How To Make Social Decisions in a Heterogeneous Society?

Dongsu Lee [1] **Minhae Kwon** [1]

Abstract

Understanding cognitive processes in multi-agent interactions is a primary goal in cognitive science. It can guide the direction of artificial intelligence (AI) research toward social decision-making in heterogeneous multi-agent systems. In this paper, we introduce *episodic future thinking (EFT) mechanism* of a reinforcement learning (RL) agent by benchmarking the cognitive process of animals. To achieve future thinking functionality, we first train a *multi-character policy* that reflects heterogeneous characters with an ensemble of heterogeneous policies. An agent's *character* is defined as a different weight combination on reward components, thus explaining the agent's behavioral preference. The future thinking agent collects observation-action trajectories of the target agents and uses the pre-trained multi-character policy to infer their characters. Once the character is inferred, the agent predicts the upcoming actions of the targets and simulates the future. This capability allows the agent to adaptively select the optimal action, considering the upcoming behavior of others in multi-agent scenarios. To evaluate the proposed mechanism, we consider the multi-agent autonomous driving scenario in which autonomous vehicles with different driving traits are on the road. Simulation results demonstrate that the EFT mechanism with accurate character inference leads to a higher reward than existing multi-agent solutions. We also confirm that the effect of reward improvement remains valid across societies with different levels of character diversity.

[1]Soongsil University, Seoul, Korea. Correspondence to: Minhae Kwon <minhae@ssu.ac.kr>.

ICML (International Conference on Machine Learning) Workshop on Counterfactuals in Minds and Machines, Honolulu, Hawaii, USA.

1. Introduction

Understanding how humans make decisions under multi-agent interactions is a significant research topic in cognitive science. It becomes a critical clue to designing the interactions between heterogeneous AI agents in multi-agent systems. Several studies have demonstrated that humans learn from past experiences and also imagine counterfactual or future scenarios to make better decisions (Lynch et al., 1991; Redish & Mizumori, 2015; Jern et al., 2017). The *counterfactual thinking*, i.e., the ability to simulate alternative consequences of the last episode, is widely studied in multi-agent RL (MARL) (Oberst & Sontag, 2019; Foerster et al., 2018; Byrne, 2019). However, the *episodic future thinking* (Schacter et al., 2015), i.e., the ability to anticipate future episodes, is rarely considered in the literature despite its essential for handling multi-agent interactions.

We, as human beings, strive to avoid costly mistakes by contemplating the upcoming situation. To incorporate this functionality into AI, a naive solution would be leveraging a single-step prediction based on model-based RL (Janner et al., 2019; Lai et al., 2020; Mehta et al., 2022; Sutton, 1991; Lin et al., 2022; Xu et al., 2021). However, the model-based RL approaches assume that the state transition model is known or easily learnable. Such an assumption is untenable in the multi-agent system. This limitation originates from the fact that the state transition model relies not only on the agent's state-action pairing but also on other agents' action combinations. The subsequent state could considerably vary depending on the action combinations of others, even for a given state-action pair of the agent. As the number of agents increases, the number of possible action combinations of all agents grows exponentially, making it infeasible to learn the state transition model in a sizeable multi-agent system. Therefore, it would be more appropriate to use model-free RL in multi-agent systems such that the agent learns the policy without explicit knowledge of the state transition model.

We aim to build the EFT mechanism for model-free RL agents to make optimal decisions in a heterogeneous society where agents exhibit diverse characteristics. We formalize this task as a Multi-agent Partially Observable Markov Decision Process (MA-POMDP), a framework to address the RL problem that multiple agents behave under partial

observation (Oliehoek, 2012; Sutton & Barto, 2018). In this study, we define the character by reflecting the behavioral preferences of the RL agent. In a driving scenario, for example, some drivers pursue safety as a top priority, but others prioritize their speed. These behavioral preferences can be parameterized by assigning weights to reward components (e.g., high weight on the safety component and low weight on the speed component of the reward function versus high weight on the speed and low weight on the safety). Therefore, each character has a different reward function, leading to heterogeneous policies and behavior patterns in society.

Creating the EFT-based decision-making agent necessitates two functional modules: a multi-character policy and a character inference module. The initial step is constructing a multi-character policy embedding behavioral patterns corresponding to the characters. To facilitate practicality, we permit the agent to observe continuous state information partially. We also train the policy to handle a hybrid action space consisting of discrete and continuous, which cannot be tackled with a naive deployment of existing RL algorithms. The character inference module is built leveraging the core idea of the inverse rational control (IRC) (Kwon et al., 2020). This module infers the target agent's characters by maximizing the log-likelihood of the target's observation-action trajectories. This work extends the accessibility of the IRC from the continuous to the hybrid action space. The agent is finally equipped with the EFT functionality that can predict the upcoming actions of the surrounding agents, combining two modules.

To perform the EFT mechanism, the agent first plays the role of an observer, i.e., it collects observation-action trajectories of target agents. Using the character inference module and the collected trajectories, the agent infers the target agents' characters. Subsequently, the agent predicts the actions of the others leveraging the inferred characters and multi-character policy, then simulating its future observation. In this mental simulation, the agent's action is fixed as 'no action.' Only target agents take the predicted actions. This allows the agent to estimate the observation at the time point when all target agents have taken actions, but the agent still needs to (i.e., has yet to). Finally, the agent selects the best action corresponding to the estimated future observation. To sum up, the EFT mechanism allows the agent to behave proactively under heterogeneous multi-agent interactions.

Summary of contributions:

- We introduce character diversity in a multi-agent system by parameterizing the reward function. We propose to build the multi-character policy and allow the agent to be equipped with it to infer the character of the target agent (Section 3).
- We propose the EFT mechanism as a model-free prediction approach in that the agent with the multi-character policy predicts the future actions of others, simulates the corresponding future observation, and performs foresighted action selection. This mechanism enables the agent to consider the multi-agent interactions in the decision-making process (Section 4).
- We verify the proposed mechanism by increasing character diversity in society. Extensive experiments confirm that the proposed mechanism enhances group rewards no matter how high a character diversity level exists in society. (Section 5).

