



university of
 groningen

faculty of mathematics and
 natural sciences

artificial intelligence

CUTTING CHERYL'S BIRTHDAY CAKE: MODELING THEORY OF MIND ORDERS IN PUBLIC ANNOUNCEMENT LOGIC

Master's Project Thesis

Andreea Minculescu
 a.minculescu@student.rug.nl

July 7, 2024

Main Supervisor: Prof. L.C. Verbrugge
 Secondary Supervisors: dr. H.A. de Weerd,
 J.D. Top, MSc

Artificial Intelligence
 University of Groningen, The Netherlands

In multi-agent systems, effective interaction, from coordination and negotiation to outright competition, is especially crucial for high-stake applications. For example, consider rescue missions, where robots evaluate and maneuver risky operations while humans rescue the victims: In such situations, there is simply no time for humans to explicitly instruct the robots; instead the robots need to anticipate and understand the needs of the human rescuers and of the human victims. To this end, it is often helpful to specify agents in terms of their mental states, such as knowledge, beliefs, plans and intentions, as well as to recursively specify the mental states of others in relation to oneself. But before we can implement such a comprehensive reasoning theory, we must first understand it within humans.

Theory of mind (ToM) is defined as the ability to reason about the behaviour of others and oneself by attributing mental states, such as beliefs, desires and knowledge. While humans are able to apply ToM recursively (e.g., “I know that you believe that they think...”), past research has shown that human recursive ToM use is limited in strategic games and that it often does not exceed second-order ToM. One possible explanation for this limitation of recursive ToM use is that embedded beliefs are processed serially through intermediate reasoning steps that are eventually sent to the long-term memory for later retrieval and retrieval from the long-term memory has been shown to take long and to be prone to errors - this is also known as the *serial processing bottleneck* hypothesis.

Dynamic epistemic logic has been historically used to formalize ToM concepts because 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 do not account for the upper limit to recursive ToM use, as revealed through behavioural research. At the same time, most behavioural research does not test assumptions about recursive ToM use on tasks that can easily be modeled in dynamic epistemic logic. The present study aimed to address this gap in the literature by: i) proposing and conducting a novel experiment on an epistemic puzzle and ii) proposing a novel modeling paradigm in public announcement logic (a variant of dynamic epistemic logic), whereby parts of complex ToM statements are sequentially removed until the new statement can be processed by the agent.

The first goal of the study was achieved by conducting an experiment where participants were asked to solve a series of *Cheryl's Birthday* epistemic puzzles that were set in different contexts (hereon, referred to as “scenarios”) and required different orders of ToM reasoning (specifically, first-order through fourth-order ToM reasoning) as the only viable strategy to reach the correct solution. I showed that: i) the time to solve a puzzle differed significantly across every two ToM orders, except for third-order and fourth-order puzzles (first-order: M=130.91 seconds; second-order: M=194.01 seconds; third-order: M=258.02 seconds; fourth-order: M=261.69 seconds) and ii) accuracy was significantly different across ToM orders (first-order: 82.1% correct answers; second-order: 51.2% correct answers; third-order: 34.5% correct answers; fourth-order: 2.4% correct answers). In other words, lower orders of ToM were associated

with lower solving time and higher accuracy (and vice-versa for higher order of ToM). Additionally, I showed that there was no effect of scenario on solving time and accuracy, respectively.

The second goal of the study was achieved by proposing a novel modeling paradigm and implementing it in public announcement logic: When participants encounter a high-order ToM statement that they cannot process, they sequentially remove knowledge operators, either from the front or from the back of the statement, until the ToM order is low enough to process the statement. I showed that a model that reduces higher-order ToM statements to second-order ToM statements captures systematic patterns in the behaviour of human participants that deviate from a model of perfect recursive ToM reasoning and a model of informed random behaviour. Additionally, I showed that removing operators from the front (left side) of a statement matches the behaviour of participants better than removing operators from the back (right side) of the statement.

Overall, the proposed experimental design allows one to distinguish theory of mind recursive reasoning at different recursive orders from other strategies in the epistemic puzzle of *Cheryl's Birthday* and I showed that differences in performance across different ToM orders are most likely due to limitations in recursive reasoning rather than the contextual information of the puzzles. Additionally, I showed that the epistemic puzzles can be modeled in public announcement logic and the proposed modeling paradigm sheds a critical light on alternative explanations of reasoning about complex ToM statements (such as ignoring those statements or probabilistic guessing). Two potential avenues for future research include investigating whether higher-order ToM reasoning abilities on the *Cheryl's Birthday* puzzle can be enhanced through training and considering ways how to formalize the cutting operation within public announcement logic.

ACKNOWLEDGEMENTS

The successful completion of this thesis would not have been possible without the support and guidance of others. Firstly, I would like to thank my supervisors, Rineke Verbrugge, Harmen de Weerd and J.D. Top, for their words of encouragement and constructive feedback. I am particularly grateful towards Rineke Verbrugge for her seemingly never-ending well of *knowledge* (no pun intended), towards Harmen de Weerd for his expertise on statistical methodologies and towards J.D. Top for his continuous stream of new ideas. Working with all of you has truly been a great experience and I will forever make use of the skills I honed under your tutelage.

Secondly, I would like to thank the University of Groningen for providing the funds and equipment necessary for running the “Cheryl’s Puzzle” experiment. Without this funding, the experiment would likely not have attracted such a large number of participants.

Last (but definitely not least), I would like to thank my parents, Mariana and Mircea, and my friends, David, Jakub, Ruhi, Andrada, Jeroen and Satchit, for providing the necessary moral support without which undertaking this project would have seemed a gruelling task. I would like to extend additional gratitude to my partner, David, for his much-needed words of wisdom that would often inspire me to overcome challenges at all stages of the project - this includes, but is not limited to, the intuition behind the cutting model.

1	Introduction	9
1.1	Limits of ToM Reasoning	9
1.2	ToM and Formalism	10
1.3	Current Study	12
2	Experimental Methods	15
2.1	Participants	15
2.2	Puzzles	15
2.2.1	“Cheryl’s Birthday” First-Order ToM	16
2.2.2	“Cheryl’s Birthday” Second-Order ToM	16
2.2.3	“Cheryl’s Birthday” Third-Order ToM	17
2.2.4	“Cheryl’s Birthday” Fourth-Order ToM	17
2.2.5	All Puzzle Variations	18
2.3	Procedure	19
2.3.1	Comprehension Questions	21
2.3.2	Cheryl’s Puzzles	21
2.3.3	Background Form	23
3	Results of the Empirical Evidence	25
3.1	Main Effects	25
3.1.1	Time to Reach Answers	25
3.1.2	Accuracy of Answers	27
3.2	Background Form	29
3.3	P-Beauty Contest	30
3.4	Conclusion	30
4	Modeling	33
4.1	Public Announcement Logic	33
4.2	Cheryl’s Puzzle Formalism	34
4.3	Models	35
4.3.1	Epistemic Model	35
4.3.2	Cutting Model	41
4.4	RFX-BMS	48

4.5	Additional Metric	50
5	Results Modeling	51
5.1	Model Configurations	51
5.2	Participant Answers	52
5.3	Performance Metrics	54
5.4	Results	55
5.4.1	Which of the five models described in Section 5.1 fits the highest proportion of the data?	56
5.4.2	Which of the five models described in Section 5.1 fits the highest proportion of the correct answers?	57
5.4.3	Which of the five models described in Section 5.1 fits the highest proportion of the wrong answers?	58
5.4.4	Which cutting direction fits the highest proportion of the third-order and fourth-order puzzle answers?	59
5.4.5	Which cutting direction fits the highest proportion of the third-order puzzle answers?	61
5.4.6	Which cutting direction fits the highest proportion of the fourth-order puzzle answers?	62
5.5	Conclusion	64
6	Discussion	65
6.1	Summary of Study	65
6.2	Critical Perspective	66
6.3	Future Research	69
6.3.1	A Cognitive Perspective	69
6.3.2	A Modeling Perspective	69
6.3.3	A Statistical Perspective	70
6.4	Conclusion	70
A	Experiment Materials	75
A.1	Consent Form	76
A.2	Interface	77
A.2.1	Comprehension Questions	79
A.2.2	Cheryl's Puzzle	83
A.2.3	Background Form	89

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 each other. Your decision depends on whether you want certain recipients to *know* that other recipients *know* the content of the message - perhaps your coworkers 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 and Woodruff, 1978].

In this chapter, I introduce the research context and the background, as well as discuss the focus of the current study: quantifying and modeling limits to the recursive use of ToM in an (epistemic) puzzle. In Section 1.1, I present past research that suggests that ToM can be applied recursively up to a certain limit and introduce a possible explanation for the existence of this cognitive limit. In Section 1.2, I introduce epistemic logic as the logic historically used to formalize knowledge. Lastly, in Section 1.3, I introduce the main focus points of the current study.

In Chapter 2, I present and justify the exper-

imental design of the (epistemic) puzzles used to measure ToM abilities. In Chapter 3, I present some interesting effects found in the dataset, which can motivate further research. In Chapter 4, I propose a novel strategy modeled in epistemic logic that may explain why participants give incorrect answers when faced with higher-order theory of mind. In Chapter 5, I discuss the quality of fit of the aforementioned model on the participants' data and compare this fit against two other alternative explanations. Lastly, in Chapter 6, I summarize the findings of the study, I discuss points of improvement for both the experimental design, the modeling approach, and the statistical methodologies and I suggest potential avenues for future research.

1.1 Limits of ToM Reasoning

Theory of mind, the ability to reason about the mental states of others and oneself, 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 or oneself. For example, in the sentence “Albert knows that Bernard thinks that the sky is red”, Albert uses first-order ToM to reason about another agent using zero-order ToM, namely Bernard, to reason about a world fact. It is then said that we, the readers, make a second-order attribution to Albert and a first-order attribution to Bernard.

Past research suggests that there is a limit to the number of times humans can apply ToM recur-

sively, as they tend to mostly use 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 [Stiller and Dunbar, 2007, Kinderman et al., 1998] and in the Mod game (through a training regime) [Veltman et al., 2019].

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 as cited by Verbrugge, 2009]. Arslan et al. [2017b] investigated whether children that cannot apply second-order ToM may favour zero-order or first-order ToM strategies in a false belief task. To this end, the authors built two cognitive models (an instance-based model in ACT-R [Anderson, 2007] and a reinforcement learning model) that reasoned about another agent as if that agent was reasoning using the same strategy as the model itself but at one ToM order lower (and this way of reasoning about other agents was applied recursively). The children’s data confirmed the predictions of the instance-based model: When children failed to use second-order ToM, they used first-order ToM strategies significantly more often than zero-order ToM.

Unfortunately, assessing ToM abilities seems to be highly dependent on the task domain. In Flobbe et al. [2008], the authors investigated the late development of second-order ToM in children between 8 and 10 years old in three different task domains: a strategic game, a grammatical task and a false belief task. While almost all children succeeded in the false belief tasks, their performance dropped noticeably for the other two tasks, even though all three tasks required the same order of ToM reasoning. This suggests that there is a gap between a child’s intentional understanding of second-order reasoning and their ability to exert it in an applied setting.

The task-dependent nature of the successful application of higher-order theory of mind lends itself to several explanations. One highly likely possibility is that there is a higher processing cost associated with higher-order cognition and, as a result, the already high demands of certain tasks may interfere with the successful application of the re-

quired order of theory of mind [Verbrugge, 2009].

To this end, Arslan et al. [2017a] investigated the role of working memory in the development of second-order false belief tasks. The authors brought forth the *serial processing bottleneck* hypothesis [see e.g., Verbrugge, 2009]. Processing serially embedded beliefs requires intermediate reasoning steps that are temporarily kept in the working memory. For example, processing the sentence “Albert knows that Bernard thinks that the sky is red” involves understanding that i) “the sky is red” is a world fact, ii) Bernard believes in that world fact and iii) Albert is aware of Bernard’s belief. Since the working memory is limited with regard to the amount of information that can be stored and processed at a time [Miller, 1956], these intermediate processing steps would need to be sent to the long-term memory for later retrieval, if necessary. However, retrieving information from the long-term memory often takes longer and is prone to errors [Anderson and Schooler, 2000 as cited by Arslan et al., 2017a].

In support for the serial processing bottleneck hypothesis, Arslan et al. [2017a] showed that the children’s scores in a complex working memory task were highly correlated with the accuracy of their answers in the false belief task: Children with high scores in the working memory task also performed better in the second-order belief task. Moreover, when children failed the second-order false belief task, they most often provided first-order answers, as opposed to zero-order answers. More generally, this suggests that failures at a certain ToM order are typically one order below the target order of (false belief) reasoning.

1.2 ToM and Formalism

Epistemic logic has been historically used to formalize ToM concepts because it allows one to model reasoning about *knowledge*, through statements of the form “I know that you know...”. Generally, models of epistemic logic encode two aspects related to an agent: i) *facts* it considers true about the state of the world and ii) its *knowledge* about other agents’ knowledge (and about the other agents’ knowledge about other agents’ knowledge, etc). Since knowledge is often not stationary, *dynamic epistemic logic* (DEL) [van Ditmarsch et al., 2007]

provides a way to model how knowledge changes in response to new information. This new information is modeled through a variant of DEL called public announcement logic (PAL) [Plaza, 1989, 2007]: If a public announcement φ occurs, everything that contradicts φ is not considered possible anymore. For example, if Amy is told that in a drawer there is either one banana or one apple, then she initially considers it possible that either fruit is in the drawer. However, if Ben later tells her that there are no bananas in the house, then Amy’s knowledge is accordingly updated: She now no longer considers it possible that a banana is in the drawer.

Two important assumptions of classic DEL are that: i) all agents are perfect reasoners, meaning that they always reason correctly given a set of rules and they know everything about every *known* aspect of the world and ii) everybody knows (a.k.a., it is common knowledge) that all agents are perfect reasoners. However, behavioural research has shown that there is a cognitive limit to human recursive reasoning [de Weerd et al., 2018, Devaine et al., 2014, Nagel, 1995]. While formal DEL research does not account for the upper-bound limit in ToM, behavioural research does not test its predictions on tasks that can be easily modeled in epistemic logic, namely epistemic puzzles [Top et al., 2023]. Therefore, further research is necessary to bridge the gap between DEL formalism and bounded cognition.

A first attempt to model upper-bounded ToM reasoning in DEL was conducted in Kaneko and Suzuki [2002] [as cited by Top et al., 2023], where the authors define the *epistemic depth* of a formula based on the nesting of its knowledge operators. Arthaud and Rinard [2023] built upon this framework in their purely theoretical study by defining several logics of public announcements where agents could understand a *limited* number of nested knowledge operators. Importantly, the authors assumed that *any* nested knowledge operator increases a formula’s epistemic depth.

Contrary to the Arthaud and Rinard [2023] approach, Cedegao et al. [2021] and Top et al. [2023] assumed that a formula’s epistemic depth is increased only by switching between knowledge operators for *different* agents - that is, there is a qualitative difference between reasoning about one’s own knowledge and reasoning about the knowledge of others. In essence, Cedegao et al. [2021] proposed

an upper-bounded simulated agent that has access to only a subset of the entire epistemic logic model associated with the logical puzzle. The size of the subset is proportional to its epistemic order: The higher the epistemic order, the more information the agent can access. A public announcement has the expected effect: Any information in the subset that contradicts the public announcement is simply removed. The agent is said to have solved the epistemic puzzle when only one option remains in the subset because all other options have been removed through public announcements. Additionally, Cedegao et al. introduced levels of stochasticity to the upper-bounded agent, in order to account for human participants’ guessing behaviour: For example, the subset is updated after a public announcement given a certain probability.

Top et al. [2023] built upon the Cedegao et al. [2021] approach by proposing a different way of reasoning about other agents’ knowledge. Top et al. claimed that the Cedegao et al. [2021] approach allows for an *infinite* number of switches of perspectives* between different agents, and that this contradicts theories of recursive ToM limitations. Instead, Top et al. explicitly modeled a perspective switch event from *agent*₁ to *agent*₂ in the following way: *agent*₁ associates with *agent*₂ an epistemic order smaller by one than its own; then *agent*₁ reasons for *agent*₂ using its own knowledge base and the lower epistemic order. Crucially, since an agent cannot allocate to another anything lower than zero-order ToM, the number of perspective switches is then *finite*.

The Cedegao et al. [2021] and Top et al. [2023] approaches were both validated on human data, on the same “Aces and Eights” dataset using different statistical methods: Maximum Likelihood Estimation [Cedegao et al., 2021] and Random-Effects Bayesian Model Selection [Top et al., 2023]. Several issues are worth mentioning with regards to these two approaches. Firstly, Top et al. [2023] have identified fundamental issues with the “Aces and Eights” experimental design: An agent could solve games that require higher-order epistemic reasoning by always claiming that they do not know the cards. Therefore, many participants who truth-

*A switch of perspective can be understood as “placing oneself in somebody’s else shoes” and reasoning from their perspective. Both agents are assumed to be using the same reasoning strategy, which is not always true in practice.

fully reported not knowing the answer could have been misclassified as being high-order ToM reasoners in Cedegao et al. [2021]. Moreover, in the game of “Aces and Eights” there is a considerably high chance of probabilistic guessing (i.e., 50% for guessing that the cards are not known) - this makes it difficult to differentiate those participants who are capable of using higher-order reasoning from participants that use other strategies, such as guessing. Secondly, one aspect of the Top et al. [2023] modeling approach seems rather counter-intuitive: Zero-order ToM agents assume that other agents do not consider anything to be possible and, therefore, that other agents vacuously consider everything to be possible, including contradictions. Thirdly, the game of “Aces and Eights” is only set within the “playing cards” context and, as discussed above, ToM abilities are task-domain dependent [e.g., Flobbe et al., 2008]; therefore, one must be cautious of drawing general conclusions about ToM strategies used by humans. Lastly, both studies only model truthful and public announcements and, therefore, do not account for the possibility that the participants may be lying or for private announcements.

Unfortunately, there are only a few behavioural experiments on classic epistemic puzzles [Cedegao et al., 2021, Jonker and Treur, 2003, Hayashi, 2002] and, for those that do exist, it is often the case that the reliability of the dataset is called into question. Therefore, the aim of the current study is to tackle the aforementioned issues of the Cedegao et al. [2021] and Top et al. [2023] approaches by proposing a novel experimental design, in which different orders of ToM reasoning in epistemic puzzles can be properly distinguished from other strategies. Additionally, as an alternative to Top et al. [2023], I propose a novel way of modeling statements that go beyond certain ToM limits: Instead of simply ignoring such statements (as done in Top et al., 2023), I propose a model that removes parts of the complex ToM statement until it can be processed.

1.3 Current Study

The present study aims to achieve two goals:

1. *Propose a novel experimental design and generate a dataset that will become 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 - please refer to Chapter 2 for a more detailed description of the experimental design. 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]. 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.

Overall, the aim of the experiment was to ensure that the potential 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 the eight puzzles requiring varying orders of ToM reasoning and set in different contexts in a randomized order, in order to account for potential learning effects.

2. *Investigate whether participants’ behaviour (and, specifically, mistakes) can be explained by a model that removes parts of complex ToM statements.*

Suppose that a model can process only up to x -order ToM statements. Now suppose that the same model encounters an $(x + 2)$ -order statement. In Top et al. [2023], the latter statement would simply be ignored since it cannot be processed by the model. I propose an alternative strategy: Knowledge operators are sequentially removed (along with their preceding negations, if applicable), either from the left side or from the right side of the statement,

such that the initial statement becomes an x -order statement and, therefore can now be processed by the model - please refer to Chapter 4 for a more detailed description. This design choice is supported by the findings in Arslan et al. [2017b] and Arslan et al. [2017a], where it is suggested that failures at a certain order of ToM typically lead to answers at a lower ToM order, as opposed to a different strategy.

This model was compared against a perfect reasoning model and a random model and the best fit was assessed using group-level random-effects Bayesian model comparison (RFX-BMS), as proposed by Stephan et al. [2009] - please refer to Chapter 5 for a more detailed description. This would reveal the existence of a second-order ToM cognitive limit, as suggested by the literature: If a model that cuts statements down to second-order is found to explain most of the data, this would suggest that participants (perhaps unconsciously) use a similar strategy.

In this chapter, I present and justify the experimental design. In Section 2.1, I discuss the demographic data of the participants involved in the experiment. In Section 2.2, I present the solutions of the four unique “Cheryl’s Birthday” puzzles used in the study and explain how to generate sixty variations based on these four original puzzles. Lastly, in Section 2.3, I discuss in detail the setup of the experiment.

The code can be found at <https://github.com/AndreeaMinculescu/Cheryl-Puzzle>. For a detailed explanation of the experimental procedure, see Section 2.3.

2.1 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. Initially, one additional Master’s student participated in the experiment, presumably without carefully reading the participation requirements mentioned in the advertisement. This participant was later excluded from any further analyses.

The participants had reported to have 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 may not have solved the “Cheryl’s Birthday” puzzle before, out of concern that they might look it up online. In spite of this, most participants reported not having heard about the “Cheryl’s Birthday” puzzle prior to the exper-

iment (45 participants) or any similar puzzle (43 participants). On a scale from 1 to 10, 10 being equivalent to “I feel very happy”, participants reported a mean mood score of 6.94.

2.2 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]. All puzzle texts and their respective solutions are discussed below. In Section 2.2.5, it is discussed how variations of the four unique puzzle types shown during the experiment were obtained.

For the discussion of the solutions below, consider the following notations: Let Albert be A, Bernard be B and Cheryl be C. Additionally, let “days” refer to the numbers on C’s list (14-18), let “months” refer to the calendaristic months on C’s list (May-September) and let “birthday” refer to a ⟨month, day⟩ combination on C’s list (e.g., May, 15). Let a *unique* element of a set be defined as an element that occurs no more than once in that set - for example, in the set $\{a, b, a, c, b, a\}$, c is a unique element.

2.2.1 “Cheryl’s Birthday” First-Order ToM

The text of the puzzle is presented below:

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: 15,16
- June: 17,18
- July: 14,16
- August: 14,17
- September: 16,18

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:

Bernard: “I know when Cheryl’s birthday is.”

When is Cheryl’s birthday?

The solution to the puzzle above can be formulated as follows:

B claims that he knows the birthday. Since B knows only the day and nothing else, it must be the case that the birthday is associated with a *unique* day (i.e., a day that occurs only once in the list of options given by Cheryl). Consider the alternative: if C’s birthday was on a day that is not unique on the list (e.g., 18), then B would not be able to differentiate between multiple months associated with that date (i.e., June and September) without more information. Since 15 is the only unique day, the solution to the puzzle must be May, 15.

This puzzle requires first-order ToM because the participant must perform one perspective switch to find the solution: reasoning from B’s perspective.

2.2.2 “Cheryl’s Birthday” Second-Order ToM

The text of the puzzle is presented below:

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: 17,18
- June: 14
- July: 16,18
- August: 15,16,17
- September: 14,15

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?

The solution to the puzzle above can be formulated as follows:

A claims that he does not know C’s birthday and, therefore, it must mean that C’s birthday is not in a month associated with only one day. Consider the following line of reasoning: if C’s birthday was in a month with multiple days associated (e.g., September), then A would not be able to differentiate between the multiple days associated with that month (i.e., 14 and 15) without more information. Therefore, C’s birthday cannot be on June, 14 (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 and have as leftover options:

- May: 17, 18;
- July: 16, 18;
- August: 15, 16, 17;
- September: 14, 15.

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 has already been eliminated as an option).

This puzzle requires second-order ToM because the participant must reason from B's perspective about A's perspective: without A's statement, B would not have been able to find the solution because initially there are no unique days.

2.2.3 “Cheryl’s Birthday” Third-Order ToM

The text of the puzzle is presented below:

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: 15,18
- June: 15,17
- July: 14,16
- August: 14,16
- September: 15,16

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 know that you don't know when Cheryl's birthday is.”

Bernard: “I didn't know at first, but now I know.”

When is Cheryl's birthday?

The solution to the puzzle above can be formulated as follows:

A claims that he knows that B does not know. For A to know that, it means that the month he was told is not associated with a unique day. Consider the alternative: if C's birthday was in a month

associated with a unique day (e.g., May, where 18 is a unique day), then A would not be able to differentiate between the two days (15 and 18) and, by extension, would not be able to confidently claim that B does not know the birthday (B could potentially have been told 18 and, in that case, B would have known the birthday) without further information. Since both 17 and 18 are unique days, C's birthday cannot be in May or June.

Next, B claims that A's statement helped him find C's birthday. As a perfect reasoner, B must have completed the steps described above correctly and have as leftover options:

- July: 14, 16;
- August: 14, 16;
- September: 15, 16.

For B to now know the solution, C's birthday must be associated with a unique day. Therefore, the day must be 15 and C's birthday must be September, 15.

This puzzle requires third-order ToM because the participant must reason from B's perspective about A's reasoning about B's perspective: without A's statement, B would not have been able to differentiate between May, June, and September, which all have 15 as an option.

