Evaluating Theory of Mind in Question Answering

Aida Nematzadeh
DeepMind
nematzadeh@google.com

Kaylee Burns
UC Berkeley
kayleeburns@berkeley.edu

Erin Grant
UC Berkeley
eringrant@berkeley.edu

Alison Gopnik
UC Berkeley
gopnik@berkeley.edu

Thomas L. Griffiths
Princeton University
tomg@princeton.edu

Abstract

We propose a new dataset for evaluating question answering models with respect to their capacity to reason about beliefs. Our tasks are inspired by theory-of-mind experiments that examine whether children are able to reason about the beliefs of others, in particular when those beliefs differ from reality. We evaluate a number of recent neural models with memory augmentation. We find that all fail on our tasks, which require keeping track of inconsistent states of the world; moreover, the models' accuracy decreases notably when random sentences are introduced to the tasks at test.[1]

1 Reasoning About Beliefs

Possessing a capacity similar to human reasoning has been argued to be necessary for the success of artificial intelligence systems (*e.g.*, Levesque et al., 2011). One well-studied domain that requires reasoning is question answering, where simply memorizing and looking up information is often not enough to correctly answer a question. For example, given the very simple scenario in Table 1, searching for the word "Mary" and returning a nearby word is not a correct strategy; instead, a model needs to recognize that Mary is currently at the second location (office and not the bathroom).

Recent research has focused on developing neural models that succeed in such scenarios (Sukhbaatar et al., 2015; Henaff et al., 2017). As a benchmark to evaluate these models, Weston et al. (2016) released a dataset – Facebook bAbi – that provides a set of toy tasks, each examining a specific type of reasoning. For example, the scenario in Table 1 evaluates the capacity to reason using a single supporting fact. However, the bAbi tasks are already too simple for the current models. Only a few years after their release, existing

---

[1] Code to generate dataset and replicate results is available at github.com/kayburns/tom-qa-dataset.

models fail at only one or two (out of 20) tasks (Rae et al., 2016; Santoro et al., 2017). Moreover, all except two of the reasoning tasks in this dataset only require transitive inference (Lee et al., 2016).

Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A: office

Table 1: A task from the bAbi dataset (Weston et al., 2016).

People reason not just about their own observations and beliefs but also about others' mental states (such as beliefs and intentions). The capacity to recognize that others can have mental states different than one's own – *theory of mind* – marks an important milestone in the development of children and has been extensively studied by psychologists (for a review, see Flavell, 2004). Artificial intelligence (AI) systems will also require a similar reasoning capacity about mental states as they are expected to be able to interact with people (*e.g.*, Chandrasekaran et al., 2017; Grant et al., 2017; Rabinowitz et al., 2018).

However, the bAbi dataset does not include tasks that evaluate a model's ability to reason about beliefs. Grant et al. (2017) created a bAbi-style dataset inspired by an influential experiment on the theory of mind called the Sally-Anne task (*e.g.* Baron-Cohen et al., 1985). Their goal was to examine whether the end-to-end memory network (Sukhbaatar et al., 2015) can answer questions such as "where does Sally think the milk is?" in situations that Sally's belief about the location of milk does not match the reality. For example, Sally thinks that the milk is in the fridge but the milk is actually on the table.

The dataset of Grant et al. (2017) provides a first step in designing benchmarks to evaluate the mental-state reasoning capacity of question-answering models, but it is still limited in the types of reasoning it probes. For example, it

only considered first-order beliefs (*e.g.*, Sally's belief about the location of milk). People also reason about second-order (and higher-order) beliefs (*e.g.*, Anne's belief about Sally's belief about the location of the milk). More importantly, similarly to the bAbi dataset, success in each task is defined as correctly answering one question. This does not guarantee that a model has an understanding of the state of the world; in fact, even in developmental theory-of-mind experiments, children are asked a few questions (*e.g.*, "where is milk really?") to ensure that their correct answer reflects their understanding and is not simply due to chance.

In this paper, we address these shortcomings by designing a new dataset that enables us to evaluate a model's capacity to reason about different types of beliefs as well as whether it maintains a correct understanding of the world. To this end, we evaluate a number of different models that perform well on the bAbi tasks: the end-to-end memory network (Sukhbaatar et al., 2015), the multiple observer model (Grant et al., 2017), the recurrent entity network (Henaff et al., 2017), and Relation-Network (Santoro et al., 2017). We find that none of these models succeed at our tasks, suggesting that they are not able to keep track of inconsistent states of the world, in particular when someone's belief does not match the history or reality of a situation.