2. Related Works

Episodic Future Thinking. Cognitive neuroscience aims to understand how humans use memory in decision-making. Interestingly, the trend of the brain's regional neural activation regarding counterfactual reasoning (i.e., simulating alternative consequences of the last episode) and future thinking (i.e., simulating episodes that may occur in the future) is similar (Schacter et al., 2015). In (Thorstad & Wolff, 2018), the authors study the relationship between future thinking and decision-making and confirm that humans perform future-oriented decision-making. The decision-making abilities, such as strategy formulation, are also significant in scenarios that require multi-agent interactions, e.g., social decision-making. Several prior studies have proposed methods to anticipate the action of other agents and the next state (Yasdi, 1999; Pan et al., 2013; Yang et al., 2021). In (Yasdi, 1999) and (Pan et al., 2013), the authors forecast the next state, not the behavior of each agent, from a macroscopic standpoint. In (Yang et al., 2021), the authors predict the behavior of an agent through a deep Bayesian network considering the dynamics and the previous surrounding environment information. Even though these studies can infer future information, no strategy formulation is suggested. Similarly, interactive POMDP-based research (Han & Gmytrasiewicz, 2019; Doshi et al., 2020; Schwartz et al., 2022) predict the action of other agents and make a decision adaptively. Still, it does not contemplate the future state or is addressed by model-based RL and dynamic programming, which requires a known dynamic model. All these approaches naively establish the strategy without consideration of the surrounding agents' upcoming actions and the next state. In this study, we propose the ETF mechanism can predict future observations based on the current state and predicted actions of surrounding agents. Consequently, the agent equipped with this mechanism can select a foresighted action corresponding to the anticipated future observation.

False Consensus Effect. The psychologist found that humans have a cognitive bias to assume their character, belief, and actions are relatively widespread throughout the general population (Folli & Wolff, 2022; Engelmann & Strobel,

2000; 2012). This is referred to as the *False Consensus Effect (FCE)* (Dawes, 1989; Marks & Miller, 1987; Ross et al., 1977). Recent research suggests that AI may adopt false beliefs (Rabinowitz et al., 2018). To highlight the importance of character inference in heterogeneous scenarios, in this paper, we compare the performance of the EFT mechanism with two types of agents; one is the proposed agent which is equipped with the character inference module. The other one is the FCE-based agent which assumes that target agents have the same character as the agent.

Machine Theory of Mind. Human beings make decisions in the social context by considering multiple perspectives of others, including emotions and personalities. This capacity is known as the Theory of Mind (ToM) in cognitive science (Premack & Woodruff, 1978; Baron-Cohen, 1997; Langley et al., 2022). The ToM is primarily related to deducing internal models of others and secondarily predicting the future action of others. Research to provide AI with this capability, which can impact its stability and performance, gets the spotlight, e.g., machine ToM (Rabinowitz et al., 2018), inverse learning (Ng et al., 2000; Ratliff et al., 2006; Dvijotham & Todorov, 2010), and Bayesian ToM (Lucas et al., 2014). All these approaches aim to reconstruct the target agent's belief, reward function, or dynamic model by leveraging its trajectories. To elaborate, the machine ToM is a meta-learning strategy for learning the reasoning method explicitly and can be used for prediction and AI-human collaboration. However, these methods have a general purpose problem regarding the target setting and application space. The IRC has mitigated the former, but the regulation of action space still needs to be resolved. In this work, we adopt the IRC and extend the action space from the continuous to the hybrid.

Model-based Reinforcement Learning. The model-based RL uses the system dynamic model, and model-free RL does not explicitly consider it. Model-based RL can be again classified into two approaches. One approach is to assume that the agent knows the dynamic model, and the other is for the agent to learn the dynamic model in the training process (Janner et al., 2019; Lai et al., 2020; Moerland et al., 2023). In model-based RL, the agent can predict future states based on an understanding of dynamic models and use it for single and multiple-step prediction (Sutton, 1991; Lin et al., 2022; Xu et al., 2021). These tasks can work on the single agent scenario, assuming the agent can observe complete state information. However, if the agent can only observe partial noisy information on the state, it is challenging to ascertain the dynamic model. Additionally, building the dynamic model in a multi-agent scenario is extremely complicated because the state transition depends not only on the state-action pair of the agent but also on the action combinations of others. The prediction-based MARL studies (Marinescu et al., 2017), a representative example of model-free prediction, do not consider the behavior of other agents but only predict sequential changes of the state in a non-stationary environment. To overcome the limitations, this work introduces the single-step prediction for model-free RL.

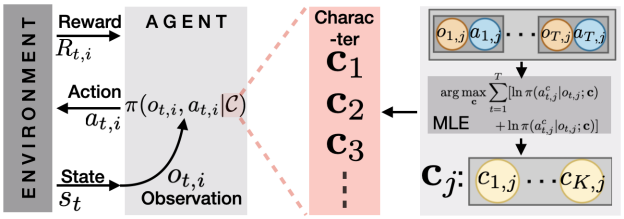

Figure 1. A block diagram of an agent i with a multi-character policy. The agent can infer the character of others by using the maximum likelihood estimation.

3. Character Inference Using Multi-character Policy

We aim to build an agent to make optimal decisions under multi-agent interactions. It requires the agent to be able to anticipate the near future by predicting other agents' actions. The agent should possess the ability to infer the others' characters, leveraging observation of their behaviors. Accurate character inference is a prerequisite for the EFT mechanism since the character is a crucial clue to predicting future action. Therefore, this section proposes two functional modules for character inference: a multi-character policy and character inference. An illustrative explanation of these functionalities is presented in Figure 1.

3.1. Problem formulations for multi-agent decision-making

We consider multi-agent scenarios where RL agents adaptively behave to each other. All agents have to make decisions and execute actions simultaneously, unlike the extensive-form game (Roth & Erev, 1995; Owen, 2013) in which the agents alternate executing the actions.

