# Memory-Augmented Theory of Mind Network

**Dung Nguyen, Phuoc Nguyen, Hung Le, Kien Do, Svetha Venkatesh, Truyen Tran**

Applied Artificial Intelligence Institute ($A^2I^2$), Deakin University, Geelong, Australia
{dung.nguyen,phuoc.nguyen,thai.le,k.do,svetha.venkatesh,truyen.tran}@deakin.edu.au

### Abstract

Social reasoning necessitates the capacity of theory of mind (ToM), the ability to contextualise and attribute mental states to others without having access to their internal cognitive structure. Recent machine learning approaches to ToM have demonstrated that we can train the observer to read the past and present behaviours of other agents and infer their beliefs (including false beliefs about things that no longer exist), goals, intentions and future actions. The challenges arise when the behavioural space is complex, demanding skilful space navigation for rapidly changing contexts for an extended period. We tackle the challenges by equipping the observer with novel neural memory mechanisms to encode, and hierarchical attention to selectively retrieve information about others. The memories allow rapid, selective querying of distal related past behaviours of others to deliberatively reason about their current mental state, beliefs and future behaviours. This results in `ToMMY`, a theory of mind model that learns to reason while making little assumptions about the underlying mental processes. We also construct a new suite of experiments to demonstrate that memories facilitate the learning process and achieve better theory of mind performance, especially for high-demand false-belief tasks that require inferring through multiple steps of changes.

## Introduction

Human social interactions necessitate a skill known as theory of mind (ToM) to infer the mental states of others without having access to their latent characteristics, internal states and computation processes. Instead, we can rely on social cues and past behaviours to construct *models* of others, thereby attributing mental states to them, for example, inferring their beliefs and intentions. The models need not perfectly match with the true hidden internal mental states but facilitate accurate social prediction and planning (Premack and Woodruff 1978; Gallese and Goldman 1998; Rusch et al. 2020; Langley et al. 2022).

Since often we can only have access to others' past behaviours and current observable context, it is plausible that we need memory to store and represent the past of others, to contextualise the present, to draw analogies between the present and the related past, and to reason about possibilities (Grant, Nematzadeh, and Griffiths 2017). Cognitive scientists have employed memory of structured representation of tasks to enable analogical reasoning in ToM, for example, to recognise false beliefs (Rabkina et al. 2017). Likewise, the work of (Nguyen and Gonzalez 2021) uses instance matching to model the human's ToM ability. This cognitive model assumes that the observer, who is constructing a model of the actor, also has access to rewards for each experience of the actor. These works rely on either domain knowledge to construct the task structure or information that can be inaccessible to the observer.

In this paper, we take an alternative road to *learn the memory mechanisms* to build the computational ToM capability into artificial social agents. Here mentalising and predicting behaviours of an actor in a partially observable environment are treated as a task to be learnt (Rabinowitz et al. 2018). In the learning phase, the ToM observer first acquires a general prior mental model from the observed behavioural episodes of training actors. In the execution phase, upon seeing an actor and its partial episode, the observer rapidly updates the specific posterior about the actor. Realising this strategy, we equip the observer with a new memory-augmented ToM network dubbed `ToMMY` (**T**heory **o**f **M**ind with **M**emor**Y**). Central to this architecture is the memory module, arranged as key-value pairs storing the past behaviours of the actor. The analogy-making capability is learned from training data, and once trained, it works by selectively querying relevant memory keys for a given context to retrieve corresponding predictive values. The retrieved values, combined with the selective few states of the current episode, constitute the posterior of the actor's mental state, which serves as an input for predicting its future behaviours. Learning memory-augmented neural networks is a powerful technique for multi-step reasoning (Sukhbaatar et al. 2015; Graves et al. 2016), handling rare events (Kaiser et al. 2017), meta-learning (Santoro et al. 2016) and rapid reinforcement learning (Le et al. 2021). However, little work has been done in the area of mentalising other agents in social settings.

To assess the performance of `ToMMY` we introduce a new *false-belief* test to evaluate the ToM ability under high reasoning demand. False-belief tasks determine if the actor is maintaining an outdated belief about something that no longer holds. A classic example is the Sally-Anne Test (Wimmer and Perner 1983; Baron-Cohen, Leslie, and Frith 1985), in

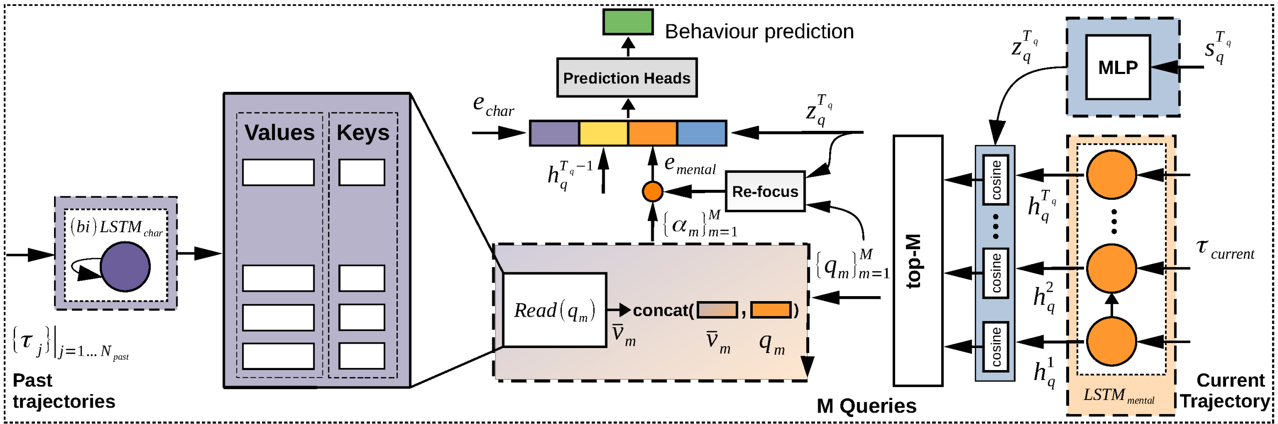



Figure 1: Architecture of memory-based theory of mind agent (ToMMY). Past trajectories of the actor are encoded into memory of (key, value) pairs. A selective set of events in current trajectory are embedded into mental states, which used to query the memory. The prediction heads then generate future prediction using the retrieved information, the current mental state and the world state.