2.2.4 “Cheryl’s Birthday” Fourth-Order ToM

The text of the puzzle is presented below:

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: 15,18
- June: 14,15
- July: 17,18
- August: 16,17
- September: 16,17

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:

Benard: "I know that you know that I don't know when Cheryl's birthday is."

Albert: "I didn't know at first but now I know."

When is Cheryl's birthday?

The solution to the puzzle above can be formulated as follows:

B claims that he knows that A knows that B does not know. Similarly to Section 2.2.3, A knows that B does not know if A was told a month that is not associated with a unique day. Since 14 is unique, it means that C's birthday is not in June. After all, if C's birthday had been in June, A would not have been able to exclude June, 14 as a possibility, in which case, B would have known the birthday. Since B knows all of this, it means that B was not told one of the days in June. Consider the alternative: if the day was 14, then B would know the birthday from the beginning (it is a unique day) and that contradicts the first statement in the dialogue. If the day was 15, then B would consider it possible for June, 15 to be the answer. If June were the correct month, then A would consider it possible for June, 14 to be C's birthday and, by extension, A would consider it possible for B to know the answer from the beginning. This, again, contradicts the first statement in the dialogue. Therefore, May, 15 is also not possible.

Next, A claims that B's statement helped him find C's birthday. As a perfect reasoner, A must have completed the steps described above correctly and have as leftover options:

May: 18;
July: 17, 18;
August: 16, 17;
September: 16, 17.

For A to now know the solution, C's birthday must be in a month associated with only one day. Therefore, the month must be May and C's birthday must be May, 18.

This puzzle requires fourth-order ToM because the participant must reason from A's perspective

about B's reasoning about A's reasoning about B's perspective: without B's statement, A would not have been able to differentiate between May, 15 and May, 18.

2.2.5 All Puzzle Variations

A puzzle was designed following a 4x4x4 design:

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 Albert and Bernard. Take the following dialogue example:

Albert: "I don't know when Cheryl's birthday is."

Bernard: "I didn't know at first, but now I know."

This is an example of a second-order ToM puzzle. The participant would need to perform two perspective switches to understand the dialogue: Bernard can use Albert's perspective to find the answer. Interestingly, it has been shown that *ignorance* (i.e., statements of the form "I don't know") can also provide useful information about the state of a game (see [van der Hoek and Verbrugge, 2002] for examples in Game Theory). This is the case here as well: the fact that Albert does not initially know the answer helps Bernard solve the puzzle.

The puzzles range from first to fourth-order ToM. Since the focus of the study was to investigate the manner in which participants process perspective switches while reasoning from other people's perspective, zero-order ToM, which would require zero perspective switches, was excluded due to its mostly trivial nature.

2. *Scenario:* The puzzle examples discussed so far have been showcasing one so-called "scenario": finding Cheryl's **birthday**. It is important to keep ToM reasoning as the main focus of the study and to isolate the ToM performance from other potential confounding factors: Their familiarity with the intuitive relation between, on one hand, days and months

and, on the other hand, birthdays, could enhance the participants' performance. Additionally, past research has shown that ToM reasoning is task-dependent [Flobbe et al., 2008] and, therefore, alternating scenarios is a first step towards accounting for this constraint and supporting more general conclusions about ToM underlying mechanisms.

To this end, three other scenarios were introduced, where only the target properties (i.e., month and day) were changed. In the **drink** scenario, Albert and Bernard are challenged to find out how Cheryl likes to have her coffee: she tells Albert the size of the coffee (e.g., large, small) and Bernard the temperature (e.g., iced, lukewarm). In the **toy** scenario, Albert and Bernard are challenged to find Cheryl's favourite childhood toy in her room: she tells Albert the location of the toy (e.g., on the armchair, on the windowsill) and Bernard the type of toy (e.g., doll, clown). Finally, in the **hair** scenario, Albert and Bernard are challenged to locate Cheryl's friend, Diane, in a busy train station and Cheryl describes her hair: she tells Albert the hair style (e.g., curly, straight) and Bernard 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 one hand and the corresponding scenario's target properties on the other hand. Specifically:

⟨May, June, July, August, September⟩
and
⟨14, 15, 16, 17, 18⟩

are replaced with, in this exact order:

⟨Extra small, Small, Regular, Large, Extra large⟩
and
⟨hot, lukewarm, room temperature, cold, iced⟩,

for the drink scenario;

⟨On the table, On the bed, On the floor, On the armchair, On the windowsill⟩
and
⟨doll, bunny, clown, cat, train⟩,

for the toy scenario; and

⟨Curly, Spiky, Straight, Pixie, With bangs⟩
and
⟨green, blue, purple, pink, orange⟩,

for the hair scenario.

For example, *⟨June, 17⟩* becomes *⟨Small, cold⟩* in the drink scenario, *⟨On the bed, cat⟩* in the toy scenario and *⟨Spiky, pink⟩* in the hair scenario.

For examples of the exact rephrasing of the puzzle texts for the other three scenarios, please refer to Appendix A.

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⟩
becomes
⟨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 (see Section 2.3.2).

2.3 Procedure

Participants were instructed that the entire experimental procedure would last 45 to 60 minutes. Upon arrival, participants were seated at a desk in front of a computer in a quiet lab room at the University of Groningen. The code was written in Python 3.10 and run on HP Z2 Mini G3 Workstations.

The experiment was run in six different rooms, but in the same quiet environment. In one case,

the code was run on a different machine. namely an Alienware m15 R7 AMD*. Otherwise, the experimental procedure was followed faithfully.

Four objects had been placed on the desk prior to the arrival of the participants: an informed consent form, a pile of eight A4 papers numbered from one to eight respectively (referred to hereon as “puzzle notes”), an empty A4 envelope and a pen. After preliminary introductions, each participant was asked to read and sign the informed consent form, explaining the purpose of the experiment and data processing regulations. Afterward, the participants were instructed to start the experiment on the computer. Figure 2.3.1 shows the workflow of the experiment. The participants were under continuous supervision throughout the entirety of the experimental procedure and were encouraged to ask questions for clarification of instructions at any point.

The participants were first asked to enter their allocated ID, which was written on the envelope before their arrival. This was done to ensure reproducibility: Eight puzzles were randomly allocated per participant prior to the start of the trial and each allocation was stored in a database, with a unique ID.

Next, they were shown a welcome message and were given additional instructions pertaining to the materials on the table and the payment method. Participants were strongly encouraged (but not forced) to use the materials on the table, consisting of the puzzle notes, the envelope, and the pen, for note-taking while solving the puzzles. The puzzle notes consisted of eight blank A4 pieces of paper which were numbered on both sides from one to eight. The participants were instructed to use one piece of paper per puzzle (hence, eight double-sided A4 pieces of paper for eight puzzles) and to place in the envelope all written pieces of paper before moving to the next puzzle. This setup achieved two purposes: i) it reduced the working memory load for intermediate steps, which has been shown to influence performance [Arslan et al., 2017a], and ii) it discouraged participants from looking at previous answers (which they were also explicitly warned against doing), such that they would not find patterns amongst the puzzles generated based on the four original Cheryl’s birthday puzzles discussed in

Section 2.2.

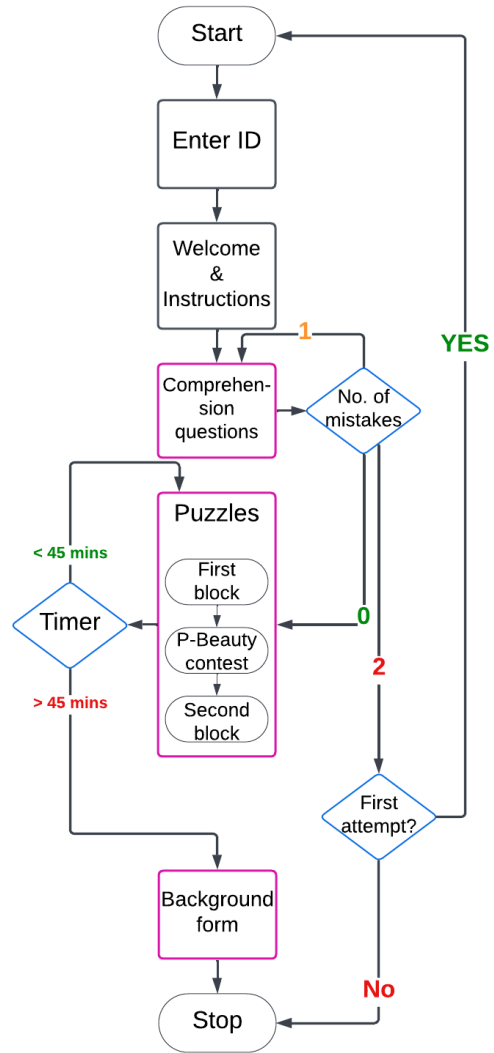


Figure 2.3.1: Workflow of the experiment. The main stages are marked with pink squares. Decision points are marked with blue diamonds.

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€). The bonus monetary reward was allocated according to the following rule: The bonus was applied if the answer to one randomly selected puzzle from the puzzles solved by the participant within the time limit was correct. This was in order to incen-

*However, note that the code complexity is low and, in principle, the code can be run on any functional device.

tivize participants to solve correctly as many puzzles as possible: the more puzzles solved correctly, the higher the chance of selecting a correct puzzle.

The three main stages of the experiment (marked with pink squares in Figure 2.3.1) are discussed in the remainder of this section.

2.3.1 Comprehension Questions

The participants were shown the following text:

You will be asked to solve a series of unrelated puzzles revolving around three friends: Albert, Bernard and Cheryl. Albert, Bernard and Cheryl are all in the same room. Albert and Bernard just met Cheryl and they want to get to know her better - for the sake of this example, suppose they want to know when her birthday is. Cheryl has a very playful personality and hates giving answers outright so, instead, she says:

“I will not tell you but I will give you some hints.” She writes down a list of four dates and shows the list to Bernard and Albert:

January 11 - January 12

February 11 - February 12

“My birthday is one of these dates.” Cheryl says.

Then Cheryl says that she will whisper in Albert’s ear the month of her birthday, and nothing else. To Bernard, she will whisper the day of her birthday, and nothing else. She does as she said. Albert sees her whisper in Bernard’s ear but cannot hear what was said (the same is the case for Bernard). However, it is well known by anybody that Albert, Bernard and Cheryl have perfect reasoning abilities and never lie or purposefully deceive each other, so both boys are convinced Cheryl did as she said.

“Can you figure out my birthday”? Cheryl asks them. Now, the following dialogue ensues.

Albert: “Bernard, I don’t know when Cheryl’s birthday is.”

Bernard: “I don’t know when Cheryl’s birthday is either.”

This practice text is more detailed than the standard puzzle texts (see Section 2.2). This is meant to introduce concepts of perfect logicians (e.g., “no reasoning mistakes”) and common knowledge (“it

is well known by anybody...”) in an intuitive way. This, together with the detailed description of the setting, serves to suggest that there are no “tricks” hidden within the puzzles and that they are meant to be solvable through logical inferences. Additionally, note that the question “When is Cheryl’s birthday?” is missing. This is purposefully done in order to avoid priming the participants for ways to solve the puzzle before the start of the actual experiment.

Instead, the participants’ attention was redirected toward text comprehension, through targeted questions. Participants were tested on their *reading comprehension* (“What does Albert know about Cheryl’s birthday right before his dialogue with Bernard?”), “Can Cheryl’s birthday be on the 15th of May?”), *understanding of epistemic concepts* (perfect logicians: “Can Bernard make false claims?”; common knowledge: “Does Albert know that Cheryl only speaks the truth?”) and *dialogue comprehension* (“Who knows the birthday after the whispering and before the dialogue between Albert and Bernard?”, “Who knows the birthday after the dialogue between Albert and Bernard?”). Participants were allowed at most one mistake before the experiment was prematurely ended. In that case, the experimenter would explain the instruction text again and answer the participant’s questions. Then, the participant would be allowed one last attempt before being excluded from the experiment. All participants passed the comprehension stage within the two attempts.

2.3.2 Cheryl’s Puzzles

Next, the participants were asked to solve eight puzzles ranging in difficulty. As explained before, they were encouraged to use the puzzle notes to write down intermediate steps. The puzzles were presented in two blocks of four puzzles each.

In total, sixty-four puzzles had been generated (see Section 2.2 for a detailed description). For all puzzles, there was always a unique solution (i.e., one of the options on Cheryl’s list) and the dialogue between Albert and Bernard contained no more and no less than the information necessary to reach the solution. It should be clear that theory of mind is the only strategy appropriate for solving the puzzles, because only the dialogue between Albert and Bernard, which is epistemic in nature,

conveys any relevant clues and the chance of simply guessing the correct answer is low (7.7% or 1 in 13 possible answers). Therefore, the puzzles are appropriate for investigating recursive theory of mind limitations.

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 - the toy scenario could, in principle, be paired with any of the four ToM orders, so 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 (see Section 2.2.5 for more information).

In each block, each scenario and each ToM order were shown exactly once. Consider the following $\langle \text{scenario} \times \text{ToM order} \rangle$ series examples:

- $\langle \text{toy} \times 4, \text{hair} \times 1, \text{birthday} \times 2, \text{hair} \times 3 \rangle$ - this is not a valid configuration for a block, because the hair scenario occurs more than once.
- $\langle \text{toy} \times 4, \text{hair} \times 1, \text{birthday} \times 2, \text{drink} \times 1 \rangle$ - this is not a valid configuration for a block, because first-order ToM occurs more than once.
- $\langle \text{toy} \times 4, \text{hair} \times 1, \text{birthday} \times 2, \text{drink} \times 3 \rangle$ - this is a valid configuration for a block, because each scenario and ToM order occurs exactly once. Note that the order in which the scenarios and the ToM order occur is randomized.

Additionally, if one $\langle \text{scenario} \times \text{ToM order} \rangle$ configuration occurs in the first block, then it is not allowed to also occur in the second block. This was done in order to ensure that no two puzzles looked too similar to each other: If two puzzles only differ by configuration, there is a higher risk that the participants might realize that there is a connection between these puzzles and, as a result, they might develop different strategies for solving the puzzles. Take the following example:

Suppose that the following configuration occurs in the first block:

$\langle \text{toy} \times 4, \text{hair} \times 1, \text{birthday} \times 2, \text{drink} \times 3 \rangle$

Consider the following configurations for the second block:

- $\langle \text{birthday} \times 2, \text{toy} \times 3, \text{drink} \times 1, \text{hair} \times 4 \rangle$ - this is not a valid configuration for the second block because $\langle \text{birthday} \times 2 \rangle$ also occurs in the first block.
- $\langle \text{birthday} \times 4, \text{toy} \times 3, \text{drink} \times 1, \text{hair} \times 2 \rangle$ - this is a valid configuration for the second block because no $\langle \text{scenario} \times \text{ToM order} \rangle$ configuration occurs in both blocks.

In the end, due to a bug in the code that was not noticed in time, in the first block, each scenario was instead 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. The second block was processed correctly, as explained above. However, due to the constraint that the same $\langle \text{scenario} \times \text{ToM order} \rangle$ may not occur in both blocks, each scenario could only be associated with three ToM orders in the second block (for example, the toy scenario could only be associated with second, third and fourth-order ToM because, in the first block, it was always associated with first-order ToM). Nonetheless, it is important to note that each participant was shown each ToM order and scenario exactly twice in total and exactly once in each block, as originally intended.

The participants were instructed to read the puzzle text carefully and to select the solution from a list of thirteen options: the ten options on Cheryl's list, "I don't know", "No solution" and "Multiple solutions". "I don't know" was set as the default option for the drop-down menu and was recorded as the answer when no option in the drop-down menu had been selected.

The participants had 45 minutes to solve as many of the eight puzzles as possible. If the time limit had passed before they finished solving all eight puzzles, they were automatically redirected to the Background Form (see Section 2.3.3), and the remainder of the unsolved puzzles were skipped. Similarly, if they finished solving all eight puzzles before the time limit had passed, they were also redirected to the Background Form.

A "p-Beauty Contest" [Nagel, 1995] was presented after the first block and before the second block. In a p-Beauty Contest, participants are

asked to pick a number between 1 and 100 that they think might be closest to p (here, $p = \frac{2}{3}$) times the average of all participants' choices. The p-Beauty Contest was meant to prevent boredom and to disrupt any learning effects that may have begun forming after solving the first four puzzles: Since the aim was to measure inherent ToM cognitive abilities, it is important to account for temporary performance enhancements due to practice. Participants were told that the "winner" of the contest would receive a chocolate bar as a prize.

The data collected at this stage was later used to answer the main research questions of the study. The following information was recorded:

- The ID associated with a puzzle series in the database.
- A unique ID per participant, generated using the `uuid` Python package. This is different from the ID associated with the puzzle series. Its purpose is to act as a fail-safe in case multiple participants are given the same puzzle series due to human error: There would always be a unique identifier for each participant. In the end, the fail-safe proved to be unnecessary as each participant received a unique puzzle series ID.
- The index of the current puzzle and the associated answer chosen from the drop-down menu. If no answer was selected, then "I don't know" was the default recorded answer.
- The time spent on a puzzle, from the moment the puzzle was first shown on the screen until the press of the "Submit" button. Note that this includes the time to read the puzzle text and, with the current experimental setup, it would be impossible to separate the reading time from the puzzle-solving time.
- The p-Beauty Contest value.

2.3.3 Background Form

Finally, the participants were asked to report their contact information (name, email address), demographic data (age, gender), educational background (study program, formal training in logic), and overall experience with the experimental procedure (see Chapter 3 for an analysis of the answers). Note

that the contact information was stored separately from everything else and was only used to contact the participants about the monetary compensation. The answers to the puzzles were processed under anonymity regulations.

All answers except for the contact information were stored together with the puzzle answers, as described in Section 2.3.2. The contact information was stored in a separate dataframe, together with: i) the unique participant ID (see Section 2.3.2), ii) the monetary reward sum and iii) the index of the puzzle randomly selected to determine whether the bonus monetary sum should be awarded.

In this chapter, I present some interesting effects found in the “Cheryl’s Puzzle” dataset. In Section 3.1, I present the findings for the main effects: whether time to reach an answer and accuracy vary across orders of theory of mind (ToM) and scenarios. In Section 3.2, I discuss external factors associated with the sample of participants that might have an effect on accuracy. In Section 3.3, I present the results of the p-Beauty contest. In Section 3.4, I summarize all results.

The code for the analysis can be found at <https://github.com/AndreeaMinculescu/Cheryl-Puzzle> and was written in Python 3.10 and R 4.2.2.

3.1 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.

In the following sub-sections, I discuss the effects that scenario and order of ToM have on the time needed to solve a puzzle and accuracy, respectively.

3.1.1 Time to Reach Answers

As a reminder, 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?*

The initial assumption was that the higher the order of ToM, the more difficult solving a puzzle should be and the more processing operations it should require. As a result, higher orders of ToM should take longer to reach an answer.

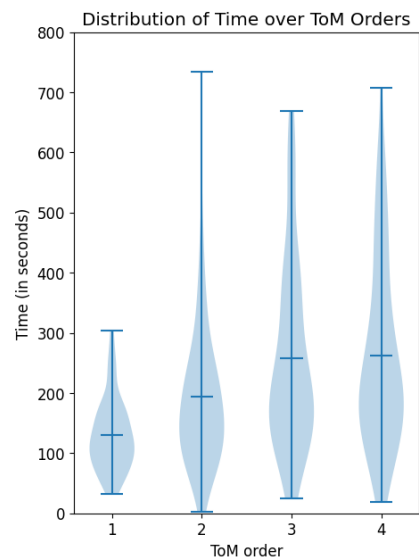


Figure 3.1.1: Violin plots of the time to solve a puzzle across the four ToM orders for the forty-two participants who completed all eight trials. The upper and lower horizontal bars show the two extremes and the middle bar shows the mean. The x-axis shows the ToM order and the y-axis shows the time in seconds.

Figure 3.1.1 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, SD=60.05), followed by second-order (M=194.01, SD=121.53), and lastly, third-order (M=258.02, SD=154.13) and fourth-order (M=261.69, SD=153.09) puzzles. There does not seem to be a noticeable difference between the distribution associated with third-order and fourth-order puzzles. Additionally, the distribution of the second-order puzzles contains one distinctively low value (namely, 3.01 seconds), which could suggest the presence of an outlier. However, upon closer inspection of the data point, there was no suggestion of a time-recording malfunction and there seems to be no research to establish the minimum time required to solve a “Cheryl’s Birthday” puzzle. Therefore, I decided against removing the data point on the grounds of it being a potential outlier.

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, were significantly different from each other in terms of solving time (see Table 3.1.1).

Comparison	Z	p-value
First vs Second order	-3.63	<i>0.001*</i>
First vs Third order	-6.41	<i><0.001*</i>
Second vs Third order	-2.78	<i>0.01*</i>
First vs Fourth order	-6.60	<i><0.001*</i>
Second vs Fourth order	-2.96	<i>0.009*</i>
Third vs Fourth order	-0.18	<i>0.8</i>

Table 3.1.1: Results of the Dunn test for the time to solve a puzzle across ToM orders. Every two ToM orders were compared against each other and the results were corrected for multiple comparisons using Holm’s method [Holm, 1979]. Starred p-values are below the significance threshold ($\alpha = 0.05$).

Overall, the visualization of the data and the associated statistical analysis support the hypothesis: lower orders of ToM are associated with decreased solving time and this result is statistically significant. Interestingly, there is virtually no difference between the solving time associated with third-order and fourth-order puzzles.

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

The initial assumption was that any potential difference in performance is only due to the order of ToM associated with a puzzle and, therefore, varying other design aspects (scenario, in this case) will not have a significant influence on overall performance.

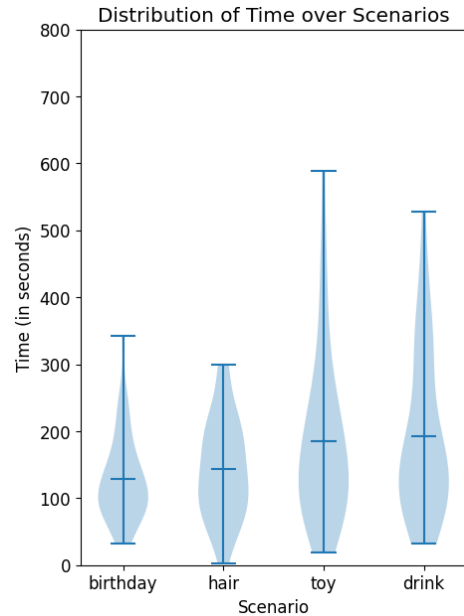


Figure 3.1.2: Violin plots of the time to solve a puzzle across the four scenarios in the second block for the forty-two participants who completed all eight trials. The upper and lower horizontal bars show the two extremes and the middle bar shows the mean of each distribution. The x-axis shows the scenario and the y-axis shows the time in seconds.

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.

Figure 3.1.2 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: $M=128.76$, $SD=63.28$; hair: $M=143.08$, $SD=72.89$; toy: $M=184.65$, $SD=123.40$; drink: $M=192.85$, $SD=120.65$. Additionally, the distribution of the hair puzzles contains a distinctively low value, which corresponds to the same data point flagged as an outlier in Figure 3.1.1.

A Shapiro-Wilk test on the within-group variability revealed that the (log-transformed) time was not normally distributed ($W=0.94$, $p < 0.001$). Therefore, a Kruskal-Wallis test was conducted on the solving time, which did not reveal statistical significance ($\chi^2(3) = 7.38$, $p = 0.060$).

Overall, the visualization of the data and the associated statistical analysis support the hypothesis: the scenario type does not significantly influence the time interval required to solve a puzzle. The lack of a significant effect is evidence that the bug in the code (i.e., the unbalanced design of the two blocks) had no adverse effect on the results of the study.

3.1.2 Accuracy of Answers

Let us define accuracy, as a percentage, as follows:

$$\text{accuracy} = \frac{\text{the number of correct answers}}{\text{the total number of answers}} \times 100, \quad (3.1.1)$$

where the total number of answers is eight since only participants who finished all eight trials are considered.

Figure 3.1.3 shows the frequency distribution of accuracy over the forty-two participants who finished all eight trials. Twenty-one participants answered half or more puzzles correctly (or, equivalently, had an accuracy of 50% or higher). Importantly, five participants failed to answer any puzzle correctly, which resulted in an accuracy lower than the probability of simply guessing the correct answer (the gray dotted line in the figure). The chance of guessing is computed as $\frac{1}{13} * 100 \approx 7.7\%$, where

13 is the total number of possible answers for any puzzle. This suggests that five participants were notably unskilled with respect to solving any puzzle.

