The Minds of Many: Opponent Modelling in a Stochastic Game

Friedrich Burkhard von der Osten and **Michael Kirley** and **Tim Miller**
School of Computing and Information Systems
The University of Melbourne, Australia
fvon@student.unimelb.edu.au, {m.kirley,tmiller}@unimelb.edu.au

Abstract

The *Theory of Mind* provides a framework for an agent to predict the actions of adversaries by building an abstract model of their strategies using recursive nested beliefs. In this paper, we extend a recently introduced technique for opponent modelling based on Theory of Mind reasoning. Our extended *multi-agent Theory of Mind* model explicitly considers multiple opponents simultaneously. We introduce a *stereotyping* mechanism, which segments the agent population into sub-groups of agents with similar behaviour. Sub-group profiles guide decision making. We evaluate our model using a multi-player stochastic game, which presents agents with the challenge of unknown adversaries in a partially-observable environment. Simulation results demonstrate that the model performs well under uncertainty and that stereotyping allows larger groups of agents to be modelled robustly. The findings show that Theory of Mind modelling is useful in many artificial intelligence applications.

1 Introduction

Opponent modelling is a technique to recognize the strategy of an opponent and make predictions about their behaviour. In AI, opponent modelling is used for decision making by exploiting sub-optimal opponents [Ganzfried and Sandholm, 2011].

One technique that can be used in opponent modeling is *Theory of Mind* (ToM) [Goldman, 2012]. ToM is a concept from psychology that describes the higher cognitive mechanism of attributing unobservable mental content, such as beliefs, desires and intentions to others. ToM is used by humans to form models for predicting the behaviour, decisions and actions of other individuals. ToM is a bounded rationality model [Brennan and Lo, 2012], and as such is suitable to simulate human decision making in agents. Typically, this is a done recursively using nested beliefs – a process referred to as *higher-order* ToM. That is, an agent A can not only form beliefs about what another agent B would do, but also about what the other agent B would believe about agent A or what the other agent B believes the agent A to know about them, etc.

In this paper, we propose a multi-agent ToM model (MToM) that extends an existing computational ToM model. In contrast to De Weerd *et al.,* [2013b], who model one adversary, our model allows multiple opponents to be considered at the same time. Thus, agents have to model other agents' views of them and their views of other agents. Humans have limited ability to model multiple opponents, so they resort to stereotyping [Fiske, 2000]. We introduce a mechanism to model stereotyping that segments opponents into categories, which are then treated as single opponents in MToM.

We validate our MToM model using a stochastic multi-player game. ToM models are particularly useful in stochastic games due to the fact that recognizing the strategies of opponents gives a player a significant advantage. The stochastic game used in the experiments is the Common Pool Resource Game (CPRG) [Sethi and Somanathan, 1996; von der Osten *et al.*, 2017], a large-scale social dilemma. This game is particularly suitable as a demonstration for MToM, as in reality individual behaviour is dependent on beliefs about the behaviour of other individuals [Cavalcanti *et al.*, 2010]. Outcomes of the game depend on the cumulative actions of all agents, hence it is critical to predict other agents' actions. Furthermore, the CPRG is a difficult game in the context of AI, as it presents agents with multiple challenges: non-deterministic game dynamics, partial observability and unknown adversaries.

We show the MToM model can be used to predict agents' actions and demonstrate the model's behaviour under uncertainty by detailing the impact of perfect and imperfect knowledge. Furthermore we show the effect of using different orders of ToM. We establish that MToM agents are able to estimate the the sophistication of their opponents' strategy, especially in non-homogeneous populations.

2 Related Work

Research in psychology has dealt with the topic of ToM quite extensively [Perner, 1999; Goldman, 2012]. Psychologists investigate ToM as a tool of human cognition that equips them with an evolutionary advantage. As an integrated part of a more general model of bounded rationality it helps explain human behavior accurately [Robalino and Robson, 2012; Brennan and Lo, 2012]. ToM is often viewed from an evolutionary perspective [Kimbrough *et al.*, 2014] and it was found that natural selection favours intention recognition, which enables the evolution of cooperation [Pereira *et al.*, 2012].

Such concepts have been formalised in computational models, such as in interactive POMDPs [Gmytrasiewicz and Doshi, 2005], in which agents explicitly contains a nested model of other agents in the environment. These have been used to successfully model human-agent settings, such as in PsychSim [Pynadath and Marsella, 2005].

De Weerd *et al.* [2015] apply ToM to the tacit communication game to support the hypothesis that ToM facilitates cooperative behaviour. This application of ToM shows its potential in the field of AI. Other approaches [Yoshida *et al.*, 2008; Pereira *et al.*, 2011] keep a model to learn beliefs about opponents' actions and a separate model to infer the complexity of opponent strategies, whereas De Weerd *et al.* [2013b] integrate beliefs about opponent behaviour with inference about the sophistication of their strategies (see integration of different orders of ToM in Section 3). Furthermore, a game theoretic advantage of ToM is that it allows agents to infer the opponents degree of sophistication [Yoshida *et al.*, 2008].

