Theory of Mind with Guilt Aversion Facilitates Cooperative Reinforcement Learning

Dung Nguyen dung.nguyen@deakin.edu.au
Svetha Venkatesh svetha.venkatesh@deakin.edu.au
Phuoc Nguyen phuoc.nguyen@deakin.edu.au
Truyen Tran truyen.tran@deakin.edu.au
Applied Artificial Intelligence Institute, Deakin University, Geelong, Australia

Editors: Sinno Jialin Pan and Masashi Sugiyama

Abstract

Guilt aversion induces experience of a utility loss in people if they believe they have disappointed others, and this promotes cooperative behaviour in human. In psychological game theory, guilt aversion necessitates modelling of agents that have theory about what other agents think, also known as Theory of Mind (ToM). We aim to build a new kind of affective reinforcement learning agents, called Theory of Mind Agents with Guilt Aversion (ToMAGA), which are equipped with an ability to think about the wellbeing of others instead of just self-interest. To validate the agent design, we use a general-sum game known as Stag Hunt as a test bed. As standard reinforcement learning agents could learn suboptimal policies in social dilemmas like Stag Hunt, we propose to use belief-based guilt aversion as a reward shaping mechanism. We show that our belief-based guilt averse agents can efficiently learn cooperative behaviours in Stag Hunt Games.

1. Introduction

People in a group may be willing to give more and take less. This may appear irrational from the individual perspective, but such behaviour often enables the group to achieve higher returns than acting selfishly. Therefore, in building artificial multi-agent systems, it is important to construct social inductive biases about the reasoning of other agents - also known as the Theory of Mind (ToM) (Rabinowitz et al., 2018; Shum et al., 2019). Theory of mind enables individuals to cooperate and this often results in optimal group rewards (Shum et al., 2019; Takagishi et al., 2010).

A mechanism to encourage social cooperation is maintaining *fair* outcomes for members of the group, and agents who do so are termed 'inequity averse' (Hughes et al., 2018). Other mechanisms stem from *guilt* (Haidt, 2012), requiring one to put themselves in the others' shoes (Chang et al., 2011; Morey et al., 2012). To be *guilt averse, the agent needs higher-order ToM - i.e.* be able to estimate what others will do (0-order ToM), and what others believe the agent itself will do (1-order ToM) (Albrecht and Stone, 2018). Inequity aversion, on the other hand, is conceptually different from guilt aversion (Nihonsugi et al., 2015) and does not require theory of mind. We focus on the computational mechanisms to control the interplay between the greedy tendencies of an individual and the inferred needs of others in a reinforcement learning (RL) setting. In (Moniz Pereira et al., 2017; Rosenstock and

O'Connor, 2018), authors analysed the evolutionary dynamics of agents with guilt, but did not include theory of mind. There has been early work to integrate theory of mind and guilt aversion in a psychological game setting (Battigalli and Dufwenberg, 2007). The first work to examine social dilemmas in a deep reinforcement learning setting is (Hughes et al., 2018; Peysakhovich and Lerer, 2018a) in which the authors incorporate knowledge from behavioural game theory when training the agents. However, guilt aversion, which plays a central role in moral decisions (Haidt, 2012) has not been considered.

Our paper addresses the open challenges of integrating theory of mind and guilt aversion into Multi-Agent Reinforcement Learning (MARL) (Littman, 1994) and studies the evolution of cooperation in such agents in self-play settings. We name the agent ToMAGA, which stands for *Theory of Mind Agent with Guilt Aversion*. In our agents, learning is driven by not only material rewards but also psychological loss due to the feeling of guilt if an agent believes that it has harmed others. Our computational model of theory of mind extends the work of (De Weerd et al., 2013) to build agents with beliefs about cooperative behaviours rather than just primitive actions. Our reinforcement learning agent uses a value function to make sequential decisions. At each learning step, after observing the other agents' actions, the agent updates its beliefs about the other agents, including what they might think about it. Then it computes psychological rewards using a guilt averse model, followed by an update of the value function. In other words, this implements a reward shaping strategy, where the additional reward is from the intrinsic social motivation of being fair to others. In reinforcement learning, reward shaping helps to guide the exploration and increase the convergence speed of the algorithm. Different from (Devlin and Kudenko, 2011) in which the reward shaping function was defined over the state space, our reward shaping function, on the other hand, is defined over actions space. To help understand how this reward shaping is effective in social dilemmas, we construct a theoretical argument to show that guilt aversion implemented as reward shaping can change the Stag Hunt game from having two pure Nash equilibria into having one pure Nash equilibrium that is Pareto efficient. In addition, our agents are able to cooperate in the grid-world Stag Hunt Games, in which the rewards given to each agent depend on the sequence of actions (at the policy level), not just on one action like in matrix-form games. We build several environments, both in a one-step decision game and in a multi-step grid-world. Our extensive suite of experiments demonstrates that modelling guilt with explicit theory of mind helps reinforcement learning agents to cooperate better than those without theory of mind, encouraging faster learning towards cooperative behaviours. At last, we demonstrate the efficiency of our reward shaping mechanism on more complex rewards structure and action space environments. We also demonstrate that the mechanism can handle the case in which there are more than two agents.

Our contribution is to design and test a framework that brings the psychological concept of guilt aversion into multi-agent reinforcement learning, and in effect it connects social psychology, psychological game theory (Geanakoplos et al., 1989), multi-agent systems and reinforcement learning. For the first time, we explore and establish a computational model for embedding guilt aversion coupled with theory of mind on reinforcement learning framework and study it in the extended Markov Games.

	C	U
C	h, h	g, c
U	c, g	m, m

Table 1: The structure of Stag Hunt ($h > c > m > g$).