which Anne secretly moves a toy out of the original box, causing Sally to falsely believe that the toy is still there. A version of the Sally-Anne Test for testing artificial theory of mind agents introduced in (Rabinowitz et al. 2018; Nguyen and Gonzalez 2021) take the form of grid-worlds. When the actor tries to achieve the sub-task before reaching a goal, the position of the goal will be changed, a so-called *swap event*. This event induces a false belief in the actor, and the observer needs to take the actor's perspective to understand the actor's false belief. This test is simple to some extent, requiring a low-demand mental capacity of social reasoning to pass. In our proposal, there are multiple steps in the trajectory of the agent with hidden information; hence the ToM observer needs to keep track of these steps in order to infer the actor's belief. More concretely, consider a robot helper: It knows more than the human in a particular area; however, it cannot observe all events in other areas where both (the robot and human) are not currently present. To pass our test, the ToM agents must refer to the states in the last visits to areas to understand whether the actor has a false belief and combine it with past behaviours to predict future behaviours. Therefore, in this high-demand false-belief task, the memory mechanism is the key to achievement. This matches with the prediction of developmental psychology that theory of mind in later development requires memory when humans need to flexibly execute belief reasoning and deal with a complex situation (Apperly and Butterfill 2009).

## Related Work

The ToMMY agent makes little assumption about the underlying mental structure of others. This differs from cognitive science works such as those in (Baker, Saxe, and Tenenbaum 2011; Baker et al. 2017), hypothesising that human will maximise their own utility when mentalising about others. In AI, ToM is traditionally studied in plan recognition (Geib and Goldman 2009; Sohrabi, Riabov, and Udrea 2016), assuming structural knowledge of the domain. Recent works have been studied in the NLP domain (Nematzadeh et al. 2018; Le, Boureau, and Nickel 2019; Arodi and Cheung 2021), focusing on second-order false-belief tasks. Instead, we take a *meta-learning* stand, similar to that in (Rabinowitz et al. 2018), updating the posterior of latent characters for unseen actors, starting from the prior learnt from seen actors. This work compresses the entire history of an actor into a vector representing its character, which is then combined with the current episode and state to predict goal, intention, action, and successor representations. Such compression rapidly forgets information from the far past, making it difficult to reason about rare and distantly separated situations. Our memory mechanisms effectively tackle this forgetting problem.

There are various methods to measure the ToM ability of humans in developmental psychology (Beaudoin et al. 2020), e.g. either direct assessment as Sally-Anne Test or indirect assessment via the violation of expectation (VoE) (Onishi and Baillargeon 2005), anticipatory looking (Clements and Perner 1994) and active helping (Knudsen and Liszkowski 2012). In (Gandhi et al. 2021) and (Shu et al. 2021), the authors used the VoE to evaluate the ToM ability. (Rabinowitz et al. 2018; Nguyen and Gonzalez 2021) constructed a test to mimic the Sally-Anne Test in artificial intelligence.

Our false-belief testbed is designed for more complex situations that demand long-term memory. Measuring the cognitive load of a task is an active field in cognitive science (Zheng 2017). Here, we intuitively add distractors to increase the cognitive load of the false-belief task. Measuring the difficulty of the task is a direction that requires investigation as it is necessary for assessing the ability of artificial agents.

## Problem Formulation

We study the general setting under the partially observable Markov decision processes (POMDPs) framework. The *observer* or *theory of mind (ToM) agent* first observes a set of $N_{past}$ *past trajectories* $\{\tau_j\}$ of an *actor* in multiple environments for $j = 1, \ldots, N_{past}$. Each past trajectory $\tau_j$ is a sequence of state-action pairs $\left(s_j^t, a_j^t\right)$ for $t = 1, 2, ..., T_j$. Upon seeing the current trajectory $\tau_q$ which is a sequence of $\left(s_q^t, a_q^t\right)$ from $t = 1$ up to time $T_q - 1$, and the current state $s_q^{T_q}$ at time $T_q$, the ToM agent will predict the goal (or preference) and the future behaviours of the actor, including the intention, the next action, and the future visit states of the actor (via successor representations (Dayan 1993)).