Despite the Vygotskian intelligence hypothesis suggesting that higher-order ToM allows individuals to cooperate more effectively [Moll and Tomasello, 2007], higher-order ToM has an advantage over lower-order ToM, but becomes ineffective beyond the second order [De Weerd *et al.*, 2013b; 2013a], *i.e.,* person A estimates what person B estimates person A estimates person B will do [zero order: person A estimates what person B will do; first order: person A estimates what person B estimates person A will do]. The reason for this might lie in the biological cost in humans (brain power) on one hand, and the mathematical limitations of the model (over-fitting when dealing with lower orders of ToM) on the other hand. However, the limited sophistication of ToM in humans is not necessarily a consequence of biological costs, but can be explained by cooperation favouring lower orders of ToM [Devaine *et al.*, 2014].

We are not the first to consider ideas of segmentation. Albrecht *et al.* [2016] use agent *types* to generalise agents with similar behaviour. The difference between our stereotyping and that of Albrecht *et al.* is that we create a behavioural profile of a type, rather than assigning a probability that an agent is of that particular type. Felli *et al.*. [2015] propose agent models, and define 'stereotypical reasoning' about individuals or groups of agents. Similar to Albrecht *et al.* though, they do not aim to learn the segments, but merely to prescribe them manually.

3 Model

3.1 Multi-ToM

The model used in this paper is an extension of the De Weerd *et al.,* [2013b] model, conceptualized for one player and one opponent. In principle this model can be used for any order of ToM, but for the sake of simplicity, we limit the description to the first two orders.

A zero-order theory of mind (ToM$_0$) agent i starts out with a set of beliefs b about the other agent j, which is a set of probabilities that indicate the likelihood that the opposing agent j takes a certain action $a_j \in \mathcal{A}_j$ in a particular state $s \in \mathcal{S}$ of the game:

$$b^{(0)}(a_j; s) \geq 0 \quad \forall a_j \in \mathcal{A}_j \\ \text{with} \sum_{a_j \in \mathcal{A}_j} b^{(0)}(a_j; s) = 1 \quad \forall s \in \mathcal{S} \tag{1}$$

where $\mathcal{A}_j$ is the set of actions performed by agent j and $\mathcal{S}$ is the set of possible game states (see Section 4.1).

In the multi-agent extension of De Weerd's ToM model, an agent i has to keep zero-order beliefs not just about one opponent, but about multiple:

$$b^{(0)} = \left(b_1^{(0)}, b_2^{(0)}, \ldots, b_{n-1}^{(0)}\right)^T \tag{2}$$

in which n is the total number of agents participating in the game and $b_j^{(0)}$ is the belief of agent i about agent j. Using this set of beliefs, agent i can assign a value Φ to playing a certain action itself given the likelihood of the opposing agents j playing a particular action, based on the pay-off π_i agent i will get. This is done by summing over all possible permutations of actions of all other agents:

$$\Phi_i(a_i; b^{(0)}, s) = \\ \sum_{\sigma_k \in \mathcal{K}_{n-1}} \left(\left(\prod_{a_j \in \sigma_k} b^{(0)}(a_j; s) \right) \cdot \pi_i(s, (a_i, \sigma_k) \right) \tag{3}$$

with permutation σ_k (also called permutation with repetition/of a multi-set or tuple) consisting of actions $a_k \in \mathcal{A}$ out of the set of all permutations $\mathcal{K}_{n-1}$. The agent will then choose the action a_i that maximizes this value Φ with a decision function t_i^*:

$$t_i^*(b^{(0)}; s) = \arg\max_{a_i \in \mathcal{A}_i} \Phi_i(a_i; b^{(0)}, s) \tag{4}$$

A first-order theory of mind (ToM$_1$) agent keeps its zero-order beliefs $b^{(0)}$, but also has first-order beliefs $b^{(1)}$; that is, a set of probabilities that describe agent i's estimate of an agent j's zero-order beliefs. The beliefs $b^{(1)}$ thus describe what agent i believes agent j believes about agent k (where k can also be agent i). The first-order beliefs do not only include the zero-order beliefs of other agents j about agent i, but also about each other:

$$b^{(1)} = \begin{pmatrix} b_{1,i}^{(1)} & b_{1,2}^{(1)} & \cdots & b_{1,n-1}^{(1)} \\ b_{2,i}^{(1)} & b_{2,1}^{(1)} & \cdots & b_{2,n-1}^{(1)} \\ \vdots & \vdots & \vdots & \vdots \\ b_{n-1,i}^{(1)} & b_{n-1,1}^{(1)} & \cdots & b_{n-1,n-2}^{(1)} \end{pmatrix} \tag{5}$$

