How Can Memory-Augmented Neural Networks Pass a False-Belief Task?

Erin Grant, **Aida Nematzadeh**, and **Thomas L. Griffiths**
University of California, Berkeley
{eringrant, nematzadeh, tom_griffiths}@berkeley.edu

Abstract

A question-answering system needs to be able to reason about unobserved causes in order to answer questions of the sort that people face in everyday conversations. Recent neural network models that incorporate explicit memory and attention mechanisms have taken steps towards this capability. However, these models have not been tested in scenarios for which reasoning about the unobservable mental states of other agents is necessary to answer a question. We propose a new set of tasks inspired by the well-known *false-belief* test to examine how a recent question-answering model performs in situations that require reasoning about latent mental states. We find that the model is only successful when the training and test data bear substantial similarity, as it memorizes how to answer specific questions and cannot reason about the causal relationship between actions and latent mental states. We introduce an extension to the model that explicitly simulates the mental representations of different participants in a reasoning task, and show that this capacity increases the model's performance on our theory of mind test.

Keywords: language understanding, question answering, theory of mind, *false-belief* test

Introduction

Question answering poses difficulties to artificial intelligence systems because correctly answering a query often requires sophisticated reasoning and language understanding capacities, and so simply memorizing the answer or searching in a knowledge base is not enough. Despite this challenge, recent neural network models that make use of attention mechanisms in combination with an explicit external memory can successfully answer questions that require more complex forms of reasoning than before (*e.g.,* Sukhbaatar, Weston, Fergus, et al., 2015; Henaff, Weston, Szlam, Bordes, & LeCun, 2017). The benchmark dataset for such tasks has become the Facebook bAbi dataset (henceforth, bAbi) (Weston, Bordes, Chopra, & Mikolov, 2016), which is a collection of question-answering tasks in the form of simple narrative episodes – termed *stories* – that are accompanied by questions about the state of the world described in the stories. (See Figure 1 for an example story from this dataset.)

Although bAbi is a start towards enumerating the requirements for human-like reasoning capabilities, it lacks tasks for testing the ability to reason about mental states, which is also necessary for correctly answering questions of the sort that humans encounter regularly. Consider the following:

Sally and Ann are in the kitchen.
Sally placed the milk in the pantry.
Sally exited the kitchen.
Ann moved the milk to the fridge.

For a model to correctly answer questions such as *Where would Sally/Ann search for the milk?* it need not only recognize that Sally and Ann have mental representations of the state of the world but also that these representations are inconsistent: Sally believes that the milk is in the pantry while Ann thinks it is in the fridge.

Psychologists have used a task similar to this scenario – termed the *false-belief* task – to examine children's development of *theory of mind*: the capacity to reason about the mental states of oneself and others (Premack & Woodruff, 1978). Most 3-year-old children, after observing such a scenario, answer that Sally would search for the milk in the fridge because they cannot infer Sally's belief about the location of the milk, which is inconsistent with their own knowledge (*e.g.,* Baron-Cohen, 1989; Baron-Cohen, Leslie, & Frith, 1985). However, most older children are able to identify, correctly, that Sally's belief is different from theirs in that she thinks that the milk is the pantry.

To answer questions about situations like those that occur in a *false-belief* task, a model needs to use the observed actions in the scenario to infer the mental states of Sally and Ann. In this work, we investigate whether the End-to-End Memory Network (henceforth MemN2N), a recent neural question-answering model (Sukhbaatar et al., 2015) that solves most of the bAbi tasks, is able to answer questions of the same structure as a *false-belief* task. We formulate scenarios to capture different possible causal relations among actions and beliefs, and examine the performance of the model therein. We find that the MemN2N model performs well only in the presence of strong supervision – when the training and test data share the same casual structure. This result suggests that the model is able to memorize the training data but is unable to learn to reason about mental states and how they cause and are caused by actions.

Furthermore, to simulate the (perhaps inconsistent) beliefs of the participants in a story, we extend the MemN2N model to include a separate memory representation for each participant. We show that this extension improves model performance, suggesting that explicitly modeling agents' knowledge in a disentangled manner is in part sufficient for more human-like reasoning on a *false-belief* task.

