How Well Do People Perform on Novel Logic Puzzles Requiring Higher-Order Theory of Mind?

Andreea Minculescu*[12], Jakob Dirk Top[2], Rineke Verbrugge[2], Harmen de Weerd[2]
[1]Department of Intelligent Systems, Delft University of Technology, Delft, Netherlands
[2]Bernoulli Institute, University of Groningen, Groningen, Netherlands
A.Minculescu@tudelft.nl; {J.D.Top, L.C.Verbrugge, Harmen.de.Weerd}@rug.nl

Abstract

Theory of mind (ToM) refers to the ability to reason about the behaviour of others and oneself by attributing internal mental states, such as knowledge, desires and intentions. ToM can be applied recursively – for example, "Amy thinks that Bernard knows that it is raining" is said to be a second-order ToM statement from the reader's perspective. Past research suggests that there is a limit to the number of times humans can apply ToM recursively – for example, they tend to use up to second-order ToM reasoning in strategic games. In the present study, we propose and conduct a novel human experimental design, in which different orders of ToM reasoning in the logic puzzle "Cheryl's Birthday" can be distinguished. Results show that higher-order ToM reasoning is associated with longer times to solve the puzzle(s) and a higher rate of mistakes.

Keywords: theory of mind, logic puzzles, recursive reasoning

Introduction

Imagine you want to send an email to your coworkers concerning a complaint raised by a customer. You just finished writing the body of the email detailing the situation and now you are about to insert the email addresses of the recipients. You are now faced with one important decision: whether to use CC or BCC – like CC, BCC sends copies of the email to additional recipients but, unlike CC, these recipients are not visible to one another. Your decision depends on whether you want certain recipients to *know* that other recipients *know* the content of the message – perhaps your co-workers are part of different departments and you do not wish to create unnecessary conflict. Therefore, you have to understand, from their perspective, how this knowledge could potentially affect their behaviour. This ability to reason about the behaviour of others and oneself by attributing internal mental states, such as knowledge, desires and intentions, is known as *theory of mind* (ToM) (Dennett, 1971; Premack & Woodruff, 1978).

Recursive ToM Reasoning and Its Limits

Theory of mind can be applied recursively (see Verbrugge, 2009 for an overview). Zero-order ToM describes world facts, such as "The sky is red" (note that the truth value of the statement is irrelevant here), while $(x+1)$-order reasoning attributes x-order reasoning to the other agent. For example, in the sentence "Albert knows that Bernard thinks that the sky is red", Albert uses first-order ToM to reason about Bernard who is using zero-order ToM, reasoning about a world fact. Now we, the authors, just made a second-order attribution to Albert and a first-order attribution to Bernard.

Previous research suggests that there is a limit to the number of times humans can apply ToM recursively, as they tend to use up to second-order ToM in strategic games (de Weerd et al., 2018; Devaine et al., 2014; Nagel, 1995). However, higher orders of recursive ToM use have been experimentally observed – for example, evidence of fourth-order ToM use has been found in story comprehension tasks (Kinderman et al., 1998; Stiller & Dunbar, 2007) and in the Mod game (through a training regime) (Veltman et al., 2019). Despite its complexity, fourth-order ToM is vital for some real-life activities, from reading literature (Zunshine, 2006) to achieving hidden war goals as a double spy (Barbier, 2007).

Children begin to distinguish their own beliefs from those of others between three and five years of age and only later, between six and nine years old, do they begin making correct second-order attributions (Perner, 1988). Arslan, Taatgen, and Verbrugge (2017) showed that five-year-old children that cannot yet apply second-order ToM predominantly use first-order ToM, not zero-order ToM. More generally, this suggests that wrong answers in ToM tasks are typically one order below the target order of ToM reasoning.

Assessing ToM abilities seems to be highly dependent on task domain. Flobbe et al. (2008) showed that most 8–10 year old children who succeeded in a second-order false belief task could not yet apply second-order ToM in a strategic game or a grammatical task. One explanation for this gap between children's *understanding* of second-order ToM and their ability to *apply* it may be a *serial processing bottleneck* (Verbrugge, 2009): Processing embedded beliefs requires intermediate reasoning steps that are temporarily kept in working memory. Due to working-memory limitations, the intermediate results are sent to long-term memory for later retrieval, which is slow and error-prone (Anderson and Schooler, 2000). Supporting the serial processing bottleneck hypothesis, Arslan, Hohenberger, and Verbrugge (2017) showed that children with high scores in a complex working memory task performed better in a second-order false belief task.

