Soongsil University

NEURAL INFORMATION PROCESSING SYSTEMS

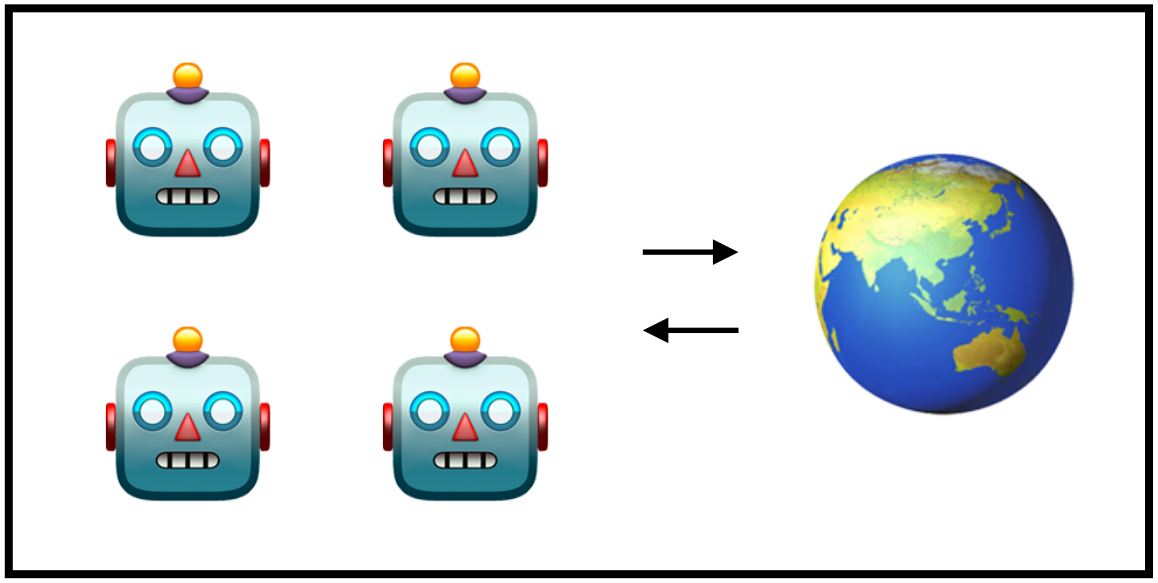
Episodic Future Thinking Mechanism for Multi-agent Reinforcement Learning

Dongsu Lee and Minhae Kwon

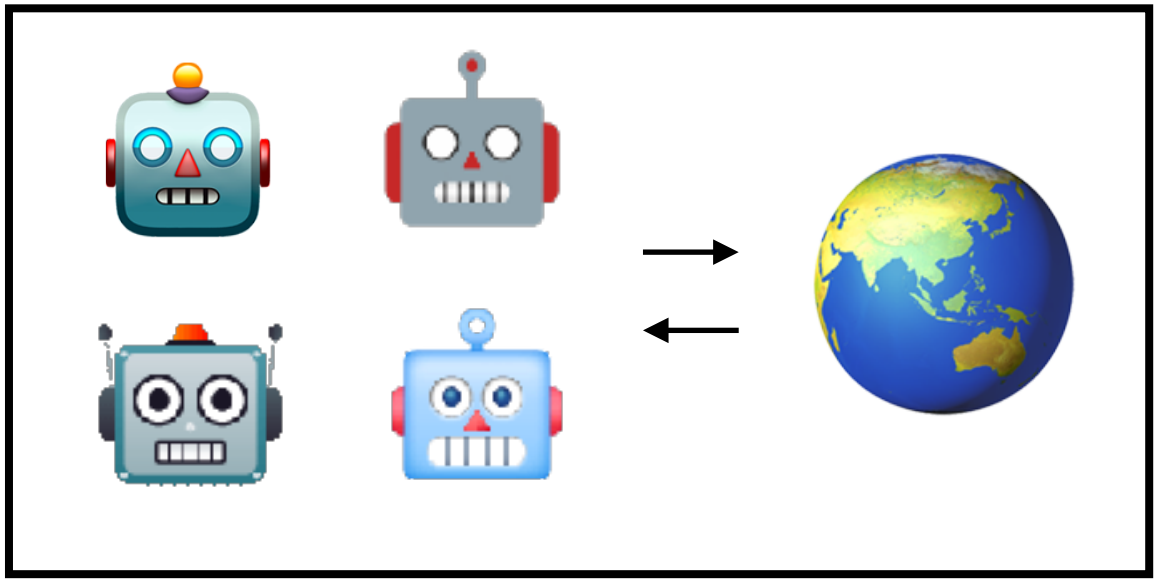
Brain and Machine Intelligence Lab. (BMIL)
Soongsil University
Seoul, Republic of Korea

