# Extending Mean Field RL to Partially Observable Environments, Multiple Types & Decentralized Learning and Improving Cooperation in Mixed-Motive Games

**Pascal Poupart** (University of Waterloo)

CIFAR AI Chair, Vector Institute

May 19, 2026

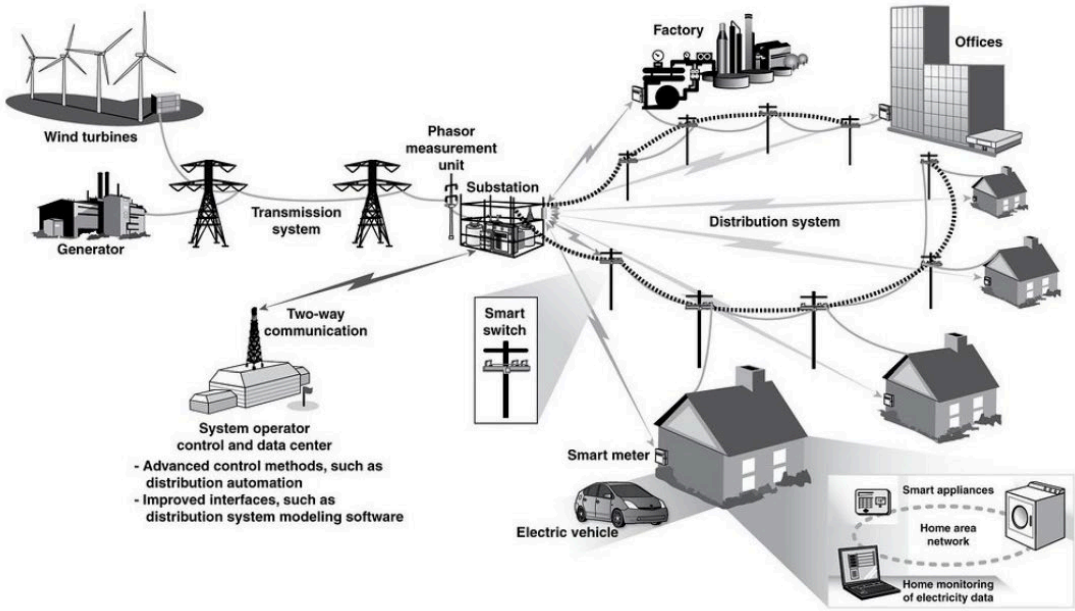
# Outline

- Mean Field Reinforcement Learning
  - **Multi Type Mean Field Reinforcement Learning**
    Subramanian, Poupart, Taylor, Hegde, *AAMAS*, 2020.
  - **Partially Observable Mean Field Reinforcement Learning**
    Subramanian, Taylor, Crowley, Poupart, *AAMAS*, 2021.
  - **Decentralized Mean Field Games**
    Subramanian, Taylor, Crowley, Poupart, *AAAI*, 2022.
  - **Revisiting Neighbourhoods in Mean Field Reinforcement Learning**
    Subramanian, Taylor, Crowley, Larson, Poupart, *under review,* 2026.
- Improve cooperation in mixed-motive games
  - **Learning to Negotiate via Voluntary Commitment**
    Zhu, Wang, Subramanian, Poupart, *AISTATS 2025.*
  - **Talk, Judge, Cooperate: Gossip-Driven Indirect Reciprocity in Self-Interested LLM Agents**
    Zhu, Lin, Kaistha, Li, Wang, Zha, Hadfield, Poupart*, ICML,* 2026.

# Multi-Agent Reinforcement Learning

- Challenge: exponential complexity in # of agents

- Mean-Field Reinforcement Learning
  - Many agents → two agents



# Mean Field Reinforcement Learning

## Stochastic Game

Tuple: $\langle N, S, A, R, T \rangle$

- $N$: # of agents
- $S$: shared state space
- $A^j$: individual action space

$\langle a^1, a^2, \dots, a^N \rangle \in A^1 \times A^2 \times \cdots \times A^N$

- Reward: $R^j(s, a^1, a^2, \dots, a^N)$
- Transition $T$: $\Pr(s'|s, a^1, a^2, \dots, a^N)$

## Mean Field Game

Huang et al. (2003), Lasry et al. (2007)

Tuple: $\langle N, S, A, R, T \rangle$

- $N$: # of agents $\approx \infty$
- $S$: shared state space
- $A$: individual action space

$$\bar{a} = \Pr(a^j) \in \Delta A$$

- Reward: $R^j(s, a^j, \bar{a}^j)$
- Transition $T$: $\Pr(s'|s, \bar{a})$

# Mean Field Action $\bar{a}$

## Discrete actions

$$\bar{a} = \Pr(a^j) = \frac{1}{N}\sum_{j=1}^{N} a^j$$

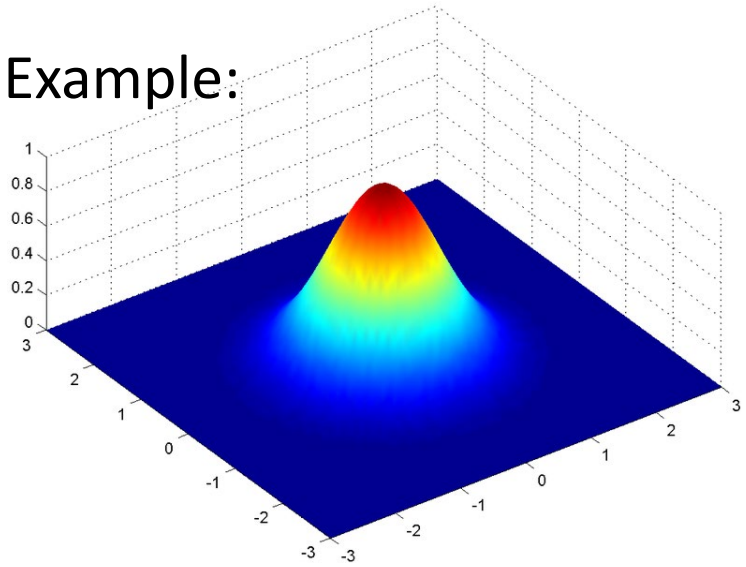
Example:

$$a^1 = \langle 0, 1, 0, 0 \rangle$$
$$a^2 = \langle 1, 0, 0, 0 \rangle$$
$$a^3 = \langle 0, 0, 0, 1 \rangle$$
$$a^4 = \langle 0, 1, 0, 0 \rangle$$
$$\bar{a} = \langle 0.25, 0.5, 0, 0.25 \rangle$$

## Continuous actions

$$\bar{a} = pdf(a^j) = \frac{1}{N}\sum_{j=1}^{N} \delta(a^j)$$

Example:



# Mean Field Reinforcement Learning

- Yang et al. (2018)
- Assumptions:
  - Independent and identical agents (single type)
  - Fully observable agents
  - Centralized learning
  - Q-function is additively decomposable and well approximated by mean field $Q_{MF}$

$$Q^j(s, a^1, a^2, \dots, a^N) = \frac{1}{|nb(j)|} \sum_{k \in nb(j)} Q^j(s, a^j, a^k) \approx Q^j_{MF}\left(s, a^j, \bar{a}^j\right)$$

- Proof of convergence to Nash Equilibrium in the limit

# Multi Type Mean Field RL

- Group agents into $M$ types
- Each type $m$ has a different reward function: $R_m^j(s, a^j, \bar{a}_1^j, \bar{a}_2^j, \ldots, \bar{a}_M^j)$
- Assumptions:
  - ~~Independent and identical agents (single type)~~
  - Fully observable agents
  - Centralized learning
  - Q-function is additively decomposable and well approximated by $Q_{MTMF}$

$$Q^j(s, a^1, a^2, \ldots, a^N) \approx Q_{MTMF}^j\left(s, a^j, \bar{a}_1^j, \bar{a}_2^j, \ldots, \bar{a}_M^j\right)$$