Dynamic epistemic logic (van Ditmarsch et al., 2007) was designed to formalize ToM concepts, as it allows one to model how (recursive) reasoning about knowledge changes in response to new information. However, classical approaches to modeling ToM in dynamic epistemic logic typically assume common knowledge of perfect rationality, which does not account for the upper limit to recursive ToM use that is commonly found in behavioural research. In recent years,

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

several formal systems where the agents *do* have limitations on their ToM reasoning have been proposed, such as Arthaud and Rinard (2023), Top et al. (2023), and Zhang et al. (2021). To verify whether such systems accurately model human behaviour, we need experimental data of humans performing tasks that are easily modelled in dynamic epistemic logic, such as epistemic puzzles. However, very little data on such tasks seems to exist, and there are some caveats to the data that is available. For example, Jonker and Treur (2003) use one participant only, while Hayashi (2002) presents puzzles in a fixed order, allowing order effects to occur. Furthermore, the Wise Men puzzle (McCarthy, 1990) is used, which has only two answers, making probabilistic guessing a feasible strategy. Similarly, the *Aces and Eights* puzzle used in Zhang et al. (2021) has only four answer options. In addition, many of the higher-level puzzles in Aces and Eights can be solved by answering "I don't know", as explained in Top et al. (2023). In the remainder of the paper, we show how our experimental design addresses such shortcomings.

Current Study

The present study aims to achieve two goals:

1. *Propose a novel experimental design and generate a dataset for use as a benchmark for future ToM studies*.
   To this end, an experimental setup was designed such that ToM reasoning was the only viable strategy to reach the correct answer. As part of the experiment, participants were asked to solve a series of eight puzzles inspired by "Cheryl's Birthday" (e.g., van Ditmarsch et al., 2017).[1] In the "Cheryl's Birthday" puzzle, the participant has to determine Cheryl's birthday from a list of possible dates, based on conversational clues provided by Cheryl's two friends, Albert and Bernard. Since the conversation between Albert and Bernard pertains to their own limited knowledge of Cheryl's birthday, only by reasoning about the boys' knowledge (i.e., applying ToM reasoning) can a participant reach the correct answer.

   The aim of the experiment was to ensure that the upper bound of recursive ToM use can be measured: One should expect that puzzles requiring ToM orders higher than the bound would not be solvable by the majority of participants. All participants solved (up to) eight puzzles requiring varying orders of ToM reasoning and set in different contexts in a randomized order. Each puzzle had ten possible answers, making probabilistic guessing a poor strategy.

2. *Investigate whether the limit to recursive ToM use can be identified in the generated dataset*.
   To this end, we conducted a statistical analysis on the time required to solve the puzzles and on accuracy. We expected to find that the time to solve a puzzle increases and that the accuracy decreases with the ToM order, since higher-order ToM requires more reasoning steps to solve correctly.

[1] A modern variant of Freudenthal's 'Sum and Product' puzzle (van Ditmarsch et al., 2005).

Methods

Participants

Forty-nine Bachelor's students (32 female; mean age 20.18, ranging from 18 to 24) at the University of Groningen participated in exchange for monetary compensation. As (international) university students, it is assumed that all participants demonstrated high linguistic proficiency and working memory capacity, although no explicit tests were conducted. Initially, one Master's student accidentally participated in the experiment and was excluded from all further analyses.

The participants reported that they had no formal training in modal/epistemic logic and had taken no Game Theory courses at the time of the experiment. It was not stated explicitly as a requirement that participants should not have solved the "Cheryl's Birthday" puzzle before, out of concern that they might look it up online. Most participants reported not having heard about the "Cheryl's Birthday" puzzle or similar puzzles (45 resp. 43 participants) prior to the experiment. On a scale from 1 to 10, 10 being equivalent to "I feel very happy", participants reported a mean mood score of 6.94.

Puzzles

Four unique "Cheryl's Birthday" puzzle texts, one for each order of theory of mind (ToM) from one to four, were adapted from van Ditmarsch et al. (2017). The puzzles were presented in a randomized order, in order to account for potential learning effects. The puzzle text and the solution for the second-order ToM puzzle are presented below. The formulation for the other puzzles only differs in terms of the set of birthday options and the dialogue between Albert and Bernard (for the latter, see Table 1).[2]

"Cheryl's Birthday" Second-Order ToM Version:

Albert and Bernard just became friends with Cheryl, and they want to know when her birthday is. Cheryl writes down a list of 10 possible dates and tells them that one of them is her birthday: May 17, May 18; June 14; July 16, July 18; August 15, August 16, August 17; September 14, September 15.