Collaboration *without* prior coordination

Training



Execution



In real-world settings, such interactions often involve **heterogeneous agents**, which has different behavioral pattern

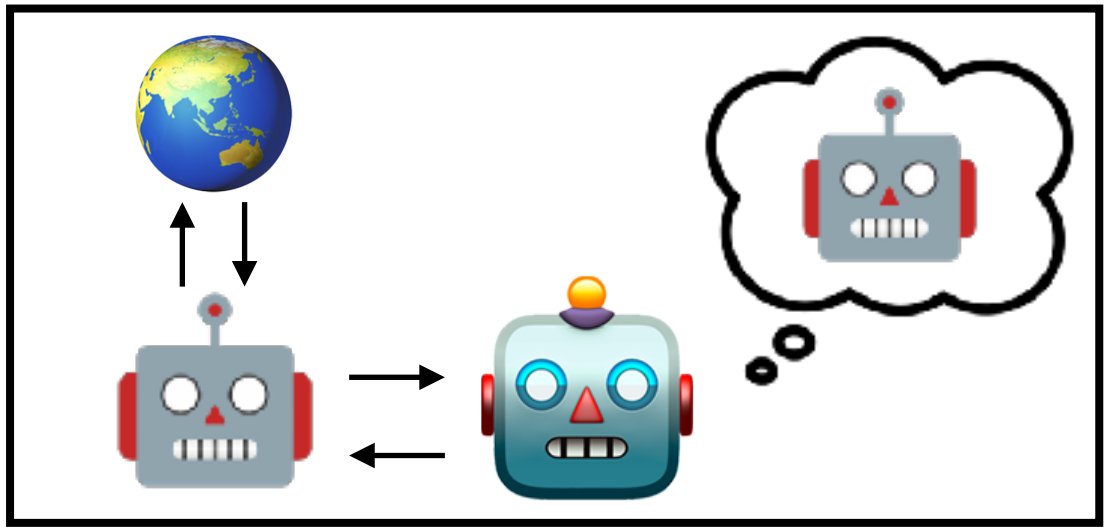
→ limited initial knowledge about other agents

→ partially observability and restricted communication

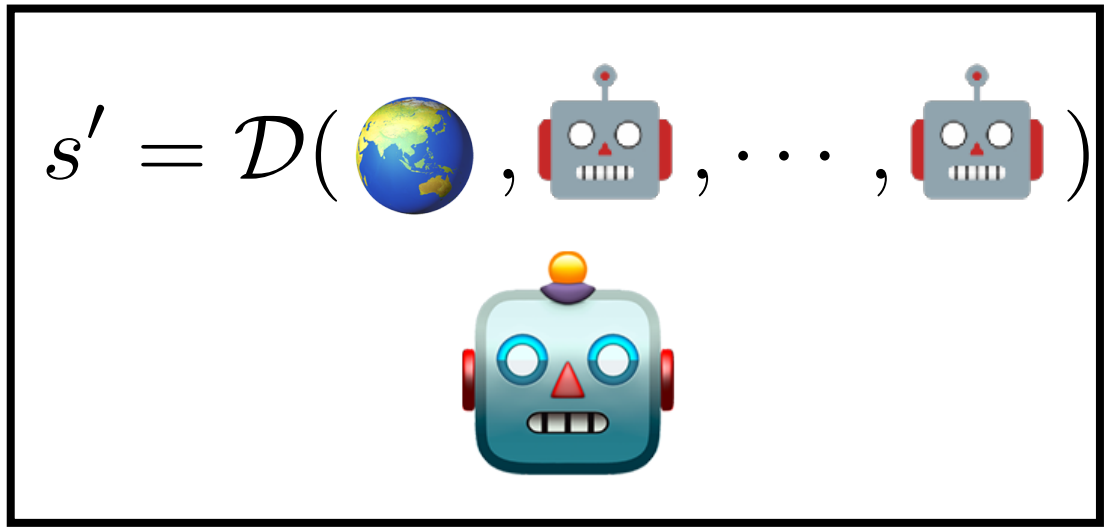
Adaptive decision-making in multi-agent system

How does an agent make an adaptive decision
in a multi-agent system with heterogeneity?