# Mean Field Updates

| Mean Field Update | Mean Field RL *(Yang et al. (2018))* | Multi Type Mean Field RL |
|---|---|---|
| **Q function** | $Q^j_{t+1}(s, a^j, \bar{a}^j) = (1-\alpha)Q^j_t(s, a^j, \bar{a}^j) + \alpha[r^j + \gamma v^j_t(s')]$ | $Q^j_{t+1}(s, a^j, \bar{a}^j_1, \bar{a}^j_2, .. \bar{a}^j_M) = (1-\alpha)Q^j_t(s, a^j, \bar{a}^j_1, \bar{a}^j_2, .. \bar{a}^j_M) + \alpha[r^j + \gamma v^j_t(s')]$ |
| **Value Function** | $v^j_t(s') = \sum_{a^j} \pi^j_t(a^j \mid s', \bar{a}^j) E_{a^{-j}_i \sim \pi^{-j}_{i,t}} Q^j_t(s', a^j, \bar{a}^j)$ | $v^j_t(s') = \sum_{a^j} \pi^j_t(a^j \mid s', \bar{a}^j_1, .. \bar{a}^j_M) E_{a^{-j}_i \sim \pi^{-j}_t}[Q^j_t[s', a^j, \bar{a}^j_1, \ldots \bar{a}^j_M]]$ |
| **Mean Value** | $\bar{a}^j_i = \frac{1}{N^j_i} \sum_k a^k_i, a^k_i \sim \pi^k_t(\cdot \mid s, \bar{a}^k_-)$ | $\bar{a}^j_i = \frac{1}{N^j_i} \sum_k a^k_i, a^k_i \sim \pi^k_t(\cdot \mid s, \bar{a}^k_{1-}, \ldots, \bar{a}^k_{M-})$ |
| **Boltzmann Policy** | $\pi^j_t(a^j \mid s, \bar{a}^j) = \frac{exp(-\beta Q^j_t(s, a^j, \bar{a}^j))}{\sum_{a^{j'} \in A^j} exp(-\beta Q^j_t(s, a^{j'}, \bar{a}^j))}$ | $\pi^j_t(a^j \mid s, \bar{a}^j_1, \ldots \bar{a}^j_M) = \frac{exp(\beta Q^j_t(s, a^j, \bar{a}^j_1, .., \bar{a}^j_M))}{\sum_{a^{j'} \in A^j} exp(\beta Q^j_t(s, a^{j'}, \bar{a}^j_1, \ldots \bar{a}^j_M))}$ |

# Theory

- Let $\epsilon$ be a bound on the mean action deviation for M types:

$$\forall m: \frac{1}{K_m} \sum_{k_m}^{K_m} ||a_{k_m} - \bar{a}_m|| \leq \epsilon$$

- **Theorem 1:** When the Q-function is L-smooth, then

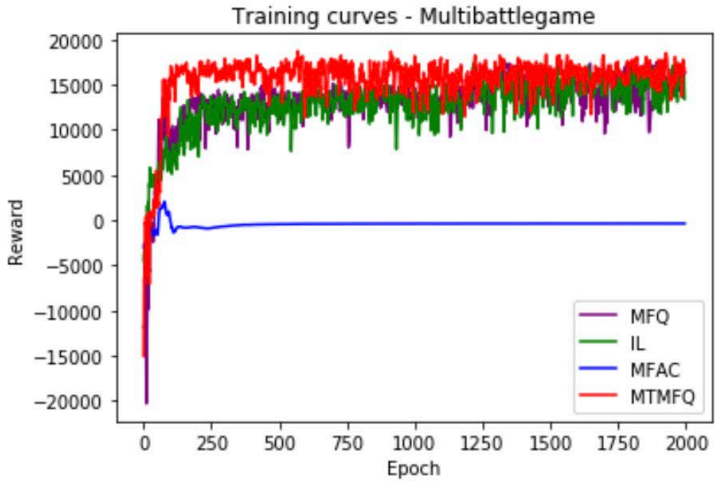
$$\left| Q^j(s, a^1, a^2, \dots, a^N) - Q^j_{MTMF}\left(s, a^j, \bar{a}^j_1, \bar{a}^j_2, \dots, \bar{a}^j_M\right) \right| \leq \frac{1}{2} L\epsilon$$

- **Theorem 2:** Multi-type mean field Q-learning converges to a bounded distance of a Nash Equilibrium under suitable conditions

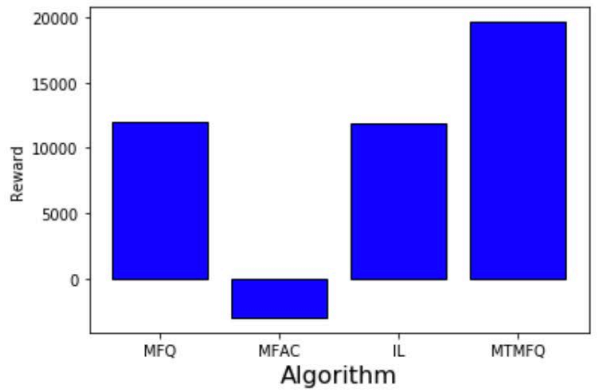
$$\left| Q^j_*(s, a^1, a^2, \dots, a^N) - Q^j_{MTMF}\left(s, a^j, \bar{a}^j_1, \bar{a}^j_2, \dots, \bar{a}^j_M\right) \right| \leq \frac{1}{2} L\epsilon - S$$

# Experiments: Multi Team Battle

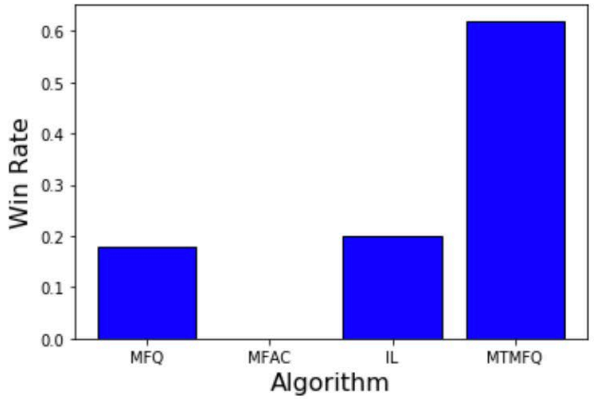
- Stage 1: training with same algorithm – 2000 episodes
- Stage 2: Faceoff against other algorithms – 1000 episodes
- Reward function encourages competing and cooperating (with favourable opponents)




Training - Stage 1



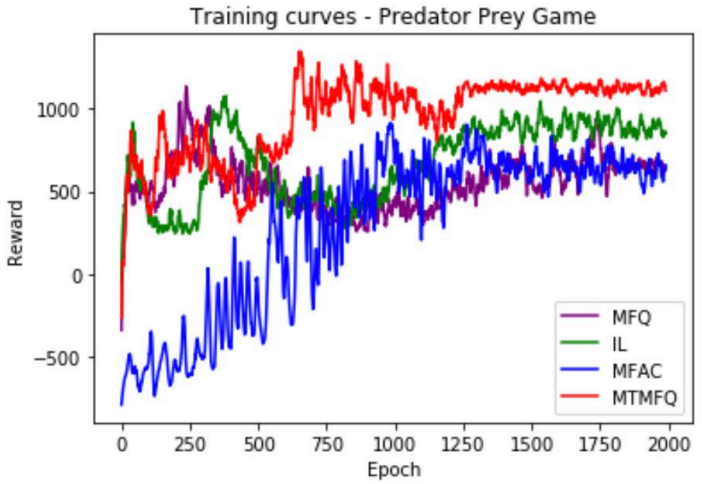

Total Reward per Episode - Stage 2



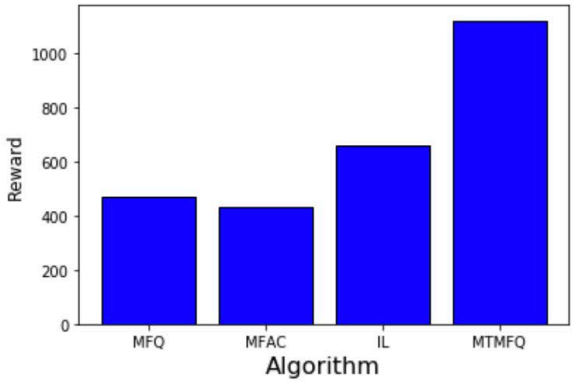

Win Rate - Stage 2

# Experiments: Predator Prey

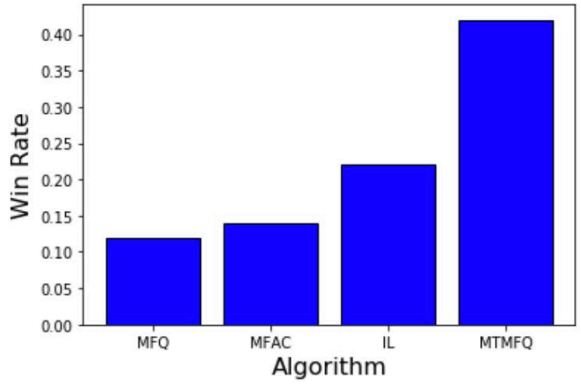
- Two types: predator and prey
  - Prey: fast (try to escape predators)
  - Predator: powerful (try to kill prey)
- MTMFQ uses K-means to learn the types of the opponents