where $b_{j,k}^{(1)}$ represents the belief agent i has that agent j believes about k. Given this set of first-order beliefs, agent i can make a prediction $\hat{a}_j^{(1)}$ of agent j's action by mimicking agent j's decision process with its decision function t^* given its beliefs $b^{(1)}$ about agent j's zero-order beliefs:

$$\hat{a}_j^{(1)} = t_j^*(b^{(1)}; s) = \arg\max_{a_j \in \mathcal{A}_j} \Phi_j(a_j; b^{(1)}, s) \tag{6}$$

The agent i now has a prediction $\hat{a}_j^{(1)}$ of each of its opponents actions based on its first-order beliefs and its own zero-order belief $b^{(0)}$ indicating the probability of the opponent playing a certain action. These two are integrated to form a combined estimate of opponent behaviour. An agent i has a certain confidence c_1^j – the extent to which ToM$_1$ governs their decisions – in its first-order beliefs $b^{(1)}$. An agent has a different confidence for each first-order belief about agent j, and uses that confidence to weight the influence of its first-order beliefs when integrating zero-order beliefs and first-order prediction. The integration function U produces an integrated zero-order belief that weighs in the first-order prediction:

$$U(b^{(0)}, \hat{a}_j^{(1)}, c_1)(a_j; s) = \begin{cases} (1 - c_1^j) \cdot b^{(0)}(a_j; s) & \text{if } a_j \neq \hat{a}_j^{(1)} \\ (1 - c_1^j) \cdot b^{(0)}(a_j; s) + c_1^j & \text{if } a_j = \hat{a}_j^{(1)} \end{cases} \tag{7}$$

Agent i can now use this integrated belief to make a decision with its decision function t^*:

$$t_i^*(U(b^{(0)}, \hat{a}_j^{(1)}, c_1); s) = t_i^*(U(b^{(0)}, t_j^*(b^{(1)}; s), c_1); s) \tag{8}$$

Given the actual action $\tilde{a}_j$ of opponent j and agent i's own actual action $\tilde{a}_i$, the confidence c_1^j is updated with

$$c_1^j := \begin{cases} (1 - \lambda) \cdot c_1^j & \text{if } \tilde{a}_j \neq \hat{a}_j^{(1)} \\ (1 - \lambda) \cdot c_1^j + \lambda & \text{if } \tilde{a}_j = \hat{a}_j^{(1)} \end{cases} \tag{9}$$

with nested beliefs being updated as follows:

$$b^{(d)}(a_j; s) := U(b^{(d)}, \tilde{a}_j, \lambda)(a_j; s) \quad \forall a_j \in \mathcal{A}_j \tag{10}$$

where λ is the learning speed of an agent. Belief integration not only includes updating first-order beliefs about agent i but also about all other agents, using their corresponding observed actions $\tilde{a}$, i.e., agent i observing agent j's action does not only prompt an update of $b_j^{(0)}$ but also $b_{1,j}^{(1)}$, $b_{2,j}^{(1)}$, etc. Note that for the experiments in this paper we limit the order of ToM to first-order beliefs. For a visualization of how MToM is applied, see Figure 2.

In the experiments in Section 4, agents either have complete knowledge (CI) of the game dynamics, or the pay-off π is not known to agents, but approximated by means of Q-Learning [Watkins and Dayan, 1992]. Agents learn their expected pay-off given other agents' contributions and their own decision in a given game state. The resulting Q-values are a qualitative representation of the pay-off magnitude and can serve as a pay-off estimation in place of the known pay-off function.

Correspondingly, the action selection mechanism consists of either choosing the action with the best value (t^*, see Equation 4) when the pay-off is known, or Boltzmann/Softmax selection when the pay-off is unknown; i.e., actions are selected with a probability proportional to their value:

$$Pr(a_i \mid b^{(0)}, s) = \frac{e^{\Phi(a_i; b^{(0)}, s)}}{\sum\limits_{k \in \mathcal{A}} e^{\Phi(a_k; b^{(0)}, s)}} \tag{11}$$

