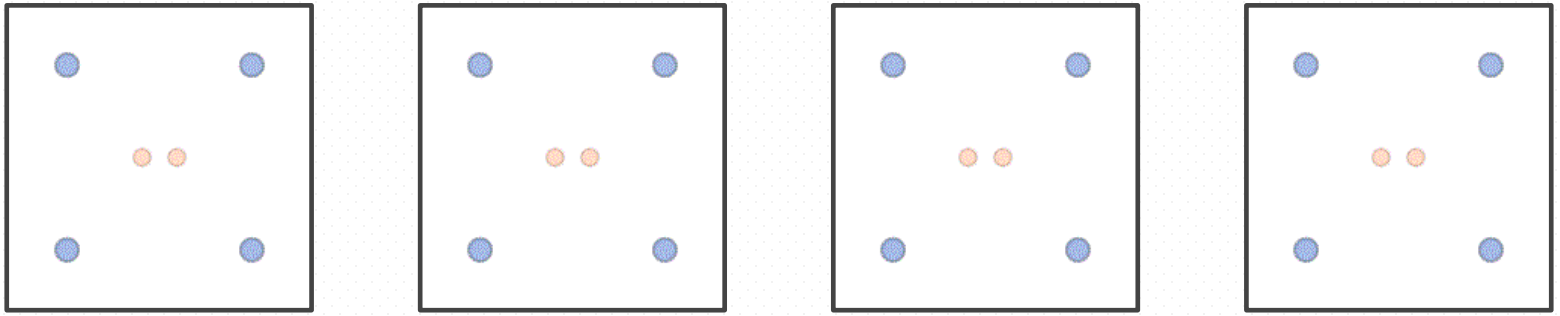


(c) Number of learned solutions in PMR-C

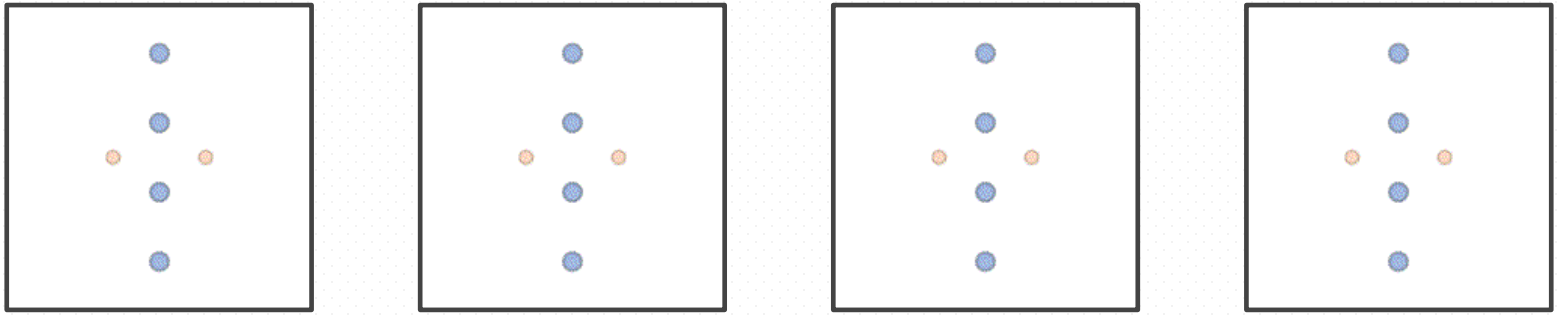
(d) Number of learned solutions in PMR-L

Qualitative results (PMR, point mass rendezvous)

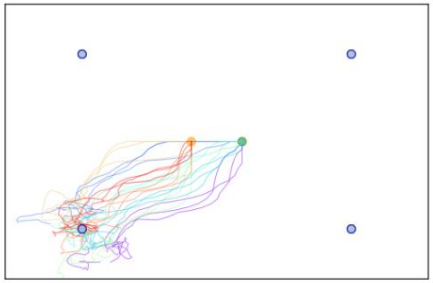
Behaviors of 4 agents produced by a single run of LIPO in PMR-C.