## Memory-Augmented Theory of Mind Network

We now describe `ToMMY` (**T**heory **o**f **M**ind with **M**emor**Y**), our theory of mind (ToM) agent, who observes an actor acting in a partially observable environment. `ToMMY` is implemented as a memory-augmented neural network as illustrated in Fig. 1. `ToMMY` maintains an episodic memory of the past trajectories of the actor, which consist of information reflecting the general latent character of the actor. From the current incomplete trajectory, `ToMMY` selectively chooses several events to serve as queries to retrieve relevant past events from memory. The retrieved events are then re-focused and combined with the queries, the character embedding, and the world state to form a *mental posterior attributed to the actor.*

### Character Embedding

To utilise all information presented in the history of the actor, we encode $\tau_j$ to forward hidden states ($\overrightarrow{h}_j^t$) and backward hidden states ($\overleftarrow{h}_j^t$) using the bidirectional long-short term memory (Bi-LSTM)

$$\overrightarrow{h}_j^t = \text{LSTM}_{\rightarrow}\left(\left[s_j^t, a_j^t\right], \overrightarrow{h}_j^{t-1}\right), \tag{1}$$

$$\overleftarrow{h}_j^t = \text{LSTM}_{\leftarrow}\left(\left[s_j^t, a_j^t\right], \overleftarrow{h}_j^{t+1}\right). \tag{2}$$

By this structure, at time step $t$, the observer can retain and reason about both the past and the future. Two sequences of forward and backward states are employed as augmented information for the value of the memory, as will be described in the next section.

Since the character of the actor is constructed by the history, we summarised the last hidden state of the forward LSTM over all past trajectories into the *character embedding* of the actor as

$$e_{char} = \frac{1}{N_{past}} \sum_{j=1}^{N_{past}} \text{ReLU}\Big(\text{MLP}\left(\overrightarrow{h}_j^{T_j}\right)\Big). \tag{3}$$

This character $e_{char}$ is supposed to govern the current behaviour of the actor. To realise this, we treat the character embedding as an additional input for another LSTM that processes the current unfinished trajectory as follows:

$$h_q^t = \text{LSTM}\left(\text{concat}\left(e_{char}, \left[s_q^t, a_q^t\right]\right), h_q^{t-1}\right). \tag{4}$$

These LSTM states will later serve as raw materials for attributing the mental states to the actor at each time step.

### Selective Attention

The information extracted from $N_{past}$ past trajectories are first stored in the key-value memory module $\mathcal{M} = \left\{\left(k_j^t, v_j^t\right)\right\}$ for $j = 1, \ldots, N_{past}$, and $t = 1, \ldots, T_j$. The memory *key* is a function of the forward state $k_j^t = g\left(\overrightarrow{h}_j^t\right)$. The associated *value* $v_j^t$ can be either (a) the forward state (which contains *actual* future information), or (b) the concatenation of the forward state and the backward state or other information computed from the rest of the trajectory from time $t+1$ (which contains *predictive* information of the future). Let's denote $\mathcal{M}.key = \left\{k_j^t\right\}_{j=1\ldots N_{past},}^{t=1\ldots T_j}$ and $\mathcal{M}.value = \left\{v_j^t\right\}_{j=1\ldots N_{past}}^{t=1\ldots T_j}$ is the set of all keys and the set of all values in the memory, respectively.

We then construct $M$ *queries* to read out from the memory based on selective events in the current unfinished trajectory. Let $z_q^{T_q} = \text{MLP}\left(s_q^{T_q}\right)$ be the embedding of the current world state. Given the set of hidden states $\mathcal{H} = \left\{h_q^t\right\}^{t=1\ldots T_q}$ of all observable events in the current trajectory computed in Eq. (4), we collect from this set $M$ selective events that are most similar to $z_q^{T_q}$ as queries $\{q_m\}$ for $m = 1...M$ by using cosine similarity, i.e. $\{q_m\}_{m=1\ldots M} = \left\{h_q^{t'} \mid h_q^{t'} \in \mathcal{H}, d_{z^{T_q} h_q^{t'}} \in \text{top-M}_{h_q^t \in \mathcal{H}}\left(d_{z^{T_q} h_q^t}\right)\right\}$ with $d_{z^{T_q} h_q^t} = \text{cosine}\left(z_q^{T_q}, h_q^t\right)$. Here, $\text{top-M}_{h_q^t \in \mathcal{H}}\left(d_{z^{T_q} h_q^t}\right)$ is a function that returns the set of $M$ highest values $d_{z^{T_q} h_q^t}$ given the set $\mathcal{H}$. The read head uses the queries in parallel to retrieve memory content as:

$$\bar{v}_m = \sum_{v_j^t \in \mathcal{M}.value} \text{attn}\left(q_m, k_j^t\right) v_j^t, \tag{5}$$

where $\text{attn}\left(q_m, k_j^t\right)$ is the soft attention score which is computed as

$$\text{attn}\left(q_m, k_j^t\right) = \frac{e^{d_{mjt}/\beta}}{\sum_{k_{j'}^{t'} \in \mathcal{M}.key} e^{d_{mj't'}/\beta}}, \tag{6}$$

with the temperature $\beta > 0$ and the distance $d_{mjt} = \text{cosine}\left(q_m, k_j^t\right)$.

### Mental Attribution

**Re-focusing on selective events** The `ToMMY` does not treat all the selective events equally, instead, it will re-weight these selective events based on the embedding of the current world state $z_q^{T_q}$ and $\{q_m\}$ via a set of *attention weights*

$$\alpha_m = \frac{e^{\delta_m/\beta}}{\sum_{m'=1\ldots M} e^{\delta_{m'}/\beta}}, \text{ for } m = 1\ldots M$$

where $\delta_m$ is a metric measuring the relationship between a recent event $z_q^{T_q}$ and the selective event $q_m$. The distance $\delta_m$ is generated by a neural network $\text{MLP}\left(\left[z_q^{T_q}, q_m\right]\right)$ which attempts to learn a metric to measure the importance of a selective event $q_m$ to the behaviour predictions made at the recent event. This mechanism captures the re-focusing process on a smaller and more selective set of events in the trajectory.

