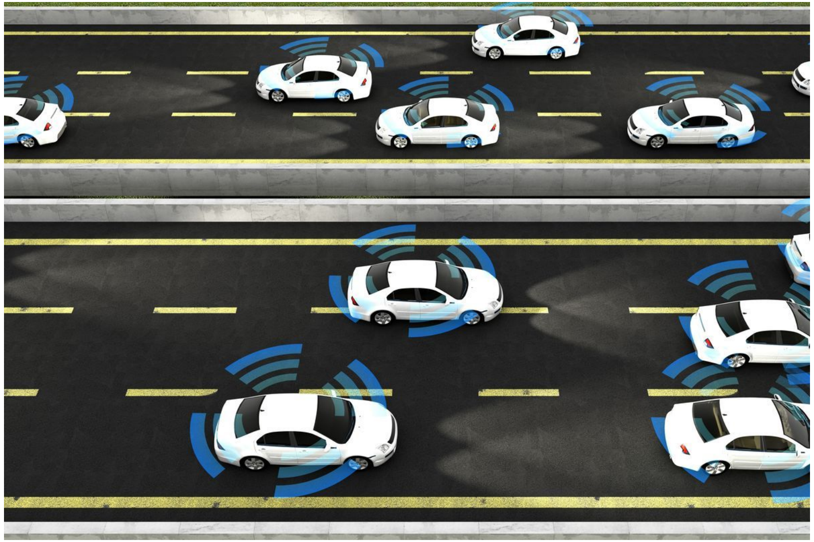
Model-Free Opponent Shaping

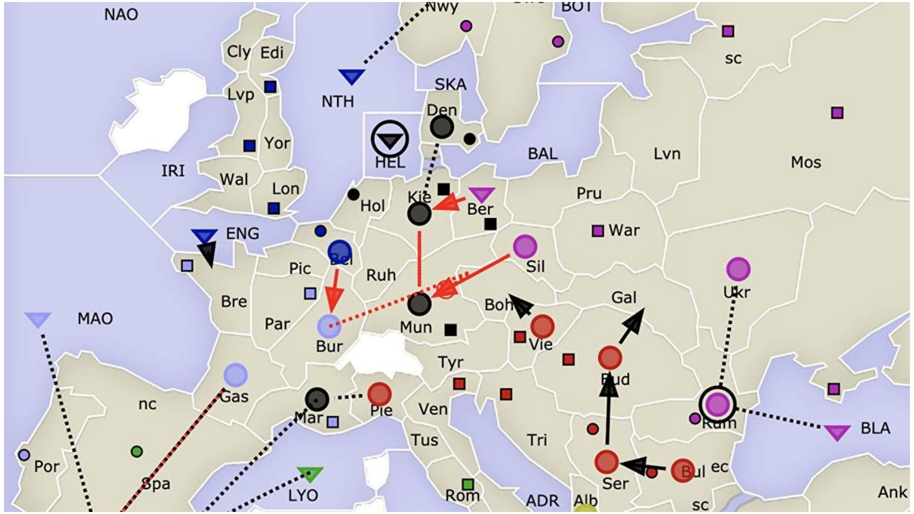
Chris Lu, Timon Willi, Christian Schroeder de Witt, Jakob Foerster

FLAIR Foerster Lab for AI Research

UNIVERSITY OF OXFORD

General-Sum Games





Iterated Matrix Games: IPD

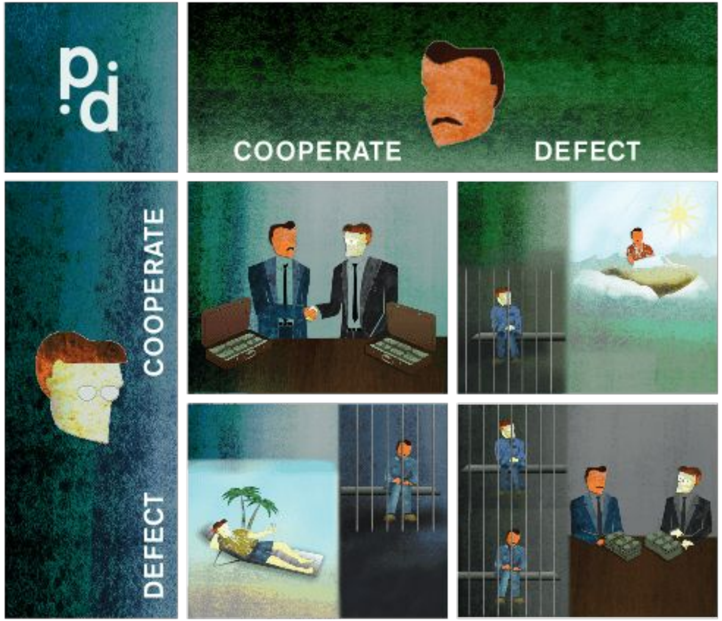


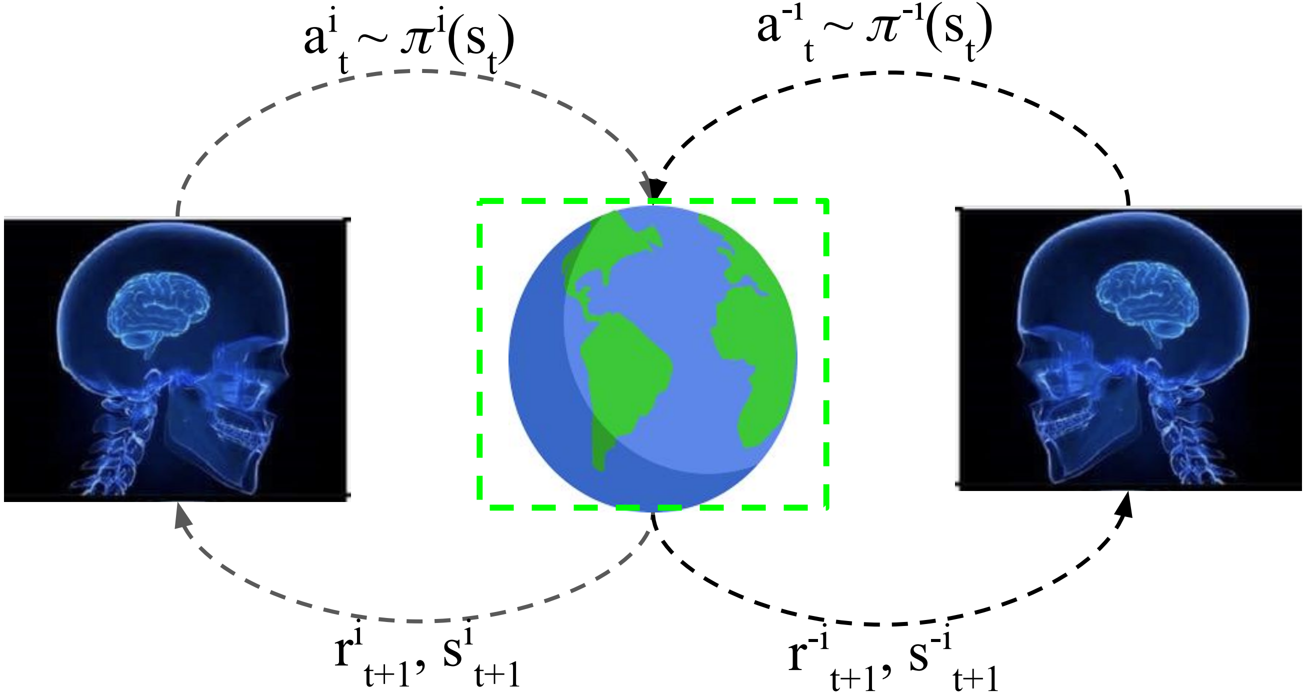
Table 1. Payoff Matrix for the Prisoner's Dilemma

