university of groningen / faculty of science and engineering / artificial intelligence

How Theory of Mind outperforms associative learning

Bachelor's Project Thesis

Marin Krebbers, s5122155, m.krebbers@student.rug.nl,
Supervisors: Dr H. de Weerd

Abstract: Theory of Mind (ToM) is the ability to attribute mental states to others. This ability is fundamental to human social cognition, but its presence in animals remains an open question, as observed signs of ToM might instead be explained by associative learning. This study investigates whether ToM reasoning provides a performance advantage over purely associative strategies in unpredictable environments. Using a negotiation game in the Coloured Trails environment, we compare Deep-Q Network (DQN) agents (zero-order ToM) with first-order ToM agents that simulate their opponent's decisions using an internal DQN model. Despite limited success in accurately inferring opponent goals, ToM agents generally achieve higher cumulative rewards than DQN agents across varied conditions. While the advantage is not consistent across all scenarios, these results indicate that predictive ToM reasoning can provide a meaningful advantage in certain contexts. This supports the idea that ToM-like strategies may have emerged gradually from associative learning, which would make it easier to believe that animals have ToM.

1 Introduction

Human interactions rest on the foundational assumption that others are like ourselves: they think, hold beliefs and act with intentions. This assumption is the basis of theory of mind (ToM) (Premack & Woodruff, 1978), the ability to reason about the mental states of others, even though those states are not directly observable. The sophistication of ToM reasoning can be measured in orders. An entity with zero-order ToM is unable to infer mental states at all and responds only to the directly observable actions of others. For example, a zero-order ToM agent may play chess by only observing the current board state and evaluating the best move purely based on the current board state and the objective consequences of moves, without considering what the opponent knows, believes, or intends. First-order ToM refers to the ability to attribute mental processes to other entities, which it then bases its own actions on. A first-order ToM agent would recognize that its chess opponent wants to win. Therefore, this ToM agent might employ tactics like leaving an important piece undefended, where taking this piece would analytically lead to a worse position for the opponent. This could bait the opponent into taking this piece and gain a more advantageous position, successfully predicting the opponent's desires and using this to its advantage. Employing a second-order ToM would entail reasoning about how other agents attribute mental states to others, and how these attributions influence their actions. Such an agent might recognize that the undefended piece could be "too good to be true", prompting a reassessment based on the possibility that the move is intended to exploit a false sense of security. This understanding would necessitate thinking about how another may think of you.

While closed analytical games make it easier to paint a picture of how ToM operates, research demonstrates that it plays a very broad role in everyday human cognition. ToM has been shown to be integral to social understanding (Baron-Cohen et al., 1985; Decety & Jackson, 2004), cooperative behaviour (Sally & Hill, 2006), empathy (Decety & Jackson, 2004), and even deception (Sodian & Frith, 1992). These skills are essential for interpersonal human life. ToM appears to be necessary for

many of the fundamental human behaviours that define us as social creatures, so it is unsurprising that scholars have argued for ToM being an evolutionary adaptation central to the success of the human species (Byrne, 1996; Tomasello et al., 2005). ToM has been found to be useful in counterfactual reasoning (CFR) settings (Perner & Rafetseder, 2011), which deals with reasoning about realistic hypothetical situations that have not actually happened, such as: if I would have studied, I would have passed the exam. Both ToM and CFR require the ability to keep in mind multiple different mental models of reality.

Although there is strong evidence that ToM is fundamental to human cognition, its presence in non-human animals remains less clear. Some studies suggest that certain animals, such as great apes and birds, may exhibit rudimentary ToM-like abilities (Kano et al., 2019; Ostojić et al., 2013). However, there is ongoing debate about whether these behaviours truly reflect mental state attribution or are better explained by associative learning mechanisms (Penn & Povinelli, 2007; van der Vaart & Hemelrijk, 2014). To help clarify this distinction, I examine the behaviour of artificial agents: Deep-Q Network (DQN) agents that learn through association, and first-order ToM agents capable of reasoning about others' goals. Importantly, the ToM agent is not trained from scratch but builds on an underlying DQN model, allowing it to simulate the opponent's decisions using the same learning mechanism. By comparing their performance, I aim to evaluate whether ToM provides distinct advantages over purely associative strategies in unpredictable environments.

De Weerd et al. (2022) demonstrated that ToM offers particular advantages in unpredictable environments. Building on this insight, I use the same coloured-trails environment, but with a key difference: rather than hand-designed reasoning heuristics, I explore how agents that learn their behaviour from a more fundamental level, both with and without ToM capabilities, perform. This approach is a more reasonable reflection of associative learning, which allows me to assess whether the benefits of ToM in an unpredictable environment persist when strategies are acquired through experience rather than pre-programmed knowledge.