2 Theory of Mind Experiments

Behavioral research shows that children gradually develop a theory of mind (for a review, see Gopnik and Astington, 1988). At the age of two, most children have an understanding of others' desires and perceptions – if someone wants something, they will try to get it and if something is in their sight, they can see it. Children begin to understand others' beliefs around the age of three, but this understanding is still limited. For example, they might not be able to reason that someone's actions are a result of their beliefs. By the age of five, most children have a unified theory of mind and are able to represent and reason about others' desires, perceptions, and beliefs. Developmental psychologists have designed various experimental paradigms to examine to what extent children are able to reason about others' mental states. We use these experiments as guidelines for designing tasks to evaluate the reasoning capacity of question-answering models. We first explain these experiments.

2.1 The Sally-Anne Experiment

The Sally-Anne false-belief experiment, proposed by Baron-Cohen et al. (1985), examines children's ability to reason about others' *false beliefs*, *i.e.*, when someone's belief does not match the reality. In this experiment, the participants observe two agents, Sally and Anne, with their containers, a basket and a box. After putting a marble in her basket, Sally leaves the room (and is not able to observe the events anymore). After Sally's departure, Anne moves the marble to her box. Then, Sally returns to the room (see Figure 1). The participants are asked the following questions:

- "Where will Sally look for her marble?" *(belief question)*
- "Where is the marble really?" *(reality question)*
- "Where was the marble in the beginning?" *(memory question)*

The first question tests the participants' ability to reason about Sally's belief about the location of her marble. Interestingly, most children before the age of 3 answer this question incorrectly and say that Sally will look at the box (where the marble really is) instead of the basket (where Sally thinks the marble is). These children are not able to reason about Sally's belief which is different from the reality of the world. The reality and memory questions are used to confirm that children's correct answer to the belief question is not due to chance; but because they have a correct understanding of the state of world and others' beliefs.
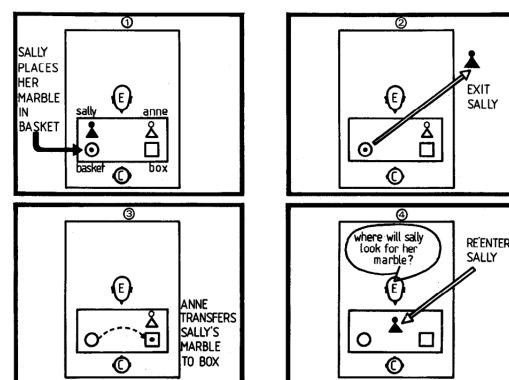

Figure 1: The Sally-Anne experiment setup from Baron-Cohen et al. (1985).

2.2 The Icecream Van Experiment

The Sally-Anne experiment examines the ability to reason about another person's belief or a *first-order belief*. People are also able to reason about

beliefs about beliefs, for example, Anne thinks that Sally believes that the marble is in basket. Perner and Wimmer (1985) performed a set of experiments to examine children's reasoning capacity about such higher-order beliefs. In their set-up Mary and John together see an ice cream van in the park, and the icecream man tells them that he will be in the park until later in the afternoon. Mary leaves the park and goes home. A bit after she leaves, the icecream man decides to leave the park and tells John that he is going to the church. On his way to the church he runs into Mary and informs her that he will be selling icecreams close to the church all afternoon. The participants are then asked the following second-order question: "Where does John think Mary goes to get icecream?" Note that John does not know that Mary has been told about the new location of the icecream van; he has a *second-order false belief* about Mary's belief. The participants are also asked a few control questions (*e.g.*, "does Mary know that the van is in the church?") to ensure that they do not correctly answer the second-order question by chance. Perner and Wimmer (1985) found that 6- and 7-year old children are able to answer the second-order questions, suggesting that reasoning about higher-order beliefs (as compared to a first-order belief) is a harder cognitive task.