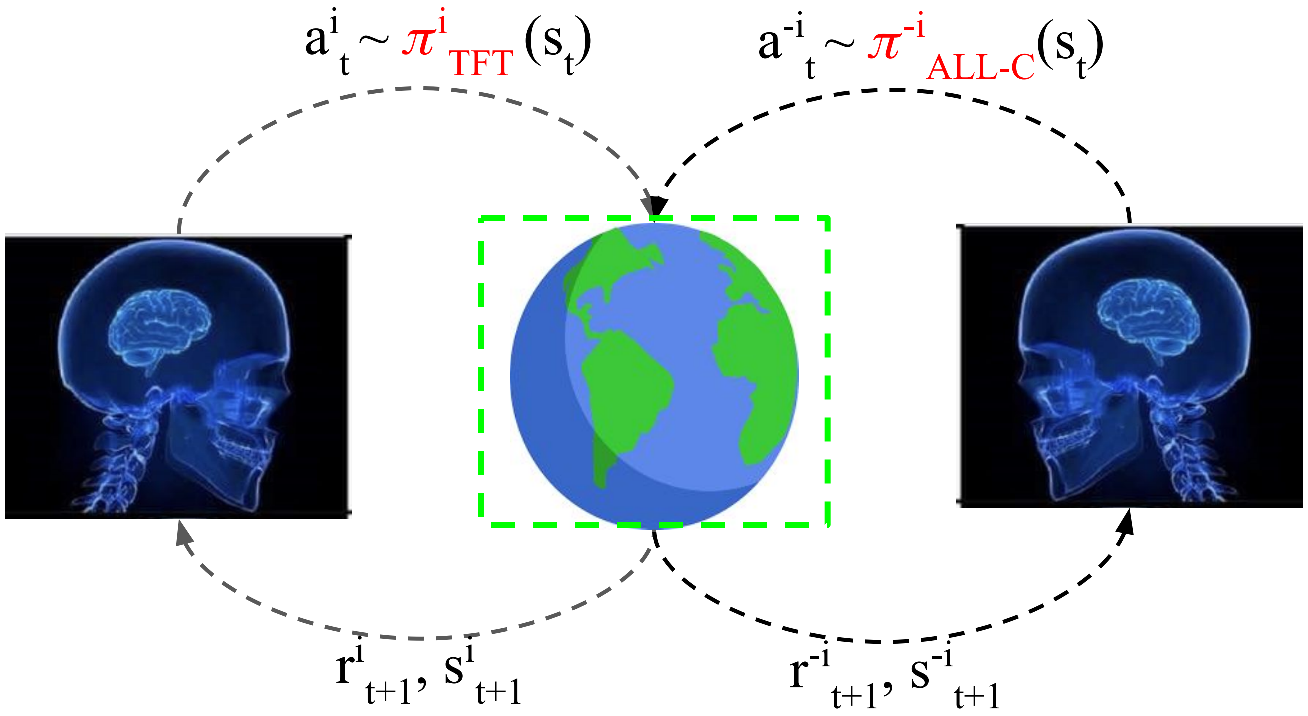
	C	D
C	(-1, -1)	(-3, 0)
D	(0, -3)	(-2, -2)

Iterated Prisoner's Dilemma:

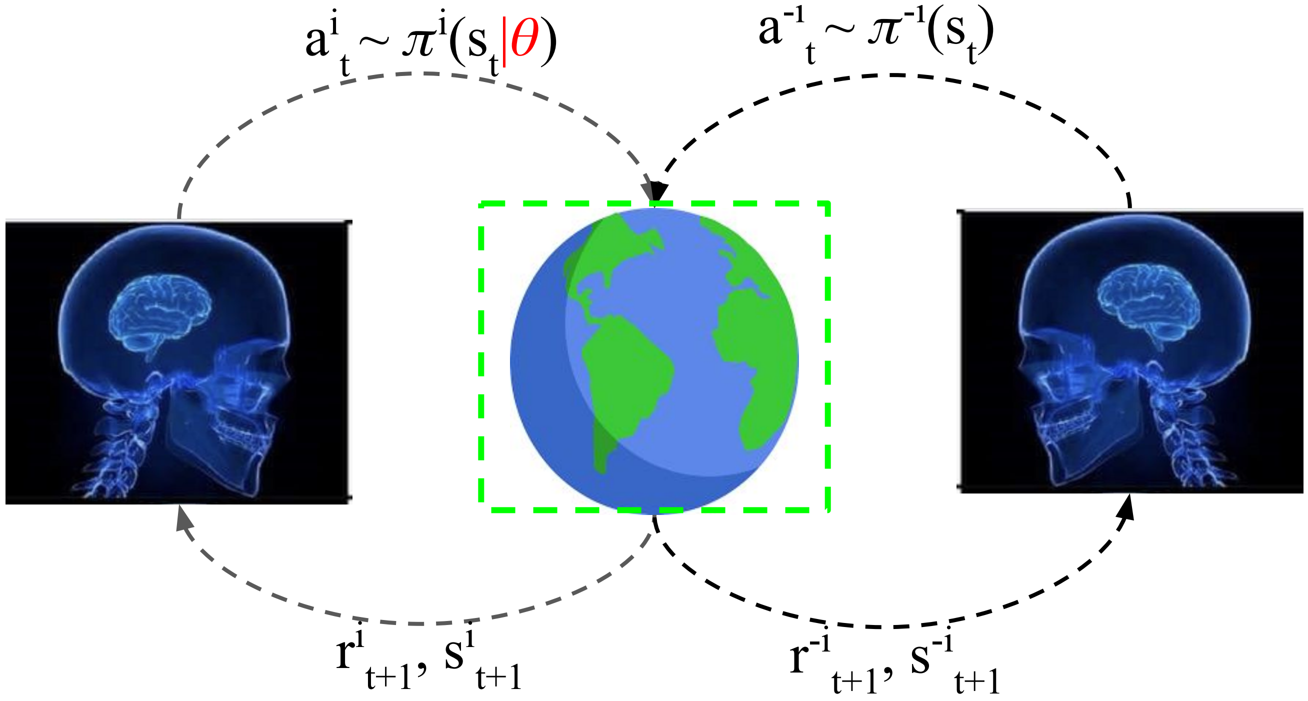
- We consider the iterated game
- Game states are the outcome of the **previous** round
  - P0, CC, CD, DC, DD