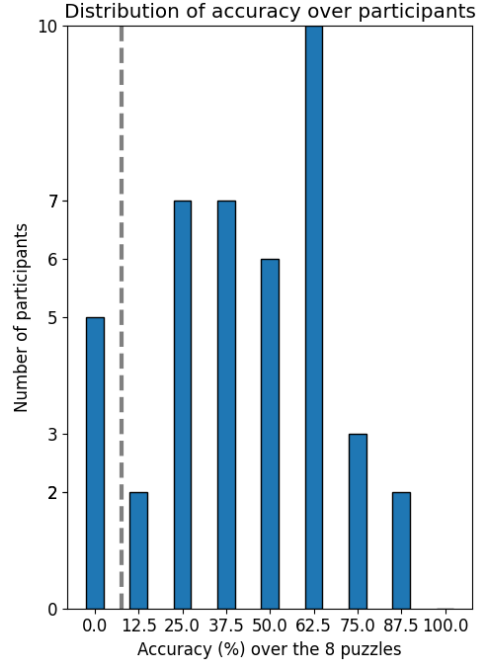


Figure 3.1.3: Distribution of accuracy over the forty-two participants who completed all eight trials. The y-axis shows the accuracy computed according to Equation 3.1.1. The x-axis shows the number of participants with the same accuracy. The dashed line shows the chance level (namely $\frac{1}{13} * 100 \approx 7.7$).

In order to conduct the following statistical analyses, I generated a binary variable `is.correct` that takes the value 1 if the answer to a puzzle given by a participant is the same as the correct answer and 0 otherwise. Two questions of interest arise:

1. *Does accuracy differ significantly across orders of ToM?*

The initial assumption was that the higher the order of ToM, the more difficult it is to solve a puzzle and, as a result, the more likely a participant is to make an error.

As shown in Figure 3.1.4, 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 forty-two 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 notably unskilled with respect to solving these puzzles. Again, the chance of guessing is computed as $\frac{1}{13} * 100 \approx 7.7\%$, where 13 is the total number of possible answers for any puzzle.

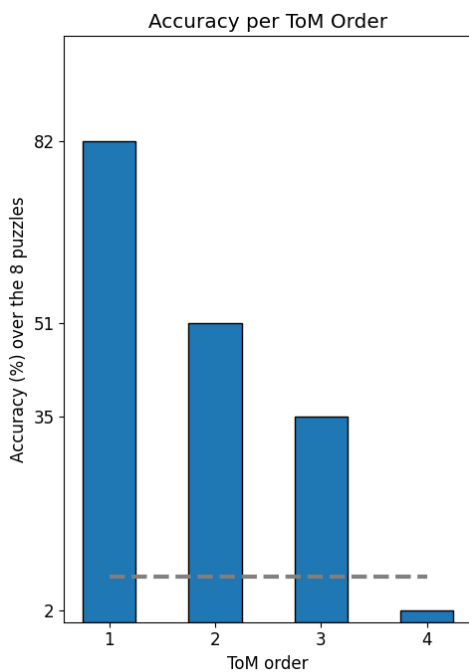


Figure 3.1.4: Bar chart of the accuracy associated with each ToM order for the forty-two participants who completed all eight trials. The x-axis shows the ToM order and the y-axis shows the accuracy, as percentage. The dashed line shows the chance level (namely $\frac{1}{13} * 100 \approx 7.7$).

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 initial assumption was that any potential difference in performance is only due to the order of ToM associated with a puzzle and, there-

fore, varying other design aspects (scenario, in this case) will not have a significant influence on overall performance.

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 3.1.5, participants solved the birthday scenario with an accuracy of 40.4%, the hair scenario with an accuracy of 31%, the drink scenario with an accuracy of 39.3% and the toy scenario with an accuracy of 59.5%. For all scenarios, the accuracy was higher than the probability of simply guessing the correct answer (the gray line in the figure).

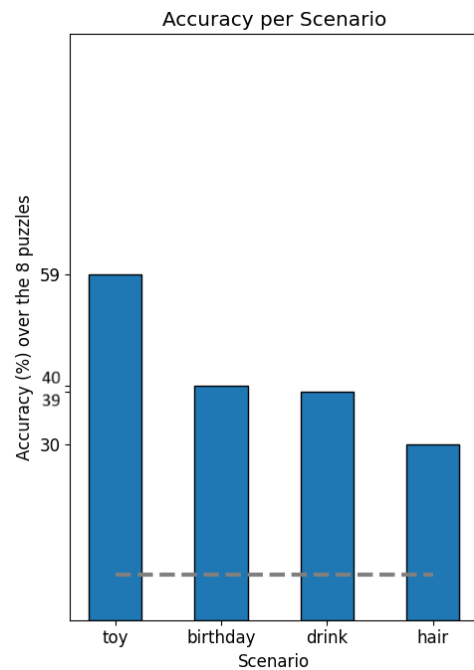


Figure 3.1.5: Bar chart of the accuracy associated with each scenario in the second block for the forty-two participants who completed all eight trials. The x-axis shows the scenario and the y-axis shows the accuracy, as percentage. The dashed line shows the chance level (namely $\frac{1}{13} * 100 \approx 7.7$).

As expected, a Chi-square test revealed that accuracy was not significantly different be-

tween scenarios ($\chi^2(3) = 6.34, p = 0.096$). Since 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 Chapter 2 for more details). Therefore, in the second block, the toy scenario was never associated with fourth-order ToM, while the hair scenario was never associated with first-order ToM. As shown before, participants solved first-order ToM puzzles correctly significantly more often than fourth-order ToM puzzles, hence a possible explanation for the reason why the accuracy of the toy scenario is higher than the accuracy of the hair scenario.

Nonetheless, the lack of a significant effect is evidence that the bug in the code (i.e., the unbalanced design of the two blocks) had no adverse effect on the results of the study.

3.2 Background Form

Another point of interest is to understand whether external factors, unrelated to the design of the experiment, may influence the accuracy of the answers. This may indicate whether the sample of participants used for this experiment may have introduced bias, for example, due to similar background knowledge, educational background and socio-economical status. The following statistical tests were conducted on all forty-nine participants and the results are summarized in Table 3.2.1.

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 ($\chi^2(1)=3.69, p=0.027$).

Participants were also 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 answer significantly better than participants who reported a score lower than or equal to five ($\chi^2(1)=1.13, p=0.123$).

In the same background form, participants were 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 answered correctly significantly less often than the participants who reported a score lower than or equal to five ($\chi^2(1)=12.128, p < 0.001$).

Lastly, 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 ($\chi^2(1)=0.02, p=0.439$).

Variable	χ^2	Df	p-value
Already knew puzzle before experiment?	3.69	1	<i>0.027*</i>
Perceived difficulty of instructions	2.35	1	<i>0.062</i>
Perceived difficulty of puzzles	12.12	1	<i><0.001*</i>
Perceived enjoyment of puzzle solving	0.02	1	<i>0.439</i>

Table 3.2.1: Results of the Chi-square tests performed on the accuracy of puzzle answers ($N = 381$) over various variables of interest collected through the background forms. Starred p-values are below the significance threshold ($\alpha = 0.05$).

Overall, accuracy is significantly higher if participants already knew “Cheryl’s Birthday” or a similar puzzle, or if participants perceived the puzzles as easy. Surprisingly, it seems that the perceived difficulty of the instructions and the perceived enjoyment of solving the puzzles do not have a significant effect on accuracy - possibly because the monetary compensation acted as a motivator. Note that for the proportion tests involving a scale rat-

ing, the threshold between the two categories was always set to five, for consistency.

3.3 P-Beauty Contest

Figure 3.3.1 shows the distribution of the p-Beauty answers across all forty-nine participants. The data has been binned in order to improve readability. The figure should be understood as follows: Eight participants selected a value larger than 36 and smaller or equal to 42.

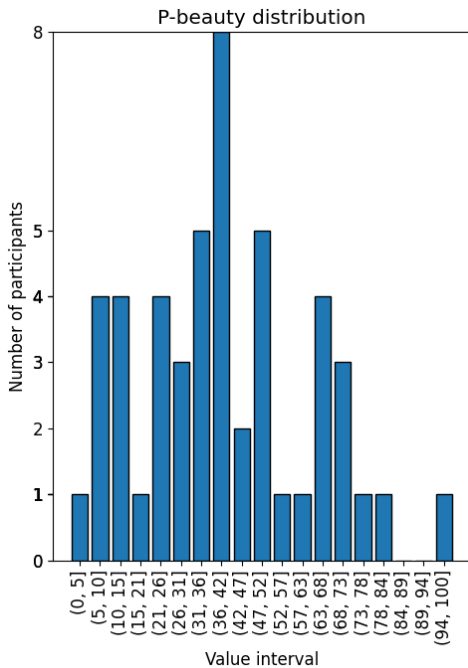


Figure 3.3.1: The distribution of the binned p-Beauty values chosen by all forty-nine participants. The y-axis shows the binned p-Beauty value. The x-axis shows the number of participants that chose a value in that specific interval.

The p-Beauty contest qualitatively shows the *level* of iterated reasoning exhibited by the participants. It is important to note here the subtle difference between models of iterated reasoning and dynamic models of theory of mind: In non-repeated single-shot games (like the p-Beauty contest) agents typically assume that all other agents use exactly one step of iterated reasoning less than themselves, while in repeated game settings (like

the Cheryl’s Birthday puzzles), agents typically adjust their order of recursive reasoning in response to the behaviour of others and consider it possible that other agents use *any* number of iterated reasoning steps up to the number of steps they use themselves [de Weerd et al., 2017]. Therefore, the average number of iterated steps found in the p-Beauty contest results is no more than an approximation of the ToM order that participants may have used while solving the epistemic puzzles proposed in the current study. In the remainder of the section, I will use “level” to describe iterated reasoning, as opposed to “order” for ToM reasoning.

As a reminder, in the p-Beauty contest, participants were instructed to select a number between 1 and 100 that they thought might be closest to p (here, $p = \frac{2}{3}$) times the average of all participants’ choices. A zero-level participant would not consider the choices of other participants and would therefore most likely select a random number. A level-one participant would believe that all other participants use level-zero reasoning, would consider fifty to be the average of the other participants’ choices and would therefore most likely select a number close to $\frac{2}{3} \times 50 \approx 33$. Similarly, a level-two participant would select a number close to $\frac{2}{3} \times (\frac{2}{3} \times 50) \approx 22$ etc. More generally, a participant who can process up to level- n ToM statements would select a number close to $\frac{2}{3}^n \times 50$.

In this case, the mean and the median of the distribution were both 40, which suggests that most participants most likely used level-zero and level-one iterated reasoning.

3.4 Conclusion

In this chapter, I discussed some interesting effects present in the “Cheryl’s Puzzles” dataset. Firstly, I presented the effects of ToM orders and scenarios on solving time and accuracy, respectively. Importantly, the analyses were conducted on a subset of the data pertaining to those participants who finished all eight trials within the allocated time. Additionally, for the analyses pertaining to scenarios, the dataset was further restricted to only the second block. For future research, it could be interesting to further restrict the dataset to only correct answers for the analysis of solving time effects - this was not possible here due to lack of data (for

fourth-order ToM puzzles only two answers were correct).

The results can be summarized as follows: lower orders of ToM are associated with decreased solving time and higher accuracy (and vice-versa for higher orders of ToM), while no effect of scenario type was found. This latter finding is a strong indication that, for further analysis, it is possible to aggregate over scenarios with (almost) no loss of information and it is possible to use both blocks, despite the unbalanced distribution between these two blocks.

Secondly, I discussed whether external factors specific to the sample of participants may have an effect on accuracy. A series of proportion tests suggest that participants answer correctly significantly more often if they had already heard of the “Cheryl’s Birthday” puzzle (or a similar puzzle) prior to the experiment and if they overall perceived the puzzles as easy. Contrary to expectations, accuracy was not improved if the puzzle instructions were perceived as easy to understand or if the puzzles were perceived as enjoyable - it is possible that the monetary compensation acted as a motivator for high accuracy, which countered potential negative effects.

Lastly, the results of the p-Beauty contest were analysed and discussed. On average, number 40 was selected, which suggests that most participants used up to 1 step of iterated reasoning. Note that this is subpar performance according to Camerer et al. [2004], who claim that humans use an average of 1.5 iterative steps for many games.

In this chapter, I present and justify the design choices for the models used to explain the underlying theory of mind (ToM) mechanisms present in the “Cheryl’s Puzzle” dataset. In Section 4.1, I introduce Public Announcement Logic (PAL) as the logic used to formalize knowledge. In Section 4.2, I show how PAL can be applied to the second-order puzzle (and, by extension, to all puzzles). In Section 4.3, I present (variations of) two models: i) one epistemic model that implements PAL to always solve puzzles correctly and ii) one cutting model that reduces the ToM order of a statement down to a pre-determined lower (and, thus, easier to process) order. In Section 4.4, I introduce group-level random-effects Bayesian model selection (RFX-BMS) as the statistical method used to determine the goodness of fit of the aforementioned models. Lastly, in Section 4.5, I present *coherence* as an additional metric used to assess the goodness of fit of the models.

4.1 Public Announcement Logic

Public Announcement Logic (PAL) [Plaza, 1989, 2007] is an extension of epistemic logic that models how agents’ knowledge changes after a public announcement has been made. Let us define A as the finite set of agents and P as a countable set of atoms. Following van Ditmarsch et al. [2007], let us inductively define the language of PAL by the following Backus–Naur form:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_i\varphi \mid [\varphi]\varphi,$$

where $p \in P$ and $i \in A$. K_i is a knowledge operator and $K_i\varphi$ is read as “agent i knows that φ ”. Constructs of the form $[\varphi]\psi$ are specific to PAL and are read as “after every (public and truthful) announcement φ , it holds that ψ ”. The effect of such a public announcement is the restriction of the epistemic model to all the states (and all connections between these states) where formula ψ holds. Note that the operators \vee , \rightarrow and \perp will also be used throughout this chapter and have the usual interpretation.

Now, let us introduce epistemic (Kripke) models in the $S5$ system [see Priest, 2008 for an overview]. Given the same A and P , an epistemic model \mathcal{M} takes the form $\mathcal{M} = \langle S, \sim, V \rangle$, where:

- S is a non-empty set of states or “worlds” (i.e., the domain)
- $\sim: A \rightarrow \mathcal{P}(S \times S)$ is the accessibility relation. For brevity, let us write $\sim(i)$ as \sim_i for $i \in A$. Given two worlds $s_1, s_2 \in S$ and $i \in A$, we say that $s_1 \sim_i s_2$ if agent i considers the two worlds equally possible, given the information that i currently holds to be true (or known). Note that as a consequence of placing the model in the $S5$ system, the \sim relation is reflexive, symmetric and transitive.
- $V: P \rightarrow \mathcal{P}(S)$ is the valuation function. For brevity, let us write $V(p)$ as V_p for $p \in P$. V essentially associates to each atomic proposition $p \in P$ the set of worlds in which p holds true.

Finally, let us define the semantics of the language. Given $\mathcal{M} = \langle S, \sim, V \rangle$ for agent $i \in A$, atom $p \in P$ and a world $s \in S$, the following hold:

$$\begin{aligned} \mathcal{M}, s \models p & \quad \text{iff } s \in V_p \\ \mathcal{M}, s \models \neg\varphi & \quad \text{iff } \mathcal{M}, s \not\models \varphi \\ \mathcal{M}, s \models \varphi \wedge \psi & \quad \text{iff } \mathcal{M}, s \models \varphi \text{ and } \mathcal{M}, s \models \psi \\ \mathcal{M}, s \models K_i\varphi & \quad \text{iff for all } t \in S \\ & \quad \text{we have } \mathcal{M}, t \models \varphi \\ \mathcal{M}, s \models [\varphi]\psi & \quad \text{iff } \mathcal{M}, s \models \varphi \text{ implies } \mathcal{M} \upharpoonright \varphi, s \models \psi, \end{aligned}$$

where $\mathcal{M} \upharpoonright \varphi = \langle S', \sim', V' \rangle$ is defined as follows:

- $S' := \llbracket \varphi \rrbracket_{\mathcal{M}} := \{s \in S \mid \mathcal{M}, s \models \varphi\}$ is the subset of states where φ holds
- $\sim'_i := \sim_i \cap (\llbracket \varphi \rrbracket_{\mathcal{M}} \times \llbracket \varphi \rrbracket_{\mathcal{M}})$ for all $i \in A$ is the subset of relations that connects only states in S' .
- $V'_p := V_p \cap \llbracket \varphi \rrbracket_{\mathcal{M}}$ for all $p \in P$ is the subset of valuations associated with the states in S' .

4.2 Cheryl's Puzzle Formalism

Let us now consider the ‘‘Cheryl’s Birthday’’ epistemic puzzles, as described in Section 2.2. Following the formalism in van Ditmarsch et al. [2008], let us define the following sets:

- set of possible months:
 $M = \{\text{May, June, July, August, September}\}$
- set of possible days:
 $D = \{14, 15, 16, 17, 18\}$
- set of agents:
 $A = \{a, b\}$, where a stands for Albert and b stands for Bernard. Note that Cheryl is not considered an agent in this system, because her role is rather to ensure that Albert and Bernard have access to all the background information necessary to solve the puzzle. The participant could be considered a third agent in the system by attributing to them the universal relation. However, this is trivial and, therefore, for simplicity, the participant was excluded from the system.

Let us now only consider the second-order puzzle - the formalism for the other three puzzles should be easily inferred. We now define an additional set:

- set of birthday options on Cheryl’s list:
 $S_c = \{(\text{May}, 17), (\text{May}, 18), (\text{July}, 16), \dots, (\text{September}, 15)\}$.
Note that we have $S_c \subset M \times D$

Based on the formulation of the puzzle, we need to find $(m, d) \in S_c$, such that Cheryl’s birthday is in month m and on day d . Consider the variable m . If its value is May, then we can represent this as the truth value of the atomic proposition ‘‘ $m = \text{May}$ ’’. Alternatively, let us define this as the propositional letter m_{May} . More generally, let us define the following finite set of atoms: $\{m_i \mid (i, j) \in S_c\} \cup \{d_j \mid (i, j) \in S_c\}$.

The proposition ‘‘Albert knows that Cheryl’s birthday is on September, 14’’ can be represented as $K_a(m_{\text{September}} \wedge d_{14})$. More generally, the proposition ‘‘Albert knows Cheryl’s birthday’’, abbreviated as $K_a(m, d)$, can be represented as a disjunction over all possible birthday options on Cheryl’s list, namely:

$$\begin{aligned} K_a(m, d) & := K_a(m_{\text{May}} \wedge d_{17}) \vee K_a(m_{\text{May}} \wedge d_{18}) \\ & \quad \vee \dots \vee K_a(m_{\text{September}} \wedge d_{15}) \\ & = \bigvee_{(i,j) \in S_c} K_a(m_i \wedge d_j), \end{aligned}$$

because we know that Cheryl’s birthday is on at least one (in reality, on exactly one) of these dates. Note that, even though the participant could give ‘‘No solution’’ as an answer to the puzzles, it does not mean that Cheryl’s birthday cannot be on any of those dates (in fact, it is explicitly mentioned that the birthday *is* on the initial list of ten options) but rather that the information from the dialogue between Albert and Bernard is inconsistent with the setting in some way (which, by design of the puzzles, is never the case). Similarly, ‘‘Bernard knows Cheryl’s birthday’’ can be represented as $K_b(m, d) := \bigvee_{(i,j) \in S_c} K_b(m_i \wedge d_j)$.

The conversation between Albert and Bernard can be modeled as a series of public announcements, as follows:

1. **Albert:** “I don’t know when Cheryl’s birthday is.”: $\neg K_a(m, d)$
2. **Bernard:** “I didn’t know at first, but now I know.”: $\neg K_b(m, d) \wedge [\neg K_a(m, d)]K_b(m, d)$

Note that the “didn’t know” in announcement 2 by Bernard refers to the *initial* epistemic model rather than the epistemic state *resulting* from the announcement made by Albert. It can be verified easily that, given the initial epistemic state, Bernard could not have known the birthday from the beginning - this renders the first part of the announcement obsolete. Therefore, as we are interested in the resulting epistemic state and Bernard’s knowledge changed in the meantime, we can simplify the second to public announcement to $K_b(m, d)$ without any loss of information.

Now take a Kripke model $\mathcal{MC} = \langle S_c, \sim, V \rangle$ consisting of a domain of all pairs (m, d) with $(m, d) \in S_c$; with accessibility relations \sim_a and \sim_b such that for a : $(m, d) \sim_a (m', d')$ iff $m = m'$ and for b : $(m, d) \sim_b (m', d')$ iff $d = d'$; and with valuation V such that $V_{m_i} = \{(m, d) \in S_c \mid m = i\}$ and $V_{d_j} = \{(m, d) \in S_c \mid d = j\}$. For the second-order puzzle, the solution (namely, September 14) can be modeled as follows in PAL:

$$\mathcal{MC} \models [\neg K_a(m, d)][K_b(m, d)](m_{\text{September}} \wedge d_{14})$$

In the formula above, the sequence of two announcements can be truthful only in state (September, 14). This is because applying the two announcements consecutively results in a model with only one epistemic state (\mathcal{MC}, s) , within which $m_{\text{September}} \wedge d_{14}$ holds. Clearly, this can only be the case for $s = (\text{September}, 14)$.

For all other states, at least one of the announcements becomes false. For example, take $s' = (\text{June}, 14)$. Suppose that $\mathcal{MC}, s' \models \neg K_a(m_{\text{June}} \wedge d_{14})$. By definition of the $\not\models$ operator, we have $\mathcal{MC}, s' \not\models K_a(m_{\text{June}} \wedge d_{14})$. However, we have only $t = (\text{June}, 14)$ such that $s' \sim_a t$ and we have $\mathcal{MC}, t \models (\text{June}, 14)$. Therefore, we have $\mathcal{MC}, s' \models K_a(m_{\text{June}} \wedge d_{14})$. This leads to a contradiction or, more formally, $\mathcal{MC}, (m_{\text{June}} \wedge d_{14}) \models [\neg K_a(m, d)]\perp$.

Similarly, take $s'' = (\text{September}, 15)$. First, let us show that $\mathcal{MC}, s'' \models \neg K_a(m_{\text{September}} \wedge d_{15})$. By definition of the $\not\models$ operator, we have $\mathcal{MC}, s'' \not\models K_a(m_{\text{September}} \wedge d_{15})$. Now take $v =$

(September, 14) with $(\text{September}, 15) \sim v$. We then have that $\mathcal{MC}, (\text{September}, 14) \not\models m_{\text{September}} \wedge d_{15}$ (trivially, September 15th is not the birthday in a world where September 14th is the birthday). Then, by definition of the K operator, we indeed have that $\mathcal{MC}, s'' \models \neg K_a(m_{\text{September}} \wedge d_{15})$. Now suppose that $\mathcal{MC}, s'' \models K_b(m_{\text{September}} \wedge d_{15})$. By definition of the K operator, we then have that for all $w \in S_c$ such that $(\text{September}, 15) \sim_b w$, $\mathcal{MC}, w \models m_{\text{September}} \wedge d_{15}$. If we take the case $w = (\text{August}, 15)$, we then have that $\mathcal{MC}, (\text{August}, 15) \not\models m_{\text{September}} \wedge d_{15}$ (trivially, September 15th is not the birthday in a world where August 15th is the birthday). This leads to a contradiction or, more formally, $\mathcal{MC}, (\text{September}, 15) \models [\neg K_a(m, d)][K_b(m, d)]\perp$.

The above can similarly be shown for all states $s \in S_c \setminus (\text{September}, 14)$.

4.3 Models

In this section, I present (variations of) the two models used in the analysis. These models will then be fit on the participant data (see Chapter 5) using the RFX-BMS algorithm (see Section 4.4). I first intuitively describe the strategy used by each model and then I (formally) show how the models process each puzzle until they reach an answer. For an overview of the answers given by all (variations of) models for each puzzle, the reader is advised to skip ahead to Table 4.3.1.

4.3.1 Epistemic Model

The epistemic model processes each line of dialogue between Albert and Bernard as a series of public announcements. If multiple public announcements exist, then they are processed sequentially: The second public announcement is applied to the Kripke model restricted to the states (and connections between states) where the formula associated with the first public announcement holds. After all public announcements have thus been applied to the initial Kripke model, the answer is extracted as follows: If only one state remains, then that is the answer; otherwise, if no state remains, then the answer is “No solution”; otherwise, if multiple states remain, then the answer is “Multiple solutions”.

In the remainder of the section, I will show how

the epistemic model reaches the answer for each of the four puzzles discussed in Section 2.2, following the formalism and notations described in Section 4.2. Note that, by design, the epistemic model always reaches the correct solution for all puzzles.

1. First-order puzzle

Figure 4.3.1 shows the initial Kripke model associated with the first-order puzzle, where the states are the ten dates given as options for Cheryl's birthday and the arrows are the accessibility relations corresponding to Albert and Bernard, respectively. As a reminder, Albert knows the month and Bernard knows the day and this leads to the following conversation:

Bernard: "I know when Cheryl's birthday is."

