Higher-order theory of mind in negotiations under incomplete information: Online appendix

Harmen de Weerd, Rineke Verbrugge, Bart Verheij

Institute of Artificial Intelligence, University of Groningen

Appendix A Theory of mind agents in Colored Trails

In this section, we describe the formal model and mathematical description of zero-order, first-order, and higher-order theory of mind agents. We consider the case in which the Colored Trails game involves two players. In our setting, a Colored Trails game is a tuple $\mathcal{CT} = \langle N, \mathcal{D}, L, \pi \rangle$, where:

- $N = \{1, 2\}$ is the set of agents, where agent 1 is the focal agent, and agent 2 is his trading partner;
- $\mathcal{D}$ is the set of possible distributions of chips over agents;
- L is the set of possible goal locations; and
- $\pi : L \times \mathcal{D} \to \mathbb{R}$ is the score function, such that $\pi(l, D)$ denotes the score of the focal agent when his goal location is $l \in L$ and the chips are distributed according to distribution $D \in \mathcal{D}$.

Note that this representation does not address the task of finding routes between locations. Instead, the score function π specifies the maximum score an agent can achieve with a given set of chips. That is, we assume that agents make no mistakes in finding routes between locations and do not consider the possibility that mistakes could be made in finding these routes.

To compare distributions $D_1, D_2 \in \mathcal{D}$, we use the relation $<_j$, such that $D_1 <_j D_2$ iff distribution D_1 assigns fewer chips of color j to the focal agent than distribution D_2. We also use the notation D^* for a distribution $D \in \mathcal{D}$ from the perspective of the trading partner. That is, for each color j, the distribution D^* assigns as many chips of color j to the focal agent as distribution D assigns to his trading partner, and vice versa.

In addition to the public information in the form of game $\mathcal{CT}$, the initial distribution of chips $D^0 \in \mathcal{D}$ is publicly announced to the agents. The focal agent and his trading partner are also assigned a goal location $l_1, l_2 \in L$. Agents know their own goal location, but do not know the goal location of the other agent.

A run of the game $\mathcal{CT}$ is an ordered list of offers $O^1, \ldots, O^t, \cdots \in \mathcal{D}$, where O^t is the offers made at time t. Agents follow a negotiation protocol in which they alternate in making offers until an agreement is reached. After the first offer has been made, the agent who received the last offer O^t decides whether to accept the offer O^t, withdraw from negotiations, or make an offer O^{t+1} of his or her own. The theory of mind agents presented here are inspired by the theory of mind agents used by [1] to investigate the effectiveness of theory of mind in competitive settings. In Section A.1, we describe an agent that does not model unobservable mental content of others, and is therefore unable to make use of theory of mind. Sections A.2 and A.3 explain how agents can make use of increasing orders of theory of mind to predict the behaviour of their trading partner.

A.1 Zero-order theory of mind agent

We assume that agents make use of a zero-order theory of mind model to predict the behaviour of other agents. A zero-order theory of mind (ToM_0) agent does not make use of unobservable mental states, but forms beliefs $b^{(0)} : \mathcal{D} \rightarrow [0,1]$ based on past experiences, so that $b^{(0)}(O)$ specifies the subjective probability that the agent assigns to his trading partner accepting a given offer $O \in \mathcal{D}$. Note that the ToM_0 agents we model do not have explicit memories of previous offers in the run, but only form beliefs about the future behaviour of their trading partner.

Whenever the ToM_0 agent receives an offer O from his trading partner, the ToM_0 agent updates his zero-order beliefs to reflect that he considers it less likely that his trading partner would accept certain offers. More precisely, the ToM_0 agent decreases his belief that his trading partner will accept an offer O' when offer O' assigns more chips of some color j to him than offer O does. For example, suppose that the trading partner makes an offer O that assigns 4 blue chips to the focal agent. The focal agent then decreases his belief that the trading partner will accept any offer that assigns 5 or more blue chips to the focal agent.