**Mental Posterior** Based on the attention weights, the retrieved memory contents, combined with the queries, constitute the current mental state of the actor:

$$e_{mental} = \frac{1}{M}\sum_{m=1}^{M} \alpha_m \text{concat}\left(\bar{v}_m, q_m\right).$$

This mental state, together with the last hidden state $h_q^{T_q-1}$, the character embedding $e_{char}$ in Eq. (3) and the representation of the current state $z_q^{T_q}$, form the *mental posterior* $\mathbf{e}_p = \text{concat}\left(e_{mental}, h_q^{T_q-1}, e_{char}, z_q^{T_q}\right)$ that serves as input for the prediction heads.

We use four prediction heads for predicting preference (or goal), one-step ahead intention, one-step ahead action, and the successor representations.

## Training

Our neural network is trained by minimising a combined loss for preference (or goal) prediction ($\mathcal{L}_{\text{pref}}$), intention prediction ($\mathcal{L}_{\text{intent}}$), action prediction ($\mathcal{L}_{\text{action}}$), and the successor representations prediction ($\mathcal{L}_{\text{SR}}$) as

$$\mathcal{L} = \mathcal{L}_{\text{pref}} + \mathcal{L}_{\text{intent}} + \mathcal{L}_{\text{action}} + \mathcal{L}_{\text{SR}}.$$

The first three component losses are negative log-likelihoods of the corresponding targets and are computed as follows:

$$\mathcal{L}_{\text{pref}} = \sum_{\text{pref}} -\log P\left(\text{pref}\,|\mathbf{e}_p\right),$$

$$\mathcal{L}_{\text{intent}} = \sum_{\text{intent}} -\log P\left(\text{intent}\,|\mathbf{e}_p\right),$$

$$\mathcal{L}_{\text{action}} = -\log P\left(a_t\,|\mathbf{e}_p\right)$$

where the $\mathbf{e}_p$ is the mental posterior. To compute the successor representation loss $\mathcal{L}_{\text{SR}}$, we first compute the empirical successor representation (SRs) by $SR_\gamma^{(t)}(s) = \frac{1}{Z_t}\sum_{t'=0}^{T-t}\gamma_{SR}^{t'} I\left(s_{t+t'}=s\right)$ where $T$ is the episode length, $t$ is the time at which the successor representation is computed, $\gamma_{SR} \in (0,1)$ is the discount factor, $Z_t = \sum_{s\in\mathcal{S}}\sum_{t'=0}^{T-t}\gamma_{SR}^{t'} I\left(s_{t+t'}=s\right)$ is a normalisation constant ($\mathcal{S}$ is the state space), and $I\left(s_{t+t'}=s\right)$ is an indicator function, which returns 1 if $s_{t+t'}=s$ and 0 otherwise. We then use the cross-entropy loss to obtain the SR loss

$$\mathcal{L}_{\text{SR}} = \sum_{\gamma_{SR}}\sum_{s} -SR_{\gamma_{SR}}^{(t)}(s)\log \widetilde{SR}_{\gamma_{SR}}^{(t)}(s).$$

By training, `ToMMY` learns a *prior model* of others from observing and predicting actors' behaviours, which is captured in network weights and the analogy-making capability. When mentalising about an actor, `ToMMY` updates the posterior upon seeing some of its behaviours.

# Experiment Results

We evaluate `ToMMY` on multiple tasks, including predicting preference, intention, action, and successor representations as well as assessing false belief understanding. For simplicity, we set the memory keys to the forward LSTM states ($k_j^t = \overrightarrow{h}_j^t$ of Eq. (1)). Similarly, the memory values are set to either the forward LSTM states ($v_j^t = \overrightarrow{h}_j^t$ of Eq. (1)), or the concatenation of both forward and backward LSTM states ($v_j^t = \left[\overrightarrow{h}_j^t, \overleftarrow{h}_j^t\right]$ of Eq. (1) and Eq. (2)). The latter is called `Bi-ToMMY`. The number of queries is set as $M = 10$.

In practice, some actions (such as `pick-up`) happen far less frequently than others in the whole sequence. Thus we use a replay buffer to balance the class of actions in training. As the relay buffer plays the role of episodic memory in the learning process, we call this balancing strategy action-based episodic memory (AEM). For comparison, we implemented a recent representative neural ToM network called ToMnet (Rabinowitz et al. 2018).

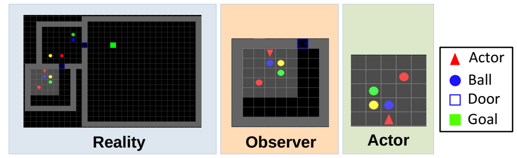



Figure 2: A multi-light-room environment with three adjoining rooms. The observer has a different perspective from the actor, and both can only partially observe reality.

## Light-Room Environment

To study ToM models, we created a multi-light-room environment using the *gym-minigrid* framework (Chevalier Boisvert, Willems, and Pal 2018) (see Fig. 2). The observer (ToM agent) can only see the lit room where the actor is in. After the actor leaves a room, the light will be turned off, and the observer will not see what happens in this room afterwards, e.g. swapping keys. As a result, the observer needs to memorise what happened in all rooms to read the current mind of the actor and to predict the actor's behaviours correctly.