2. Preliminaries

2.1. Two-player Markov Games

A two-player fully observable Markov Game is a tuple $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P} \rangle$, where $\mathcal{N} = \{1, 2\}$ denotes the set of two players, $\mathcal{S}$ is the state space, $\mathcal{A} =\mathcal{A}_1 \times \mathcal{A}_2$ is the joint action space, $\mathcal{R} =\mathcal{R}_1 \times \mathcal{R}_2$ is the reward space with $\mathcal{R}_1, \mathcal{R}_2 : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$, $\mathcal{P}$ is the transition function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ denotes the probability distribution over $\mathcal{S}$. Each agent i takes an action $a_i \in \mathcal{A}_i$ based on its policy $\pi_i : \mathcal{S} \mapsto \mathcal{A}_i$. Denote by Π_i the set of all policies available to the player i. The set of joint policies is $\Pi = \Pi_1 \times \Pi_2$.

Definition 1. A joint policy $\pi = (\pi_1, \pi_2) \in \Pi$, denoted as $\pi^C = (\pi_1^C, \pi_2^C)$, is a *cooperative joint policy* iff

$$\pi = \text{argmax}_{\pi_1 \in \Pi_1, \pi_2 \in \Pi_2} \mathbb{E}_{a_1 \sim \pi_1, a_2 \sim \pi_2, s_{t+1} \sim \mathcal{P}} \left[R \right], \text{ for}$$
$$R = \sum_{t=0}^{\infty} \gamma^t \left[r_1(a_1, a_2, s_t) + r_2(a_1, a_2, s_t) \right]$$

If two agents follow a *cooperative joint policy* π^C, we say that two agents have *cooperative behaviours.* We denote Π^C as a set of *cooperative joint policies.*

Definition 2. A policy π_i is a *cooperative policy* iff

$$\exists j \in \mathcal{N} \backslash i, \pi_j \in \Pi_j : (\pi_1, \pi_2) \in \Pi^C.$$

We denote Π_i^C as a set of cooperative policies. A policy $\pi_i \in \Pi_i \backslash \Pi_i^C$ of agent i is called an *uncooperative policy.* We denote $\Pi_i^U = \Pi_i \backslash \Pi_i^C$ as a set of uncooperative policies.

The definition says that if the agent i follows policy π_i and there exists at least one policy π_j of agent j that their joint policy $\pi = (\pi_1, \pi_2)$ is a cooperative joint policy, then the policy π_i a cooperative policy.

2.2. The Stag Hunt Game

Stag Hunt is a coordination game of two persons hunting together (Macy and Flache, 2002). If they hunt stag together, they can both obtain a large reward h. However, one can choose to trap hare gaining a reward, sacrificing the other's benefit. The reward matrix is shown in Table 1. The game has two pure Nash equilibria: (1) both hunting stag, which is Pareto optimal; (2) or both hunting hare. If one player thinks the other will choose to hunt hare, her best response will be hunting hare. This is because in the worst case, if hunting hare, the player will receive a reward m. This amount is larger than g which is the worst case if she

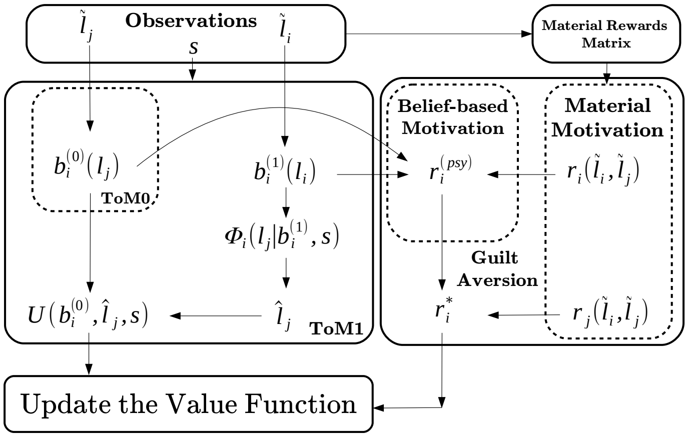

Figure 1: The learning process in Theory of Mind Agents with Guilt Aversion (ToMAGAs).

hunts stag. Therefore, both hunting hare is the *risk-dominant* equilibrium (Harsanyi et al., 1988). Here, the dilemma is that the *risk-dominant* Nash equilibrium is not the Pareto optimal. There is one mixed Nash equilibrium but its common outcome is not Pareto optimal. Because both will receive the highest collective rewards when jointly hunting stag, both hunting stag is a joint cooperative policy. Therefore, hunting stag is a cooperative policy that is also Pareto efficient.

3. Theory of Mind Agents with Guilt Aversion

We present our agent model named *Theory of Mind Agent with Guilt Aversion* (ToMAGA). The internal working process of the agent is illustrated in Fig. 1. It has a ToM module that is augmented with a guilt aversion (GA) component. We detail these parts as follows.

3.1. Settings

In our setting, an agent learns: (1) to predict whether the other agent follows a cooperative policy or an uncooperative policy; and (2) a cooperative policy. These objectives are also a part of things to learn in a fully observable games in the area of Multi-agent Learning (MAL) (Shoham et al., 2007). During the k^{th} iteration, agent i follows its policy $\pi_i^{(k)}$. For readability, we omit the superscript (k). At each time step t of an iteration in the game, two agents simultaneously take actions, hence, generating a trajectory of *experiences* $\tau = \left(s^{(t)}, a_{1,}^{(t)}a_{2}^{(t)}, r_1^{(t+1)}, r_2^{(t+1)}, s^{(t+1)}\right)_{t=0}^{T-1}$. We make the following assumptions about the observations of agents and the reward structure of the training environment

Assumption. *(about the observations) In any iteration k, a policy π_i for $i \in \mathcal{N}$ belongs to the set of joint cooperative policy Π_i^C or the set of uncooperative policies Π_i^U. Both agents can observe this information.*

We denote by $l_i = C$ the event that at iteration k the policy of agent i is a *cooperative policy* $\pi_i \in \Pi_i^C$ and by $l_i = U$ the event that the policy of agent i is an *uncooperative policy* $\pi_i \in \Pi_i^U$.