The belief update as a result of receiving an offer O from the trading partner is represented by $U^+(b^{(0)}, O)$, which is defined as

$$U^+(b^{(0)}, O)(O') = (1-\lambda)^m \cdot b^{(0)}(O') \qquad \text{with } m = |\{j | O <_j O'\}|, \tag{1}$$

where $\lambda \in [0,1]$ is an agent-specific learning speed.

The agent's learning speed represents the degree to which the agent adjusts his beliefs concerning the behaviour of his trading partner. An agent with a high learning speed strongly believes that his trading partner is unwilling to give up chips. Such an agent is more likely to withdraw from negotiations or accept the offer of his trading partner. On the other hand, an agent with learning speed $\lambda = 0$ will keep making the same offer, and expects that this will eventually lead to a successful trade.

Once the ToM_0 agent has replaced his zero-order beliefs $b^{(0)}$ with updated beliefs $U^+(b^{(0)}, O)$, he decides how to respond to the offer O. If the agent were to accept the offer O, the new distribution of chips becomes O and the game ends. To the ToM_0 agent, this option has a value of $\pi(l_1, O) - \pi(l_1, D^0)$, which is the change in score between the distribution resulting from offer O and the current distribution D.

Alternatively, the agent can choose to reject the offer O and make an offer of his own. Using his zero-order beliefs $b^{(0)}$, the ToM_0 agent can calculate the zero-order expected value $EV^{(0)}(O'; b^{(0)}, l_1, D^0)$, which is the expected change in score as a result of making the offer $O' \in \mathcal{D}$. This function weights the expected increase in score by making the offer O' with the time costs incurred for an additional round of negotiation. That is,

$$\begin{aligned} EV^{(0)}(O'; b^{(0)}, l_1, D^0) &= b^{(0)}(O') \cdot \Big(\pi(l_1, O') - \pi(l_1, D^0)\Big) + (1 - b^{(0)}(O')) \cdot (-1) \\ &= b^{(0)}(O') \cdot \Big(\pi(l_1, O') - \pi(l_1, D^0) + 1\Big) - 1. \qquad (2) \end{aligned}$$

If the *ToM*$_0$ agent chooses to make an alternative offer O', he makes the offer that maximizes the expected value $EV^{(0)}(O'; b^{(0)}, l_1, D^0)$.

If the score of the *ToM*$_0$ agent would decrease by accepting the trading partner's current offer O, and if he assigns a negative value of every alternative offer $O' \in \mathcal{D}$, the *ToM*$_0$ agent can choose to withdraw from negotiations. This action does not change the current distribution of chips, but also does not incur the cost of another round of negotiation. The value assigned to terminating negotiations is therefore zero.

The *ToM*$_0$ agent chooses the option that he assigns the highest value. That is, the *ToM*$_0$ agent's response to offer O is

$$\mathit{ToM}_0(O; b^{(0)}, l_1, D^0) = \begin{cases} O' & \text{if } EV^{(0)}(O'; b^{(0)}, l_1, D^0) \geq V > 0 \\ \text{accept} & \text{if } \pi(l_1, O') - \pi(l_1, D^0) \geq V > 0 \\ \text{withdraw} & \text{if } V \leq 0 \end{cases}$$

where $V = \max\{EV^{(0)}(O'; b^{(0)}, l_1, D^0), \pi(l_1, O') - \pi(l_1, D^0)\}$.

If the *ToM*$_0$ agent chooses to continue negotiations by making a new offer O', and if his trading partner does not accept this offer, the *ToM*$_0$ agent makes an additional update to his beliefs. This update is similar to the one shown in Equation (1), and causes the *ToM*$_0$ agent to decrease his belief that his trading partner will accept an offer O'' for every color j for which the offer O'' does not assign fewer chips of color j to the *ToM*$_0$ agent than the rejected offer O'. For example, if the trading partner refuses an offer O' that assigns 4 blue chips to the focal agent, the focal agent decreases his belief that the trading partner will accept any offer that assigns 4 or more blue chips to the focal agent.