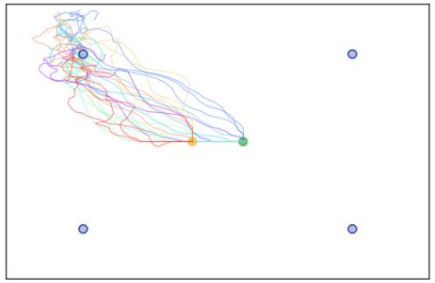
Behaviors of 4 agents produced by a single run of LIPO in PMR-L.



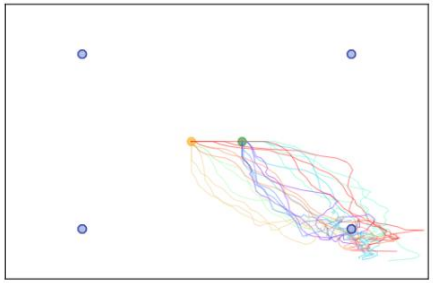
MI objective helps induce local variation



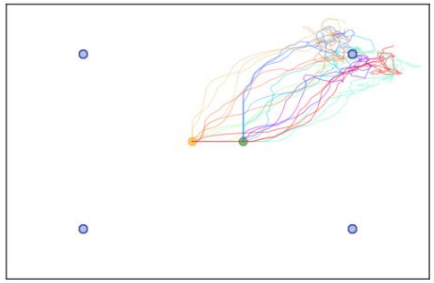
(a) $\lambda_{\text{MI}} = 0.5$



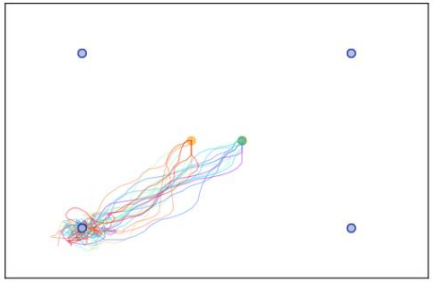
(b) $\lambda_{\text{MI}} = 0.5$



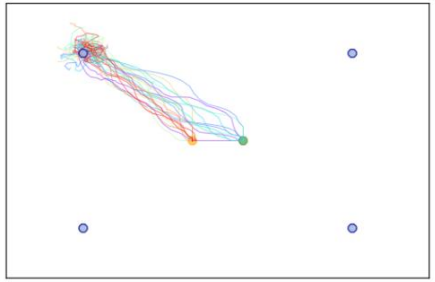
(c) $\lambda_{\text{MI}} = 0.5$



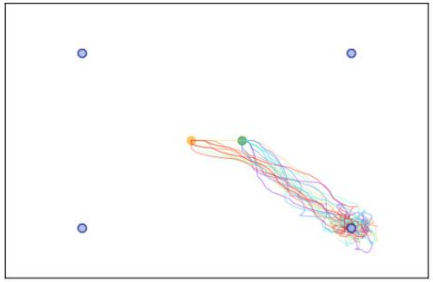
(d) $\lambda_{\text{MI}} = 0.5$