This thesis is structured as follows. Chapter two describes the experimental setup, including the negotiation environment, agent designs, and implementation details of both the Deep Q-Network (DQN) and Theory of Mind (ToM) models. Chapter three presents the results of the experiments, comparing performance across different conditions and strategies. Chapter four discusses the implications of these findings in the context of associative learning and ToM.

2 Methods

In this section, I will specify what environment, associative learning (zero-order ToM) model, first-order ToM model and exprimental setup I used and how they operate. For the full implementation of these methods, see: `https://github.com/PotatoPulse/coloured_trails`.

Environment

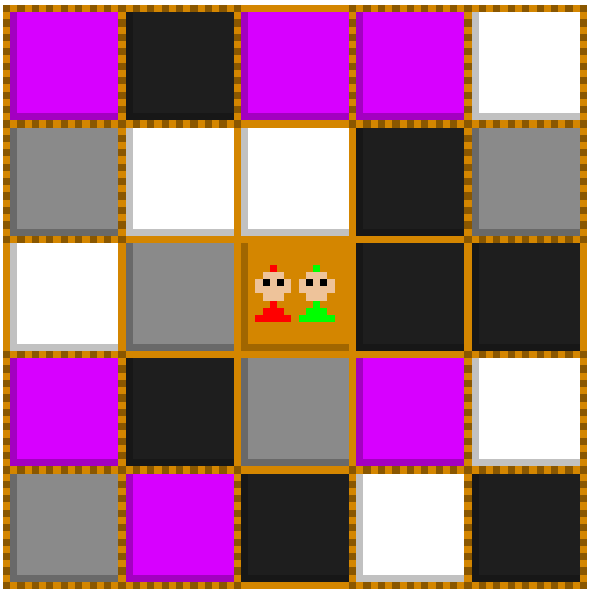

Figure 2.1: A 5×5 Coloured Trails board. Each tile is a distinct colour (white, black, gray, purple), and the two agents (red and green) are positioned on the central tile. Tiles with dotted borders indicate possible goal locations.

The environment consists of a 5×5 board filled with coloured tiles. There are four different colours used: white, black, gray and purple. This board is implemented as a 5×5 list of strings, where each

string represents a colour. At the start of each game, a new board is generated by randomly shuffling its rows. This results in 5! = 120 possible board configurations, introducing variability across games, which provides a balance between maintaining a small, manageable state space and encouraging generalisation.

For each possible tile colour, there are two corresponding chips with the same colour, creating a fixed chip set: (white, white, black, black, gray, gray, purple, purple). Each game begins with both players being randomly assigned four chips from the fixed chip set, as well as a goal location, following De Weerd et al. (2017). The players know what their own goal location is, but the goal location of their opponent is hidden information. Only corner tiles and tiles adjacent to the corners are valid goal locations (tiles with dotted borders in Figure 2.1), ensuring that any path towards a goal will have a minimum length of three. Players can move to any of the horizontally and vertically adjacent tiles, as long as they own a chip with the same colour as that tile. After a move, the coloured chip is handed in.

Players take turns proposing offers, which the opponent can either accept or reject. This process follows a fixed order: the "initiator" always makes the first offer, and the "responder" always replies first. This decision is based on prior research indicating that the initial offer can significantly influence the rest of the negotiation process (Engler & Page, 2022). The game ends as soon as a player accepts an offer or decides to terminate negotiations.

Player scores are calculated using the following formula:

$$steps\_to\_goal * 100 + chips * 50 + win * 500$$

Where:

steps_to_goal is the number of steps the agent took toward its goal location.

chips is the number of chips not used to move on the board

win has a value of one if the goal was reached, and zero otherwise.

2.1 Zero-order model - DQN

A zero-order ToM agent is incapable of attributing mental states such as goal states to others. In this thesis, I used a Deep Q-Network (DQN) (Mnih et al., 2013) implemented with torch. The Neural Network (NN) I used has an input size of 148, a hidden layer of size 128 and an output size of 81. This hidden layer size was selected to reflect the reduced complexity of the state space relative to the original DQN paper (Mnih et al., 2013), which contained high-dimensional visual input from Atari games using two hidden layers of 256 units each.

The input of the NN represents the state of the player. The first 12 neurons encode the goal location using one-hot encoding over 12 possible goals. The following eight neurons encode the current chip distribution. If one of these neurons has a value of one, the chip is assigned to the player; if it is zero, the opponent owns the chip. The next eight neurons encode the previous offer in the same way, where the chip distribution is seen as if the previous offer had been accepted. Following De Weerd et al. (2017), the agent only considers the most recent offer that the opponent has made and has no explicit memory about offers made further in the past. For the very first turn, this previous offer will be the initial chip distribution. The last 120 neurons encode the current board using one-hot encoding.