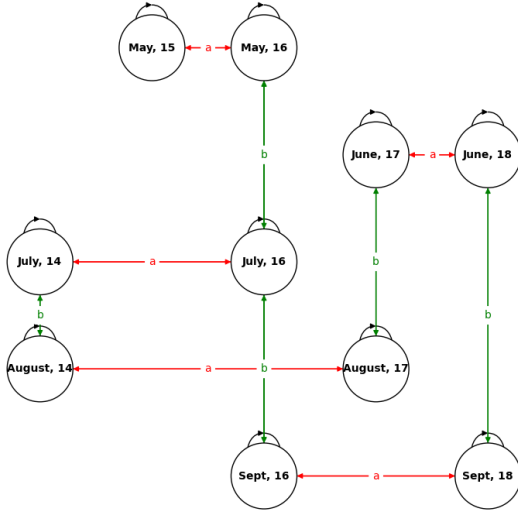


Figure 4.3.1: The initial Kripke model for the first-order puzzle. Transitive arrows are omitted for readability. Red arrows labeled with "a" mark the accessibility relation for Albert and green labeled with "b" for Bernard. Black (reflexive) arrows with no label are accessibility relations for both Albert and Bernard.

Only one public announcement occurs, which can be modeled as $K_b(m, d)$. This announcement restricts the Kripke model in Figure 4.3.1 to only those states (and connections between states) where the formula $K_b(m, d)$ holds. As

can be seen in Figure 4.3.2, this formula only holds in state (May, 15).

Let us show this fact, namely $\mathcal{MC}, (\text{May}, 15) \models K_b(m, d)$. By definition of the \vee operator, it is enough to show that $\mathcal{MC}, (\text{May}, 15) \models K_b(m_{\text{May}} \wedge d_{15})$. From the definition of the K -operator, we then have that for all $t \in S_c$ such that $(\text{May}, 15) \sim_b t$, $\mathcal{MC}, t \models m_{\text{May}} \wedge d_{15}$. We can only have $t = (\text{May}, 15)$, where it does follow that $\mathcal{MC}, (\text{May}, 15) \models m_{\text{May}} \wedge d_{15}$ (trivially, May 15th is the birthday in a world where May 15th is the birthday). Therefore, we have $\mathcal{MC}, (\text{May}, 15) \models K_b(m_{\text{May}} \wedge d_{15})$ and, further, we have $\mathcal{MC}, (\text{May}, 15) \models K_b(m, d)$.

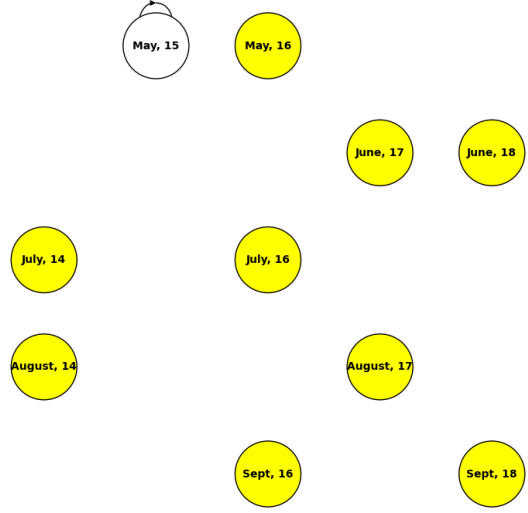


Figure 4.3.2: The Kripke model for the first-order puzzle after the first (and only) public announcement. Black (reflexive) arrows with no label are accessibility relations for both Albert and Bernard. States marked with yellow are removed after applying the announcement.

For completion, let us show that no other state s' meets the constraint that $\mathcal{MC}, s' \models K_b(m, d)$. To show that $\mathcal{MC}, s' \not\models K_b(m, d)$, we need to show that for all $w \in S_c \setminus \{(\text{May}, 15)\}$, $\mathcal{MC}, s' \not\models K_b w$. Take $s' = (\text{May}, 16)$ as an example and show that $\mathcal{MC}, (\text{May}, 16) \not\models K_b(m_{\text{May}} \wedge d_{16})$. Additionally, take $t = (\text{July}, 16)$ with $(\text{May}, 16) \sim_b t$. We then have that $\mathcal{MC}, (\text{July}, 16) \not\models m_{\text{May}} \wedge d_{16}$ (trivially, May 16th is not the birthday in a

world where July 16th is the birthday). Thus, by definition of the K operator, we have that $\mathcal{MC}, (\text{May}, 16) \not\models K_b(m_{\text{May}} \wedge d_{16})$. The same can be shown for all other states s' [except (May, 15)].

Since only one state remains after applying all public announcements, May 15 is the answer given by the epistemic model.

2. Second-order puzzle

Figure 4.3.3 shows the initial Kripke model associated with the second-order puzzle, where the states are the ten dates given as options for Cheryl's birthday and the arrows are the accessibility relations corresponding to Albert and Bernard, respectively. As a reminder, Albert knows the month and Bernard knows the day and this leads to the following conversation:

Albert: "I don't know when Cheryl's birthday is."

Bernard: "I didn't know at first, but now I know."

The first public announcement can be modeled as $\neg K_a(m, d)$. This announcement restricts the Kripke model in Figure 4.3.3 to only those states (and connections between states) where the formula $\neg K_a(m, d)$ holds. As can be seen in Figure 4.3.4, this formula does not hold in state (June, 14).

Let us show this fact, namely that $\mathcal{MC}, (\text{June}, 14) \not\models \neg K_a(m, d)$. By definition of $\not\models$ and \neg , we have $\mathcal{MC}, (\text{June}, 14) \models K_a(m, d)$. By definition of the \vee operator, it is enough to show that $\mathcal{MC}, (\text{June}, 14) \models K_a(m_{\text{June}} \wedge d_{14})$. By definition of the K operator, we then have that for all $t \in S_c$ such that $(\text{June}, 14) \sim_a t$, $\mathcal{MC}, t \models m_{\text{June}} \wedge d_{14}$. We only have $t = (\text{June}, 14)$, where it indeed follows that $\mathcal{MC}, (\text{June}, 14) \models m_{\text{June}} \wedge d_{14}$ (trivially, June 14th is the birthday in a world where June 14th is the birthday). This concludes the explanation.

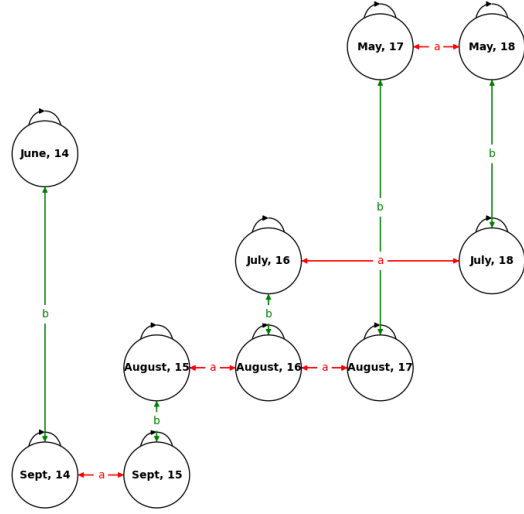


Figure 4.3.3: The initial Kripke model for the second-order puzzle. Red arrows labeled with "a" mark the accessibility relation for Albert and green labeled with "b" for Bernard. Black (reflexive) arrows with no label are accessibility relations for both Albert and Bernard.

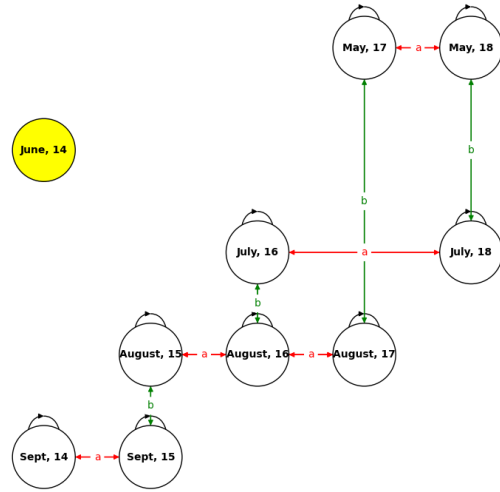


Figure 4.3.4: The Kripke model for the second-order puzzle after the first announcement. Red arrows labeled with "a" mark the accessibility relation for Albert and green labeled with "b" for Bernard. Black (reflexive) arrows with no label are accessibility relations for both Albert and Bernard. States marked with yellow are removed after applying the announcement.

The second public announcement can be modeled as $K_b(m, d)$. This announcement restricts the Kripke model in Figure 4.3.4 to only those states (and connections between states) where the formula $K_b(m, d)$ holds. As can be seen in Figure 4.3.5, this formula holds only in state (September, 14) - the explanation for this is similar to the one presented for the first-order puzzle. Since only one state remains after applying all announcements to the initial Kripke model, September 14 is the answer given by the epistemic model.

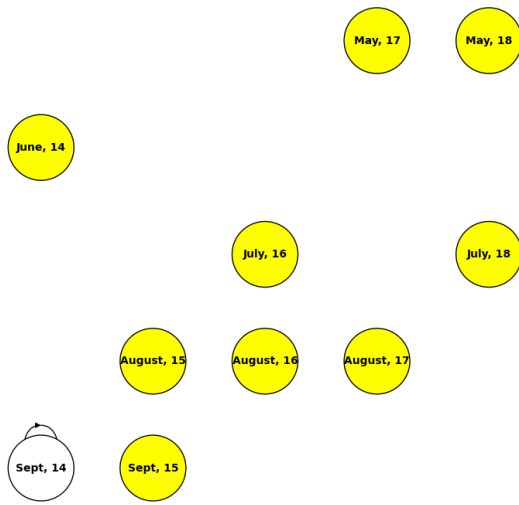


Figure 4.3.5: The Kripke model for the second-order puzzle after the second announcement. Black (reflexive) arrows with no label are accessibility relations for both Albert and Bernard. States marked with yellow are removed after applying the announcement.

3. Third-order puzzle

Figure 4.3.6 shows the initial Kripke model associated with the third-order puzzle, where the states are the ten dates given as options for Cheryl's birthday and the arrows are the accessibility relations corresponding to Albert and Bernard, respectively. As a reminder, Albert knows the month and Bernard knows the day and this leads to the following conversation:

Albert: "I know that you don't know when Cheryl's birthday is."

Bernard: "I didn't know at first, but now I know."

The first public announcement can be modeled as $K_a \neg K_b(m, d)$. This announcement restricts the Kripke model in Figure 4.3.6 to only those states (and connections between states) where the formula $K_a \neg K_b(m, d)$ holds. As can be seen in Figure 4.3.7, this formula does not hold in the following states: (May, 15), (May, 18), (June, 15) and (June, 17).

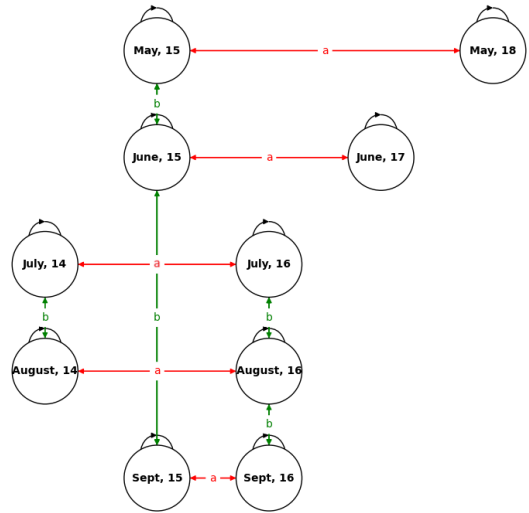


Figure 4.3.6: The initial Kripke model for the third-order puzzle. Transitive arrows are omitted for readability. Red arrows labeled with "a" mark the accessibility relation for Albert and green labeled with "b" for Bernard. Black (reflexive) arrows with no label are accessibility relations for both Albert and Bernard.

For $s' = (\text{May}, 15)$, let us show that indeed $\mathcal{MC}, (\text{May}, 15) \not\models K_a \neg K_b(m, d)$. Suppose instead that $\mathcal{MC}, (\text{May}, 15) \models K_a \neg K_b(m, d)$ and try to reach a contradiction. Now by definition of the K operator, we have that for all $t \in S_c$ such that $(\text{May}, 15) \sim_a t$, $\mathcal{MC}, t \models \neg K_b(m, d)$. We then have $t = \{(\text{May}, 15), (\text{May}, 18)\}$. Take $t = (\text{May}, 18)$ and reach a contradiction by showing that $\mathcal{MC}, (\text{May}, 18) \not\models \neg K_b(m, d)$. By definition of the $\not\models$ and \neg operators, we then have that $\mathcal{MC}, (\text{May}, 18) \models K_b(m, d)$. By definition of the \vee operator, it is then enough

to show that $\mathcal{MC}, (\text{May}, 18) \models K_b(m_{\text{May}} \wedge d_{18})$. By definition of the K operator, we then have that for all $w \in S_c$ such that $(\text{May}, 18) \sim_b w$, $\mathcal{MC}, w \models m_{\text{May}} \wedge d_{18}$. We can only have $w = (\text{May}, 18)$ and we indeed have $\mathcal{MC}, (\text{May}, 18) \models m_{\text{May}} \wedge d_{18}$ (trivially, May 18th is the birthday in a world where May 18th is the birthday). This concludes the explanation.

It can similarly be shown that $\mathcal{MC}, (\text{June}, 15) \not\models K_a \neg K_b(m, d)$. By the veridicality axiom* of S5, $\neg K_b(m, d)$ is entailed from the first public announcement. It is then possible to show that $\mathcal{MC}, (\text{May}, 18) \not\models \neg K_b(m, d)$ and $\mathcal{MC}, (\text{June}, 17) \not\models \neg K_b(m, d)$ by using the definition of $\not\models$ and following the explanation for the first-order puzzle.

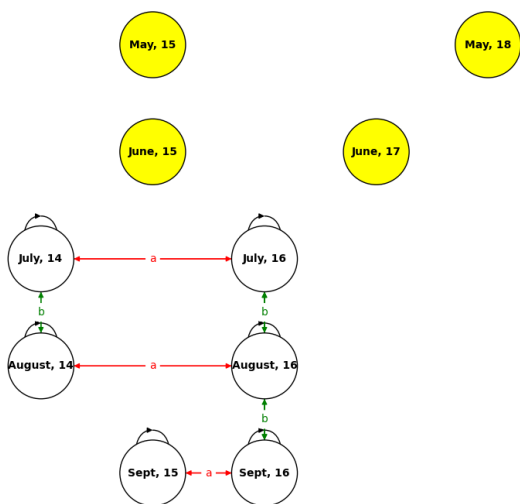


Figure 4.3.7: The Kripke model for the third-order puzzle after the first announcement. Transitive arrows are omitted for readability. Red arrows labeled with “a” mark the accessibility relation for Albert and green labeled with “b” for Bernard. Black (reflexive) arrows with no label are accessibility relations for both Albert and Bernard. States marked with yellow are removed after applying the announcement.

*The veridicality axiom of S5 formally states that only true formulae are known by agents or, more formally, that $K_i\varphi \rightarrow \varphi$, for all agents $i \in A$.

The second public announcement can be modeled as $K_b(m, d)$. This announcement restricts the Kripke model in Figure 4.3.7 to only those states (and connections between states) where the formula $K_b(m, d)$ holds. As can be seen in Figure 4.3.8, this formula holds only in state (September, 15) - the explanation for this is similar to the one presented for the first-order puzzle. Since only one state remains after applying all announcements to the initial Kripke model, then September 15 is the answer given by the epistemic model.

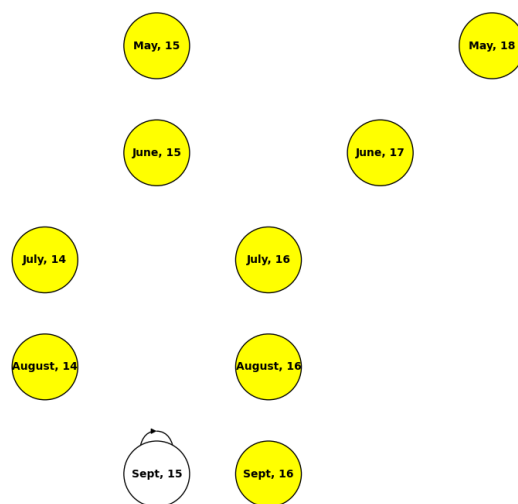


Figure 4.3.8: The Kripke model for the third-order puzzle after the second announcement. Black (reflexive) arrows with no label are accessibility relations for both Albert and Bernard. States marked with yellow are removed after applying the announcement.

4. Fourth-order puzzle

Figure 4.3.9 shows the initial Kripke model associated with the fourth-order puzzle, where the states are the ten dates given as options for Cheryl’s birthday and the arrows are the accessibility relations corresponding to Albert and Bernard, respectively. As a reminder, Albert knows the month and Bernard knows the day and this leads to the following conversation:

Bernard: “I know that you know that I don’t know when Cheryl’s birthday is.”

Albert: “I didn’t know at first but now I know.”

The first public announcement can be modeled as $K_b K_a \neg K_b(m, d)$. This announcement restricts the Kripke model in Figure 4.3.9 to only those states (and connections between states) where the formula $K_b K_a \neg K_b(m, d)$ holds. As can be seen in Figure 4.3.10, this formula does not hold in the following states: (May, 15), (June, 14) and (June, 15).

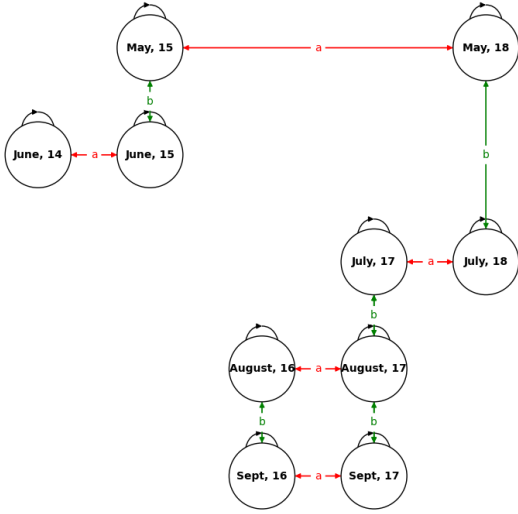


Figure 4.3.9: The initial Kripke model for the fourth-order puzzle. Transitive arrows are omitted for readability. Red arrows labeled with “a” mark the accessibility relation for Albert and green labeled with “b” for Bernard. Black (reflexive) arrows with no label are accessibility relations for both Albert and Bernard.

For $s' = (\text{May}, 15)$, let us show that indeed $\mathcal{MC}, (\text{May}, 15) \not\models K_b K_a \neg K_b(m, d)$. Suppose that $\mathcal{MC}, (\text{May}, 15) \models K_b K_a \neg K_b(m, d)$ and try to reach a contradiction. By definition of the K operator, we have that for all $t \in S_c$ such that $(\text{May}, 15) \sim_b t$, $\mathcal{MC}, t \models K_a \neg K_b(m, d)$. We can only have $t = \{(\text{May}, 15), (\text{June}, 15)\}$, so let us take $t_1 = (\text{June}, 15)$. We now have $\mathcal{MC}, (\text{June}, 15) \models K_a \neg K_b(m, d)$. By definition of the K operator, we have that for

all $w \in S_c$ such that $(\text{June}, 15) \sim_a w$, $\mathcal{MC}, w \models \neg K_b(m, d)$. We can only have $w = \{(\text{June}, 15), (\text{June}, 14)\}$, so let us take $w_1 = (\text{June}, 14)$. We reach a contradiction if we show that $\mathcal{MC}, (\text{June}, 14) \not\models \neg K_b(m, d)$. By definition of $\not\models$ and \neg , we have that $\mathcal{MC}, (\text{June}, 14) \models K_b(m, d)$. By definition of the \vee operator, it is then enough to show that $\mathcal{MC}, (\text{June}, 14) \models K_b(m_{\text{June}} \wedge d_{14})$. By definition of the K operator, we then have that for all $v \in S_c$ such that $(\text{June}, 14) \sim_b v$, $\mathcal{MC}, v \models m_{\text{June}} \wedge d_{14}$. We can only have $v = (\text{June}, 14)$ and indeed we have $\mathcal{MC}, (\text{June}, 14) \models m_{\text{June}} \wedge d_{14}$ (trivially, June 14th is the birthday in a world where June 14th is the birthday). This concludes the explanation.

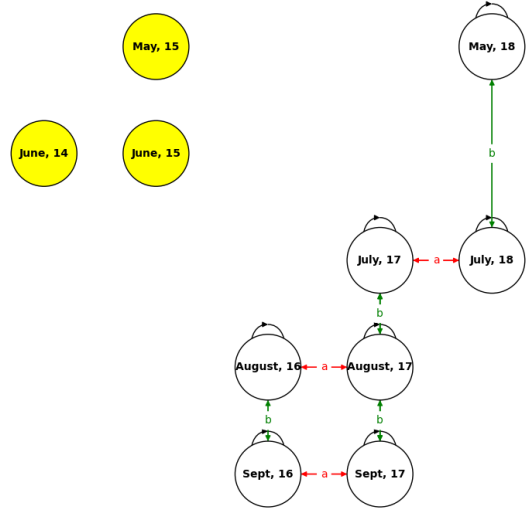


Figure 4.3.10: The Kripke model for the fourth-order puzzle after the first announcement. Transitive arrows are omitted for readability. Red arrows labeled with “a” mark the accessibility relation for Albert and green labeled with “b” for Bernard. Black (reflexive) arrows with no label are accessibility relations for both Albert and Bernard. States marked with yellow are removed after applying the announcement.

By the veridicality axiom of S5, $K_a \neg K_b(m, d)$ is entailed from the first public announcement. It is then enough to show that $\mathcal{MC}, (\text{June}, 15) \not\models K_a \neg K_b(m, d)$, which has been shown as part of the explanation for

$s' = (\text{May}, 15)$. Further, by the veridicality axiom of S5, $\neg K_b(m, d)$ follows from $K_a \neg K_b(m, d)$. Thus, it is enough to show that $\mathcal{MC}, (\text{June}, 14) \not\models \neg K_b(m, d)$, which has also been shown as part of the explanation for $s' = (\text{May}, 15)$.

The second public announcement can be modeled as $K_a(m, d)$. This announcement restricts the Kripke model in Figure 4.3.10 to only those states (and connections between states) where the formula $K_a(m, d)$ holds. As can be seen in Figure 4.3.11, this formula holds only in state $(\text{May}, 18)$ - the explanation for this is similar to the one presented for the first-order puzzle. Since only one state remains after applying all announcements to the initial Kripke model, then May 18 is the answer given by the epistemic model.

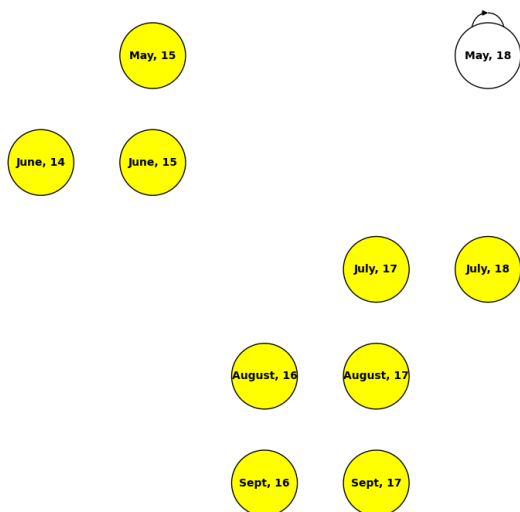


Figure 4.3.11: The Kripke model for the fourth-order puzzle after the second announcement. Transitive arrows are omitted for readability. Black (reflexive) arrows with no label are accessibility relations for both Albert and Bernard. States marked with yellow are removed after applying the announcement.

4.3.2 Cutting Model

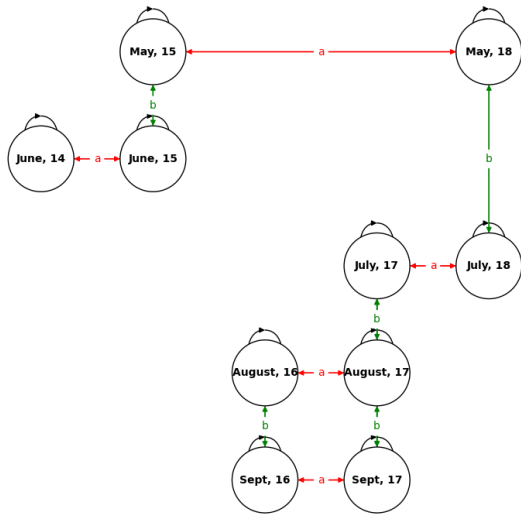
The cutting model is constructed in such a way that it is able to process statements up to a pre-determined ToM order and all public announcements exceeding this ToM order undergo certain transformations. To this end, the cutting model receives two parameters that determine its behaviour: a maximum ToM order and a cutting direction. The ToM order determines the maximum order of ToM that the model is capable of processing: For all statements with a higher ToM order, K operators are sequentially removed (along with the negation preceding them, if applicable) until the new statement is of the maximum ToM order; otherwise, the statements remain unchanged. The cutting direction determines the order in which K operators are removed. Once all public announcements have been reduced to (at most) the required ToM order, the puzzle is solved the same way as the epistemic model (see Section 4.3.1). The four possible configurations are detailed below:

1. **Cut model 2-lr:**

This model cuts public announcements down to (at most) second-order ToM statements and operators are removed sequentially from the left-most operator to the right-most (hence, the “lr” parameter).