The multi-agent decision-making problem can be formalized as a MA-POMDP $M = \langle E, \mathcal{S}, \{\mathcal{O}_i\}, \{\mathcal{A}_i\}, \mathcal{T}, \{\Omega_i\}, \{R_i\}, \gamma \rangle_{i \in E}$ that includes an index set of agents $E = \{1, 2, \cdots, N\}$, continuous states $s_t \in \mathcal{S}$, continuous observations $o_{t,i} \in \mathcal{O}_i$, hybrid actions $a_{t,i} = \{a^c_{t,i}, a^d_{t,i}\} \in \mathcal{A}_i$, where continuous action $a^c_{t,i} \in \mathcal{A}^c_i$ and a discrete action $a^d_{t,i} \in \mathcal{A}^d_i = \{w : |w| \leq W, w \in \mathbb{Z}, W \in \mathbb{N}\}$, where the size of discrete action space is $|\mathcal{A}^d_i| = 2W + 1$, $\mathbb{Z}$ denotes the set of integers, and $\mathbb{N}$ denotes the set of natural numbers. Let $\mathcal{A} := \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_N$.

Subsequently, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the state transition probability; $\Omega_i : \mathcal{S} \to \mathcal{O}_i$ is the observation probability; $R_i : \mathcal{S} \times \mathcal{A}_i \times \mathcal{S} \to \mathbb{R}$ denotes the reward function that evaluates the agent's action $a_{t,i}$ for a given state s_t and the outcome state s_{t+1}; $\gamma \in [0, 1)$ is the temporal discount factor.

An unordered set of the actions of all agents at time t is denoted as

$$\begin{aligned}\mathbf{a}_t &= \langle a_{t,i} \rangle_{i \in E} \\ &= \langle a_{t,1}, \cdots, a_{t,i}, \cdots, a_{t,N} \rangle = \langle a_{t,i}, \mathbf{a}_{t,-i} \rangle\end{aligned} \tag{1}$$

where subscript $-i$ represents the indices of all agents in E except i. Thus, $\mathbf{a}_{t,-i} = \langle a_{t,1}, \cdots, a_{t,i-1}, a_{t,i+1} \cdots, a_{t,N} \rangle$ represents a set of all agents' actions at time t without $a_{t,i}$. The state transition probability denotes $\mathcal{T}(s_{t+1}|s_t, \mathbf{a}_t)$. Note that state transition is based on the action combination of all agents $\mathbf{a}_t$, and not on the action of a single agent $a_{t,i}$.

Next, $\mathbf{c}_i = \{c_{i,1}, c_{i,2}, \cdots, c_{i,K}\} \in \mathcal{C} \in \mathbb{R}^K$ denotes a K-dimensional character vector for the agent i. Character $\mathbf{c}_i$ can parameterize the reward function of the agent i, i.e., $R_{t,i} = R_i(s_t, a_{t,i}, s_{t+1}; \mathbf{c}_i)$. The agent aims to learn the optimal policy that returns the optimal action $a^*_{t,i} \sim \pi^*(\cdot|o_{t,i}; \mathbf{c}_i)$ given observation and character. Specifically, the objective of the agent aims to maximize the expected discounted cumulative reward $\mathcal{J}(\pi) = \mathbb{E}_\pi\Big[\sum_t \gamma^t R_i(s_t, a_{t,i}, s_{t+1}; \mathbf{c}_i)\Big]$ by building the best policy π. This defines the state-action value function $Q^\pi(s, a; \mathbf{c}_i) = \mathbb{E}_\pi\Big[\sum_t \gamma^t R_i(s_t, a_{t,i}, s_{t+1}; \mathbf{c}_i) | s_0 = s, a_0 = a\Big]$. In the next section, we discuss the details of the multi-character policy in terms of neural network design and its training.

3.2. Training a multi-character policy

The multi-character policy includes inputs in continuous space (e.g., observation $o_{t,i}$ and character $\mathbf{c}_i$) and outputs in hybrid space (e.g., action $a_{t,i}$). To build the policy generalized over continuous space, actor-critic architecture is used. It approximates the policy $\pi_\phi(\cdot|o_{t,i}; \mathbf{c}_i)$ and Q-function $Q_\theta(o_{t,i}, a_{t,i}; \mathbf{c}_i)$, where ϕ denotes parameters of the actor network and θ denotes the parameters of the critic network.

The loss functions used to train the actor and critic networks are $\mathcal{L}(\phi) = -\sum Q_\theta(o_{t,i}, \pi_\phi(\cdot|o_{t,i}; \mathbf{c}_i))$, and $\mathcal{L}(\theta) = \sum(y - Q_\theta(o_{t,i}, \pi_\phi(\cdot|o_{t,i}; \mathbf{c}_i)))^2$, respectively. Herein, $y = R_{t,i} + Q_{\theta'}(o_{t+1,i}, \pi_{\phi'}(\cdot|o_{t+1,i}; \mathbf{c}_i))$ represents the Temporal Difference (TD) target, where θ' and ϕ' denote the target networks.

Next, we propose a post-processor $g(\cdot)$ to handle hybrid action space. Let a proto-action $\bar{a}^d_{t,i}$ be the output of the actor-network. The post-processor $g(\cdot)$ performs quantization process by discretizing the continuous proto-action $\bar{a}^d_{t,i}$ into discrete post-action $a^d_{t,i}$, i.e.,

$$\begin{aligned}a^d_{t,i} &= g(\bar{a}^d_{t,i}, W) \\ &= \min\Big(\Big\lfloor \frac{2W+1}{2W}\Big(\bar{a}^d_{t,i} + \frac{W}{2W+1}\Big)\Big\rfloor, W\Big),\end{aligned} \tag{2}$$

where $\lfloor \cdot \rfloor$ denotes a floor function. The derivation of (2) is presented in Appendix D.