3 The Theory of Mind Task Dataset

Inspired by the theory-of-mind experiments explained in Section 2 and building on the work of Grant et al. (2017), we created a dataset based on three tasks designed to capture increasingly complex theory-of-mind reasoning: *true-*, *false-*, and *second-order false-belief* tasks. Examples of each task type are given in Figure 2. In the true-belief task, Sally observes the world and as a result she has a first-order *true belief* about the location of the milk – her belief matches reality. In the false-belief task, Sally's first-order belief differs from reality (*i.e.*, she has a *false belief*) because she was absent when the state of the world changed. In the second-order false-belief task, Sally observes the new location of the milk; thus, she has a *true belief* about the milk's location. However, Anne's belief about Sally's mental state does not match reality because Anne does not know that Sally has observed the change in the environment. As a result, Anne has a *false belief* about Sally's beliefs.

These tasks are more challenging than the bAbI scenarios, because a model needs to learn whether each agent has a true or false belief about a given world state to succeed, where the world state now includes the mental states of each agent.

Note that we assume all containers are transparent in the underlying world; whenever an agent enters a location, they become aware of the objects true location. We made this decision to keep the tasks as structurally similar as possible. This prevents models from simply learning to produce a specific answer for a task type when a sentence like "Sally looks inside the pantry" is present in the story. The container-transparency property is consistent throughout all task-question pairs.

Question types. To examine the reasoning capacity of each model about beliefs and second-order beliefs, we employ four question types inspired by theory-of-mind experiments discussed in Section 2; see Table 2 for examples of these question types. These questions enable us to test whether a model can reason about first-order and second-order beliefs, and at the same time, knows the initial and current correct location of an object; thus, we can distinguish between when a model answers a question by chance and when it actually understands the entire state of the world.

Table 3 gives the answers for the 12 combinations of task type and question. Given a *true-belief* or *false-belief* task, the answers to the first-order and second-order questions are the same (*e.g.*, "pantry" in the *true-belief* condition and "fridge" in the *false-belief* condition for the tasks in Figure 2). However, they are different in the *second-order false belief* task because Anne has a false belief about Sally's belief.

Dataset variants. We use these tasks to generate two datasets: *ToM* and *ToM-easy*. The primary difference between these two datasets is that, in *ToM-easy*, each story has only one task, while *ToM* can have multiple tasks within a single story. Each dataset contains a training set with 10 000 examples with each of the 12 combinations of task and question types.

In *ToM*, the tasks are randomly grouped into sets of 5 to form stories, which is the same number used in the bAbI dataset. In the test set for *ToM*, each story contains 4 tasks, but there is only one question present at the end. Because questions that come closer to the beginning of a story have

2394

True Belief
Anne entered the kitchen.
Sally entered the kitchen.
The milk is in the fridge.
Anne moved the milk to the pantry.

False Belief
Anne entered the kitchen.
Sally entered the kitchen.
The milk is in the fridge.
Sally exited the kitchen.
Anne moved the milk to the pantry.

Second-order False Belief
Anne entered the kitchen.
Sally entered the kitchen.
The milk is in the fridge.
Sally exited the kitchen.
Anne moved the milk to the pantry.
Anne exited the kitchen.
Sally entered the kitchen.

Figure 2: An example story from each of the three task types.

Memory	Where was the milk at the beginning?
Reality	Where is the milk really?
First-order	Where will Sally look for the milk?
Second-order	Where does Anne think that Sally searches for the milk?

Table 2: Examples of the four question types.

	TB	FB	SOFB
Memory	first	first	first
Reality	second	second	second
First-order	second	first	second
Second-order	second	first	first

Table 3: The correct answer to each question for *true-belief* (TB), *false-belief* (FB), and *second-order false-belief* (SOFB) tasks. Here, "first" and "second" are the initial and actual locations of the object of interest, respectively (*e.g.*, fridge and pantry in Figure 2).

fewer distracting sentences (*i.e.*, potential answer words) that may confound a model, they are easier to answer. We found that this testing procedure gave us a more precise understanding of the performance of the model by separating the difficulty of a question due to its position in a story from the inherent difficulty of the question itself.

Generating the data. Each reasoning task in Weston et al. (2016) can be formalized with a grammar. The training and test data are then the derivations of this grammar. We refer to each derivation as a story (*e.g.*, Figure 2). We follow Grant et al. (2017) in writing grammars for our new tasks. In particular, all the task grammars consist of a set of *entities* (people and objects in the stories) and *predicates* that take entities as subject or object. The grammars also specify the *properties* of entities – which predicates take them as subjects or objects. A predicate can include *actions* that are ways an agent interact with the world (*e.g.*, *place, move, enter, exit*) and *beliefs* that are mental state terms (*e.g.*, *believe, think*). As an example, *Sally* with the property *is agent* can perform the action *displace* on *apple* with the property *is object*. Similar to the previous work, we use a restricted set of *action* and *belief* predicates.