Phase 1: Understanding agent



Phase 2: Future thinking



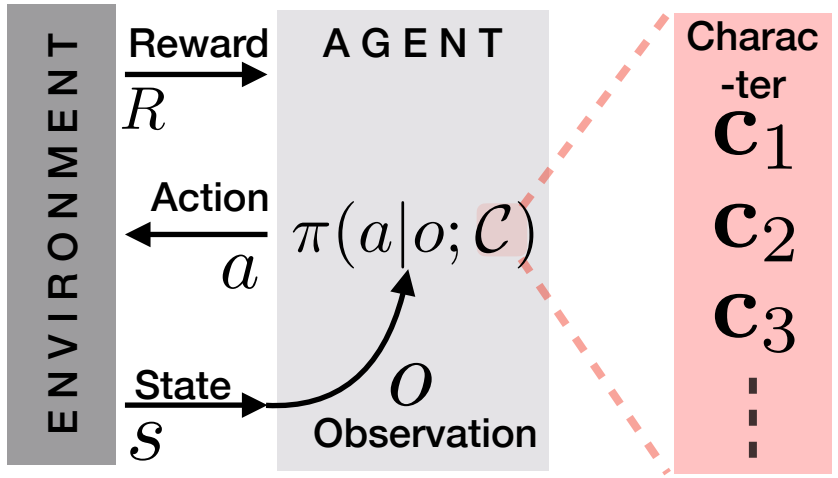
Goal: Understand other agents, then predict future

Agent modeling in reinforcement learning

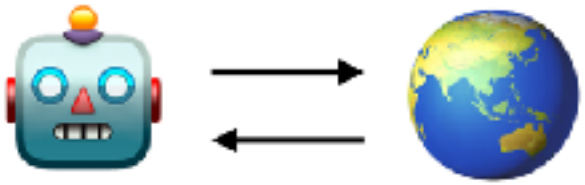
How do we parameterize the optimal agent efficiently?

→ Parameterize the reward function in forward process

$$R(s_t; a_t; \mathbf{c}) = \sum_{n=1}^{N} c_n \mathcal{R}_n = c_1\mathcal{R}_1 + c_2\mathcal{R}_2 + \cdots + c_N\mathcal{R}_N$$