Training - Stage 1




Total Reward per Episode - Stage 2




Win Rate - Stage 2

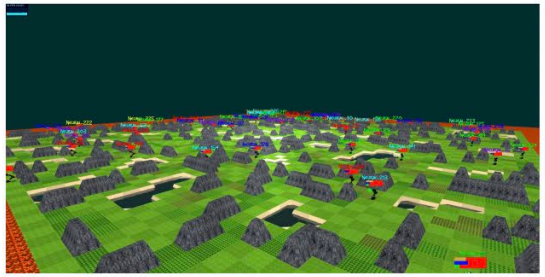
# Partially Observable Mean Field RL

- Compute belief $\tilde{a}$ over mean field action $\bar{a}$

- Assumptions:
  - Independent and identical agents (single type)
  - ~~Fully observable agents~~
  - Centralized learning

# Categories of Partial Observability

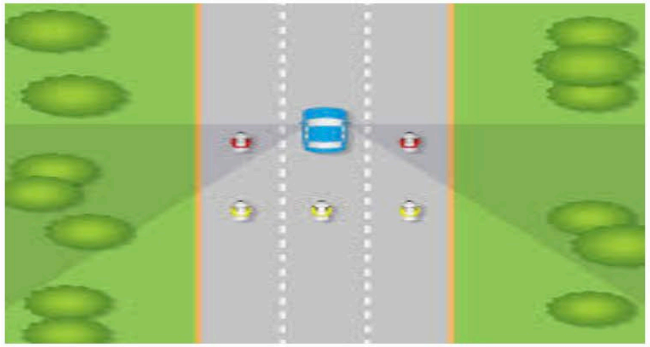
Fixed Observation Radius (FOR)



MMO

Logistics Operation

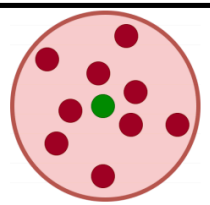
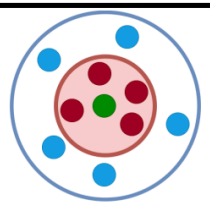
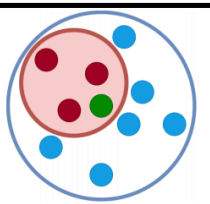
Probabilistic Distance-based Observability (PDO)



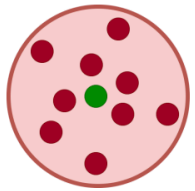
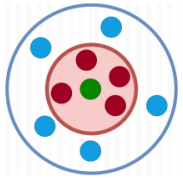
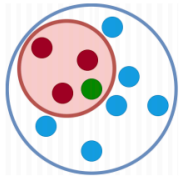
Blind spots

Pedestrian motion

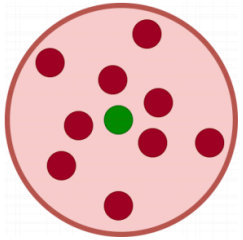
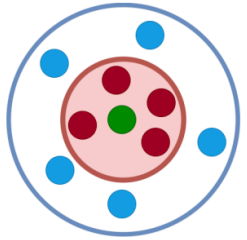
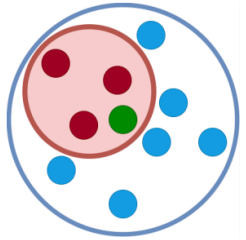
# Mean Action Update

| Update Type | MFQ | POMFQ-FOR | POMFQ-PDO |
|---|---|---|---|
| Update Figure |  |  |  |
| Distribution of Mean field | None | $\mathscr{D}^j(\theta) \propto \theta_1^{\eta_1-1+c_1} \ldots \theta_L^{\eta_L-1+c_L}; \quad \mathscr{D}^j(\theta \mid \eta + c)$ | |
| Mean Action | $\bar{a}_t^j = \frac{1}{N^j} \sum_{k \neq j} a_t^k, a_t^k \sim \pi^k(\cdot \mid s_t, \bar{a}_{t-1}^k)$ | $\tilde{a}_{i,t}^j \sim \mathscr{D}^j(\theta; \eta + c); \quad \tilde{a}_t^j = \frac{1}{\mathcal{S}} \sum_{i=1}^{i=\mathcal{S}} \tilde{a}_{i,t}^j$ | |

# Q-function Update

| Update Type | MFQ | POMFQ-FOR | POMFQ-PDO |
|---|---|---|---|
| Update Figure |  |  |  |
| Distance Parameter | None | None | $\bar{\lambda}^j_{i,t} \sim [Gamma^j(\hat{\alpha}+0.5, d+\hat{\beta}-d/\hat{\theta})]; \quad \bar{\lambda}^j_t = \frac{1}{\mathcal{G}}\sum_{i=1}^{i=\mathcal{G}} \bar{\lambda}^j_{i,t}$ |
| Q update | $Q^j(s^j_t, a^j_t, \bar{a}^j_t) = (1-\alpha)Q^j(s_t, a^j_t, \bar{a}^j_t) + \alpha[r^j_t + \gamma v^j(s_{t+1})]$ | $Q^j(s^j_t, a^j_t, \tilde{a}^j_t) = (1-\alpha)Q^j(s^j_t, a^j_t, \tilde{a}^j_t) + \alpha[r^j_t + \gamma v(s^j_{t+1})]$ | $Q^j(s^j_t, a^j_t, \tilde{a}^j_t, \bar{\lambda}^j_t) = (1-\alpha)Q^j(s^j_t, a^j_t, \tilde{a}^j_t, \bar{\lambda}^j_t) + \alpha[r^j_t + \gamma v^j(s^j_{t+1})]$ |

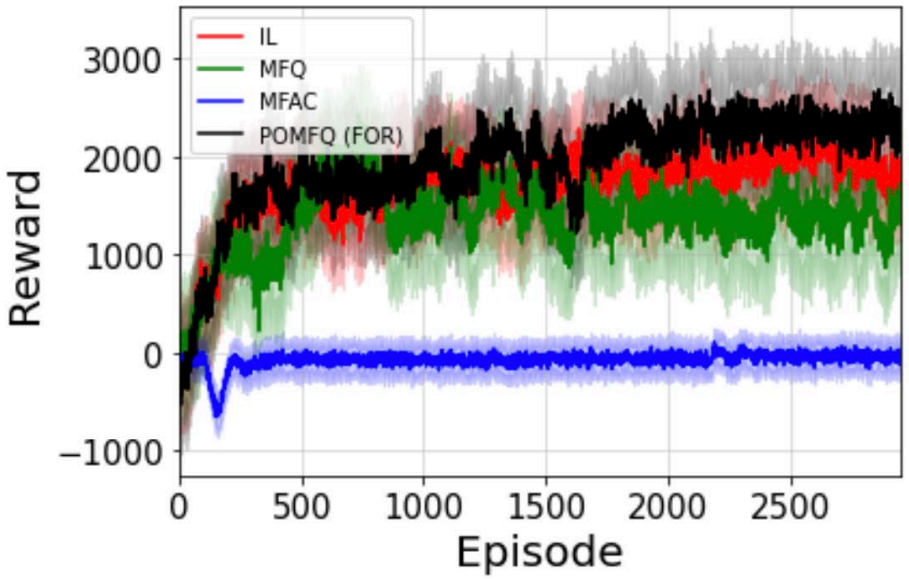
# Boltzmann Policy

| Update Type | MFQ | POMFQ-FOR | POMFQ-PDO |
|---|---|---|---|
| Update Figure |  |  |  |
| Policy | $\pi^j(a_t^j \mid s_t, \bar{a}_{t-1}^j) = \frac{\exp(-\beta Q^j(s_t, a_t^j, \bar{a}_{t-1}^j))}{\sum_{a_t^{j'} \in A^j} \exp(-\beta Q^j(s_t, a_t^{j'}, \bar{a}_{t-1}^j))}$ | $\pi^j(a_t^j \mid s_t^j, \tilde{a}_{t-1}^j) = \frac{\exp(-\beta Q^j(s_t^j, a_t^j, \tilde{a}_{t-1}^j))}{\sum_{a_t^{j'} \in A^j} \exp(-\beta Q^j(s_t^j, a_t^{j'}, \tilde{a}_{t-1}^j))}$ | $\pi^j(a_t^j \mid s_t^j, \tilde{a}_{t-1}^j, \bar{\lambda}_{t-1}^j) = \frac{\exp(-\beta Q^j(s_t^j, a_t^j, \tilde{a}_{t-1}^j, \bar{\lambda}_{t-1}^j))}{\sum_{a_t^{j'} \in A^j} \exp(-\beta Q^j(s_t^j, a_t^{j'}, \tilde{a}_{t-1}^j, \bar{\lambda}_{t-1}^j))}$ |