We procedurally generated the actor's behaviours as follows. At each step, the actor chooses one amongst three intentions `find()`, `goto()`, `pickup()` and executes the intention by choosing between four primitive actions {`turn-left`, `turn-right`, `move-forward`, `pickup`}. To find an object, the actor first hypothesises an arbitrary position in the room and then walks to this position to verify. If the actor could not find the object, it will make another hypothesis. We call this type of actor a *hypo-actor*. After seeing any object, the actor will hold a belief about the position of this object. This belief can be changed if the actor recognises that the object no longer exists in the original position. Each actor can have a small field of view as $3 \times 3$, e.g. it can observe a square of $3 \times 3$ in front of it or has a larger field of view as $5 \times 5$ (the left-most figure of Fig. 2).

## Preference, Action and Intention Prediction

We carry out experiments where the observer uses its knowledge about the past and current trajectories of the actor to predict the actor's preference, actions and intentions. In these settings, the actor has its own preference for the ball and tries to navigate to collect the ball in its preferred colour in the first light-room. It will try to navigate to the goal located in

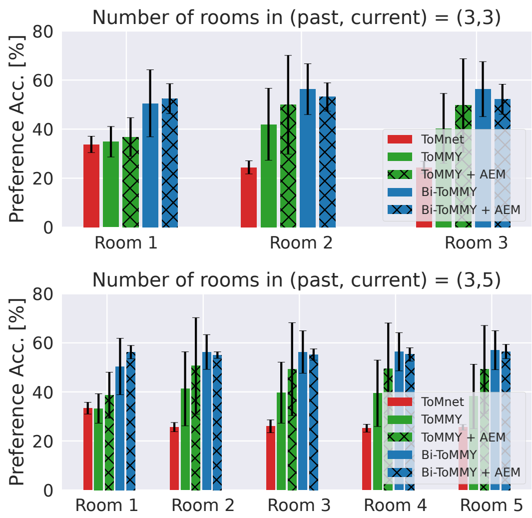



Figure 3: Preference prediction (mean and std.) of ToMnet and `ToMMY`s across rooms in two scenarios: (top) both current and past trajectories have three rooms; and (bottom) there are three and five rooms in the past and current trajectories, respectively. The performance is measured when the agent presents in each room during the current trajectory.

the final room. In other light-rooms between the first and the final room, balls of different colours exist. Since the actor has already picked up the preferred ball, it will no longer pick up other balls; therefore, these other balls can be considered as distractors to the observer. We trained ToM models in episodes with three light-rooms and test models under different conditions: (1) past trajectories are in three light-rooms, and the current trajectory is in three light-rooms; (2) past trajectories are in three light-rooms, and the current trajectory is in five light-rooms.

**Preference Prediction** Fig. 3 shows the performance of ToMnet (Rabinowitz et al. 2018) and `ToMMY`s on preference prediction tasks. The mean and standard deviation (std.) are computed over 4 runs for each model. Since ToMnet uses LSTM to compress the entire history of the actor into a single character embedding vector, it will struggle to remember details of the long past, which are critical to giving the correct answer in this situation. In our experiment, `ToMMY`s give better answers than the ToMnet during the episode. This is because `ToMMY`s effectively querying past trajectories by the memory mechanisms.

**Action and Intention Prediction** Fig. 4 shows the performance of ToM models on action and prediction tasks. Since `ToMMY`s can efficiently use past information, it can predict better over ToMnet when the actor changes its direction to look for an object. The models that used bidirectional long-short term memory to process the past trajectory (`Bi-ToMMY`) can improve the performance in action prediction when the actor changes its direction. In all settings, when the actor picks up an object–a rare event–only the methods augmented with the action-based episodic memory can learn to predict correctly.

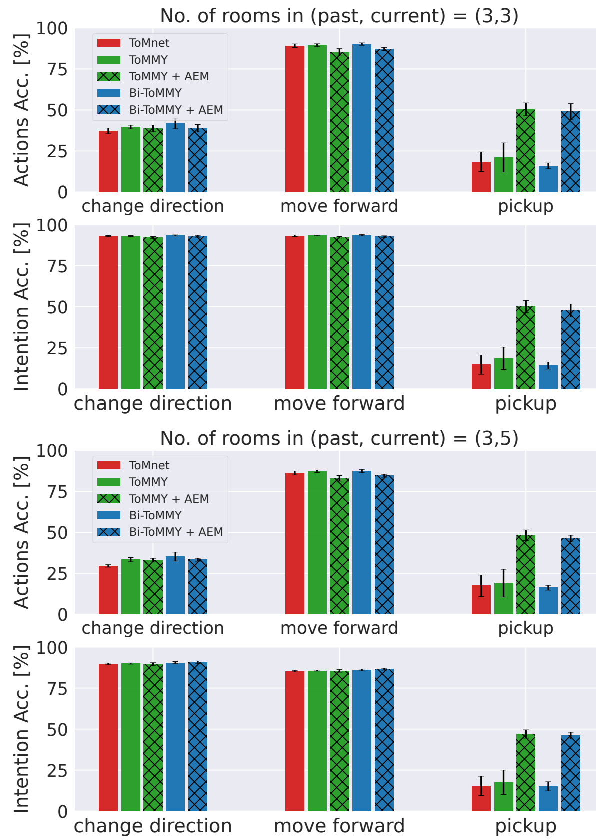



Figure 4: Performance of ToMnet and `ToMMY`s on the action and intention prediction (mean and std.). The x-axis shows three groups of actions: (1) change direction or (`turn-left` or `turn-right`), (2) `move-forward`, and (3) `pickup`.