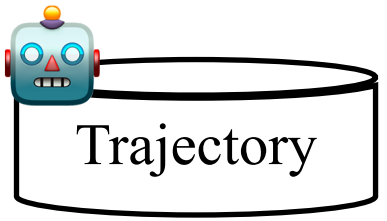


Forward modeling



$\pi(\cdot|s;\mathbf{c})$

Inverse modeling



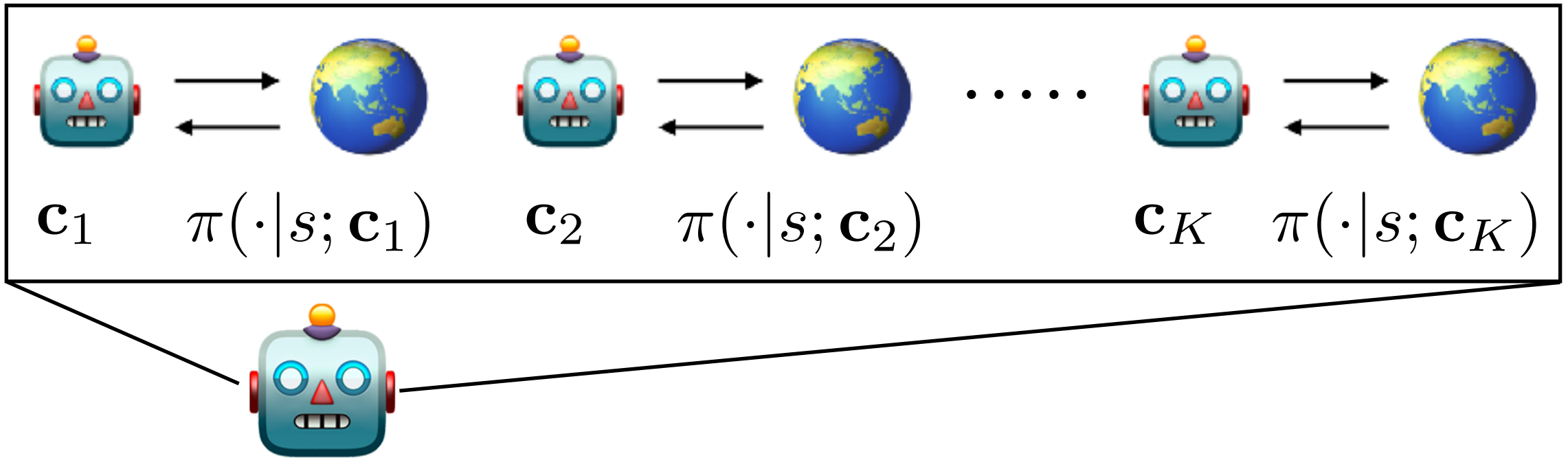
→ $\mathbf{c}$

Advantages of reward parameterization

- Low-sized vector
- Explainability

Multi-character agent for inverse modeling

Parameterize the optimal agent with the character parameters



$\Pi_{\mathcal{C}}(\cdot|s;\mathbf{c}_k)$

Multi-character agent

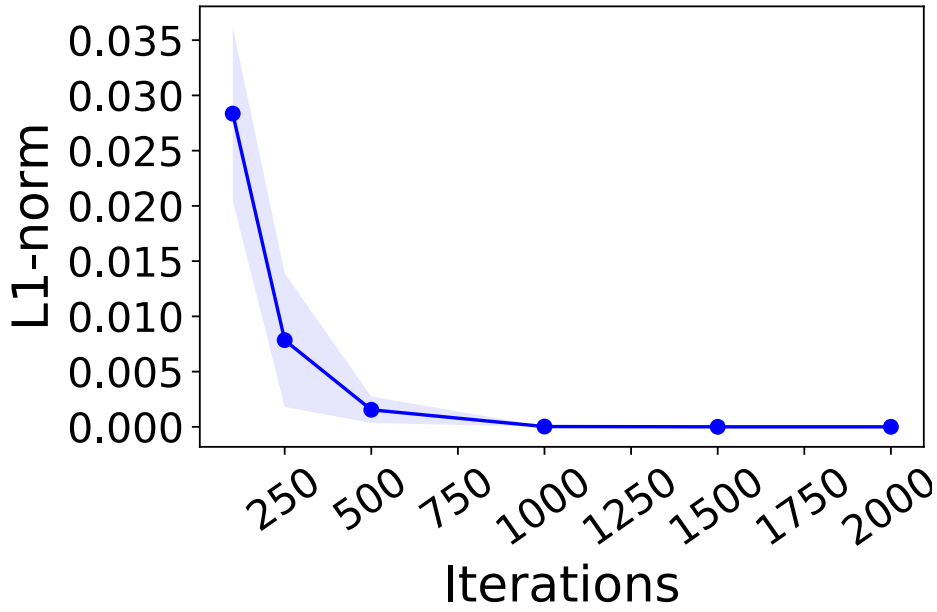
$\mathbf{c}_k \in \mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_K\}$

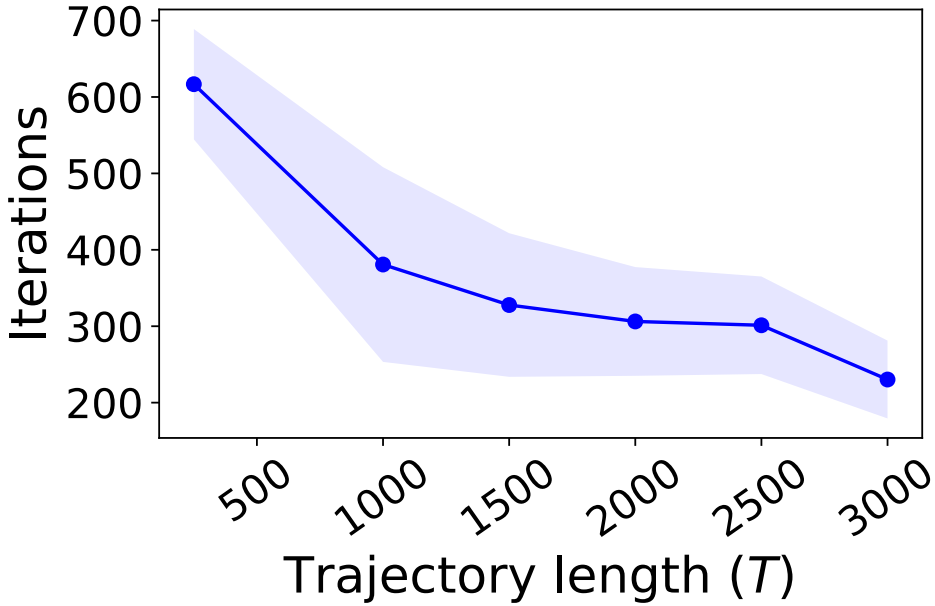
Multi-character agent for inverse modeling

Goal: find a character that explains the observation-action pair of a target agent best