The 81 output neurons of the NN represents the Q-values of the actions the player can take. In this case, there are $3^4 = 81$ different offers possible, because there are four unique chip variants with a duplicate, so each player can get either no, one, or two of each of the four chips.

An offer is accepted when the agent outputs the exact same offer as the previous one, which it receives as the previous offer in its input. Withdrawing from negotiations is indicated by proposing the initial chip distribution. Any other output is interpreted as a rejection of the incoming offer.

At the start of each game, if a new board configuration is encountered, a reward table is generated for every possible chip distribution using a recursive pathfinding algorithm. The reward corresponding to the initial chip distribution is stored as a property for future reference.

When the agent needs to make an offer, it calculates the current state by encoding the goal location, current chip distribution, previous offer, and board configuration into a single input vector for the neural network. The agent then applies an epsilon-greedy strategy (Sutton et al., 1998), where it computes a value for epsilon based on the current

step count, decay rate, and epsilon bounds. If a randomly sampled float is less than epsilon, a random action (i.e., offer) is selected. Otherwise, the action with the highest Q-value (action associated with the output neuron with the highest activation) is chosen. The selected action, along with the state, is stored in the agent's transition memory, and the corresponding offer is raised.

When the agent receives an offer and must decide whether to accept or reject it, the NN input state is recalculated using the incoming offer as the new "previous offer". An action is again selected using the network. If this action matches the incoming offer, the offer is accepted. Otherwise, it is rejected. The agent must learn this behaviour through experience. The resulting input state is stored as the next_state in the transition memory, finalising the transition containing: $[s, a, r, s']$ (i.e., [state, action, reward, next_state]).

After each negotiation outcome, the agent evaluates the reward. If the opponent accepted the agent's offer, the reward is computed as the difference between R_{terminal}, the reward when the game has ended, and R_{start}, the reward of the initial state. If the offer was rejected, the reward is zero. This is unlike De Weerd et al. (2017), where agents have explicit time pressure. Instead, it's expected that this time pressure is eventually implicitly encoded in the network weights.

If the agent's memory buffer contains more than 30 transitions, an optimisation step is performed. A batch of transitions is randomly sampled, and their associated state, action, reward, and next state tensors are extracted from the stored transitions for further calculations. The policy network estimates Q-values for the current state-action pairs, while the target network computes the maximum Q-value for the next states. Terminal states, in which negotiations are ended, are excluded from this computation. Expected Q-values are calculated using the Bellman equation, incorporating immediate rewards and discounted future values (Sutton et al., 1998). The network is updated by minimising the Huber loss between predicted and expected Q-values. Gradient clipping (Pascanu et al., 2013) is applied to improve stability, and a single step of gradient descent is used to update the model parameters.

2.2 Theory of Mind model

The Theory of Mind (ToM) model is built around an internal DQN agent that operates exactly as the DQN agent that was previously described. This model is inspired by prior work on recursive ToM in negotiation scenarios (De Weerd et al., 2017). De Weerd et al. (2017) notes that the ToM model may perform particularly poorly when it functions as the initiator. For this reason, the ToM model is only assigned the responder role in this thesis. Transitions that it encounters are stored during the game and passed to this internal DQN so its network keeps up to date.

When the ToM agent is asked to respond to the initiator's offer by raising a new offer, it first obtains the current states from its DQN puppet. This state is passed to its action selection, which follows a sort of epsilon greedy-algorithm. An epsilon is calculated just like before. If a randomly generated float is larger than this epsilon, an action is predicted using ToM, otherwise, a random action is selected. This way the ToM agent can still explore unseen actions. The ToM algorithm loops over each possible action (i.e. all 81 possible offers) that it could take and constructs the opponent's state for that offer, the current board, its guessed goal and the current chip distribution. This state is given to the DQN puppet to simulate what the opponent would do in response to the potential offer. Now, the value of the offer is calculated for all different cases. If the offer already withdraws from negotiations to begin with, it has a value of 0. For the values that take into account the simulated response from the opponent, I incorporated a term v for the possibility that the opponent may behave differently. The base values are calculated follows if we predict the opponent withdraws or accepts:

$$Q_{\text{offer}}(\text{withdraw}) = 0$$

$$Q_{\text{offer}}(\text{accept}) = R_{\text{accept}} - R_{\text{start}}$$

where:

- R_{accept} is the reward after the offer is accepted and chips have been traded,
- R_{start} is the reward in the initial state of the game.