Assumption. *(about the reward structure) At the last time step T of an iteration, after reaching the termination state of the game, the agents receive material rewards $r_1^{(T)}(l_i, l_j)$ and $r_2^{(T)}(l_i, l_j)$. The rewards follow the structure of the Stag Hunt Game described in Table 1.*

3.2. First-order Theory of Mind (ToM1) Agent

We construct ToM1 agents as in (De Weerd et al., 2013). Agent i maintains two beliefs: (1) zero-order belief $b_i^{(0)}(l_j)$ for $l_j \in \{C, U\}$ which is a probability distribution over events that agent $j \neq i$ follows a cooperative or an uncooperative policy; and (2) first-order belief $b_i^{(1)}(l_i)$ for $l_i \in \{C, U\}$, which is a recursive belief, representing what agent i thinks *about the belief of agent j's belief (the probability distribution over events that agent i follows a cooperative or an uncooperative policy).* At the end of each iteration, agent i observes a trajectory τ and the information about whether the executed policies were cooperative or uncooperative, i.e. $\left\{\tilde{l}_i, \tilde{l}_j\right\}$. Agent i first predicts whether agent j uses a cooperative or an uncooperative policy. The prediction is based on the current first-order belief $b_i^{(1)}(l_i)$ as follows

$$\hat{l}_j = \text{argmax}_{l_j \in \{C,U\}} \mathit{\Phi}_{ij}(l_j) \text{ where}$$
$$\mathit{\Phi}_{ij}(l_j) = \sum_{l_i \in \{C,U\}} b_i^{(1)}(l_i) \times r_j^{(T)}(l_i, l_j),$$

where $\mathit{\Phi}_{ij}(l_j)$ is the value function agent i thinks agent j will have if agent j greedily maximises its material reward. Now, agent i has two guesses about the agent j: the zero-order belief $b_i^{(0)}(l_j)$ and policy type $\hat{l}_j$. To combine these two pieces of information into the belief about the action of agent j, called a belief integration function $BI(l_j)$. To do this, agent i maintains and updates a confidence $c_{ij} \in [0, 1]$ about its ToM1 as follows:

$$c_{ij} \qquad (1-\lambda)c_{ij} + \lambda\delta\left[l_j = \hat{l}_j\right]$$

for learning rate $\lambda \in [0, 1]$ and identity function $\delta[\cdot]$. After updating the confidence, agent i then computes its belief integration function

$$BI(l_j) \qquad (1 - c_{ij})\, b_i^{(0)}(l_j) + c_{ij}\delta\left[l_j = \hat{l}_j\right]$$

for all $l_j \in \{C, U\}$. Now the agent i can update its zero-order belief as

$$b_i^{(0)}(l_j) \qquad BI(l_j),$$

for all $l_j \in \{C, U\}$ and first-order belief as

$$b_i^{(1)}(l_i) \qquad (1 - c_{ij})\, b_i^{(1)}(l_i) + c_{ij} \times \delta\left[l_i = \tilde{l}_i\right],$$

for all $l_i \in \{C, U\}$.

3.3. Guilt Aversion (GA)

The guilt averse agent i will experience a utility loss if it thinks it lets the other agent down. The utility loss is realised through reward shaping. More concretely, once beliefs are updated, the agent i first computes an expected material value experienced by the agent j:

$$\phi_j = \sum_{l_i, l_j \in \{C, U\}} b_i^{(0)}(l_j) \times b_i^{(1)}(l_i) \times r_j^{(T)}(l_i, l_j) \tag{1}$$

where $r_j^{(T)}(l_i, l_j)$ is the material reward received after the last time step T. In addition, the agent experiences a psychological reward of "feeling guilty", caring about how much it lets the other down, as (Battigalli and Dufwenberg, 2007):

$$r_i^{(psy)}(\tilde{l}_i, \tilde{l}_j) = -\theta_{ij} \max\left(0, \phi_j - r_j^{(T)}(\tilde{l}_i, \tilde{l}_j)\right) \tag{2}$$

where guilt sensitivity $\theta_{ij} > 0$. The reward is then shaped as:

$$r_i^* = r_i^{(T)}(\tilde{l}_i, \tilde{l}_j) + r_i^{(psy)}(\tilde{l}_i, \tilde{l}_j). \tag{3}$$

This computation is based on an assumption that a guilt averse agent *does not* know whether the other is guilt averse.

3.4. Update the Value Function

Algorithm 1: ToMAGA i

```
Input  : $K$ is the number of iterations
         $T_{max}$ is the maximum timesteps per iteration
Output: The policy $\pi_i$
for $k$   $0$ to $K-1$ do
    Reset $\tau^{(k)}$;
    for $t \leftarrow 0$ to $T_{max}$ do
        Takes action $\tilde{a}_i^{(t)}$ based on the value function;
        Add the new experience to $\tau$;
        if $s_t$ is the termination state then
            break;
        end
    end
    Get information about policies $\left\{\tilde{l}_i, \tilde{l}_j\right\}$;
    Update beliefs $b_i^{(0)}(l_j)$ and $b_i^{(1)}(l_i)$ ;
    Compute psychological reward $r_i^{(psy)}(\tilde{l}_i, \tilde{l}_j)$;
    forall experiences in $\tau$ do
        Update the value function by using $r_i^*$ in Eq. 3;
    end
end
```

	C	U
C	$h + r^{(psy)}, h + r^{(psy)}$	$g + r^{(psy)}, c + r^{(psy)}$
U	$c + r^{(psy)}, g + r^{(psy)}$	$m + r^{(psy)}, m + r^{(psy)}$

Table 2: The reward structure of Stag Hunt games after having psychological reward factor defined in Eq. 2.

Given the shaped reward in Eq. (3), the reinforcement learning agent learns by updating the value function as follows.

Matrix-form Stag Hunt Because the size of the state space $|\mathcal{S}| = 1$, the strategy of agent i reduces to select action a_i, and agent i updates its value function based on temporal difference algorithm TD(1):

$$\begin{aligned} V_i(\tilde{a}_i) \quad & V_i(\tilde{a}_i) + \alpha\Delta_i, \text{ where} \\ \Delta_i = r_i^* + \gamma \max_{a_i} & \sum_{a_j} b_i^{(0)}(a_i) r_i(a_i, a_j) - V_i(\tilde{a}_i) \end{aligned}$$

General Stag Hunt with Deep Reinforcement Learning In the general Stag Hunt games, we parameterise the value function and policy by deep neural networks trained by the Proximal Policy Optimization (PPO) (Schulman et al., 2017). The training algorithm is shown in Algorithm 1.