We summarize the multi-character policy training process in Algorithm 1. In the next subsection, we introduce the character inference module that infers the characters of other agents.

Algorithm 1 Multi-character policy training

Initialization: Actor-network ϕ, critic network θ
Require: Total episode K, total time steps per episode T, discrete action space parameter W, agent i
for episode $k = 1, K$ **do**
  Reset s_1 and get $o_{1,i} \sim \Omega_i(\cdot|s_1)$
  Sample character $\mathbf{c}_i \sim \mathcal{C}$
  for timestep $t = 1, T$ **do**
    Get proto-action $\{a^c_{t,i}, \bar{a}^d_{t,i}\} \sim \pi_\phi(\cdot|o_{t,i}; \mathbf{c}_i)$
    Get post-action
    $a^d_{t,i} \leftarrow g(\bar{a}^d_{t,i}, W)$
    Execute $a_{t,i} = \{a^c_{t,i}, a^d_{t,i}\}$, Update s_{t+1}
    Receive $R_{t,i}$, Get $o_{t+1,i} \sim \Omega_i(\cdot|s_{t+1})$
    Calculate $\mathcal{L}(\phi), \mathcal{L}(\theta)$, Update ϕ, θ
  end for
end for
return ϕ, θ

Algorithm 2 Character inference module

Require: Trained actor network ϕ, length of trajectories T, observation-action trajectories $o_{1:T,j}, \{a^c_{1:T,j}, a^d_{1:T,j}\}$, and initial $\mathbf{c} \sim \mathcal{C}$, target agent j
repeat
  Reset $\mathcal{U}(\mathbf{c}) = 0$
  for $t = 1, T$ **do**
    $\mathcal{U}(\mathbf{c}) \leftarrow \mathcal{U}(\mathbf{c}) + \ln \pi(a^c_{t,j}|o_{t,j}; \mathbf{c}) + \ln \pi(a^d_{t,j}|o_{t,j}; \mathbf{c})$
  end for
  Update $\mathbf{c} \leftarrow \mathbf{c} + \alpha \nabla_{\mathbf{c}} \mathcal{U}(\mathbf{c})$
until $\mathbf{c}$ converges
return $\hat{\mathbf{c}}_j \leftarrow \mathbf{c}$

3.3. Inferring character of target agent

After completing the training on the multi-character policy, our next objective is to infer the character $\mathbf{c}_j$ of the target agent $j \in E$. The agent first collects observation-action trajectories of the target for character inference. Subsequently, it utilizes the multi-character policy to identify the charac-

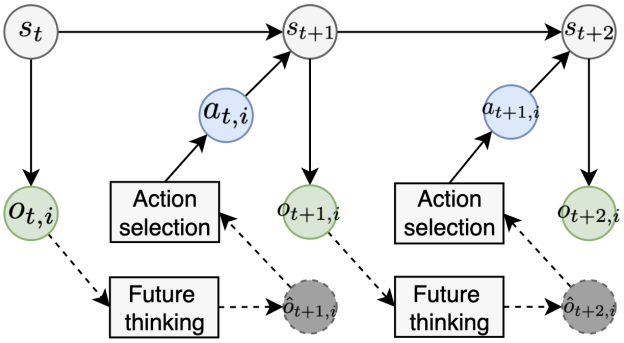

Figure 2. Diagram of POMDP with EFT mechanism. The future thinking and action selection modules are included to obtain action from the observation. The solid line and circles represent the actual event in the real world. The dashed line and circles depict the virtual event in the simulated world of the agent i.

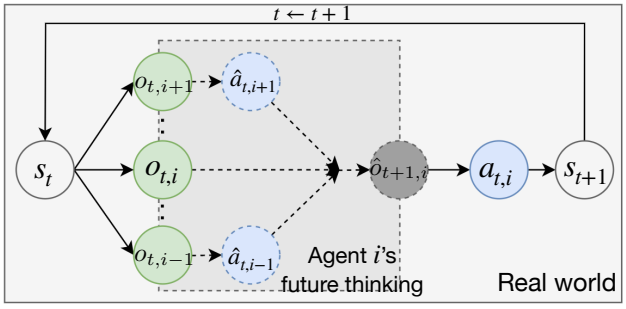

Figure 3. Illustrative example of the EFT mechanism of the agent i. The gray area with a dashed line indicates the mental simulation of the agent about the future step. The brighter area with a solid line indicates the real-world transitions from the perspective of the agent i.

ter $\mathbf{c}_j$ that best explains the collected data. To elaborate, $\mathbf{c}_j$ can be estimated by maximizing the log-likelihood of observation-action trajectories $\ln P(o_{1:T,j}, a_{1:T,j}|\mathbf{c}_j)$. This can be formulated as follows.

$$\begin{aligned}\hat{\mathbf{c}}_j &= \arg\max_{\mathbf{c}} \ln P(o_{1:T,j}, a_{1:T,j}|\mathbf{c}) \\ &= \arg\max_{\mathbf{c}} \sum_{t=1}^{T} \left[\ln \pi(a^c_{t,j}|o_{t,j};\mathbf{c}) + \ln \pi(a^d_{t,j}|o_{t,j};\mathbf{c})\right]\end{aligned} \tag{3}$$

The derivation of (3) can be found in Appendix E.

To efficiently perform the inference task, we use the gradient ascent method. It runs the iteration by changing $\mathbf{c}$ toward the direction to increase $\mathcal{U}(\mathbf{c}) = \ln \pi(a^c_{t,j}|o_{t,j};\mathbf{c}) + \ln \pi(a^d_{t,j}|o_{t,j};\mathbf{c})$, which is summarized in Algorithm 2.[1]

4. Foresight Action Selection Based on Episodic Future Thinking Mechanism

This section presents the proposed EFT mechanism that enables the agent to simulate the subsequent observations and select a foresighted action. The proposed EFT mechanism comprises a future thinking module and an action selection module as described in Figure 2.