4 The Models

We briefly describe the models that we evaluate in this paper. We chose these models based on the novelty in their architecture or their near state-of-the-art results in the bAbi tasks. More specifically, given 10k examples and joint-training on the bAbi tasks, the best end-to-end memory network (Sukhbaatar et al., 2015) and relation network (Santoro et al., 2017) fail at 6 and 2 tasks, respectively. Given the same training dataset and per-task training, the recurrent entity network succeed at all tasks (but the authors do not report the results of joint-training). Recall that the bAbi tasks are structured as a set of sentences followed by a question about them (*e.g.*, a story from Figure 2 followed by a question from Table 2).

The End-to-End Memory Network. Sukhbaatar et al. (2015) proposed a neural memory-augmented model, the end-to-end memory network (MemN2N), that extends the memory network architecture (Weston et al., 2014). Similarly to its predecessor, MemN2N has a memory component in which sentences are embedded and an attention function that weights the embedded sentences based on their similarity to a given question. The MemN2N model introduces multiple layers of memory (hops) by stacking memory components such that the question embedding at layer $k + 1$ is the sum of the output and question embedding of layer k.

The Multiple Observer Model. To perform well on the *false-belief* and *second-order false-belief* conditions, a model needs to identify that agents have experienced different events, and, as a result, have differing knowledge about the state of the world. Although having multiple layers of memory in the MemN2N model enables it to combine attention to different memory slots (*i.e.*, embedded sentences) at each layer, the model does not have access to each agent's unique perspective. For example, the model is not explicitly told that Sally does not observe the change of the lo-

cation of the milk in the *false-belief* condition. To address this, Grant et al. (2017) propose the Multiple Observer model that integrates MemN2N with individual memory modules for each agent in the story. An agent's memory only receives the sentences for which the agent is present and observes the world. Their model has an additional attention function that weighs the memory modules based on their relevance to the question. The model is expected to learn to attend to Sally's memory module if the question is about her belief about a state of the world.

The Recurrent Entity Network. Henaff et al. (2017) propose a memory-augmented architecture, EntNet, with two interesting properties; first, their model is a recurrent neural network and thus can capture the sequential nature of the events in a story. Second, instead of keeping a whole sentence embedding in a memory slot, their model can learn the important entities of a story (*e.g.*, a person) and their properties (*e.g.*, location) through a set of gated recurrent units and two weight matrices.

The Relation Network. Santoro et al. (2017) propose a neural model for relational reasoning. Their model consider the possibility of a relation among each two possible pairs of objects. To model the bAbi tasks, they consider each pair of sentences together with the question as inputs to their relation network.

5 Experimental Results

Experiment set-up We train all models jointly over all task types without noise, but evaluate them independently on different task and question pairs. We choose the best-performing models by selecting hyperparameters on the validation set. Similarly to Sukhbaatar et al. (2015), we consider a model successful only when its accuracy exceeds 95% across the entire task suite.

MemN2N and Multiple Observer Models. We first examine how each model performs across a range of parameter and initialization values. MemN2N models are very sensitive to the network initialization and for each set of parameters, the best result out of 10 runs is reported (Sukhbaatar et al., 2015). We first visualize the accuracy of all runs as a box plot to identify how sensitive each model is to random initialization (of parameters and internal states) and thus difficult to train. We also report the results for the best run in each experiment. We use a memory size of 50, the same

as experiments of Sukhbaatar et al. (2015), to ensure that the memory contains all sentences of a given story.

EntNet. We report results averaged over 3 initializations because we observed little randomness due to initialization. We selected the learning rate on a held out validation set separately for *ToM-easy* and *ToM*; all otehr the same hyperparameters as Henaff et al. (2017): 20 memory slots and an embedding size of 100 We trained until the training error hit zero, which occurred around 50 epochs for both datasets. .

RelNet. We report results using a single seed because we saw little randomness due to initialization; this is in accordance with the authors' findings (Santoro et al., 2017). We selected model hyperparameters on a held-out validation set separately for each of the *ToM* and *ToM-easy* datasets.

### 5.1 Overall Performance on *ToM-easy*

We expect the models perform well on this dataset, given that there is only one task in the memory during both training and test and as a result unrelated events do not interfere with a model's reasoning in this condition.