3.2 Segmentation

With an increasing number of agents, not only does the computational complexity of a simulation rise exponentially (see Figure 1), the model also becomes increasingly unrealistic in modelling human behaviour. When confronted with too much information to process at once, humans resort to *stereotyping* to generalize details and still make reasonably accurate predictions [Oakes *et al.*, 1994; Fiske, 2000; Kashima *et al.*, 2000]. Stereotyping is the generalization of individual behaviour to a group to simplify decisions and evaluations. For this reason we introduce a mechanism to segment agents into groups according to their observed history of actions. There are several ways to approach this, and to demonstrate MToM we have chosen a straightforward approach [Camerer *et al.*, 2015; Kashima *et al.*, 2000]: an agent j's action $a_{j,r}$ in round r, averaged over the last Δ_t rounds of the game, is assumed to be representative of its behaviour:

$$\bar{a}_j = \sum_{r=t-\Delta_t}^{t} a_{j,r} / \Delta_t \tag{12}$$

Given a predefined number of categories $\mathcal{C} \in \mathbb{N}^*$, an agent falls into category v when

$$\frac{v-1}{\mathcal{C}} \|\mathcal{A}\| \leq \bar{a} < \frac{v}{\mathcal{C}} \|\mathcal{A}\| \tag{13}$$

When agents are segmented, the corresponding groups are treated as single agents, thus reducing the computational cost and simulating stereotyping as described above. This process is repeated periodically throughout the game to be able to assess and reassess opponent behaviour accurately.

3.3 Complexity Analysis

Theorem 1. *The space complexity of MToM is $\mathcal{O}(n^2)$ with $\mathcal{O} = \Theta = \Omega$.*

Proof. There are two data structures used by the model: the zero-order beliefs $b^{(0)}$ about each agent, thus of size n, and the first-order beliefs $b^{(1)}$ about each agent and their beliefs about each other (bar their beliefs about themselves), thus of size $n - 1 \times n - 2$ (see Equations 2 & 5 respectively). Hence, the total size of the data structures used is $n \times n = n^2$. Furthermore, the data structures are static, making the worst case space complexity the same as the best and average cases. ■

Theorem 2. *The time complexity of MToM is $\mathcal{O}(nm^n)$ with $m = \|\mathcal{A}\|$.*

Proof. For zero-order MToM, the expected pay-offs using zero-order beliefs $b^{(0)}$ about other agents are calculated by summing over all permutations of actions a of all agents n multiplied. The number of permutations is then $\|\mathcal{A}\|^n$ (see Equation 3). For first-order MToM, predictions for each other agent's action are made by mimicking their zero-order decision process given $b^{(1)}$, thus repeating zero-order MToM $n - 1$ times: $n - 1\|\mathcal{A}\|^n$ (see Equation 6). The integration of first and zero-order beliefs takes one iteration of the belief table $b^{(0)}$, i.e., n operations (see Equation 7), and the update process of both belief tables takes n^2 and n operations

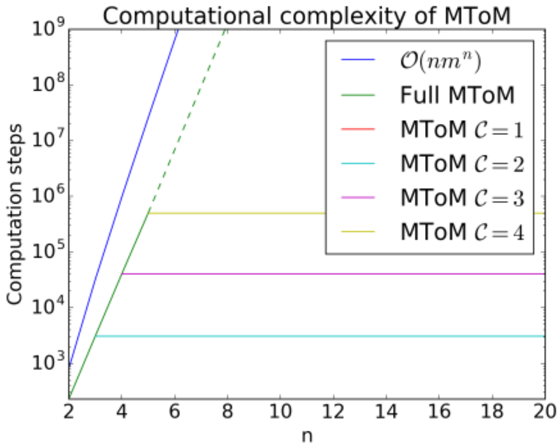

Figure 1: The number of computation steps vs the number of agents n. The plot shows the calculated worst case values for the MToM model and MToM with segmentation (increasing numbers of categories). The results indicate that segmentation reduces the computational complexity to a constant.

respectively (see Equation 9 & 10). The overall action selection uses the integrated beliefs, taking an additional $n-1$ times: $n-1\|\mathcal{A}\|^n$ operations (see Equation 8), resulting in $2(n-1)\|\mathcal{A}\|^n + n^2 + 2n$ operations for an agent decision. Figure 1 presents the computation steps measured in the experiments. For segmentation the time complexity is reduced to: $\mathcal{O}(\mathcal{C}m^{\mathcal{C}})$. For d orders of ToM, the time complexity is $\mathcal{O}(mn^n)$ to the d-th power of n. ■

4 Experimental Design

4.1 Game Model

Stochastic games are Markov Decision Processes (MDP) in which state transitions and pay-offs depend on the behaviour of other agents [Condon, 1992]. A stochastic game is typically defined by the quadruple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$ with a finite set of players [Shapley, 1953; Mertens and Neyman, 1981]. As a case study, we use an instance of a stochastic game – the Common Pool Resource Game (CPRG). Figure 2 illustrates the thought processes of agents using MToM in a representative CPRG (an abstract fisheries game).