Iterated Matrix Games: IPD

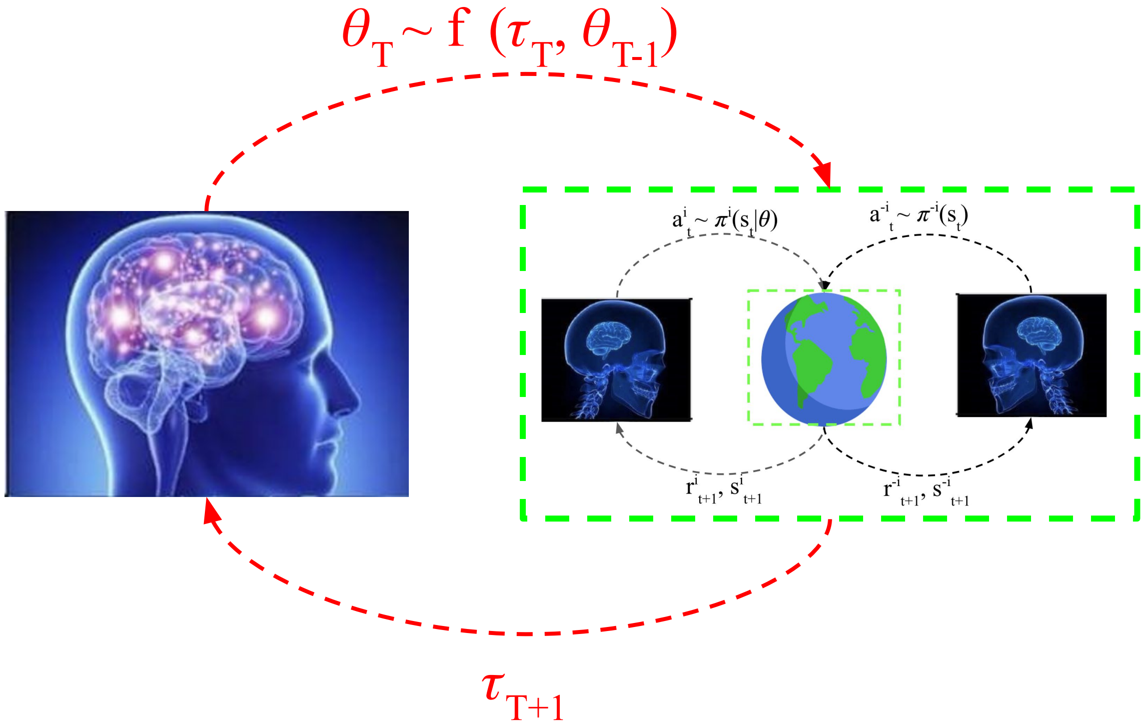




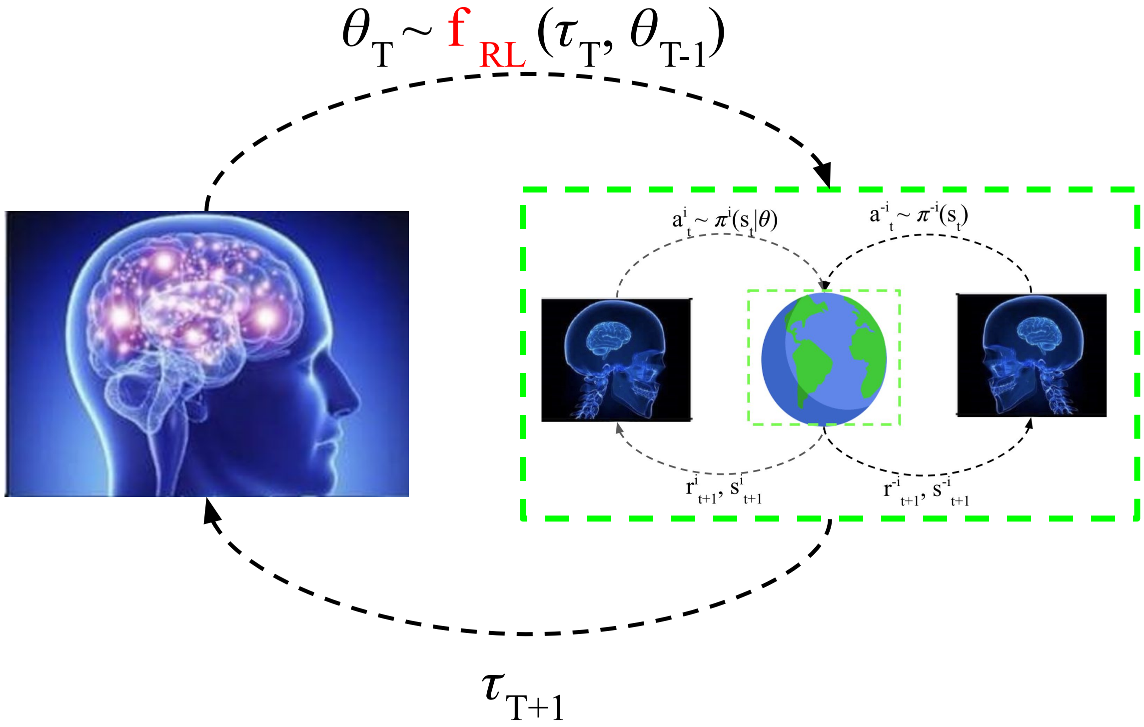
$a^i_t \sim \pi^i_{TFT}(s_t)$
$a^{-i}_t \sim \pi^{-i}_{ALL\text{-}C}(s_t)$
r^i_{t+1}, s^i_{t+1}
$r^{-i}_{t+1}, s^{-i}_{t+1}$

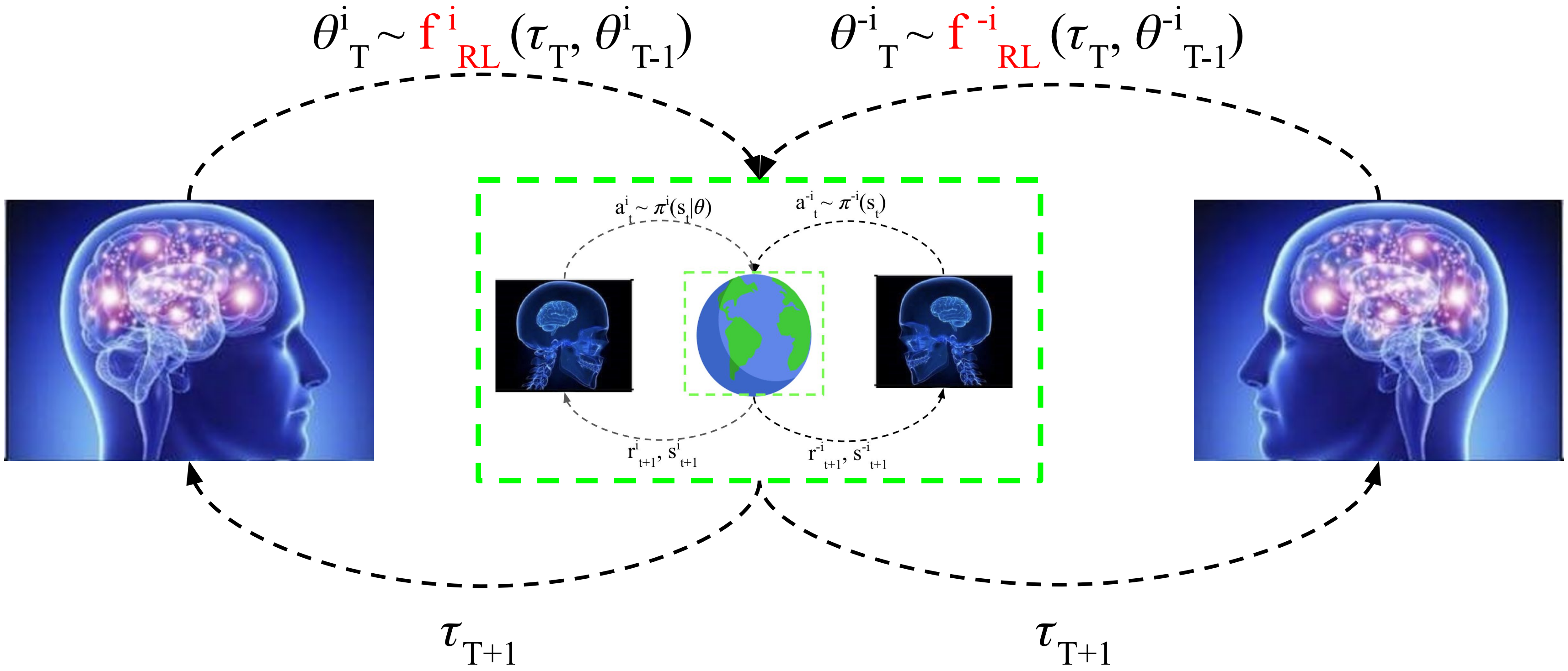


$a^i_t \sim \pi^i(s_t|\theta)$
$a^{-i}_t \sim \pi^{-i}(s_t)$
r^i_{t+1}, s^i_{t+1}
$r^{-i}_{t+1}, s^{-i}_{t+1}$