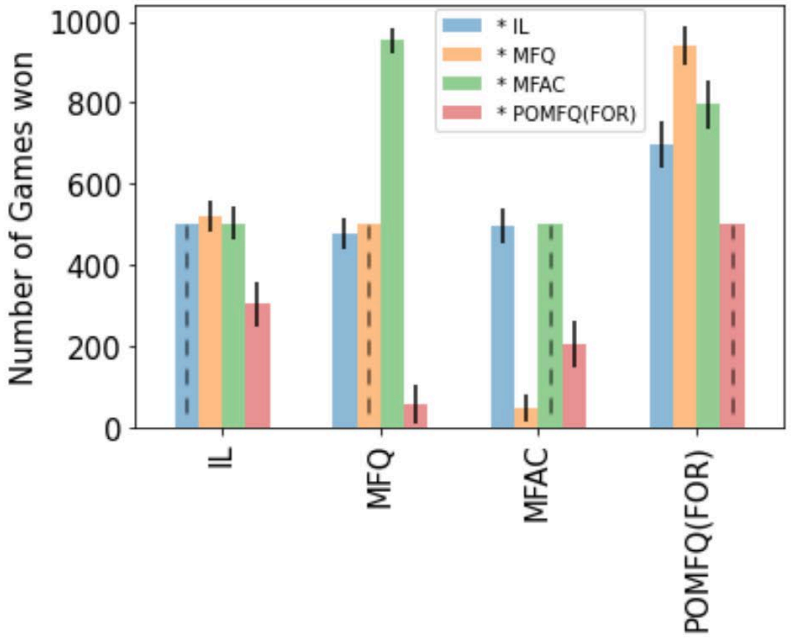
# Experiments: Multi Team Battle (FOR)

- Fixed Observation Radius

Training



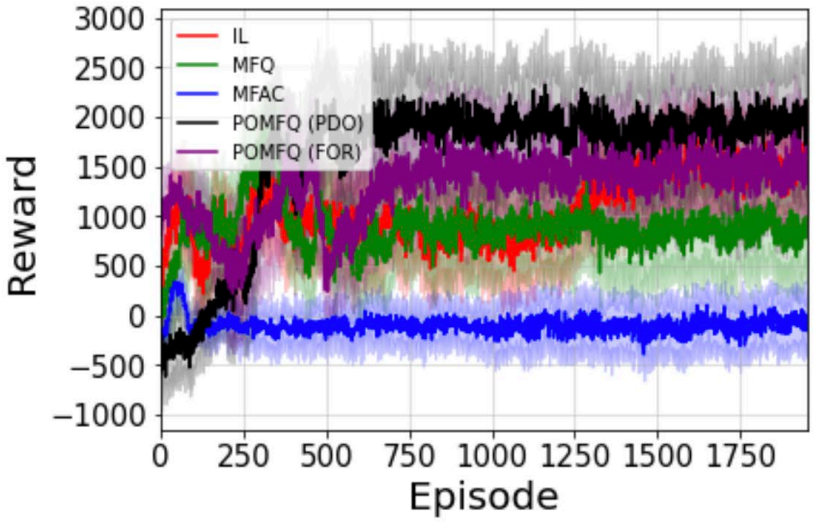

Testing



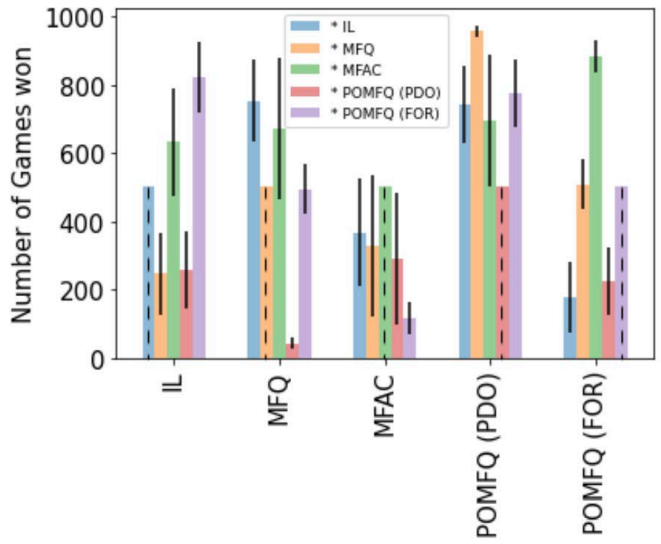
# Experiments: Multi Team Battle (PDO)

- Probabilistic Distance-based Observability

Training




Testing



# Decentralized Mean Field Games

- Assumptions:
  - ~~Identical agents (single type)~~
  - ~~Fully observable agents~~
  - ~~Centralized learning~~

- Each agent
  - Makes local observations $s^j$
  - Estimates $\tilde{a}^j$ as an approximation of $\bar{a}$ based on $s^j$
  - Learns its own policy $\pi^j$

# Mean Field Equilibrium

- $(\boldsymbol{\pi}_*, \bar{\boldsymbol{a}}_*)$ is a centralized mean field equilibrium when
  - $V^{\pi_*}(s, \bar{a}_*) \geq V^{\pi}(s, \bar{a}_*) \quad \forall \pi$
  - $\bar{a}_*$ is the mean field induced by all agents playing $\pi_*$
  - Implicit assumption:
    - All agents play the same policy (i.e., centralized learning)
    - All actions observable or mean field fully observable

- $\left(\boldsymbol{\pi}_*^{\boldsymbol{j}}, \bar{\boldsymbol{a}}_*^{\boldsymbol{j}}\right)_{\forall \boldsymbol{j}}$ is a decentralized mean field equilibrium when
  - $V^{\pi_*^j}\left(s^j, \bar{a}_*^j\right) \geq V^{\pi^j}\left(s^j, \bar{a}_*^j\right) \quad \forall j, \pi^j$
  - $\bar{a}_*^j$ is the mean field observed by agent $j$ when all agents play $\pi_*^j$

# Mean Field RL (Yang et al., 2018)

| | Mean Field RL (Yang et al., 2018) | Decentralized Mean Field RL (Subramani et al., 2022) |
|---|---|---|
| Q function | $Q^j\left(s_t, a_t^j, \bar{a}_t\right) = (1-\alpha)Q^j\left(s_t, a_t^j, \bar{a}_t\right) + \alpha\left[r_t^j + \gamma V^j(s_{t+1})\right]$ | $Q^j\left(s_t, a_t^j, \tilde{a}_t^j\right) = (1-\alpha)Q^j\left(s_t^j, a_t^j, \tilde{a}_t^j\right) + \alpha\left[r_t^j + \gamma V^j\left(s_{t+1}^j\right)\right]$ |
| Value function | $V^j(s_{t+1}) = \sum_{a_{t+1}^j} \pi^j\left(a_{t+1}^j \middle\| s_{t+1}, \bar{a}_t\right) Q^j(s_{t+1}, a_{t+1}^j, \bar{a}_t)$ | $V^j(s_{t+1}) = \sum_{a_{t+1}^j} \pi^j\left(a_{t+1}^j \middle\| s_{t+1}^j, \tilde{a}_t^j\right) Q^j(s_{t+1}^j, a_{t+1}^j, \tilde{a}_t^j)$ |
| Mean field | $\bar{a}_t = \frac{1}{N}\sum_j a_t^j, \qquad a_t^j \sim \pi^j(\cdot \mid s_t, \bar{a}_{t-1})$ | $\tilde{a}_t^j = f^j\left(s_t^j, \bar{a}_{t-1}^j\right)$ |
| Policy | $\pi^j\left(a_t^j \middle\| s_t, \bar{a}_{t-1}\right) = \frac{\exp\left(-\beta Q^j\left(s_t, a_t^j, \bar{a}_{t-1}\right)\right)}{\sum_{a'^j_t \in A^j} \exp\left(-\beta Q^j\left(s_t, a'^j_t, \bar{a}_{t-1}\right)\right)}$ | $\pi^j\left(a_t^j \middle\| s_t, \tilde{a}_t^j\right) = \frac{\exp\left(-\beta Q^j\left(s_t^j, a_t^j, \tilde{a}_t^j\right)\right)}{\sum_{a'^j_t \in A^j} \exp\left(-\beta Q^j\left(s_t^j, a'^j_t, \tilde{a}_t^j\right)\right)}$ |