In the CPRG, a group of individuals harvest a resource (the state space $\mathcal{S}$) and each individual decides their effort x at a cost c to invest in harvesting (the action space $\mathcal{A}$). The combined effort of the group X, along with the state of the resource N, determines their pay-off π (the reward $\mathcal{R}$). The state of the resource is in turn dependent on the harvest H, as well as its growth G (described by the transition probabilities between states $\mathcal{P}$). The goal of the population as a whole is to harvest the resource sustainably, whereas the individual goal is to maximize the pay-off from harvesting. As the population grows larger, it becomes increasingly difficult to harvest both sustainably and profitably. The CPRG is formulated as a stochastic game as follows. The state space $\mathcal{S}$ describes the level of the resource N:

$$\mathcal{S} : s = \frac{N}{N_{max}}\|\mathcal{S}\| \text{ with } N = [0, 1000] \in \mathbb{R}$$

The action space $\mathcal{A}$ describes the actions available to an individual agent. In the CPRG, an action is a level of investment x into harvesting the resource. The range of actions depends on the number of agents n participating in the game:

$$\mathcal{A} : a = \frac{x - x_{min}}{x_{max} - x_{min}}\|\mathcal{A}\| \text{ with } x = [\frac{X_{min}}{n}, \frac{X_{max}}{n}]$$

Note that both state and action space are discretized into finite sets. The transition between states is represented as probabilities $\mathcal{P}$ dependent on the current state and action. In the CPRG, these transitions depend on the actions of all agents, and are governed by coupled differential equations describing resource dynamics, particularly the harvest H and growth G:

$$\begin{aligned} \mathcal{P}_a(s, s') = \;& Pr(s_{t+1} = s' | s_t = s, a_t = a) : \mathcal{S} \times \mathcal{A} \\ \text{with} \;& N_t = N_{t-1} - H(X_{t-1}, N_{t-1}) + G(N_{t-1}) \\ \text{and} \;& X = \sum_{i=1}^{n} x_i \end{aligned}$$

Finally, agents receive a reward determined by the reward function $\mathcal{R}$, depending on the invested effort X of all agents and the level of the resource N:

$$\mathcal{R}_a(s, s') = 0.5s' + 0.5\pi(s, a) \text{ with } \pi = \frac{x}{X}H(N, X) - cx$$

with the pay-off π being the agents proportional fraction of harvest according to its invested effort x with resource level N. The cumulative pay-off $A = \sum_{t=0}^{t_{max}} \pi$ and the resource level N are used as performance measures for the CPRG.

The harvest returns and resource growth are determined by a set of coupled differential equations and not usually known to the participants of the game. A detailed description of the game dynamics can be found in von der Osten *et al.* [2017]. The uncertainty in estimating pay-offs and the dynamic resource behaviour makes this an interesting application of opponent modeling. Furthermore there are usually multiple participants in this type of game, resulting in decision making becoming more complex as populations increase.

4.2 Setup

In the first experiment, we demonstrate the efficacy of MToM as an opponent modelling mechanism, under the assumptions of either complete information or uncertainty. The population is fixed size ($n = 5$, independent variable). Different zero ($MToM_0$) and first-order ($MToM_1$) agents are compared with an optimal baseline (OPT) — which knows the pay-offs and actions of all agents — and a population of random agents (RND). In the second experiment, we examine the scalability and relative robustness of the segmentation process. The population ranged from $n = 10$ to $n = 50$ with increments of 10, and the number of categories ranging from $\mathcal{C} = 1$ to $\mathcal{C} = 4$. The third set of experiments is used to determine the success of stereotyping facilitated by the inference of opponent sophistication in MToM given non-homogeneous populations. The population size is fixed ($n = 10$) using $\mathcal{C} = 1$ category for segmentation. Populations

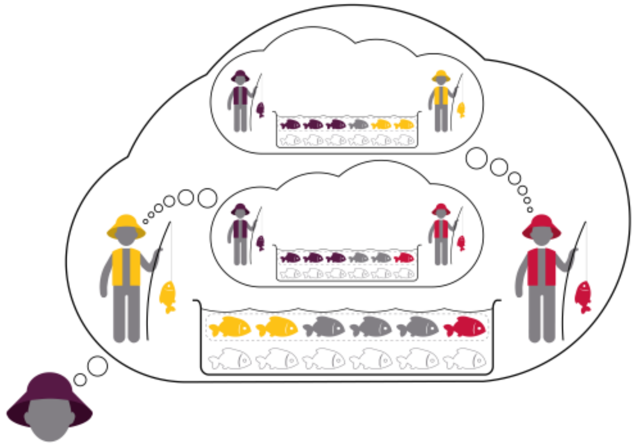

Figure 2: The thought processes of agents using MToM in a CPRG (fisheries example). The fish tank depicts the resource N (here $N = 12$). The dashed box denotes the fish that are replenished after the harvest (determined by the growth function G) as part of the transition probabilities. Agents are thinking about the amount of resource to extract (number of fish they attempt to catch). Agents also consider the thought process of their opponents. The actual harvest return (fish caught) is determined by the production function H.