3.5. Theoretical Analysis

We now show that guilt aversion implemented as reward shaping can change the Stag Hunt game from having two pure Nash equilibria into having one pure Nash equilibrium that is Pareto efficient. We recall that ToMAGAs play the game with a new pay-off matrix as shown in Table 2. We then establish the following observations.

Observations: **(1)** *If there exists a sequence of trajectories leading to $\phi_j > m$ and $\theta_{ij} > \frac{m-g}{min(\phi_j,c)-m}$ with $i, j \in \{1, 2\}, i \neq j$, this game will have only one pure Nash equilibrium, in which both players choose to cooperate (C, C); and* **(2)** *ToMAGA with higher guilt sensitivity θ_{ij} will have a higher chance of converging to this pure Nash equilibrium in self-play setting.*

Proof. This game will have only one pure Nash equilibrium (NE), in which both players choose to cooperate (C, C), when two conditions hold:

$$\begin{aligned} &\text{(C1) } h - \theta_{ij}\max(0, \phi_j - h) > c - \theta_{ij}\max(0, \phi_j - g) \\ &\text{(C2) } \ g - \theta_{ij}\max(0, \phi_j - c) > m - \theta_{ij}\max(0, \phi_j - m) \end{aligned}$$

for $h > c > m > g$, the sensitivity $\theta_{ij} > 0$, and the expected material value experienced by other agent $\phi_j \in [g, h]$ described in Eq. 1. (C1) holds within the structure of the Stag Hunt game. When $\phi_j \in (c, h]$, (C2) is satisfied iff $\theta_{ij} > \frac{m-g}{c-m}$. When $\phi_j \in (m, c]$, (C2) is satisfied iff $\theta_{ij} > \frac{m-g}{\phi_j - m}$. Therefore, the first observation is proved. To prove the second observation, we

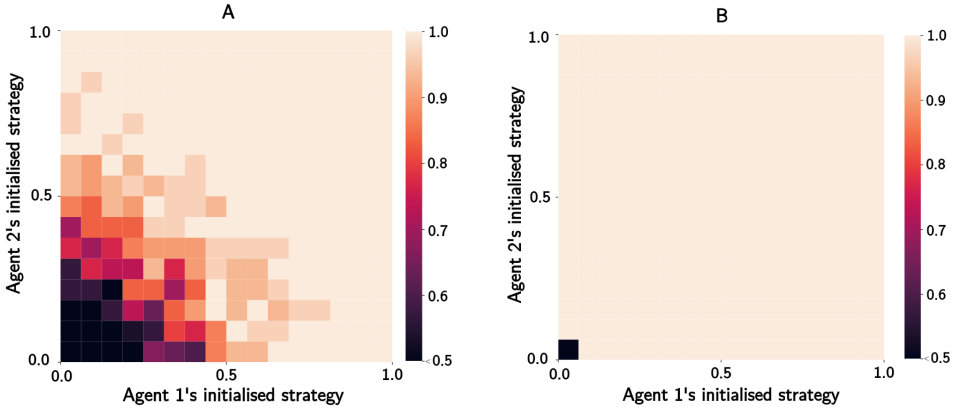

Figure 2: Initial probability of the second player following cooperative strategy (y-axis) vs Initial probability of the first player following cooperative strategy (x-axis). The colour shows the probability (lighter values indicate higher probability) of the first player following cooperative strategy after 500 timesteps of (A) Guilt averse agents without theory of mind and (B) ToMAGAs.

consider the case when $\phi_j \in (m, c]$, the condition $\theta_{ij} > \frac{m-g}{\phi_j - m} \Leftrightarrow \phi_j > \left(m + \frac{m-g}{\theta_{ij}}\right) \triangleq f(\theta_{ij})$ implies $\phi_j \in (f(\theta_{ij}), c]$. Because $f(\theta_{ij})$ is a decreasing function, the chance of ϕ_j belongs to $(f(\theta_{ij}), c]$ is increasing when θ_{ij} is increasing, i.e. the second observation is proved. □

By introducing the psychological rewards, we increase the probability of changing the game from two Nash equilibria to one Nash equilibrium, which intuitively helps the reinforcement learning algorithm converge to Pareto efficient Nash equilibrium. In other words, the higher guilt sensitivity the agents have, the more chance that they will converge to the cooperative behaviour. During the exploration, both agents need to obtain higher beliefs about the event that other will choose a cooperative policy. If both agents believe that the expectation of other are higher than the outcome that players received when they are at risk-dominant Nash equilibrium, i.e. $\phi_j > m$, then higher θ_{ij} will lead to higher chance to converge to NEs. However, initially, if both agents believe that the expectation of other are equal or lower than the inefficient outcome, i.e. $\phi_j \leq m$, the agents *need to* increase this expectation during the training process.

4. Experiments

We test our ToMAGAs in three environments: (1) Matrix-form Stag Hunt Games; (2) Grid-world Stag Hunt Games; and (2) The modified version of Stag Hunt Games called *Island* with the more complex reward structure and action space.

4.1. Matrix-Form Stag Hunt Games

In this experiment, we aim to answer two questions:

Q1: *How does ToM model affect cooperative behaviour in the self-play setting?* We compare the behaviour of ToMAGAs and GA agents *without* ToM that do not update first order beliefs. All agents have the guilt sensitivity $\theta_{ij} = 200$. The initial probabilities of each agents to follow a cooperative strategy constitute the grid index in Figure 2). We measure the probability of the agents following cooperative policy after 500 timesteps of playing the matrix-form games with $h = 40, c = 30, m = 20, g = 0$. Figure 2 shows that ToMAGAs promote cooperation better than the guilt averse agents without theory of mind. This is more pronounced in settings where agents are initialised with a low probability of following cooperative strategy (to the left bottom corner of Figure 2-A and 2-B).

