THEORY OF MIND WITH GUILT AVERSION FACILITATES COOPERATIVE REINFORCEMENT LEARNING

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

ABSTRACT

Guilt aversion induces experience of a utility loss in people if they believe they have disappointed others, which promotes cooperative behaviour in human. In psychological game theory, a branch of behavioural economics, guilt aversion necessitates modelling agents having 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 a social dilemmas (SDs) like Stag Hunt, we propose to use belief-based guilt aversion as a reward shaping mechanism. We show that our ToMAGA 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 individuals alone. In building artificial multi-agent systems to model such behaviours, it is important to construct such social inductive biases about the reasoning of other agents - also known as the Theory of Mind (ToM) (Rabinowitz *et al.*, 2018; Gopnik and Wellman, 1992). ToM enables individuals to cooperate and often results in optimal group rewards (Shum *et al.*, 2019; Takagishi *et al.*, 2010).

Maintaining *fair* outcomes for members of the group results in greater community good, 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 to guilt aversion (Nihonsugi *et al.*, 2015) and does not require ToM. 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 ToM. Initial work has examined integrating ToM and guilt aversion in a psychological game setting (Battigalli and Dufwenberg, 2007). The first work to examine social dilemmas in a deep RL setting is (Hughes *et al.*, 2018; Peysakhovich and Lerer, 2018) who incorporate knowledge through behavioural game theory when training the agents. However, guilt aversion, which plays a central role in moral decisions (Haidt, 2012) has not been considered.

This paper addresses the open challenges of integrating ToM and guilt aversion in 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 construction of ToM extends the work of (De Weerd *et al.*, 2013) to build agents with beliefs about cooperative behaviours rather than just primitive actions. Our RL agent uses a value function to make sequential decisions.

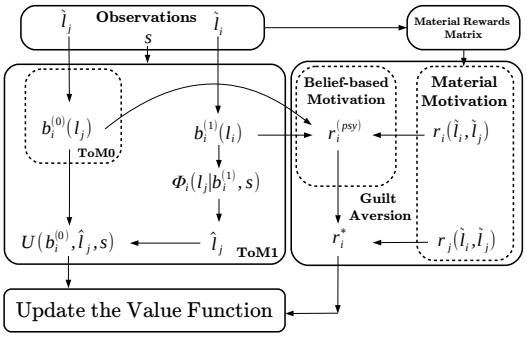

Figure 1: The learning process in ToMAGA.

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.

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 hunts stag. Therefore, both hunting hare is the *risk-dominant* equilibrium. 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. 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 environments in a one-step decision game and in a multi-step grid-world. Our experiments demonstrates that modelling guilt with explicit ToM helps RL agents to cooperate better than those without ToM, encouraging faster learning towards cooperative behaviours.

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

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

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

2 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. 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 agent j's belief*. We construct ToM1 agents as in (De Weerd *et al.*, 2013) (detailed in Supplementary Section A). The guilt averse agent i will experience a utility loss if it thinks it lets the other agent down, which 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, which is $\phi_j = \sum_{l_i, l_j \in \{C,U\}} b_i^{(0)}(l_j) \times b_i^{(1)}(l_i) \times r_j^{(T)}(l_i, l_j)$ where

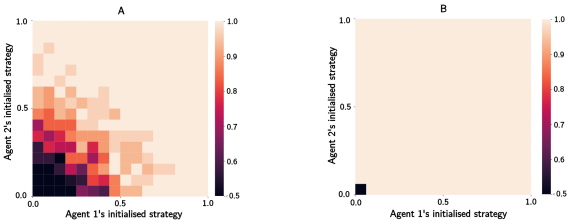

Figure 2: Initial probability of second player following cooperative strategy (y-axis) vs Initial probability of first player following cooperative strategy (x-axis). The colour shows the probability (lighter values indicate higher probability) of first player following cooperative strategy after 500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

$r_j^{(T)}(l_i, l_j)$ 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):

$$r_i^{(psy)}(\tilde{l}_i, \tilde{l}_j) = -\theta_{ij} \max\left(0, \phi_j - r_j^{(T)}(\tilde{l}_i, \tilde{l}_j)\right) \tag{1}$$

where guilt sensitivity $\theta_{ij} > 0$. The reward is then shaped as $r_i^* = r_i^{(T)}(\tilde{l}_i, \tilde{l}_j) + r_i^{(psy)}(\tilde{l}_i, \tilde{l}_j)$. This computation is based on an assumption that a guilt averse agent *does not* know whether the other is guilt averse. Given the shaped reward, the RL agent learns by updating its value function based on temporal difference algorithm TD(1) on the matrix-form Stag Hunt game. 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).