$$\mathbf{c}^* = \arg\max_{\mathbf{c}} \ln \Pi(o, a_{i,1:T} | \mathbf{c})$$

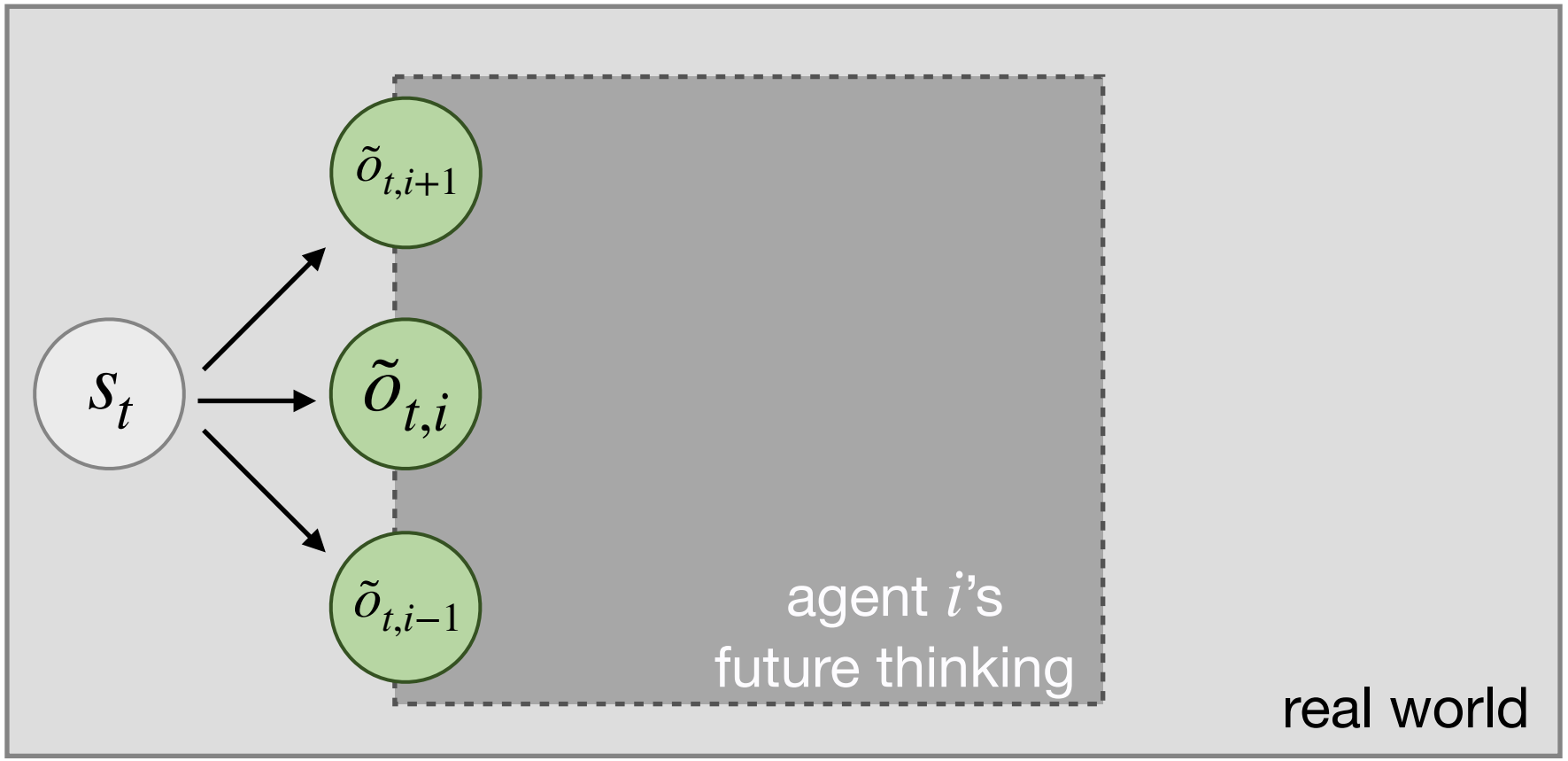
Method: Maximum likelihood estimation with gradient ascent [Kwon2020]