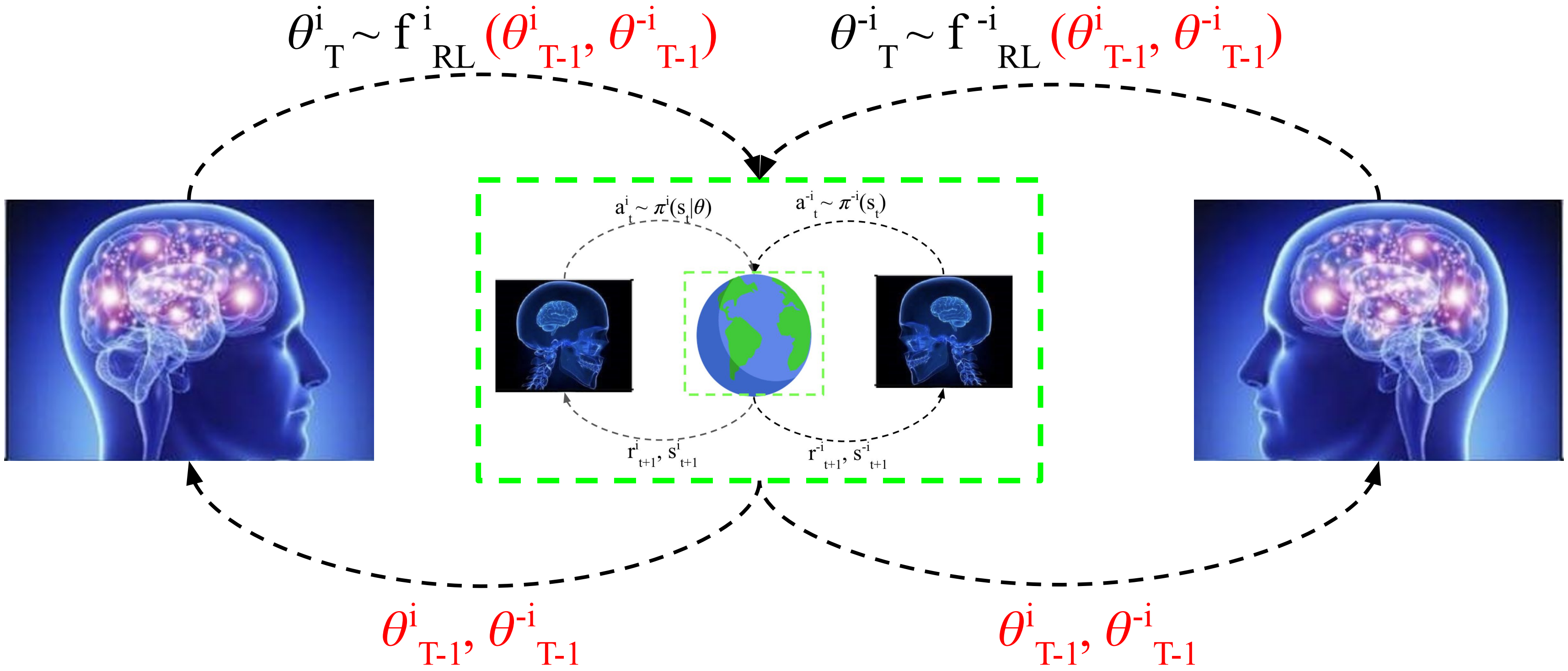


$\theta_T \sim f(\tau_T, \theta_{T-1})$
$a^i_t \sim \pi^i(s_t|\theta)$
$a^{-i}_t \sim \pi^{-i}(s_t)$
r^i_{t+1}, s^i_{t+1}
$r^{-i}_{t+1}, s^{-i}_{t+1}$
τ_{T+1}


$\theta_T \sim f_{RL}(\tau_T, \theta_{T-1})$
$a^i_t \sim \pi^i(s_t|\theta)$
$a^{-i}_t \sim \pi^{-i}(s_t)$
r^i_{t+1}, s^i_{t+1}
$r^{-i}_{t+1}, s^{-i}_{t+1}$
τ_{T+1}



$\theta^{i}_{T} \sim f^{i}_{RL}(\tau_T, \theta^{i}_{T-1})$
$\theta^{-i}_{T} \sim f^{-i}_{RL}(\tau_T, \theta^{-i}_{T-1})$
$a^{i}_{t} \sim \pi^{i}(s_t|\theta)$
$a^{-i}_{t} \sim \pi^{-i}(s_t)$
r^{i}_{t+1}, s^{i}_{t+1}
$r^{-i}_{t+1}, s^{-i}_{t+1}$
τ_{T+1}
τ_{T+1}



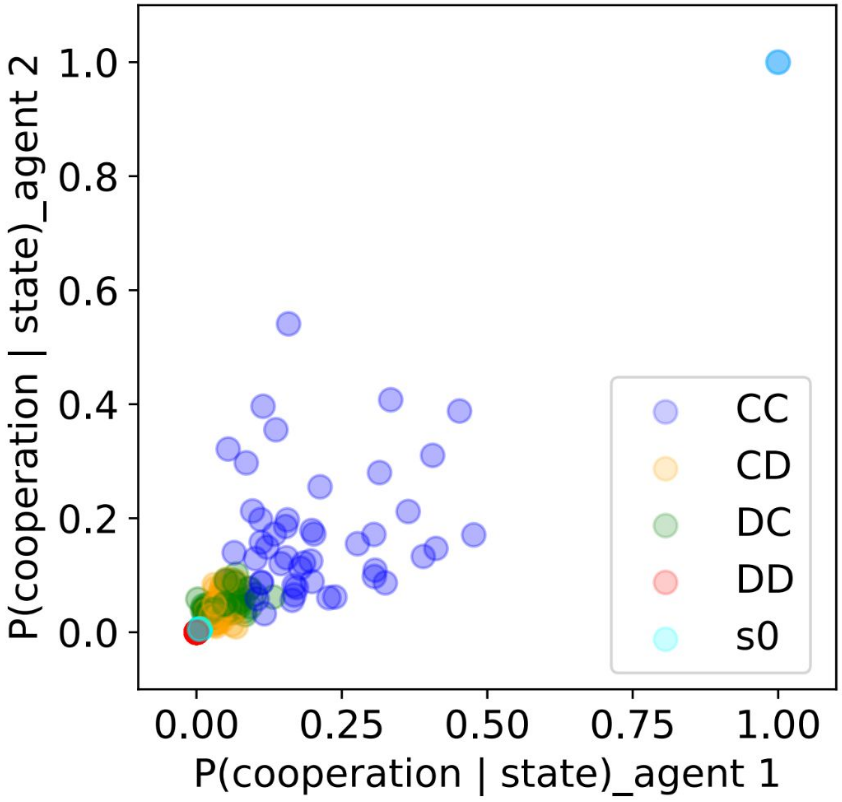
$\theta^{i}_{T} \sim f^{i}_{RL}(\theta^{i}_{T-1}, \theta^{-i}_{T-1})$
$\theta^{-i}_{T} \sim f^{-i}_{RL}(\theta^{i}_{T-1}, \theta^{-i}_{T-1})$
$a^{i}_{t} \sim \pi^{i}(s_t|\theta)$
$a^{-i}_{t} \sim \pi^{-i}(s_t)$
r^{i}_{t+1}, s^{i}_{t+1}
$r^{-i}_{t+1}, s^{-i}_{t+1}$
$\theta^{i}_{T-1}, \theta^{-i}_{T-1}$
$\theta^{i}_{T-1}, \theta^{-i}_{T-1}$