Cheryl then tells only to Albert the month of her birthday, and tells only to Bernard the day of her birthday. (And Albert and Bernard are aware that she did so.) Everybody knows that Albert, Bernard and Cheryl don't make any reasoning mistakes and never lie.

Albert and Bernard now have the following conversation:
***Albert**: "I don't know when Cheryl's birthday is."*
***Bernard**: "I didn't know at first, but now I know."*

When is Cheryl's birthday?

Here follows the solution to the above puzzle: Henceforth, let Albert be A, Bernard B, and Cheryl C. Let "days" refer to the numbers on C's list (14-18), let "months" refer to the calendar

[2] Note that ToM order makes a real difference in difficulty, whilst change of scenario and configuration keeps puzzles isomorphic.

Puzzle type	Dialogue between A and B
First-order	B: "I know when C's birthday is."
Second-order	A: "I don't know when C's birthday is." B: "I didn't know at first, but now I know."
Third-order	A: "I know that you don't know when C's birthday is." B: "I didn't know at first, but now I know."
Fourth-order	B: "I know that you know that I don't know when C's birthday is. A: "I didn't know at first but now I know."

Table 1: A and B's dialogue structures for all puzzle types.

months on C's list (May-September), and let "birthday" refer to a ⟨month, day⟩ combination on C's list (e.g., May, 15).

Now, A claims that he does not know C's birthday. This means that C's birthday is not in a month associated with only one day, because then A would have known C's birthday. Therefore, C's birthday cannot be on June, 14 – recall that June is the only month associated with only one day. Next, B claims that A's statement helped him find C's birthday. As a perfect reasoner, B must have completed the step described above correctly. For B to now know the solution, C's birthday must be associated with a unique day. Therefore, the day must be 14 and C's birthday must be September, 14, as June, 14 had already been eliminated as an option.

All Puzzle Variations We use a 4x4x4 design[3]:

1. *ToM order*: The ToM order required to solve a puzzle is modeled based on the number of perspective switches a participant would need to perform to process the dialogue between A and B. The puzzle discussed above is an example of a second-order ToM puzzle because the participant would need to perform two perspective switches to understand the dialogue, namely by using B's perspective who uses A's perspective to find the answer. The dialogue between A and B for all puzzles can be found in Table 1.

2. *Scenario*: The puzzle examples discussed so far showcased one scenario, namelyfinding C's **birthday**. Three other scenarios were introduced, where only the target properties (i.e., month and day) were changed. In the **drink** scenario, A and B are challenged to find out how C likes to have her coffee: She tells A the size of the coffee (e.g., large, small) and B the temperature (e.g., iced, lukewarm). In the **toy** scenario, A and B have to find C's favourite childhood toy in her room: She tells A the location of the toy (e.g., on the armchair, on the windowsill) and B the type of toy (e.g., doll, clown). Finally, in the **hair** scenario, A and B have to locate C's friend D, in a busy train station and C describes D's hair: She tells A the hair style (e.g., curly, straight) and B the hair color (e.g., orange, blue).

   The three scenarios were generated based on the birthday scenario by creating a one-to-one correspondence between the days and months on the one hand and the other scenario's target properties on the other. For example, in the drink scenario, ⟨*May, June, July, August, September*⟩ were replaced, in this exact order, with ⟨*On the table, On the bed, On the floor, On the armchair, On the windowsill*⟩ and ⟨*14, 15, 16, 17, 18*⟩ with ⟨*doll, bunny, clown, cat, train*⟩.

3. *Configuration*: Additional puzzles can be generated by mirroring properties while keeping the remainder of the puzzle text unchanged. For example, when mirroring (only) the month property, ⟨*May, June, July, August, September*⟩ is replaced, in this exact order, with ⟨*September, August, July, June, May*⟩, while keeping the day property constant. Thus, for each ⟨*scenario* × *ToM order*⟩ combination, four configurations were generated: i) the original configuration, ii) mirroring of only the first property, iii) mirroring of only the second property, and iv) mirroring of both properties. This was done to increase the drawing pool of available puzzles.

[3]All text, code, and puzzle variants used in the experiment can be found at https://github.com/AndreeaMinculescu/Cheryl-Puzzle

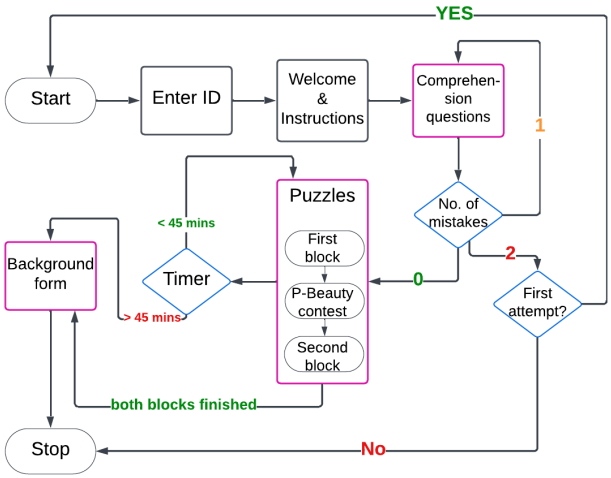

Figure 1: Experimental workflow. Main stages are marked with pink squares, decision points with blue diamonds.