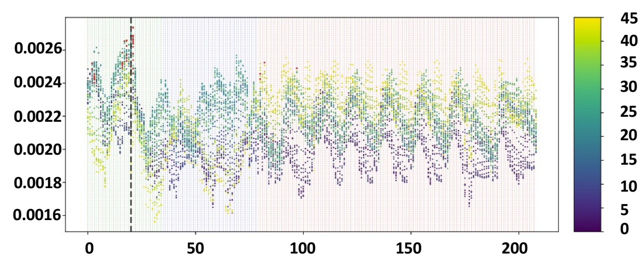


Figure 5: Visualisation of the reading weights from the beginning of the current trajectory up to the moment the actor picks up the ball (here is at step 45). The heatmap colour indicates the time in the current trajectory at which the query is conducted. The red dots are the peak weights of trajectories. The background colours indicate the rooms that the actor was in. Rooms 1, 2 and 3 are coded as green, blue, and red. The vertical black dash line indicates when the actor picked up the ball in the past trajectory. Our query mechanism generates the weights with high values when the actor is in room 1, especially when it picks up the ball, e.g. high peak at the vertical black dash line.

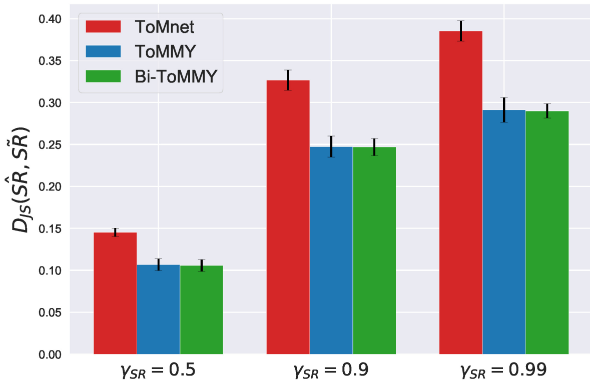



Figure 6: Performance of theory of mind agents (ToMnet, ToMMY and Bi-ToMMY) in high demand false-belief task measured by the Jensen-Shannon divergence (mean and std.) between the successor representations ($\gamma_{SR} = 0.5$, $\gamma_{SR} = 0.9$, $\gamma_{SR} = 0.99$) predicted by the models and the ground truth (the lower the better).

**Visualisation** Fig. 5 shows the attention weights (in Eq. (6)) over experiences in past trajectories. The attention weights generated by the network are relatively higher during the period when the actor was in the first room, especially when it picked up the ball. This means the ToM agent learnt to correctly attends to moments that express the actor's preference.

## High Demand False-belief Assessment

We constructed a three-light-room environment that demands a high mental load for theory of mind. This setting is inspired by the emerging human-machine teaming scenarios in which a ToM-equipped robot companion follows a human actor. This robot has the privilege of seeing more than what the human can see in the current room, but may not know what happens in the other room, e.g. the keys are swapped unseen by the observer. Hence, the robot needs to remember the past well to recognise human's false beliefs and then uses this information to provide proper help. A scenario is illustrated in Fig. 7.

In this setting, there are three rooms: (1) The **key room**, (2) the **ball room** as distractors, and (3) the **box room** where the goal is revealed. The actor has an initiated belief about the position of keys located in the **key room** (stage 1). The actor starts from the **key room**, passes by the **ball room** (stage 2), and arrives in the **box room**. Here the actor finds the box whose colour matches with that of the key which the actor should collect (stage 3). The actor then comes back to the **key room** with the goal is to collect the right key. When the actor passes by the **box room** the second time, the position of the keys can be changed without the actor knows. This induces false beliefs in the actor (stage 4). When the actor comes back to the **key room** the second time, we ask the observer to predict the successor representations of the actor (stage 5). At this stage, the observer does not have false beliefs since it can observe the current position of the keys. However, to know that the actor has false beliefs, the observer needs to refer to the position of the keys in the **key room** at the beginning of the episode. Also, to infer the actor's goal, the observer needs to recall the event in the **box room**.

In this test, the past trajectories do not provide any information about the goal or preference like in previous experiments but reveal the behaviour of the actors. For example, it would help the observer predict the distance the actor can see the object. The current trajectory contains information about the actor's goal in the middle of the trajectory. This setting challenges the ToM models without the ability to recall the key's position to predict the actor's behaviours.

**Results** ToMMYs predict more accurately the successor representations at the time the actor comes back to the key room, as shown in Fig. 6. At this moment, to answer correctly, the theory of mind agents must recall the position of keys at the beginning of the episode when the actor was at the key room the *first time* and the colour of the box in the box room. These important events are divided twice times by periods when the actor is in the distractor rooms. Failing to recall this information will lead to incorrect predictions.

**Visualisation** Fig. 8 shows that ToMMY attends to the period in the current trajectory that the actor is in the box room when the goal is revealed. It also refers to the beginning of the episode to recall the keys' original position to know whether keys were swapped. In case the weights according to top-M selective events are not highlighted on events where the actor was in the key room and the box room, the model is able to re-generate weights based on this small set of events. As shown in Figure 8.a, although the two highest weights which are generated by the cosine similarity metric are on important events when agents are in the box room and the key room, the re-focusing mechanism still gently corrects the attention of ToMMY by decreasing the weights of other events in the distractor room. Especially, when the attention over top-M selective events highly rises up in the distractor events, as in Figure 8.b, this mechanism is crucial to correct and help the theory of mind agent re-focuses on the important events. Hence, ToMMY understands whether the actor may have false beliefs.