We establish two 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.* The proof is provided in Supplementary Section B.

3 Experiments

3.1 Matrix-Form Stag Hunt Games

In this experiment we aim to answer the question: *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 shown that ToMAGAs promote cooperation better than the GA agents without ToM. 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).

3.2 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. 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. We are interested in two situations: At the beginning of the training process, agents

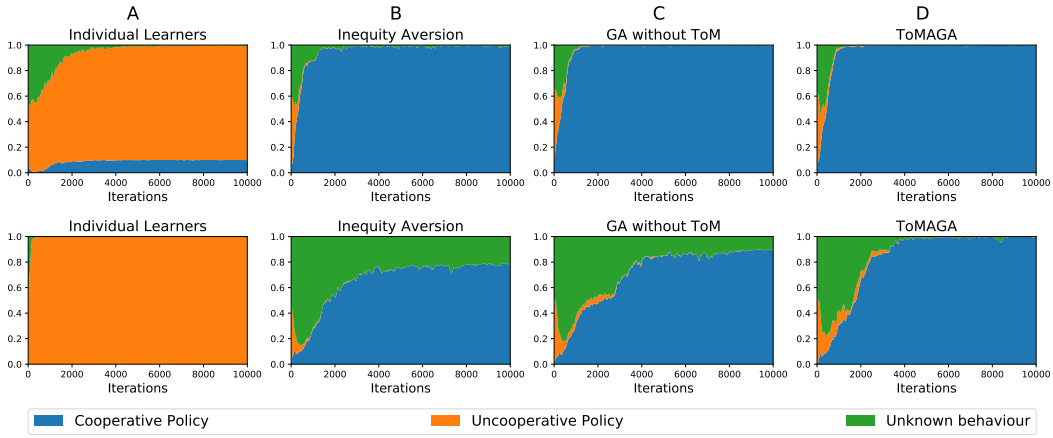

Figure 3: Policies of individual learners (column A), agents with inequity aversion (column B), GA agents without ToM (column C), and ToMAGAs (column D) when they start nearby the stag (the first row) and nearby hares (the second row). Proportion of following cooperative (blue), uncooperative (orange), unknown (green) behaviours (y-axis) vs Iterations.

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 stag if they are put nearby stag at the beginning. After each iteration, each policy will be labelled as follows: (1) when both hunt the stags, 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 RL 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. IA agents are agents that have a shaping reward $r_i^{(psy)} = -\frac{\theta_{ad}}{N-1} \times \sum_{j \neq i} \max(r_i - r_j, 0) - \frac{\theta_{dis\_ad}}{N-1} \times \sum_{j \neq i} \max(r_j - r_i, 0)$, where N is the number of agents, θ_{ad} and θ_{dis_ad} are advantageous and disadvantageous sensitivity, respectively (Fehr and Schmidt, 1999).

Fig. 3 shows the policies of deep RL agents over the training process when they start nearby the stag (the first row of Fig. 3) and nearby the hares (the second row of Fig. 3). In both cases, the individual learners i.e. deep RL 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 RL agents with social preferences can learn to cooperate in both cases. When the agents are put nearby stag at the beginning, the performance of IA agents is comparable to GA agents without ToM and the ToMAGA (the first row of columns B, C, and D of Fig. 3). However, when initialised nearby hares, GA agents without ToM and ToMAGA learn to cooperate faster than the IA agents (the second row of columns B, C, and D of Fig. 3). 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 Conclusion

We present a new emotion-driven multi-agent reinforcement learning framework in which reinforcement learning agents are not only equipped with theory of mind, but also guilt aversion. 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 two environments for Stag Hunt games. Our results demonstrate the effectiveness of the method over methods which are not belief-based guilt aversion.

REFERENCES

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

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.

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.

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

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.

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, 2018.

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.

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.

SUPPLEMENTARY

A THE FIRST-ORDER THEORY OF MIND AGENT

The ToM1 agents is constructed as in (De Weerd *et al.*, 2013). It worth noting that it essentially is a special case of the autoregressive integrated moving average (ARIMA$(0, 1, 1)$) model. After interacting with the environment, 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

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

where $\Phi_{ij}(l_j)$ 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 $U(l_j)$. To do this, agent i maintains and updates a confidence $c_{ij} \in [0, 1]$ about its ToM1 as follows:

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

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

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

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

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

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

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

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

B PROOF OF THE OBSERVATIONS

Recall that we establish the following 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:

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

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 section. (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 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. □