Procedure

Figure 1 shows the procedure employed when running the experiments. Prior to the experiment, the participants were informed that they would earn 7.5€ for their participation. Additionally, they *could* earn a bonus monetary reward of 2.5€ (so, in total, 10€) if their answer to one randomly chosen puzzle had been correct. This was done following Azrieli et al. (2018) in order to incentivize participants to solve correctly as many puzzles as possible.

Comprehension Questions Participants were tested on their reading comprehension and understanding of epistemic concepts (e.g., perfect logicians and common knowledge). Importantly, at this stage, participants were not asked to solve any example puzzles in order to avoid priming them prior to the actual experimental phase.

Participants were allowed one mistake in the comprehension stage. In that case, the experimenter explained the instruction text again and answered any questions. All participants passed the comprehension stage within two attempts.

Puzzle Blocks Next, the participants were asked to solve eight puzzles ranging in difficulty, presented in two blocks of four puzzles each. Participants received and were encouraged to make use of pen and paper to write down intermediate steps, in order to decrease their working-memory load.

In total, 64 puzzles had been generated. For each puzzle, there was one unique solution, and the dialogue between A and B contained precisely the information needed to reach the solution. ToM reasoning at the specified order is necessary for solving the puzzles, because only the dialogue between A and B, which involves explicit knowledge attribution, conveys any relevant clues. Furthermore, we posit that any alternative strategies are unfeasible: Participants did not follow any (dynamic) epistemic logic courses and random guessing was unlikely to be consistently successful, as the guessing chance was low (8.3% or 1 in 12 possible answers).

The puzzles were allocated such that each participant would encounter each order of ToM and each scenario exactly twice overall but not necessarily in the same pairing – one participant could see a second-order puzzle of the toy scenario while another participant could see a fourth-order puzzle of the toy scenario. Once the ToM order and scenario had been selected, a random puzzle was selected amongst the four possible types of mirroring configurations.

In each block, each scenario and each ToM order were shown exactly once. In the first block, each scenario was associated with one unique ToM order. Thus, in the first block, each participant would encounter one first-order toy puzzle, one second-order drink puzzle, one third-order birthday puzzle, and one fourth-order hair puzzle, with only the mirroring configurations varying. In the second block, there was no such restriction (i.e., each ToM order could, in principle, occur with any scenario).[4]

Participants were instructed to carefully read the puzzle text and to select the solution from a list of twelve options: the ten options on C's list, "No solution" and "Multiple solutions". Furthermore, "I don't know" was set as the default option and was recorded when the participant did not select any option in the drop-down menu. Participants had 45 minutes to solve as many of the eight puzzles as possible. If the time limit had passed or they finished all eight puzzles, they were automatically redirected to the Background Form.[5]

Background Form Finally, participants were asked to report their contact details (name, email address), demographic data (age, gender), educational background (study program, formal training in logic), and overall experience with the experiment (see the Results section for an analysis of the answers). To ensure anonymity, all contact information was stored separately, whereas all other data was anonymized.

[4]This imbalanced design was the result of a bug in the original code, but since no effect of scenario was found (see Results), we argue that the bug did not affect the conclusions of the study.

[5]A "p-Beauty Contest" (Nagel, 1995) was also presented after the first and before the second block. Since participants did not exhibit unexpected patterns, we do not discuss the results here.

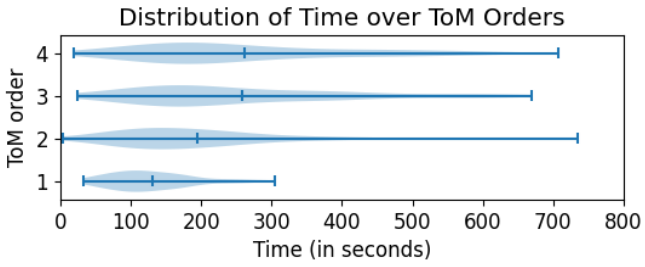

Figure 2: Violin plots of the time to solve a puzzle across the four ToM orders for the 42 participants who completed all eight trials. The left-most and right-most vertical bars show the two extremes and the middle bar shows the mean.