Theory of Mind and the *False-Belief* Task

A theory of mind is integral for an agent to predict and explain the behavior that is caused by the mental representations of other agents, and therefore succeed on tasks such as the *false-belief* task. For children, this capacity is acquired gradually over the course of development. In particular, children undergo several milestones before they develop an adult-like theory of mind: By age two, they can distinguish between external states of the world and internal mental states possessed by cognitive agents (*e.g.,* Meltzoff, Gopnik, & Repacholi,

Mary got the milk there.
John moved to the bedroom.
Sandra went back to the kitchen.
Mary traveled to the hallway.
Q: Where is the milk? **A:** hallway

Figure 1: An example task from the bAbi dataset.

1999). By age four, they can distinguish between consistent and inconsistent mental states (*e.g.,* Perner, Leekam, & Wimmer, 1987), which allows them to identify a false belief.

Previous computational works have modeled human performance on the *false-belief* task. Some focus on modeling the development of theory of mind by instantiating a model that initially fails but eventually passes the *false-belief* test (Van Overwalle, 2010), while others study the settings in which a model can succeed on the task by varying the input data or the model architecture (O'Laughlin & Thagard, 2000; Triona, Masnick, & Morris, 2002; Goodman et al., 2006). However, none of these models use natural language sentences, despite the fact that the psychological *false-belief* task is usually administered verbally in the form of a natural language reasoning problem.

Furthermore, natural language is known to interact with the development of theory of mind. For example, use of mental state terms in child-directed speech (*e.g.,* Slaughter & Gopnik, 1996), engagement in pretend play (Youngblade & Dunn, 1995), storybook reading (Rosnay & Hughes, 2006), and reflection on events in the child's past (Nelson, 2007) serve to accelerate its developments, while, in turn, a greater grasp of theory of mind leads to increased linguistic ability (Milligan, Astington, & Dack, 2007). In this work, we examine whether a model can learn from natural language about the causal relationship between actions and beliefs, in order to be able to answer questions that require reasoning about mental states.

Memory Networks

The MemN2N model of Sukhbaatar et al. (2015) comprises an external memory cache and mechanisms to read and write to this memory. The model is trained to write a sequence of stories into its external memory and to answer questions about the stories by reading its memory and emitting the correct vocabulary item. At test time, the model is evaluated by the extent to which it can correctly answer questions about a held-out set of test stories.

Formally, the model ingests a sequence of input sentences $(x_1, \ldots, x_n)$ and produces, for each input item x_i, both a memory representation m_i and a context representation c_i, which are stored in memory. The model is then presented with a question q_k about the story, for which it produces an internal representation u_k. To answer the question, the model computes a normalized association score p_{ik} between the question representation and each of its stored memory representations:

$$p_{ik} = \frac{\exp\left\{u_k^{\mathrm{T}} m_i\right\}}{\sum_j \exp\left\{u_k^{\mathrm{T}} m_j\right\}}. \tag{1}$$

This weight can be interpreted as an attention mechanism that defines where in memory the model will look for information relevant to the given question.

The model then produces an output representation by way of a linear combination of its context representations, weighted by the attention computed in Equation (1):

$$o_k = \sum_i p_{ik} c_i. \tag{2}$$

The output representation is combined with the query representation and decoded by some function f to produce the predicted answer $\hat{a}$:

$$\hat{a} = f(o_k + u_k). \tag{3}$$

Learning model parameters at training time is done by way of stochastic gradient descent in cross entropy error.

Simulation 1: MemN2N Model

We evaluate the model introduced in the previous section on a set of novel textual reasoning tasks inspired by the *false-belief* task. Our tasks take the form of a sequence of natural language sentences – termed a *story* – and an associated question about the story.

Since we aim to create tasks that, for humans to solve, involve reasoning about other agents' beliefs, we design various story templates that simulate how different actions give rise to different beliefs, and conversely how different beliefs result in different actions. These stories differ in whether or not the agent who is the subject of the question has observed a change in the state of the world (*i.e.,* the agent has a true belief), or has not (*i.e.,* has a false belief). The stories further differ in whether the belief is observable (*i.e.,* the story explicitly contains sentences such as *Sally believes the milk is in the pantry*) or whether only actions are observable. When the agent harbors a false belief, and the model is asked to predict the action of the agent without explicit reference to the beliefs of the agent in the story, we recover a simulation of the classic *false-belief* task.

With this experimental design, we aim to determine whether the MemN2N model can reason about how actions cause beliefs and vice versa, and how much information needs to be revealed to enable the model to succeed.