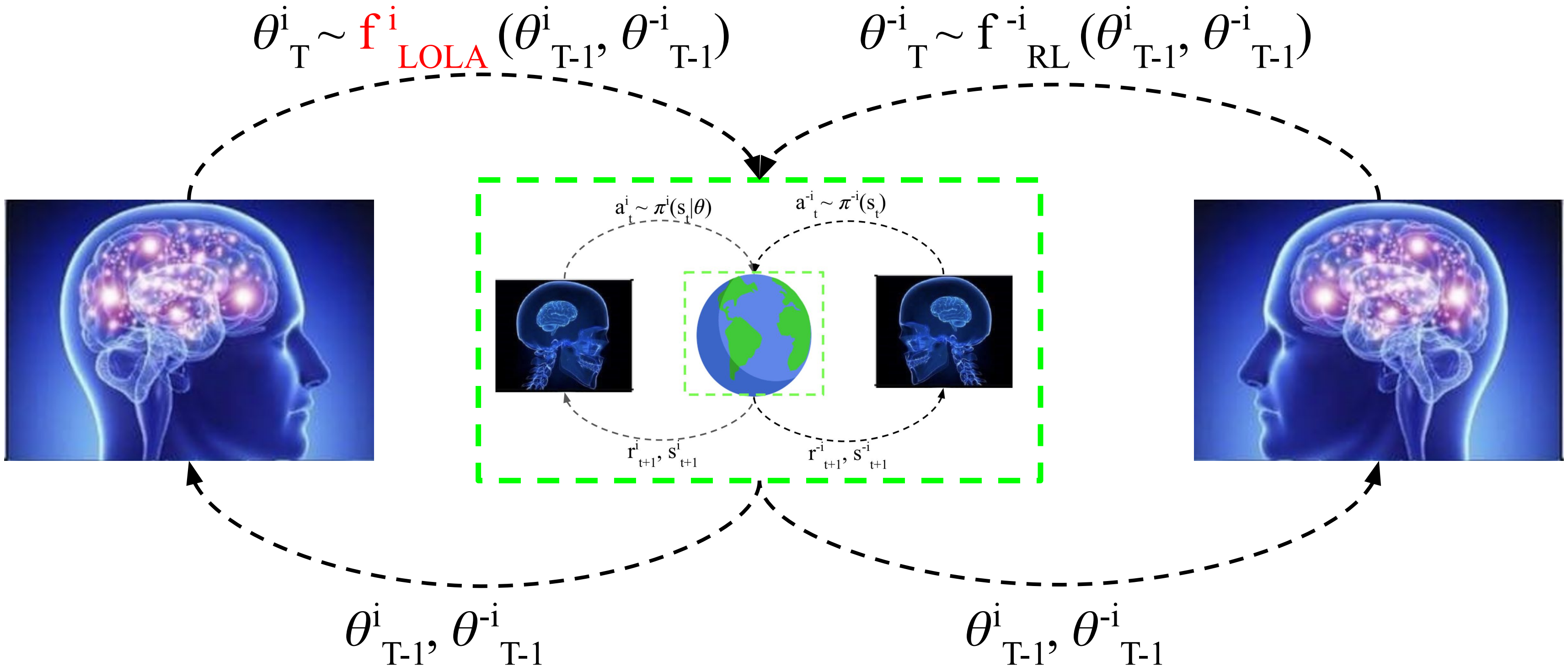
Naive Update

$$\theta^{i}_{T} \sim f_{RL}(\theta^{i}_{t-1}, \theta^{-i}_{t-1})$$

$$f_{RL}(\theta^{i}_{t-1}, \theta^{-i}_{t-1}) = \theta^{i}_{t-1} + \alpha \triangledown_{\boldsymbol{\theta}} V^{i}(\theta^{i}_{t-1}, \theta^{-i}_{t-1})$$

Naive Learning Results



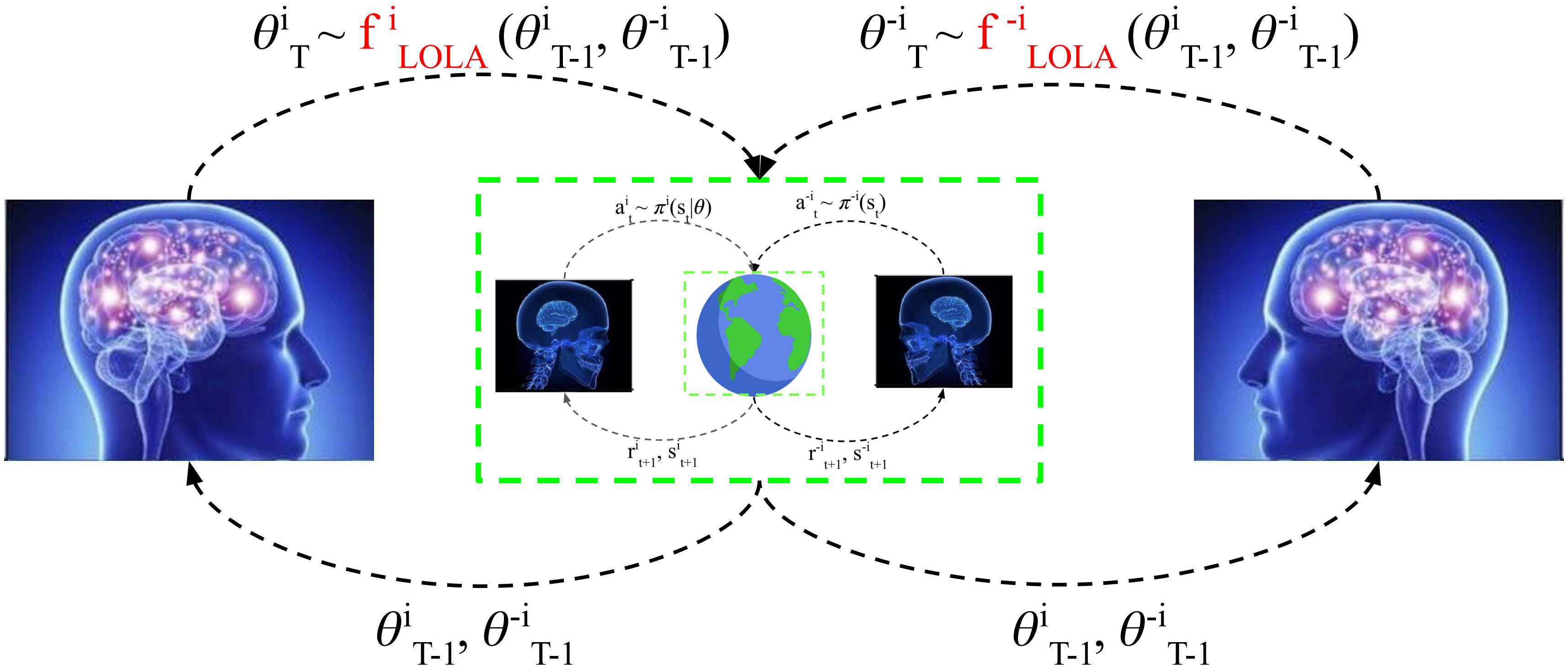


$\theta^i_T \sim f^i_{LOLA}(\theta^i_{T-1}, \theta^{-i}_{T-1})$
$\theta^{-i}_T \sim f^{-i}_{RL}(\theta^i_{T-1}, \theta^{-i}_{T-1})$
$a^i_t \sim \pi^i(s_t|\theta)$
$a^{-i}_t \sim \pi^{-i}(s_t)$
r^i_{t+1}, s^i_{t+1}
$r^{-i}_{t+1}, s^{-i}_{t+1}$
$\theta^i_{T-1}, \theta^{-i}_{T-1}$
$\theta^i_{T-1}, \theta^{-i}_{T-1}$