Results

The analysis code was written in Python 3.10 and R 4.2.2.

Main Effects

The statistical analyses discussed in this section were conducted only on the data pertaining to participants who finished all eight trials within the allocated forty-five minutes (42 out of 49 participants). This was done to ensure an equal distribution over all orders of ToM and all scenarios.

Next, we discuss the effects scenario and order of ToM have on puzzle solving time and accuracy, respectively.

Time to Reach Answers The time to reach an answer was measured from the moment the puzzle text was first shown to the participant until the click of the "Submit answer" button. Two questions of interest arise:

1. Does the time to solve a puzzle differ significantly across orders of ToM?

Figure 2 shows the distribution of the solving time across the four orders of ToM. Participants required the lowest amount of time to solve first-order puzzles (M=130.91), followed by second-order (M=194.01), and lastly, third-order (M=258.02) and fourth-order (M=261.69) puzzles.

A Shapiro-Wilk test on the within-group variability revealed that the (log-transformed) time was not normally distributed (W=0.95, p <0.001). Therefore, a Kruskal-Wallis test was conducted on the solving time, which revealed statistical significance ($\chi^2(3) = 57.29$, $p < 0.001$). A post-hoc Dunn test, adjusted for multiple comparisons (Holm, 1979), revealed that every two ToM orders, except for the third and fourth orders ($p = 0.8$), were significantly different from each other in terms of solving time.

Interestingly, we find no evidence for an interaction between accuracy and ToM order on (log-transformed) solving time (ANOVA test: $F(3) = 0.87$, $p = 0.453$), which suggests that puzzles of increasing ToM orders do not increase

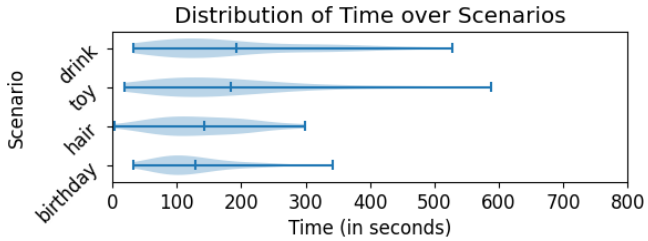

Figure 3: Violin plots of the time to solve a puzzle across the four scenarios in the second block for the 42 participants who completed all eight trials. The left-most and right-most vertical bars show the extremes and the middle bar the mean.

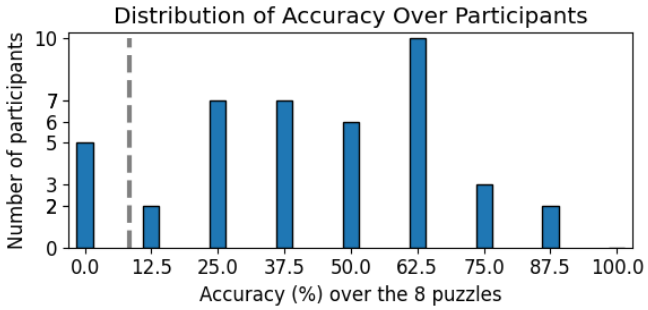

Figure 4: Distribution of accuracy over the 42 participants who completed all eight trials. The dashed line shows the chance level (namely $\frac{1}{12} * 100 \approx 8.3\%$).

the difference in solving time between correct and incorrect answers.

2. Does the time to solve a puzzle differ significantly across scenarios?

The following analysis was conducted only on the data from the second block due to the inconsistent design of the two blocks. Figure 3 shows the distribution of the solving time across the four scenarios. Participants seem to require a similar amount of time to solve puzzles of each scenario type — birthday: M=128.76; hair: M=143.08; toy: M=184.65; drink: M=192.85. A Shapiro-Wilk test on the within-group variability revealed that the (log-transformed) time was not normally distributed (W=0.94, p <0.001). A Kruskal-Wallis test did not reveal statistically significant differences in solving time across scenarios ($\chi^2(3) = 7.38$, p =0.060).

Accuracy of Answers Let us define a participant's accuracy as the percentage of correct answers.