## Conclusion

Aiming at equipping artificial agents with new social capacities we introduced ToMMY, a new neural theory of mind model that utilises the power of external memory and hierarchical attention for mentalising over complex behaviours of other agents in POMDPs settings. The memory facilitates meta-learning from prior experiences the analogy-making capability in social situations without the need of explicit domain knowledge or task structures. This capability is then refined when ToMMY sees an actor and its past and current behaviours. We also introduced a new high-demand false-belief task to assess the theory of mind ability to understand if others wrongly believe in things that no longer hold. Our experiments showed that memory facilitates the learning process and achieves better social understanding, especially in theory of mind tasks that demand a high cognitive load.

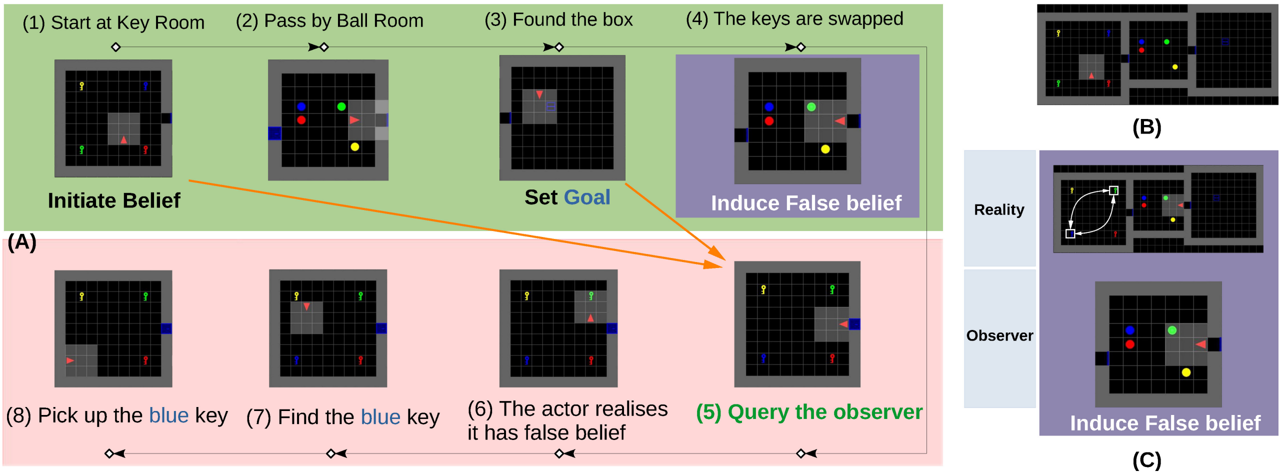
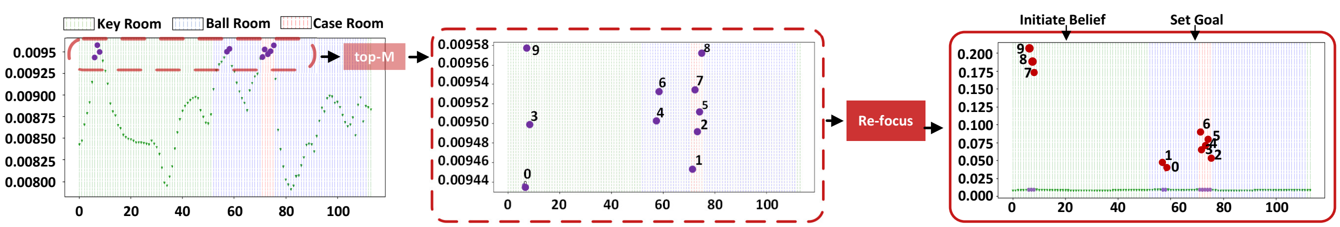
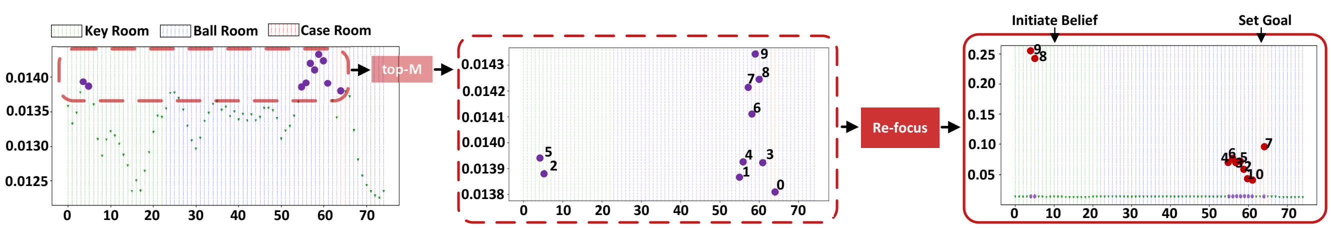



Figure 7: High demand false-belief task in light-room environment. (A) The sequence of observations given to the theory of mind agent (from stage 1 to 5) and the subsequence events in the key room when the keys are swapped (stage 6 to stage 8); (B) The full observation of the environment and the reality at the initial state; (C) The moment when keys are swapped. At stage 5, to predict the successor representations, the observer must remember the actor's goal that is revealed at stage 3 and recall the positions of the keys seen at the beginning of the episode/stage 1 (two orange arrows).




(a) Weak Correction




(b) Strong Correction

Figure 8: Each figure shows the attention weights over the current trajectory right after the actor returns to the key room. The key room, ball room, box room are coded as green, blue, and red, respectively. The purple circles indicate the top ten attention weights. The red circles indicate the top-M attention weights after re-focusing. The number over the circles shows the rank of the weights magnitude, i.e., the higher number, the more critical. When asked to predict future behaviour, the observer recalls when the actor was in the key room (to the original position of the keys) and box room (the actor reveals its goal). Figures in the left-most column are the attention weights generated by the cosine similarity. The middle column is a closer look at the top-M selective events $\{q_m\}_{m=1...M}$. The right-most column presented the weights over top-M selective events reshaped by the re-focusing mechanism $\{\alpha_m\}_{m=1...M}$. In the first case (a), the re-focusing mechanism only needs to provide weak correction since events that have the highest attention weights after the cosine similarity are events when the actor is in the box room and the key room. However, in the second case (b), the strong correction from the re-focusing mechanism is important.