Despite the large variance in accuracy across runs, the MemN2N models often succeed at the memory, reality, and second-order questions (Figure 3a). Note that the median accuracy (the dark blue line in box) is close to 1 for these questions. However, the model often fails (median accuracy around 0.5) given the first-order question ("where will Sally look for the milk") and the *false-belief* and *second-order false-belief* tasks. This pattern is different from the empirical findings on people; the second-order question is harder than the first order one for people. We observe that the performance of the Multiple Observer model is similar to the MemNet for most question-task pairs (expect one) but there is less variation in the accuracy values. Interestingly, the median accuracy is close to one for the first-order question and the *second-order false-belief* task but the Multiple Observer model still performs poorly for this question on the *false-belief* task.

Why is the first-order question harder for the MemN2N model? To investigate this, we look more closely at our task-question pairs. As shown in Table 3, the answers to the first-order question are different for the *false-belief* and *second-order false-belief* tasks but are the same for the

Figure 3: Memory Network and Multiple Observer Model Performance Across Task and Question Types.
Pink indicates that the answer to the question is the first container that contained the object in that task. Blue indicates that the answer is the last container that contained the object before the question was asked. Grey indicates that the answer was the first container that contained the object in the entire story which may or may not be the same as the pink.
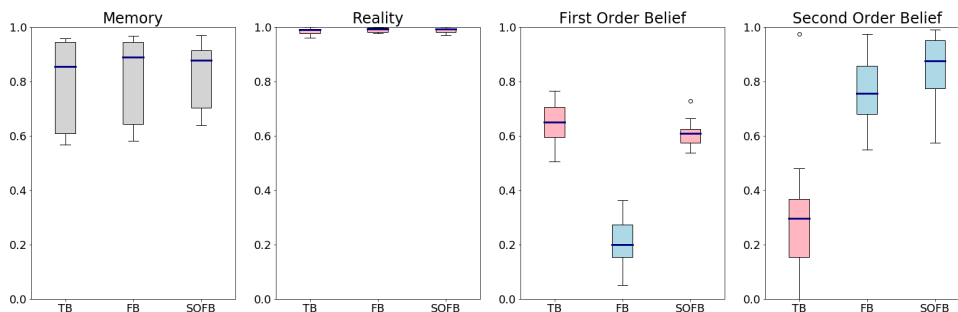
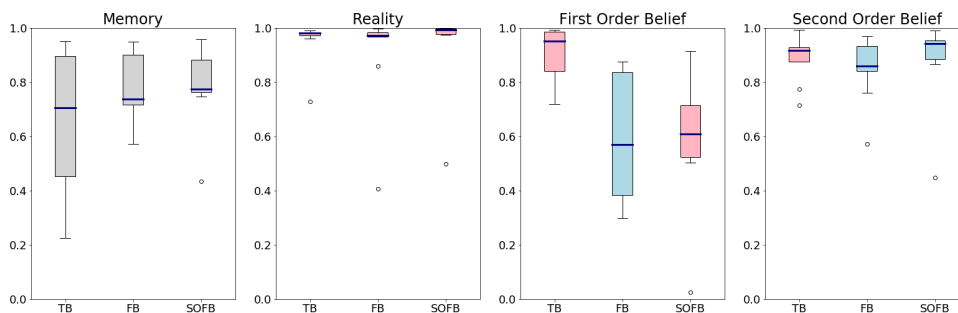(a) Memory Network with memory size 50 evaluated on the *ToM-easy* dataset.
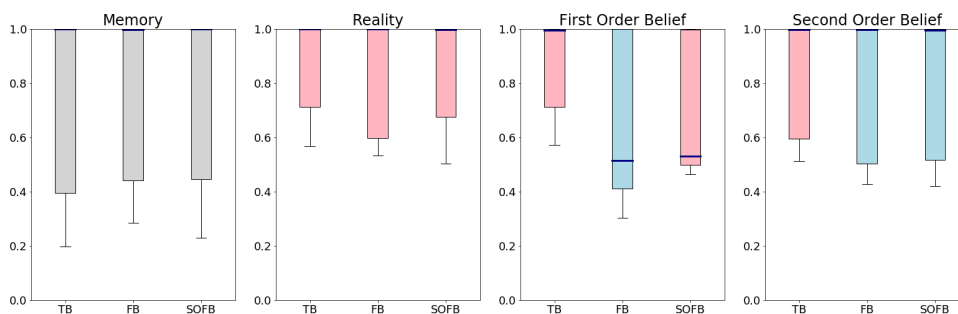
(b) Multiple Observer Model with memory size 50 evaluated on the *ToM-easy* dataset.
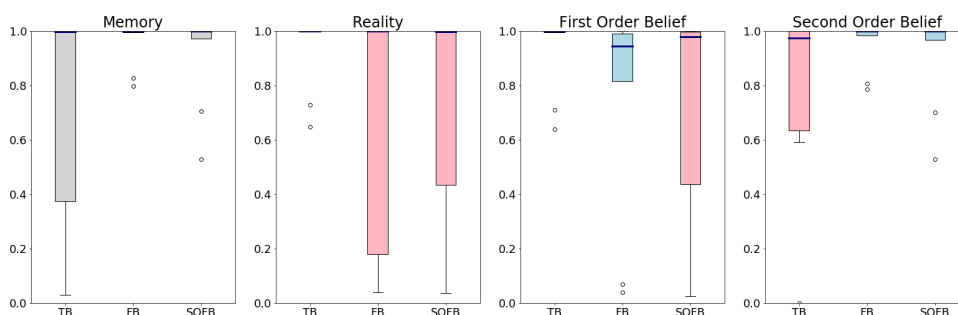
(c) Memory Network with memory size 50 evaluated on the *ToM* dataset.