Figure 4 shows the frequency distribution of accuracy over the 42 participants who finished all eight trials. Twenty-one participants answered 50% or more of the puzzles correctly. Importantly, five participants failed to answer any puzzle correctly, so their accuracy was lower than the probability of simply guessing the correct answer (the gray dotted line in the figure). The chance of guessing is computed as $\frac{1}{12} * 100 \approx 8.3\%$, where 12 is the total number of possible

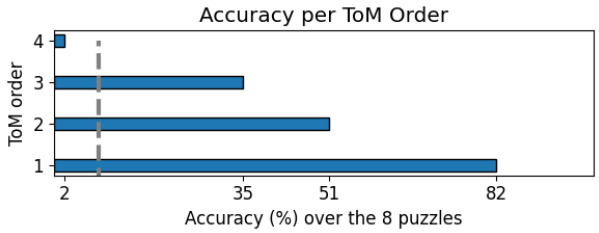

Figure 5: Bar chart of the accuracy associated with each ToM order for the 42 participants who completed all eight trials. The dashed line shows the chance level ($\frac{1}{12} * 100 \approx 8.3\%$).

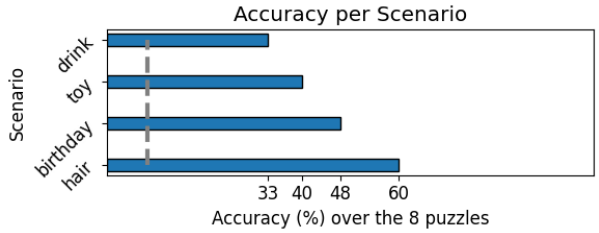

Figure 6: Bar chart of the accuracy associated with each scenario in the second block for the 42 participants who completed all eight trials. The dashed line shows the chance level ($\frac{1}{12} * 100 \approx 8.3\%$).

answers for any puzzle. Two questions of interest arise:

1. Does accuracy differ significantly across orders of ToM?

As shown in Figure 5, participants solved first-order puzzles with an accuracy of 82.1%, second-order puzzles with an accuracy of 51.2%, third-order puzzles with an accuracy of 34.5% and fourth-order puzzles with an accuracy of 2.4%. Given the subset of 42 participants, the accuracy associated with the fourth order is lower than the probability of simply guessing the correct answer (the gray line in the figure), which suggests that these participants were mostly unsuccessful in solving these puzzles. As expected, a Chi-square test revealed that accuracy was significantly different between orders of ToM ($\chi^2(3) = 114.09$, $p < 0.001$).

2. Does accuracy differ significantly across scenarios? The following analysis was conducted only on the data from the second block due to the inconsistent design of the two blocks: In the first block, each scenario was associated with exactly one ToM order.

As shown in Figure 6, participants solved the birthday scenario with an accuracy of 47.6%, the hair scenario with an accuracy of 59.5%, the drink scenario with an accuracy of 33.3% and the toy scenario with an accuracy of 40.5%. For all scenarios, the accuracy was higher than the probability of guessing the correct answer (gray line in the figure). As expected, a Chi-square test revealed that accuracy was not significantly different between scenarios ($\chi^2(3) = 6.34$, $p =$ 0.096). As the difference in accuracy between scenarios is not significant, it can likely be explained by the design of the

second block: Each scenario was associated with only three out of four ToM orders, while the remaining ToM order occurred only in the first block (see Methods for details). For example, in the second block, the hair scenario was never associated with fourth-order ToM, which can explain why its accuracy is qualitatively higher than for the other scenarios. As shown before, participants solved first-order ToM puzzles correctly significantly more often than fourth-order ones.

Background Form

The following statistical tests were conducted on all forty-nine participants. In the background form, participants were asked to indicate whether, before the experiment, they had encountered the "Cheryl's Birthday" puzzle or a similar puzzle. Five participants indicated that they indeed had and, based on a proportion test, these participants answered correctly significantly more often than participants who had not come in contact with "Cheryl's Birthday" or any similar puzzle before the experiment ($\chi^2(1)$=3.69, p =0.027) [6].

Participants were asked to rate on a scale from one to ten how difficult they found the instructions shown on the interface, where ten meant that they easily understood all instructions. Thirty-nine participants reported a score higher than five (i.e., the instructions were reasonably easy) but, based on a proportion test, they did not perform significantly better than participants who reported a score lower than or equal to five ($\chi^2(1)$=1.13, p =0.123).

Participants were then asked to rate on a scale from one to ten how difficult they found the puzzles, where ten meant that they found the puzzles very difficult to solve. Thirty-five participants reported a score higher than five (i.e., the puzzles were reasonably difficult) and, based on a proportion test, these participants performed correctly significantly less often than the participants who reported a score lower than or equal to five ($\chi^2(1)$=12.128, p <0.001).

Finally, participants were asked to rate on a scale from one to ten how much they enjoyed solving the puzzles, where ten meant that they greatly enjoyed solving the puzzles. Forty-one participants reported a score higher than five (i.e., they enjoyed solving the puzzles) but, based on a proportion test, these participants did not perform significantly better than the participants who reported a score lower than or equal to five ($\chi^2(1)$=0.02, p =0.439).