Data Generation

To generate stories and corresponding questions, we emulate the bAbi (Weston et al., 2016) dataset generation procedure. We define a world of *entities*, which are the people and objects described in the stories, and possible *predicates* that take entities as subject and, optionally, object. Each entity has *properties* that define the predicates of which it can be subject or object. For example, a world may contain *Sally* with

	BA	AB	A(B)A
True Belief	Anne moved the milk to the fridge. Sally believes the milk is in the fridge. **Q:** Where did Sally search for the milk? **A:** fridge	Sally placed the milk in the pantry. Anne moved the milk to the fridge. **Q:** Where does Sally believe the milk is? **A:** fridge	Sally placed the milk in the pantry. Anne moved the milk to the fridge. **Q:** Where did Sally search for the milk? **A:** fridge
False Belief	Sally believes the milk is in the pantry. Sally exited the kitchen. Anne moved the milk to the fridge. Sally entered the kitchen. **Q:** Where did Sally search for the milk? **A:** pantry	Sally placed the milk in the pantry. Sally exited the kitchen. Anne moved the milk to the fridge. Sally entered the kitchen. **Q:** Where does Sally believe the milk is? **A:** pantry	Sally placed the milk in the pantry. Sally exited the kitchen. Anne moved the milk to the fridge. Sally entered the kitchen. **Q:** Where did Sally search for the milk? **A:** pantry

Figure 2: Examples of the training data, with the predicates of interest underlined. Note that the *true-belief* and *false-belief* test tasks are of the same form as the top and bottom items, respectively, in the last column.

the property *is agent* and *apple* with the property *is object*. Our rules permit *Sally* to perform the action *displace* on the *apple*.

In this work, we consider a restricted set of *action* and *belief* predicates. Our actions define simple interactions of an agent with the world (*e.g., place, move, enter, exit*) and our beliefs correspond to mental state terms (*e.g., believe, think*), inspired by the terms that children gradually learn to understand and use correctly over the course of development (*e.g.,* Bretherton & Beeghly, 1982; Johnson & Wellman, 1980). Our templates manipulate the order of action and belief predicates to test how the model reasons about the causal relations between them.

Experimental Conditions

Story Template We define a set of templates that correspond to the type of story that we wish to generate. Each template fixes a sequence of predicates and therefore puts constraints on the entities that may fill the template. For example, a template could be the sequence (*drop*, *pick up*, *exit*). Completion of the template entails sampling valid entities from the world to fill the subject and object positions of the predicates, producing, for example, the story (*Sally dropped the ball*, *Sally picked up the ball*, *Sally exited the room*).

We consider three different template types:

- **BA**: observable beliefs (*e.g., Sally believes the milk is in the pantry*) give rise to observable actions (*e.g., Sally searches the pantry*);
- **AB**: observable actions (*e.g., Sally places the milk in the pantry*) give rise to observable beliefs (*e.g., Sally believes milk is in the pantry*); and
- **A(B)A**: observable actions (*e.g., Sally places the milk in the pantry*) give rise to observable actions (*e.g., Sally searches the pantry*) by way of unobserved beliefs (*e.g., Sally believes the milk is in the pantry*).

Note that the **AB** and **A(B)A** conditions are different in that in **AB**, the question explicitly asks about Sally's belief; in **A(B)A**, on the other hand, the question is about Sally's action, which has been brought about by Sally's unobserved belief.

True vs. False Belief In addition to the type of template, for each story we manipulate whether the agent about whom the question is asked (*i.e.,* Sally) has a true belief or a false belief about the state of the world. In the case that the agent has a true belief, the agent observes all changes in the state of the world and thus their beliefs are consistent with the world. On the other hand, in the case that the agent has a false belief, the agent does not observe one or more changes in the state of the world (because, for instance, Sally may exit the room), and thus has a belief that is inconsistent with the world.

Training Conditions We have six possible story types as a results of crossing the template types with the true and false belief story types; examples of each of the story types are given in Figure 2. We sample from these story types to produce our training conditions, in the following manner:

- When the training condition is such that $p(\text{false belief}) = 0$ or 1, we sample only stories with true beliefs or false beliefs, respectively, and when $p(\text{false belief}) = 0.5$, we sample half of our stories with true beliefs and half with false beliefs.
- We sample stories from five different possible groups of templates: **BA**, **AB**, **AB+BA**, **A(B)A**and **AB+BA+A(B)A**.

