



**HAL**  
open science

# MindGames: Targeting Theory of Mind in Large Language Models with Dynamic Epistemic Modal Logic

Damien Sileo, Antoine Lernoùld

► **To cite this version:**

Damien Sileo, Antoine Lernoùld. MindGames: Targeting Theory of Mind in Large Language Models with Dynamic Epistemic Modal Logic. 2023. hal-04098588

**HAL Id: hal-04098588**

**<https://inria.hal.science/hal-04098588v1>**

Preprint submitted on 16 May 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# MindGames: Targeting Theory of Mind in Large Language Models with Dynamic Epistemic Modal Logic

Damien Sileo<sup>1</sup> and Antoine Lernoald<sup>2</sup>

<sup>1</sup>Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRISTAL, F-59000 Lille, France

<sup>2</sup>Univ. Lille, Lille, France

[damien.sileo@inria.fr](mailto:damien.sileo@inria.fr)

## Abstract

Theory of Mind (ToM) is a critical component of intelligence, yet accurately measuring it continues to be a subject of debate. Prior research has attempted to apply human ToM assessments to natural language processing models using either human-created standardized tests or rule-based templates. However, these methods primarily focus on simplistic reasoning and require further validation. In this study, we utilize dynamic epistemic logic, which has established overlaps with ToM, to generate more intricate problems. We also introduce novel verbalization techniques to express these problems using natural language. Our findings indicate that some language model scaling (from 70M to 6B and 350M to 174B) does not consistently yield results better than random chance. While GPT-4 demonstrates superior epistemic reasoning capabilities, there is still room for improvement. Our code and datasets are publicly available<sup>1</sup>.

## 1 Introduction

Theory of Mind (ToM) is the cognitive ability to attribute mental states, such as beliefs, desires, and intentions, to oneself and others, allowing individuals to understand and predict behavior based on these inferred mental states. It is an important requirement for general text understanding, or general artificial intelligence (Navarro et al., 2020). Animal studies have highlighted that claiming Theory of Mind (ToM), or the absence of it, can be methodologically problematic and prone to being biased by human expectations (de Waal, 2016). Kosinski (2023) recently sparked debate by showing that scaling large language models (LLMs) improves performance at standardized tests designed to measure theory of mind. However, such standardized tests were widely discussed in academic research and might have leaked into the training corpora of

LLM. Other previous work generated synthetic examples instead, extending the bAbi (Weston et al., 2016) framework. Nematzadeh et al. (2018) proposed a dataset of fixed templates based on the Sally-Anne problem (Baron-Cohen et al., 1985):

*Sally puts a marble in a box while Anne is with her. Sally leaves for a moment and Mary puts the marble in a basket. Where will Sally look for the marble?* [ANSWER=BOX]

Le et al. (2019) deem these problems simplistic and extend them to track second-order beliefs (e.g. the belief of Sally about Anne’s beliefs).

In this study, we create dynamic epistemic logic (DEL) problems and develop verbalizations to transform them into natural language inference problems. Dynamic epistemic logic is a type of modal logic that facilitates reasoning about agents’ knowledge of facts or other agents’ knowledge. This logic also enables reasoning about the impact of consecutive public announcements:

*Alice and Bob have mud on their head. Their father says that at least one of them is muddy, and asks them if they are muddy. Do they know that they are muddy?* [ANSWER=NO] *They answer that they don’t know. Do they now know that they are muddy?* [ANSWER=YES]

DEL serves as one method to formalize certain ToM problems, making it a valuable perspective for ToM assessment. The problems we create can necessitate tracking multiple agents’ beliefs and reasoning about higher-order beliefs<sup>2</sup>. Our dataset encompasses numerous variations of the *Muddy Children* and *Drinking Logicians* problems (van Eijck, 2014). This controlled test bench offers deeper insights into language model scaling and presents the first dataset with adequate complexity to challenge supervised learning models. The dataset and the scripts to generate them are publicly available<sup>1</sup>.

<sup>2</sup>For example, Anne’s belief about Sally’s belief about Anne’s belief about Mary’s belief.