(d) Multiple Observer Model with memory size 50 evaluated on the *ToM* dataset.

second-order one. We suspect that it is harder for the MemN2N model to learn two distinct answers for the same question given the the similarities of the two false-belief tasks. To test this hypothesis, we altered the data such that the answers to first-order question are (incorrectly) the

same for both *false-belief* and *second-order false-belief* tasks (and also the second-order question). We observe that the median accuracy is close to 1 for all conditions suggesting that the model can learn the distinction between two of the tasks but not all three.

We observe that EntNet and RelNet models are not too sensitive to the initialization value, and thus just report the result on best-performing models. Both EntNet and RelNet best models succeed at the *ToM-easy* tasks; their mean error is 0.

5.2 Overall Performance on *ToM*

This dataset is more similar to the bAbi dataset in that, during both training and test, the memory contains a story with multiple tasks; as a result, it is harder for the model to identify the entities relevant to a given question.

As shown in Figure 3c, the MemN2N performs worse on all the questions with the exception of the reality question, where performance is slightly worse but has lower variance. (*cf.* the *ToM-easy* dataset). The first- and second-order questions are, overall, harder for the model. The performance of the Multiple Observer model is better for most of the questions (see Figure 3d), especially the second-order, but it is slightly worse for the memory question. This suggests that adding agent-specific memory modules is not enough to succeed on all the *ToM* tasks, and that the increased complexity of inference with multiple memory modules harms performance on another task.

We also look at the best-performing MemN2N and Multiple Observer models over all question-task pairs. These models are selected based on their performance on the validation dataset. We observe that none of these models succeed on the *ToM* tasks, but, interestingly, the Multiple Observer model performs better overall as compared to the original MemN2N model (see Table 4).

***ToM* and bAbi.** How are *ToM* and bAbi tasks similar? The combination of our *true-belief* task and the reality question is very similar to bAbi task 1 (single supporting fact). To correctly answer the reality question ("where is the milk really?", a model need to use a single fact from the story ("Anne moved the milk to the pantry."). The MemN2N model succeeds at both bAbi task 1 and the reality question given the *true-belief* task. However, the correct answer to the memory question ("where was the milk at the beginning?") for

the *true-belief* task also requires a single fact ("the milk is in the fridge."). Interestingly, the error of MemN2N on the memory question is much higher than the reality question. The model (unlike people) cannot learn two representations (initial and current location) for an object. This result demonstrates the importance of representing alternative states of the world, whether it be past states of reality (*i.e.*, where the "milk" used to be) or mental states about the world (*i.e.*, where an agent believes the "milk" to be).

EntNet and RelNet. We also report results of two relevant memory-augmented neural network models on our tasks, EntNet (Henaff et al., 2017) and RelNet (Santoro et al., 2017), in Table 4. Again, because we did not observe sensitivity to initialization for these models, only average performance of their best-performing model is reported. We see that even though these models succeed on the *ToM-easy* dataset, they fail on the *ToM* tasks, suggesting that these models cannot simultaneously deal with inconsistent (*i.e.*, past, present, and mental) states of the world.