Whenever the opponent denies the offer of the agent, the agent must be able to evaluate the resulting state. In this thesis, I consider three ways

of evaluating this state resulting from the predicted offer of the trading partner (s').

$$Q_{\text{offer}}(\text{deny}) = \max_{a'} Q_{\text{next}}(s', a')$$

$$Q_{\text{offer}}(\text{deny}) = \max(0, Q_{\text{response\_offer}}(\text{accept}))$$

$$Q_{\text{offer}}(\text{deny}) = \begin{cases} Q_{\text{response\_offer}}(\text{accept}), & \text{if we accept} \\ 0, & \text{otherwise} \end{cases}$$

Which I have respectively given the following names:

- Raw: The raw Q-value from the action we will be able to take after the action of the opponent.
- Max: The maximum of zero and the reward we would receive if we accept the response offer. This basically assumes that we will not accept an offer in the future that leads to a negative score
- 1-lookahead: If we accept the predicted response, we use the value we obtain from this. Else, the reward is zero.

For Raw and 1-lookahead, we use the ToM agent's own zero-order predicted response (obtained from a forward pass of the internal DQN) to the opponent's predicted response. From these values, the final values for the possible offers are calculated as follows:

$$\text{Value(end)} = \upsilon \cdot Q_{\text{offer}}(\text{accept}) + \upsilon \cdot Q_{\text{offer}}(\text{deny})$$

$$\text{Value(accept)} = (1-2\upsilon) \cdot Q_{\text{offer}}(\text{accept}) + \upsilon \cdot Q_{\text{offer}}(\text{deny})$$

$$\text{Value(deny)} = \upsilon \cdot Q_{\text{offer}}(\text{accept}) + (1-2\upsilon) \cdot Q_{\text{offer}}(\text{deny})$$

where $\upsilon \in [0, 1]$ models uncertainty in the opponent's behaviour. Over all possible offers, the offer with the highest resulting value will be raised by the ToM agent.

When the ToM agent has to decide whether to accept or deny an incoming offer, it does so just like the DQN agent, sampling an action with the updated previous offer. If this offer matches the incoming offer, the agent accepts, otherwise it rejects the offer. With this incoming offer, a prediction is made about the goal location of the opponent. To do this, the agent contains a goal guess distribution, which starts out as a 1/12 for each goal. At each incoming offer, the agent loops over all possible goal locations and constructs the opponent's state for each goal location. An action over each possible goal is sampled from the DQN puppet. If this action aligns with the perceived action the opponent actually took, the probability of the opponent having that goal location increases, otherwise it decreases. The used goal prediction is the goal with the maximum probability in this distribution and will be used to predict future actions of the opponent.

2.3 Model parameters

epsilon_start The initial value of epsilon

epsilon_end The value epsilon decays toward over time

epsilon_decay The rate at which epsilon decays

prediction_epsilon* The chance that the opponent may behave differently in ToM action selection (υ in this paper)

gamma The discount factor in the Bellman equation, which controls how much future rewards are worth relative to immediate rewards.

lr The learning rate controls how much the network weights change in response to the loss.

goal_lr* The learning rate for goal estimation controls how much the assigned likelihood of a possible opponent's goal changes based of observations.

batch_size The number of transitions sampled per optimisation step

board The board object currently in use

name A string identifier for the agent

DQN_agent* An instance of a DQN agent that will serve as a puppet. If this is not passed, a new DQN agent will be made from the other given parameters.

* This parameter is unique to the ToM model

epsilon_start (ϵ_{start}), epsilon_end (ϵ_{end}) and epsilon_decay (ϵ_{decay}) are used to regulate the

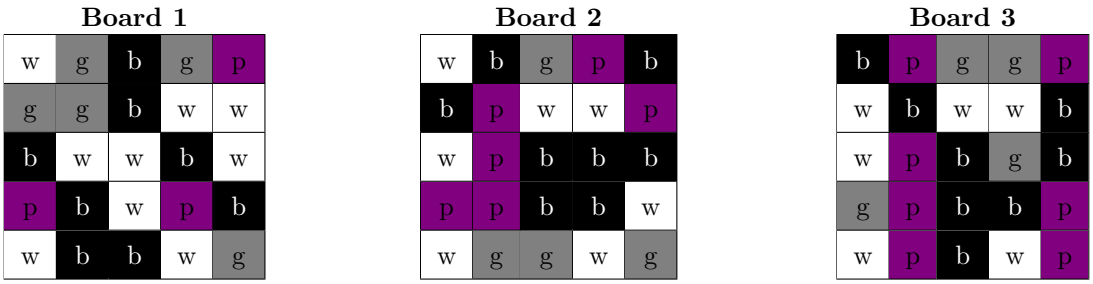

Figure 2.2: Side-by-side display of the three randomly selected 5×5 game boards.

epsilon decay for each agent. The chance to take a random action ϵ is calculated as follows:

$$\epsilon = \epsilon_{\text{end}} + (\epsilon_{\text{start}} - \epsilon_{\text{end}}) \cdot e^{-1 \cdot \text{steps}/\epsilon_{\text{decay}}}$$

Where steps is the amount of actions sampled up until this point.