LOLA Update

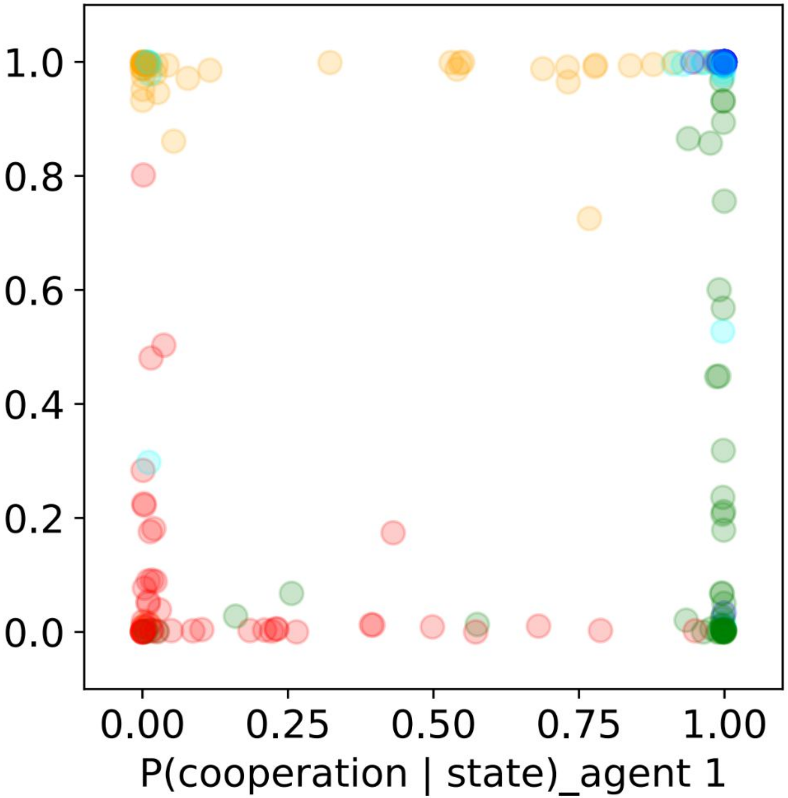
$$\theta^{i}_{T} \sim f_{LOLA}(\theta^{i}_{t-1}, \theta^{-i}_{t-1})$$

$$f_{LOLA}(\theta^{i}_{t-1}, \theta^{-i}_{t-1}) = \theta^{i}_{t-1} + \alpha \bigtriangledown_{\boldsymbol{\theta}} V^{i}(\theta^{i}_{t-1}, f_{RL}(\theta^{-i}_{t-1}, \theta^{i}_{t-1}))$$



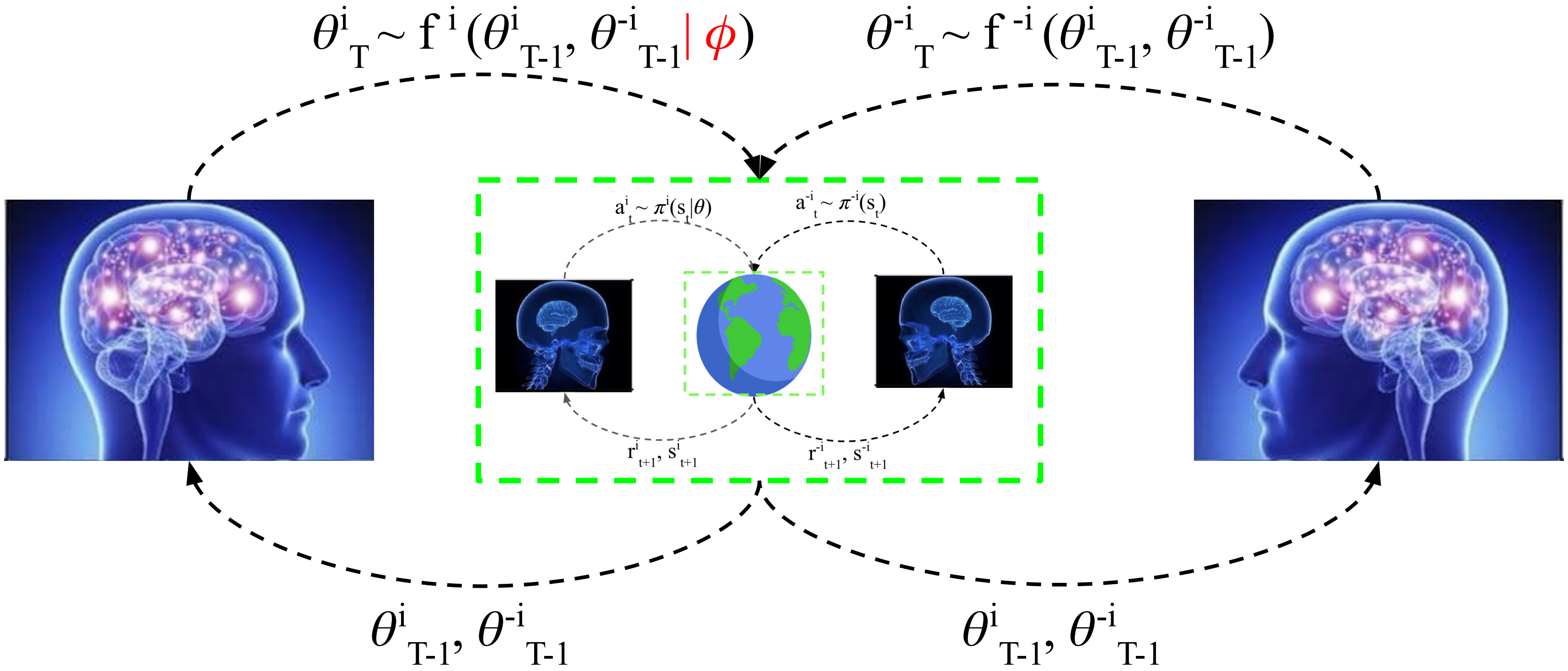
$\theta^{i}_{T} \sim f^{i}_{LOLA}(\theta^{i}_{T-1}, \theta^{-i}_{T-1})$
$\theta^{-i}_{T} \sim f^{-i}_{LOLA}(\theta^{i}_{T-1}, \theta^{-i}_{T-1})$
$a^{i}_{t} \sim \pi^{i}(s_t|\theta)$
$a^{-i}_{t} \sim \pi^{-i}(s_t)$
r^{i}_{t+1}, s^{i}_{t+1}
$r^{-i}_{t+1}, s^{-i}_{t+1}$
$\theta^{i}_{T-1}, \theta^{-i}_{T-1}$
$\theta^{i}_{T-1}, \theta^{-i}_{T-1}$