The public announcements associated with the first, second and third-order puzzles remain unchanged, as they already are at most second-order ToM statements - this also implies that the cut 2-lr and epistemic models provide the same answers for these puzzles.

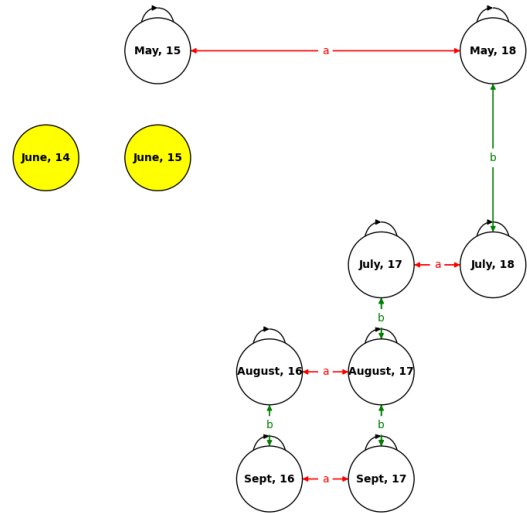
For the fourth-order puzzle, the first public announcement $K_b K_a \neg K_b(m, d)$ is first reduced to $K_a \neg K_b(m, d)$ before being applied to the initial Kripke model associated with the puzzle, while the second public announcement remains unchanged. As a result, the cut 2-lr model answers “No solution”, as shown in Figure 4.3.12.



(a) Initial Kripke model for the fourth-order puzzle.

Figure 4.3.12: The cut 2-1r model answers “No solution” for the fourth-order puzzle. Transitive arrows are omitted for readability. Red arrows labeled with “a” mark the accessibility relation for Albert and green labeled with “b” for Bernard. Black (reflexive) arrows with no label are accessibility relations for both Albert and Bernard. States marked with yellow are removed after applying the announcement.

nouncement to the Kripke model, all remaining states are removed. This is because, in all of these states, Albert cannot differentiate the state itself from (at least) another adjacent one and, therefore, in none of these states would he know the birthday - this contradicts the formula associated with the public announcement.

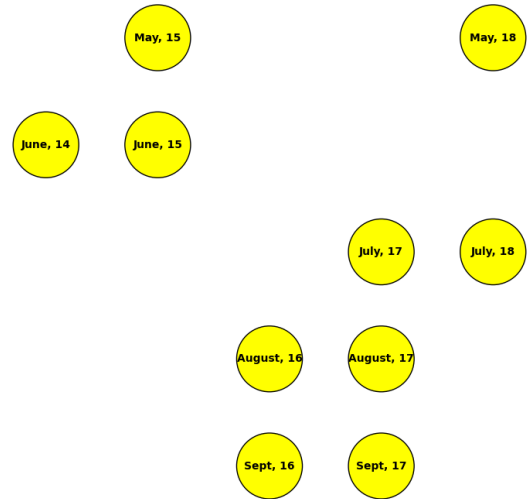


(b) Kripke model after the first public announcement: $K_a \neg K_b(m, d)$.

Briefly, it can be shown that $\mathcal{MC} \models [K_a \neg K_b(m, d)][K_a(m, d)] \perp$ as follows:

After applying the first public announcement to the Kripke model, states $\{(June, 14), (June, 15)\}$ are removed. State (June, 14) is removed because, in this state, Bernard does not consider another state possible (other than the state itself) and, therefore, in this state, he would know the birthday - this contradicts the formula associated with the public announcement (by the veridicality axiom of S5). State (June, 15) is removed because, from this state, Albert can access state (June, 14) and in state (June, 14) Bernard knows the birthday - this contradicts the formula associated with the public announcement.

After applying the second public an-



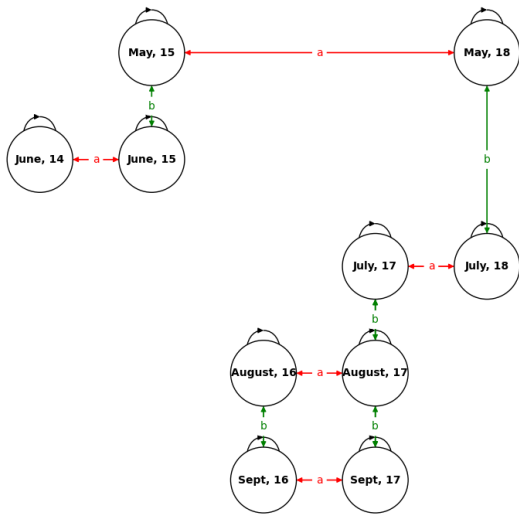
(c) Kripke model after the second public announcement: $K_a(m, d)$.

2. **Cut model 2-rl:**

This model cuts public announcements down to (at most) second-order ToM statements and operators are removed sequentially from the right-most operator to the left-most (hence, the “rl” parameter).

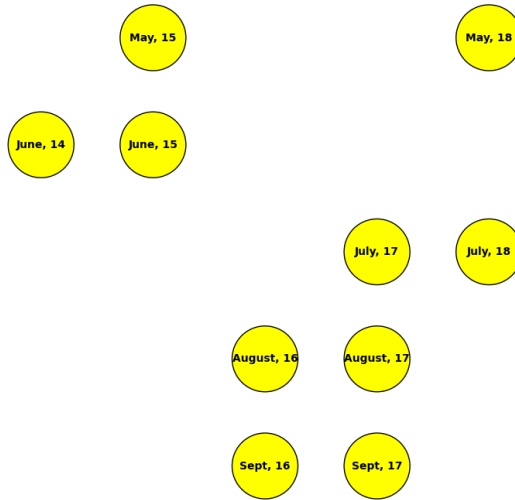
The public announcements associated with the first, second and third-order puzzles remain unchanged, as they already are at most second-order ToM statements - this also implies that the cut 2-rl and epistemic models provide the same answers for these puzzles.

For the fourth-order puzzle, the first public announcement $K_bK_a\neg K_b(m, d)$ is first reduced to $K_bK_a(m, d)$ before being applied to the initial Kripke model associated with the puzzle, while the second public announcement remains unchanged. As a result, the cut 2-rl model answers “No solution”, as shown in Figure 4.3.13.

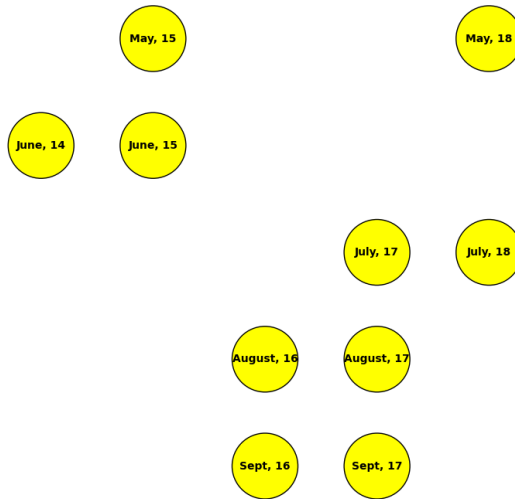


(a) Initial model for the fourth-order puzzle.

Figure 4.3.13: The cut 2-rl model answers “No solution” for the fourth-order puzzle. Transitive arrows are omitted for readability. Red arrows labeled with “a” mark the accessibility relation for Albert and green labeled with “b” for Bernard. Black (reflexive) arrows with no label are accessibility relations for both Albert and Bernard. States marked with yellow are removed after applying the announcement.



(b) Kripke model after the first public announcement: $K_bK_a(m, d)$.



(c) Kripke model after the second public announcement: $K_a(m, d)$.

Briefly, it can be shown that $\mathcal{MC} \models [K_bK_a(m, d)][K_a(m, d)]\perp$ as follows:

After applying the first public announcement to the Kripke model, all states are removed. This is because, in all of these states, Albert cannot differentiate the state itself from (at least) another adjacent one and, therefore, in none of these states would he know the birthday - this

contradicts the formula associated with the public announcement (by the veridicality axiom of S5).

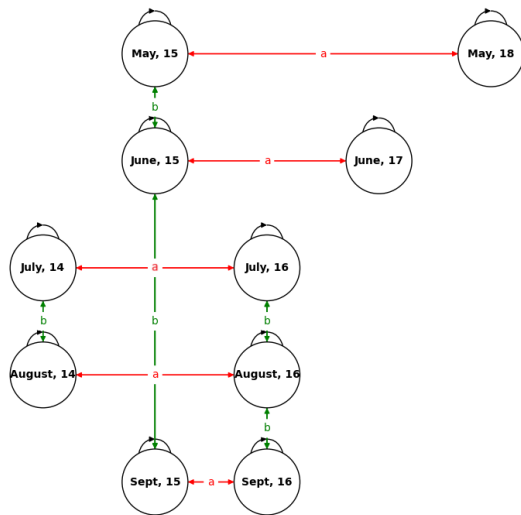
The second public announcement is applied on an empty Kripke model and, therefore, it has no effect.

3. Cut model 1-lr:

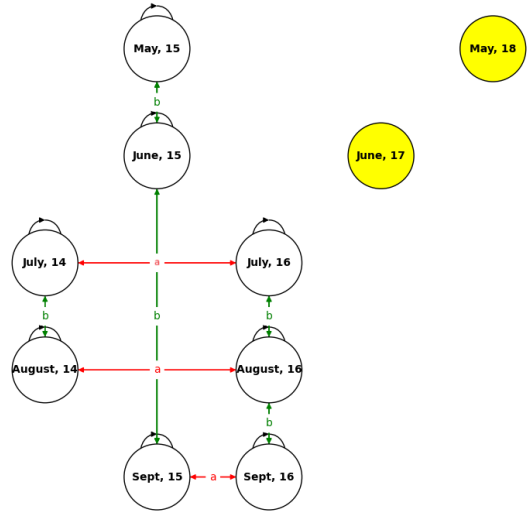
This model cuts public announcements down to first-order ToM statements and operators are removed sequentially from the left-most operator to the right-most (hence, the “lr” parameter).

The public announcements associated with the first and second-order puzzles remain unchanged, as they already are first-order ToM statements - this also implies that the cut 1-lr and epistemic models provide the same answers for these puzzles.

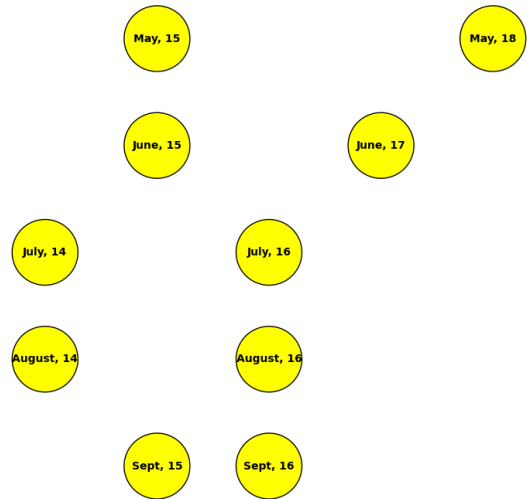
For the third-order puzzle, the first public announcement $K_a \neg K_b(m, d)$ is first reduced to $\neg K_b(m, d)$ before being applied to the initial Kripke model associated with the puzzle, while the second public announcement remains unchanged. As a result, the cut 1-lr model answers “No solution”, as shown in Figure 4.3.14.



(a) Initial Kripke model for the third-order puzzle.



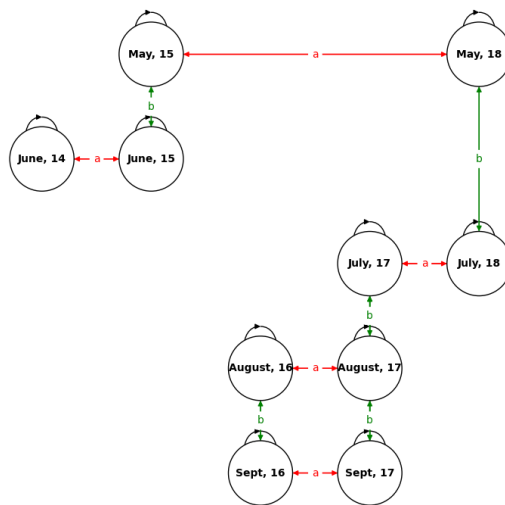
(b) Kripke model after the first public announcement: $\neg K_b(m, d)$.



(c) Kripke model after the second public announcement: $K_b(m, d)$.

Figure 4.3.14: The cut 1-lr model answers “No solution” for the third-order puzzle. Transitive arrows are omitted for readability. Red arrows labeled with “a” mark the accessibility relation for Albert and green labeled with “b” for Bernard. Black (reflexive) arrows with no label are accessibility relations for both Albert and Bernard. States marked with yellow are removed after applying the announcement.

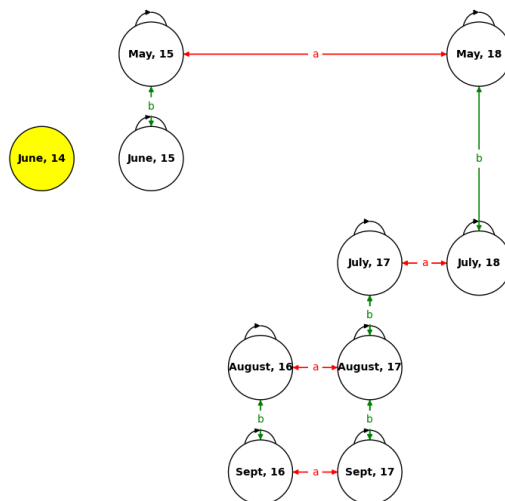
Briefly, it can be shown that $\mathcal{MC} \models [\neg K_b(m, d)][K_b(m, d)]\perp$ as follows:



(a) Initial Kripke model for the fourth-order puzzle.

After applying the first public announcement to the Kripke model, states $\{(May, 18), (June, 17)\}$ are removed. This is because, in both of these states, Bernard does not consider another state possible (other than the state itself) and, therefore, in both of these states, he would know the birthday - this contradicts the formula associated with the public announcement.

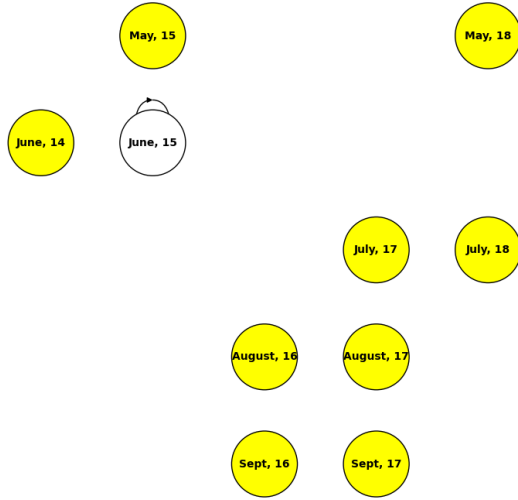
After applying the second public announcement to the Kripke model, all remaining states are removed. This is because, in all of these states, Bernard cannot differentiate the state itself from (at least) another adjacent one and, therefore, in none of these states would he know the birthday - this contradicts the formula associated with the public announcement.



(b) Kripke model after the first public announcement: $\neg K_b(m, d)$.

Figure 4.3.15: The cut 1-lr model answers “June, 15” for the fourth-order puzzle. Transitive arrows are omitted for readability. Red arrows labeled with “a” mark the accessibility relation for Albert and green labeled with “b” for Bernard. Black (reflexive) arrows with no label are accessibility relations for both Albert and Bernard. States marked with yellow are removed after applying the announcement.

For the fourth-order puzzle, the first public announcement $K_b K_a \neg K_b(m, d)$ is first reduced to $\neg K_b(m, d)$ before being applied to the initial Kripke model associated with the puzzle, while the second public announcement remains unchanged. As a result, the cut 1-lr model answers “June, 15”, as shown in Figure 4.3.15.



(c) Kripke model after the first public announcement: $K_a(m, d)$.

Briefly, it can be shown that $\mathcal{MC} \models [\neg K_b(m, d)][K_a(m, d)](m_{\text{June}} \wedge d_{15})$ as follows:

After applying the first public announcement to the Kripke model, state (June, 14) is removed. This is because, in this state, Bernard does not consider another state possible (other than the state itself) and, therefore, in this state, he would know the birthday - this contradicts the formula associated with the public announcement.

After applying the second public announcement to the Kripke model, all states except one are removed. This is because, in all of these states, Albert cannot differentiate the state itself from (at least) another adjacent one and, therefore, in none of these states would he know the birthday - this contradicts the formula associated with the public announcement.

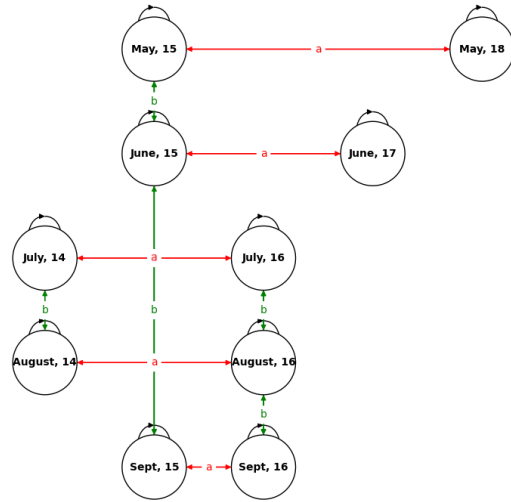
4. Cut model 1-rl:

This model cuts public announcements down to first-order ToM statements and operators are removed sequentially from the right-most operator to the left-most (hence, the “rl” parameter).

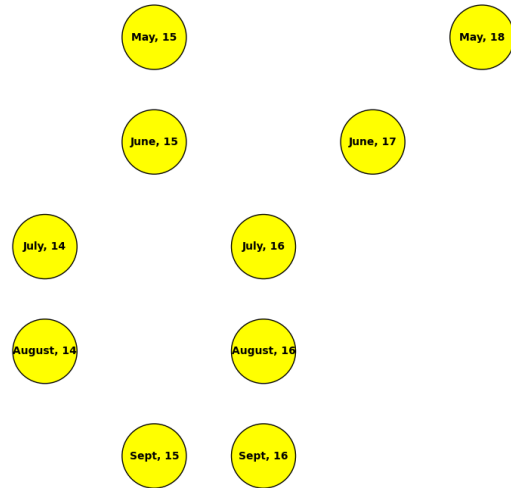
The public announcements associated with the first and second-order puzzles remain un-

changed, as they already are first-order ToM statements - this also implies that the cut 1-rl and epistemic models provide the same answers for these puzzles.

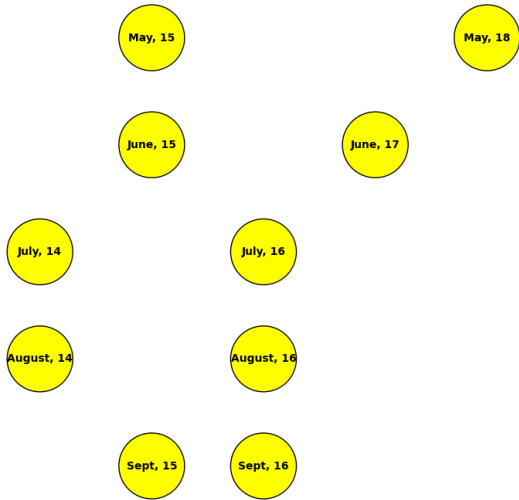
For the third-order puzzle, the first public announcement $K_a \neg K_b(m, d)$ is first reduced to $K_a(m, d)$ before being applied to the initial Kripke model associated with the puzzle, while the second public announcement remains unchanged. As a result, the cut 1-rl model answers “No solution”, as shown in Figure 4.3.16.



(a) Initial Kripke model for the third-order puzzle.



(b) Kripke model after the second public announcement: $K_b(m, d)$.



(c) Kripke model after the first public announcement: $K_a(m, d)$.

Figure 4.3.16: The cut 1-rl model answers “No solution” for the third-order puzzle. Transitive arrows are omitted for readability. Red arrows labeled with “a” mark the accessibility relation for Albert and green labeled with “b” for Bernard. Black (reflexive) arrows with no label are accessibility relations for both Albert and Bernard. States marked with yellow are removed after applying the announcement.

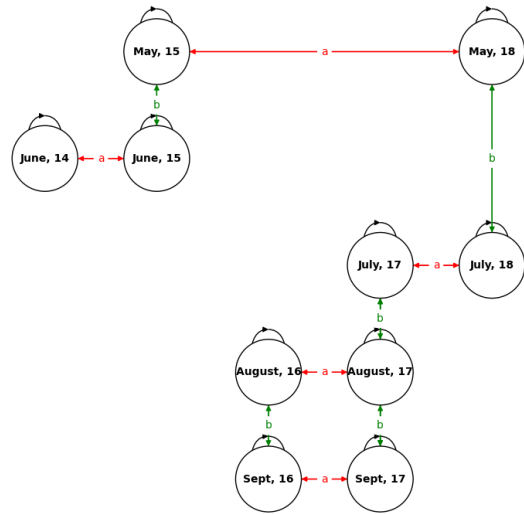
Briefly, it can be shown that $\mathcal{MC} \models [K_b(m, d)][K_a(m, d)]\perp$ as follows:

After applying the first public announcement to the Kripke model, all states are removed. This is because, in all states, Albert cannot differentiate the state itself from (at least) another adjacent one and, therefore, in none of these states would he know the birthday - this contradicts the formula associated with the public announcement.

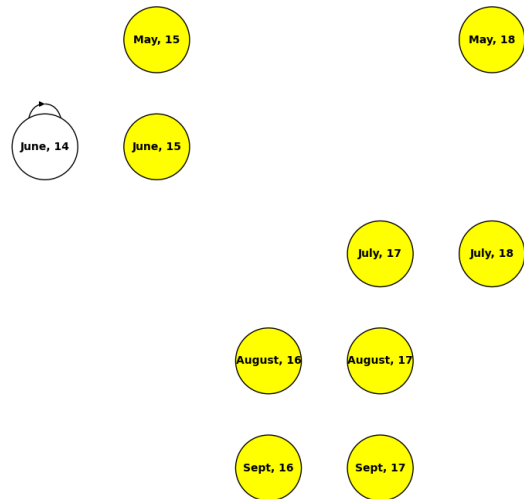
The second public announcement is applied on an empty Kripke model and, therefore, it has no effect.

For the fourth-order puzzle, the first public announcement $K_b K_a \neg K_b(m, d)$ is first reduced to $K_b(m, d)$ before being applied to the initial Kripke model associated with the puzzle,

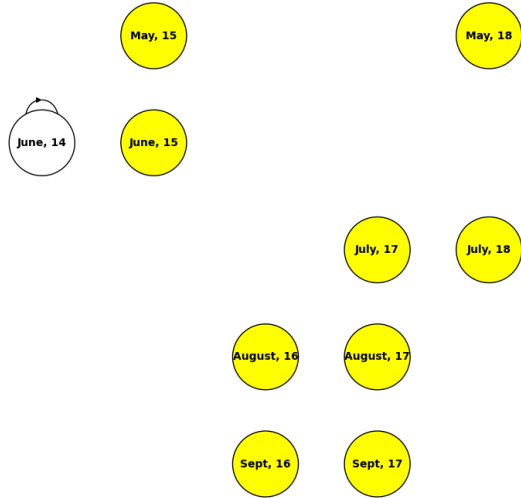
while the second public announcement remains unchanged. As a result, the cut 1-rl model answers “June, 14”, as show in Figure 4.3.17.



(a) Initial Kripke model for the fourth-order puzzle.



(b) Kripke model after the first public announcement: $K_b(m, d)$.



(c) Kripke model after the first public announcement: $K_a(m, d)$

Figure 4.3.17: The cut 1-rl model answers “June, 14” for the fourth-order puzzle. Transitive arrows are omitted for readability. Red arrows labeled with “a” mark the accessibility relation for Albert and green labeled with “b” for Bernard. Black (reflexive) arrows with no label are accessibility relations for both Albert and Bernard. States marked with yellow are removed after applying the announcement.