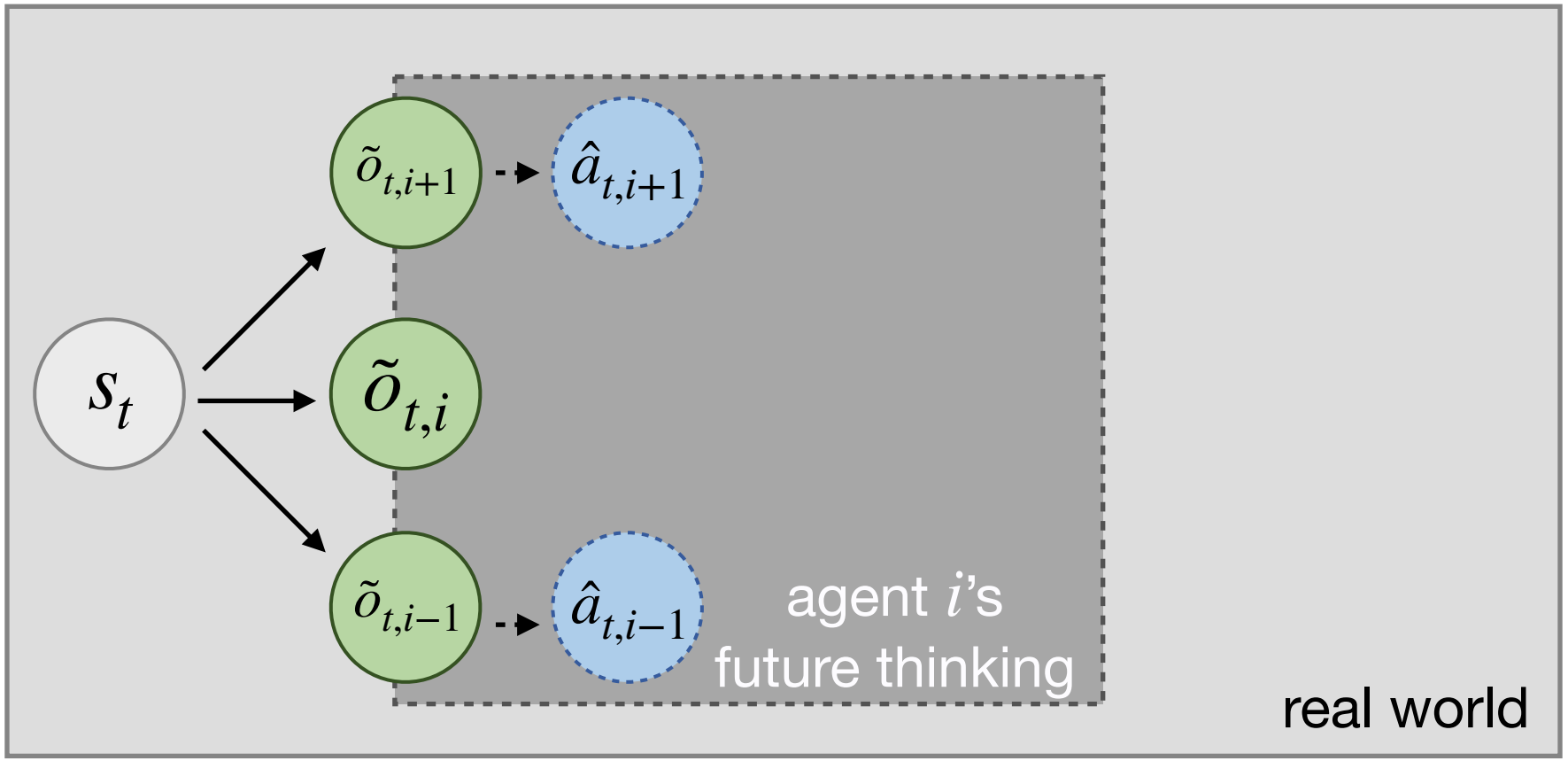


[Kwon2020] M. Kwon et al., "Inverse Rational Control with Partially Observable Continuous Nonlinear Dynamics," NeurIPS2020

Decision-making with episodic future thinking



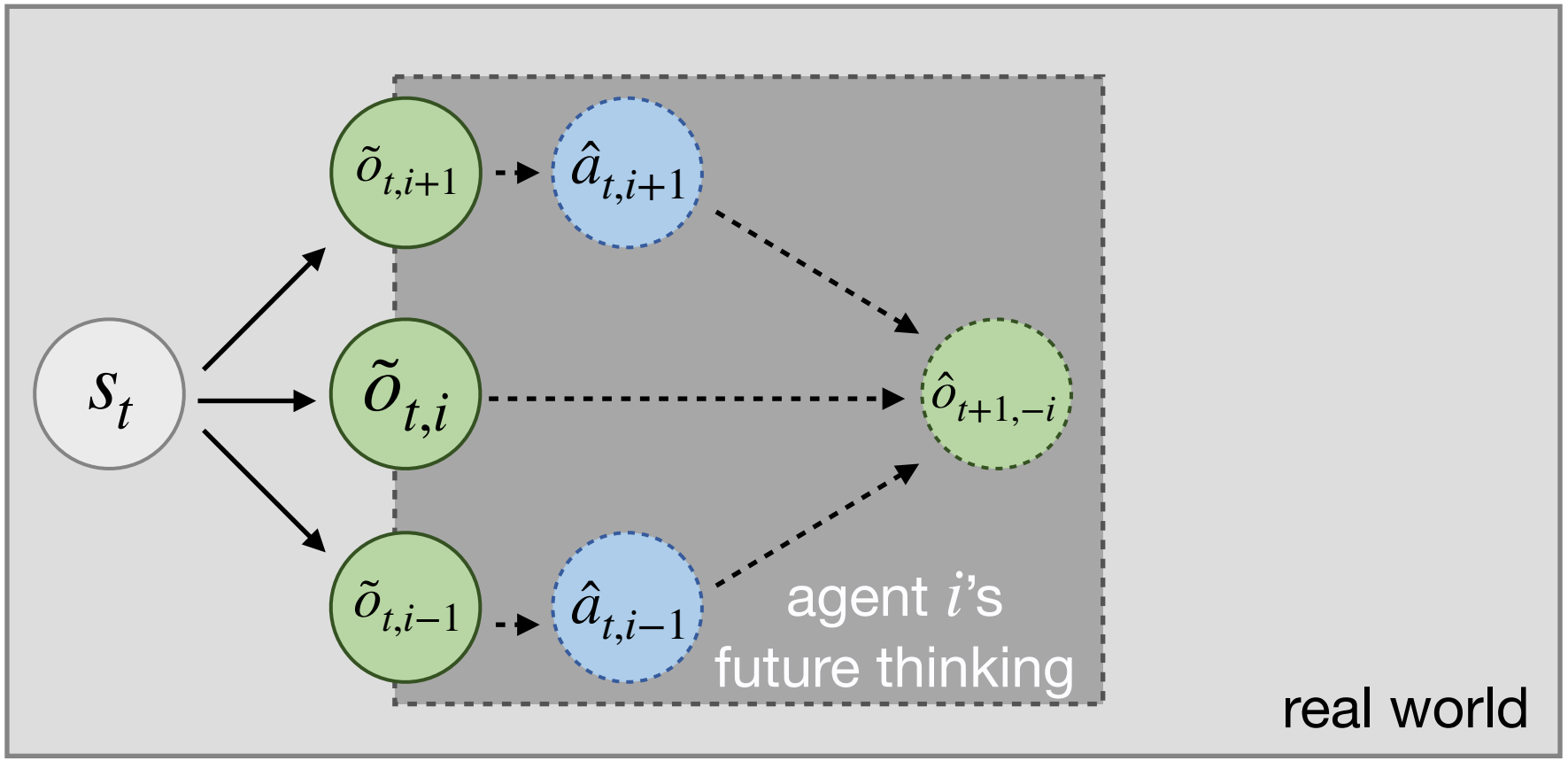
Decision-making with episodic future thinking