The **AB+BA** and **AB+BA+A(B)A** conditions provide the model with training data that better approximates the variety of possible scenarios in the world. In these cases, the model observes more ways in which actions and beliefs interact, and thus we would expect it to be able to better generalize to new scenarios. Moreover, **AB+BA** provides the model with the opportunity to learn transitive inference – given that an action (*e.g.,* placing milk in the pantry) results in a belief (*e.g.,* the milk is in the pantry), and a belief (*e.g.,* the milk is in the pantry) can cause an action (*e.g.,* searching for milk in the pantry), a model that reasons about actions and beliefs could learn that an action (*e.g.,* searching for milk in the pantry) is

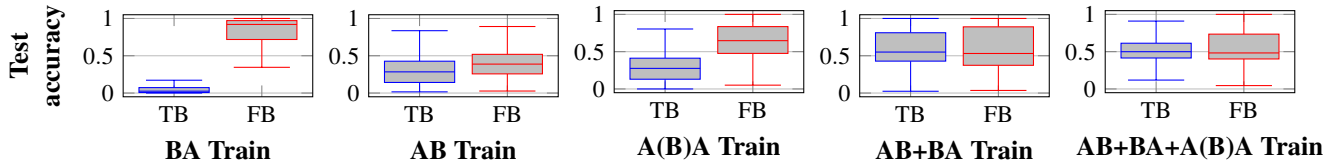

Figure 3: **Accuracy in Simulation 1.** Test accuracies for the *true-belief* (TB) and *false-belief* (FB) tests across training conditions in Simulation 1. We report results for $p(\text{false belief}) = 0.5$, since varying this parameter did not affect results except in the few cases discussed in the text.

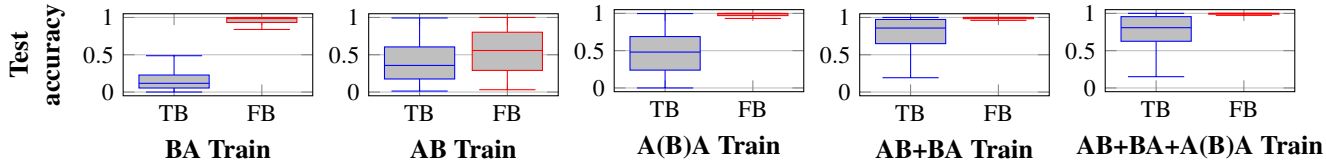

Figure 4: **Accuracy in Simulation 2.** Test accuracies for the *true-belief* (TB) and *false-belief* (FB) tests across training conditions in Simulation 2. As in Figure 3, we report results only for $p(\text{false belief}) = 0.5$.

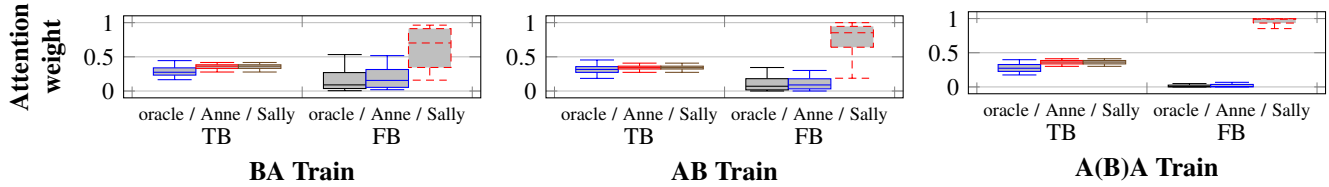

Figure 5: **Attention in Simuation 2.** Visualisation of the attention weighting over memory caches for the *true-belief* (TB) and *false-belief* (FB) tests. We omit the visualization for the **BA+AB** and **BA+AB+A(B)A** training conditions because the test accuracy distribution in Simulation 2 for these conditions is very similar to the **A(B)A** training condition (see Figure 4).

a consequence of an unobservable belief brought about by a preceding action (*e.g.,* placing milk in the pantry).