Q2: *How does ToMAGA promote cooperative behaviour in a group of agents?* The experiments are designed similarly to the tournament commonly used in studies of how cooperation could evolve (Axelrod and Hamilton, 1981). In each round, agents are randomly matched and each pair plays the matrix-form of Stag Hunt game with $h = 5, c = 4, m = 2, g = 1$. We report the average common reward of the last 100 rounds after 5000 rounds of interaction.

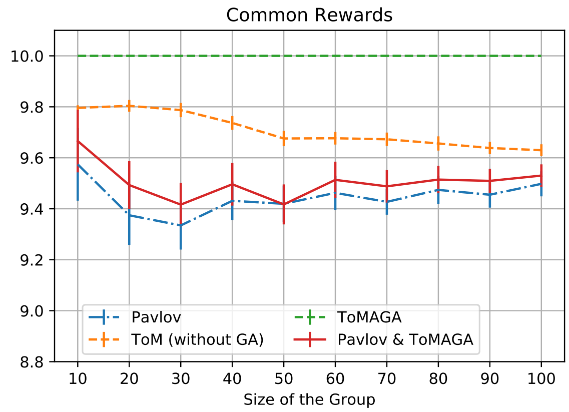

Figure 3: Common Reward (y-axis) vs Size of Group (x-axis). ToMAGAs encourage the cooperation in both homogeneous and heterogeneous groups. Common reward is higher when a group contains ToMAGAs.

There are two types of groups: homogeneous group and heterogeneous group; and two types of agents: ToMAGA and Pavlov agent. We compare behaviours of ToMAGAs in groups with a general version of Win-Stay-Lose-Shift (WSLS) strategy, called Pavlov agent, which is a popular strategy for solving Stag Hunt. A Pavlov agent (Kraines and Kraines, 1996) chooses to hunt stag with probability $p_n = \frac{i}{n}$ with $0 \leq i \leq n$ and updates the strategy based on the outcome it received and actions that both took in the last interaction. The probability cooperatively hunt stag p_n is increased when two players matched their behaviours, and p_n is decreased otherwise. In the heterogeneous group of N agents, there are $(N - 1)$ Pavlov agents and a ToMAGA. From the structure of Stag Hunt games, if one group has more agents with cooperative behaviours, they will obtain higher average common rewards. Fig. 3 shows that the homogeneous group of ToMAGAs cooperate better than the homogeneous group of Pavlov agents. As the size of heterogeneous groups is small, having one ToMAGA will enhance the cooperation and help to obtain higher common rewards. When the size of the group increases, the homogeneous group of ToM agents without guilt aversion converge much slower than the group of ToMAGA. This leads to the homogeneous group of ToM agents without guilt aversion has lower common rewards than the homogeneous group of ToM agents after 5000 rounds of interaction.

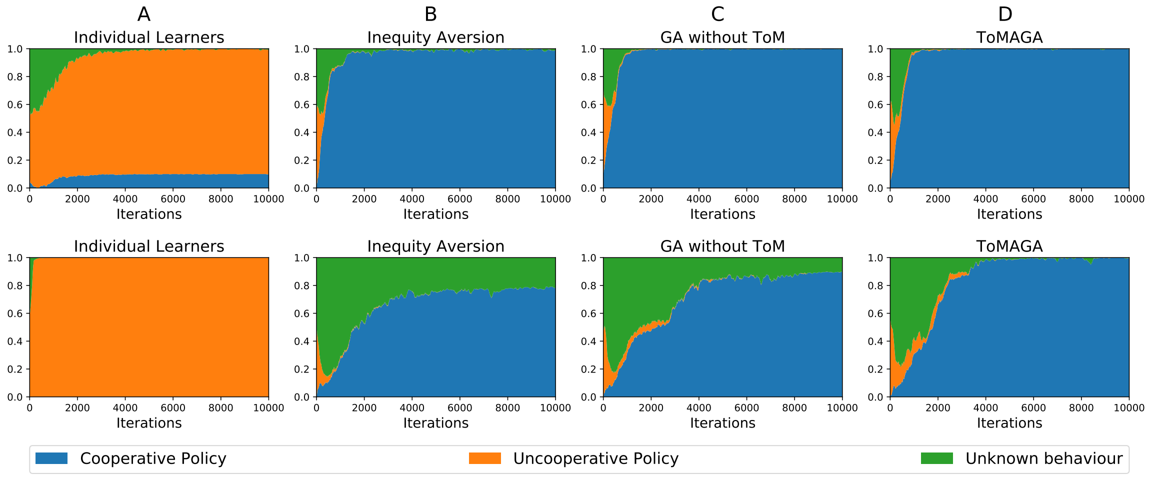

Figure 5: Policies of individual learners (column A), agents with inequity aversion (column B), GA agents without ToM (column C), and ToMAGAs (column D) when they start nearby the stag (the first row) and nearby hares (the second row). Proportion of following cooperative (blue), uncooperative (orange), unknown (green) behaviours (y-axis) vs Iterations (x-axis).

4.2. Grid-World Stag Hunt Games

In the grid-world Stag Hunt games, two players simultaneously move in a fully observable 4×4 grid-world, and try to catch stag or hare by moving into their squares (see Fig. 4). Every timestep, each player can choose among 5 actions {`left, up, down, right, stay`}. While the players need to cooperate to catch the stag, i.e. both move to the position of the stag at the same time, each player can decide to catch the hare alone. The rewards given to agents follow the reward structure of the Stag Hunt games. In detail, if two players catch the stag together, the reward given to each player is 4.0. If two players catch the hares at the same time, the reward given to each player is 2.0. Otherwise, the player catching the hare alone will receive a reward of 3.0, and the other will receive 0.0. The game is terminated when at least one player reaches the hare, two players catch the stag, or the time T_{max} runs out.

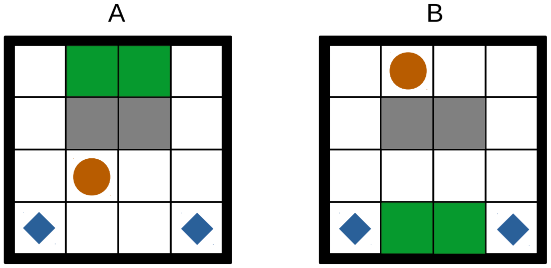

Figure 4: The grid-world version of Stag Hunt Games. Two agents (blue diamonds) learn to hunt the moving stag (brown circle) or the static hares (green cell) while avoiding the obstacles (in gray). Agents start (A) start nearby the stag, and (B) nearby the hares.