Parameter	Interpretation	Value
λ	Learning parameter	0.5
Δ_t	Timescale for assessing agents	10
X_{min}	Minimum effort	100
X_{max}	Maximum effort	500
$\|\mathcal{A}\|$	# of actions available to agents	10
$\|\mathcal{S}\|$	# of game states agents recognize	5
n	# of agents in the game	[2,10]
$\mathcal{C}$	# of categories for segmentation	[1,4]
h	% of RND agents in the population	[20,50]

Table 1: Parameters used in the experiments

are made up of $MToM_0$/$MToM_1$ agents and 20/30/40/50% RND agents (independent variable h). Each simulation is run for 5000 rounds, with each setting repeated 50 times. We ran one-way ANOVA with subsequent post-hoc Tukey HSD tests. Table 1 shows the parameters and variables used in the experiments.

Agents use MToM as their strategy to play the stochastic game (see Figure 2). The success of the agents is defined by the outcome of the game. Therefore, the measures taken are: (1) the level of the resource N that describes the sustainability of the system; and (2) the average agent assets $\bar{A}$, which is symptomatic of an agent's goal of profit maximization. Ideally, the resource N stays on a stable positive level (see the baseline optimal agent type), whereas the assets A would grow. It is difficult for agents to harvest the resource in a stable and profitable manner (only 7% of the action space $\mathcal{A}$ are profitable and sustainable [von der Osten *et al.*, 2017]), even more so with increasing population, as the impact of an individual action on the overall outcome decreases.

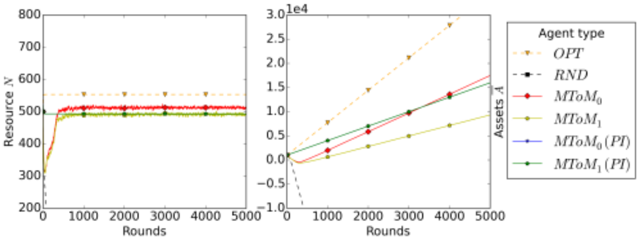

Figure 3: Full MToM: each agent models each opponent, $n = 5$. Random (RND) and optimal (OPT) agents are shown for reference.

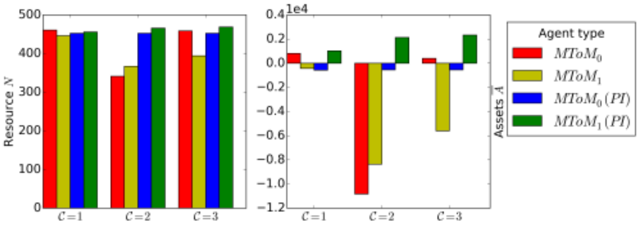

Figure 4: MToM models using segmentation with different granularity, $n = 10$.

5 Results and Discussion

5.1 MToM

The first experiment demonstrates the feasibility and performance of MToM in a stochastic game. Figure 3 shows the development of the resource N and the average agent assets $\bar{A}$ in a time series over 5000 rounds. Agents with complete knowledge of the pay-offs (CI) satisfy both performance measures. Note results for differing levels of $MToM(CI)$ are exactly the same, hence they are not distinguishable on the graph. When presented with uncertainty, their pay-offs decrease slightly, and a difference in the results for alternative orders of ToM becomes visible: $MToM_0$ performs better than $MToM_1$. $MToM_0$ surpasses both models with complete information (possibly due to overestimating adversary behaviour).

5.2 MToM with Segmentation

The second experiments confirms the appropriateness of segmentation as a means of scaling the MToM model for increasing numbers of agents. Figure 4 shows a population of 10 agents playing the CPRG using segmentation with different granularity. Results of these experiments unveil several insights: Agents with perfect information generally perform better than agents without. Furthermore, even with complete information, only the $MToM_1$ agents manage to increase their assets, even though all agent types can maintain the resource. This is contrary to the results of the first experiment, where no differences between the $MToM_0$ and $MToM_1$ levels were apparent for agents with complete information.

More interesting, however, is the development with increasing $\mathcal{C}$. For $MToM_1$ with complete information, more categories are beneficial, whereas $MToM_0$ and $MToM_1$ are better off with treating the entire population as one opponent.

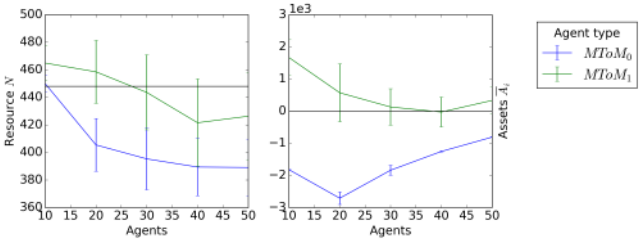

Figure 5: MToM with increasing population size, $\mathcal{C} = 2$.