$\Pi_C(\cdot|s;\mathbf{c}_k)$

Predict neighbor agents' actions based on the inferred characters

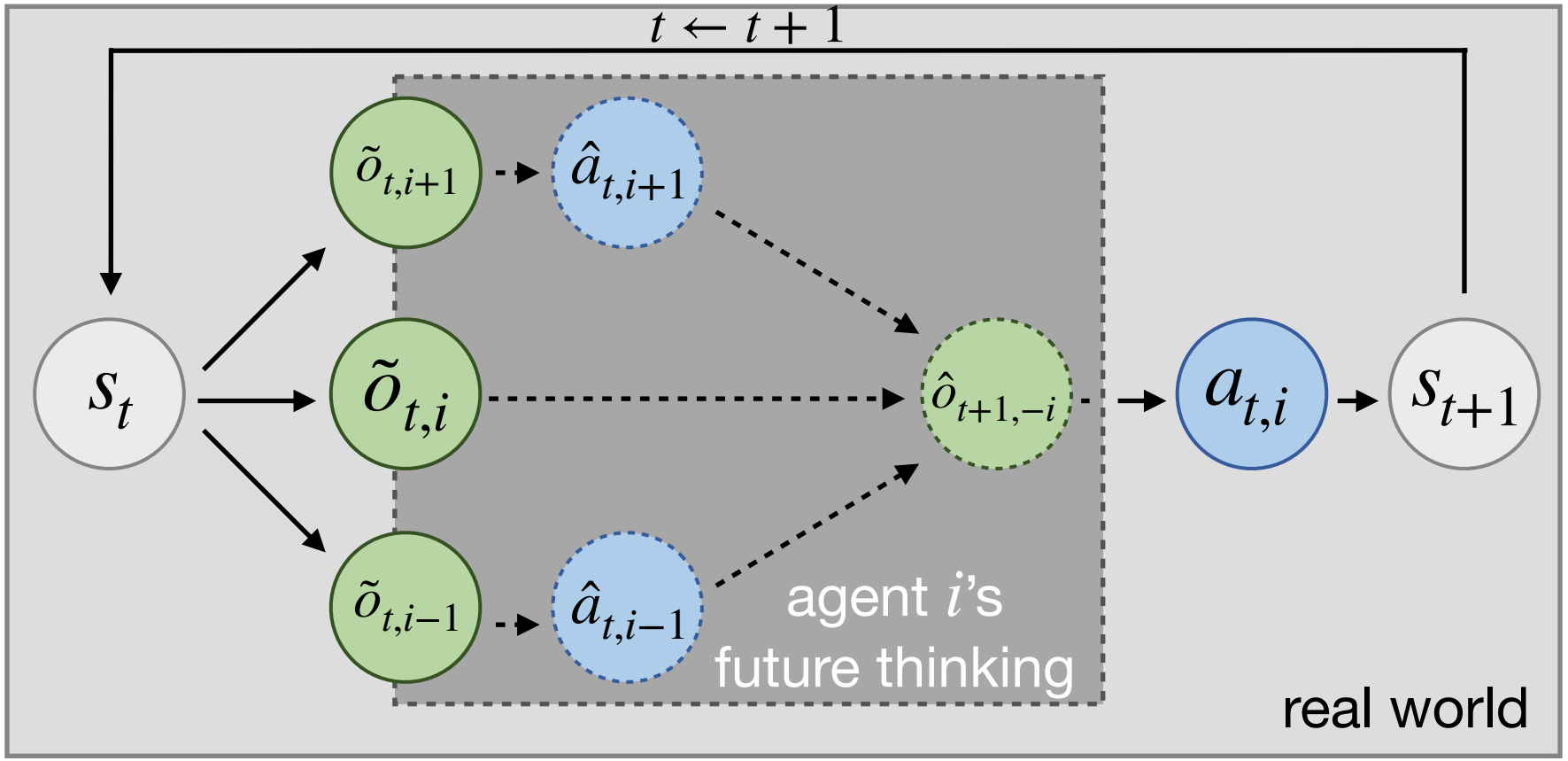
Decision-making with episodic future thinking