Recall in the section 2.2 that both players cooperatively catching the stag result in policies that are Pareto efficient. We are interested in two situations: At the beginning of the training process, agents are put (A) nearby the stag and far from the hares, and (B) put nearby the hares and far from the stag. We hypothesise that it is easier for agents to learn to cooperatively catch the stag if they are put nearby the stag at the beginning. After each iteration, each policy will be labelled as follows: (1) when both hunt the stag, labels are ($\tilde{l}_i = C, \tilde{l}_j = C$); (2) when one hunts hare and other hunts stag, labels are U and C, respectively; (3) when both hunt hare, labels are U; and (4) if the game is terminated, the policies of the agent who does not hunt hare or stag will be considered as unknown behaviours.

We construct deep reinforcement learning agents having both value network and policy network trained by PPO (Schulman et al., 2017). We compare the behaviours of four types of agents: (1) the individual learners; (2) the agents with inequity aversion (IA) (Hughes et al., 2018); (3) the GA agents without ToM; and (4) the ToMAGAs. Individual learners are agents that behave self-interest and only optimise their rewards. Inequity averse agents are agents that have a shaping reward $r_i^{(psy)} = -\frac{\theta_{ad}}{N-1} \times \sum\limits_{j \neq i} \max(r_i - r_j, 0) - \frac{\theta_{dis\_ad}}{N-1} \times \sum\limits_{j \neq i} \max(r_j - r_i, 0)$, where N is the number of agents, θ_{ad} and θ_{dis_ad} are advantageous and disadvantageous sensitivity, respectively (Fehr and Schmidt, 1999).

Fig. 5 shows the policies of deep reinforcement learning agents over the training process when they start nearby the stag (the first row of Fig. 5) and nearby the hares (the second row of Fig. 5). In both cases, the individual learners i.e. deep reinforcement learning agents without social preferences cannot learn to cooperate and even learn the uncooperative behaviours (to individually catch hares) since the very early stage of training process if they start nearby hares. In contrast, the deep reinforcement learning agents with social preferences can learn to cooperate in both cases. When the agents are put nearby stag at the beginning, the performance of inequity averse agents is comparable to GA agents without ToM and the ToMAGA (the first row of columns B, C, and D of Fig. 5). However, when initialised nearby hares, GA agents without ToM and ToMAGA learn to cooperate faster than the inequity averse agents (the second row of columns B, C, and D of Fig. 5). Also, in this case, our ToMAGAs can learn to cooperate faster than the GA without ToM because the GA without ToM does not update its first-order belief, leading to wrong predictions about the expectation of others.

4.3. The Island: ToMAGA in a Complex Environment

This suite of experiments aims to demonstrate the performance of our reward shaping mechanism on more complex environments, we consider the behaviours of ToMAGA on the Island and its extended version with more agents (Large-Island) (Wang et al., 2020). These environments are modified versions of the Stag Hunt game with more complex rewards structure and action space. In Island game, there are two agents and a beast in a 10 × 10 environment. Instead of only moving around by choosing `left, up, down, right, stay` as in the grid-world Stag Hunt games, the agent must get close and choose `attack` to kill the beast. Similarly, the beast also can move and attack the agents which are in its attack range. The beast and agents will have its own energy which will be reduced if they are attacked, and one will be killed if their energy is equal 0. The agents will receive a reward 300 if they kill

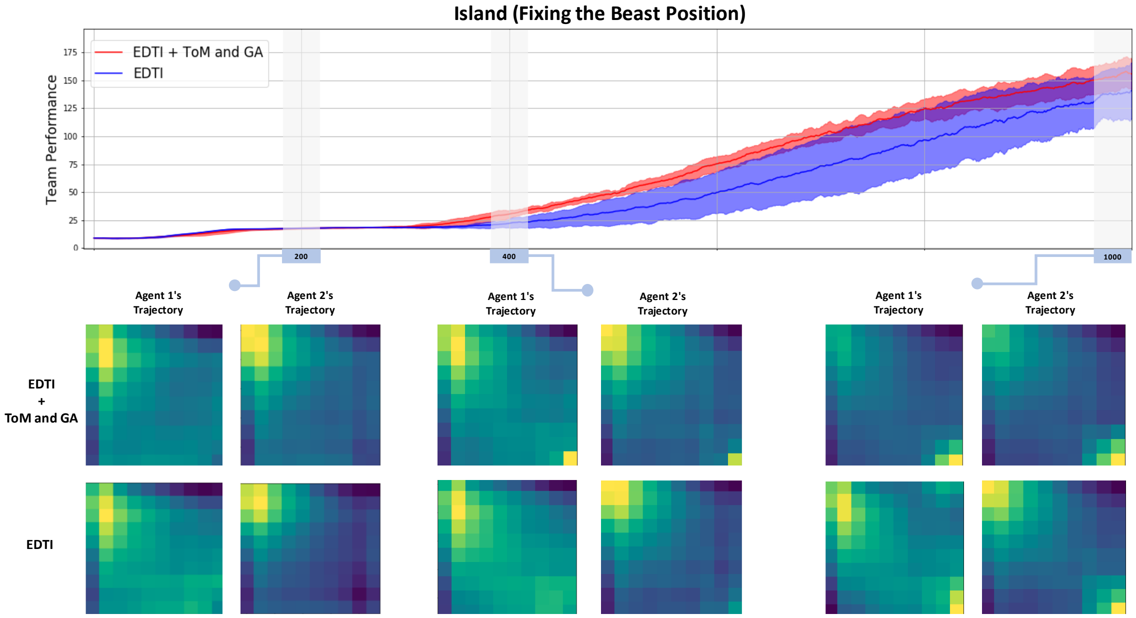

Figure 6: The behaviours of the EDTI agents and EDTI agents augmented with theory of mind and guilt aversion in the 10 × 10 Island environment with static beast. The upper part is Team Performance (y-axis) vs Number of Updates (x-axis). The lower part is the visitation count of the agents overtime. A cell with lighter colour means that it is visited more frequently by the agents. EDTI agents augmented with theory of mind and guilt aversion learns to cooperative attack the beast faster and tend to maintain equity in the team.