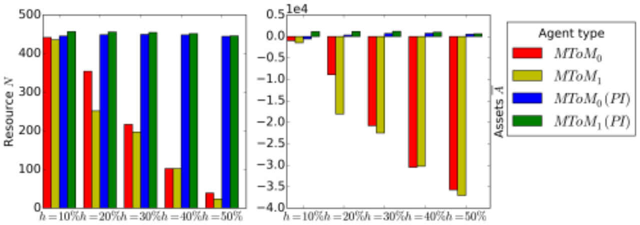

Figure 6: MToM models in non-homogeneous populations with proportions of random adversaries, $n = 10, \mathcal{C} = 1$.

Without complete information, $MToM_0$ performed better than $MToM_1$, indicating that the gains from higher-order ToM depends on how much information is available. Moreover, the results illustrate that either one category, or an increasing granularity of categories promises better results. Note that the drop in assets for two categories is related to the threshold of the categories lying between the observed behaviour of agents, making their assessment inaccurate.

Figure 5 show an increasing population of agents using a fixed segmentation ($\mathcal{C} = 2$) under complete information. This demonstrates the ability for segmentation to scale to larger populations. While the agents are able to maintain the resource, their profitability decreases initially. With larger population sizes, first-order MToM agents are able to make profit. This behaviour is attributed to resource dynamics.

5.3 MToM in Non-Homogeneous Populations

In the final set of experiments, the performance of agents is tested against unknown adversaries. Figure 6 presents the results of simulations with 10 agents, a certain fraction h of which are random (RND) agents. The results indicate that in a non-homogeneous scenario, complete information leads to pronounced differences. As expected, populations perform worse when RND agents are present, however, the effect is more distinct when facing uncertainty in pay-offs. Nonetheless, $MToM_0$ and $MToM_1$ agents scale linearly with the number of unknown opponents. The difference between them becomes less distinct when the population is more well-mixed, otherwise $MToM_0$ performs better. This result is consistent with the previous experiment. It should be noted that RND agents perform badly when compared to the other models (see Figure 3). This means that having even a small fraction of RND agents in the game makes the profitability metric almost impossible to achieve.

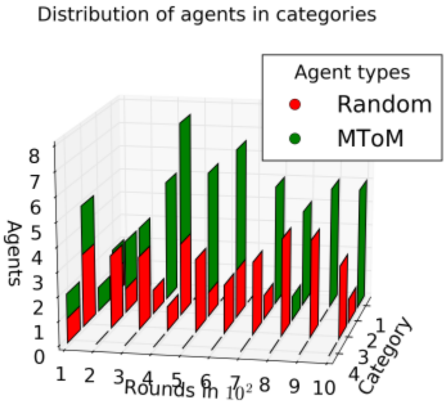

Figure 7: A sample view of the segmented population using the $MToM_1$ model with $\mathcal{C} = 4$. Here, $n = 10$ including 4 random (RND) agents. The game is run for 1000 rounds and categories are reassessed every 100 rounds.

Figure 7 provides important insights into the segmentation process. The plot shows the distribution of agents in categories after every assessment. As RND agents select actions uniformly, they would usually fall in the middle of the action space. MToM agents tend to take restrained actions. While the groups are mixed in the beginning, agents are more distinctly segmented into RND (mostly categories 2/3) and MToM (mostly category 1) opponents, demonstrating the integrated approach of opponent modelling and strategy sophistication works well with MToM.

6 Conclusion

This paper has examined an interesting application of ToM in groups using a stochastic game as a case study. The results demonstrate that MToM is a useful approach to use when modelling multiple opponents. MToM performs robustly under uncertainty and it scales with increasing numbers of agents provided it is combined with segmentation into groups based on stereotyping. When agents have complete information about game pay-offs, $MToM_1$ outperforms $MToM_0$. In contrast, when there is uncertainty in game pay-offs lower $MToM_0$ performs better. This suggests that higher-order ToM only gains an advantage over lower-order ToM, or other simpler strategies, if enough information is available. A similar distinction in model performance was evident when segmentation granularity was considered. As the number of categories increases, ToM under uncertainty performs worse, whereas ToM with complete information performs better in terms of the stochastic game metrics.

Two research directions will be explored in future work. Firstly, we will explore alternative segmentation mechanisms. Secondly, we aim to apply MToM applications such as security games, trading, or other stochastic and economic games.

References

[Albrecht *et al.*, 2016] Stefano V Albrecht, Jacob W Crandall, and Subramanian Ramamoorthy. Belief and truth in hypothesised behaviours. *Artificial Intelligence*, 235:63–94, 2016.