The future thinking module includes two steps: action prediction and the next observation simulation. With these two steps, the agent can foresee the next observation. This process is illustrated in Figure 3. Subsequently, the action selection module enables the agent to decide the current action corresponding to the simulated next observation.

4.1. Future thinking: step I - action prediction

In this step, the agent with the multi-character policy predicts the actions of the neighbor agents by using pre-inferred characters and observations. The agent can predict the action of the target agent j ($\in E, j \neq i$)[2] using the trained multi-character policy π_ϕ and inferred character $\hat{\mathbf{c}}_j$, i.e., $\hat{a}_{t,j} \sim \pi_\phi(\cdot|o_{t,j};\hat{\mathbf{c}}_j)$. Therefore, the predicted action set of others $\hat{\mathbf{a}}_{t,-i}$ is as follows.

$$\begin{aligned}\hat{\mathbf{a}}_{t,-i} = \langle &\pi_\phi(o_{t,1};\hat{\mathbf{c}}_1), \cdots, \pi_\phi(o_{t,i-1};\hat{\mathbf{c}}_{i-1}), \\ &\pi_\phi(o_{t,i+1};\hat{\mathbf{c}}_{i+1}), \cdots, \pi_\phi(o_{t,N};\hat{\mathbf{c}}_N)\rangle\end{aligned} \tag{4}$$

4.2. Future thinking: step II - next observation simulation

In this step, we introduce how the agent simulates its next observation by using the predicted action $\hat{\mathbf{a}}_{t,-i}$. Note that this prediction is the result of the mental simulation of agent i, when $a_{t,i} = \emptyset$ is satisfied. Herein, $\emptyset$ denotes null action, meaning that no action is performed. This is to simulate the observation of the time point when all target agents have performed the action but the agent has not yet.

The simulated next observation $\hat{o}_{t+1,i}$ can be determined based on the predicted action set $\hat{\mathbf{a}}_{t,-i}$ and the current observation $o_{t,i}$. The function of the next observation simulation $\mathcal{D}(\cdot)$ is defined as follows:

$$\hat{o}_{t+1,i} = \mathcal{D}(o_{t,i}, \hat{\mathbf{a}}_{t,-i}, a_{t,i} = \emptyset).$$

The action selection using the simulated next observation $\hat{o}_{t+1,i}$ allows the agent to ignore the influence of the others'

[1] By specifying the distribution of π, (3) can be reformulated. In Appendix F, an example of the Gaussian distribution of continuous action $\pi(a^c_{t,j}|o_{t,j};\mathbf{c})$ and the Dirac delta distribution of discrete action $\pi(a^d_{t,j}|o_{t,j};\mathbf{c})$ is provided.

[2] If the agent i cannot observe the entire set of agents, a subset of the agent can be the targets of agent i, i.e., $E_{\mathcal{O}_i} \subset E$.

Algorithm 3 Episodic future thinking mechanism

Require: Trained actor-network ϕ, discrete action space parameter W, set of inferred characters $\hat{\mathbf{c}}_{-i}$, character of agent $\mathbf{c}_i$, initial state s_1
for $t = 1, T$ **do**
  Get observation $o_{t,i} \sim \Omega_i(s_t)$
  `// Start future simulation //`
  for $j = 1, N(j \neq i)$ **do**
    Get observation $o_{1,j} \sim \Omega_j(s_t)$
    Predict action of agents j
    $$\hat{a}_{t,j} \sim \pi_\phi(\cdot|o_{t,j}; \mathbf{c}_j)$$
    Store $\hat{a}_{t,j}$ in predicted action set $\hat{\mathbf{a}}_{t,-i}$
  end for
  Simulate future observation of agent i
  $$\hat{o}_{t+1,i} = \mathcal{D}(o_t, \hat{\mathbf{a}}_{t,-i}, a_{t,i} = \emptyset)$$
  `// End simulation //`
  Get proto-action $\{a^c_{t,i}, \bar{a}^d_{t,i}\} \sim \pi_\phi(\cdot|\hat{o}_{t+1,i}; \mathbf{c}_i)$
  Get post-action $a^d_{t,i} \leftarrow g(\bar{a}^d_{t,i}, W)$
  Execute $a_{t,i} = \{a^c_{t,i}, a^d_{t,i}\}$, Update s_{t+1}
end for

actions. This is because the next state is determined solely by its own action $a_{t,i}$ in the agent's mental simulation, as $\hat{o}_{t+1,i}$ has already applied the other agents' actions $\hat{\mathbf{a}}_{t,-i}$.

4.3. Action selection

Once the agent has simulated the next observation $\hat{o}_{t+1,i}$, the agent can make a foresighted decision. The agent uses the multi-character policy π_ϕ with the input of the simulated next observation $\hat{o}_{t+1,i}$ and its own character $\mathbf{c}_i$, and finally gets the action $a_{t,i} = \{a^c_{t,i}, \bar{a}^d_{t,i}\} = \pi_\phi(\cdot|\hat{o}_{t+1,i}; \mathbf{c}_i)$. In other words, the agent can select an adaptive action with consideration for other agents' upcoming behaviors. The decision-making procedure with the proposed EFT mechanism is summarized in Algorithm 3.

5. Experiments

To select a suitable task that can verify the effectiveness of the proposed solution, we consider the following requirements. There should be multiple approaches to achieving character diversity, as well as interactions between agents. The agent should have only partial observations of the state, and the action space should be both continuous and discrete.

We chose the autonomous driving task, which has numerous automated vehicles on the road. The task can consider the driving character of the agent based on driving preferences (e.g., one agent prioritizes safety and the other prioritizes speed) (Rosbach et al., 2019; Eboli et al., 2017; Cooper et al., 2002). Additionally, it is realistic for a driver to behave under the partial observation of the road state, and the driver makes a decision in a hybrid action space. To implement this task, we use the FLOW framework (Vinitsky et al., 2018).