We further investigate which questions are the hardest for each best model; see Table 5. We observe that each of the MemN2N, Multiple Observer and RelNet models perform poorly on some combination of the first- and second-order questions, but are successful at answering the reality question. We hypothesize that this phenomenon occurs because the reality question is the most similar to the bAbi tasks. In addition, all models fail the memory question for each task type. While this is to be expected for EntNet due to its recurrent nature and therefore bias towards recency, it is surprising that the other models, which exhibit only a small positional bias, cannot correctly represent a past state of the world in order to answer the memory question correctly.

5.3 Experimenting with Noise

We examine to what extent each model's architecture is sensitive to the position of sentences in each story. We do so by adding a novel sentence at random locations in each story at test time. For any setting of the noise, p, there is a $p\%$ probability of a noise sentence occurring before each sentence in the story. Noise sentences cannot follow other noise sentences in the story. In this paper, we report results with $p = .1$. We observe that the accuracy of all best models decreases notably in the

Model	*ToM-easy*	*ToM-easy* (noised)	*ToM*	*ToM* (noised)
MemN2N (Sukhbaatar et al., 2015)	100.00%	90.28%	82.38%	77.00%
Multiple Observer (Grant et al., 2017)	100.00%	**95.81**%	91.11%	87.43%
EntNet (Henaff et al., 2017)	100.00%	94.61%	93.67%	**88.63**%
RelNet (Santoro et al., 2017)	100.00%	87.82%	**94.31**%	76.84%

Table 4: A comparison of model performance. All models succeed on the *ToM-easy* dataset without noise. Multiple Observer model performs best on *ToM-easy* with noise, RelNet performs best on *ToM*, and EntNet performs best on *ToM* with noise.

Model	True-belief	False-belief	Second-order False-belief
MemN2N (Sukhbaatar et al., 2015)	M (94.4); FOB (78.2); SOB (42.9)	M (94.3); FOB (17.3)	M (92.8); FOB (56.4)
Multiple Observer (Grant et al., 2017)	M (93.2)	M (90.5); FOB (56.4); SOB (92.5)	M (93.0); FOB (90.3); SOB (90.3)
EntNet (Henaff et al., 2017)	M (74.0)	M (76.1)	M (74.25)
RelNet (Santoro et al., 2017)	M (39.7%); FOB (71.37%)	M (78.5%)	M (34.9%); FOB (81.8%)

Table 5: Model accuracy on failed questions given the *ToM* task (without noise); M, R, FOB, and SOB are the memory, reality, first- and second-order questions, respectively. The number in the parentheses is the accuracy for that question on that task.

presence of noise (see Table 4). This result is particularly interesting as it shows that none of the models are able to use the semantics of the sentences in a story in their reasoning – they are all sensitive to the presence of distractor sentences.

Interestingly, the RelNet model is the best performer amongst the models we considered on the *ToM* dataset, yet it is also the most sensitive to noise. Moreover, the Multiple Observer model – with explicit memories for each agent – is the most robust to noise; it has the minimum decrease in accuracy between each dataset and its noised version.

5.4 Experimenting with Memory

In the experiments of Sukhbaatar et al. (2015) the memory size is fixed to 50, which is necessary to capture the entire story in memory (e.g. the answer to the memory question in *ToM* may rely on information at the beginning of a story). We observed that smaller memory sizes artificially improved the performance of the MemN2N and Multiple Observer model on *ToM* tasks. For example, using a memory size of 10, our best MemN2N model performance boosts on the hardest task of *ToM* (FB task with first order belief question) from 5.1% to 97.5% and on the easiest task from 98.3% to 100.0% (SOFB task with reality question). This result is not surprising because given a small memory size, *ToM* and *ToM-easy* are very similar tasks; the memory size of 10 allows for at most two full tasks in memory.

6 Related Work

Recent research has emphasized the importance of modeling and understanding people's mental states for AI systems. Eysenbach et al. (2016) created a dataset of scene-description pairs where each scene is a set of visual frames and some frames include people with mistaken beliefs.[2] The authors build a regression model for identifying a mistaken belief and the person who has such a belief in a given frame. Our work differs with theirs in that we are interested in understanding whether a model can reason about people's true and false beliefs to correctly answer questions as opposed to identifying mistaken beliefs.

