Putting yourself in another's ruse: Theory of Mind in the Bridge Crossers game

Bachelor's Project Thesis

Mink Weel, s4377672, m.v.k.weel@student.rug.nl,
Supervisor: Dr. H.A. de Weerd

Abstract: Theory of mind allows humans to reason about the mental states of others, including their beliefs, desires, goals, and emotions. It can become increasingly complex, where higher-order theory of mind can be employed to think about mental states of those who use theory of mind themselves. In this paper an agent-based model is used to examine the benefit of first- and second-order theory of mind in a 3-player competitive setting, where agents can increase their chance of winning by predicting their opponents' actions. This extends an existing model of 2-player interactions to allow for 3-player interactions. Theory of mind agents will be added to a population initially consisting of behavior-based agents. Evolutionary dynamics will show whether theory of mind can invade the population, providing insight into why/how theory of mind evolved for humans. Generally the performance of second-order theory of mind proves to be the best, but all strategies have merit. The results are highly dependent on the amount of time agents are given to learn their opponents' strategies, and on the frequency of agents taking a random action.

1 Introduction

Theory of mind (ToM)—also known as 'mind-reading' (Apperly, 2010)—allows one to reason about the mental state of another. This mental state can include beliefs, desires, goals, and emotions (Premack & Woodruff, 1978). It is therefore no wonder that theory of mind has been shown to overlap with the capacity to feel empathy (Cerniglia et al., 2019): to 'put oneself in another's shoes.' Besides having certain cooperational benefits, theory of mind is useful in competitive situations (de Weerd & Verheij, 2011; de Weerd et al., 2013). In these settings theory of mind can be used to reason about other's thoughts in order to deduce their intentions: to 'put oneself in another's ruse.'

The focus of this paper lies on the competitive side of theory of mind. A simple competitive setting is rock paper scissors (RPS). A player *not* using ToM may take random actions, or take a behavior-based approach. If the player plays randomly, it will win 50% of the time. Instead the behavior-based approach could be used to attempt to predict the opponent's move by considering previous patterns, as humans have been shown to do (Falk & Konold, 1997). Following an opponent's rock in a prior round, the player may opt for paper, expecting the opponent to repeat their move. A player could also utilize theory of mind. Such a player would not merely consider previous behavior of an opponent to predict their action. Instead it would attribute a mental state to the opponent. A ToM player would expect the opponent to use a strategy themselves, and would account for that. Suppose the opponent lost with rock against the player's paper last round. Then the opponent might expect the player to choose paper again (belief), and play scissors (intention). The ToM player can use this deduced intention to choose rock. If the player is correct about the intention of the opponent, the theory of mind strategy would win.

But the ToM strategy can itself be beaten by another, more complex ToM strategy (de Weerd et al., 2013). The theory of mind strategy the player applied in the RPS example was (only) of the first order (the behavior-based strategy of the opponent can be seen as zero-order ToM). There is also second-order ToM, where a player forms a mental state of the opponent in which this opponent has its own mental state of the player. While the first-order ToM player before picked rock to counter scissors, a second-order ToM player would consider that the opponent expects that. The player then infers that the opponent will play paper, and therefore the player picks scissors.

There is also theory of mind of the third order and up. Humans have been shown to employ higher orders (Perner & Wimmer, 1985; Sullivan et al., 1994; Miller, 2009; Goodie et al., 2012; Arslan et al., 2012; Verbrugge et al., 2018; Arslan et al., 2020) reliably up to fourth-order ToM (Kinderman et al., 1998; Stiller & Dunbar, 2007). Groups of animals such as other primates also show some capacity of mentalization (Kano et al., 2019), However, currently no evidence shows that any animal other than humans has the same ability of higher-order theory of mind. And it is still debated whether any is fully capable of first-order theory of mind (Penn & Povinelli, 2007; Carruthers, 2008; van der Vaart et al., 2012; Martin & Santos, 2014).

The skill of theory of mind is cognitively demanding, but humans still possess it. Therefore there should be an evolutionary advantage for (higher-order) ToM reasoning (Aiello & Wheeler, 1995). Previous research shows that higher-order theory of mind can be beneficial in cooperative (de Weerd et al., 2015), competitive (de Weerd & Verheij, 2011; de Weerd et al., 2013), as well as mixed-motive situations (Verbrugge, 2009; de Weerd et al., 2017, 2014a,b). The Machiavellian intelligence hypothesis states that competitive skills such as deception and manipulation gave humans the evolutionary advantage since the largest challenge was dealing with and competing against companions (Whiten & Byrne, 1997).

In this paper an agent-based model will be used to examine the effectiveness of zero-, first- and second-order theory of mind in a competitive setting. The goal is to find out whether a first- and/or second-order theory of mind strategy can ‘invade’ a population of computational agents relying on a simple behavior-based strategy in a competitive setting, in a one-dimensional lattice population structure. In contrast to RPS, the setting will have 3 competitors. This will likely influence the effectiveness of theory of mind as it will require agents to think about what others are thinking about others. The RPS shows that a ToM strategy of a certain order can directly counter that of the preceding order. Hence it is expected that first-order ToM agents will succeed in completely invading a population of agents with a behavior-based strategy, even with three players per game. When second-order ToM agents are introduced however, the interplay between the three strategies becomes harder to predict. While second-order ToM should be able to beat first-order ToM, it might ‘overthink’ when competing against zero-order ToM. This makes the competition between the three strategies in the population a game of rock paper scissors in and of itself.

The remainder of this thesis is structured as follows. In section 2 the Bridge Crossers game is explained, as well as the three orders of ToM reasoning that agents use and the evolutionary dynamics of the population. Section 3 presents results of the agent evolution for a subset of populations, each with different assumptions on the values of important parameters. Finally in section 4 an analysis of the significance of the results is given, in particular its relevance to ToM evolution in humans.

2 Methods

2.1 The Competitive Setting

Agents will participate a game called ‘Bridge Crossers,’ which is displayed in figure 3.5. All agents start on a center platform. Three bridges are connected to this platform, each leading to a certain number of coins—either 1, 3, or 5. The game consists of multiple rounds. Every round, all agents pick a bridge and cross simultaneously. When multiple agents choose the same bridge, it collapses and none of these agents receives any coins. Meanwhile agents alone on a bridge receive the number of coins their bridge leads to. The agent with the most coins after all rounds wins.

The game is mainly inspired by the Game of Chicken (Rapoport & Chammah, 1966), in which two drivers are on a collision course with each other and one of them has to break off for both of them to survive. But the one who breaks off is called the ‘chicken’ and receives a lower reward than the other driver. In Bridge Crossers this Game of Chicken would correspond to choosing the 5-coin bridge or not. There is a high reward for being the only one who selects that bridge, but only when no oppo-

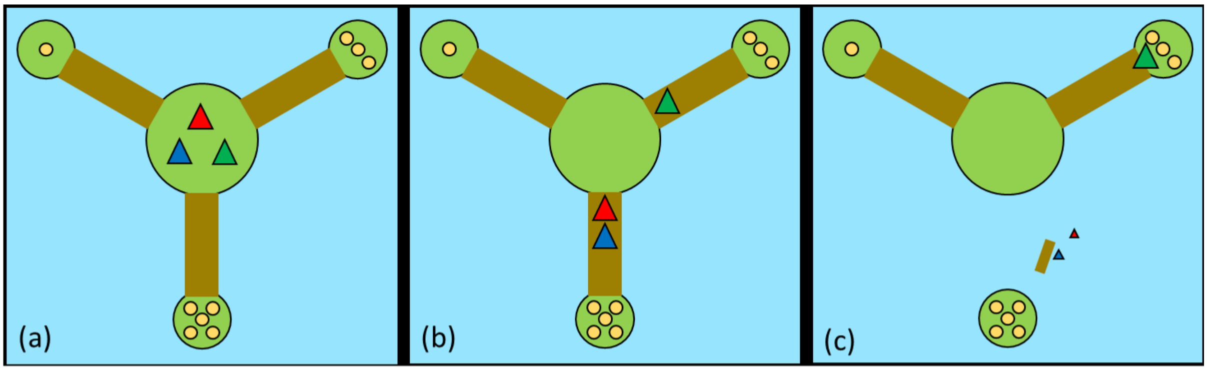

Figure 2.1: One round of the 3-player Bridge Crossers game. At the start of a round a) all players (triangles) are placed on the central platform. Connected to this position are three bridges, each leading to a number of coins (1, 3 and 5). All players choose a bridge, and then b) simultaneously cross their selected bridge. Finally, c) the coins are distributed. When a bridge has to support more than one player (in this case the bridge leading to 5 coins, which both red and blue picked), it collapses and the corresponding players get no coins. But if a player is alone on a bridge (green), that player receives all coins the bridge leads to.

nents do. The risk of opponents choosing the same bridge is tied to the number of coins it leads to.