Experimental setup

For experiments, I will use a randomly generated parent board that is saved. Chips and goals are selected for agents as described above. At the start of each game, the rows of the parent board will be randomly shuffled. On this board setup, a DQN agent trains against an agent that always accepts an incoming offer to initialise parameters for 500 games, which will initialise the weights of the DQN with more appropriate values for making offers. This DQN agent will be saved and loaded as two new agents for the next round of training. These DQN clones will play against each other for 30,000 games on the same parent board, again with shuffled rows, as the initialisation run. Within these 30,000 games, the DQN will learn strategies to effectively play against one another. These 30,000 DQN vs. DQN games will serve as pre-training for the following DQN vs. ToM runs. The responder DQN agent is saved and then loaded twice. One of these loaded agents will serve as an internal DQN for the ToM model. The ToM model and the other loaded DQN agent will play against each other for 250 games on the same parent board, with row-shuffling, the underlying DQN pre-trained on to preserve learned behaviour. During these games, the agents' cumulative scores are stored as well as the predicted probability that the ToM agent ascribed to the actual goal of its opponent. These metrics are recorded for 20 different runs of the 250 games for the same initial agents with stochastic strategies. These 250 games are performed for three different variations of calculating $Q_{\text{offer}}(\text{deny})$, which tries to value offers according to how good they will be in the future, after the current turn. This whole process is repeated for three different random boards.

3 Results

The three randomly generated boards can be seen in Figure 2.2.

Two DQNs played against each other on these three boards for 30,000 games as pre-training. These training results can be seen in Figure 3.1. On Board one (see Figure 2.2), the responder beat the initiator throughout the games, ending with a more than 2,000,000 point advantage. The initiator obtained a negative score in this simulation, which means that this agent has not learned to withdraw from negotiations, which would give it a score of zero. On Board two, the initiator started off on the winning hand for the first 10,000 games, after which the responder overtook the initiator, ending in another win for the responder. The point difference in the end is around 1,000,000 this time. On the third board, the agents start out equally well for the first 10,000 games. Eventually, the initiator ends up winning this game, again with a more than 2,000,000 point difference. It's important to notice that these results are tied to the stochasticity within the agents' strategies, so running a new experiment on these same boards may lead to completely different results.

Negotiation can be understood as if the agents are trying to divide a pie between themselves (Raiffa, 1982). An agent can try taking a bigger portion of this pie for themselves, or the agents can cooperate to increase the size of the pie itself. Changes in the cumulative score graph that happen symmetrically for both the initiator and responder are a consequence of redistributing the pie, because one agent loses points so the other can gain points.

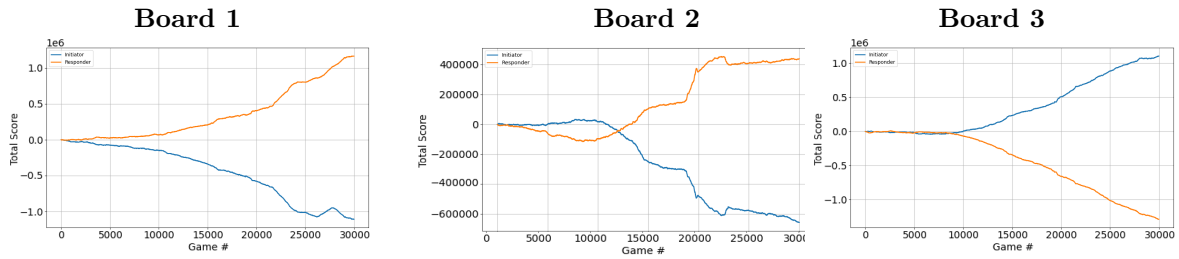

Figure 3.1: Cumulative score of DQN vs DQN across board configurations over 30,000 games. The blue graphs represent the initiator and the orange graphs represent the responder.

An example of this is on Board two at around game 20,000. Asymmetric changes in the graphs point to a change in the size of the pie. For example, in Board one around game 27,500, the initiator gains points without the cumulative score of the responder decreasing. In this instance, they have likely found a cooperative strategy from which they both gain points. In the pre-training, we can see that competition and cooperation both show up in this environment with learned agents.