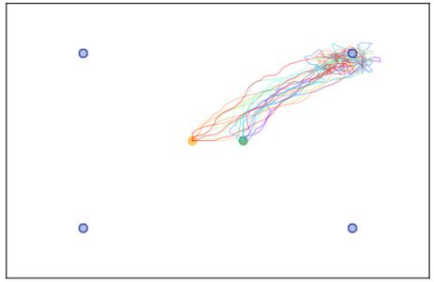
(e) $\lambda_{\text{MI}} = 0$



(f) $\lambda_{\text{MI}} = 0$

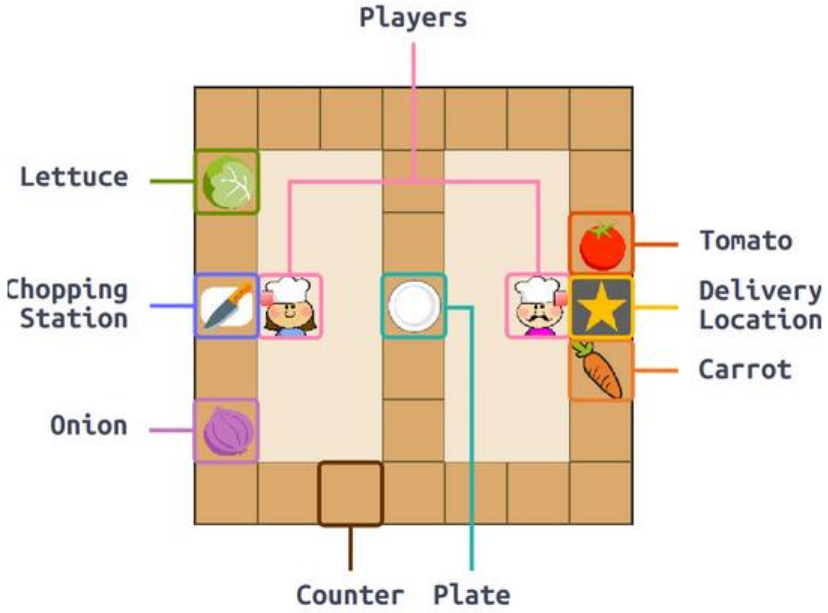


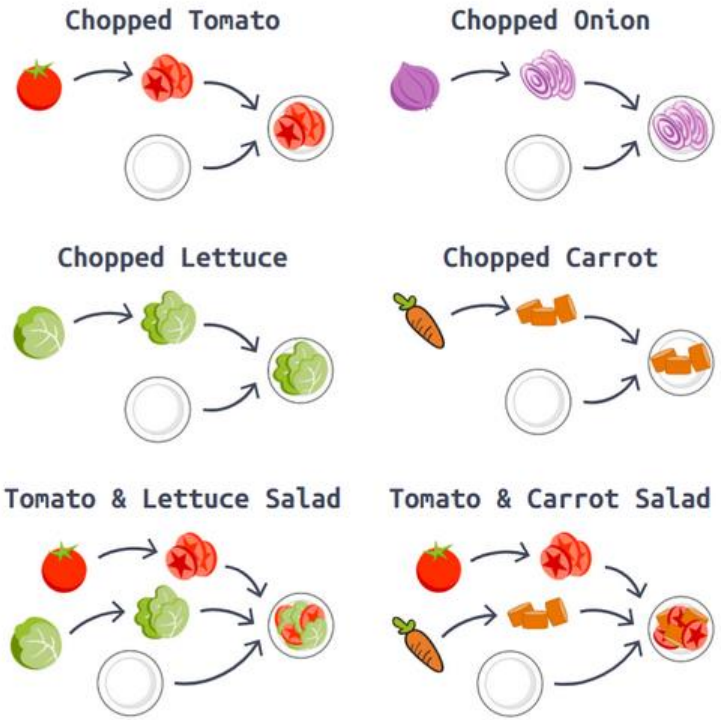
(g) $\lambda_{\text{MI}} = 0$



(h) $\lambda_{\text{MI}} = 0$

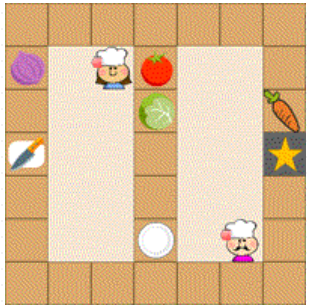
Multi-recipe Overcooked



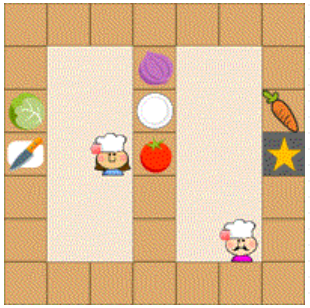


Qualitative results (multi-recipe Overcooked)

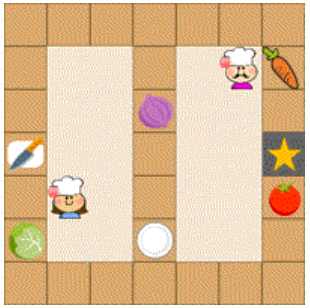
Behaviors of 8 agents produced by a single run of LIPO