Another similar game theory experiment is the prisoners dilemma (Poundstone, 1992). In this thought experiment two prisoners have no way to contact each other, but they are presented with a choice: betray the other and be set free, or cooperate and both get a reduced sentence. If both prisoners decide to betray however, both get the full sentence. This is similar to how two agents on the same bridge in Bridge Crossers receive no coins (they crossed each other). More specifically, Bridge Crossers resembles the *iterated* prisoners dilemma, where the prisoners dilemma is repeated for several iterations. This allows for more complex dynamics such as trust (Axelrod & Hamilton, 1981). In this variation, a successful method proved to be the tit-for-tat strategy, where a prisoner copies the previous move of the other (Harrald & Fogel, 1996; Kuhn, 2019).

Bridge Crossers is also at its core a rock paper scissors game. Each agent has three options, and if agents choose the same option, they tie. But there are two differences. One, there are more than two players. Considering the Machiavellian intelligence hypothesis, competition between more than two persons must have played a big role during evolution of human cognition. In the Bridge Crossers game specifically it has the following effect: Even if two players tie, another can still get coins. This means that the players who tied actually end up both losing.

The second difference with RPS is that each action has a unique potential reward. Going for 5 coins can bring a higher reward, which also brings higher risk. Since a game spans multiple rounds, this can introduce interesting dynamics. For example an agent can go for 5 coins even though it predicts that others will too, either to prevent anyone else from taking over the lead, or because it is the only way left to win. Two players might even work together against the leading player by taking turns in taking the fall to prevent the leader from getting coins. By introducing action-dependent reward potentials, the competitive setting is more closely related to the real world competitive situations which caused humans to evolve theory of mind.

These two additions to RPS are present with any number of players higher than two. Therefore the chosen number in this study is three, to limit computational complexity and allow for easier inspection of individual games. As for the number of bridges, this should be equal to the number of players, for the following reason. An agent tries to predict which bridges its opponents will cross. If there are fewer bridges than players, then even if an agent correctly predicts all actions of its opponents, there might not be any bridge left to cross. Conversely, if there are more bridges than players, this increases the probability that an agent incorrectly guesses the choices of the others but still by chance crosses an empty bridge. Three bridges for three players maximally rewards agents for correctly predicting the moves of others (with theory of mind), which is what this study aims to inspect.

There are three rounds per game. This should be sufficient for agents to be affected by the described additions to RPS. In particular, three rounds is the minimum which allows for all players to win regardless of the first (or any singular other) round. This means agents have the opportunity to adapt to their opponents within a single game. For the exact rewards of the bridges there are numerous viable options. The rewards should be relatively close to each other, to prevent options being so unprofitable that they are never picked. The rewards should also be compatible with the number of rounds, so that all rewards have use. With 3 rounds, the rewards are set to 1, 3, and 5 coins. This choice assures that 1) each reward, not only the highest, is useful. Receiving 3 coins twice wins over 5 coins once. And 1 coin can still result in a slight advantage over agents with otherwise the same score. 2) There is a clear difference between the rewards. This lowers randomness in actions of players (the extreme case being where all rewards are identical).

2.2 Zero-Order Theory of Mind

The experiment will use an agent-based model. These models can show how interactions between individuals may cause the emergence or necessity of a certain type of behavior (Nowak & May, 1992; Epstein, 1999; Macy & Willer, 2002; Gilbert, 2007). In this case, the interactions are the games being played and the behavior is theory of mind. While three players in Bridge Crossers start in same position, they all have their own strategy—either zero-, first- or second-order theory of mind. Zero-order theory of mind (ToM$_0$) agents do not use any theory of mind. This still means that they can be implemented in many ways. To correctly determine whether theory of mind indeed gives evolutionary advantage, the behavior-based ToM$_0$ strategy needs to be as good as it can be. De Weerd (2013) gave these agents for each of their opponents a set of zero-order beliefs: numerical values for each action which represent the likelihood of the opponent taking that action. This worked well for rock paper scissors, but the addition of action-dependent reward potentials makes it so that not only these likelihoods matter, but also the potential rewards of actions. It may be acceptable to select the 5-coin bridge even though there is a higher probability that a opponent will too. Something more complex is required for zero-order ToM.

Instead of beliefs, ToM$_0$ agents use reinforcement learning (RL). Specifically they will use Q-learning to find state-action values that represent how likely each action is to lead to a good outcome at the end of the game. The first piece of information agents need for this approach is all possible states of the game. A state in the Bridge Crossers game is the unique combination of the round number and the number of coins each player has. An example: s_x = $\{turn_x,\{coins_x\}\}$ = $\{2,\{3,0,5\}\}$. The order of the coins matters. The first number represents the player, while the following coins are those of the opponents. $\{2,\{3,5,0\}\}$ would be a different state. This allows an agent to prefer giving one opponent 5 coins over another (perhaps the one it has found to be less challenging). Since there are 3 turns, all situations where it is the start of turn 4 are either the 'win,' 'loss,' or 'tie' state. Furthermore, situations where there are turns left but there is no way to tie or win, are also the loss state. And situations where the player will win regardless of what happens are included in the win state.

All states are connected through actions. Going for 5 coins can only lead to states where either the player has 5 more coins, or the player has the same number of coins (in case an opponent chooses the same bridge). While playing, Q-learning will assign and update state-action values—Q-values—to each performed action in each visited state. The initial values are equal to the number of coins an action might lead to. This encourages agents to pick bridges leading to more coins at first, as they are more likely to lead to higher rewards when no information about the opponents is available, and there is no reason yet to leave 5 coins for others.

Every action in each state leads to a numeric reward. A possible reward scheme for actions is simply the number of coins they result in. Then the action of going for 5 coins leads to either a reward of 5 or 0. But there is a problem with this. Consider the case where an agent receives 3 coins while an opponent receives 5 coins. This would give the agent a reward of 3, while in reality it lost by 2 coins. Therefore the reward instead uses the difference in coins between the agent and the opponent with the most coins: the (possibly negative) *lead*. The reward is equal to the increase in lead due to an action. There is one exception, which is when an action leads to the end of the game. If an action wins the game, the reward is equal to the maximum coins a player can get in one turn: the maximum possible gain in lead (+5). This is because it does not matter with how much of a lead an agent wins. And contrarily, the reward for an action resulting in a loss is the maximum possible loss in lead (-5). Finally, a tie results in a reward of 0. This results in the following equation for the reward:

$$R(a, s_t) = \begin{cases} 5 & \text{if } s_{t+1} = \text{win} \\ -5 & \text{if } s_{t+1} = \text{loss} \\ 0 & \text{if } s_{t+1} = \text{tie} \\ p(s_{t+1}) - o(s_{t+1}) & \text{otherwise} \\ -(p(s_t) - o(s_t)) & \end{cases}, \tag{2.1}$$

where $p(s)$ is the number of coins the player has in state s and $o(s)$ is the number of coins of the opponent with the most coins in state s.