All in all, we see a variety of behaviours and outcomes across these three boards, which will make conclusions about the final experiments more robust, because they have been obtained using differing DQN agents.

In Figure 3.2, we can see the performance of the DQN agent against the ToM agents across boards and Q_{offer}(deny) strategies. There is a visible spread across the 20 runs, which is caused by stochasticity inherent to the agents' ϵ-greedy strategy together with randomness in the environment: random row shuffling, random goal locations and random initial chip distributions. It is noticeable that ToM outperforms DQN on average in every circumstance, ending up with a higher cumulative score than the DQN agent. Even though the ToM agents outperforms DQN on average, there were still turns in which the DQN overtook the ToM agent in score at some point during the run. This is signified by the overlap in standard deviation.

Looking at the different Q_{offer}(deny) strategies. To evaluate the strategies, I will look at both cumulative score value and the spread as standard deviation. A lower standard deviation points to a more robust strategy, because it gained a more consistent outcome across different situations. On Board one, the Max strategy outperformed the others, achieving a higher average cumulative score with low spread at the first half of the runs (see Figure 3.2). For Board two, the 1-lookahead strategy had a similar average score to the other strategies, but with less spread and minimal overlapping spread, meaning the strategy is more robust against its opponent. For the final board, the 1-lookahead strategy ended up with a higher average cumulative score and had a lower spread. The spread in the first 25 games was minimal, which is probably due to the fact that the used DQN network had a big vulnerability (see results for Board three in Figure 3.1). The 1-lookahead strategy may have taken advantage of this vulnerability immediately, because it only looks at how good the next state will be for itself, while the other strategies' Q_{offer}(deny) calculation may have been mislead by this exploit themselves, leading to a smaller advantage.

Finally, I recorded the predicted probability the ToM assigned to the DQN's actual goal (see Figure 3.3). Comparing the goal location predictions to the prior probability of 1/12, I can conclude that the ToM did not meaningfully predict the opponents' goal location. The lack of spread observed for the 1-lookahead strategy on Board three can be explained by the previously described strategy of exploiting a vulnerability in the DQN network immediately, such that the ToM did not have enough information to base any prediction on.

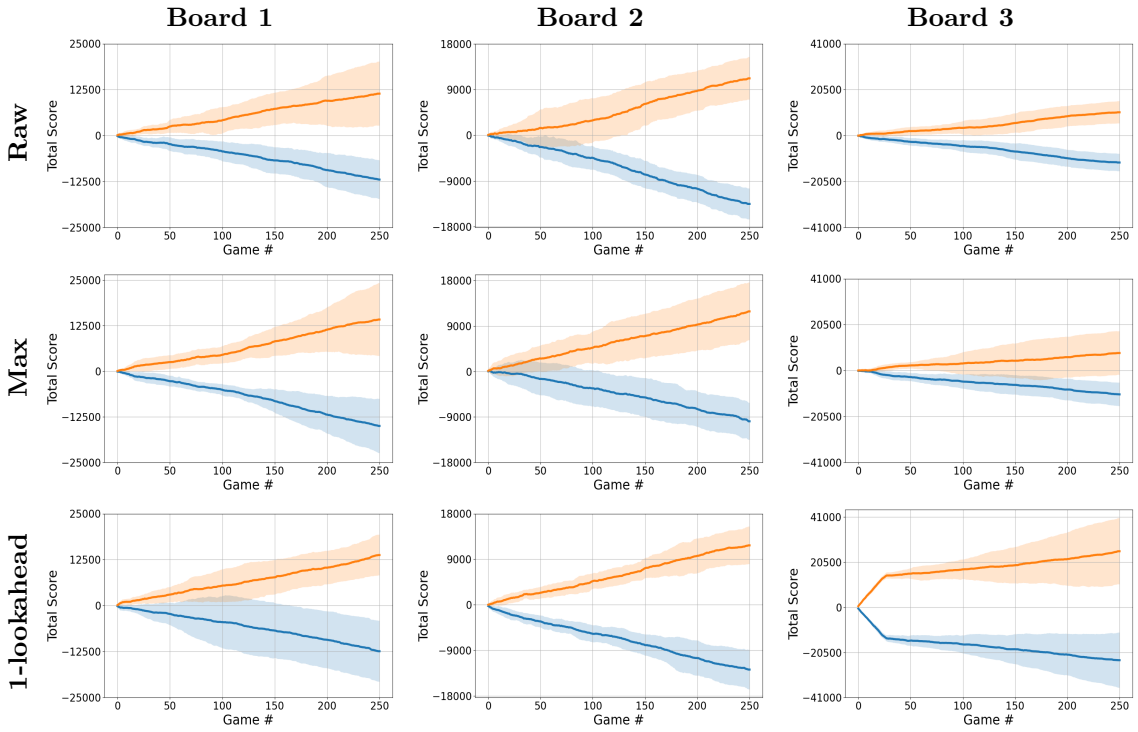

Figure 3.2: Cumulative score of DQN vs ToM agents across strategies and board configurations. The orange line represents the responder (ToM), while the blue line represents the intiator (DQN). The shaded area represents one standard deviation difference from the mean.