LOLA Results: IPD

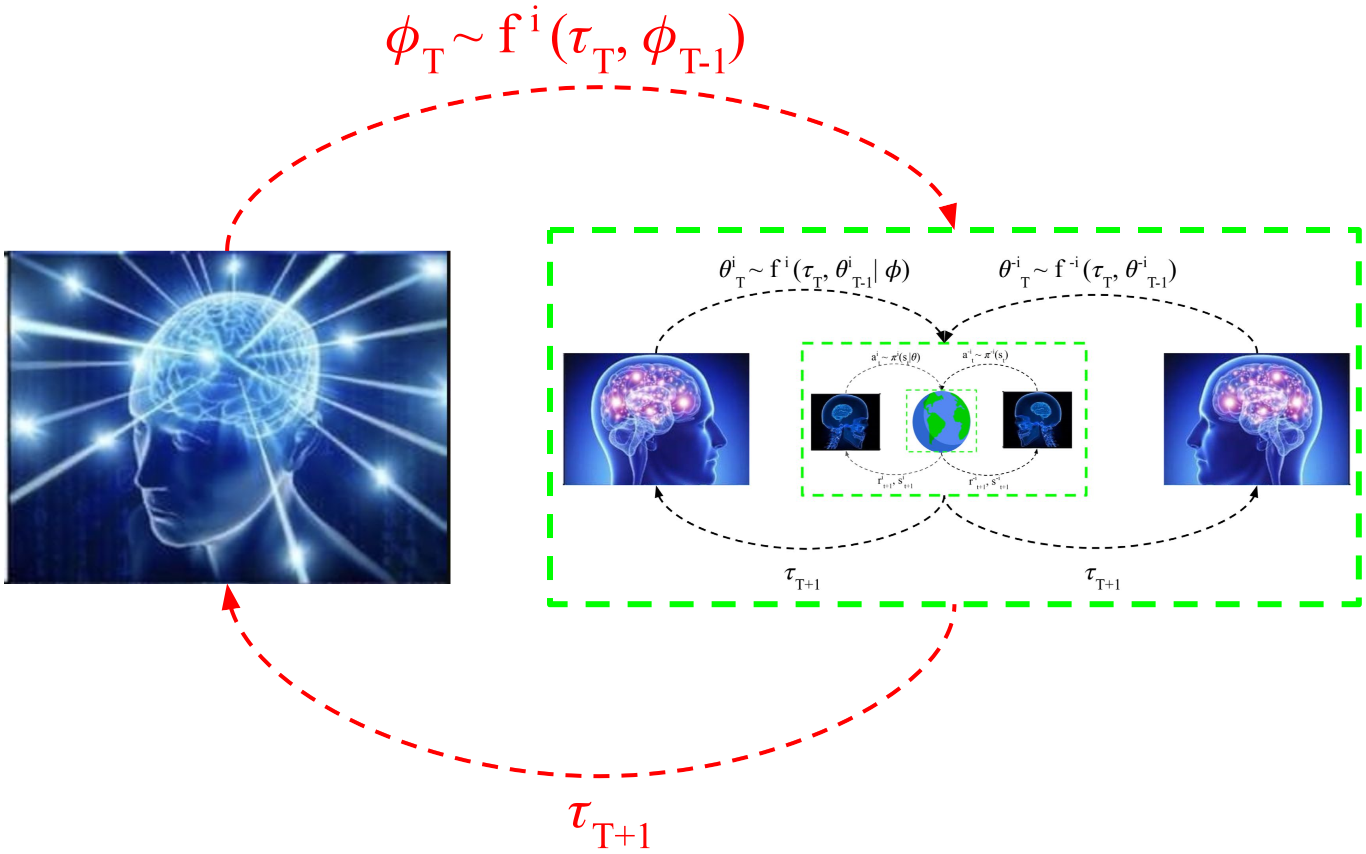


Issues with LOLA

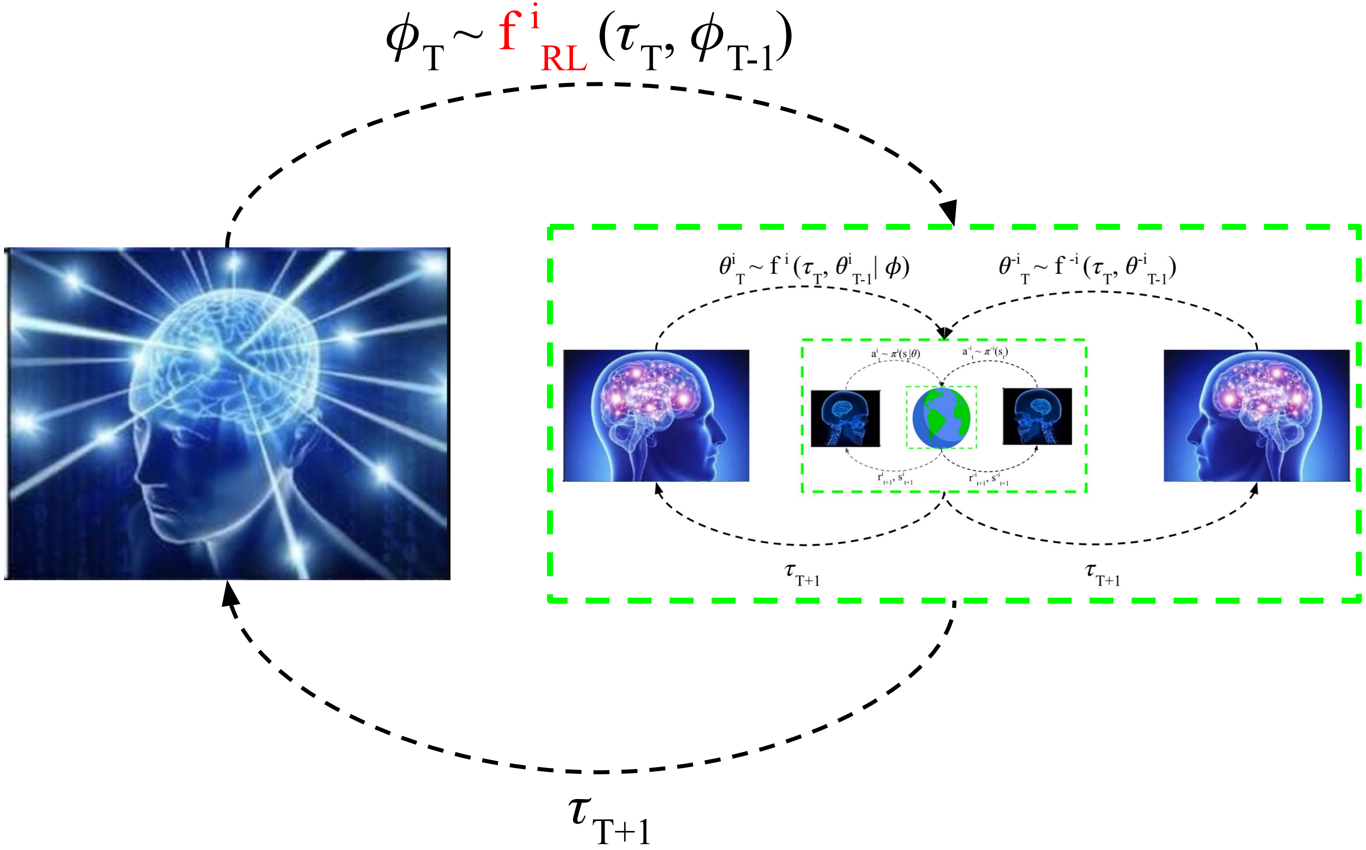
1. Myopic: Only shapes the opponent's next step
2. Inconsistent: Explicitly assumes the opponent is a naive learner
3. Unstable: Uses higher-order derivatives, which can be difficult to estimate



$\theta^{i}_{T} \sim f^{i}(\theta^{i}_{T-1}, \theta^{-i}_{T-1} | \phi)$
$\theta^{-i}_{T} \sim f^{-i}(\theta^{i}_{T-1}, \theta^{-i}_{T-1})$
$a^{i}_{t} \sim \pi^{i}(s_t|\theta)$
$a^{-i}_{t} \sim \pi^{-i}(s_t)$
r^{i}_{t+1}, s^{i}_{t+1}
$r^{-i}_{t+1}, s^{-i}_{t+1}$
$\theta^{i}_{T-1}, \theta^{-i}_{T-1}$
$\theta^{i}_{T-1}, \theta^{-i}_{T-1}$