ToM$_0$ agents generally take the action with the highest Q-value, but there is a probability that they make a random choice instead. The agents follow the ϵ-greedy exploration strategy. The ϵ parameter is set to 0.2. With 3 turns, this leads to a ≈50% probability per game of performing at least one random move (which is not the action with the highest Q-value). This allows for exploration necessary for RL, and also makes it harder for other agents to correctly predict the agent's actions. At the same time, the ToM$_0$ agent's strategy is not sacrificed too heavily as there is also a 50% probability of playing a game without any random actions.

Besides allowing exploration and unpredictability, the ϵ parameter has a third function. Due to the nature of the agent-based simulation, each ToM$_0$ agent has the exact same strategy. By including some randomness in their actions, two ToM$_0$ agents are prevented from always choosing the same bridge, as their initial Q-values and Q-value update steps are the exact same. The Q-learning state-action value update rule is

$$Q(s_t, a_t) := Q(s_t, a_t) + \alpha[r_t + \gamma \max_{a \in A} Q(s_{t+1}, a) - Q(s_t, a_t)], \tag{2.2}$$

where r_t is the reward given after taking action a_t, α is the learning rate, and γ the discount factor. These last two parameters are the same for all agents, meaning agents update state-action values identically if they select the same action. An ϵ probability of selecting a 'sub-optimal' action makes up for the simplicity of the agents. It simulates in a very abstract way the many other considerations humans make which make them choose for seemingly sub-optimal actions, requiring for the theory of mind strategies to have a more sophisticated way to predict actions. The ϵ parameter will have a large effect on the results, so several values around 0.2 will be tested: 0.05, 0.1, 0.2, 0.3, and 0.4.

The discount factor for Q-learning is set to 0.9. With 3 turns, this means that the agents learn to plan ahead as none of the rewards are discounted too heavily. And planning ahead should be very possible with only 3 turns. Finally, the learning rate is 0.5. This allows an agent to quickly change its Q-values when facing ToM agents of a higher order, which themselves will be able to change their strategy during the game. It reflects the dynamic nature of the game where agents have to stay one step ahead of their opponents not to choose the same bridge.

2.3 First-Order Theory of Mind

A ToM$_0$ agent considers how likely each action is to lead to a good outcome based only on its own Q-values. But a ToM$_1$ agent first thinks about what it would do were it in the situation of each of its opponents. The agent uses their Q-values to make predictions of their actions, in order to inform its own action. Therefore a ToM$_1$ agent keeps track of not just its own Q-values, but those of every player in the game.

Each action can lead to multiple different states, depending on which actions the opponents take. The Q-value of an action is simply put a weighted 'average' of the rewards it has led to in the past. This means that it implicitly includes the probabilities of which actions the opponents will take. For example, the 5 coins action Q-value likely lowers over time since the opponents will take the same action, returning a negative reward or 0. By first explicitly predicting the opponents' actions with theory of mind, a ToM$_1$ agent can compute a more accurate utility of the action rather than relying on the Q-value.

ToM$_1$ agents use three steps to select an action, based on a similar study by De Weerd (2013). (1) The agent tries to predict the actions of its opponents. (2) The agent uses the predictions to compute the ToM utilities of its own actions. (3) The agent integrates the ToM utilities with its own (zero-order) Q-values.

(1) To predict an opponent's action, a ToM$_1$ agent first assumes that the opponent is a ToM$_0$ agent. This is the only case where a prediction is possible, as a ToM$_1$ agent cannot form a mental state of another ToM$_1$ agent; only a ToM$_{\geq 2}$ agent can. Assuming that the opponent has no theory of mind of its own, the ToM$_1$ agent simply places itself in the state of the opponent, and uses the opponent's Q-values to select one optimal action $\hat{a}_j^{(1)}$ in the same way a ToM$_0$ agent would. There are three cases where this leads to an incorrect prediction. First, the opponent could take a random action due to the ϵ parameter. Second, the Q-values might be too different from the actual Q-values of the opponent. Third, the opponent is not a ToM$_0$ agent.

To solve the first problem, the ToM$_1$ agent does not assign a probability of 100% to the predicted action of the opponent. Instead it creates a probability distribution over all the opponent's possible actions. This distribution, called beliefs b, is formed with the agent's own ϵ value:

$$b_j^{(1)}(a) = \begin{cases} 1-\epsilon, & \text{if } a = \hat{a}_j^{(1)} \\ \epsilon/2, & \text{if } a \neq \hat{a}_j^{(1)} \end{cases}. \tag{2.3}$$

As for the problem of mismatching Q-values, this solves itself over time, assuming the learning rates of both agents are similar. The third problem of the opponent not being a ToM$_0$ agent is addressed during the integration step (3).

(2) The reward of each action depends on the specific combination of both opponents' actions. Therefore the ToM$_1$ agent has to combine the beliefs of the opponents' actions into probabilities for each *opponent action pair*, given by

$$P_{tom}(a_j, a_k) = b_j(a_j) \cdot b_k(a_k), \tag{2.4}$$

where a_j is one of the three possible actions of opponent j, and a_k that of opponent k. Next, the agent assigns a utility to each of its own actions according to theory of mind, taking into account the likelihood of each opponent action pair:

$$\phi_{tom}(a_i) = \sum_{a_j \in A} \sum_{a_k \in A} P_{tom}(a_j, a_k) \cdot U(a_i, a_j, a_k), \tag{2.5}$$

where a_i is the action of the agent itself, and the utility U of that action is given by

$$U(a_i, a_j, a_k) = R(a_i, a_j, a_k, s_t) + \max_{a \in A} Q(s_{t+1}, a_i), \tag{2.6}$$

which depends on the actions a_j and a_k of the opponent. This equation sums the immediate payoff (reward) with the highest Q-value in the next state (that actions a_i, a_j, and a_k lead to), allowing the agent to plan ahead.

(3) Now ϕ is the preference for actions according to ToM$_1$. But a ToM$_1$ agent does not completely rely on its theory of mind predictions. It also uses its own Q-values, which represent the agent's ToM$_0$ strategy. To know how much the agent should rely on the theory of mind predictions, it has to determine how confident it is that the opponent is a ToM$_0$ agent, or rather how confident it is that first-order theory of mind works against that opponent. For this reason, while playing, the agent keeps track of whether ToM$_1$ correctly predicts the opponent's actions, and assigns to opponent j a confidence value $0 \leq c_j^{(1)} \leq 1$ for ToM$_1$ (more in section 2.5). There is a separate confidence value for both opponents. Similar to the beliefs, the two confidence values need to be combined into a general confidence in theory of mind:

$$c_{tom} = c_j^{(1)} \cdot c_k^{(1)}. \tag{2.7}$$

With this confidence, the ToM utilities ϕ_{tom} can be integrated with the Q-values:

$$I(a) = (1 - c_{tom}) \cdot Q(s, a) + c_{tom} \cdot \phi_{tom}(a). \tag{2.8}$$

This results in a utility for each action which is based on the agent's Q-values and the predictions of first-order theory of mind. Like ToM$_0$ agents, ToM$_1$ agents pick the action with maximum value with a probability of 1-ϵ.

2.4 Second-Order Theory of Mind

Just like ToM$_1$ agents expand upon the strategy of ToM$_0$ agents, ToM$_2$ expands on ToM$_1$ with one difference: ToM$_2$ agents have the ability to predict actions of ToM$_1$ agents. They can place themselves in the position of their opponent, and select an action just like a ToM$_1$ agent would. But ToM$_2$ agents can also still place themselves in the mental state of a ToM$_0$ agent. Rather than having for opponent j one confidence value $c_j^{(1)}$ and one predicted action $\hat{a}_j^{(1)}$, ToM$_2$ agents have these for both first- and second-order ToM. This changes a few details in the three steps explained for ToM$_1$ agents.