<sup>1</sup>[\[code:GitHub\]](#) [\[data:HF-datasets\]](#)

## 2 Related Work

**Logical Reasoning in NLP** Logic shares profound connections with NLP. Early systems were built around logic, and more recent approaches incorporate logical reasoning into neural networks (Hamilton et al., 2022). Another line of research investigates the logical capabilities of NLP text encoders using textual datasets. RuleTaker (Clark et al., 2020) explores this area with propositional logic, while LogicNLI addresses first-order logic (Tian et al., 2021). Richardson and Sabharwal (2022) examine the satisfiability problem in natural language. Sileo and Moens (2022) targets probabilistic logic. Our study is the first to focus on modal logic, specifically epistemic logic, in natural language.

**Theory of Mind in NLP** To measure ToM capabilities of NLP models, Nematzadeh et al. (2018) created examples using Sally-Ann templates, and Le et al. (2019) added complexity to the data by incorporating second-order knowledge. Both studies framed their examples as question-answering tasks. Kosinski (2023) employed handcrafted tests to evaluate language models’ next-word prediction capabilities. The Social-IQA dataset (Sap et al., 2019) addresses social commonsense in general, which involves some theory of mind aspects along with other challenges such as knowledge of desires and emotions. Cohen (2021) investigated whether natural language inference models captured veridicality with epistemic verbs like *know* and *think*, using handcrafted patterns. This task was incorporated into the BIG-Bench framework (Srivastava et al., 2022) as the *epistemic-reasoning* task, but it targets only one shallow aspect of epistemic reasoning.

**Epistemic Logic and ToM** Van Ditmarsch and Labuschagne (2007) examined the connections between DEL and ToM, while Dissing and Bolander (2020) demonstrated DEL’s applicability in robotics. Van De Pol et al. (2018) explored the feasibility of using epistemic logic to explain theory of mind by investigating its theoretical computational tractability.

## 3 Dynamic Epistemic Logic Problem Generation and Verbalization

### 3.1 Problem definition

Our objective is to simultaneously create dynamic epistemic logic problems and their corresponding

textual representations, allowing us to develop natural language problems in a (PREMISE, HYPOTHESIS, LABEL) format.

An epistemic logic problem can be constructed using the following components:

**Agents:** A set of  $N$  individuals, each assigned an arbitrary name.

**Predicates:** A set of boolean predicates. In this case, we use  $N$  predicates, one corresponding to each agent (e.g., *Alice has mud on her head*).

**Observabilities** describe each agent’s initial knowledge of predicate values. We represent observabilities with an  $N \times N$  matrix,  $\mathcal{O}$ . If  $\mathcal{O}_{i,j}=1$ , it means that agent  $i$  initially knows whether predicate  $j$  is true.

**Announcements:** A list of expressions (predicates or agent knowledge about predicates) that are shared to all agents. Announcements are made sequentially, and each new announcement can change what the agents know, even if the same announcement is repeated twice.

**Hypothesis:** An expression that may contain predicates and knowledge of agents about particular expressions after the announcements, given the agents, observabilities, and announcements combined into a premise.

### 3.2 Setups: connecting predicate and observabilities

The choice of predicates dictates the observabilities structure. For example, the predicate *"Alice has mud on her head"* is observable by agents other than Alice, but *"Alice has mud on her hand"* could be observable by everyone. We combine predicates and observabilities into what we call *setups* to generate textual descriptions. We define the following setups:

#### Forehead-mud setup

PREDICATE <sub>$i$</sub> :  $\langle \text{AGENT}_i \rangle$ 's forehead is muddy.  
 $\mathcal{O} : \text{ONES}(N) - \text{IDENTITY}(N)$

#### Forehead-mud-mirror setup

PREDICATE <sub>$i$</sub> :  $\langle \text{AGENT}_i \rangle$ 's forehead is muddy.  
 $\mathcal{O} : \text{ONES}(N)$

OBSERVATION: *There is a mirror in the room.*

#### Thirst setup

PREDICATE <sub>$i$</sub> :  $\langle \text{AGENT}_i \rangle$ 's is thirsty.  
 $\mathcal{O} : \text{IDENTITY}(N)$

