Revisiting the Evaluation of Theory of Mind through Question Answering

Matthew Le **Y-Lan Boureau** **Maximilian Nickel**
Facebook AI Research
New York, NY
{mattle,ylan,maxn}@fb.com

Abstract

Theory of mind, i.e., the ability to reason about intents and beliefs of agents is an important task in artificial intelligence and central to resolving ambiguous references in natural language dialogue. In this work, we revisit the evaluation of theory of mind through question answering. We show that current evaluation methods are flawed and that existing benchmark tasks can be solved without theory of mind due to dataset biases. Based on prior work, we propose an improved evaluation protocol and dataset in which we explicitly control for data regularities via a careful examination of the answer space. We show that state-of-the-art methods which are successful on existing benchmarks fail to solve theory-of-mind tasks in our proposed approach.

1 Introduction

Humans interact and communicate with other people in a highly efficient way, as described for instance as Grice's cooperative principle (Grice et al., 1975). Inferring other people's mental state is a crucial component of cognitive development, playing a role in how humans learn the meaning of words (Bloom, 2002), distinguish beliefs from reality (Gopnik and Astington, 1988), predict people's behavior (Wimmer and Perner, 1983), understand what others refer to, and reduce ambiguity in conversation (Clark, 1981; Keysar et al., 2000). This ability to reason about the mental state of other agents, called *theory of mind*, is thus an important component of an intelligent system aiming to emulate and interact with humans.

In developmental psychology, classic tests such as the *Sally-Anne test* (Baron-Cohen et al., 1985) have been used to assess the ability to infer false beliefs in others. Recently, Grant et al. (2017) and Nematzadeh et al. (2018) proposed to adapt these tests to evaluate the capability of machine learning models to form a theory of mind. The key insight of Grant et al. (2017) was to cast them as question answering tasks, where a system is given a story and has to answer questions about the beliefs of agents in that story. This allows to adapt the bAbi benchmarking protocol (Weston et al., 2016) to evaluate theory of mind capabilities of modern neural network architectures: stories are automatically generated so that a suitably large number of examples can be provided for training.

We believe this is a promising approach for advancing research on theory of mind, as it decouples its evaluation from problems such as multi-agent systems, game theory, and meta-learning, which are all components of other evaluation methods (Rabinowitz et al., 2018; Bard et al., 2019). However, artificial data runs the risk of displaying hidden artifacts and biases that correlate with the prediction task and can be exploited by models (Jabri et al., 2016; Gururangan et al., 2018; Poliak et al., 2018). The highly systematic nature of data generation puts synthetic benchmarks at an even higher risk of overestimating the target competence of interest.

This paper shows that current theory-of-mind QA benchmarks do indeed suffer from data biases, and are perfectly solvable without theory of mind. To overcome this, we propose an improved evaluation method and dataset[1]. We then show that state-of-the-art memory-augmented models – which are successful on existing benchmarks – fail to solve theory-of-mind tasks in our improved approach.

2 Theory of Mind Benchmarks

This section briefly reviews the Sally-Anne test and related paradigms for evaluating theory of mind.

First-Order Beliefs The so-called *Sally-Anne test* (Baron-Cohen et al., 1985) examines children's

[1] Our code and dataset will be made available at https://github.com/facebookresearch/ToMi

ability to reason about other agents' false beliefs. The child observes the actions of two agents (Sally, Anne). First, Sally puts an object into a container. Second, Anne moves the object without Sally observing this action. Third, the child is asked multiple questions about reality and the agents' beliefs:

First-Order Belief: Where will Sally look for the marble?

Reality: Where is the marble really?

Memory: Where was the marble in the beginning?

The first question tests the ability of the child to infer the correct mental state of Sally, i.e., that she has the false belief of the marble being in the basket. The reality and memory questions ensure that the child has a correct understanding of the state of the world and is not responding at chance.

Second-Order Beliefs "Sally-Anne"-type questions are well-suited to evaluate *first-order* beliefs. Nematzadeh et al. (2018) proposed to also evaluate *second-order* theory of mind, i.e., the ability to infer beliefs about beliefs. Perner and Wimmer (1985) proposed a set of experiments to test such second-order beliefs in children. The experiments can be summarized as follows: Two agents (Mary, John) see an icecream van in the park. The vendor tells them that he will be in the park all afternoon. After Mary leaves the park, the vendor decides to leave the park and tells John he is going to the church. On the way to the church, the vendor meets Mary and tells her also that he will be at the church. The child is then asked the following question:

Second-Order Belief: Where does John think Mary will go to get ice-cream?

As control, children are also asked memory, reality, and first-order belief questions.