(1) To predict opponent j's action, a ToM$_1$ agent assumes that the opponent is a ToM$_0$ agent. ToM$_2$ agents do the same, and predict action $\hat{a}_j^{(1)}$. But they also predict what the opponent would do if it were a ToM$_1$ agent, resulting in $\hat{a}_j^{(2)}$. While equation 2.8 computes a ToM$_1$ agent's final action utilities by integrating $\hat{a}_j^{(1)}$ with the agent's Q-values, for ToM$_2$ agents $\hat{a}_j^{(2)}$ needs to be integrated as well.

Since there are now two different action predictions, there are also two sets of beliefs: the ToM$_1$ belief $b_j^{(1)}$ given by equation 2.3, and the ToM$_2$ belief $b_j^{(2)}$ which instead uses $\hat{a}_j^{(2)}$:

$$b_j^{(2)}(a) = \begin{cases} 1-\epsilon, & \text{if } a = \hat{a}_j^{(2)} \\ \epsilon/2, & \text{if } a \neq \hat{a}_j^{(2)} \end{cases}. \tag{2.9}$$

(2) For ToM$_1$ agents each opponent action pair has one ToM probability given by equation 2.4. This probability depends on the ToM$_1$ beliefs $b_j^{(1)}$ and $b_k^{(1)}$. But there are now 2 sets of beliefs for each opponent. Therefore every opponent action pair has 4 different ToM probabilities: one for each combination of opponent j's possible ToM order o and opponent k's possible ToM order p:

$$P_{tom(o,p)}(a_j, a_k) = b_j^{(o+1)}(a_j) \cdot b_k^{(p+1)}(a_k). \tag{2.10}$$

(3) Similarly there are 4 ToM utilities for each action of the agent

$$\phi_{tom(o,p)}(a_i) = \sum_{a_j \in A} \sum_{a_k \in A} P_{tom(o,p)}(a_j, a_k) \cdot U(a_i), \tag{2.11}$$

and 4 combine confidence values

$$c_{tom(o,p)} = c_j^{(o+1)} \cdot c_k^{(p+1)}. \tag{2.12}$$

To integrate the ToM utilities with their Q-values, ToM$_2$ agents use the same function as ToM$_1$ agents (equation 2.8), except 4 times in sequence, with

$$I(a) := (1 - c_{tom(o,p)}) \cdot I(a) + c_{tom(o,p)} \cdot \phi_{tom(o,p)}(a), \tag{2.13}$$

where I(a) starts as the action's Q-value. The ToM utilities are integrated in the following order: $\phi_{tom(0,0)}$, $\phi_{tom(0,1)}$, $\phi_{tom(1,0)}$, $\phi_{tom(1,1)}$. This order matters and will influence the results, but this influence is not examined in this paper. This order simply integrates $\phi_{tom(0,0)}$ first like a ToM$_1$ agent would, and then integrates the three utilities computed with ToM$_2$ in an arbitrary order.

2.5 Learning over Games

A ToM$_0$ agent updates the Q-values of each state-action pair with Q-learning after each turn. A ToM agent keeps track of the Q-values of its opponents as well, so it updates three state-action values after a turn. Additionally, a ToM agent updates the confidence values in theory of mind for each of its opponents. These represent how sure the agent is that a ToM order n works against opponent j. ToM$_2$ agents therefore have two confidence values per opponent. These are updated with

$$c_j^{(n)} = \begin{cases} \lambda + (1-\lambda) \cdot c_j^{(n)}, & \text{if } a_j = \hat{a}_j^{(n)} \\ c_j^{(n)}, & \text{if } n = 2 \text{ and} \\ & a_j = \hat{a}_j^{(1)} = \hat{a}_j^{(2)} \\ (1-\lambda) \cdot c_j^{(n)}, & \text{otherwise} \end{cases}, \tag{2.14}$$

where λ is the *learning speed* with the value of 0.5. This allows an agent to quickly adapt the confidence in its ToM orders, which is useful when an opponent can change its strategy as well (by adapting its own confidence values). The confidence value for ToM$_2$ only increases if ToM$_1$ did not already correctly predict the action. This prevents an agent from using ToM$_2$ in the case where ToM$_1$ performs just as well, so that the agent doesn't overestimate the theory of mind of its opponents.

A ToM_2 agent has to keep track of not only its own confidence values, but also the ToM_1 confidence of the opponents about itself. This is necessary for when the agent forms the mental state of ToM_1 agent. To update this value, in each round a ToM_2 agent 'predicts' what its own action would be if it were a ToM_0 agent. The agent assumes that the opponents update this confidence identically, so there is only one value $c_i^{(1)}$. All confidence values are initialized to 0 so that theory of mind is only used when deemed necessary. Empirical results show that this does not have a large effect, likely due to the high learning speed.

2.6 Population and Evolution

To determine whether the ToM_1 and/or ToM_2 strategy can invade a population of ToM_0 agents, first 1000 ToM_0 agents enter the population. Since theory of mind evolved in a large population of humans, this number should not be too small. Before the evolution starts, one of the ToM_0 agents becomes a ToM_1 agent. Then in each epoch, every agent competes in the Bridge Crossers game according to a one-dimensional lattice population structure, where each agent has two neighbours. Since there are 3 players per game, each agent plays 3 games per epoch. Once in the 'middle' against its two neighbours, and in two games as an 'opponent.'

An agent takes over (evolves into) the ToM order of the winner of the game where the agent faced its own two neighbours. This evolution happens at the end of an epoch, after all agents have played 3 games. This prevents a possible cascade where all ToM_0 evolve into ToM_1 agents in epoch 1. Instead, it will take at least 500 epochs for all ToM_0 to disappear, which will make the results more interpretable as epochs will resemble passing time. The population structure likely has a large influence on the results. In the one-dimensional lattice, a strategy spreads from a single point. In e.g. a well-mixed population it would likely spread more quickly.

A ToM_1 agent may have a non-zero probability of losing against two ToM_0 agents. Even if it is a small probability, this may cause the ToM_1 strategy to immediately disappear from the population. Therefore in each subsequent epoch, if there are no ToM_1 agents, one ToM_1 agent replaces the agent where the first ToM_1 agent started. From epoch 100, a ToM_2 agent is placed here instead, as long as there are no ToM_2 agents already in the population. This gives the ToM_1 strategy some time to spread, but not to take over the entire population, which would prevent ToM_2 agents from having to compete against ToM_0 agents. The ToM_2 agent is placed in the same position as the initial ToM_1 agent, because ToM_2 is meant to counter ToM_1 and would therefore emerge where the ToM_1 strategy exists.

As many epochs will be ran as necessary for the population to stabilize (the percentage of agents for each strategy stops changing). It is difficult to apply a strict rule for this, so first an (over)estimation is made. If the population is still visibly changing when evolution stops, a larger number of epoch will be ran. The population is assumed to be stabilized when the strategy percentages show a periodic pattern for more epochs than when there wasn't such a pattern. For the result graphs this periodic pattern will be trimmed down.

In one epoch, each agent should compete in 3 games. However, there are only 10 possible configurations of strategies ({0,0,0}, {0,0,1}, ... {2,2,2}). To reduce computation time, instead of simulating identical games over and over, each of these configuration is assigned a win percentage distribution, just once. For example, the configuration {0,0,1} could be assigned a 10% probability of ToM_0 winning, and a 90% probability of ToM_1 winning. These percentages are computed before the first epoch starts. Then during the population algorithm, agents use the win percentages as a probability distribution to decide which strategy to evolve into.