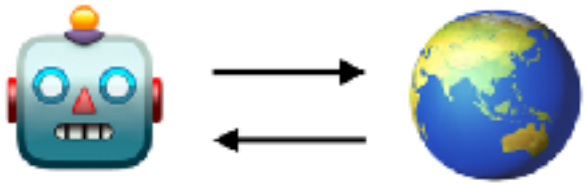


$\mathcal{D}(o_{t,i}, \hat{\mathbf{a}}_{t,-i}, a_{t,i} = \emptyset)$

Predict the next observation based on its current observation and predicted neighbor's actions

Decision-making with episodic future thinking



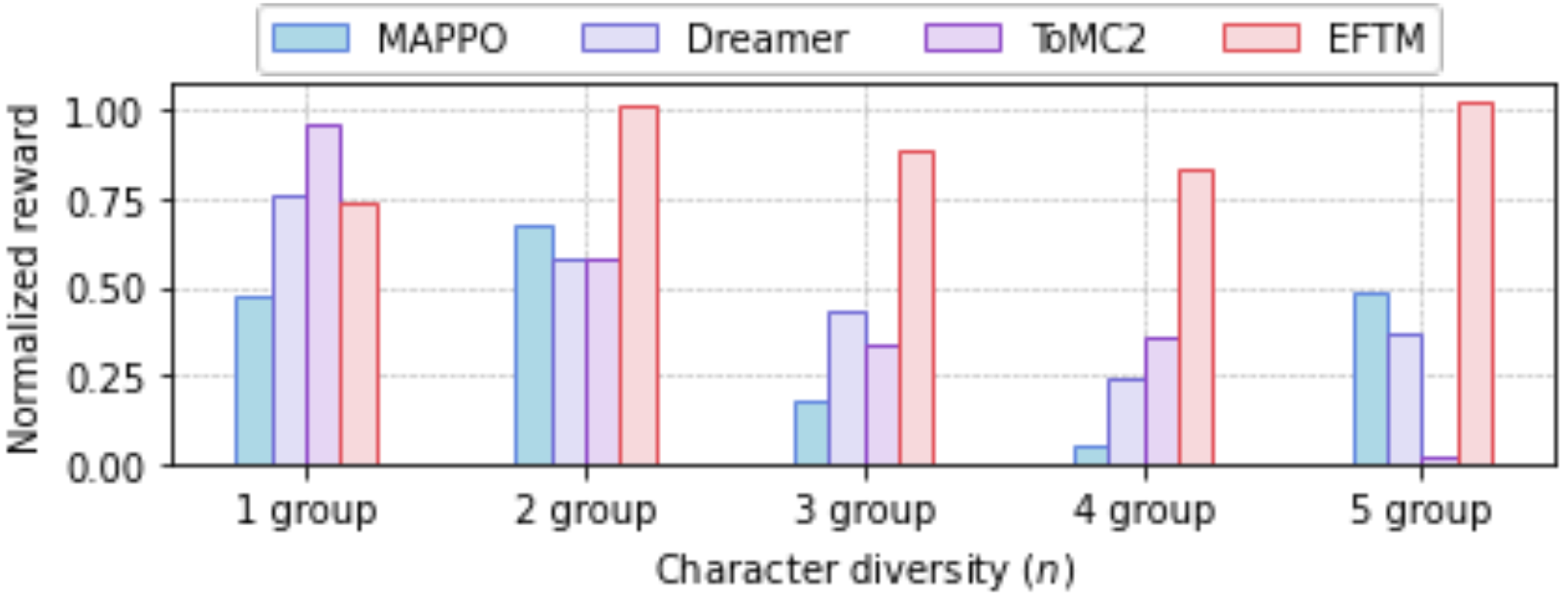


Choose an action at a given predicted observation, which updates environment

Conclusions

EFTM optimizes multi-character policy through reward parameterization
→ Model a character of other agents from their behavioral pattern
→ Predict upcoming future to make an adaptive decision

This strategy improves MARL robustness as character diversity increases!



Project website