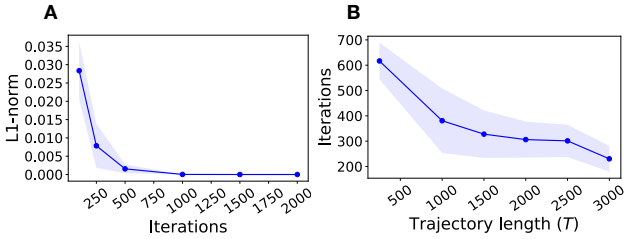

Figure 4. The performance of the character inference module. **A**. L1-norm between estimated and true characters over the number of iterations ($T = 1000$). **B**. The number of required iterations for convergence over the length of the observation-action trajectory T.

The scenario includes multiple automated vehicles on the multi-lane roundabout road. The number of agents $|E| = 21$, and each agent decides on acceleration and lane change control given their observation. Here we express the driving character using weights of three reward terms. Thus, the dimension of the character vector is three, i.e., $\mathbf{c}_i = [c_{i,1}, c_{i,2}, c_{i,3}]$.[3] The target agent j is limited to the vehicles located in the observable area of the agent. All results in this section are averaged results of over 10 independent experiments. The markers indicate the average value, and the shaded area represents the confidence interval within one standard deviation.

5.1. Performance evaluation: character inference

To make the EFT mechanism more effective, an accurate character inference should be preceded. In this subsection, we investigate the character inference module with two questions:

- How many iterations does it require to achieve an accurate inference (in terms of repetition in Algorithm 2)?
- How long should the agent collect the observation-action trajectories of target agents (in terms of trajectory length T in Algorithm 2)?

In Figure 4, the performance of the character inference module is presented. To ignore the effect of the initial point in convergence, the initial point of the character is randomly selected. More results regarding the initial point are provided in Appendix I.

Figure 4**A** illustrates the convergence of the estimated character to the true one. The inaccuracy of inference is evalu-

[3]Details regarding the experiments are presented in Appendix G.

ated based on the L1-norm between the estimated character and the true one. Thus, a lower L1-norm implies higher inference accuracy. As the number of iterations increases, the L1-norm quickly decreases to approximately zero, meaning that the estimated value quickly converges to the true one. Specifically, if the number of iterations is set to over 500, high accuracy of the character inference can be achieved.

Figure 4**B** shows the trade-off between the length of observation-action trajectory T and the number of iterations required for the convergence. The convergence criterion is set to L1-norm $\leq 5 \times 10^{-4}$. The results demonstrate that the number of iterations for convergence decreases as longer trajectories are provided. Thus, the length of trajectories and the number of iterations can be jointly determined by considering system requirements.

5.2. Ablation study: character inference and EFT modules

We investigate the impact of two main modules (the character inference module and the EFT module) on performance by increasing character diversity levels of the heterogeneous society. The following three cases are compared.

- Proposed: the agent enables the EFT with the inferred character of other agents based on the character inference module.
- FCE-EFT: the agent experiences the FCE by assuming that all other agents have equal character to itself (i.e., $\mathbf{c}_j = \mathbf{c}_i, \forall j \in E$). So no character inference is required. The agent performs the EFT, but action prediction is performed based on the same character $\mathbf{c}_i$.
- without EFT (Fujimoto et al., 2018): the agent performs neither character inference nor the EFT mechanism. It treats the problem as a single agent RL and selects the best action given observation. The policy is trained based on the TD3.

In Figure 5, the average rewards of entire agents are presented over increasing the number of character groups.[4] The higher number of character groups means that more diverse characters coexist in society, and the higher reward implies better performance. Because the number of agents is fixed to $|E| = 21$, the number of members per group is $|E|/n$, where n denotes the number of groups. The members belonging to the same group have the same character $\mathbf{c}$. Note that each group character is randomly sampled from character space $\mathcal{C}$ in every independent experiment.

[4]Each market point is the average value of 10 independent test experiments. To obtain all results presented in Figure 5, we ran $7 \times 3 \times 10 = 210$ test experiments.

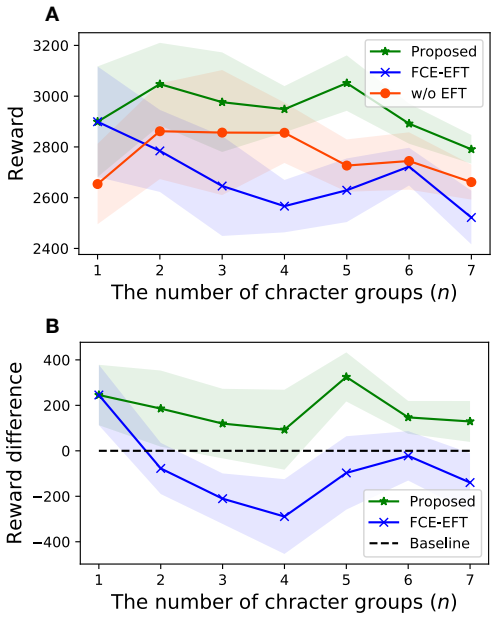

Figure 5. **A**: The average reward of entire agents over an increasing number of character groups. **B**: The amount of reward enhancement for two EFT approaches by setting without EFT as a baseline (i.e., reward of other approaches - reward of without EFT).

Figure 5**A** shows the average reward of entire agents. In a single group scenario (i.e., the entire agents have the same characters), the results of both the proposed and the FCE-EFT solutions are identical. This is because all agents have homogeneous characters, which allows the FCE agent to have the accurate characters of others. The reward of without EFT is lower than two solutions in a single group scenario. This confirms that the proposed EFT mechanism can help the agent to consider multi-agent interactions. Next, in two or more group scenarios, the proposed solution consistently achieves the highest reward, and the FCE-EFT consistently achieves the lowest reward.