In 3 of these configurations all agents use the same strategy, so one strategy has a win probability of 100%. For the other 7 each, n x 100 games are played with agents using the strategies of that configuration. During the first n-1 games, the agents update their Q-values and confidence values according to section 2.5, but the outcomes of the games do not matter—these are practice games, necessary for reinforcement learning and for theory of mind. Afterwards, one more game is played. The strategy of the winner of the nth game receives a point.

As for ties, each tying agent *could* be given a point. If in a game with two ToM_1 agents and one ToM_0 agent the two ToM_1 agents tie for first place, it might seem logical to award a point to the ToM_1 strategy. But the goal of agents is to win, without taking into consideration the possibility of a tie. The reward of a tie is 0, so it is not given points here either. This is not merely an implementation choice, but it is because in this study ToM is examined strictly in a competitive setting. If the ToM_1 strategy were given a point, the setting would become mixed-motive as the sum of pay-offs would vary. Agents in the same strategy would become a 'team' while the setting is supposed to be a competition between the individual agents. Even though the agents will not be able to *learn* to become a team, as the strategy points are separate to agent rewards, the results will not be those of a competitive setting.

After n games, the Q-values and confidence values are reset to their initial values. This process of playing n games and resetting values is repeated 99 more times, resulting in 100 - ties points distributed over the participating strategies. The distribution of points becomes the probability distribution that agents use when deciding which strategy to evolve into. For example, if n is 10 and in the configuration {0,0,1} ToM_0 wins the 10th game 20 times, ToM_1 wins the 10th game 40 times, and there are 40 ties in the 10th game, this results in a probability distribution of 33.3% for ToM_0 and 66.6% for ToM_1.

The number of games n is an important variable, because theory of mind may only be useful before/after a certain number of games. In order to form correct mental states of opponents, their strategy has to be learned. The number of games will also influence the performance of ToM_0 because of reinforcement learning. Therefore the results of the population evolution will be inspected for several values of n: 2, 3, 4, 5, 7, 10, 15, 20, 25, and 50. This does not include 1 as the results would be too random; each state would be visited for the first time. Additionally, several ϵ values will be tested: 0.05, 0.1, 0.2, 0.3 and 0.4. This parameter may also have a large impact on the results and will show how robust the theory of mind strategies are against randomness. Or how much they need not to continuously take the same action as another agent with the same ToM order.

The code for the simulation is published on GitHub here: https://github.com/The-Mink/BridgeCrossersToM

3 Results

The results prove to be highly dependent on the ϵ value and the number of played games n. Each of the three orders of theory of mind performs well in at least a few of the tested combinations of n and ϵ. Only some cases are shown in this section.

3.1 Second-Order Theory of Mind

For most values of ϵ and n the final population is comprised of only ToM_2 agents, and in many others ToM_2 still performs better than the other strategies. The evolution of agent strategies when ϵ is 0.2 and n is 25 is displayed in figure 3.1. Before ToM_2 agents are introduced, only 2 configurations (matchups of ToM orders) matter: 2 ToM_0 agents against 1 ToM_1 agent and 1 ToM_0 agent against 2 ToM_1 agents. As seen in table 3.1, in the first configuration ToM_1 wins 56/100 games. In the second, ToM_1 wins 57/100 games. Meanwhile ToM_0 wins only 13/100 and 19/100 games. Figure 3.1 reflects this with a steep decrease in ToM_0 agents and increase in ToM_1 agents.

When at epoch 100 ToM_2 agents enter the population, the number of ToM_0 agents keeps decreasing at the same rate, while the increase of ToM_1 slows down. The ToM_2 strategy starts invading at a slightly slower rate than ToM_1. Around epoch 1150 there are no more ToM_0 agents, and the ToM_1 population immediately starts declining. Now only the configurations {1,1,2} and {1,2,2} matter, where ToM_2 wins 43/100 and 50/100 games respectively, and ToM_1 only 30/100 and 17/100. The difference here is smaller than that between ToM_0 and ToM_1 agents, and so is the rate of ToM_2 taking over ToM_1. Still, after epoch 2650 only ToM_2 agents remain. Like for all following results, the graph ends when the number of agents for each strategy stops changing (as explained in section 3.3). The periodic pattern at the end (in this case horizontal lines) is trimmed down without a specific rule.

Figure 3.2 shows a situation where the ToM_2 still successfully invades, but not the entire population. In this specific population ϵ is 0.05 and n is 10, although the same pattern is present when $4 \leq n \leq 15$). ToM_1 has a 83/100 win rate in the configuration {0,0,1} as seen in table 3.2. But the win rate in the configuration {0,1,1} is only 35/100, lower than that of ToM_0 (59/100). Even though ToM_1 per-

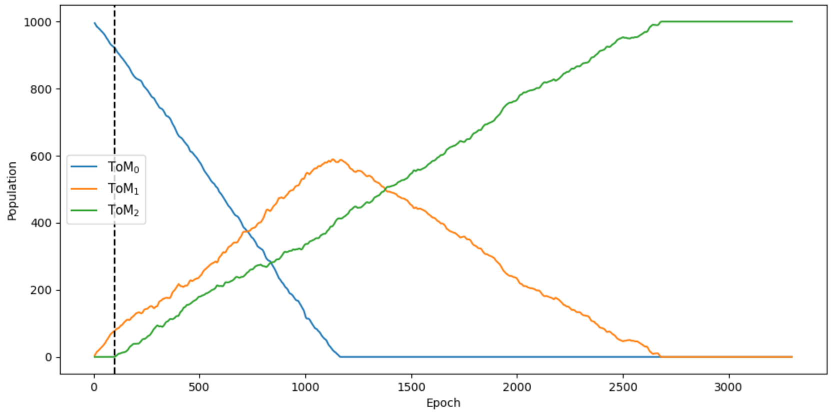

Figure 3.1: The number of zero-, first- and second-order agents in the population of 1000 during epochs, where the probability of a random action was 20% and the agents learned their opponents' strategies over 25 consecutive games. ToM$_2$ agents entered the population at epoch 100 (visualized with a dashed line). The results are smoothed with a moving average over 10 epochs.

	ToM$_0$	ToM$_1$	ToM$_2$
0,0,1	13	56	
0,1,1	19	57	
0,0,2	22		51
0,2,2	21		42
1,1,2		30	43
1,2,2		17	50
0,1,2	20	17	33

Table 3.1: The number of wins (out of 100) in the 25th game for each order of theory of mind in each configuration of strategies, where the probability of a random action was 20%.

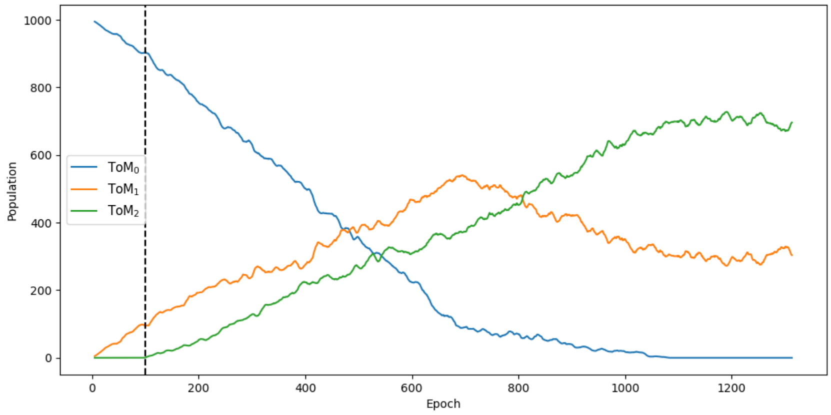

Figure 3.2: The number of zero-, first- and second-order agents in the population of 1000 during epochs, where the probability of a random action was 5% and the agents learned their opponents' strategies over 10 consecutive games. ToM$_2$ agents entered the population at epoch 100 (visualized with a dashed line). The results are smoothed with a moving average over 10 epochs.