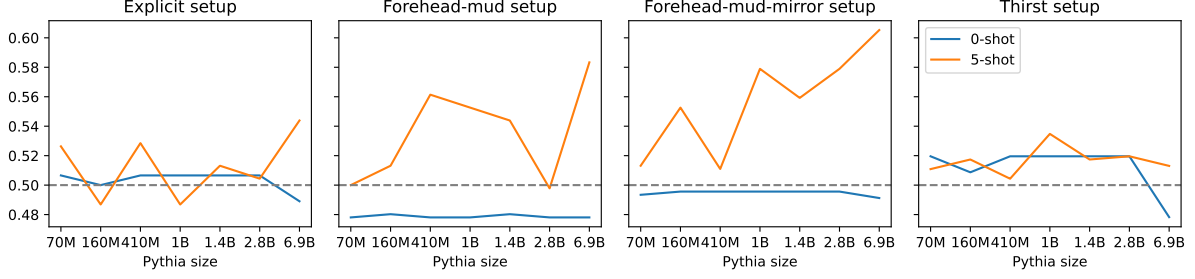


Figure 1: Accuracy of Pythia language models on MindGames setups.

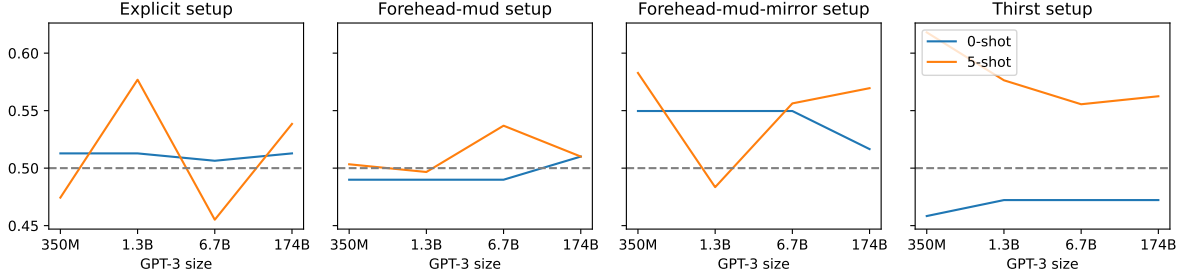


Figure 2: Accuracy of GPT-3 family (ada, cabbage, curie, davinci) language models on MindGames setups.

### Explicit setup

PREDICATE<sub>*i*</sub>:  $\langle \text{AGENT}_i \rangle$  picked a red card.

$\mathcal{O} : \text{RANDBOOL}(N, N), \mathbb{E}(\text{sum}(\mathcal{O})) = N$

OBSERVATION: *Each person draws a card, face unrevealed (red or black).  $\langle \text{AGENT}_j \rangle$  card is revealed to  $\langle \text{AGENT} \rangle_i$ . for all  $i, j$  where  $\mathcal{O}_{i,j} = 1$*

### 3.3 Problem verbalization

We then construct a problem for a setup with the following natural language template:

**[Premise]** *There are  $\langle N \rangle$  persons. Everyone is visible to others.  $\langle \text{SETUP-OBSERVATION} \rangle$  It is publicly announced that someone  $\langle \text{SETUP-PREDICATE} \rangle$   $\langle [0 - N] \text{ ANNOUNCEMENTS} \rangle$*

**[Hypothesis]**  $\langle [1 - K]^{\text{th}} \text{ ORDER BELIEF} \rangle$

We restrict announcements to first-order beliefs. A first-order belief has the following structure:  $\langle \text{AGENT} \rangle$  (*can know whether | can know that | cannot know that | cannot know whether*) ( $\langle \text{SETUP-PREDICATE} \rangle$  |  $\langle \text{NEGATED-SETUP-PREDICATE} \rangle$ ), e.g. *Mary cannot know whether Paul is muddy*. We use the *can* verb to account for the fact that sometimes an agent *can* theoretically infer an expression, but that the reasoning might not be obvious enough for any agent. A  $K^{\text{th}}$  order belief is a first-order belief about a  $(K-1)^{\text{th}}$  order belief. We consider *everyone*, *not everyone*, and *nobody* as possible subjects to setup predicates. Subjects

are uniformly sampled among these quantifiers and individual agents. We transform abstract problem representations into natural language and code that can be fed to a model checker to determine whether a hypothesis is entailed by the premise. We utilize the SMCDEL model checker (Bentham et al., 2018), an announcement logic based on the S5 modal logic. This implementation is the most cited publicly available epistemic logic as of April 2023. We discard examples where the premise contains a contradiction<sup>3</sup>. To generate diverse and gender-balanced random English surnames, we use CensusName<sup>4</sup> (Qian et al., 2022).