Crossing template types **BA**, **AB**, **A(B)A**, **AB+BA**, **AB+BA+A(B)A** with $p(\text{false belief}) = \{0.0, 0.5, 1.0\}$ produces our 15 training conditions. We run 10 simulations for each training condition and for each configuration of parameter settings of the MemN2N model.[1]

Test Conditions We aim to evaluate the model on tasks that require reasoning about latent mental states, in analogy to the classic *false-belief* task; however, such a capacity should apply not only in cases when an agent has a belief that is inconsistent with the state of the world (*i.e.,* a false belief) but also when they have a true belief about the world. We therefore consider two test conditions: a *true-belief* (TB) and a *false-belief* (FB) task. All examples in both of these test conditions share the **A(B)A** template type, but the conditions differ in that the *true-belief* task contains only examples with true beliefs (*i.e.,* $p(\text{false belief}) = 0$), and the *false-belief* task contains only false belief examples (*i.e.,* $p(\text{false belief}) = 1$).

[1]We vary the dimensionality of the memory and word embedding, the number of *computational hops* (accesses to the memory cache to answer a single question), the number of training and testing examples (1000 vs. 10000), and the size of the world from which the dataset of stories is generated (5 vs. 10 vs. 30 entities per entity type, which correspond to the objects, container, etc. in the story).

Results

As noted by Sukhbaatar et al. (2015), the MemN2N model exhibits large variance in performance across simulations, and so we show performance by plotting the distribution of test accuracies in boxplot format. In Figure 3, we report accuracy on both test conditions (the *true-belief* (TB) and *false-belief* (FB) tasks) across the training conditions, for $p(\text{false belief}) = 0.5$. The results for $p(\text{false belief}) \in \{0, 1\}$ were similar except in the case of the **AB** story template; we compare this case with the **BA** condition in Figure 6 and discuss in the following. Note that success at test time corresponds to achieving 1.0 accuracy in **both** the TB and FB test conditions.

Training Condition BA: Beliefs to Actions The model fails on the TB task in the **BA** training condition, while succeeding on the FB task. This is true no matter the value of $p(\text{false belief})$ (as depicted in Figure 6). To understand why this occurs, consider the following example of a **BA** training story when the false belief occurs:

Sally believes the milk is in the pantry. Sally exited the kitchen. Anne moved the milk to the fridge. Sally entered the kitchen.

Additionally, consider the **BA** training story when the false belief does not occur:

Anne moved the milk to the fridge. Sally believes the milk is in the fridge.

To answer the training question *Where did Sally search for the milk?* the model seems to learn that it should look for the sentence containing *Sally* and a container entity (*i.e., Sally believes the milk is in the fridge*).

This strategy works for the *false-belief* test (see Figure 2, last column, bottom row), because Sally believes that the milk is in the pantry – the location in which she originally placed it – and thus the sentence containing *Sally* and the identity of a container always proviedes the correct answer. However, this strategy fails on the *true-belief* test (again, see Figure 2, last column, top row), because Sally observes that the milk has been moved, and so no longer believes that the milk is in fridge. This suggests that the model is unable to infer that an observable action changes the mental state of Sally.

Training Condition AB: Actions to Beliefs The model is unable to achieve good performance on both the TB and FB tests in the **AB** condition. When the model performs better, it is in cases where the test is very similar to the training condition, *i.e.,* the *false-belief* test with $p(\text{false belief}) = 1$ in training and *true-belief* test with $p(\text{false belief}) = 0$ in training.

Training Condition AB+BA: Transitive Inference The model fails on both test tasks in the **AB+BA** training condition. This is evidence that the model cannot reason about the causal relationships between actions and beliefs to perform transitive inference.

Training Condition A(B)A: Equivalent to TB/FB Test The model achieves best performance on **A(B)A** in the $p(\text{false belief}) = 0.5$ condition. This again happens because the test and training conditions are similar: the model observes examples of both the FB and TB test tasks in training, and thus receives supervision to give the correct answer at test. However, the model performs well only on the TB task in the $p(\text{false belief}) = 0$, and on the FB task in the $p(\text{false belief}) = 1$ condition. This is because the model does not observe examples like one or the other test condition at training time.

Notably, the performance is not high even in the $p(\text{false belief}) = 0.5$ condition (the median is approximately 55% on both test tasks), despite the fact that the model is given test-like examples at training time. It is therefore not clear that the model is robustly able to solve a conditional reasoning task in which the correct answer is dependent on whether or not the observer sees the movement of the object and thus has a false or true belief. This, along with the model's failure in the other training scenarios, motivates an extension to the model, which we consider in the next section.