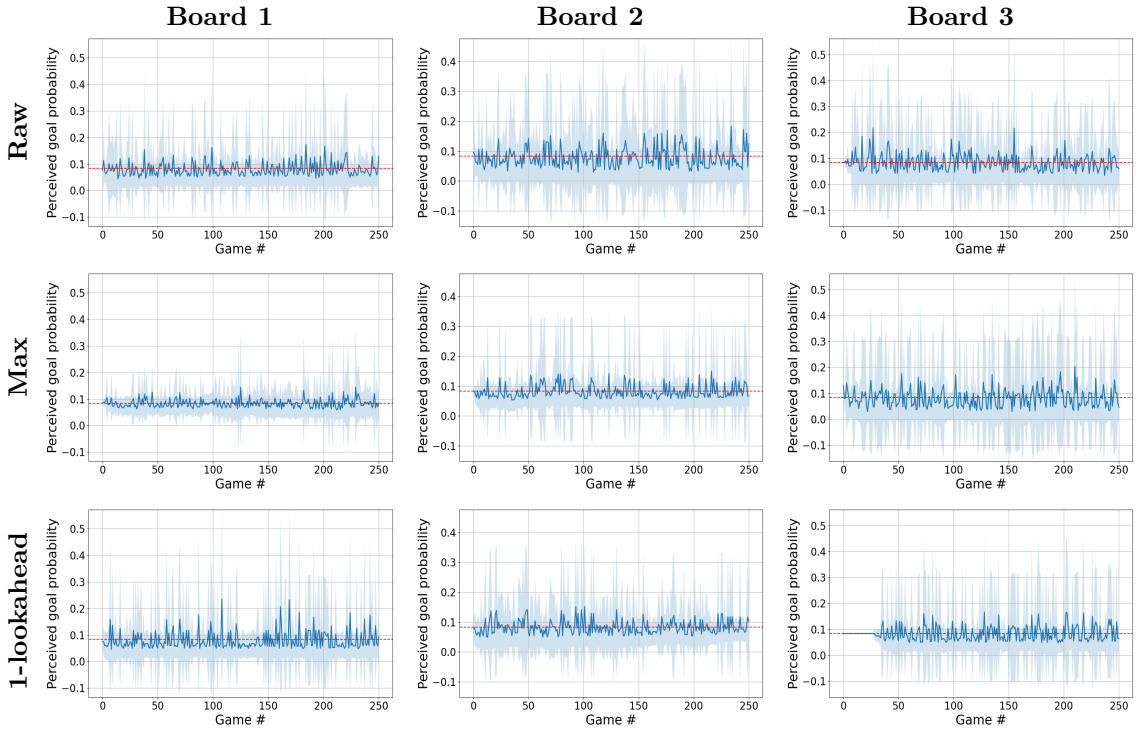

Figure 3.3: ToM goal estimation probability of opponent's actual goal (red dashed line shows the prior probability of 1/12). The shaded area represents one standard deviation difference from the mean.

4 Discussion

This thesis set out to investigate whether agents equipped with Theory of Mind (ToM) capabilities achieve a strategic advantage over agents that rely solely on associative learning, specifically in unpredictable environments. By comparing the performance of Deep Q-Network (DQN) agents and ToM agents built on top of DQN across multiple game configurations, the goal was to assess whether reasoning about others' goals and decision making leads to more effective negotiation behaviour.

The results demonstrate that ToM agents achieved higher cumulative scores than DQN agents across all board configurations and offer evaluation strategies. While performance varied across the 20 runs due to the agents' stochastic decision-making processes, ToM consistently outperformed DQN on average. Among the three $Q_{\text{offer}}(\text{deny})$ strategies tested (Raw, Max, and 1-lookahead), the 1-lookahead strategy generally resulted in the highest average scores and the lowest variance, particularly on the third board. The cumulative score plots showed that in some runs the DQN agent temporarily outscored the ToM agent, but the overall trend favoured ToM. Goal inference by the ToM agent remained close to chance level, which raises questions about how the ToM agent achieved its performance advantage.