The belief update as a results of the trading partner rejecting an offer O' is represented by $U^-(b^{(0)}, O')$, which is defined as

$$U^-(b^{(0)}, O')(O'') = (1 - \lambda)^m \cdot b^{(0)}(O'') \qquad \text{with } m = |\{j | O'' \not<_j O'\}|. \qquad (3)$$

A.2 First-order theory of mind agent

The first-order theory of mind (*ToM*$_1$) agent considers the possibility that his trading partner has beliefs and goals, which determine whether or not the trading partner will accept an offer. Specifically, the *ToM*$_1$ agent is able to consider the perspective of his trading partner, and calculate her response to his offer as outlined in the previous subsection. However, the *ToM*$_1$ agent does not have access to the actual zero-order beliefs $b^{(0)}$ held by his trading partner or her goal location l_2. Instead, the ToM_1 agent forms beliefs about his trading partner's goal location and beliefs.

The goal location of the trading partner is represented as a probability distribution $p^{(1)} : L \rightarrow [0, 1]$, such that the ToM_1 agent believes that the likelihood of his trading partner having goal location $l \in L$ is $p^{(1)}(l)$. The ToM_1 agent also forms first-order beliefs $b^{(1)} : \mathcal{D} \rightarrow [0, 1]$, which specify that the agent believes that his trading partner assigns a subjective probability $b^{(1)}(O)$ to the event of the ToM_1 agent accepting a given offer $O \in \mathcal{D}$. However, although the ToM_1 agent explicitly models the zero-order beliefs of his trading partner through $b^{(1)}$, he does not attempt to model the trading partner's learning speed λ. Instead, whenever the ToM_1 agent updates his first-order beliefs $b^{(1)}$, he makes use of his own learning speed. This means that, unless the focal agent and his trading partner have the same learning speed, the focal agent will have an incorrect representation of his trading partner's beliefs.

When the ToM_1 agent receives an offer $O \in \mathcal{D}$ from his trading partner, he updates his zero-order beliefs as in Equation (1). Next, the ToM_1 agent updates his location beliefs. By putting himself in the position of his trading partner, the ToM_1 agent believes it to be impossible that his trading partner has a goal location $l \in L$ for which $\pi(l, O^*) \leq \pi(l, D^{0*})$. For other locations $l \in L$, the agent adjusts his beliefs based on the expected increase in the score of the trading partner if the offer O would be accepted, so that

$$p^{(1)}(l) := \begin{cases} 0 & \text{if } \pi(l, O^*) \leq \pi(l, D^{0*}) \\ \beta \cdot p^{(1)}(l) \cdot \frac{1+EV^{(0)}(O^*; b^{(1)}, l, D^{0*})}{1+\max_{\tilde{O} \in \mathcal{D}} EV^{(0)}(\tilde{O}; b^{(1)}, l, D^{0*})} & \text{otherwise,} \end{cases} \tag{4}$$

where β is a normalizing constant.

At the same time, the ToM_1 updates his confidence in first-order theory of mind c_1 to reflect how well the ToM_1 agent feels the first-order theory of mind model fits the behaviour of his trading partner. This is achieved through

$$c_1 := (1 - \lambda) \cdot c_1 + \lambda \cdot \sum_{l \in L} p^{(1)}(l) \cdot \frac{1 + EV^{(0)}(O^*; b^{(1)}, l, D^{0*})}{1 + \max_{\tilde{O} \in \mathcal{D}} EV^{(0)}(\tilde{O}; b^{(1)}, l, D^{0*})}.$$

Once the ToM_1 agent has updated his beliefs, he decides whether to accept the offer O made by his trading partner, to make an alternative offer of his own, or to withdraw from negotiations. Unlike the ToM_0 agent, the ToM_1 agent takes into account how his trading partner will react and change her behaviour as a result of his offer O'. The ToM_1 agent predicts his trading partner's behaviour by determining his own response from what he believes to be her perspective. According to the ToM_1 agent, the expected value of making an offer O', given that the trading partner's goal location is l, therefore becomes