Discussion

Future Research

Our 4 *ToM orders* × 4 *scenarios* × 4 *configurations* design allowed for some interesting investigations into recursive ToM use. Firstly, we could clearly distinguish between performance at different orders of recursive ToM use. Secondly, the four-scenarios design mitigated the issue of task dependency (Flobbe et al., 2008) and is a first step towards a more robust measuring of ToM reasoning abilities in epistemic puzzles. Finally, the configuration design provides a simple way of generating a large number of seemingly different puzzles, which require the same processing steps to solve correctly.

A possible confound is that higher-order puzzles use more complex dialogue and require more reasoning steps. This increased demand on working memory (WM) could explain some of our results. Future work could fix the amount of reasoning steps required to see whether our results still hold. Apperly et al. (2010), Bradford et al. (2018), and Meijering et al. (2013) indeed show that participants are slower and less accurate when a task uses ToM framing, even if conditions are otherwise identical.

Many open questions remain. For instance, Verbrugge et al. (2018) showed that a stepwise training regime, in which items are presented in increasing order of difficulty, leads to improved performance in second-order versions of the Matrix Game. Thus, it would be interesting to investigate whether participants' performance would be improved by implementing a similar training regime — for example, instead of randomizing the order in which puzzles are shown, the puzzles could instead be grouped by the required ToM order.

It has lately been of interest whether large language models (LLMs) have recursive ToM reasoning abilities (Strachan et al., 2024; van Duijn et al., 2023; Verma et al., 2024). Testing an LLM's recursive ToM reasoning abilities on an extensively studied epistemic puzzle can prove problematic, because the LLM might have have been trained on similar puzzles, which could artificially enhance its ToM performance. Since the current study introduces a novel experimental design, we posit that (variations of) our epistemic puzzles could be used to safely investigate ToM abilities in LLMs.

Finally, it would be interesting to investigate the underlying mechanisms involved in recursive ToM reasoning for these epistemic puzzles. This can be done by, for example, building computational models of human cognition and testing their predictions on the human data collected as part of the current study. In fact, we, the authors of the paper, are currently working on implementing a computational model based on epistemic logic, strongly inspired by the works of Top et al. (2023) and Zhang et al. (2021). Other epistemic logics with ToM limitations, such as the one in Arthaud and Rinard (2023), can also be verified using our novel data.

Conclusion

In this study, we introduced a novel experimental design that can be used to distinguish theory of mind (ToM) reasoning at different recursive orders from other strategies in the epistemic puzzle of *Cheryl's Birthday*. This experimental design allows one to generate a variety of logically equivalent puzzles for each ToM order, presented in different contextual scenarios. Moreover, the design allows experimenters to determine whether participants can correctly apply higher-order theory of mind reasoning, but also to investigate what happens if participants fail to do so.

[6] We also ran our analysis without these participants. However, none of the findings presented changed significantly, so we do not report those results here.

Acknowledgements

This research was funded by the project 'Hybrid Intelligence: Augmenting Human Intellect', a 10-year Gravitation programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, grant number 024.004.022. Participants were paid using the University of Groningen education budget.

References

Anderson, J., & Schooler, L. (2000). The adaptive nature of memory. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 557–570). Oxford University Press.

Apperly, I. A., Carroll, D. J., Samson, D., Humphreys, G. W., Qureshi, A., & Moffitt, G. (2010). Why are there limits on theory of mind use? Evidence from adults' ability to follow instructions from an ignorant speaker. *Quarterly Journal of Experimental Psychology*, *63*(6), 1201–1217.

Arslan, B., Hohenberger, A., & Verbrugge, R. (2017). Syntactic recursion facilitates and working memory predicts recursive theory of mind. *PLoS ONE*, *12*(1), e0169510.

Arslan, B., Taatgen, N., & Verbrugge, R. (2017). Five-year-olds' systematic errors in second-order false belief tasks are due to first-order theory of mind strategy selection: A computational modeling study. *Frontiers in Psychology*, *8*, 275.

Arthaud, F., & Rinard, M. (2023). Depth-bounded epistemic logic. In R. Verbrugge (Ed.), *Proceedings of the 19th conference on Theoretical Aspects of Rationality and Knowledge (TARK 23)* (pp. 46–65).

Azrieli, Y., Chambers, C., & Healy, P. (2018). Incentives in experiments: A theoretical analysis. *Journal of Political Economy*, *126*(4), 1472–1503.

Barbier, M. (2007). *D-Day deception: Operation Fortitude and the Normandy invasion*. Greenwood Press.