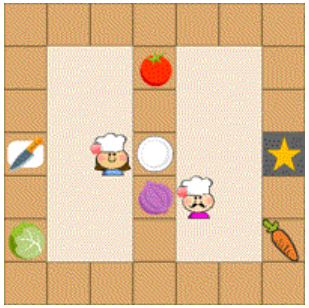
Joint policy preference:
Tomato & Carrot Salad



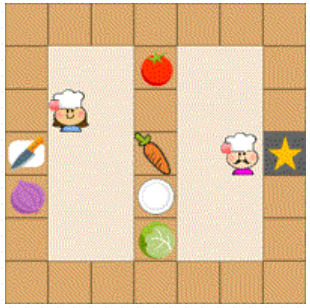
Joint policy preference:
Single-ingredient recipes



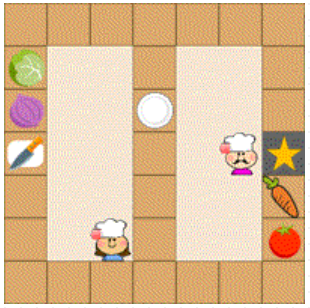
Joint policy preference:
Chopped Lettuce



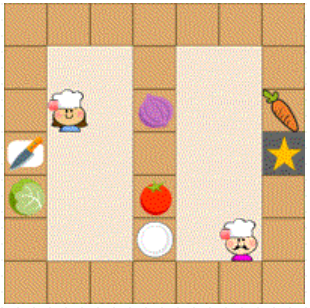
Joint policy preference:
Chopped Lettuce or Tomato