Figure 5**B** highlights the amount of reward enhancement or degradation by equipping the proposed modules. As a baseline solution, without EFT is used. The proposed approach consistently outperforms the baseline, and the FCE-EFT is inferior to the baseline when character diversity exists. These results verify that the EFT mechanism with accurate character inference always enhances the reward. However, the naive employment of the EFT mechanism with the incorrect character degrades the reward. This is because incorrect character inference leads to incorrect action prediction and next observation simulation, which leads to improper action selection of the agent, leading to low reward. Therefore, accurate character inference is crucial in the EFT mechanism.

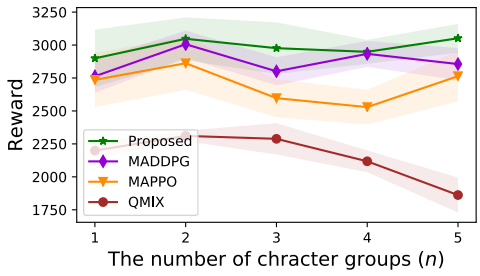

Figure 6. The performance comparison for the average reward between algorithms over increasing the number of character groups.

5.3. Performance comparison: multi-agent reinforcement learning algorithms

We compare the performance of the proposed solution to the following popular MARL algorithms. In MARL algorithms, we go through independent policy training regarding the diversity level of society.[5] This is the effort to have a fair comparison by allowing the best performance of comparing algorithms. Note that the proposed method does not need plural training for different heterogeneity settings.

- MADDPG (Lowe et al., 2017): It is a multi-agent version of Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2016). In training, it uses a centralized Q-function that uses observations and actions of all agents.
- MAPPO (Yu et al., 2022): It is a multi-agent version of Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017). It considers a centralized critic that uses the local observations across all agents.
- QMIX (Rashid et al., 2018): It uses a mixer and individual Q-networks. The mixer network uses the Q-values (output of individual Q-network) of all agents as inputs and calculates a global Q_{tot} as an output. Since it can only handle the discrete action space, we quantize the continuous actions.

Figure 5**B** shows the average reward of the entire agents as the number of character groups increases. This figure verifies that the proposed solution outperforms all popular MARL algorithms. Note that the MARL algorithms assume centralized training, which requires access to the observations and actions of all agents in policy training. In contrast, our solution trains the policy with only local observations and actions, which can be a more practical solution. The QMIX has the lowest performance since it operates in a discrete action space, whereas our task is in a hybrid action space. A summary of numerical results of Figure 5 are presented in Table 1.

[5]For each algorithm, five independent trainings are performed since five heterogeneity settings are considered, i.e., $n = [1, 2, 3, 4, 5]$.

Table 1. A summary of numerical results (avg. reward $\pm$ 1 std)

Algorithm	The number of character groups (n)				
	1	2	3	4	5
Proposed	2899 $\pm$217	3047 $\pm$162	2976 $\pm$196	2948 $\pm$91	3051 $\pm$109
FCE-EFT	2899 $\pm$217	2784 $\pm$161	2646 $\pm$196	2566 $\pm$103	2629 $\pm$125
w/o EFT	2653 $\pm$158	2861 $\pm$188	2856 $\pm$246	2855 $\pm$119	2726 $\pm$103
MADDPG	2763 $\pm$126	3006 $\pm$103	2800 $\pm$106	2933 $\pm$98	2856 $\pm$121
MAPPO	2753 $\pm$206	2862 $\pm$201	2597 $\pm$144	2529 $\pm$131	2763 $\pm$190
Q-MIX	2199 $\pm$56	2310 $\pm$39	2288 $\pm$118	2118 $\pm$82	1861 $\pm$132

6. Conclusions

In this paper, we propose the EFT mechanism, which is a social decision-making approach for a multi-agent scenario. The EFT mechanism enables the agent to behave by considering current and near-future observations. To achieve this functionality, we first build a multi-character policy that is generalized over character space. Then, the agent with the multi-character policy can infer others' characters using the observation-action trajectory. Next, the agent predicts the others' behaviors and simulates its future observation based on the proposed EFT mechanism. In the simulation result, we confirm that the proposed solution outperforms existing solutions across all diversity levels of the heterogeneous society.

The proposed EFT idea paves the way for research on multi-agent scenarios. The proposed method enables the agent to simulate other agents' upcoming actions, which is analogous to humans' decision-making. Furthermore, we believe the proposed method can be broadened by combining counterfactual thinking, current information, and future thinking.

Even though this work shows promising results with a novel method, there are a few limitations to tackle. In our experiments, there is only one EFT agent, and all other agents do not have the EFT functionality. This is an inevitable setting to make the problem tractable. Additionally, the character inference module relies on an iterative method, which hinders the solution from running in a real-time manner. This can be alleviated by selecting an efficient optimization solution or a deep learning based inference approach.

7. Acknowledge

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support programs (IITP-2022-2020-0-01602; No. 2021-0-00739, Development of Distributed/Cooperative AI based 5G+ Network Data Analytics Functions and Control Technology) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation)

References

Baron-Cohen, S. *Mindblindness: An essay on autism and theory of mind*. MIT press, 1997.

Byrne, R. M. Counterfactuals in eXplainable Artificial Intelligence (XAI): Evidence from human reasoning. In *IJCAI*, 2019.

Cooper, R. A., Thorman, T., Cooper, R., Dvorznak, M. J., Fitzgerald, S. G., Ammer, W., Song-Feng, G., and Boninger, M. L. Driving characteristics of electric-powered wheelchair users: how far, fast, and often do people drive? *Archives of Physical Medicine and Rehabilitation*, 83(2):250–255, 2002.

Dawes, R. M. Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology*, 25(1):1–17, 1989.

Doshi, P., Gmytrasiewicz, P., and Durfee, E. Recursively modeling other agents for decision making: A research perspective. *Artificial Intelligence*, 279:103202, 2020.

Dvijotham, K. and Todorov, E. Inverse optimal control with linearly-solvable MDPs. In *ICML*, 2010.

Eboli, L., Mazzulla, G., and Pungillo, G. How drivers' characteristics can affect driving style. *Transportation Research Procedia*, 27:945–952, 2017.