Simulation 2: Multiple-Observer Model

We now propose a model that is given information about whether each agent in the story observes each sentence in the story. In general, this must also be inferred from context, but here we assume such annotations are available to the model as we simply attempt to investigate the effect of this information on the model's predictions.

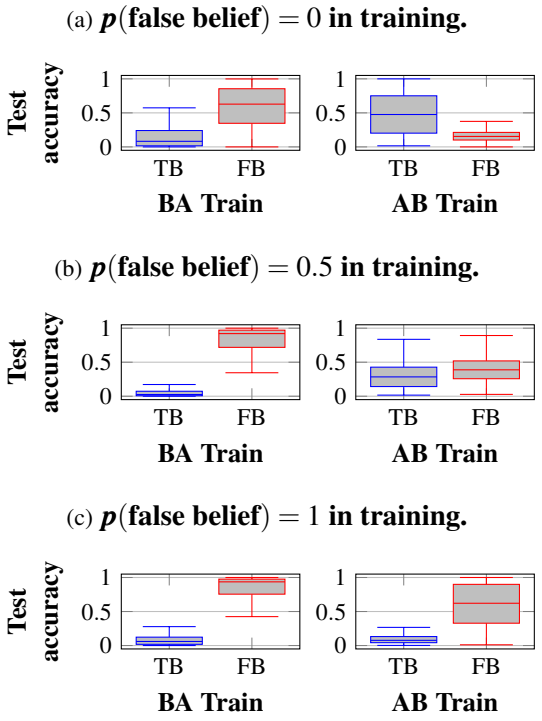

Figure 6: From **Simulation 1**. The test accuracy in the **AB** condition is dependent on the value of $p(\text{false belief})$, but not in the **BA** condition.

Formally, for a story of N input items that describes a situation with M agents, we provide the model with an N-by-$(M+1)$ *observer annotation matrix* S such that $S_{ij} = 1$ if input item x_i is observable to agent j and 0 otherwise, where we assign the oracle observer (who observes all input items) to the first index. These annotations are used to mask the input such that $M+1$ (possibly different) stories are produced, each of which corresponds to the story that a particular agent observes. Memory representations, attention over each memory cache, and output representations are computed separately for each observer, and so $M+1$ output representations are computed, each corresponding to the output of a distinct observer's memory.

The model then computes an attention weighting over each of the observer memory caches (*cf.* Equation (1)):

$$r_{k\ell} = \frac{\exp\left\{u_k^{\mathrm{T}} o_{k\ell}\right\}}{\sum_n \exp\left\{u_k^{\mathrm{T}} o_{kn}\right\}}. \tag{4}$$

This attention over memory caches is used to compute a weighted combination of the output representations that correspond to the memory cache for each agent (*cf.* Equation (3)):

$$\hat{a} = f(u_k + \sum_{\ell} r_{k\ell} o_{k\ell}). \tag{5}$$

Note that the model considered in Simulation 1 is exactly this model extension with $r_{k0} = 1$ and $r_{km} = 0, \forall m \neq 0$ (*i.e.,* attention is given only to the oracle memory cache).

In this extension, the model is given explicit information about which observations in a story are available to each agent, by way of the annotation matrix S. However, it must *learn* to reason about this information in order to arrive at the correct answer, as before with how to write to memory and read from memory, and now with how to select over which observer's knowledge of the story is relevant to answer the question.

Results

We report results of the model extension on the TB and FB tests in Figure 4, as well as a visualization of the attention weights in Figure 5. Our simulated data is composed of scenarios with only two agents, and therefore the extended model attends over three memory caches (one for the oracle that observes everything, one for Anne, and one for Sally, about whom the question is asked).

The extended model achieves higher accuracy across all training conditions. Notably, the model performs near perfectly (*i.e.,* both TB and FB are close to 1) in the **AB+BA+A(B)A** case, meaning that the model can learn to ignore irrelevant training stimuli. This suggests that awareness of agent's knowledge about the state of the world helps in a task of reasoning about latent mental states.

Furthermore, the attention plots show that the model learns to attend to the memory representation of Sally in the FB test, which contains the information about how to answer questions about Sally's actions and beliefs. On the other hand, in the TB test, the model does not attend differently to the different memory caches, because the observations stored in all caches are the same.