Briefly, it can be shown that $\mathcal{MC} \models [K_b(m, d)][K_a(m, d)](m_{\text{June}} \wedge d_{14})$ as follows:

After applying the first public announcement to the Kripke model, all states except one are removed. This is because, in all of these states, Bernard cannot differentiate the state itself from (at least) another adjacent one and, therefore, in none of these states would he know the birthday - this contradicts the formula associated with the public announcement.

After applying the second public announcement to the Kripke model, only state (June, 14) remains because, in this state, Albert does not consider another state possible (other than the state itself) and, therefore, he would know the birthday - this follows the formula associated with the public announcement.

Table 4.3.1 shows an overview of the answers given by all models for all four unique puzzles presented in Section 2.2. The puzzles were aggregated over scenario and configuration. This was possible because no effect was found on scenario with regards to accuracy and solving time (see Chapter 3) and configuration was not meant to be a condition in this experiment in the first place, but rather a way to generate a large variety of fundamentally equivalent puzzles.

Model	Order puzzle	Model answer	Correct answer
Epistemic	1	May 15	May 15
	2	Sept. 14	Sept. 14
	3	Sept. 15	Sept. 15
	4	May 18	May 18
Cut 1-lr	3	No solution	Sept. 15
	4	June 15	May 18
Cut 1-rl	3	No solution	Sept. 15
	4	June 14	May 18
Cut 2-lr	4	No solution	May 18
Cut 2-rl	4	No solution	May 18

Table 4.3.1: The answers given by all models for all four unique puzzles. For the Cut models, puzzles were omitted from the table because the answer given in those cases are identical to the answers given by the epistemic model for the same puzzle. “Sept” stands for “September”.

4.4 RFX-BMS

To estimate the ToM order of the population participants were drawn from, I apply a statistical method known as group-level random-effects Bayesian model selection (RFX-BMS), first proposed by Stephan et al. [2009]. Unlike fixed-effects Bayesian model selection, which assumes that one single strategy fits all participants, RFX-BMS treats strategies s as random effects that generate pieces of evidence $p(y | s)$, representing the probability that strategy s generated some observed data y (see Equation 4.4.1). Under the assumption that each participant is drawn from a fixed distribution of pre-defined strategies (i.e., the strategies used by the models described in Section 4.3), the aim is to estimate the frequency of the different ToM strategies within the population of participants. Unlike

Maximum Likelihood Estimation, RFX-BMS allows one to make more general claims about the distribution of strategies and is more sensitive to small differences between model predictions and participant data [Stephan et al., 2009, de Weerd et al., 2018, Veltman et al., 2019, Top et al., 2023].

The algorithm for RFX-BMS was implemented according to the following steps:

1. The Models described in Section 4.3 generate unique answers for each of the four puzzles (see Section 2.2).
2. Assume that participants follow one strategy s associated with one of the models. Strategy s should be “good enough” to explain *most* of the data but allow for a certain rate p (see Equation 4.4.1) of deviation from the predictions.
3. Decide which strategy s best fits a participant by following equation (14) in Stephan et al. [2009], which maximizes the log-likelihood that each participant uses each strategy by iteratively updating the strategy frequencies until some convergence point. This log-likelihood is computed as follows:

$$\begin{aligned} \text{likelihood} &= \ln(p(y | s)) = \\ &= n(1 - e) \ln(1 - e) + n \cdot e \ln(p \cdot e), \end{aligned} \quad (4.4.1)$$

where n is the total number of decision points for a participant (i.e., the number of trials completed), p is the probability of choosing an incorrect answer (here, $p = \frac{1}{12}$, because there are 13 answers in total and, therefore, 12 incorrect answers) and e is the error rate, which is computed as follows:

$$e = \frac{\text{number of incoherent predictions}}{n}, \quad (4.4.2)$$

where a predicted answer is *coherent* if it matches exactly the answer of the participant and otherwise it is *incoherent*.

4. By design, RFX-BMS will always classify data as belonging to one of the pre-defined strategy, even if the fit is extremely poor. For this reason, in addition to the models and associated

strategies described in Section 4.3, I also included a random model in the analysis meant to be chosen only when the other models (completely) fail to explain the answers given by some participants. The random model is fit to the entire population and follows the algorithm below:

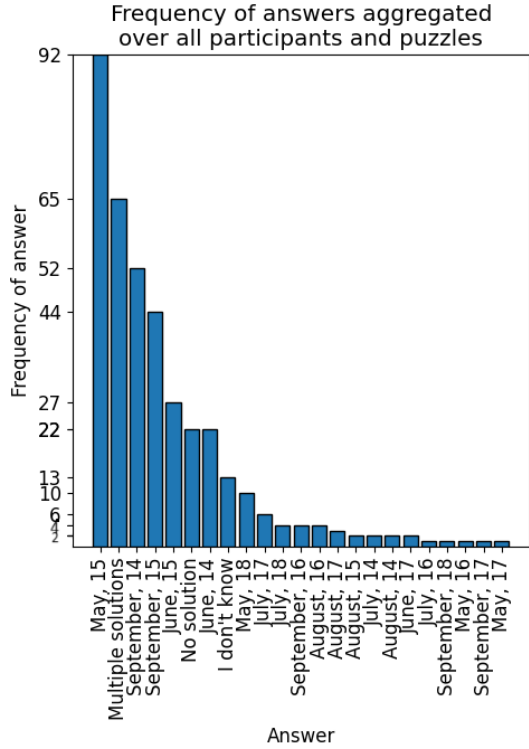


Figure 4.4.1: Frequency of answers aggregated over all participants and all puzzles for the entire dataset (i.e., both correct and incorrect answers, $N = 381$). The puzzles and participant answers were first translated to the original four unique puzzles and then aggregated over. Note that the frequency of answers is used in the computation of the likelihood and coherence for the random model.

- (a) Compute the distribution of frequency of answers for the population (i.e., aggregated over all participants and all puzzles) - see Figure 4.4.1.
- (b) For each participant, compute the log-likelihood of the random model according to the following formula:

$$\text{likelihood} = \sum_{i=1}^n \log(p_{answer}^i), \quad (4.4.3)$$

where n is the total number of decision points for a participant (i.e., the number of trials completed) and p_{answer}^i is the probability of encountering the participant's i -th answer in the population. For example, $p_{(\text{May},15)} = \frac{92}{381} \approx 0.24$, where 381 is the total number of answers aggregated over all participants and all puzzles.

4.5 Additional Metric

RFX-BMS estimates a distribution over a set of pre-defined strategies. By design, the algorithm is constrained to selecting one best fit amongst these strategies for each participant and, as a direct consequence, it may disregard how *closely* these strategies actually fit the data. Therefore, it is important to compute the *coherence* (a.k.a., correct rate or error rate) of the seemingly best-fitting model, namely the proportion of model predictions that exactly match the participant's data. For the models described in Section 4.3, the coherence is computed as in Equation 4.4.2. For the random model, the coherence is computed as follows:

$$\text{coherence} = \frac{\sum_{i=1}^n p_{answer}^i}{n}, \quad (4.5.1)$$

where n is the total number of decision points for a participant (i.e., the number of trials completed) and p_{answer}^i is the probability of encountering the participant's i -th answer in the population. As before, the computation of p_{answer}^i is based on the distribution of frequency of answers for the entire population of participants. Additionally, note that the division by n essentially restricts the coherence to the range $[0, 1]$.

Intuitively, the larger the coherence, the better the fit of the participant's data to the strategy used by a model - a coherence value of one means that the model perfectly predicts all the answers of the participant.

In this chapter, I discuss the goodness of fit for the computational models described in Chapter 4 on the data of the forty-nine participants that took part in the experiment described in Chapter 2. In Section 5.1, I remind the reader of the model configurations investigated in the current study. In Section 5.2, I present the distribution of answers given by the participants for the four puzzles and qualitatively compare these against the answers given by the models. In Section 5.3, I explain how to quantitatively assess the fit of the models to the participants' answers through a group-level random-effects Bayesian model selection (RFX-BMS) analysis, as proposed in Chapter 4. In Section 5.4, I present the results of the RFX-BMS analysis and distributions over coherence values for the proposed models in specific settings. Lastly, in Section 5.5, I summarize the findings and discuss implications.

5.1 Model Configurations

The following five model configurations, representing different possible reasoning strategies by adults as outlined in Chapter 4, have been included in the analysis. Note that the following model naming conventions will be adhered to throughout the chapter:

1. Epistemic model: The conversation between Albert and Bernard is modeled as a series of public announcements and each public announcement restricts the (Kripke) model. This model always finds the correct solution to all puzzles.
2. Cut 1-lr model: Knowledge operators (along with their preceding negations if applicable) are sequentially removed **from left to right**, until all public announcements are **first-order** theory of mind (ToM) statements (i.e., containing no chained knowledge operators referring to different agents). For example, the statement $K_d K_e \neg K_f p$ becomes $\neg K_f p$, for some agents d, e and f and propositional atom p . Otherwise, the public announcements are processed identically to the epistemic model.
3. Cut model 1-rl: Knowledge operators (along with their preceding negations if applicable) are sequentially removed **from right to left**, until all public announcements are **first-order** ToM statements (i.e., containing no chained knowledge operators referring to different agents). For example, the statement $K_d K_e \neg K_f p$ becomes $K_d p$, for some agents d, e and f and propositional atom p . Otherwise, the public announcements are processed identically to the epistemic model.
4. Cut model 2-lr: Knowledge operators (along with their preceding negations if applicable) are sequentially removed **from left to right**, until all public announcements are (at most) **second-order** ToM statements (i.e., containing at most two chained knowledge operators referring to different agents). For example, the statement $K_d K_e \neg K_f p$ becomes $K_e \neg K_f p$, for some agents d, e and f and propositional atom p . Otherwise, the public announcements are processed identically to the epistemic model.

5. Random model: It randomly selects an answer from the list of possible answers, based on the answers of the participant pool - the more commonly chosen answer x is by the sample of participants, the more likely the random model is to answer x (see Section 4.4 for more information).

Importantly, the Cut 2 model always gives the same answers to all puzzles whether the cutting is done from the left to the right side or from the right to the left side - see Cut 2-lr and Cut 2-rl in Table 4.3.1. Therefore, as these two models are indistinguishable from each other in terms of the answers given to the puzzles in this study, only the Cut 2-lr model is considered for the analysis.

5.2 Participant Answers

Let us now take a closer look at the distribution of answers selected by the forty-nine participants. Figure 5.2.1 shows the same information as Figure 4.4.1 but separated over the four puzzles. For example, “May, 15” was selected as an answer 92 times by the sample of participants and, more specifically, 78 times as the answer to a first-order puzzle (blue bar), 3 times as the answer to a third-order puzzle (green bar) and 11 times as the answer to a fourth-order puzzle (red bar). Note that, for the purpose of this analysis, the puzzles were aggregated over scenario and configuration. This was possible because no effect was found on scenario with regards to accuracy and solving time (see Chapter 3) and configuration was not meant to be a condition in this experiment in the first place, but rather a way to generate a large variety of fundamentally equivalent puzzles. The aggregation was done by “translating” back to the original four ToM puzzles described in Section 2.2.

If we expect most participants to answer the puzzles correctly, then we would be able to notice peaks associated with the correct answer for each puzzle. This is indeed the case for some of the puzzles: The correct answer for the first-order puzzle is “May, 15” and indeed we see a blue-bar peak at that value; the correct answer for the second-order puzzle is “September, 14” and indeed we see an orange-bar peak at that value; the correct answer for the third-order puzzle is “September, 15” and indeed we see

a green-bar peak at that value - although a noticeable number of answers for the third-order puzzles seems to be “Multiple solutions”. Interestingly, this is not the case for the fourth-order puzzle: The correct answer for the fourth-order puzzle is “May, 18” and there is no red-bar peak at that value - instead the red-bar peak is at “June, 15”.

	Epistemic	Cut 1-lr	Cut 1-rl	Cut 2-lr
1st-order	82.98	82.98	82.98	82.98
2nd-order	54.74	54.74	54.74	54.74
3rd-order	36.84	7.37	7.37	36.84
4th-order	2.06	25.78	13.4	7.22

Table 5.2.1: Percentage (%) of participant answers that match the answers given by each model for each puzzle. Columns show the models and rows show the puzzle types. Bolded values mark the models that match the highest percentage of participant answers for each puzzle. As a reminder, Cut 1-lr and Cut 1-rl answer *No solution* for the third-order puzzle; Cut 1-lr answers *June, 15* for the fourth-order puzzle; Cut 1-rl answers *June, 14* for the fourth-order puzzle; Cut 2-lr answers *No solution* for the fourth-order puzzle.

Table 5.2.1 shows the percentage of the participant data that matches a model’s answers. For example, the epistemic model answers “May, 15” for the first-order puzzle, so the epistemic model matches $\frac{78}{94} \times 100 \approx 82.98$, where 94 is the total number of trials associated with first-order puzzles (approximately 49 participants \times 2 trials, as not all participants finished all eight trials within the time limit). Note that, if multiple models are associated with the same percentage for a puzzle, it means that they give the same answer (see Table 4.3.1).

As expected, the epistemic model matches a large percentage of the answers for the first-order, second-order and third-order puzzles, which is consistent with the peaks observed in Figure 5.2.1. Something interesting occurs for the fourth-order puzzles: The epistemic model matches a low percentage of the participants’ answers (only 2.06%), while the models using a cutting strategy match 46.4% of the answers and the cut 1-lr model matches 25.78% of the answers. This is again consistent with Figure 5.2.1, where red-bar peaks can be noticed at “June, 14” and “June, 15”. In the

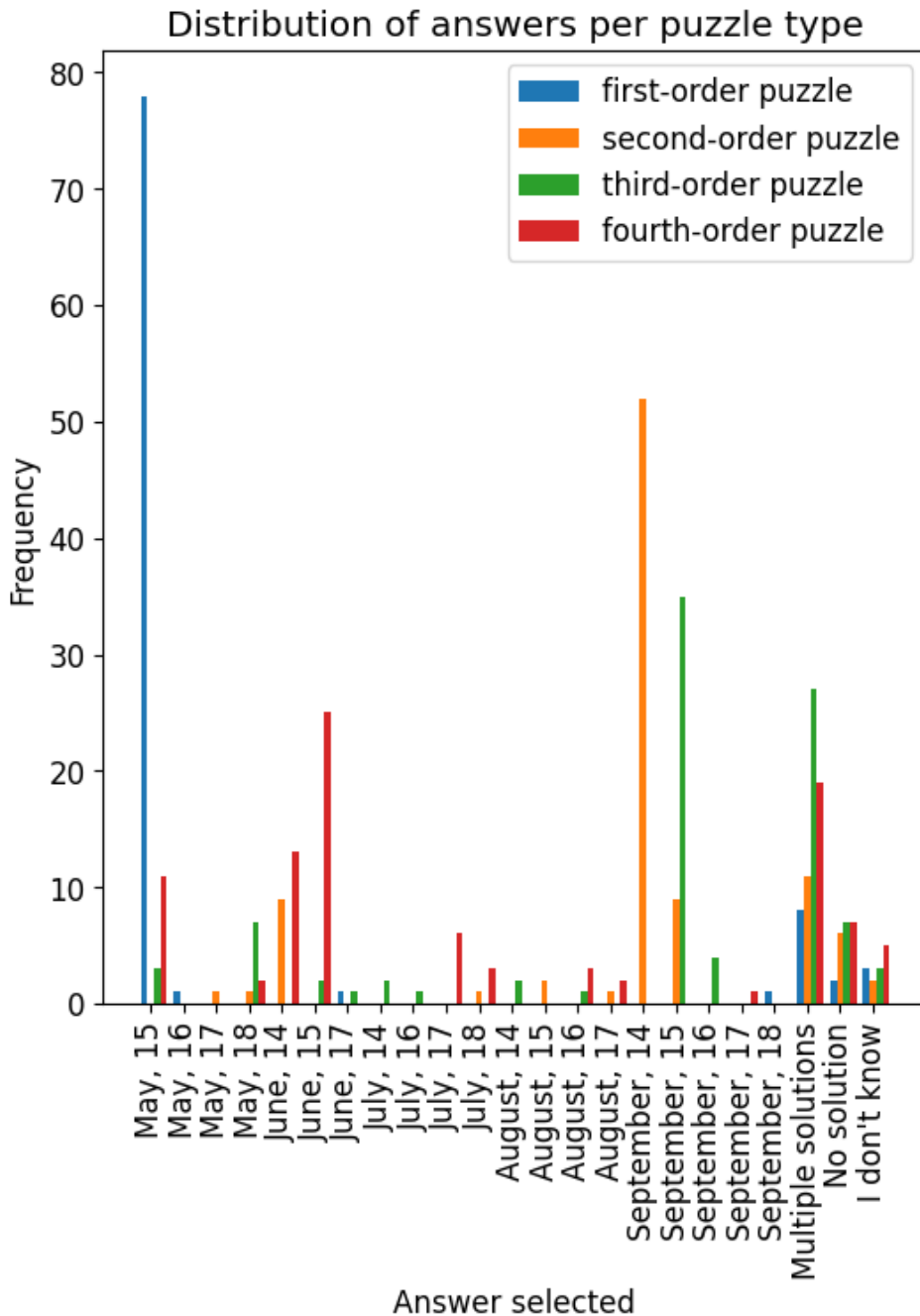


Figure 5.2.1: Distribution of answers for all forty-nine participants for each puzzle ($N = 381$ trials). The puzzles were aggregated over scenario and configuration by “translating” back to the original four ToM puzzles described in Section 2.2. The x-axis shows all answers selected by the participants and the y-axis shows the number of times that answer has been selected throughout all experiments. The bars indicate how many times a date was selected as an answer to a first-order, second-order, third-order and fourth-order puzzle, in this order. As a reminder, the correct answer for the first-order puzzle was *May, 15*; for the second-order puzzle *September, 14*; for the third-order puzzle *September, 15*; and for the fourth-order puzzle *May, 18*.

following sections, I analyze the models' goodness of fit for the four puzzles more formally.

Since the random model provides "random" answers (fitted to all participants' answers), computing the percentage of the random model's answers that match the participant data is equivalent to averaging over a large number of runs. This approach is qualitatively different from computing the exact match percentage for the other models and, for this reason, the random model was excluded from Table 4.3.1. However, it is possible to approximate this percentage of matching data by averaging over the coherence values of the random model with regards to all participants, for each subset of data associated with each puzzle type (and multiplied by 100 to get a *percentage*). Hence, the random model matches, on average, 13.39% of the answers given to the first-order puzzles; 13.34% of the answers given to the second-order puzzles; 13.36% of the answers given to the third-order puzzles; and 13.39% of the answers given to the fourth-order puzzles. Note that these values are not $\frac{1}{13} \cdot 100 \approx 7.7\%$ (as would be the case for a truly random model), because there was not an equal distribution over all thirteen possible answers within the sample of participants.

5.3 Performance Metrics

RFX-BMS, first introduced by Stephan et al. [2009], was used to estimate the frequency of the different ToM strategies within the population that participants were drawn from (see Chapter 4 for more information on the performance metrics). However, RFX-BMS does not quantify the goodness of fit of the strategies to the data. Take two arbitrary reasoning strategies A and B and suppose that strategy A explains more of the data than strategy B: RFX-BMS says nothing about *how much* of the data is explained by strategy A. Therefore, in this section, I also analyze the distribution of *coherence* between a strategy and the participant data, which is defined as the ratio of model predictions that exactly match a participant's answers. Intuitively, the larger the coherence, the better the fit of the strategy used by a model to the participant's data - a coherence value of one means that the model perfectly predicts all the answers of the participant. It is important to know that, for a

participant, it is possible to have multiple different strategies with the same "best" coherence - in the end, the coherence is computed based on at most eight decision points (i.e., trials) per participant.

In the following section, I will show the following plots:

- a) *Bar plot of proportion of model fit.* The RFX-BMS algorithm computes the relative frequency of the strategies used by the models, in the population that participants were drawn from. One bar is then plotted per model.
- b) *The coherence distribution for all non-random models.* For each participant, the coherence values for all models are computed as explained in Section 4.5. Then, for each participant, I find the best coherence values. If this "best" coherence value was generated by at least one non-random model, then I store the coherence value in a list. After iterating through all participants, I display this list through a violin plot. Additionally, I plot the coherence values for the participants whose "best" coherence value was generated by the random model (and no other model) as jittered scatter points.
- c) *The coherence distribution for the best model.* For each participant, the coherence values for all models are computed as explained in Section 4.5. Then, for each participant, I find the best coherence values. If this "best" coherence value was generated (among, possibly, other models) by the model with the highest frequency, as deemed by the RFX-BMS algorithm, then I store the coherence value in a list. After iterating through all participants, I display this list through a violin plot. If the model with the highest frequency is not the random model, then this coherence list is a subset of the coherence list in (b); in this case, I also plot the same jittered scatter points for the random model. This plot will qualitatively show whether the best non-random model is (much) better at explaining the data, compared to the other non-random models.

5.4 Results

In this section, I investigate the following:

1. *Which of the five models described in Section 5.1 fits the highest proportion of the data?* This should provide a general intuition regarding the viability of the cutting strategy: The cutting strategy should fit more of the data than the baseline models, namely the epistemic and the random models. The results are discussed in Section 5.4.1.
2. *Which of the five models described in Section 5.1 fits the highest proportion of the correct answers?* This analysis is run on a subset of the data discussed in (1) and is used as a validation method: Since the epistemic model always answers correctly, then it should fit 100% of the correct answers. However, note that RFX-BMS will most likely allocate some small proportions of the data to the other models that *sometimes* answer correctly. The results are discussed in Section 5.4.2.
3. *Which of the five models described in Section 5.1 fits the highest proportion of the wrong answers?* This analysis is run on a subset of the data discussed in (1) and it should reveal whether a cognitive limit to the recursive use of ToM can be found in this dataset: If the cutting strategy fits a high proportion of the data, then this supports the hypothesis that participants cannot process statements beyond that ToM order (either first or second order). The results are discussed in Section 5.4.3.
4. *Which cutting direction fits the highest proportion of the third-order and fourth-order puzzle answers?* It is interesting to consider whether removing knowledge operators either from the left side or from the right side of a statement is more consistent with the participants' behaviour. To this end, I compare the Cut 1-lr and Cut 1-rl models against the baseline models (i.e., the epistemic and the random models) for third and fourth-order puzzles; the first and second-order puzzles were excluded from the analysis because, in these cases, the Cut 1 and epistemic models give the same answers. The results are discussed in Section 5.4.4.

5. *Which cutting direction fits the highest proportion of the third-order puzzle answers?* This is a more in-depth analysis of the results shown in (4): I use the same setup described in (4) but I restrict the data to only third-order puzzle answers. The results are discussed in Section 5.4.5.
6. *Which cutting direction fits the highest proportion of the fourth-order puzzle answers?* This is a more in-depth analysis of the results shown in (4): I use the same setup described in (4) but I restrict the data to only fourth-order puzzle answers. The results are discussed in Section 5.4.6.

For all settings described above, I show (a selection of) the plots described in Section 5.3.

5.4.1 Which of the five models described in Section 5.1 fits the highest proportion of the data?

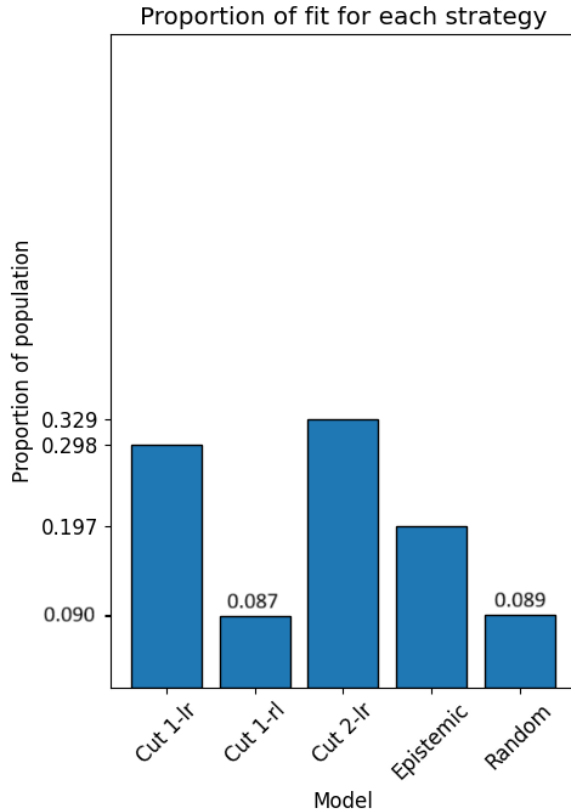


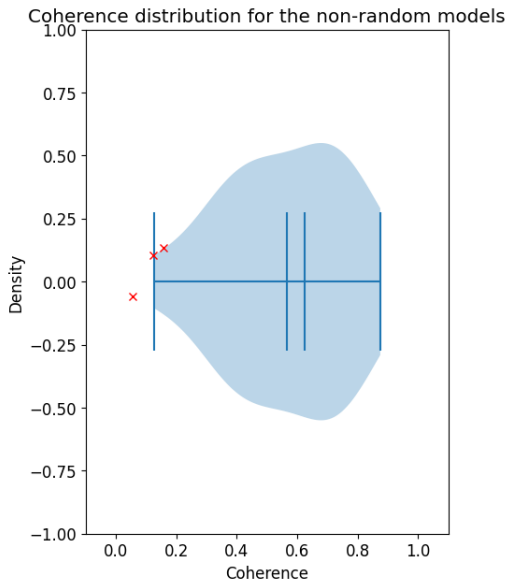
Figure 5.4.1: Results for all models over all the participant data ($N = 381$): proportion of fit to the participant population for each strategy, according to the RFX-BMS algorithm. Each bar corresponds to one model. The y-axis shows the proportion of participants that was assigned to the strategy associated with a model.