$\phi_T \sim f^i(\tau_T, \phi_{T-1})$
$\theta^i_T \sim f^i(\tau_T, \theta^i_{T-1} \mid \phi)$
$\theta^{-i}_T \sim f^{-i}(\tau_T, \theta^{-i}_{T-1})$
τ_{T+1}
τ_{T+1}
τ_{T+1}

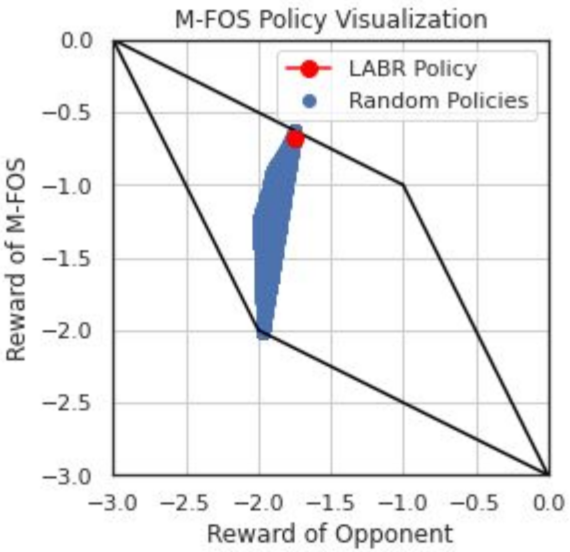


$\phi_T \sim f^i_{RL}(\tau_T, \phi_{T-1})$
$\theta^i_T \sim f^i(\tau_T, \theta^i_{T-1} | \phi)$
$\theta^{-i}_T \sim f^{-i}(\tau_T, \theta^{-i}_{T-1})$
$a^i_t \sim \pi^i(s_t|\theta)$
$a^{-i}_t \sim \pi^{-i}(s_t)$
r^i_{t+1}, s^i_{t+1}
$r^{-i}_{t+1}, s^{-i}_{t+1}$
τ_{T+1}
τ_{T+1}
τ_{T+1}

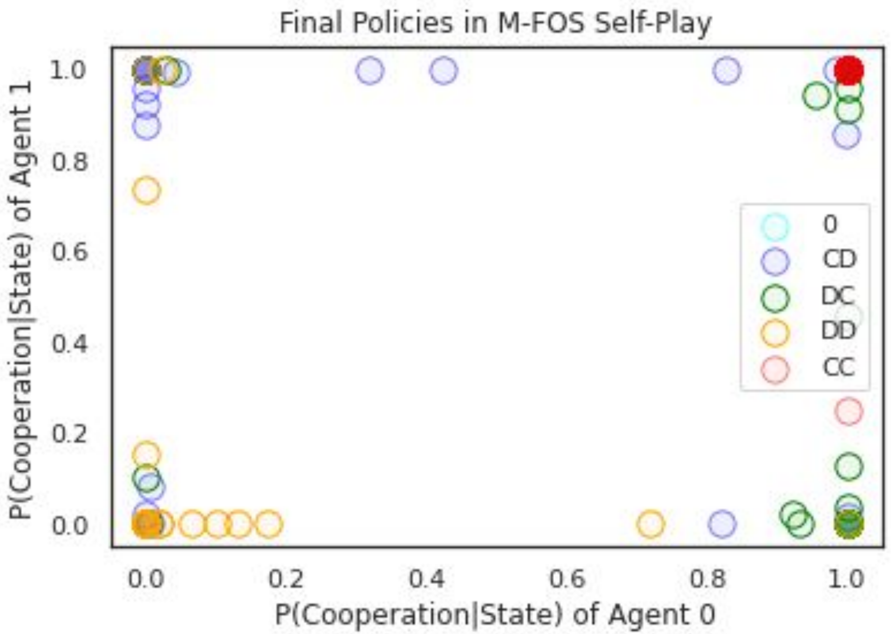
Results in IPD

	M-FOS	NL	LOLA	M-MAML
M-FOS	**-1.01**	**-0.51**	**-0.73**	**-0.67**
NL	-2.14	-1.98	-1.52	-1.28
LOLA	-2.09	-1.30	-1.09	-1.04
M-MAML	-1.86	-1.25	-1.15	-1.17

Results in IPD



M-FOS Self-Play



Results in IMP

	M-FOS	NL	LOLA	M-MAML
M-FOS	**0.0**	**0.20**	**0.19**	**0.22**
NL	-0.20	0.0	-0.02	-0.01
LOLA	-0.19	0.02	0.0	0.02
M-MAML	-0.22	0.01	-0.02	0.0

Scaling up M-FOS

- Inputting and outputting entire policies doesn’t scale!

Scaling up M-FOS

- Inputting and outputting entire policies doesn't scale!
- Solution:
    - The Meta-Agent takes as input *trajectories*, and outputs a *conditioning vector*

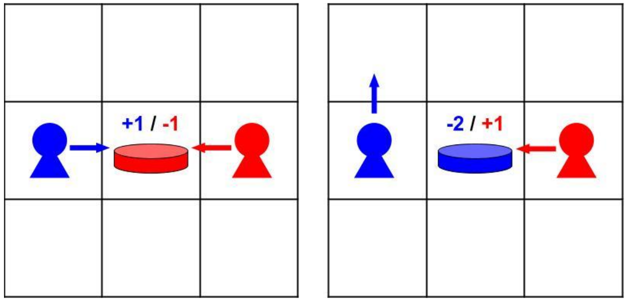
Scaling up M-FOS

- Inputting and outputting entire policies doesn't scale!
- Solution:
  - The Meta-Agent takes as input *trajectories*, and outputs a *conditioning vector*
  - The inner agent then uses this *conditioning vector* to influence its policy within the episode

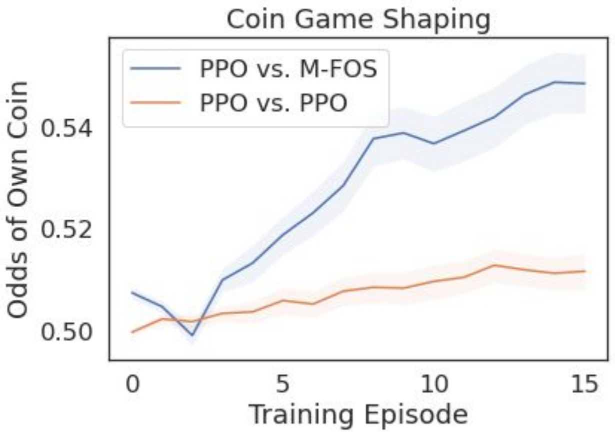
Scaling up M-FOS

- Inputting and outputting entire policies doesn’t scale!
- Solution:
  - The Meta-Agent takes as input *trajectories*, and outputs a *conditioning vector*
  - The inner agent then uses this *conditioning vector* to influence its policy within the episode
  - This is related to Hierarchical Reinforcement Learning

Scaling up M-FOS



- Two players: Red and Blue
- 3x3 Grid
- Coin has color, randomly placed on grid
- Picking up coin -> +1 reward
- IF coin opposite color, then -2 reward for opponent
- Greedy policy: Expected Reward of 0
- MFOS positively influences PPO



	M-FOS	PPO
M-FOS	**20.56**	**44.26**
PPO	-24.62	4.25

Future Work

- Can M-FOS learn to influence other learning agents over a cheap talk channel, without impacting the underlying environment dynamics?
- Can M-FOS learn to generalize against different opponents and different environments?

Thanks for Listening!