Psychology Tests as AI Benchmarks Grant et al. (2017) cast the aforementioned experiments as question answering tasks and proposed to create a bAbi-style dataset (Weston et al., 2016) to evaluate theory of mind in artificial intelligence models. For instance, this is the bAbi question-answering task version of the Sally-Anne test:

Sally puts a marble in her basket
Sally leaves the room
Anne moves the marble in her box

Q: *Where would Sally look for the marble?*
A: Basket

Nematzadeh et al. (2018) evaluate several modern neural network architectures over the resulting dataset and report that all fail on theory-of-mind tasks, especially when irrelevant sentences are introduced into stories at test time. In the following, we refer to these benchmarks as *ToM-bAbi*.

3 Evaluating Theory of Mind Evaluation

This section examines shortcomings of ToM-bAbi for evaluating a model's theory of mind abilities and proposes an improved approach.

Related Work on Dataset Artifacts It is challenging to determine what specific competence is revealed by the successful completion of a task, as illustrated by the case of *Clever Hans*, the famous horse whose skill at reading human reactions passed for arithmetic ability (Pfungst, 1911). Recent analyses of several AI benchmarks have uncovered similar difficulties when probing learning models. In the domain of visual question answering, baselines solely relying on candidate answers have shown surprisingly good performance (Jabri et al., 2016), leading to the creation of a carefully designed diagnostic dataset (Johnson et al., 2017). Similarly, natural language inference benchmarks have been shown to be vulnerable to bias exploitation (Gururangan et al., 2018; Poliak et al., 2018), and multimodal machine translation benchmarks have been demonstrated to be too simple to require multi-modality (Caglayan et al., 2019). We follow the same approach here of examining the performance of a baseline that does not make use of a type of information crucial to the target competence. Taking "reasoning about another agent" as a working definition of theory of mind, a reasonable prerequisite of benchmarks probing theory of mind competence is that they be impossible to solve without some input about the other agent.

Leveraging Dataset Biases in ToM-bAbi Reviewing the generation process of ToM-bAbi uncovers predictable regularities that allow a model to use corner-cutting heuristics instead of keeping track of agents: the stories follow a strict event sequence template for each task type, shown in Fig. 1. The ToM-bAbi dataset takes precautions to guard against the most simplistic heuristic (e.g., 'always output the location of line 4') by adding irrelevant sentences at random places as noise, but many regularities and correlations remain. These regularities make it possible to construct a parsimonious set

Algorithm 1: Rules to solve ToM-bAbi

```
Data: question $q$, story $s$
Result: location $l$
if "beginning" $\in q$ then          // Memory
    l $\leftarrow$ location of first object occurrence;
else if "really" $\in q$ then        // Reality
    l $\leftarrow$ location of last object occurrence;
else if "look" $\in q$ then          // 1st Order
    if "exited" $\in s$ after last "object is in" then
        l $\leftarrow$ location of last object occurrence before last "exited";
    else
        l $\leftarrow$ location of last object occurrence;
    end
else if "think" $\in q$ then         // 2nd Order
    if "exited" $\in s$ after last "object is in" then
        l $\leftarrow$ location of last object occurrence before first "exited";
    else
        l $\leftarrow$ location of last object occurrence;
    end
end
```

of rules that perfectly solve all tasks without ever extracting any information about the agents. As shown in the pseudo-code implementation of Algorithm 1, these rules only involve simple lexical and ordinal patterns (code provided in the supplemental material).

Towards Robust ToM Evaluation in QA To increase robustness against such regularities and correlations, we build upon the ideas of ToM-bAbi but improve data generation and evaluation in multiple ways. We refer to the supplementary material for the full pseudo-code of our dataset generator.

First, to generate a balanced dataset over story types, ToM-bAbi uses different generators for true-belief, false-belief, and second-order false-belief stories. However, the different generators add clear biases to the data which can be exploited to identify the story type (for example, lack of the word "exited" signals a true belief story – see Fig. 1). To overcome this issue, we use the same randomized generation method for all stories and keep track of which type is produced. We then sample from this randomized story generator with rejection to create a balanced dataset over all three types of stories.

Second, to decrease the amount of information that can be predicted from any given event, we add the following random distractors during data generation: actions of unrelated agents to decorrelate actions from answers, distractor statements about locations and objects to make the number of mentions less informative, randomization of the order of exit/move/re-entry actions, and randomization of the agent whose beliefs are being queried. This leads to stories that are a lot less predictable (see examples in Fig. 1).

Third, theory of mind manifests in "Sally-Anne" type tests through the understanding that an agent's belief is different from the actual state of the world, and that both coexist at the same moment. It is therefore crucial to evaluate both, the ability to infer the state of the world *and* the mental state of an agent. Although current benchmarks include reality and memory control questions, only one type of question is asked for each different, separately generated story. This can obscure revealing correlations in the models' responses e.g., that accuracy in first-order belief questions is associated with lower performance on reality questions. Since a correct answer to a false-belief question is only meaningful if the reality question is also answered correctly, it is especially important to check that the models can distinguish between states *for each story*. For this reason, we propose to systematically ask all question types for each generated story. In particular, for a single story involving agents `A`, `B`, and object `O`, we ask *all* following questions:

Reality: Where is `O`?