This section discusses the performance of the models on all the data (both correct and wrong answers) for all forty-nine participants and for all puzzles. Figure 5.4.1 shows that model Cut 2-lr performs best, as it is assigned to almost 33% of the population, followed closely by model Cut 1-lr. The random and the Cut 1-rl models are assigned to a small proportion of the data. This suggests that Cut 1-lr and Cut 2-lr capture some systematic patterns in the behaviour of the participants. Additionally,

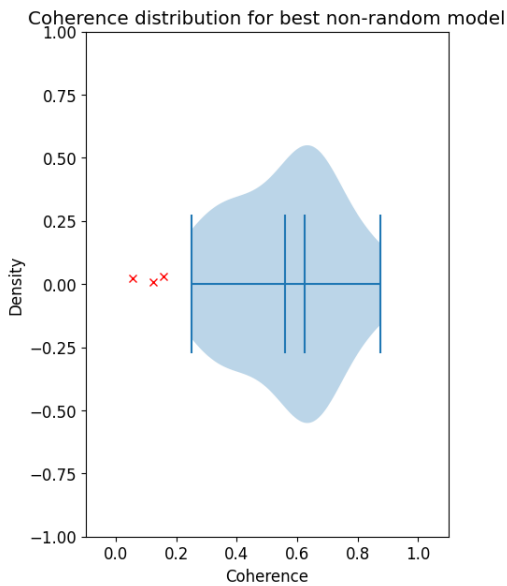
since it is assigned to a higher proportion of the population than the epistemic model, the Cut 2-lr model seems to describe consistent deviations from the epistemic strategy across participants.

Figure 5.4.2a shows a violin plot of the distribution of coherence values for the non-random models. The red crosses mark the coherence values associated with those participants that match the strategy of the random model the best. As can be seen, only three participants were classified as exhibiting behavior most similar to the random model and the associated coherence values are not larger than 0.2; this means that, on average, the random model correctly predicts no more than two out of eight answers per participant assigned to this model. However, the non-random models are associated with coherence values in the range $[0.125, 0.875]$, with mean = 0.567 and median = 0.625; this means that, on average, the non-random models correctly predict more than half the answers per participant assigned to these models.

Figure 5.4.2b shows a violin plot of the distribution of coherence values for the model that fits the data best according to the RFX-BMS algorithm (Cut 2-lr, in this case). As before, the red crosses mark the coherence values associated with those participants that match the strategy of the random model the best. The Cut 2-lr model is associated with coherence values in the range $[0.25, 0.875]$, with mean = 0.560 and median = 0.625; this means that, on average, the Cut 2-lr model correctly predicts more than half the answers per participant assigned to this model.



(a) Violin plot of coherence distribution for all non-random models.



(b) Violin plot of coherence distribution for the best model (Cut 2-lr).

Figure 5.4.2: Results for all models over all the participant data ($N = 381$): distributions of coherence. The left-most and right-most vertical lines mark the extremes and the middle vertical lines mark the mean and median, respectively. The red crosses, marking the participants found to best fit the strategy of the random model, were jittered over the y-axis for readability - thus, the y-axis has no meaning here.

5.4.2 Which of the five models described in Section 5.1 fits the highest proportion of the correct answers?

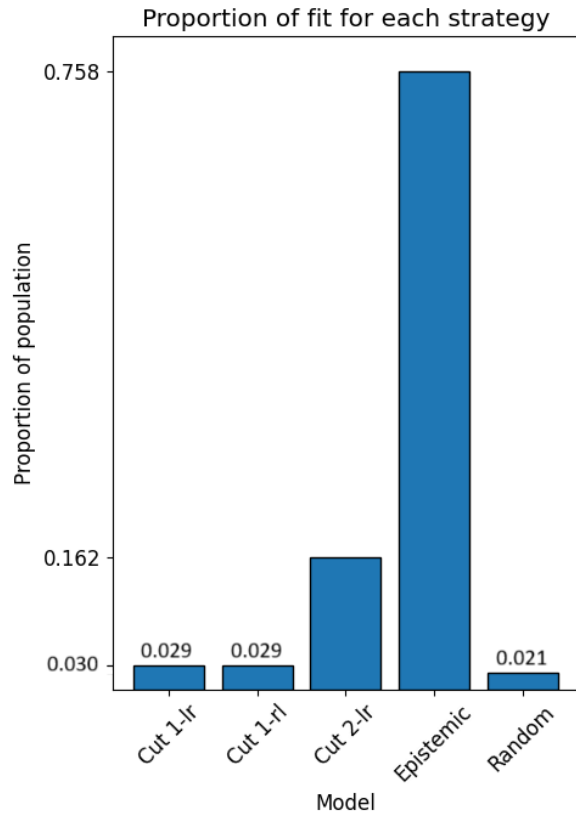


Figure 5.4.3: Results for all models over the correct answers ($N = 167$): proportion of fit to the participant population for each strategy, according to the RFX-BMS algorithm. Each bar corresponds to one model. The y-axis shows the proportion of participants that was assigned to a strategy associated with a model.

This section discusses the performance of the models on the correct answers for all forty-nine participants and for all puzzles. Figure 5.4.3 shows that the epistemic model performs best, as it is assigned to around 76% of the population - this can easily be explained, as the epistemic model always answers all puzzles *correctly*. All other models explain a negligibly small proportion of the data, except model Cut 2-lr. This latter model always answers correctly

all first, second and third-order puzzles - note that, since there are only two correct answers to the fourth-order puzzles, the Cut 2-lr model matches the vast majority of the correct answers. Trivially, the coherence values associated with the epistemic model are always one.

5.4.3 Which of the five models described in Section 5.1 fits the highest proportion of the wrong answers?

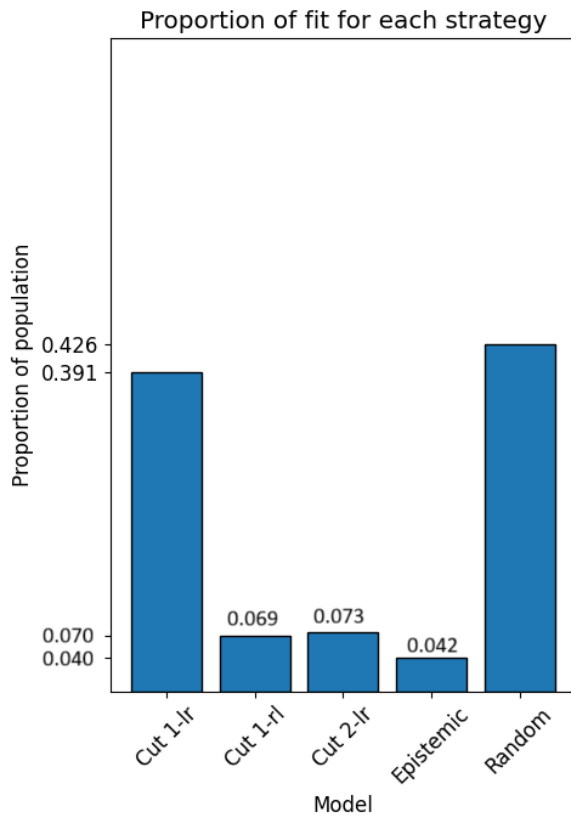
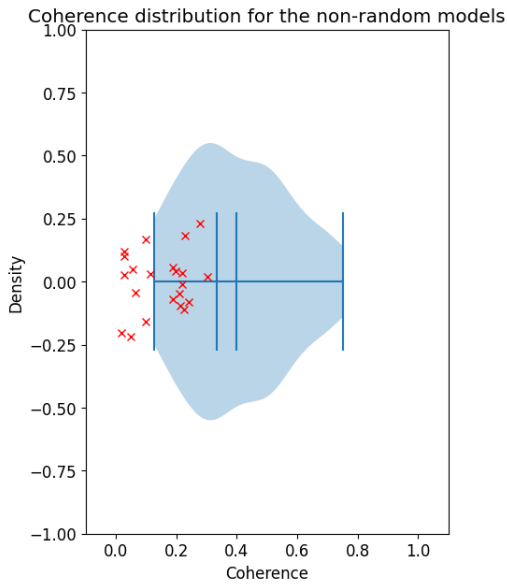


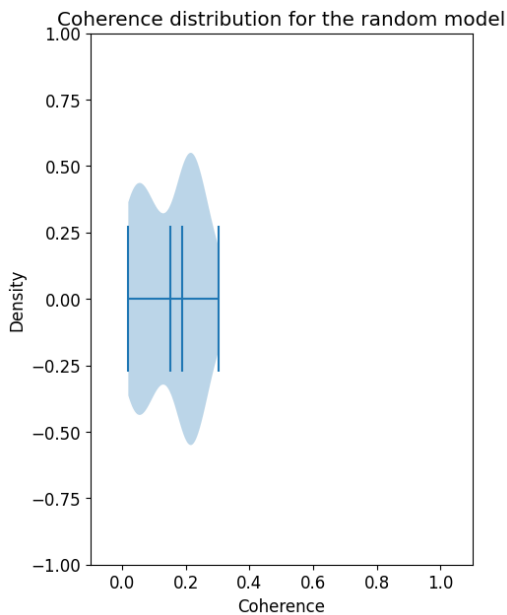
Figure 5.4.4: Results for all models over the wrong answers ($N = 214$): proportion of fit to the participant population for each strategy, according to the RFX-BMS algorithm. Each bar corresponds to one model. The y-axis shows the proportion of participants that was assigned to the strategy associated with a model.

This section discusses the performance of the models on the wrong answers for all forty-nine participants and for all puzzles. Figure 5.4.4 shows that the random model performs best, as it is assigned to almost 43% of the population, followed closely by the Cut 1-lr model. All other models are assigned to a small proportion of the data. Overall, this suggests that the cutting strategy may account for systematic errors made by part of the population.

Figure 5.4.5a shows a violin plot of the distribution of coherence values for the non-random models. The red crosses mark the coherence values associated with those participants that match the strategy of the random model the best. As can be seen, twenty-two out of forty-nine participants were classified as exhibiting behavior most similar to the random model and the associated coherence values are in the range $[0.019, 0.303]$, with mean = 0.149 and median = 0.188 (see also Figure 5.4.5b); this means that, on average, the random model correctly predicts no more than two out of eight answers per participant assigned to this model. However, the non-random models are associated with coherence values in the range $[0.125, 0.75]$, with mean = 0.399 and median = 0.333; this means that, on average, the non-random models correctly predict around three out of eight answers per participant assigned to these models.



(a) Violin plot of coherence distribution for all non-random models.



(b) Violin plot of coherence distribution for the best model (random model).

Figure 5.4.5: Results for all models over the wrong answers ($N = 214$): distributions of coherence. The left-most and right-most vertical lines mark the extremes and the middle vertical lines mark the mean and median, respectively. The red crosses, marking the participants found to best fit the strategy of the random model, were jittered over the y-axis for readability - thus, the y-axis has no meaning here.

5.4.4 Which cutting direction fits the highest proportion of the third-order and fourth-order puzzle answers?

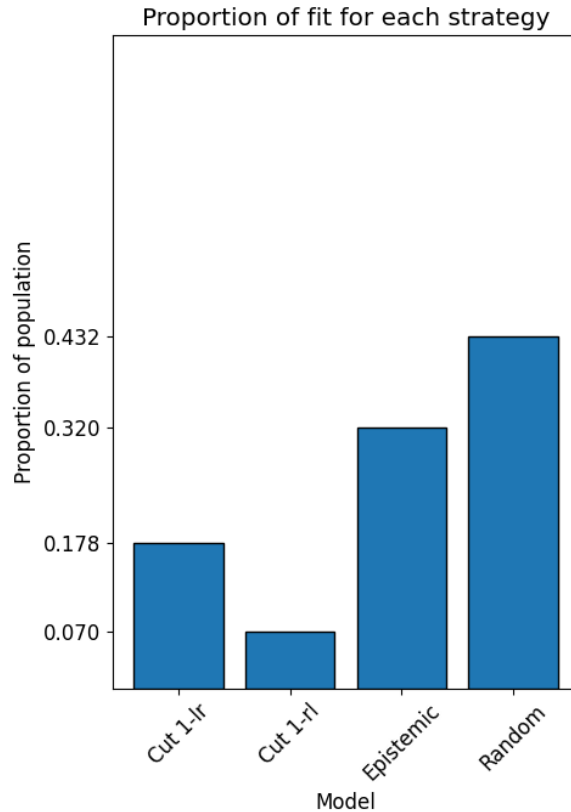


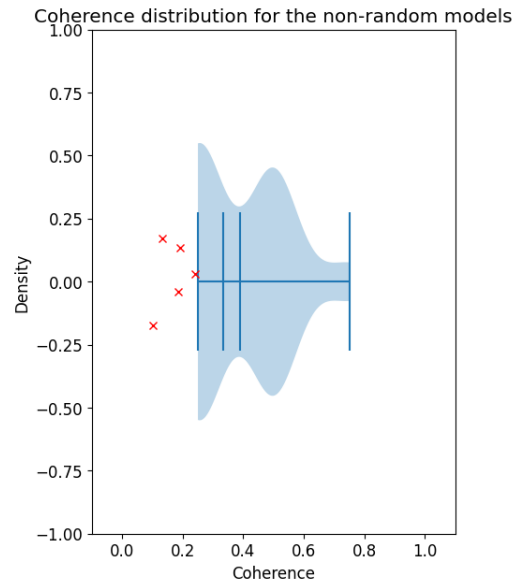
Figure 5.4.6: Results for the Cut 1-lr and Cut 1-rl models, epistemic model, and random model over all the answers for the third and fourth-order puzzles ($N = 192$): proportion of fit to the participant population for each strategy, according to the RFX-BMS algorithm. Each bar corresponds to one model. The y-axis shows the proportion of participants that was assigned to the strategy associated with a model.

This section discusses the performance of the Cut 1-lr, Cut 1-rl, epistemic and random models on all the answers for all forty-nine participants and for the third and fourth-order puzzles. Figure 5.4.6 shows that the random model performs best, as it is assigned to around 43% of the population, followed closely by the epistemic model. Interestingly, the

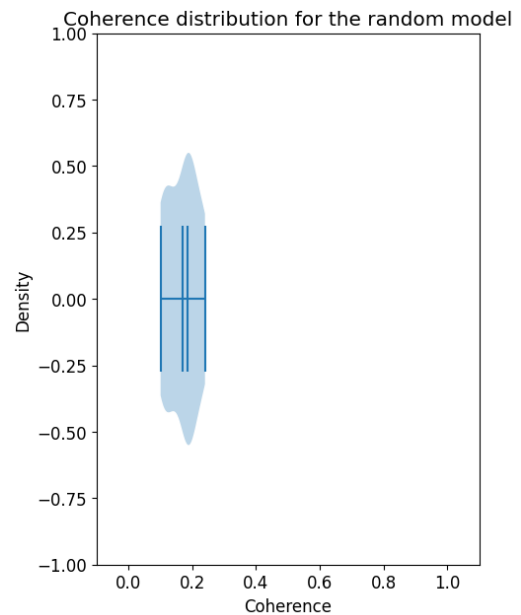
Cut 1-lr model is assigned to a higher proportion of the data than the Cut 1-rl model. This suggests that cutting from left to right is more consistent with the behaviour of the participants.

Figure 5.4.7a shows a violin plot of the distribution of coherence values for the non-random models. The red crosses mark the coherence values associated with those participants that match the strategy of the random model the best. As can be seen, five out of forty-nine participants were classified as exhibiting behavior most similar to the random model and the associated coherence values are in the range $[0.01, 0.240]$, with mean = 0.169 and median = 0.184 (see also Figure 5.4.7b); on average, the random model correctly predicts no more than two out of eight answers per participant assigned to this model. However, the non-random models are associated with coherence values in the range $[0.25, 0.75]$, with mean = 0.39 and median = 0.333; on average, the non-random models correctly predict around three out of eight answers per participant assigned to these models.

It is worth reminding the reader that the coherence violin plots for the non-random models include participants that were assigned to *at least* one of the non-random models - this explains why the random model was assigned to such a large proportion of the data even though only five participants (red crosses in Figure 5.4.7a) were uniquely assigned to the random model.



(a) Violin plot of coherence distribution for all non-random models.



(b) Violin plot of coherence distribution for the best model (random model).

Figure 5.4.7: Results for the Cut 1-lr and Cut 1-rl models, epistemic model, and random model over all the answers for the third and fourth-order puzzles ($N = 192$): distributions of coherence. The left-most and right-most vertical lines mark the extremes and the middle vertical lines mark the mean and median, respectively. The red crosses, marking the participants found to best fit the strategy of the random model, were jittered over the y-axis for readability - thus, the y-axis has no meaning here.

5.4.5 Which cutting direction fits the highest proportion of the third-order puzzle answers?

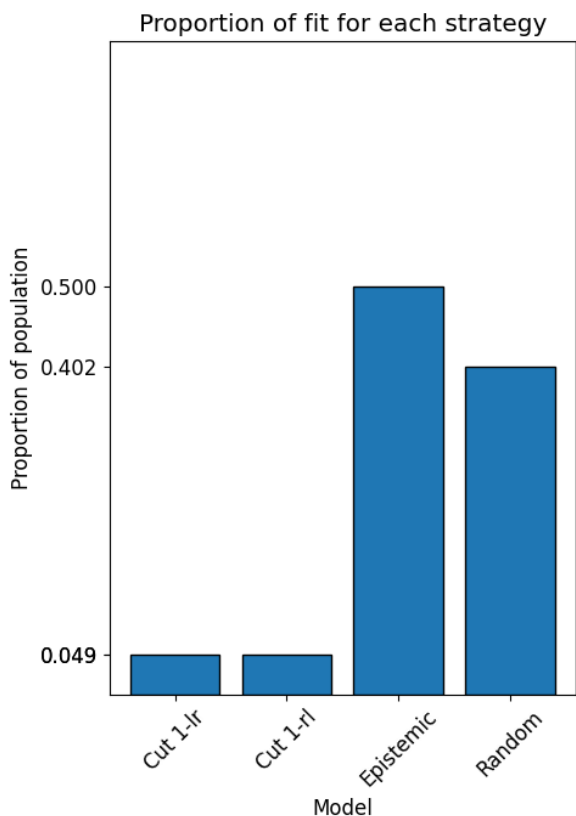


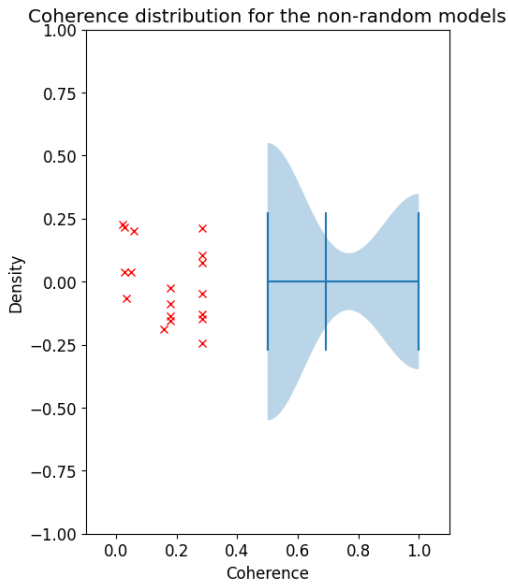
Figure 5.4.8: Results for the Cut 1-lr and Cut 1-rl models, epistemic model, and random model over all the answers for the third-order puzzles ($N = 95$): proportion of fit to the participant population for each strategy, according to the RFX-BMS algorithm. Each bar corresponds to one model. The y-axis shows the proportion of participants that was assigned to the strategy associated with a model.

This section discusses the performance of the Cut 1-lr, Cut 1-rl, epistemic and random models on all the answers for all forty-nine participants and for the third-order puzzles. Figure 5.4.8 shows that the epistemic model performs best, as it is assigned to 50% of the population, followed closely by the random model. The fact that now the epistemic model performs better than the random model compared to the results in Figure 5.4.6 is easily explain-

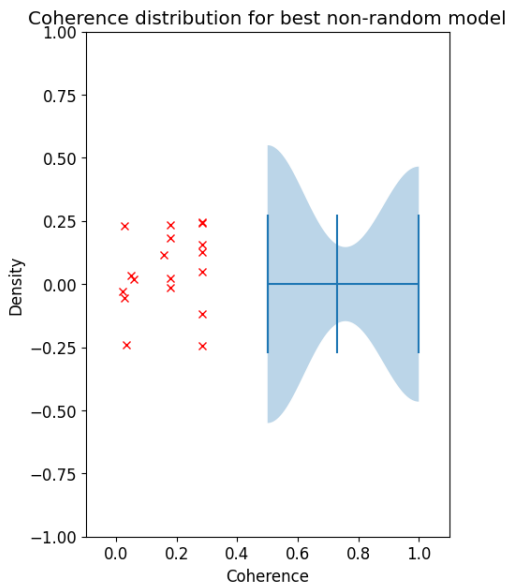
able: The third-order puzzles were associated with a higher proportion of correct answers than the fourth-order puzzles (or both combined). The Cut 1-lr and Cut 1-rl models explain an equally low proportion of the data - note that both models give the same answer (namely, “No solution”) for the third-order puzzles.

Figure 5.4.9a shows a violin plot of the distribution of coherence values for the non-random models. The red crosses mark the coherence values associated with those participants that match the strategy of the random model the best. As can be seen, eighteen out of forty-nine participants were classified as exhibiting behavior most similar to the random model and the associated coherence values are in the range $[0.021, 0.284]$, with mean = 0.171 and median = 0.179; this means that, on average, the random model correctly predicts no more than two out of eight answers per participant assigned to this model. However, the non-random models are associated with coherence values in the range $[0.5, 1]$, with mean = 0.693 and median = 0.5; this means that, on average, the non-random models correctly predict around five out of eight answers per participant assigned to these models.

Figure 5.4.9b shows a violin plot of the distribution of coherence values for the model that fits the data best according to the RFX-BMS algorithm (the epistemic model, in this case). As before, the red crosses mark the coherence values associated with those participants that match the strategy of the random model the best. The Cut 2-lr model is associated with coherence in the range $[0.5, 1]$ (mean = 0.729 and median = 0.5); this means that, on average, the epistemic model correctly predicts around six out of eight answers per participant assigned to this model.



(a) Violin plot of coherence distribution for all non-random models.



(b) Violin plot of coherence distribution for the best model (epistemic model).

Figure 5.4.9: Results for the Cut 1-lr and Cut 1-rl models, epistemic model, and random model over all the answers for the third-order puzzles ($N = 95$): distributions of coherence. The left-most and right-most vertical lines mark the extremes and the middle vertical lines mark the mean and median, respectively. The red crosses, marking the participants found to best fit the strategy of the random model, were jittered over the y-axis for readability - thus, the y-axis has no meaning here.