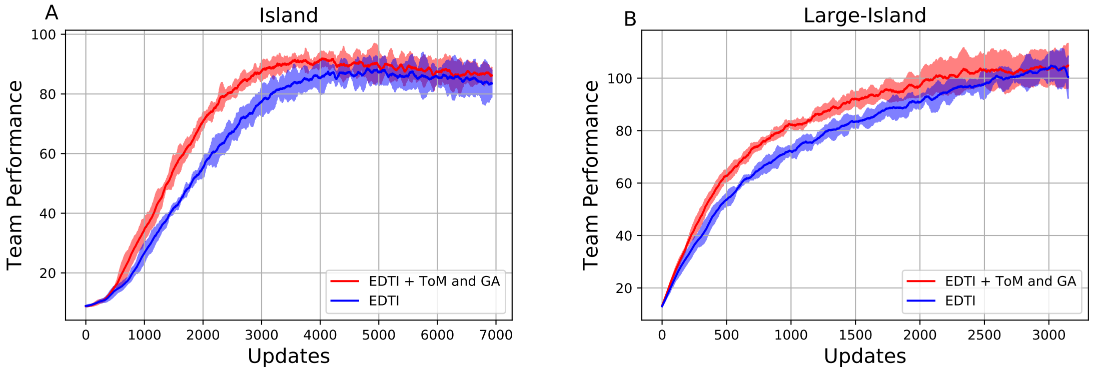

Figure 7: Team Performance (y-axis) vs Number of Updates (x-axis). The performance of EDTI agents and EDTI agents augmented with theory of mind and guilt aversion on complex environments, which are: (A) Island and (B) Large-Island.

the beast. In this island, there are treasures, which an agent only individually receives a reward of 10 if they collect. If the agents jointly attack the beast, they will kill the beast faster and reduce the chance of being killed. In the Large-Island, we aim to test our reward shaping mechanism with more than two agents. In this environment, the team with four agents will explore the 4 × 4 island.

We build our rewards shaping mechanism to the agent Exploration via Decision-Theoretic Influence with intrinsic rewards (EDTI) (Wang et al., 2020). In both environments, the labels of agents' policies will be given to agents at the end of each episode. An agent will be considered as following cooperative behaviour if this agent attacks the beast at least once during the episode. In the Large-Island, to handle the interaction of more than two agents, an agent i will have the first-order belief not only about what other believes about itself, but about what other believes about others. Therefore, the first-order beliefs of agents are implemented as described in (von der Osten et al., 2017). For example, the first-order belief of agent i is $b_i^{(1)} = \left\{ b_{j,k}^{(1)} | j \in \mathcal{N} \setminus \{i\}, k \in \mathcal{N} \right\}$, which is what agent i believes about agent j believes about agent k for all $j \in \mathcal{N} \setminus \{i\}, k \in \mathcal{N}$.

To illustrate the behaviours of the EDTI agent and EDTI agent augmented with theory of mind and guilt aversion, we fix the beast on the Island at the cell $(9^{\text{th}}, 9^{\text{th}})$ (on the right bottom corner of the environment) and the treasures at cells $(1^{\text{th}}, 1^{\text{th}})$ and $(2^{\text{th}}, 1^{\text{th}})$ (on the left top corner of the island). Figure 6 shown the team performance which is the average of team rewards and the visitation of agents (cells that have ligher colour are cells that visited more frequently by agents). In the earlier phase, both type of agents tend to visit the top left of the environment and collect the treasures. Over time, the agents start to discover the position of the beast and learn how to attack the beast. The EDTI agents augmented with theory of mind discovered the strategy of together attacking the beast and focus on visiting cells nearby the beast faster than the EDTI agents. It worth noting that the EDTI agent 1 finds the beast and attacks the beast before the EDTI agent 2, .i.e. the right bottom corner of the visitation map of the agent 1 is lighter than the visitation map of the agent 2, which shows the phenomena of inequity between two EDTI agents in this particular setting. In contrast, because the EDTI agents augmented with theory of mind and guilt aversion try to match the expectation of each other, they tend to preserve the equity. The team performance in the Island with moving beast is shown in Figure 7-A. Figure 7-B shown the efficient of augmenting EDTI with theory of mind and guilt aversion in the Large-Island. This demonstrates that our rewards shaping mechanism can be extended to the setting with more than two agents.

5. Related Works

Theory of Mind (Gopnik and Wellman, 1992; Premack and Woodruff, 1978; Gordon, 1986), or the ability of understanding that other having mental states, is crucial ability for an agent which involves in social interactions. In economics, it is studied as forecasting the forecast of other (Townsend, 1983). In multi-agent system, it is known as modelling others (Albrecht and Stone, 2018) to reduce the non-stationary problem while learning agents are updating their models simultaneously. Recent works in cognitive science and artificial intelligence have proposed several computational model of Theory of Mind such as the Bayesian Theory of Mind (Baker et al., 2011, 2017; Yoshida et al., 2008) and Machine Theory of Mind (Rabinowitz et al., 2018). In (De Weerd et al., 2013), authors used ARIMA(0,1,1) to model theory of mind as the recursive reasoning about other. The agent with Theory of Mind level-1 will hold a belief about what other agents think about its action. (von der Osten et al., 2017) extended this model to the settings in which there are more than two agents, which requires each agent holds belief about the believe of other not only about itself but also about others, i.e. thinking about other thinking about other. We used the Theory of Mind level-1 models described in (De Weerd et al., 2013; von der Osten et al., 2017) for belief about the higher level of actions to construct the belief based guilt aversion agent.