## 4 Experiments

### 4.1 Problem generation parameters

We randomly sample  $N \in \{2, 3\}$  agents, as we observed that problems were sufficiently challenging with only three agents, and we use  $K=2$  for the same reason. In announcements, we apply negations 80% of the time to encourage richer inferences (as the fact that an agent does not know something conveys information to others), and 50% of the time otherwise. We deduplicate then undersample the problems to generate 400 cases with a perfect balance of True/False labels per setup. We re-

<sup>3</sup>We identify contradictions by examining whether an unused predicate is entailed or not by the premise.

<sup>4</sup><https://pypi.org/project/censusname/>

fer to the resulting dataset as MindGames.

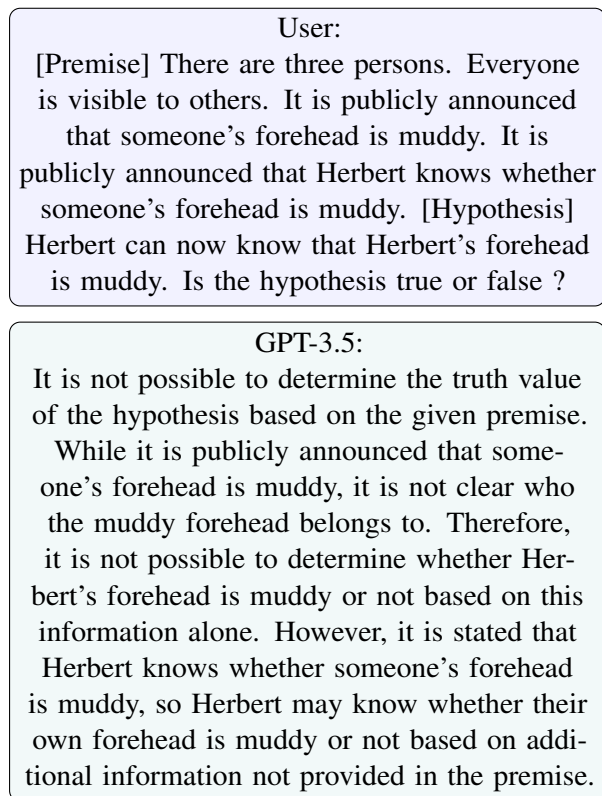


Figure 3: GPT-3.5 conversation

## 4.2 Scaling experiments

We conduct zero-shot experiments and few-shots with a range of language models. We use the `lm-eval-harness` software (Gao et al., 2021) to measure whether a language model perplexity favors the correct reasoning in a multiple-choice setting, with a natural language inference prompt from Brown et al. (2020): `<PREMISE> Question: <HYPOTHESIS> True or False ?"` with two possible continuation choices, *True* and *False*. We evaluate two families of language models:

**Pythia language models** We select the Pythia (Biderman et al., 2023) language models for our open-source scaling experiments. We use the checkpoints trained on the deduplicated corpus (deduped) with checkpoint sizes of 70M, 160M, 410M, 1B, 1.4B, 2.8B, and 6.9B.

**OpenAI API** We utilize the OpenAI GPT-3 (Brown et al., 2020) models, specifically the `ada`, `babbage`, `curie`, and `davinci` checkpoints, through the public API. We assume that their model sizes are respectively 350M, 1.3B, 6.7B, and 174B.

Figure 1 displays the results for various Pythia model sizes. We observe that scaling improves 5-

shot reasoning, but it has no impact on zero-shot reasoning. In contrast to the emergence results reported by Kosinski (2023), Figure 2 does not show a clear scaling trend for GPT-3 models on MindGames data, which suggests that the emergent behavior they observed was not due to robust epistemic logic capabilities.

## 4.3 Qualitative analysis with ChatGPT

We also run brief qualitative analyses with GPT-3.5 and GPT-4 (OpenAI, 2023), as of May 2023. On 20 randomly sampled problems, we found that GPT3 was 60% correct and GPT-4 more than 70% correct. We show a brief qualitative analysis of the respective models.