	ToM$_0$	ToM$_1$	ToM$_2$
0,0,1	16	83	
0,1,1	59	35	
0,0,2	14		82
0,2,2	42		43
1,1,2		14	72
1,2,2		60	31
0,1,2	38	19	34

Table 3.2: The number of wins (out of 100) in the 10th game for each order of theory of mind in each configuration of strategies, where the probability of a random action was 5%.

forms worse in this configuration compared to when ϵ was 0.2 and n was 25 (and ToM$_1$ won 56/100), it invades the population more quickly.

ToM$_1$ starts declining at epoch 700. This time, there are still ToM$_0$ agents in the population. The ToM$_2$ strategy keeps invading, but both the ToM$_0$ and ToM$_1$ strategies decline. As fewer ToM$_0$ agents remain in the population, the decline of the ToM$_1$ strategy slows down. When at epoch 1050 the ToM$_0$ population disappears, the decline of ToM$_1$ agents stops. Its population stabilizes along with ToM$_2$, at a ratio of about 3:7.

3.2 First-Order Theory of Mind

There are a few cases where ToM$_2$ fails to invade a population of ToM$_1$ agents. One of these is shown in figure 3.3, where ϵ is 0.3 and n is 25. In each epoch after 100, a ToM$_2$ agent enters the population if there is not already one present. Still, the strategy fails to properly enter the population, both when there are still ToM$_0$ agents and when there are not. The win rates of ToM$_2$ given in table 3.3 are much lower than the win rates of ToM$_2$ in tables 3.1 and 3.2. This allows ToM$_1$ to completely take over the population.

In even fewer cases, the final population consists of mostly ToM$_1$ agents while there are still ToM$_2$ agents. Figure 3.4 (ϵ is 0.2 and n is 2) shows one of these situations. ToM$_1$ takes over the ToM$_0$ strategy, and the ToM$_2$ strategy slowly invades as well. When there are no more ToM$_0$ agents in epoch 1000, the ToM$_2$ strategy catches up to ToM$_1$ until about a third of the population is ToM$_2$ and two thirds are ToM$_1$. Some variation does occur. For example around epoch 7500, where for a short time the ratio of ToM$_1$ agents increases. Table 3.4 shows that the configurations {1,1,2} and {1,2,2} have win rates 28/100 and 40/100 for ToM$_2$ and 51/100 and 25/100 for ToM$_1$, which evidently gives ToM$_1$ an advantage over ToM$_2$.

3.3 Zero-Order Theory of Mind

Zero-order theory of mind does not maintain the highest ratio of agents in the population with any tested combination of ϵ and n. However, it can get close. Figure 3.5 shows the evolution of the population when ϵ is 0.05 and n is 2—the situation where ToM$_0$ performs best. The ToM$_1$ strategy starts invading the ToM$_0$ population, and ToM$_2$ joins at epoch 100 at a similar rate. But the ToM$_0$ strategy never drops to zero; it stabilizes around epoch 500 with a population of 500. This coincides with the point where the ToM$_1$ strategy starts losing agents. ToM$_2$ keeps invading until epoch 1200, where it stabilizes just slightly above the ToM$_0$ strategy (the gap between the two strategies varies for other values of ϵ and n). Table 3.5 shows that the configurations {0,0,2} and {0,2,2} have win rates 4/100 and 93/100 for ToM$_0$ and 90/100 and 6/100 for ToM$_2$. Since both strategies lose almost all games where two of their agents face one of the other strategy, their average win rates drop quickly whenever they have more agents. Therefore both strategies hover around 50%.

4 Discussion

The goal of this study was to find out whether a first- and/or second-order theory of mind strategy can invade a population of agents relying on a simple behavior-based strategy in a competitive setting, in a one-dimensional lattice population structure. The used agent-based model was designed to reflect the theory of mind that humans possess, and to simulate why/how this skill may have evolved in the human population.

How well first-order theory of mind (ToM$_1$) and second-order theory of mind (ToM$_2$) evolved in the population of behavior-based agents (ToM$_0$) depends on two variables. First, the number of games that agents have to learn each other's strategies and to adapt their own: n. Second, the probability that agents take a random action instead of following their strategy: ϵ.

Figure 3.1 shows one of the cases where ToM$_2$ completely invades the population. Figure 3.3 shows one where ToM$_1$ does. Figures 3.2 and 3.4 show cases where both ToM$_1$ and ToM$_2$ remain when the population evolution stabilizes. Finally, figure 3.5 shows a case where ToM$_0$ is in the final population along with ToM$_2$. These results demonstrate that each of the three strategies is useful under certain circumstances. Both first- and second-order theory of mind can indeed invade a population of behavior-based agents in a competitive setting. But they do not always succeed in completely

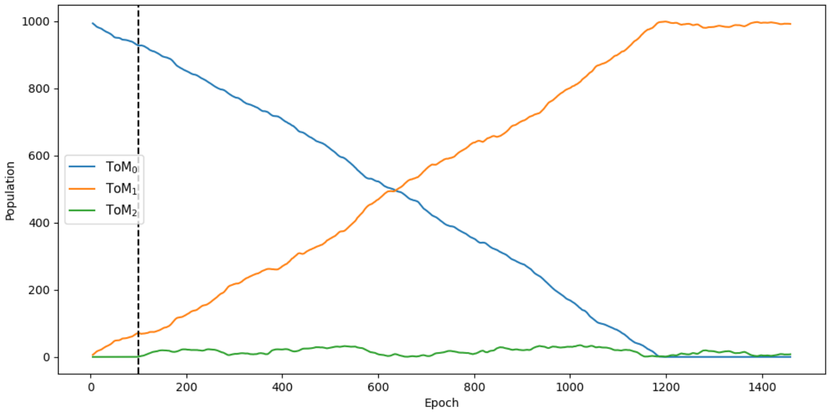

Figure 3.3: The number of zero-, first- and second-order agents in the population of 1000 during epochs, where the probability of a random action was 30% and the agents learned their opponents' strategies over 25 consecutive games. ToM$_2$ agents entered the population at epoch 100 (visualized with a dashed line). The results are smoothed with a moving average over 10 epochs.

	ToM$_0$	ToM$_1$	ToM$_2$
0,0,1	29	40	
0,1,1	24	44	
0,0,2	38		33
0,2,2	21		43
1,1,2		42	25
1,2,2		25	37
0,1,2	25	16	25

Table 3.3: The number of wins (out of 100) in the 25th game for each order of theory of mind in each configuration of strategies, where the probability of a random action was 30%.

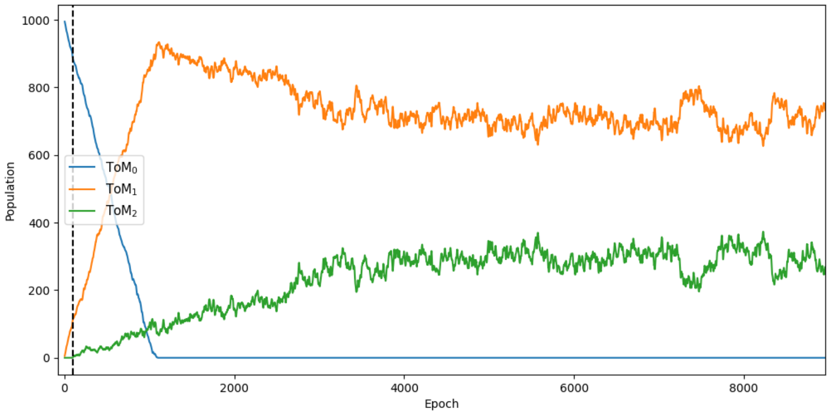

Figure 3.4: The number of zero-, first- and second-order agents in the population of 1000 during epochs, where the probability of a random action was 20% and the agents learned their opponents' strategies over 2 consecutive games. ToM$_2$ agents entered the population at epoch 100 (visualized with a dashed line). The results are smoothed with a moving average over 10 epochs.