Solving social dilemma still is a challenge for reinforcement learning agents (Peysakhovich and Lerer, 2018b). Using behavioural game theory as prior knowledge, recent works demonstrated that inequity averse agents (Hughes et al., 2018) and prosocial agents (Peysakhovich and Lerer, 2018a) can promote cooperation in social dilemma. However, belief based guilt aversion (Battigalli and Dufwenberg, 2007), which is a well known mechanism in psychological game theory to promote cooperation, has not been studied in multi-agent reinforcement learning. (Moniz Pereira et al., 2017; Rosenstock and O'Connor, 2018) only considered the guilt without theory of mind model agents on evolutionary dynamics. We instead use the guilt aversion with theory of mind model to shape the reward of reinforcement learning agents.

6. Conclusion

We present a new emotion-driven multi-agent reinforcement learning framework in which reinforcement learning agents are equipped with theory of mind and guilt aversion - the emotion faculty that induces a utility loss in an agent if it believes that its action has caused harm in others. We studied the agent behaviours in Stag Hunt games, which simulate social dilemmas, whose Pareto optimal equilibrium demands cooperation between agents making it hard for pure reinforcement learning agents. We validated the framework in three environments for Stag Hunt games. Our results demonstrate the effectiveness of belief-based guilt aversion over other methods.

References

Stefano V Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018.

Robert Axelrod and William D Hamilton. The evolution of cooperation. *Science*, 211(4489): 1390–1396, 1981.

Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Annual Meeting of the Cognitive Science Society, 33 (33)*, 2011.

Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):1–10, 2017.

Pierpaolo Battigalli and Martin Dufwenberg. Guilt in games. *American Economic Review*, 97(2):170–176, 2007.

Luke J Chang, Alec Smith, Martin Dufwenberg, and Alan G Sanfey. Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron*, 70(3):560–572, 2011.

Harmen De Weerd, Rineke Verbrugge, and Bart Verheij. How much does it help to know what she knows you know? An agent-based simulation study. *Artificial Intelligence*, 199: 67–92, 2013.

Sam Devlin and Daniel Kudenko. Theoretical considerations of potential-based reward shaping for multi-agent systems. In *The 10th International Conference on Autonomous Agents and MultiAgent Systems-Volume 1*, pages 225–232. International Foundation for Autonomous Agents and Multiagent Systems, 2011.

Ernst Fehr and Klaus M Schmidt. A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868, 1999.

John Geanakoplos, David Pearce, and Ennio Stacchetti. Psychological games and sequential rationality. *Games and Economic Behavior*, 1(1):60–79, 1989.

Alison Gopnik and Henry M Wellman. Why the child's theory of mind really is a theory. *Mind & Language*, 7(1-2):145–171, 1992.

Robert M Gordon. Folk psychology as simulation. *Mind & language*, 1(2):158–171, 1986.

Jonathan Haidt. *The righteous mind: Why good people are divided by politics and religion.* Vintage, 2012.

John C Harsanyi, Reinhard Selten, et al. A general theory of equilibrium selection in games. *MIT Press Books*, 1, 1988.

Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. Inequity aversion improves cooperation in intertemporal social dilemmas. In *NIPS*, pages 3326–3336, 2018.

David Kraines and Vivian Kraines. The threshold of cooperation among adaptive agents: Pavlov and the stag hunt. In *Intelligent Agents III, Agent Theories, Architectures, and Languages, ECAI '96 Workshop (ATAL), Budapest, Hungary, August 12-13, 1996, Proceedings*, pages 219–231, 1996.

Michael L Littman. Markov games as a framework for MARL. In *Machine learning Proceedings*, pages 157–163. Elsevier, 1994.

Michael W Macy and Andreas Flache. Learning dynamics in social dilemmas. *PNAS*, 99 (suppl 3):7229–7236, 2002.

Luis Moniz Pereira, Tom Lenaerts, Luis A Martinez-Vaquero, et al. Social manifestation of guilt leads to stable cooperation in multi-agent systems. In *Proceedings of the 16th Conference on AAMAS*, pages 1422–1430. IFAAMAS, 2017.

Rajendra A Morey, Gregory McCarthy, Elizabeth S Selgrade, Srishti Seth, Jessica D Nasser, and Kevin S LaBar. Neural systems for guilt from actions affecting self versus others. *Neuroimage*, 60(1):683–692, 2012.

Tsuyoshi Nihonsugi, Aya Ihara, and Masahiko Haruno. Selective increase of intention-based economic decisions by noninvasive brain stimulation to the dorsolateral prefrontal cortex. *Journal of Neuroscience*, 35(8):3412–3419, 2015.

Alexander Peysakhovich and Adam Lerer. Prosocial learning agents solve generalized stag hunts better than selfish ones. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, pages 2043–2044. International Foundation for Autonomous Agents and Multiagent Systems, 2018a.

Alexander Peysakhovich and Adam Lerer. Towards ai that can solve social dilemmas. In *2018 AAAI Spring Symposium Series*, 2018b.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *ICML*, pages 4215–4224, 2018.

Sarita Rosenstock and Cailin O'Connor. When it is good to feel bad: An evolutionary model of guilt and apology. *Frontiers in Robotics and AI*, 5:9, 2018.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Yoav Shoham, Rob Powers, and Trond Grenager. If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 171(7):365–377, 2007.

Michael Shum, Max Kleiman-Weiner, Michael L Littman, and Joshua B Tenenbaum. Theory of minds: Understanding behavior in groups through inverse planning. *AAAI*, 2019.

Haruto Takagishi, Shinya Kameshima, Joanna Schug, Michiko Koizumi, and Toshio Yamagishi. Theory of mind enhances preference for fairness. *Journal of experimental child psychology*, 105(1-2):130–137, 2010.

Robert M Townsend. Forecasting the forecasts of others. *Journal of Political Economy*, 91 (4):546–588, 1983.

Friedrich Burkhard von der Osten, Michael Kirley, and Tim Miller. The minds of many: Opponent modeling in a stochastic game. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3845–3851, 2017.

Tonghan Wang, Jianhao Wang, Yi Wu, and Chongjie Zhang. Influence-based multi-agent exploration. In *8th International Conference on Learning Representations*, 2020.

Wako Yoshida, Ray J Dolan, and Karl J Friston. Game theory of mind. *PLoS Comput Biol*, 4(12):e1000254, 2008.