$$\begin{aligned} & EV^{(1)}(O', l; b^{(1)}, l_1, D^0) = \\ & \qquad \begin{cases} -1 & \text{if } \mathit{ToM}_0(O'^*; b^*, l, D^{0*}) = \text{ withdraw,} \\ \pi(O', l_1) - \pi(D^0, l_1) & \text{if } \mathit{ToM}_0(O'^*; b^*, l, D^{0*}) = \text{ accept,} \\ \max\{\pi(O'', l_1) - \pi(D^0, l_1), -1\} & \text{otherwise,} \end{cases} \end{aligned}$$

where

$$O''^{*} = ToM_0(O'^{*}; b^{*}, l, D^{0*}), \text{ and} \tag{5}$$
$$b^{*} = U^{+}(U^{-}(b^{(1)}, O^{*}), O'^{*}). \tag{6}$$

The first term in this equality describes the situation when the ToM_1 agent believes that his trading partner will decide to withdraw from negotiations as a result of receiving offer O'. In this case, the ToM_1 agent suffers the cost of rejecting the trading partner's current offer O. The second term shows the situation in which the ToM_1 agent believes that his trading partner will accept his offer O'. Finally, the third term shows the situation in which the ToM_1 believes that his trading partner will reject his offer O', but counter with a new offer O''. Note that the ToM_1 agent takes into account that the behaviour of his trading partner may change as a result of the offer O' he makes himself.

Although a ToM_1 agent has access to first-order theory of mind to predict the behaviour of his trading partner, he may decide that these predictions are not better than zero-order theory of mind. The expected value that a ToM_1 agent assigns to making an offer O' consists of both his first-order and his zero-order beliefs, which are integrated to reflect the agent's confidence in first-order theory of mind c_1. The integrated expected value of a ToM_1 agent then becomes

$$EV^{(1)}(O'; b^{(1)}, l_1, D^0) = (1 - c_1) \cdot EV^{(0)}(O'; b^{(0)}, l_1, D^0) + c_1 \cdot \sum_{l \in L} p^{(1)}(l) \cdot EV^{(1)}(O', l; b^{(1)}, l_1, D^0). \tag{7}$$

Like a ToM_0 agent, a ToM_1 has three possible responses to a given offer O. First, the ToM_1 agent may decide to withdraw from negotiations and get zero score. Second, the ToM_1 agent can accept his trading partner's offer O, which increases the agent's score by $\pi(l_1, O) - \pi(l_1, D^0)$. Finally, the ToM_1 agent can choose to make an alternative offer O'. If the ToM_1 agent decides to do so, he selects the offer O' that maximizes the expected value as shown in Equation (7), and updates his first-order beliefs to reflect that he has rejected the offer O, and make a counter offer O'. That is,

$$b^{(1)} := U^{+}(U^{-}(b^{(1)}, O^{*}), O'^{*}).$$

The ToM_1 agent uses first-order theory of mind of mind in two ways. First, when receiving an offer O from his trading partner, the ToM_1 agent uses his theory of mind to learn the goal location of his trading partner. He does so by determining how consistent a possible goal is with his trading partner making the offer O. This information allows the ToM_1 agent to only make offers that are mutually beneficial. Secondly, the ToM_1 agent also takes into account how making an offer O' changes the beliefs and behaviour of his trading partner. In some cases, this may allow the ToM_1 agent to manipulate his trading partner into making an offer O'' that he is willing to accept. However, higher-order theory of mind is needed to deceive a trading partner, as we will show in Section A.3.

A.3 Higher orders of theory of mind agent

Agents that are able to use orders of theory of mind beyond the first can use this ability to attempt to manipulate the beliefs of lower orders of theory of mind to obtain an advantage. For example, a second-order theory of mind agent can use his understanding of first-order theory of mind agents to create an offer that signals his goal location to his trading partner as clearly as possible. Alternatively, the *ToM*$_2$ agent could adjust his offer to deceive his trading partner so that the trading partner believes that the goal location of the focal agent is elsewhere.