	ToM$_0$	ToM$_1$	ToM$_2$
0,0,1	22	47	
0,1,1	34	33	
0,0,2	33		45
0,2,2	44		22
1,1,2		28	40
1,2,2		51	25
0,1,2	24	14	19

Table 3.4: The number of wins (out of 100) in the 2nd game for each order of theory of mind in each configuration of strategies, where the probability of a random action was 20%.

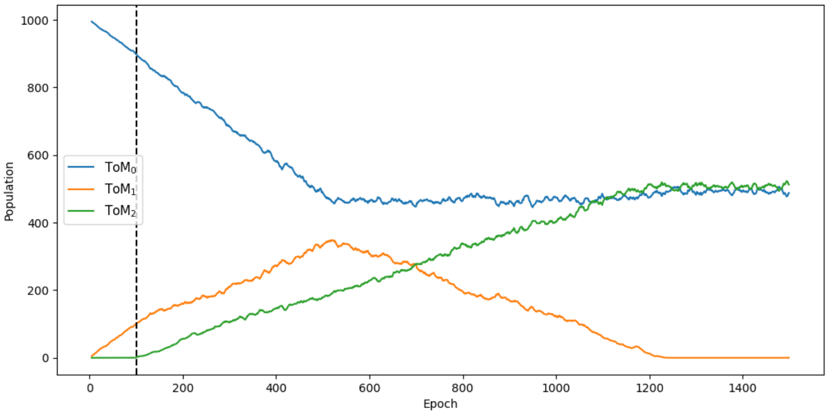

Figure 3.5: The number of zero-, first- and second-order agents in the population of 1000 during epochs, where the probability of a random action was 5% and the agents learned their opponents' strategies over 2 consecutive games. ToM$_2$ agents entered the population at epoch 100 (visualized with a dashed line). The results are smoothed with a moving average over 10 epochs.

	ToM$_0$	ToM$_1$	ToM$_2$
0,0,1	2	97	
0,1,1	81	11	
0,0,2	4		93
0,2,2	90		6
1,1,2		14	84
1,2,2		87	12
0,1,2	12	5	21

Table 3.5: The number of wins (out of 100) in the 2nd game for each order of theory of mind in each configuration of strategies, where the probability of a random action was 5%.

taking over the population.

Specifically, the ToM strategies almost always completely invade the ToM$_0$ population. In the majority of cases the ToM$_2$ takes over most of the population, but in many cases the ToM$_1$ strategy does instead. There does not seem to be a clear correlation between the two parameters (n and ϵ) and the cases where ToM$_1$ performs better than ToM$_2$.

But it is clear when the ToM strategies struggle against the behavior-based strategy. This happens when there is both a small probability of taking a random action, and agents have few games to learn their opponents' strategies (like in figure 3.5 where n is 2 and ϵ is 0.05). Because of the few games, agents do not have much time to learn, which means the ToM strategies do not have a large advantage over ToM$_0$. And since ϵ is low as well, the effect where two agents of the same order are at a disadvantage is augmented, since they will more often pick the same action. Table 3.5 demonstrates clearly that agents rarely win when one of their opponents have the same ToM order. ToM$_0$ remains in the population because it is too difficult to beat with two ToM$_1$ or ToM$_2$ agents.

It may seem like a very specific case where both these parameters have a low value, but during evolution it might have been more important to be able to compete against novel opponents than against companions. And a small probability of taking a random action might be more realistic than the higher probabilities.

4.1 Human Evolution

The model was meant to give some insight into why/how the skill of theory of mind emerged in humans. The results might implicate that theory of mind evolved because of (social) civilisation, where humans developed more relations with each other. At first the behavior-based strategy was useful because it performs well against novel opponents, but this type of first- or even second-contact competition became less and less common. As competition started to occur more between relatives in a more compact world, humans were given time to learn each other's strategies, which in turn set the stage for theory of mind to be beneficial.

But it is a leap to confidently explain with these results the evolution of theory of mind, mainly because the model is too abstract. There are many factors that will have influenced the process of theory of mind evolution, such as biological capabilities, social connections, and historical circumstances. However, the results of the model do illustrate *why* humans may have learned (higher-order) theory of mind. It beats the behavior-based strategy in many cases.

4.2 Population Structure

The results depend highly on a very important implementation detail: the population structure. And the one-dimensional lattice is not necessarily an accurate representation of the real-world human population structure, which even now is undergoing major changes (Campbell et al., 2009). The lattice is the cause of certain aspects in the population evolution graphs, for example the steady increase of the ToM$_1$ population which at one point suddenly starts decreasing. This is seen most clearly in figure 3.1, where agents play for 25 games and have a 20% probability of taking a random action. Due to the one-dimensional lattice, the ToM$_1$ strategy spreads in two directions. From epoch 1 to 100 in the population graph there is a steep increase in ToM$_1$ agents. Table 3.1 indeed shows that ToM$_1$ clearly beats ToM$_0$ when no ToM$_2$ agent is involved. But around epoch 1150 the ToM$_1$ population starts diminishing. This is because the two 'waves' of spreading ToM$_1$ agents have met; there are no ToM$_0$ agents left.

At epoch 100 the ToM$_2$ strategy is placed where the ToM$_1$ strategy originated. This makes sense from an evolutionary standpoint, as the ToM$_2$ strategy likely emerged to combat the ToM$_1$ strategy. However, with the lattice population structure this has an important consequence. The ToM$_2$ strategy starts spreading in both directions just like ToM$_1$ did. Except the ToM$_2$ agents only compete against ToM$_1$ agents. Even if some ToM$_1$ agents lose against ToM$_0$, the 'holes' that form in the spreading ToM$_1$ population will be quickly filled again with ToM$_1$ agents. This means that the ToM$_2$ strategy invades within a 'bubble' of ToM$_1$ agents and never has to compete against ToM$_0$ agents.

The one-dimensional lattice structure nullifies the game of rock paper scissors between the three ToM orders as mentioned in section 1. It was expected that ToM$_1$ beats ToM$_0$, ToM$_2$ beats ToM$_1$, and ToM$_0$ might beat ToM$_2$. Additionally, when multiple agents in a game have the same strategy, they are at a disadvantage. This is most clearly visible in table 3.5. Since in this case there is only a 5% probability of taking a random action, agents with the same strategy often take the same action, which in the Bridge Crossers game is disadvantageous. This fact combined with the rock paper scissors relationship between the strategies, could cause a population to fall into some sort of equilibrium where even for the worst performing strategy a few agents might remain. But this would only happen if the strategies were better spread throughout the population. Because of the 'waves' of evolution in the lattice, it does not.

4.3 Unrealistic Theory of Mind

In a vast majority of cases, second-order theory of mind completely takes over the population. This is not solely because ToM$_2$ generally does not have to compete against ToM$_0$ agents. For instance table 3.1, corresponding to one of the cases where ToM$_2$ invades completely, shows that ToM$_2$ would outperform ToM$_0$ even when no ToM$_1$ agents are involved.

A big reason why both ToM strategies perform better than the behavior-based strategy in almost all cases is likely the simplicity of the model. The agents use reinforcement learning, and as explained in section 2.5, they keep track of each other's state-action values. Since each agent uses the exact same computations, these values are perfect representations; the mental states that ToM agents form are without error. The disadvantage of theory of mind—the possibility of overthinking by making a series of error-prone predictions—is less present. The only variable is the probability of taking a random action, which ToM strategies can account for with confidence values. This is why theory of mind performs so well in the model (except in the cases where both n and ϵ have low values).