Memory: Where was `O` in the beginning?

First-Order Belief `A`: Where will `A` look for `O`?

First-Order Belief `B`: Where will `B` look for `O`?

Second-Order Belief `A`: Where does `A` believe `B` will look for `O`?

Second-Order Belief `B`: Where does `B` believe `A` will look for `O`?

Furthermore, we propose to count a story s as answered correctly if all questions about s are jointly correct, and to measure overall accuracy as the fraction of correctly answered stories. This ensures that a model is not acquiring theory of mind at the expense of its ability to perform other tasks and that it correctly answers reality and false-beliefs

ToM-bAbi dataset

Three types of stories

1 $\langle A \rangle$ entered $\langle L \rangle$
2 $\langle B \rangle$ entered $\langle L \rangle$
3 Phone rang. *// Distractor can appear anywhere*
4 The $\langle O \rangle$ is in $\langle C1 \rangle$.
5 $\langle B \rangle$ exited $\langle L \rangle$ *// if story type 1 or 2*
6 $\langle A \rangle$ moved the $\langle O \rangle$ to $\langle C2 \rangle$.
7 $\langle A \rangle$ exited the $\langle L \rangle$ *// if story type 2*
8 $\langle B \rangle$ entered the $\langle L \rangle$ *// if story type 2*

Example story

1 Isla entered the bathroom.
2 Benjamin entered the bathroom.
3 The cabbage is in the green_pantry.
4 Phone rang.
5 Isla moved the cabbage to the red_drawer.

Answers for each story-question pair

	Story 1	Story 2	Story 3
First Order	C1	C2	C2
Second Order	C1	C1	C2
Memory	C1	C1	C1
Reality	C2	C2	C2

ToMi dataset

Examples of stories from the ToMi dataset

1 Oliver dislikes the kitchen
2 Carter entered the porch.
3 Abigail entered the porch.
4 The potato is in the green_suitcase.
5 Abigail exited the porch.
6 Abigail entered the hall.
7 Carter moved the potato to the green_envelope.
8 Oliver entered the hall.

1 Mila entered the closet.
2 Isla entered the closet.
3 Ava entered the closet.
4 The orange is in the blue_container.
5 Isla exited the closet.
6 Isla entered the garage.
7 Ava moved the orange to the green_bathtub.

1 William entered the staircase.
2 Aiden entered the staircase.
3 Aiden exited the staircase.
4 Aria entered the staircase.
5 The potato is in the red_drawer.
6 Aiden dislikes the grapefruit
7 William moved the potato to the blue_container.
8 Aria exited the staircase.

Figure 1: *Left:*Stories from the ToM-bAbi dataset follow three strict templates, with the possible random insertion of the distractor phrase "Phone rang." This makes it easier to devise rules to locate the answers for all pairs of question and story types, as shown in Alg. 1. For example, C1 always appears in the same sentence as the first object occurrence. *Right:* the ToMi dataset we propose is generated with considerably more randomness, with distractor phrases, distractor locations, distractor characters, and shuffling of the order of actions.

questions jointly. Formally, we define this *joint accuracy* as

$$Acc_J(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \prod_{q \in \mathcal{Q}_s} \mathbb{1}(\widehat{a}_q = a_q) \qquad (1)$$

where $\mathbb{1} : \{\perp, \top\} \rightarrow \{0, 1\}$ is the indicator function, $\mathcal{S}$ the set of all stories, $\mathcal{Q}_s$ the set of all questions about story s, and $a_q, \widehat{a}_q$ are the correct and the model's answer for question q respectively.

Table 1: Accuracy on benchmark datasets. "Average" indicates the average accuracy over all questions. "Joint" indicates the joint accuracy as defined in Eq. (1).

	Average			**Joint**
	ToM-easy	ToM	ToMi	ToMi
Rules	100.0	100.0	77.5	36.5
MemNN	100.0	90.7	77.2	44.3
EntNet	100.0	94.9	90.6	66.8
RelNet	100.0	94.4	86.0	57.4

4 Experimental Evaluation

This section evaluates state-of-the-art question answering methods on the original ToM-bAbI and on our improved dataset. We follow Nematzadeh et al. (2018) and consider Memory Networks (MemNN; Sukhbaatar et al. 2015), Relation Networks (RelNet; Santoro et al. 2017, and Recurrent Entity Networks (EntNet; Henaff et al. 2017), which solve the original bAbI question answering tasks.