Each additional order of theory of mind allows an agent to consider an additional model of opponent behaviour. These models are constructed analogously to first-order theory of mind. Based on kth-order theory of mind, a *ToM*$_k$ agent specifies additional location beliefs $p^{(k)} : L \to [0, 1]$, offer acceptance beliefs $b^{(k)} : \mathcal{D} \to [0, 1]$, and a corresponding confidence c_k. Independent of beliefs of other orders of theory of mind, whenever he receives an offer O from his trading partner, a *ToM*$_k$ agent simultaneously updates his confidence in kth-order theory of mind and his location beliefs analogously to a *ToM*$_1$ agent. That is, when receiving an offer O from the trading partner, the *ToM*$_k$ agent updates his confidence in kth-order theory of mind to become

$$c_k := (1 - \lambda) \cdot c_k + \lambda \cdot \sum_{l \in L} p^{(k)}(l) \cdot \frac{1 + EV^{(k-1)}(O^*; b^{(k)}, l, D^{0*})}{1 + \max_{\tilde{O} \in \mathcal{D}} EV^{(k-1)}(\tilde{O}; b^{(k)}, l, D^{0*})},$$

and updates his location beliefs $p^{(k)}$ to become

$$p^{(k)}(l) := \begin{cases} 0 & \text{if } \pi(l, O^*) \leq \pi(l, D^{0*}) \\ \beta \cdot p^{(k)}(l) \cdot \frac{1 + EV^{(k-1)}(O^*; b^{(k)}, l, D^{0*})}{1 + \max_{\tilde{O} \in \mathcal{D}} EV^{(k-1)}(\tilde{O}; b^{(k)}, l, D^{0*})} & \text{otherwise,} \end{cases} \tag{8}$$

where β is a normalizing constant.

Based on these location beliefs, the *ToM*$_k$ agent formulates the expected value of making an offer O', given that his trading partner's goal location is l as

$$EV^{(k)}(O', l; b^{(k)}, l_1, D^0) = \\ \begin{cases} -1 & \text{if } \mathit{ToM}_{k-1}(O'^*; b^*, l, D^{0*}) = \text{ withdraw,} \\ \pi(O', l_1) - \pi(D^0, l_1) & \text{if } \mathit{ToM}_{k-1}(O'^*; b^*, l, D^{0*}) = \text{ accept,} \\ \max\{\pi(O'', l_1) - \pi(D^0, l_1), -1\} & \text{otherwise,} \end{cases}$$

where

$$O''^* = \mathit{ToM}_{k-1}(O'^*; b^*, l, D^{0*}), \text{ and} \tag{9}$$

$$b^* = U^+(U^-(b^{(k)}, O^*), O'^*). \tag{10}$$

These expected values are then integrated into the expected values of previous orders according to

$$EV^{(k)}(O'; b^{(k)}, l_1, D^0) = (1 - c_k) \cdot EV^{(k-1)}(O'; b^{(k-1)}, l_1, D^0) + \\ c_k \cdot \sum_{l \in L} p^{(k)}(l) \cdot EV^{(k)}(O', l; b^{(k)}, l_1, D^0). \tag{11}$$

This yields the kth-order theory of mind response function

$$\mathit{ToM}_k(O; b^{(k)}, l_1, D^0) = \begin{cases} O' & \text{if } EV^{(k)}(O'; b^{(k)}, l_1, D^0) > V \\ \text{accept} & \text{if } \pi(l_1, O') - \pi(l_1, D^0) \geq V \\ \text{withdraw} & \text{otherwise,} \end{cases} \tag{12}$$

where $V = \max\{EV^{(k)}(O'; b^{(k)}, l_1, D^0), \pi(l_1, O') - \pi(l_1, D^0)\}$.

Note that for each order of theory of mind, the agent considers an additional round of play. The ToM_1 agent only predicts what offer O' his trading partner may make if the agent were to make some offer O. A ToM_2 agent, on the other hand, believes that the trading partner may be a ToM_1 agent. The ToM_2 agent therefore takes into account what offer O'' his trading partner believes the ToM_2 agent to make in response to an offer O' that the trading partner may make if the ToM_2 agent were to make some offer O.

References

1. de Weerd, H., Verbrugge, R., Verheij, B.: How much does it help to know what she knows you know? An agent-based simulation study. Artif. Intell. **199–200** (2013) 67–92