4.4 Conclusion

Though there are some implementation choices which may not reflect human cognition, the model has shown that both a behavior-based strategy and (higher-order) theory of mind have use. This agrees with studies which have found that humans use theory of mind (Perner & Wimmer, 1985; Sullivan et al., 1994; Miller, 2009; Goodie et al., 2012; Arslan et al., 2012; Verbrugge et al., 2018; Arslan et al., 2020). The model expands upon a similar model used for the rock paper scissors competitive setting (de Weerd et al., 2013). There are three players instead of two players, and the same model could be employed for settings with more participants.

Additionally, due to the reinforcement learning approach, the model accounts for action-dependent reward potentials which applies to many real-world situations.

References

Aiello, L. C., & Wheeler, P. (1995). The expensive-tissue hypothesis: the brain and the digestive system in human and primate evolution. *Current Anthropology*, *36*(2), 199–221.

Apperly, I. (2010). *Mindreaders: The Cognitive Basis of "Theory of Mind"*. London: Psychology Press. doi: 10.4324/9780203833926

Arslan, B., Hohenberger, A., & Verbrugge, R. (2012). The development of second-order social cognition and its relation with complex language understanding and working memory. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 1290–1295.

Arslan, B., Verbrugge, R., Taatgen, N., & Hollebrandse, B. (2020). Accelerating the development of second-order false belief reasoning: a training study with different feedback methods. *Child Development*, *91*(1), 249–270. doi: 10.1111/cdev.13186

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–1396.

Campbell, H., Rudan, I., Bittles, A. H., & Wright, A. F. (2009, September). Human population structure, genome autozygosity and human health. *Genome Medicine*, *1*(9), 91. doi: 10.1186/gm91

Carruthers, Peter. (2008, February). Metacognition in animals: A skeptical look. *Mind & Language*, *23*, 58–89. doi: 10.1111/j.1468-0017.2007.00329.x

Cerniglia, L., Bartolomeo, L., Capobianco, M., Lo Russo, S. L. M., Festucci, F., Tambelli, R., . . . Cimino, S. (2019, September). Intersections and divergences between empathizing and mentalizing: development, recent advancements by neuroimaging and the future of animal modeling. *Frontiers in Behavioral Neuroscience*, *13*, 212. doi: 10.3389/fnbeh.2019.00212

de Weerd, H., Verbrugge, R., & Verheij, B. (2013, June). How much does it help to know what she knows you know? An agent-based simulation study. *Artificial Intelligence*, *199–200*, 67–92. doi: 10.1016/j.artint.2013.05.004

de Weerd, H., Verbrugge, R., & Verheij, B. (2014a). Agent-based models for higher-order theory of mind. In B. Kamiński & G. Koloch (Eds.), *Advances in Social Simulation* (pp. 213–224). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-39829-2_19

de Weerd, H., Verbrugge, R., & Verheij, B. (2014b, August). The effectiveness of higher-order theory of mind in negotiations. In *CEUR Workshop Proceedings* (Vol. 1208).

de Weerd, H., Verbrugge, R., & Verheij, B. (2015, January). Higher-order theory of mind in the tacit communication game. *Biologically Inspired Cognitive Architectures*, *11*, 10–21. doi: 10.1016/j.bica.2014.11.010

de Weerd, H., Verbrugge, R., & Verheij, B. (2017, March). Negotiating with other minds: the role of recursive theory of mind in negotiation with incomplete information. *Autonomous Agents and Multi-Agent Systems*, *31*(2), 250–287. doi: 10.1007/s10458-015-9317-1

de Weerd, H., & Verheij, B. (2011, July). The advantage of higher-order theory of mind in the game of limited bidding. In *CEUR Workshop Proceedings* (Vol. 751, pp. 149–164).

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*, *4*(5), 41–60. doi: 10.1002/(SICI)1099-0526(199905/06)4:5<41::AID-CPLX9 3.0.CO;2-F

Falk, R., & Konold, C. (1997). Making sense of randomness: implicit encoding as a basis for judgment. *Psychological Review*, *104*, 301–318. doi: 10.1037/0033-295X.104.2.301

Gilbert, N. (2007, January). Agent-Based models. *The Centre for Research in Social Simulation*. doi: 10.1007/978-0-387-35973-1_39

Goodie, A. S., Doshi, P., & Young, D. L. (2012). Levels of theory-of-mind reasoning in competitive games. *Journal of Behavioral Decision Making*, *25*(1), 95–108. doi: 10.1002/bdm.717

Harrald, P. G., & Fogel, D. B. (1996, January). Evolving continuous behaviors in the Iterated Prisoner's Dilemma. *Biosystems*, *37*(1), 135–145. doi: 10.1016/0303-2647(95)01550-7

Kano, F., Krupenye, C., Hirata, S., Tomonaga, M., & Call, J. (2019, October). Great apes use self-experience to anticipate an agent's action in a false-belief test. *Proceedings of the National Academy of Sciences*, *116*(42), 20904–20909. doi: 10.1073/pnas.1910095116

Kinderman, P., Dunbar, R., & Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, *89*(2), 191–204. doi: 10.1111/j.2044-8295.1998.tb02680.x

Kuhn, S. (2019). Prisoner's Dilemma. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019 ed.). Metaphysics Research Lab, Stanford University.

Macy, M., & Willer, R. (2002, August). From factors to actors: computational sociology and agent-based modeling. *Annual Review of Sociology - ANNU REV SOCIOL*, *28*, 143–166. doi: 10.1146/annurev.soc.28.110601.141117

Martin, A., & Santos, L. R. (2014, March). The origins of belief representation: Monkeys fail to automatically represent others' beliefs. *Cognition*, *130*(3), 300–308. doi: 10.1016/j.cognition.2013.11.016

Miller, S. A. (2009). Children's understanding of second-order mental states. *Psychological Bulletin*, *135*, 749–773. doi: 10.1037/a0016854

Nowak, M., & May, R. (1992, October). Evolutionary Games and Spatial Chaos. *Nature*, *359*, 826–829. doi: 10.1038/359826a0

Penn, D., & Povinelli, D. (2007, May). On the lack of evidence that non-human animals possess anything remotely resembling a 'Theory of Mind'. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *362*, 731–44. doi: 10.1098/rstb.2006.2023

Perner, J., & Wimmer, H. (1985, June). "John thinks that mary thinks that..." attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, *39*(3), 437–471. doi: 10.1016/0022-0965(85)90051-7

Poundstone, W. (1992). *Prisoner's Dilemma* (1st ed ed.). New York: Doubleday.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*, 515–526. doi: 10.1017/S0140525X00076512

Rapoport, A., & Chammah, A. M. (1966, November). The Game of Chicken. *American Behavioral Scientist*, *10*(3), 10–28. doi: 10.1177/000276426601000303

Stiller, J., & Dunbar, R. I. M. (2007, January). Perspective-taking and memory capacity predict social network size. *Social Networks*, *29*(1), 93–104. doi: 10.1016/j.socnet.2006.04.001

Sullivan, K., Zaitchik, D., & Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. *Developmental Psychology*, *30*, 395–402. doi: 10.1037/0012-1649.30.3.395

van der Vaart, E., Verbrugge, R., & Hemelrijk, C. K. (2012, March). Corvid re-caching without 'theory of mind': A model. *PLOS ONE*, *7*(3), e32904. doi: 10.1371/journal.pone.0032904

Verbrugge, R. (2009, December). Logic and social cognition. *Journal of Philosophical Logic*, *38*(6), 649–680. doi: 10.1007/s10992-009-9115-9

Verbrugge, R., Meijering, B., Wierda, S., van Rijn, H., & Taatgen, N. (2018, January). Stepwise training supports strategic second-order theory of mind in turn-taking games. *Judgment and Decision Making*, *13*(1), 79–98. doi: 10.1017/S1930297500008846

Whiten, A., & Byrne, R. W. (Eds.). (1997). *Machiavellian intelligence II: Extensions and evaluations.* New York, NY, US: Cambridge University Press. doi: 10.1017/CBO9780511525636