Table 1 shows the average accuracy over all questions for the original ToM-bAbI benchmarks (ToM-easy, ToM) and our improved dataset (ToMi). Moreover, for ToMi we also show the joint accuracy as defined in Equation (1). Algorithm 1 achieves perfect accuracy on ToM-easy and ToM. Since all QA models have the capacity to capture these simple heuristics, it is not surprising that they also perform very well on the original benchmarks, i.e., perfect accuracy on ToM-easy and over 90% percent accuracy on ToM.[2] Moreover, we found that the drop in accuracy in ToM is mostly caused

[2] These results are better than in the ToM-bAbI paper. Communication with the authors has not provided clear reasons why, so this may be due to more extensive optimization.

Table 2: Average accuracy by question type (MemNN)

	ToM-easy	ToM	ToMi
Memory	100.0	71.9	98.90
Reality	100.0	100.0	93.39
First-Order	100.0	98.6	70.72
Second-Order	100.0	95.3	64.66

Table 3: Average accuracy per question type in ToMi. "FB" and "'w/o FB" indicate question-story pairs that do and do not involve false beliefs, respectively.

		MemNN	RelNet	EntNet
w/o FB	First Order	85.45	96.42	94.29
	Second Order	82.67	95.37	85.08
	Reality	93.39	100.0	100.0
	Memory	98.90	99.90	100.0
FB	First Order	12.62	10.40	54.95
	Second Order	17.27	17.81	36.55

by *memory* questions. This is because ToM stories are significantly longer and the location of the first occurrence of the queried object can exceed the memory capacity of the models. Table 2 shows an ablation for MemNNs that illustrates this effect.

By contrast, Algorithm 1[3] and QA models fail to solve the ToMi tasks. This is especially clear when looking at the joint accuracy of Table 1, which is not inflated by easy-to-answer memory and reality questions. Our ablation in Table 3 provides further insights into these results. It lists the average accuracy per question type and further splits the results into false-belief and non-false-belief question-story pairs. All models do reasonably well on question-story pairs that do not involve false beliefs, i.e., where the mental state of an agent should coincide with the state of the world. However, for question-story pairs with false beliefs, all models fail to provide correct answers consistently. Recurrent Entity Networks, which explicitly aim to keep track of the state of the world, are performing best on false beliefs tasks, indicating that this is a useful inductive bias for QA models on theory-of-mind tasks.

5 Discussion

Theory of mind is an important component of intelligent systems which interact with humans. In the context of natural language, theory of mind is not only of interest because of its evaluation through question answering but also because it is a crucial component to understand references and reduce ambiguities. This work re-examined the evaluation of theory of mind through bAbi-style question answering tasks. We revealed exploitable regularities in the generated data of existing benchmarks, and proposed to remedy this with a new dataset and evaluation method. We also showed that existing question answering methods that were capable of solving the previously proposed benchmarks are not able to solve the new tasks anymore. In future work, we aim at developing models to solve the newly proposed tasks.

Achieving this would demonstrate some level of ability to reason about first- and second-order false beliefs. But an important point to keep in mind when using our benchmark is that it still relies on synthetic data and tasks. We chose this approach since synthetic data generation can be especially useful in novel and early-stage research efforts, as it provides a controlled environment and allows for detailed analyses of a model's ability to solve a task. However, while methods such as Recurrent Entity Networks have shown promise for keeping track of the state-of-the-world in our experiments, this is still in scenarios where the complexity of natural language is relatively simple. On real-world data, this ability would be much weaker as it would require additional competencies such as co-reference resolution, handling polysemy and ambiguities, and common-sense reasoning. Therefore, solving the tasks we propose can only be considered a prerequisite for a fully functional theory of mind which will ultimately have to be evaluated in real-world scenarios.

References

Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. 2019. The hanabi challenge: A new frontier for ai research. *arXiv preprint arXiv:1902.00506*.

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

Paul Bloom. 2002. *How children learn the meanings of words*. MIT press.

Ozan Caglayan, Pranava Swaroop Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need

[3] A more elaborate set of rules could solve the tasks, but would require to take the agents into account.

for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170.

Herbert H Clark. 1981. Definite reference and mutual knowledge. *Elements of Discourse Understanding*.

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 Development*, 59(1):26.

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 Meeting of the Cognitive Science Society, CogSci 2017, London, UK, 16-29 July 2017*.

H Paul Grice, Peter Cole, Jerry L Morgan, et al. 1975. Logic and conversation. *1975*, pages 41–58.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2017. Tracking the world state with recurrent entity networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2016. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

Boaz Keysar, Dale J Barr, Jennifer A Balin, and Jason S Brauner. 2000. Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, 11(1):32–38.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L. Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2392–2400.

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.

Oskar Pfungst. 1911. *Clever Hans:(the horse of Mr. Von Osten.) a contribution to experimental animal and human psychology*. Holt, Rinehart and Winston.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *International Conference on Machine Learning*, pages 4215–4224.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 4974–4983.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 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 *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.