As shown in Figure 3, GPT3.5 tend to answer that there is not enough information and to perform correct inferences only when it requires very shallow reasoning. GPT-4 can solve this particular example. However, some problems are still challenging, as shown in Figure 4. GPT-4 rarely answers that there is not enough information and its reasoning looks has the surface form of epistemic reasoning, but occasionally contains glaring mistakes.

## 5 Conclusion

We have developed a novel dataset aimed at evaluating epistemic logic reasoning, addressing a particular aspect of Theory of Mind (ToM). Our results reveal that this task continues to pose challenges for contemporary large-scale language models. When future models can solve MindGames for 2-3 agents, the difficulty of the task can be easily scaled up by with more agents. Future studies could explore human performance on our dataset, taking into account factors such as age and educational background. Additionally, further investigation can examine the impact of fine-tuning on other downstream tasks and assess how well Transformer circuits model Kripke structures that represent modal logic problems.

## 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.
- Johan Benthem, Jan van Eijck, Malvin Gattinger, and Kaile Su. 2018. *Symbolic model checking for dynamic epistemic logic — s5 and beyond\**. *Journal of Logic and Computation*, 28:367–402.

User:  
 [Premise] There are two persons. Everyone is visible to others. There is a mirror in the room. It is publicly announced that someone's forehead is muddy. It is publicly announced that not everyone's forehead is muddy. It is publicly announced that not everyone's forehead is muddy. [Hypothesis] Robert can now know whether or not everyone's forehead is muddy. True or false ?

GPT-4:  
 False. Based on the information provided, Robert cannot know for certain whether or not everyone's forehead is muddy. The statements provided are: There are two persons. Everyone is visible to others. There is a mirror in the room. It is publicly announced that someone's forehead is muddy. It is publicly announced that not everyone's forehead is muddy. Since it is publicly announced that not everyone's forehead is muddy, Robert can deduce that at least one person does not have a muddy forehead. However, this information alone does not allow Robert to definitively determine whether or not his own forehead is muddy.

Figure 4: GPT4 conversation

- Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. *arXiv preprint arXiv:2304.01373*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. [Transformers as soft reasoners over language](#). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3882–3890. International Joint Conferences on Artificial Intelligence Organization. Main track.
- Michael Cohen. 2021. Exploring roberta's theory of mind through textual entailment. *philarchive*.
- F. de Waal. 2016. *Are We Smart Enough to Know How Smart Animals Are?* W. W. Norton.
- Lasse Dissing and Thomas Bolander. 2020. [Implementing theory of mind on a robot using dynamic epistemic logic](#). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1615–1621. International Joint Conferences on Artificial Intelligence Organization. Main track.
- Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. [A framework for few-shot language model evaluation](#).
- Kyle Hamilton, Aparna Nayak, Bojan Božić, and Luca Longo. 2022. Is neuro-symbolic ai meeting its promises in natural language processing? a structured review. *Semantic Web*, pages 1–42.
- Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.
- Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. [Revisiting the evaluation of theory of mind through question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.
- Ester Navarro, Sara Anne Goring, and Andrew R. A. Conway. 2020. The relationship between theory of mind and intelligence: A formative g approach. *Journal of Intelligence*, 9.
- Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. [Evaluating theory of mind in question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400, Brussels, Belgium. Association for Computational Linguistics.
- OpenAI. 2023. [Gpt-4 technical report](#).
- Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*.
- Kyle Richardson and Ashish Sabharwal. 2022. Pushing the limits of rule reasoning in transformers through natural language satisfiability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11209–11219.

- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. [Social IQa: Commonsense reasoning about social interactions](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.
- Damien Sileo and Marie-Francine Moens. 2022. Probing neural language models for understanding of words of estimative probability. *arXiv preprint arXiv:2211.03358*.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. [Diagnosing the first-order logical reasoning ability through LogicNLI](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Iris Van De Pol, Iris Van Rooij, and Jakub Szymanik. 2018. Parameterized complexity of theory of mind reasoning in dynamic epistemic logic. *Journal of Logic, Language and Information*, 27:255–294.
- Hans Van Ditmarsch and Willem Labuschagne. 2007. My beliefs about your beliefs: a case study in theory of mind and epistemic logic. *Synthese*, 155:191–209.
- Jan van Eijck. 2014. Dynamic epistemic logics. *Johan van Benthem on logic and information dynamics*, pages 175–202.
- Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas 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*.