Engelmann, D. and Strobel, M. The false consensus effect disappears if representative information and monetary incentives are given. *Experimental Economics*, 3(3):241–260, 2000.

Engelmann, D. and Strobel, M. Deconstruction and reconstruction of an anomaly. *Games and Economic Behavior*, 76(2):678–689, 2012.

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., and Whiteson, S. Counterfactual multi-agent policy gradients. In *AAAI*, 2018.

Folli, D. and Wolff, I. Biases in belief reports. *Journal of Economic Psychology*, 88:102458, 2022.

Fujimoto, S., Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *ICML*, pp. 1587–1596, 2018.

Han, Y. and Gmytrasiewicz, P. IPOMDP-net: A deep neural network for partially observable multi-agent planning using interactive POMDPs. In *AAAI*, 2019.

Janner, M., Fu, J., Zhang, M., and Levine, S. When to trust your model: Model-based policy optimization. In *NeurIPS*, 2019.

Jern, A., Lucas, C. G., and Kemp, C. People learn other people's preferences through inverse decision-making. *Cognition*, 168:46–64, 2017.

Kwon, M., Daptardar, S., Schrater, P. R., and Pitkow, Z. Inverse rational control with partially observable continuous nonlinear dynamics. In *NeurIPS*, 2020.

Lai, H., Shen, J., Zhang, W., and Yu, Y. Bidirectional model-based policy optimization. In *ICML*, 2020.

Langley, C., Cirstea, B. I., Cuzzolin, F., and Sahakian, B. J. Theory of mind and preference learning at the interface of cognitive science, neuroscience, and AI: A review. *Frontiers in Artificial Intelligence*, pp. 62, 2022.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. In *ICLR*, 2016.

Lin, H., Sun, Y., Zhang, J., and Yu, Y. Model-based reinforcement learning with multi-step plan value estimation. *arXiv preprint arXiv:2209.05530*, 2022.

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O., and Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. In *NeurIPS*, 2017.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., Markson, L., and Hu, J. The child as econometrician: A rational model of preference understanding in children. *PloS one*, 9(3):e92160, 2014.

Lynch, J. G., Alba, J., and Hutchinson, J. W. Memory and decision making. *Handbook of Consumer Behavior*, pp. 1–9, 1991.

Marinescu, A., Dusparic, I., and Clarke, S. Prediction-based multi-agent reinforcement learning in inherently non-stationary environments. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 12(2):1–23, 2017.

Marks, G. and Miller, N. Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin*, 102(1):72, 1987.

Mehta, V., Paria, B., Schneider, J., Ermon, S., and Neiswanger, W. An experimental design perspective on model-based reinforcement learning. In *ICLR*, 2022.

Moerland, T. M., Broekens, J., Plaat, A., Jonker, C. M., et al. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 16(1): 1–118, 2023.

Ng, A. Y., Russell, S. J., et al. Algorithms for inverse reinforcement learning. In *ICML*, 2000.

Oberst, M. and Sontag, D. Counterfactual off-policy evaluation with Gumbel-max structural causal models. In *ICML*, 2019.

Oliehoek, F. A. Decentralized POMDPs. In *Reinforcement Learning*, pp. 471–503. Springer, 2012.

Owen, G. *Game theory*. Emerald Group Publishing, 2013.

Pan, T., Sumalee, A., Zhong, R.-X., and Indra-Payoong, N. Short-term traffic state prediction based on temporal–spatial correlation. *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1242–1254, 2013.

Premack, D. and Woodruff, G. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4): 515–526, 1978.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., and Botvinick, M. Machine theory of mind. In *ICML*, 2018.

Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., and Whiteson, S. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *ICML*, 2018.

Ratliff, N. D., Bagnell, J. A., and Zinkevich, M. A. Maximum margin planning. In *ICML*, 2006.

Redish, A. D. and Mizumori, S. J. Memory and decision making. *Neurobiology of Learning and Memory*, 117:1, 2015.

Rosbach, S., James, V., Großjohann, S., Homoceanu, S., and Roth, S. Driving with style: Inverse reinforcement learning in general-purpose planning for automated driving. In *IROS*, pp. 2658–2665, 2019.

Ross, L., Greene, D., and House, P. The false consensus phenomenon: An attributional bias in self-perception and social perception processes. *Journal of Experimental Social Psychology*, 13(3):279–301, 1977.

Roth, A. E. and Erev, I. Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and economic behavior*, 8(1): 164–212, 1995.

Schacter, D. L., Benoit, R. G., De Brigard, F., and Szpunar, K. K. Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions. *Neurobiology of Learning and Memory*, 117:14–21, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Schwartz, J., Zhou, R., and Kurniawati, H. Online planning for interactive-POMDPs using nested Monte Carlo tree search. In *IROS*, 2022.

Sutton, R. S. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 2(4):160–163, 1991.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Thorstad, R. and Wolff, P. A big data analysis of the relationship between future thinking and decision-making. *Proceedings of the National Academy of Sciences*, 115 (8):E1740–E1748, 2018.

Vinitsky, E., Kreidieh, A., Le Flem, L., Kheterpal, N., Jang, K., Wu, C., Wu, F., Liaw, R., Liang, E., and Bayen, A. M. Benchmarks for reinforcement learning in mixed-autonomy traffic. In *CoRL*, 2018.

Xu, X., Lv, K., Dong, X., Han, S., and Lin, Y. Multi-step prediction for learning invariant representations in reinforcement learning. In *HPBD&IS*, 2021.

Yang, L., Zhao, C., Lu, C., Wei, L., and Gong, J. Lateral and longitudinal driving behavior prediction based on improved deep belief network. *Sensors*, 21(24):8498, 2021.

Yasdi, R. Prediction of road traffic using a neural network approach. *Neural Computing & Applications*, 8(2):135–142, 1999.

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., and Wu, Y. The surprising effectiveness of PPO in cooperative multi-agent games. In *NeurIPS*, 2022.