## References


Apperly, I. A.; and Butterfill, S. A. 2009. Do humans have two systems to track beliefs and belief-like states? *Psychological review*, 116(4): 953.

Arodi, A.; and Cheung, J. C. K. 2021. Textual Time Travel: A Temporally Informed Approach to Theory of Mind. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4162–4172.

Baker, C.; Saxe, R.; and Tenenbaum, J. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society*, volume 33.

Baker, C. L.; Jara-Ettinger, J.; Saxe, R.; and Tenenbaum, J. B. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4): 1–10.

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

Beaudoin, C.; Leblanc, É.; Gagner, C.; and Beauchamp, M. H. 2020. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10: 2905.

Chevalier Boisvert, M.; Willems, L.; and Pal, S. 2018. Minimalistic Gridworld Environment for OpenAI Gym. https://github.com/maximecb/gym-minigrid. Accessed: 2022-03-08.

Clements, W. A.; and Perner, J. 1994. Implicit understanding of belief. *Cognitive development*, 9(4): 377–395.

Dayan, P. 1993. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4): 613–624.

Gallese, V.; and Goldman, A. 1998. Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12): 493–501.

Gandhi, K.; Stojnic, G.; Lake, B. M.; and Dillon, M. R. 2021. Baby Intuitions Benchmark (BIB): Discerning the goals, preferences, and actions of others. *Advances in Neural Information Processing Systems*, 34: 9963–9976.

Geib, C. W.; and Goldman, R. P. 2009. A probabilistic plan recognition algorithm based on plan tree grammars. *Artificial Intelligence*, 173(11): 1101–1132.

Grant, E.; Nematzadeh, A.; and Griffiths, T. L. 2017. How Can Memory-Augmented Neural Networks Pass a False-Belief Task? In *CogSci*.

Graves, A.; Wayne, G.; Reynolds, M.; Harley, T.; Danihelka, I.; Grabska-Barwińska, A.; Colmenarejo, S. G.; Grefenstette, E.; Ramalho, T.; Agapiou, J.; et al. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626): 471–476.

Kaiser, Ł.; Nachum, O.; Roy, A.; and Bengio, S. 2017. Learning to remember rare events. *ICLR*.

Knudsen, B.; and Liszkowski, U. 2012. 18-month-olds predict specific action mistakes through attribution of false belief, not ignorance, and intervene accordingly. *Infancy*, 17(6): 672–691.

Langley, C.; Cirstea, B. I.; Cuzzolin, F.; and Sahakian, B. J. 2022. Theory of Mind and Preference Learning at the Interface of Cognitive Science, Neuroscience, and AI: A Review. *Frontiers in Artificial Intelligence*, 62.

Le, H.; George, T. K.; Abdolshah, M.; Tran, T.; and Venkatesh, S. 2021. Model-based episodic memory induces dynamic hybrid controls. *NeurIPS*.

Le, M.; Boureau, Y.-L.; and Nickel, M. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of EMNLP-IJCNLP*, 5872–5877.

Nematzadeh, A.; Burns, K.; Grant, E.; Gopnik, A.; and Griffiths, T. L. 2018. Evaluating Theory of Mind in Question Answering. In *EMNLP*.

Nguyen, T. N.; and Gonzalez, C. 2021. Theory of Mind From Observation in Cognitive Models and Humans. *Topics in Cognitive Science*.

Onishi, K. H.; and Baillargeon, R. 2005. Do 15-month-old infants understand false beliefs? *Science*, 308(5719): 255–258.

Premack, D.; and Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4): 515–526.

Rabinowitz, N.; Perbet, F.; Song, F.; Zhang, C.; Eslami, S. A.; and Botvinick, M. 2018. Machine theory of mind. In *International conference on machine learning*, 4218–4227. PMLR.

Rabkina, I.; McFate, C.; Forbus, K. D.; and Hoyos, C. 2017. Towards a computational analogical theory of mind. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*.

Rusch, T.; Steixner-Kumar, S.; Doshi, P.; Spezio, M.; and Gläscher, J. 2020. Theory of mind and decision science: towards a typology of tasks and computational models. *Neuropsychologia*, 146: 107488.

Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; and Lillicrap, T. 2016. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, 1842–1850.

Shu, T.; Bhandwaldar, A.; Gan, C.; Smith, K.; Liu, S.; Gutfreund, D.; Spelke, E.; Tenenbaum, J.; and Ullman, T. 2021. Agent: A benchmark for core psychological reasoning. In *International Conference on Machine Learning*, 9614–9625. PMLR.

Sohrabi, S.; Riabov, A. V.; and Udrea, O. 2016. Plan Recognition as Planning Revisited. In *IJCAI*, 3258–3264. New York, NY.

Sukhbaatar, S.; Weston, J.; Fergus, R.; et al. 2015. End-to-end memory networks. *Advances in neural information processing systems*, 28.

Wimmer, H.; and Perner, J. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young childrenś understanding of deception. *Cognition*, 13(1): 103–128.

Zheng, R. Z. 2017. *Cognitive load measurement and application*. Routledge.