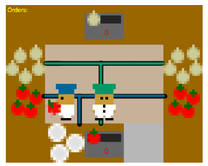
Diverse Conventions for Human-AI Collaboration

Bidipta Sarkar, Andy Shih, and Dorsa Sadigh

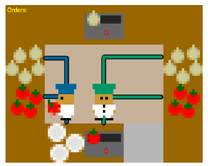
Stanford University

Motivation

MARL agents often converge to arbitrary conventions.



Split by ingredient



Split by side

Can we generate a diverse set of conventions to learn to work with people?

Diversity Definition

Want to maximize nearest-neighbors score for set of conventions, D:

$$S(D) = \mathop{\mathbb{E}}_{\pi \in D_{test}} [\max_{\pi^* \in D} J(\pi^*, \pi)]$$

For ad-hoc coordination, train a single convention-aware agent:

$$L(\hat{\pi}, D) = -J(\hat{\pi}, \hat{\pi}) - \frac{\lambda}{D} \sum_{\pi \in D} \mathop{\mathbb{E}}_{(o,a)\sim\pi} [\log(\hat{\pi}(a \mid o))]$$

To evaluate the performance of the convention-aware agent:

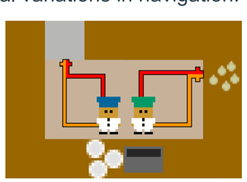
$$J(\hat{\pi}, D_{test}) = \mathop{\mathbb{E}}_{\pi \in D_{test}} [J(\hat{\pi}, \pi)]$$

Statistical Diversity

Some approaches add a term to induce statistical variations in trajectories:

$$C_{ADAP}(D) = \mathop{\mathbb{E}}_{s \in S} \left[\mathop{\mathbb{E}}_{\substack{\pi_1, \pi_2 \in D \\ \pi_1 \neq \pi_2}} \exp(-D_{\mathrm{KL}}(\pi_1(s) \mid\mid \pi_2(s)) \right]$$

These fail to identify trivial variations in navigation:



Cross-Play Minimization

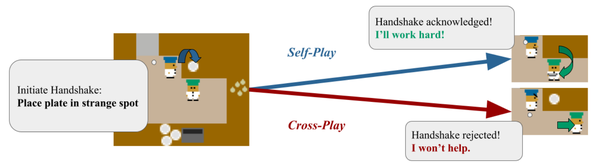
Adding a new convention π_n guarantees an increase in the score S:

$$S(D_n) \geq S(D_{n-1}) + p(\pi_n)(J(\pi_n, \pi_n) - \max_{\pi^* \in D_{n-1}} J(\pi_n, \pi^*))$$

Maximizing this lower bound gives a cross-play minimization algorithm.

Handshakes and Mixed-Play

Pure cross-play minimization incentivizes sabotages under cross-play.

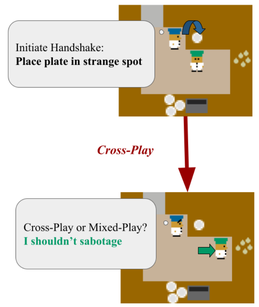


Mixed-play ensures that conventions always act in good faith.

Algorithm 1: Generating Mixed Play Buffer

Input: policies π_n, π^*, MDP M
Output: Replay buffer from running mixed-play.

$s_0, R \leftarrow s, 0$
$t \sim \text{uniform}(1, T)$
for $i \leftarrow [0, T)$ **do**
  $o^1, o^2 \leftarrow o(s_i)$
  if $i < t$ **then**
    $\pi_1^m \leftarrow$ Randomly choose π_n or π^*
    $\pi_2^m \leftarrow$ Randomly choose π_n or π^*
    $a_i^1, a_i^2 \leftarrow \pi_1^m(o^1), \pi_2^m(o^2)$
    $s_{i+1}, - \leftarrow$ Step in M with (a_i^1, a_i^2)
  else
    $a_i^1, a_i^2 \leftarrow \pi_n(o^1), \pi_n(o^2)$
    $s_{i+1}, r_i \leftarrow$ Step in M with (a_i^1, a_i^2)
Return: ReplayBuffer$(s_{t:T}, a_{t:T}^1, a_{t:T}^2, r_{t:T})$



CoMeDi: Cross-Play and Mixed-Play

Loss function to generate convention π_n

$$L(\pi_n) = -J(\pi_n, \pi_n) + \alpha J(\pi_n, \pi^*) - \beta J_{\mathsf{M}}(\pi_n, \pi^*)$$

Where π^* is the most-compatible previously discovered convention.