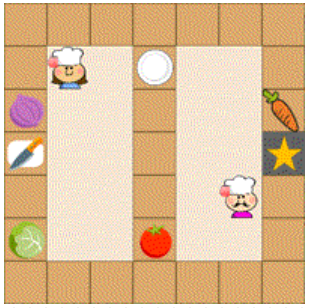
Joint policy preference:
Chopped Onion



Joint policy preference:
Tomato & Lettuce Salad

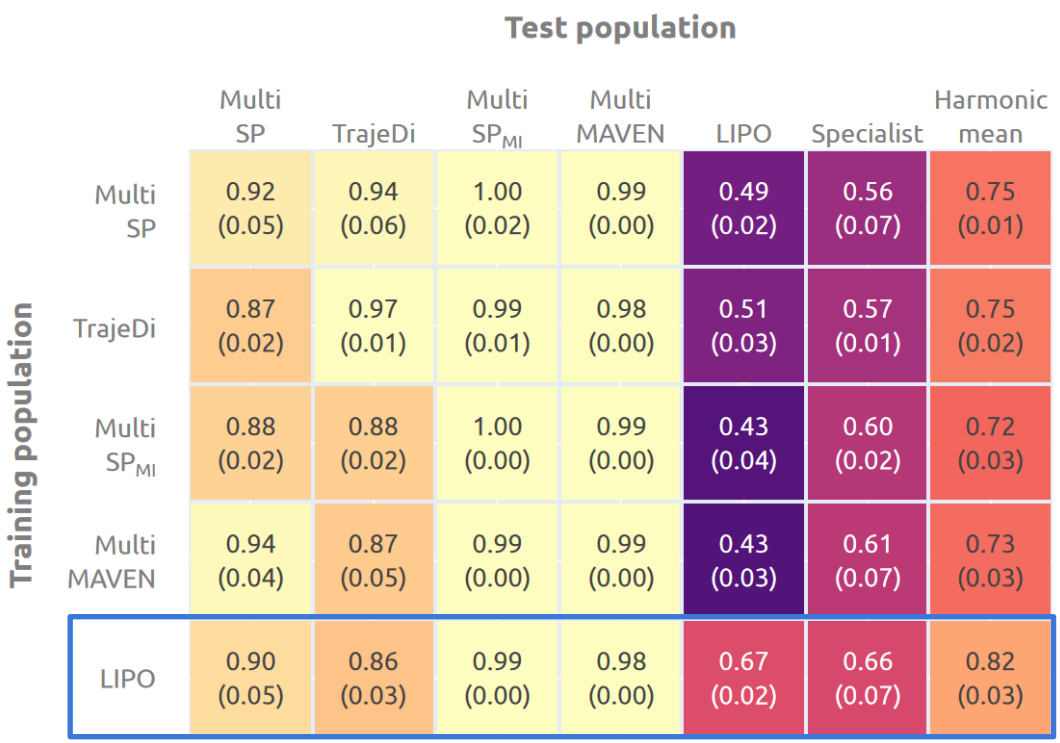


Joint policy preference:
Chopped Lettuce



Joint policy preference:
Tomato & Carrot Salad

Training generalist agents using generated agents

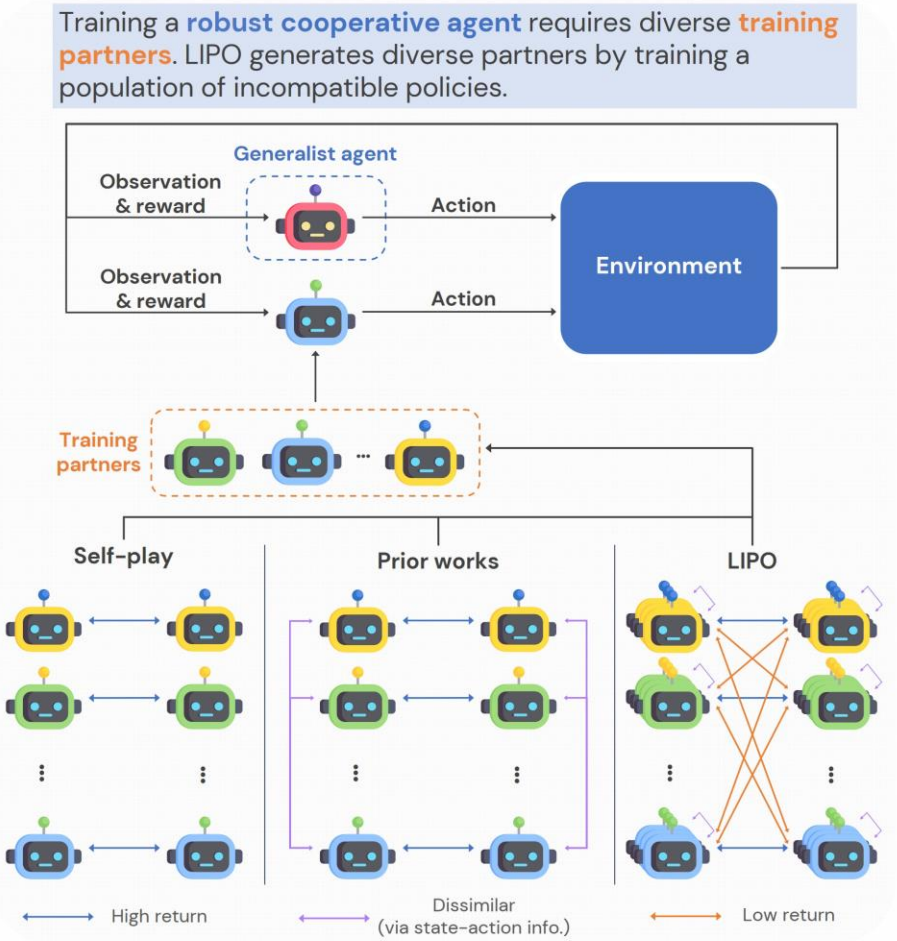


Limitations

- Computationally **expensive**
  - evaluate all possible policy pairs to minimize the cross-play return

- **Adversarial** vs other agents (addressed by **Cui et. al., 2023**)

Cui, Brandon, et al. "Adversarial Diversity in Hanabi." *International Conference on Learning Representations*. 2023.



Summary

- Use generated agents as training partners
- LIPO generates behaviorally diverse agents by **learning incompatible policies**
- On top of LIPO objective, we utilize a **mutual information** objective to diversify local behaviors
- LIPO agents are useful for training a robust **generalist agent**

Thank you!

Poster # 118

Contact: rujikorn.c_s19@vistec.ac.th

VISTEC
VIDYASIRIMEDHI
INSTITUTE OF SCIENCE AND TECHNOLOGY

SDU

IT