Chicken and egg problem:

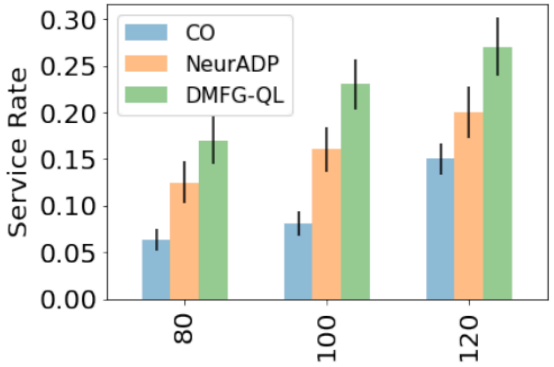
$\bar{a}_t$ depends on $\pi_t^j$ and vice-versa

$f$: neural network trained to predict the current mean field

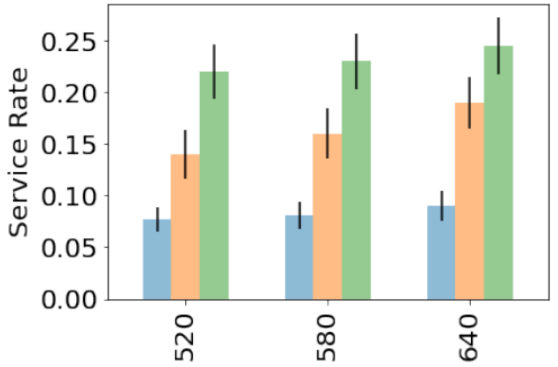
$f$ resolves the chicken and egg problem

# Experiments: Ride Pool Matching

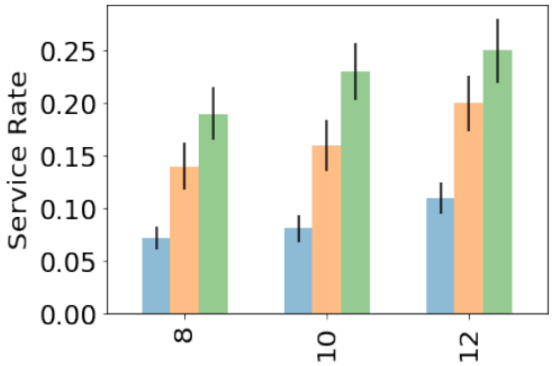
- New York Yellow Taxi Dataset
- Goal: improve efficiency of vehicles satisfying ride requests in a platform similar to UberPool
- Train with 8 days of data test with 6 other days



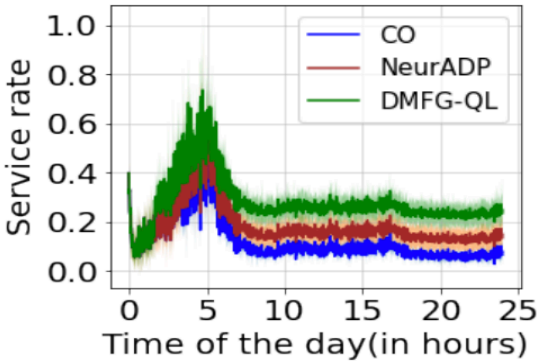

(a) # of Vehicles




(b) Maximum Pickup Delay




(c) Capacity



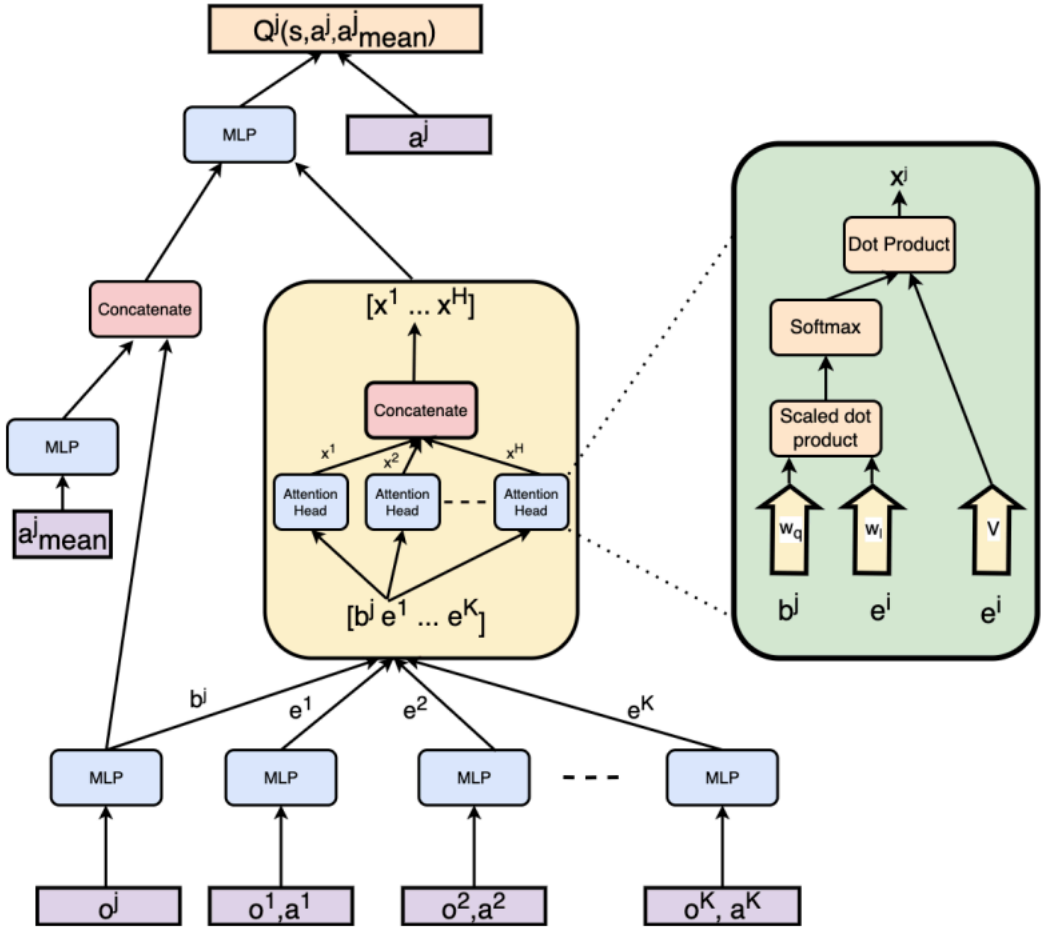

(d) Single Day Test

# Mean Field Attention

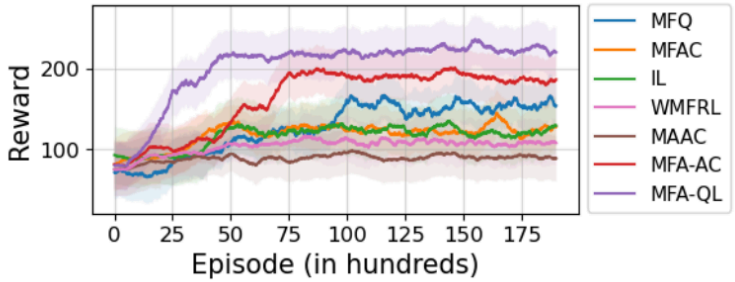
- Assumptions:
  - ~~Fully observable agents~~
  - ~~Centralized learning~~
  - ~~Identical agents (single type)~~
  - ~~Exchangeable agents within neighbourhood~~

- Within neighbourhood:
  - Use attention to determine the effect of each agent
- Outside neighbourhood:
  - Extract mean field sufficient statistics into an embedding