[Brennan and Lo, 2012] Thomas J Brennan and Andrew W Lo. An evolutionary model of bounded rationality and intelligence. *PloS One*, 7(11):e50310, 2012.

[Camerer *et al.*, 2015] Colin F Camerer, Teck-Hua Ho, and Juin Kuan Chong. A psychological approach to strategic thinking in games. *Current Opinion in Behavioral Sciences*, 3:157–162, 2015.

[Cavalcanti *et al.*, 2010] Carina Cavalcanti, Felix Schläpfer, and Bernhard Schmid. Public participation and willingness to cooperate in common-pool resource management. *Ecological Economics*, 69(3):613–622, 2010.

[Condon, 1992] Anne Condon. The complexity of stochastic games. *Inf. and Comp.*, 96(2):203–224, 1992.

[De Weerd *et al.*, 2013a] Harmen De Weerd, Rineke Verbrugge, and Bart Verheij. Higher-order theory of mind in negotiations under incomplete information. In *PRIMA*, pages 101–116. Springer, 2013.

[De Weerd *et al.*, 2013b] 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.

[De Weerd *et al.*, 2015] Harmen De Weerd, Rineke Verbrugge, and Bart Verheij. Higher-order theory of mind in the tacit communication game. *Biologically Inspired Cognitive Architectures*, 11:10–21, 2015.

[Devaine *et al.*, 2014] Marie Devaine, Guillaume Hollard, and Jean Daunizeau. Theory of mind: did evolution fool us? *PloS One*, 9(2):e87619, 2014.

[Felli *et al.*, 2015] Paolo Felli, Tim Miller, Christian Muise, Adrian R. Pearce, and Liz Sonenberg. Computing social behaviours using agent models. In *IJCAI 2015*, 2015.

[Fiske, 2000] Susan T Fiske. Stereotyping, prejudice, and discrimination at the seam between the centuries: Evolution, culture, mind, and brain. *European Journal of Social Psychology*, 30(3):299–322, 2000.

[Ganzfried and Sandholm, 2011] Sam Ganzfried and Tuomas Sandholm. Game theory-based opponent modeling in large imperfect-information games. In *AAMAS*, volume 2, pages 533–540, 2011.

[Gmytrasiewicz and Doshi, 2005] Piotr J Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005.

[Goldman, 2012] Alvin I Goldman. Theory of mind. *The Oxford handbook of philosophy of cognitive science*, pages 402–424, 2012.

[Kashima *et al.*, 2000] Yoshihisa Kashima, Jodie Woolcock, and Emiko S Kashima. Group impressions as dynamic configurations: The tensor product model of group impression formation and change. *Psychological Review*, 107(4):914, 2000.

[Kimbrough *et al.*, 2014] Erik O Kimbrough, Nikolaus Robalino, and Arthur J Robson. The evolution of 'theory of mind': Theory and experiments, 2014.

[Mertens and Neyman, 1981] J-F Mertens and Abraham Neyman. Stochastic games. *International Journal of Game Theory*, 10(2):53–66, 1981.

[Moll and Tomasello, 2007] Henrike Moll and Michael Tomasello. Cooperation and human cognition: the vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362(1480):639–648, 2007.

[Oakes *et al.*, 1994] Penelope J Oakes, S Alexander Haslam, and John C Turner. *Stereotyping and social reality.* Blackwell Publishing, 1994.

[Pereira *et al.*, 2011] Luís Moniz Pereira, Francisco C Santos, et al. Intention recognition promotes the emergence of cooperation. *Adaptive Behavior*, 19(4):264–279, 2011.

[Pereira *et al.*, 2012] Luís Moniz Pereira, Francisco C Santos, et al. Corpus-based intention recognition in cooperation dilemmas. *Artificial Life*, 18(4):365–383, 2012.

[Perner, 1999] Josef Perner. Theory of mind. *Dev. psych.: Achievements and prospects*, pages 205–230, 1999.

[Pynadath and Marsella, 2005] David V Pynadath and Stacy C Marsella. PsychSim: Modeling theory of mind with decision-theoretic agents. In *IJCAI*, volume 5, pages 1181–1186, 2005.

[Robalino and Robson, 2012] Nikolaus Robalino and Arthur Robson. The economic approach to theory of mind. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599):2224–2233, 2012.

[Sethi and Somanathan, 1996] Rajiv Sethi and Eswaran Somanathan. The evolution of social norms in common property resource use. *The American Economic Review*, pages 766–788, 1996.

[Shapley, 1953] Lloyd S Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100, 1953.

[von der Osten *et al.*, 2017] Friedrich Burkhard von der Osten, Michael Kirley, and Tim Miller. Sustainability is possible despite greed — exploring the nexus between profitability and sustainability in common pool resource systems. *Scientific Reports*, 7(1):2307, 2017.

[Watkins and Dayan, 1992] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

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