Conclusions

We investigated whether a recent language learning model that succeeds on a suite of textual reasoning tasks is able to succeed in a task that requires reasoning about latent mental states. We found that the model is unable to succeed in a set of simulated *true-belief* and *false-belief* tasks unless it has observed at training time situations that have the same structure as the test tasks, even if the diversity of the data is increased. This strongly suggests that the model is not reasoning about the state of the world, nor about mental representations thereof, but is simply memorizing its input. As a consequence, the model will not be able to succeed in a task of reasoning that differs greatly from the situations that it has observed at training time. This is in contrast to the the novelty of situations that people encounter regularly, in which they must reason about the causal relationship between events in the world and latent mental states.

However, incorporating a simple mechanism that informs the model that there may be multiple observers with differing representations of the story allows the model to achieve higher performance on the simulated *false-belief* and *true-belief* tasks. Under this modification, the model does not simply memorize the training data but also learns to use knowledge that agents have (perhaps conflicting) observations about the story in order to answer the question. We could interpret this as analogous to the development of theory of mind in that, when a child is able to reason about others' knowledge of and beliefs about the world, the child succeeds on tests of theory of mind such as the *false-belief* task. A further direction of research could investigate whether manipulating variables in the training data (*e.g.,* frequency of mental state terms) affects the model's performance in a manner similar to how a child's developmental trajectory would be affected.

References

Baron-Cohen, S. (1989). The autistic child's theory of mind: A case of specific developmental delay. *J. of Child Psychology & Psychiatry*, *30*(2), 285–297.

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

Bretherton, I., & Beeghly, M. (1982). Talking about internal states: The acquisition of an explicit theory of mind. *Developmental Psychology*, *18*(6), 906.

Goodman, N. D., et al. (2006). Intuitive theories of mind: A rational approach to false belief. In *Proceedings of Cog. Sci.* (pp. 1382–1387).

Henaff, M., Weston, J., Szlam, A., Bordes, A., & LeCun, Y. (2017). Tracking the world state with recurrent entity networks. In *Proceedings of ICLR*.

Johnson, C. N., & Wellman, H. M. (1980). Children's developing understanding of mental verbs: Remember, know, and guess. *Child Development*, 1095–1102.

Meltzoff, A. N., Gopnik, A., & Repacholi, B. M. (1999). Toddlers' understanding of intentions, desires and emotions: Explorations of the dark ages. In P. Zelazo, J. Astington, & D. Olson (Eds.), *Developing theories of intention: Social understanding and self control* (pp. 17–41). Erlbaum.

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: meta-analysis of the relation between language ability and false-belief understanding. *Child develop.*, *78*(2), 622–646.

Nelson, K. (2007). *Young minds in social worlds: Experience, meaning, and memory*. Harvard University Press.

O'Laughlin, C., & Thagard, P. (2000). Autism and coherence: A computational model. *Mind & Language*, *15*(4), 375–392.

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, *5*(2), 125–137.

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

Rosnay, M., & Hughes, C. (2006). Conversation and theory of mind: Do children talk their way to socio-cognitive understanding? *British Journal of Developmental Psychology*, *24*(1), 7–37.

Slaughter, V., & Gopnik, A. (1996). Conceptual coherence in the child's theory of mind: Training children to understand belief. *Child development*, *67*(6), 2967–2988.

Sukhbaatar, S., Weston, J., Fergus, R., et al. (2015). End-to-end memory networks. In *Proceedings of NIPS* (pp. 2440–2448).

Triona, L. M., Masnick, A. M., & Morris, B. J. (2002). What does it take to pass the false belief task? An ACT-R model. In *Proceedings of Cog. Sci.* (p. 1045).

Van Overwalle, F. (2010). Infants' teleological and belief inference: A recurrent connectionist approach to their minimal representational and computational requirements. *NeuroImage*, *52*(3), 1095–1108.

Weston, J., Bordes, A., Chopra, S., & Mikolov, T. (2016). Towards AI-complete question answering: A set of prerequisite toy tasks. In *ICLR*.

Youngblade, L. M., & Dunn, J. (1995). Individual differences in young children's pretend play with mother and sibling: Links to relationships and understanding of other people's feelings and beliefs. *Child Development*, *66*(5), 1472–1492.