Grant et al. (2017) studied whether the end-to-end memory network of Sukhbaatar et al. (2015) can pass a false-belief test – correctly answer where Sally would search for an object in false- and true-belief situations. They created a dataset inspired by the bAbi dataset to examine whether the model can reason about interaction of beliefs and actions in these situations – how actions cause beliefs and vice versa. They show that MemN2N fails at the false-belief test, and their extension of that model with separate memories for each agent and an observer outperforms MemN2N.

Rabinowitz et al. (2018) formulate the capacity to reason about others' beliefs as a meta-learning problem. They propose a neural network, ToM-net, that learns to predict the behavior of different agents given their past and current trajectory. Similarly to Grant et al. (2017), in addition to individual agents, they model an "observer" that has access to states and actions of all agents (though this information can be noisy and partial). Interestingly, their model successfully predicts an agent's behavior in a false-belief situation – the agent's behavior reflects its false-belief as opposed to the reality of the world.

Finally, Chandrasekaran et al. (2017) take a dif-

---

[2]For example, a scene where a person gets sick eating mushrooms is paired with the sentence "the couple mistakenly thinks it's ok to eat the mushrooms".

ferent approach by studying whether people can understand the "beliefs" of a visual-question answering system. More specifically, they examine whether the participants can predict when the model would fail in answering a question as well as if they can predict the model's answer. They find that even with a few examples, people get better at answering these questions.

7 Discussion

We propose a dataset for evaluating question-answering models. Our dataset – inspired by seminal theory-of-mind experiments in children – measures to what extent recently introduced neural models can reason about beliefs and states of the world that are potentially mututally inconsistent. We evaluate three of the recent neural question answering models (and an extension of one) on our tasks. We find that none of the models are able to succeed fully on a suite of tasks that requires keeping track of inconsistent beliefs or states of the world. These inconsistencies arise from differences between the past and the present, as well as the mental states of agents who may have false beliefs about the world or about the mental states of other agents.

The purpose of the dataset introduced in this work is not to test advanced language fluency; instead, consistency in the linguistic structure of the tasks allows us to isolate the performance of the models' reasoning capabilities. Even though the language is simple, the models struggle to achieve good performance. Furthermore, we note that the proposed dataset should be treated as a diagnostic tool and that good performance on similar toy tasks is not sufficient for reasoning capabilities.

References

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a theory of mind? *Cognition*, 21(1):37–46.

Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. 2017. It takes two to tango: Towards theory of AI's mind. *arXiv preprint arXiv:1704.00717*.

Benjamin Eysenbach, Carl Vondrick, and Antonio Torralba. 2016. Who is mistaken? *arXiv preprint arXiv:1612.01175*.

John H Flavell. 2004. Theory-of-mind development: Retrospect and prospect. *Merrill-Palmer Quarterly*, 50(3):274–290.

Alison Gopnik and Janet W Astington. 1988. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child develop.*, 59:26–37.

Erin Grant, Aida Nematzadeh, and Thomas L. Griffiths. 2017. How can memory-augmented neural networks pass a false-belief task. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2017. Tracking the world state with recurrent entity networks. In *International Conference on Learning Representations*.

Moontae Lee, Xiaodong He, Wen tau Yih, Jianfeng Gao, Li Deng, and Paul Smolensky. 2016. Reasoning in vector space: An exploratory study of question answering. In *International Conference on Learning Representations*.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, page 47.

Josef Perner and Heinz Wimmer. 1985. "John thinks that Mary thinks that ..." Attribution of second-order beliefs by 5-to 10-year-old children. *Journal of experimental child psychology*, 39(3):437–471.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *Proceedings of the 35th International Conference on Machine Learning*.

Jack W. Rae, Jonathan J. Hunt, Tim Harley, Ivo Danihelka, Andrew W. Senior, Greg Wayne, Alex Graves, and Timothy P. Lillicrap. 2016. Scaling memory-augmented neural networks with sparse reads and writes. In *Proceedings of 30th Conference on Neural Information Processing Systems*.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy P. Lillicrap. 2017. A simple neural network module for relational reasoning. In *International Conference on Learning Representations*.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Proceedings of 29th Conference on Neural Information Processing Systems*, pages 2440–2448.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2016. Towards AI-complete question answering: A set of prerequisite toy tasks. In *International Conference on Learning Representations*.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory Networks. *ArXiv e-prints*.