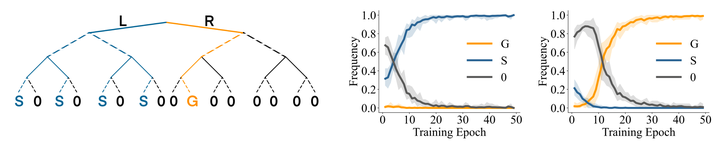
Algorithm 2: Diverse Conventions with CoMeDi

Input: Number of policies to generate n
Output: Diverse set of conventions D

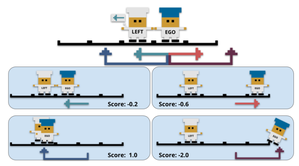
$D \leftarrow (\pi_1, \ldots, \pi_n)$, parameterized by $(\theta_1, \ldots, \theta_n)$
Train π_1 with standard self-play
for $i \in \{2, \ldots, n\}$ **do**
  while policy π_i has not converged **do**
    $\pi^* \leftarrow \arg\max_{\pi^* \in D_{1:i-1}} J(\pi_i, \pi^*)$
    $\tau_{SP} \leftarrow$ `GetRollout`(π_i, π_i)
    $\tau_{XP} \leftarrow$ `GetRollout`(π_i, π^*)
    $\tau_{MP} \leftarrow$ `MixedPlayRollout`(π_i, π^*)
    Estimate $J(\pi_i, \pi_i), J(\pi_i, \pi^*), J_M(\pi_i, \pi^*)$ with $\tau_{SP}, \tau_{XP}, \tau_{MP}$
    $\theta_i \leftarrow \theta_i + \nabla_{\theta_i}[-J(\pi_i, \pi_i) + \alpha J(\pi_i, \pi^*) - \beta J_M(\pi_i, \pi^*)]$
Return: D

Simulation Results

In Blind Bandits, CoMeDi finds both the S and G conventions:



In Balance Beam, tuning the mixed-play weight eliminates handshakes.



β	SP ↑	XP ↓	HS ↓	PX ↓	LS ↑	RS ↓
0.00	1.616	**-0.392**	0.16	**0.00**	1.128	**-0.656**
0.25	1.808	0.096	**0.00**	**0.00**	1.272	0.096
0.50	**2.000**	0.200	**0.00**	**0.00**	**1.384**	0.128
1.00	1.904	0.632	**0.00**	0.24	1.232	0.160

In Overcooked, CoMeDi's convention-aware agent has strong cross-play performance with held-out conventions.

layout	CoMeDi	XP	ADAP	SP	PPO_BC
simple	**4.55**	3.49	2.79	4.36	3.75
random1	**3.42**	1.58	0.72	2.70	2.29
random3	**1.19**	0.49	0.03	0.33	1.17
unident_s	**5.27**	3.62	1.84	2.51	2.58
random0	0.37	0.49	0.00	0.21	**1.44**

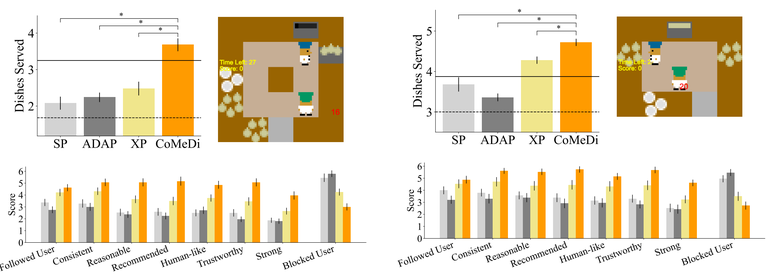
Scores with PBT

layout	CoMeDi	XP	ADAP	SP	PBT
simple	**4.91**	4.46	2.77	4.51	3.75
random1	**3.43**	1.81	1.10	2.38	2.29
random3	**1.91**	1.1	0.03	1.13	1.17
unident_s	2.94	3.16	2.59	**4.52**	2.58
random0	**1.56**	0.88	0	0.25	1.44

Scores with PPO_BC

Overcooked User Study

CoMeDi significantly outperforms baselines in terms of pure score and user opinion, even surpassing human-level performance.



Live Demo, Code, and Videos

Visit our website: https://iliad.stanford.edu/Diverse-Conventions/