The ToM agent failed to meaningfully predict the opponent's goal. There are several possible explanations for this. It could mean that the underlying neural network changes significantly during playing, making predictions based on the ToM's own neural network useless. However, you would expect to see some meaningful predictions at the start of the runs, because the neural networks would not have had time to diverge yet. A more likely possibility is that the DQN agents do not explicitly look at their goal location while calculating their next move. This may mean that the DQN agents do not encode their goal location in their raised offers, making it impossible for the ToM to predict the opponent's goal successfully. It is likely that the DQN agents found more generalised strategies, looking more so at offers that may be better in general on this board instead of looking explicitly at its assigned goal location. As such, ToM also would not need to guess the opponent's goal to gain an advantage, because private information is not relevant when the opponent does not use this information in its decision making, which is why we still see a better performance from the ToM agent. So all of the advantage of the ToM model hinges on the predictive part of ToM, since the interpretative part fails to make meaningful conclusions about the opponent's state. It could be interesting to look into agents that do successfully encode private information in their actions to see how the predictive component of ToM plays into these results.

From the overlap in the standard deviation of cumulative scores (Figure 3.2), it is evident that the DQN agent could occasionally outperform the ToM agent. Throughout this project, I observed that the exploitability of the internal DQN model works both ways. Since both agent types at least start with the exact same underlying DQN network and evaluate incoming offers in the same way, they are susceptible to similar weaknesses. This highlights how tightly coupled the current implementation is to the DQN architecture. Applying other reinforcement learning algorithms to represent associative learning or implementing a more generalized ToM framework could lead to different behavioural dynamics and potentially more robust outcomes.

The score results show that associative learning can find strategies in this unpredictable environment that both exploit others or cooperate with others. Importantly, ToM provides a clear advantage over associative learning in this setting. This advantage supports the theory that ToM gradually arose through evolution. Since we already assume that the animals in question employ associative learning (Pearce & Bouton, 2001), a clear gradual evolutionary path from associative learning towards predictive ToM capabilities can be painted by these results. The fact that a simpler version of ToM, which did not directly take into account any raw learned Q-values to guess future utility, performed the best may further reinforce this argument.

References

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46.

Byrne, R. W. (1996). Machiavellian intelligence.

Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews, *5*(5), 172–180.

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews*, *3*(2), 71–100.

De Weerd, H., Verbrugge, R., & Verheij, B. (2017). Negotiating with other minds: The role of recursive theory of mind in negotiation with incomplete information. *Autonomous Agents and Multi-Agent Systems*, *31*, 250–287.

De Weerd, H., Verbrugge, R., & Verheij, B. (2022). Higher-order theory of mind is especially useful in unpredictable negotiations. *Autonomous Agents and Multi-Agent Systems*, *36*(2), 30.

Engler, Y., & Page, L. (2022). Driving a hard bargain is a balancing act: How social preferences constrain the negotiation process. *Theory and Decision*, *93*(1), 7–36.

Kano, F., Krupenye, C., Hirata, S., Tomonaga, M., & Call, J. (2019). 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.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Ostojić, L., Shaw, R. C., Cheke, L. G., & Clayton, N. S. (2013). Evidence suggesting that desire-state attribution may govern food sharing in eurasian jays. *Proceedings of the National Academy of Sciences*, *110*(10), 4123–4128.

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning* (pp. 1310–1318).

Pearce, J. M., & Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology*, *52*(1), 111–139.

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 731–744.

Perner, J., & Rafetseder, E. (2011). Counterfactual and other forms of conditional reasoning: Children lost in the nearest possible world. In *Understanding Counterfactuals, Understanding Causation: Issues in Philosophy and Psychology* (pp. 90–109). Oxford University Press.

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

Raiffa, H. (1982). *The Art and Science of Negotiation*. Harvard University Press.

Sally, D., & Hill, E. (2006). The development of interpersonal strategy: Autism, theory-of-mind, cooperation and fairness. *Journal of Economic Psychology*, *27*(1), 73–97.

Sodian, B., & Frith, U. (1992). Deception and sabotage in autistic, retarded and normal children. *Journal of Child Psychology and Psychiatry*, *33*(3), 591–605.

Sutton, R. S., Barto, A. G., et al. (1998). *Reinforcement Learning: An Introduction* (Vol. 1) (No. 1). MIT press Cambridge.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(5), 675–691.

van der Vaart, E., & Hemelrijk, C. K. (2014). 'Theory of mind' in animals: ways to make progress. *Synthese*, *191*, 335–354.