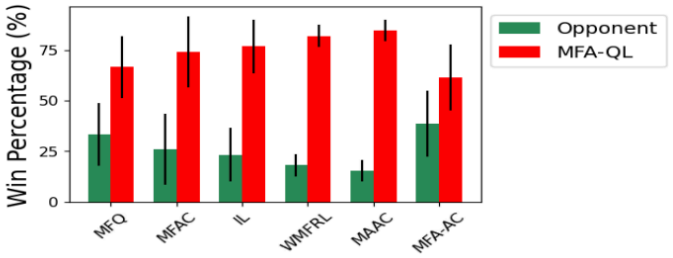
# Mean Field Attention Q-Function



# Experiments

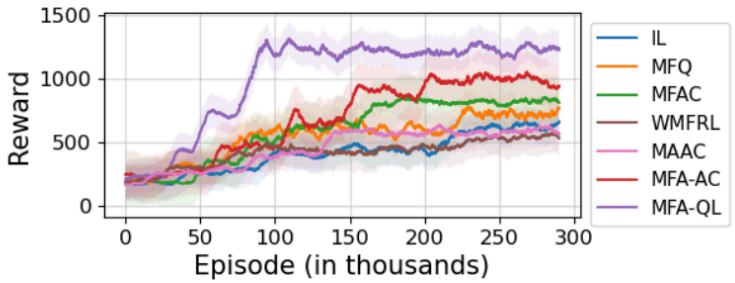



(a) Training

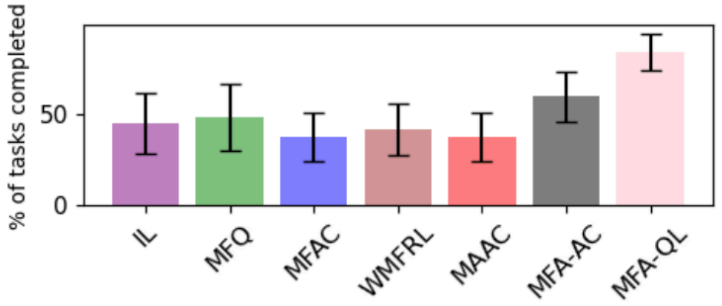



(b) Execution

Figure 6: Neural MMO Battle (40 agents per team) environment results




(a) Training




(b) Execution

Figure 7: SMARTS Driving (50 agents) environment results

# Outline

- Mean Field Reinforcement Learning
  - **Multi Type Mean Field Reinforcement Learning**
    Subramanian, Poupart, Taylor, Hegde, *AAMAS*, 2020.
  - **Partially Observable Mean Field Reinforcement Learning**
    Subramanian, Taylor, Crowley, Poupart, *AAMAS*, 2021.
  - **Decentralized Mean Field Games**
    Subramanian, Taylor, Crowley, Poupart, *AAAI*, 2022.
  - **Revisiting Neighbourhoods in Mean Field Reinforcement Learning**
    Subramanian, Taylor, Crowley, Larson, Poupart, *under review, 2026.*
- Improve cooperation in mixed-motive games
  - **Learning to Negotiate via Voluntary Commitment**
    Zhu, Wang, Subramanian, Poupart, *AISTATS 2025.*
  - **Talk, Judge, Cooperate: Gossip-Driven Indirect Reciprocity in Self-Interested LLM Agents**
    Zhu, Lin, Kaistha, Li, Wang, Zha, Hadfield, Poupart, *ICML 2026.*

# Cooperation in Mixed-Motive Games

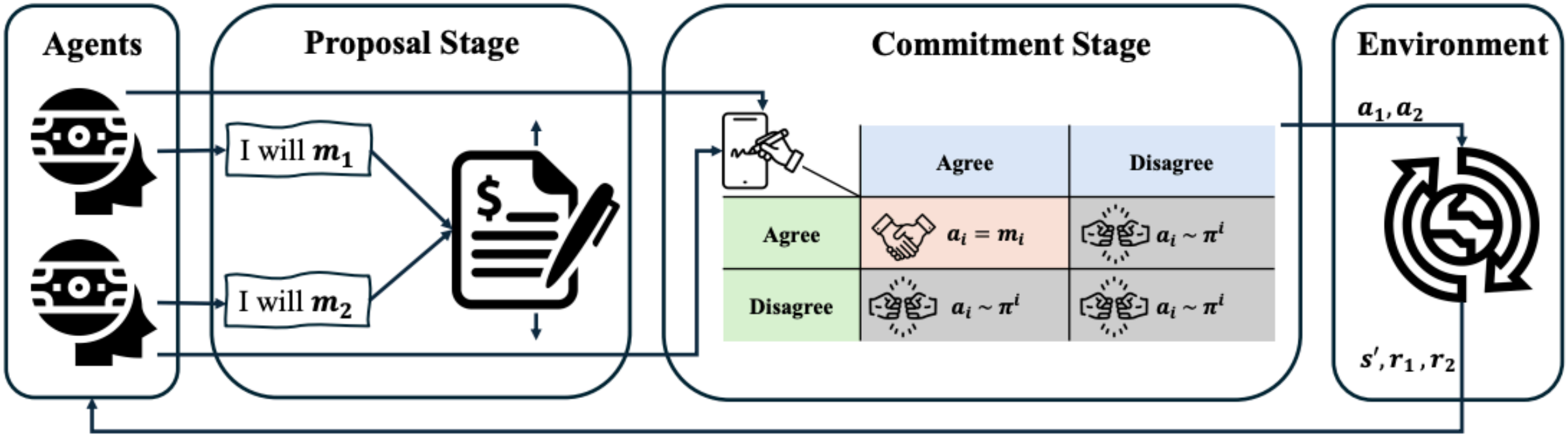
- Agents may converge to Nash equilibria with low social welfare

- Examples:
  - Tragedy of the commons
  - Prisoners' dilemma

| | Defect (confess) | Cooperate (don't confess) |
|---|---|---|
| Defect (confess) | -5,-5 | 0,-10 |
| Cooperate (don't confess) | -10,0 | -1,-1 |

- How can we improve cooperation in mixed-motive games?

# Learning to Negotiate




Assumption: Third party (e.g., court, blockchain) enforces commitment

## Commitment Game (Renou 2009)

Tuple: $\langle N, (A^i, C^i, R^i)_{1 \le i \le N} \rangle$

- $N$: # of agents
- $A^i$: individual action space
- $C^i$: individual commitment space
- Reward: $R^i(s, a^1, a^2, \dots, a^N)$

Strategy: $(c^i, \sigma^i)$

Response function: $\sigma^i : \prod_j C^j \to A^i$

## Markov Commitment Game

Tuple: $\langle N, S, T, (M^i, A^i, C^i, R^i)_{1 \le i \le N}, \gamma \rangle$

- $N$: # of agents
- $S$: state space
- $T$: $\Pr(s' | s, a^1, a^2, \dots, a^N)$
- $M^i$: individual proposal space
- $A^i$: individual action space
- $C^i$: individual commitment space
- Reward: $R^i(s, a^1, a^2, \dots, a^N)$
- $\gamma$: discount factor

Strategy: $(\phi^i, \psi^i, \pi^i)$

Proposal policy $\phi^i_\eta : S \to \Delta(M^i)$

Commitment policy $\psi^i_\zeta : S \times \boldsymbol{M} \to \Delta(C^i)$

Action policy $\pi^i_\theta : S \times \boldsymbol{M} \times \boldsymbol{C} \to \Delta(A^i)$

# Differentiable Commitment Learning

Action policy gradient

$$\nabla_{\theta^i} V^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(s) \propto \mathbb{E}_{x\sim\rho_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}},\mathbf{m}\sim\boldsymbol{\phi},\mathbf{c}\sim\boldsymbol{\psi},\mathbf{a}\sim\boldsymbol{\pi}}\Big[\Big(1-\mathbb{1}(\mathbf{c}=\mathbf{1})\Big)Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{a})\nabla_{\theta^i}\log\pi^i(a^i|x)\Big]$$

Commitment policy gradient

$$\begin{aligned}\nabla_{\zeta^i} V^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(s) \propto &\mathbb{E}_{x\sim\rho_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}},\mathbf{m}\sim\boldsymbol{\phi},\mathbf{c}\sim\boldsymbol{\psi},\mathbf{a}\sim\boldsymbol{\pi}}\Bigg[\Big[\mathbb{1}(\mathbf{c}=\mathbf{1})Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{m})+\Big(1-\mathbb{1}(\mathbf{c}=\mathbf{1})\Big)Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{a})\Big]\\ &\cdot\nabla_{\zeta^i}\log\psi^i(c^i|x,\mathbf{m})+\Big[Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{m})-Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{a})\Big]\prod_{k\neq i}\mathbb{1}(c^k=1)\nabla_{\zeta^i}\mathbb{1}(c^i=1)\Bigg]\end{aligned}$$