Bradford, E. E. F., Gomez, J.-C., & Jentzsch, I. (2018). Exploring the role of self/other perspective-shifting in theory of mind with behavioural and EEG measures. *Social Neuroscience*, *14*(5), 530–544.

de Weerd, H., Diepgrond, D., & Verbrugge, R. (2018). Estimating the use of higher-order theory of mind using computational agents. *The B.E. Journal of Theoretical Economics*, *18*(2), 20160184.

Dennett, D. (1971). Intentional systems. *The Journal of Philosophy*, *68*(4), 87–106.

Devaine, M., Hollard, G., & Daunizeau, J. (2014). The social Bayesian brain: Does mentalizing make a difference when we learn? *PLoS Computational Biology*, *10*(12), e1003992.

Flobbe, L., Verbrugge, R., Hendriks, P., & Krämer, I. (2008). Children's application of theory of mind in reasoning and language. *Journal of Logic, Language and Information*, *17*, 417–442.

Hayashi, H. (2002). Possibility of solving complex problems by recursive thinking. *The Japanese Journal of Psychology*, *73*(2), 179–185.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Jonker, C., & Treur, J. (2003). Modelling the dynamics of reasoning processes: Reasoning by assumption. *Cognitive Systems Research*, *4*(2), 119–136.

Kinderman, P., Dunbar, R., & Bentall, R. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, *89*(2), 191–204.

McCarthy, J. (1990). Formalization of two puzzles involving knowledge. *Formalizing Common Sense: Papers by John McCarthy*, 158–166.

Meijering, B., van Rijn, H., Taatgen, N. A., & Verbrugge, R. (2013). Reasoning about diamonds, gravity and mental states: The cognitive costs of theory of mind. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*, 3026–3031.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American Economic Review*, *85*(5), 1313–1326.

Perner, J. (1988). Higher-order beliefs and intentions in children's understanding of social interaction. In J. W. Astington, P. L. Harris, & D. R. Olson (Eds.), *Developing theories of mind* (pp. 271–294).

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

Stiller, J., & Dunbar, R. (2007). Perspective-taking and memory capacity predict social network size. *Social Networks*, *29*(1), 93–104.

Strachan, J., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M., & C., B. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 1–11.

Top, J. D., Jonker, C., Verbrugge, R., & de Weerd, H. (2023). Predictive theory of mind models based on public announcement logic. In N. Gierasimczuk & F. R. Velázquez-Quesada (Eds.), *Dynamic Logic. New Trends and Applications: 5th International Workshop, DaLí 2023* (pp. 85–103, Vol. 14401). Springer.

van Ditmarsch, H., Hartley, M., Kooi, B., Welton, J., & Yeo, J. (2017). Cheryl's birthday. *Electronic Proceedings in Theoretical Computer Science*, *251*, 1–9.

van Ditmarsch, H., van der Hoek, W., & Kooi, B. (2007). *Dynamic Epistemic Logic* (Vol. 337). Springer Science & Business Media.

van Ditmarsch, H., Ruan, J., & Verbrugge, L. (2005). Model checking sum and product. *AI 2005: Advances in Artificial Intelligence: 18th Australian Joint Conference on Artificial Intelligence*, 790–795.

van Duijn, M., van Dijk, B., Kouwenhoven, T., de Valk, W., Spruit, M., & van der Putten, P. (2023, December). Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In J. Jiang, D. Reitter, & S. Deng (Eds.), *Pro-*

ceedings of the 27th conference on computational natural language learning (CoNLL) (pp. 389–402). Association for Computational Linguistics.

Veltman, K., de Weerd, H., & Verbrugge, R. (2019). Training the use of theory of mind using artificial agents. *Journal on Multimodal User Interfaces*, *13*(1), 3–18.

Verbrugge, R. (2009). Logic and social cognition: The facts matter, and so do computational models. *Journal of Philosophical Logic*, *38*(6), 649–680.

Verbrugge, R., Meijering, B., Wierda, S., van Rijn, H., & Taatgen, N. (2018). Stepwise training supports strategic second-order theory of mind in turn-taking games. *Judgment and Decision Making*, *13*(1), 79–98.

Verma, M., Bhambri, S., & Kambhampati, S. (2024). Theory of mind abilities of large language models in human-robot interaction: An illusion? *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 36–45.

Zhang, C., Ham, H., & Holliday, W. (2021). Does Amy know Ben knows you know your cards? A computational model of higher-order epistemic reasoning. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, 2588–2594.

Zunshine, L. (2006). *Why we read fiction: Theory of mind and novel*. The Ohio State University Press.