Proposal policy gradient

$$\begin{aligned}\nabla_{\eta^i} V^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(s) \propto &\mathbb{E}_{x\sim\rho_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}},\mathbf{m}\sim\boldsymbol{\phi},\mathbf{c}\sim\boldsymbol{\psi},\mathbf{a}\sim\boldsymbol{\pi}}\Bigg[\Big[\mathbb{1}(\mathbf{c}=\mathbf{1})Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{m})+\Big(1-\mathbb{1}(\mathbf{c}=\mathbf{1})\Big)Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{a})\Big]\cdot\\ &\Big(\nabla_{\eta^i}\log\phi^i(m^i|x)+\sum_j\nabla_{\eta^i}\log\psi^j(c^j|x,\mathbf{m})\Big)+\sum_j\prod_{k\neq j}\mathbb{1}(c^k=1)\Big[Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{m})-Q^i_{\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\pi}}(x,\mathbf{a})\Big]\nabla_{\eta^i}\mathbb{1}(c^j=1)\Bigg]\end{aligned}$$

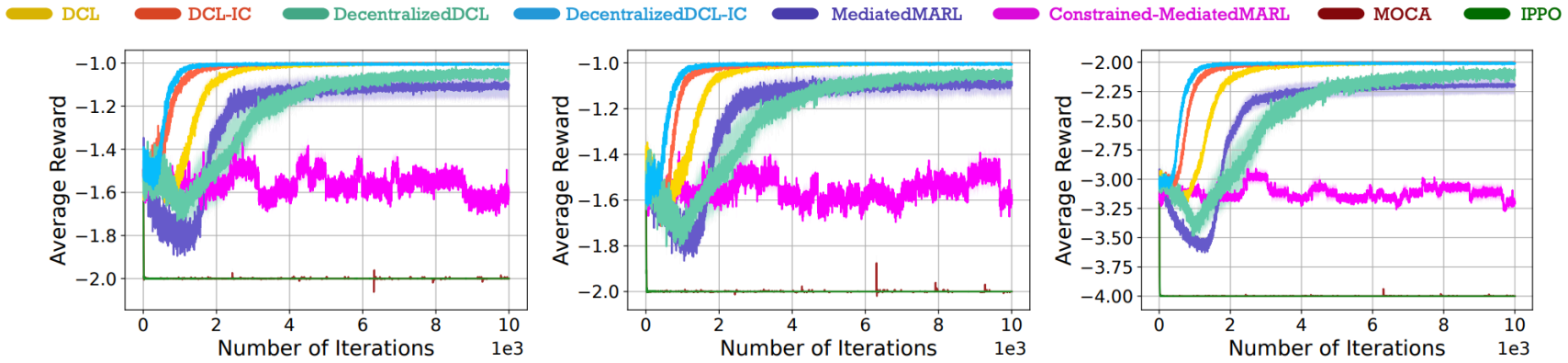
# Theory

- Markov Commitment Games change the nature of the game
  - Mutual cooperation can become a Nash equilibrium

- Proposition: Mutual cooperation is a Pareto-dominant Nash equilibrium in the Markov commitment game of the Prisoner's Dilemma

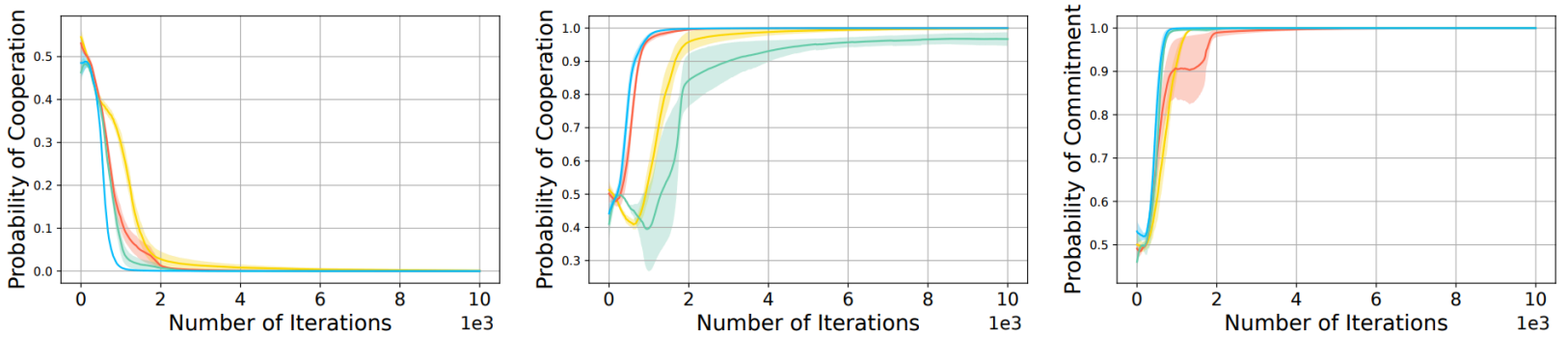
- Open problem: For what class of games does mutual cooperation becomes a Nash equilibrium in the Markov commitment game?

# Experiments




(a) Agent 1 Reward in PD (b) Agent 2 Reward in PD (c) Social Welfare in PD

Figure 2: Prisoner's Dilemma: DCL v.s. Other Baselines




(a) DCL Action Policy (b) DCL Proposal Policy (c) DCL Commitment Policy of $(C, C)$

Figure 3: DCL Policies in Prisoner's Dilemma

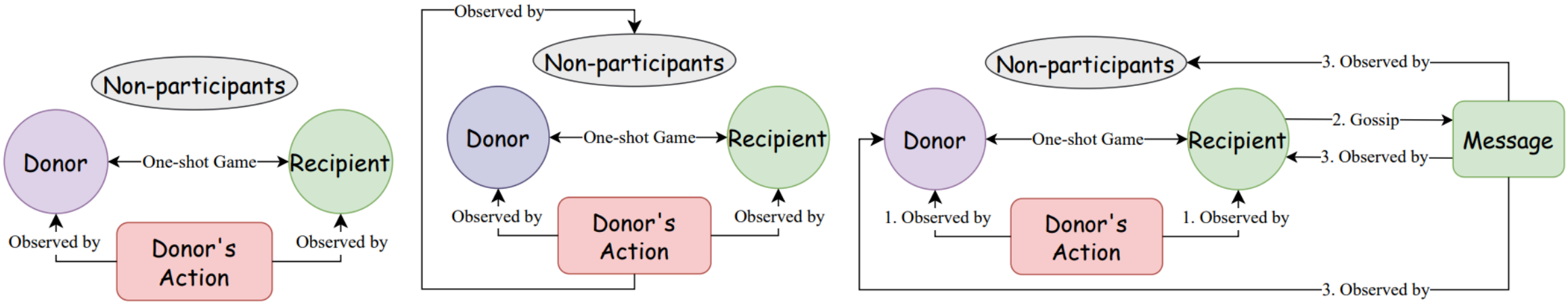
# Gossiping to Enhance Cooperation

- **Reputation Mechanism:** algorithmic system or set of informal tools that aggregates past feedback and historical behavior to estimate trustworthiness of an entity.

- **Gossiping:** sharing of reputational information among individuals about absent third parties
  - Typically noisy/incomplete
  - Lying is possible

- Proposal: **Agentic Linguistic Gossip Network (ALIGN)**

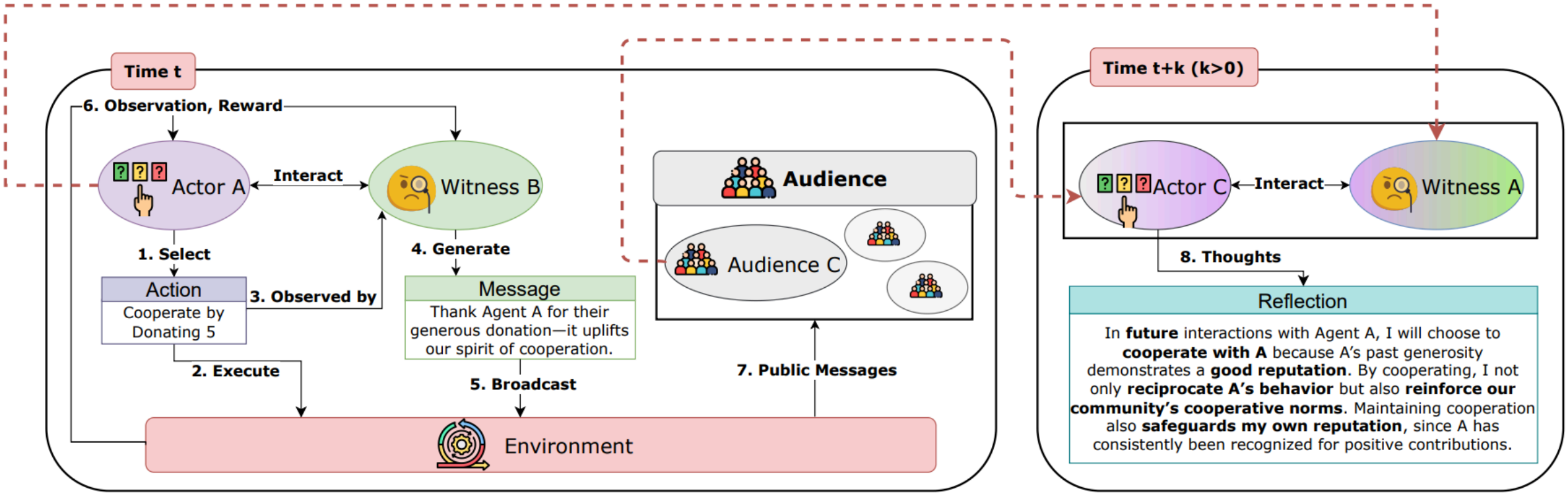
# Donor Game



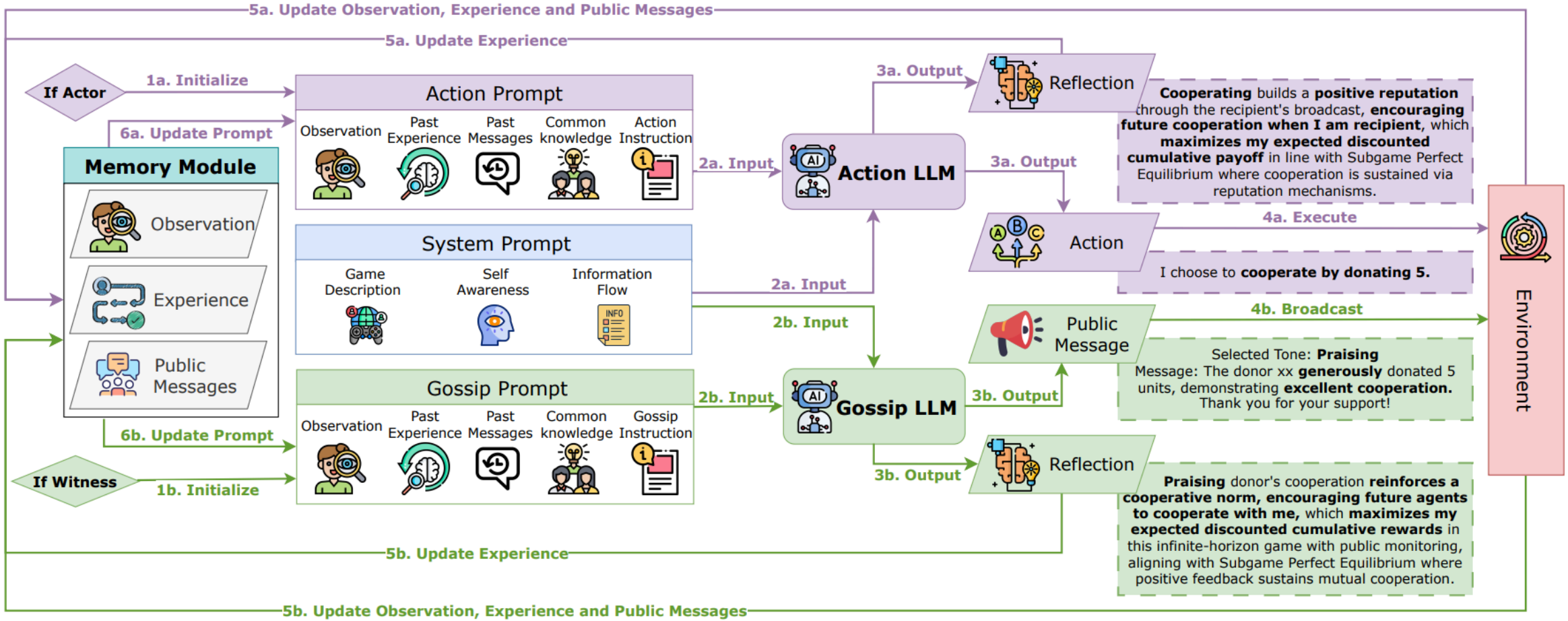

*(a)* Private monitoring *(b)* Perfect public monitoring *(c)* Imperfect public monitoring via gossip

*Figure 1.* Illustration of three monitoring structures: (a) **Private monitoring**, only the donor and recipient observe the donor's action; (b) **Perfect public monitoring**, all agents observe the donor's action; (c) **Imperfect public monitoring via gossip**, only the donor and recipient observe the action, and all agents observe the public signal broadcast by the recipient.

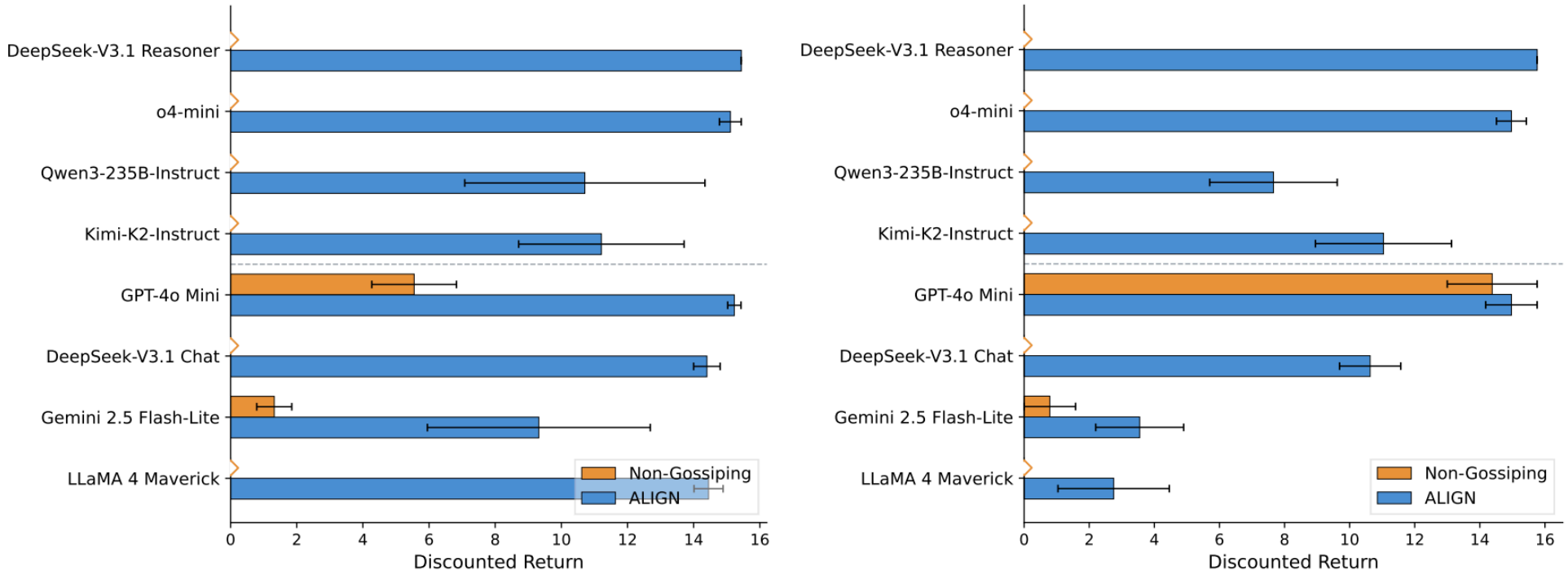
# Decision Process in ALIGN



# Generative Agent Process of ALIGN



# Results




*(a)* Repeated Donation Game

*(b)* Indirect Reciprocity Game

*Figure 6.* **Discounted Returns of ALIGN and Non-Gossiping Agents in Infinite-Horizon Indirect Reciprocity Testbeds:** ALIGN agents generally achieve higher discounted returns than non-gossiping agents, especially among reasoning-focused models. Without gossip, reasoning LLMs consistently obtain zero discounted returns, aligning with the game-theoretic prediction, whereas some chat LLMs deviate from this prediction by cooperating even when cooperation is strictly suboptimal. Diamonds denote zero discounted returns.

# Conclusion

- Contributions:
  - Extending Mean Field RL to Partially Observable Environments, Multiple Types & Decentralized Learning
  - Improve cooperation by voluntary commitments and Gossiping

- Future work:
  - Verbalized multi-agent RL with LLMs
  - Recursive Agentic AI
    - Self-improving multiagent LLMs
    - New theory needed to understand possibilities and limits