5.4.6 Which cutting direction fits the highest proportion of the fourth-order puzzle answers?

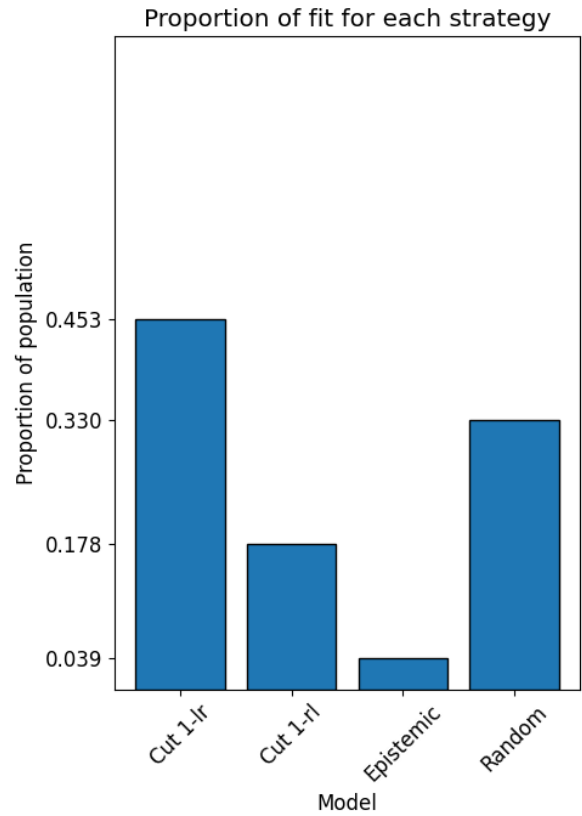


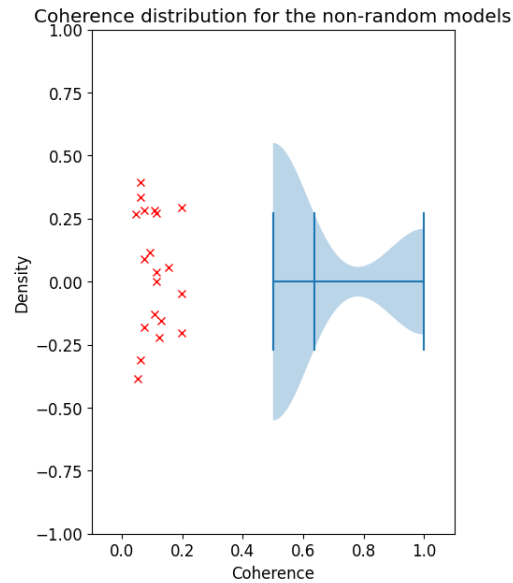
Figure 5.4.10: Results for the Cut 1-lr and Cut 1-rl models, epistemic model, and random model over all the answers for the fourth-order puzzles ($N = 97$): proportion of fit to the participant population for each strategy, according to the RFX-BMS algorithm. Each bar corresponds to one model. The y-axis shows the proportion of participants that was assigned to the strategy associated with a model.

This section discusses the performance of the Cut 1-lr, Cut 1-rl, epistemic and random models on all the answers for all forty-nine participants and for the fourth-order puzzles. Figure 5.4.10 shows that the Cut 1-lr model performs best, as it is assigned to around 45% of the population. Interestingly, the Cut 1-lr model is assigned to a noticeably higher proportion of the data than the Cut 1-rl model, which suggests that cutting from left to right is

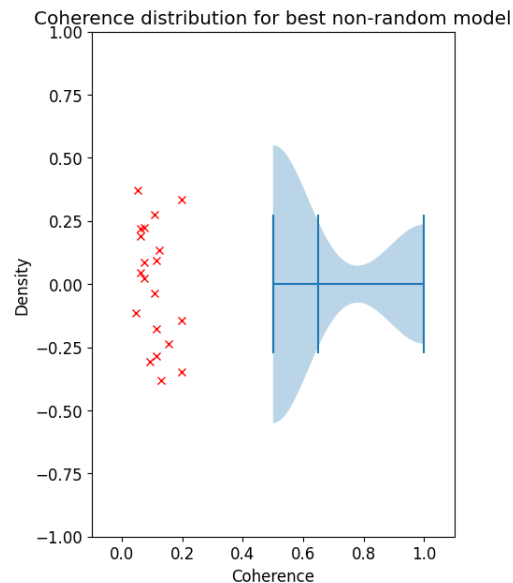
more consistent with the behaviour of the participants.

Figure 5.4.11a shows a violin plot of the distribution of coherence values for the non-random models. The red crosses mark the coherence values associated with those participants that match the strategy of the random model the best. As can be seen, twenty out of forty-nine participants were classified as exhibiting behavior most similar to the random model and the associated coherence values are in the range $[0.046, 0.196]$, with mean = 0.107 and median = 0.108; this means that, on average, the random model correctly predicts no more than two out of eight answers per participant assigned to this model. However, the non-random models are associated with coherence values in the range $[0.5, 1]$, with mean = 0.638 and median = 0.5; this means that, on average, the non-random models correctly predict around six out of eight answers per participant assigned to these models.

Figure 5.4.11b shows a violin plot of the distribution of coherence values for the model that fits the data best according to the RFX-BMS algorithm (the Cut 1-lr model, in this case). As before, the red crosses mark the coherence values associated with those participants that match the strategy of the random model the best. The Cut 1-lr model is associated with coherence values in the range $[0.5, 1]$, with mean = 0.65 and median = 0.5; this means that, on average, the epistemic model correctly predicts around six out of eight answers per participant assigned to this model.



(a) Violin plot of coherence distribution for all non-random models.



(b) Violin plot of coherence distribution for the best model (Cut 1-lr model).

Figure 5.4.11: Results for the Cut 1-lr and Cut 1-rl models, epistemic model, and random model over all the answers for the fourth-order puzzles ($N = 97$): distributions of coherence. The left-most and right-most vertical lines mark the extremes and the middle vertical lines mark the mean and median, respectively. The red crosses, marking the participants found to best fit the strategy of the random model, were jittered over the y-axis for readability - thus, the y-axis has no meaning here.

5.5 Conclusion

In this chapter, I investigated the goodness of fit of the strategies proposed in Chapter 4 on (subsets of) the participant data.

Firstly, according to the RFX-BMS algorithm, around 71% of all the participant data was assigned to the cutting strategy and around 33% of all the participant data was assigned to the Cut 2-lr model. Moreover, the predictions of the Cut 2-lr model match the participants' answers for more than half of the trials, on average. This suggests that there is merit to further investigating the cutting-operator strategy: It is assigned to a higher proportion of the data and has higher coherence than the baseline models (i.e., an epistemic model of perfect reasoning and a random model fit to the sample of participants). As expected, when testing the goodness of fit on only the correct answers, the epistemic model was found to fit the data best, because this model always answers all puzzles correctly. Lastly, when testing the goodness of fit on only the wrong answers, the random model was found to fit around 43% and the Cut 1-lr model almost 40% of the participant data, according to the RFX-BMS algorithm. This suggests that the mistakes made by participants can be attributed to incorrectly reducing the ToM statements down to a manageable order of complexity, as initially hypothesized. However, it is worth noting that the relatively high fit of the random model and the associated low coherence values may indicate that there are underlying mechanisms that the cutting strategy cannot explain and that cannot be attributed to (completely) random behaviour.

Secondly, I investigated whether the direction of removing knowledge operators for the cutting models has an impact on the goodness of fit. I compared the Cut 1-lr and Cut 1-rl models against each other (and against the epistemic and random models) on all data associated with the third and fourth-order puzzles. For the third-order puzzles, the Cut 1-lr model and the Cut 1-rl model give the same answers and, therefore, explain the same (small) proportion of the population. However, for the fourth-order puzzles, the Cut 1-lr model explains noticeably more data than the Cut 1-rl model - around 45% the data. Therefore, this suggests that cutting from left to right is more consistent with the participants' behaviour than cutting from right to left.

In this chapter, I reflect upon how the findings of the study relate to past and future research. In Section 6.1, I summarize the findings of the study and answer the research questions posed in Section 1.3. In Section 6.2, I discuss the strengths and weaknesses of the study from three perspectives: experimental design, modeling approach and statistical methodologies. In Section 6.3, I propose avenues for future research. Lastly, in Section 6.4, I draw the final conclusions.

6.1 Summary of Study

Theory of mind (ToM) is defined as the ability to reason about the behaviour of others and oneself by attributing mental states, such as beliefs, desires and knowledge [Dennett, 1971, Premack and Woodruff, 1978]. While humans are able to apply ToM recursively (e.g., “I know that you believe that they think...”), past research has shown that human recursive ToM use is limited in strategic games and that it often does not exceed second-order ToM [de Weerd et al., 2018, Devaine et al., 2014, Nagel, 1995]. However, ToM abilities seem to be highly dependent on the task domain [Flobbe et al., 2008]. For example, in story comprehension tasks, humans have been shown to score above chance on questions that require up to fourth-order ToM reasoning [Stiller and Dunbar, 2007, Kinderman et al., 1998].

One possible explanation [Arslan et al., 2017b, Arslan et al., 2017a, Verbrugge, 2009] for this limitation of recursive ToM use is that embedded beliefs are processed serially through intermediate reasoning steps that are eventually sent to the long-

term memory for later retrieval and retrieval from the long-term memory has been shown to take long and to be prone to errors [Anderson and Schooler, 2000]. This phenomenon is also known in the literature as the *serial processing bottleneck* [see e.g., Borst, 2012, Borst et al., 2010]. This hypothesis is further supported by studies showing that there is a correlation between ToM and working memory capacity in various tasks [Laillier et al., 2019, Lin et al., 2010, Mutter et al., 2006].

Formally measuring ToM limitations in humans is difficult due to one apparent shortcoming in the literature: Classical epistemic logic research does not usually account for the upper-bound limit found through behavioural research and behavioural research does not often test its predictions on tasks that can be easily modeled in epistemic logic, such as epistemic puzzles. Moreover, the few behavioural experiments on epistemic puzzles [Cedegao et al., 2021, Jonker and Treur, 2003, Hayashi, 2002] have been criticized for their experimental designs [e.g., Top et al., 2023], as it is often unclear whether participants use theory of mind reasoning or an alternative strategy to solve the puzzles. Thus, the aim of the current study was to bridge this gap, by proposing and conducting a novel experimental design: Participants were asked to solve a series of epistemic puzzles that were set in different contexts (hereon, referred to as “scenarios”) and required different orders of ToM reasoning (specifically, first-order through fourth-order ToM reasoning) as the only viable strategy to reach the correct solution (see Chapter 2 for more information on the experimental design). Additionally, I

showed that: i) theories of ToM reasoning limitations established through past research can be observed in this new “Cheryl’s Puzzle” dataset and ii) underlying mechanisms of ToM reasoning present in the dataset can be modeled in epistemic logic.

Firstly, I showed that ToM reasoning is limited by investigating the effects of ToM order and scenario on the time needed to solve a puzzle and on accuracy, respectively (see Chapter 3 for more information). A Kruskal-Wallis test and a post-hoc Dunn test revealed that the time to solve a puzzle differed significantly across every two ToM orders, except for third-order and fourth-order puzzles (first-order: $M=130.91$ seconds; second-order: $M=194.01$ seconds; third-order: $M=258.02$ seconds; fourth-order: $M=261.69$ seconds). A Chi-square test revealed that accuracy was significantly different across ToM orders (first-order: 82.1% correct answers; second-order: 51.2% correct answers; third-order: 34.5% correct answers; fourth-order: 2.4% correct answers). Overall, lower orders of ToM are associated with lower solving time and higher accuracy (and vice-versa for higher order of ToM) - this suggests that participants have trouble correctly applying higher order ToM reasoning and supports past research findings of a limitation to the human recursive ToM use. A Kruskal-Wallis test and a Chi-square test revealed that there is no effect of scenario on solving time and accuracy, respectively. This suggests that the difference in performance over the different ToM orders is more likely due to limitations in ToM recursive, rather than the contextual information of the puzzles.

Secondly, I hypothesized about a possible underlying ToM mechanism supported by past research [e.g., Arslan et al., 2017b, Arslan et al., 2017a]: When participants encounter a high-order ToM statement that they cannot process, instead of ignoring that statement (as proposed in Top et al. [2023]), they sequentially remove knowledge operators, either from the front or from the back of the statement, until the ToM order is low enough to process the statement. This cutting model was compared against two baseline models: i) an epistemic model of perfect reasoning, implemented in a variant of dynamic epistemic logic [van Ditmarsch et al., 2007] called public announcement logic [Plaza, 1989, 2007], and ii) a “random” model fit to the answers of the sample of participants. I showed that the data of the “Cheryl’s Puzzle”

dataset is best explained by the cutting model that reduces all statements of the puzzles to second-order ToM statements (Cut 2-lr) and that, on average, five to six out of eight answers are explained by this model for each participant assigned to this model. Additionally, I showed that the mistakes of the “Cheryl’s Puzzle” dataset are somewhat explained by the cutting model that reduces all statements of the puzzles to first-order ToM statements (Cut 1-lr) and that, on average, around three of eight answers are explained by this model for each participant assigned to this model. Overall, this suggests that there is merit to the cutting-operators hypothesis, because models using this strategy often out-performed both baseline models. However, it is worth noting that the relatively high fit of the random model and the associated low coherence values may indicate that there are underlying mechanisms that the cutting strategy cannot explain and that cannot be attributed to (completely) random behaviour. Lastly, I showed that removing operators from the front (left side) of a statement matches the behaviour of participants better than removing operators from the back (right side) of the statement, especially for the fourth-order puzzles. Note that this is in line with the findings of Arslan et al. [2017b] and Arslan et al. [2017a] and sheds a critical perspective on alternative explanations of reasoning about complex ToM statements (such as ignoring these statements, as proposed in Top et al. [2023]).

6.2 Critical Perspective

In this section, I reflect on the following research procedures:

- Experimental design

The “Cheryl’s Puzzle” experiment consisted of eight epistemic puzzles presented in two blocks of four puzzles each. Each puzzle had an associated ToM order and scenario - since there were four ToM orders and four scenarios, each ToM order and each scenario was presented to each participant exactly twice and exactly once per block, in a randomized order. The $\langle \text{scenario} \times \text{ToM order} \rangle$ configuration (i.e., mirroring of attributes) was also randomized but the same configuration could not occur in both

blocks (for example, if a third-order hair puzzle occurred in the first block, another third-order hair puzzle could not occur in the second block for the same participant). This $4 \text{ ToM orders} \times 4 \text{ scenarios} \times 4 \text{ configurations}$ design allows for some interesting investigations into recursive ToM use. Firstly, the four ToM orders design allows one to clearly distinguish between performance at different orders of recursive ToM use. Secondly, the four scenarios design addresses 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. Lastly, 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.

Participants were tasked with finding Cheryl’s birthday from a list of ten options, based on clues from a conversation between Cheryl’s friends, Albert, who knows the month of the birthday and Bernard, who knows the day of the birthday. In all conversations in all puzzles, Albert and Bernard simply state whether they know the birthday (or whether they know that the other person does or does not know the birthday, etc.) - no other relevant piece of information about Cheryl’s birthday was made available to the participant. This is one of the biggest strengths of the current study: Since the only clues about Cheryl’s birthday come from the conversation between Albert and Bernard and the conversation is purely epistemic in nature, this experimental design ensures that the participant *must* reason about others’ knowledge (hence, use ToM reasoning) in order to reach the correct answer. Additionally, since there are thirteen answer options for each puzzle, guessing the correct answer becomes an unfeasible strategy. This is in contrast to other epistemic puzzles, such as the *Aces and Eights* puzzle, where there are four answer options [Cedegao et al., 2021] or the *Wise Men* puzzle, where there are only two answer options [McCarthy, 1990].

It is also important to remark upon some design flaws. Firstly, the design of the two blocks was not implemented as intended due to a bug

in the original code: Only in the first block, each scenario was associated with one unique ToM order for all participants. While this did not influence the modeling approach (each participant still encountered each ToM order and each scenario exactly twice), it did hinder the statistical analysis of the effect of the scenario on solving time and accuracy: For each scenario, one ToM order was encountered around three times as often as the other three orders independently. Nonetheless, the effect of scenario on solving time and accuracy was found to be not significant, which further supports the idea that the bug did not have a strong effect on the study, in the end.

Secondly, it might be a good idea to remove “Multiple solutions” as a potential answer and instead allow the participants to select multiple dates from Cheryl’s list of options - this might be more informative for the analysis and modeling approach.

Thirdly, isolating the actual puzzle-solving time from the time to read the puzzle text (for example, by showing the possible answers separately from the puzzle text) might provide a more accurate estimation of the time it takes to solve puzzles associated with different ToM orders, as some participants might inherently take longer to read than others.

Lastly, it might be worth keeping track of the order in which the possible answers were shown in the drop-down menu in order to search for potential ordering effects - perhaps one participant always selected the second option shown in the drop-down menus so keeping track of the order would be the only way to reveal this behaviour.

- Modeling approach

The strategy proposed in this study can be summarized as follows: Given a maximum theory of mind order x that an agent can process correctly and a statement with an associated ToM order higher than x , then remove enough knowledge operators (either from the back or from the front of the statement) such that the statement now has x ToM order (see Chapter 4).

The strength of this proposed modeling approach lies in the fact that it is based on past research: Arslan et al. [2017b] and Arslan et al. [2017a] show that children who cannot process second-order ToM statements but can mostly process first-order ToM statements will approach second-order ToM statements using first-order ToM reasoning (as opposed to, for example, ignoring the statement, probabilistic guessing or using zero-order ToM reasoning).

While I showed that the models using this strategy explain a good proportion of the data, it is worth noting that the experimental design was a limiting factor, as most models give the same answer for the same puzzles. For this reason, it was not possible to formally compare the models that remove operators down to second-order ToM statements from the front (left side) and back (right side) respectively, because these two models give identical answers for all puzzles and, therefore, cannot be distinguished from each other. Similarly, the answers given by the Cut 1 models cannot be differentiated from the answers given by the epistemic model for the first-order and second-order puzzles, and similarly for the Cut 2 models and the epistemic model for all puzzles except the fourth-order puzzle. Nonetheless, a large part of the participant behaviour is consistent to the cutting strategy. Additionally, following the approaches in Cedegao et al. [2021] and Top et al. [2023], the fit of these models was compared against two baseline models of perfect reasoning and informed random behaviour. However, it is worth noting that, unlike the random models proposed by Cedegao et al. [2021], the random model proposed in this study has not stochasticity implemented but is rather fit to the answers given by the sample of participants, as also done in Top et al. [2023].

Another design flaw is that the models never answer “I don’t know”, even though, in principle, the participants had this option. On a similar note, it is worth noting that participants might have been hesitant to answer “No solution” or “Multiple solutions”, given that the puzzle text explicitly stated that Cheryl’s birthday was one of the dates on the list of

options. Again, note that the interpretation of “No solution” or “Multiple solutions” is that the information given in the puzzle text is insufficient to solve the puzzle rather than the fact that Cheryl’s birthday is none or a multiple of the dates specified, respectively - but this distinction may not have been clear to all participants. In contrast, the cutting models would often answer “No solution” to multiple puzzles (see Table 4.3.1).

- Statistical methodologies

Two statistical analyses were conducted. Firstly, I tested whether the ToM order and scenario have an effect on the time it takes to solve a puzzle and accuracy, respectively (see Chapter 3). To this end, I followed the appropriate methodology: I identified the appropriate statistical tests given the type of predictor and response variables, I investigated whether the assumptions of the selected statistical tests hold and I drew the appropriate conclusions. If the assumptions of the selected statistical tests did not hold (as was the case for the solving time analysis), I instead opted for an alternative non-parametric test. It is worth noting here that non-parametric tests generally do not hold as much statistical power as their parametric equivalent and therefore subjecting the data to a clever transformation or conducting another experiment with more participants might address this issue better.

Secondly, I investigated the goodness of fit of the proposed models to the “Cheryl’s Puzzle” dataset (see Chapter 5). I first used the RFX-BMS algorithm to estimate the frequency of all ToM strategies within the population that participants were drawn from (see Section 4.4). It is worth noting that RFX-BMS *must* assign each participant to one of the pre-defined models, irrespective of how well that model actually fits the data. Even though I introduced an epistemic model and a population-informed random model as baselines, it is possible that there are strategies that may explain the data even better. Additionally, the penalty term p used in the computation of the log-likelihood (see Equation 4.4.1) strongly affects the relative fit of the models as computed by the RFX-BMS algorithm. While there was clear

justification behind setting the penalty term to $p = \frac{1}{12}$ (i.e., the chance of selecting a wrong answer, where 12 is the number of wrong answers for all puzzles), it is worth noting that different values might lead to contradictory results.

Coherence was introduced as a metric to qualitatively assess how well a model explains the data (see Section 4.5). One limitation of this metric is that, due to there being a fixed number of possible answers, there will always be a baseline coherence value. This makes it difficult to compare performance on the *Cheryl's Birthday* puzzle against performance on other epistemic puzzles when the number of possible answers is different - for example, the puzzle of *Aces and Eights* investigated in Top et al. [2023] has four possible answers and, therefore, is associated with a higher baseline coherence value. A metric more robust to different numbers of answer options might be more desirable.

6.3 Future Research

In this section, I introduce various avenues of future research and divide them into areas of research.

6.3.1 A Cognitive Perspective

One aim of the study was to identify cognitive limitations to recursive theory of mind use in the “Cheryl’s Puzzle” dataset and to draw conclusions about underlying mechanisms involved in ToM reasoning. The latter is particularly difficult because past research has shown that ToM reasoning is task-dependent [Flobbe et al., 2008]. To this end, I introduced scenario as an experimental condition and showed that participants did not perform significantly differently when the puzzles were set in different contexts. However, it is not clear what effect different scenarios may have had on the participants’ reasoning ability and whether this experimental condition sufficiently accounts for task dependency - in the end, the overall structure of all puzzles was similar and participants reported skipping to the relevant parts of the puzzle (i.e., Cheryl’s list of options and the dialogue between Albert and Bernard) once they had gotten used

to the format. Additionally, I hypothesized that limitations in recursive ToM use may be due to working memory limitations and participants were encouraged to make use of pen and paper to relieve some of the workload placed on the working memory. However, a more thorough study is required to determine the connection between working memory workload and ToM reasoning abilities for the *Cheryl’s Birthday* puzzle - perhaps EEG/fMRI data would be particularly useful here.

It would be interesting to investigate whether participants’ performance can and does improve over time. For example, Verbrugge et al. [2018] showed that a carefully designed stepwise training regime, in which items are presented in increasing order of difficulty, leads to improved performance in second-order iterations 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, participants could be shown puzzles that gradually require higher orders of ToM to solve correctly.

With the latest developments in large language models (LLMs), it has been of interest whether LLMs have recursive ToM reasoning abilities [Strachan et al., 2024, Verma et al., 2024]. Testing an LLM’s recursive ToM reasoning abilities on an epistemic puzzle can prove problematic if that puzzle has been extensively researched before, because the LLM might have access to all information presented in those studies and this information could artificially enhance its ToM reasoning abilities. Since the current study introduces a novel experimental design, the epistemic puzzles proposed here can be safely used to investigate ToM abilities in LLMs.

6.3.2 A Modeling Perspective

A second aim of the study was to explain (some of) the mechanisms involved in ToM reasoning using public announcement logic (PAL) and I proposed a model that removes knowledge operators from the back or from the front of ToM statements that are too complex to understand otherwise. However, a more thorough analysis of the merit of the cutting strategy was partially hindered by the experimental design: Many of the models provided identical answers to many of the puzzles. Therefore, in a

follow-up study, the puzzles should be redesigned such that the answers of these models can be better distinguished from each other.

The cutting operation was introduced intuitively rather than through rigorous definitions that are consistent with logical formalism. Therefore, research is needed to determine how the cutting operation can be formalized within PAL or other forms of dynamic epistemic logic.

The modeling approach presented in this study was greatly inspired by the works of Cedegao et al. [2021] and Top et al. [2023]. In the two studies, the authors explain limitations in recursive ToM reasoning by restricting the initial Kripke graph: The lower the ToM order of an agent, the fewer the number of states it would have access to (in the sense of considering those states possible) and, therefore, the more limited the worldview of possible answers. Instead, I propose an alternative approach, whereby an agent has access to the entire initial Kripke graph but the public announcement is instead restricted. Thus, it would be interesting to investigate whether combining the two approaches could yield a more accurate model. On a similar note, it would be interesting to incorporate stochasticity into the model, following the Cedegao et al. [2021] approach.

It is worth noting that classical approaches to modeling human behaviour through logics of knowledge often suffer from making assumptions about an agent’s reasoning which are often unrealistic in real-life settings. For example, Parikh has argued that logical omniscience, the idea that an agent knows all logical consequences of her assumptions, is not a realistic account of human knowledge - for example, humans may not know some logical truths or may not be aware of some consequences of the things that they already know [Parikh, 1994, 1987]. Therefore, it would be interesting to augment formal models in epistemic logic with more realistic theories of knowledge according to cognitive research and investigate whether such an addition might provide a more complete explanation of the underlying mechanisms in the “Cheryl’s Puzzle” dataset.

6.3.3 A Statistical Perspective

The goodness of fit for the proposed models was measured using the RFX-BMS algorithm [Stephan et al., 2009], which must always assign each participant to one pre-defined model. Thus, it is possible that other strategies may explain the data even better. Perhaps participants remove knowledge operators both from the front and from the back of the ToM statements, or perhaps participants do not remove the negation along with the knowledge operators, or perhaps they use completely different strategies - for example, eye-tracking data could be used to inform more accurate logically-inspired models [Meijering et al., 2012, Top et al., 2018b, Top et al., 2018a as cited by Top et al., 2023]. On a similar note, coherence is bounded by the number of answer options of a puzzle, which makes it difficult to compare findings across different studies testing ToM abilities on different epistemic puzzles. Thus, while there was clear justification behind selecting RFX-BMS and coherence as validation methods, it might be worth investigating whether other appropriate statistical methods may confirm or contradict the findings of this study.

6.4 Conclusion

In this study, I 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 wide variety of fundamentally equivalent puzzles, presented in different contextual scenarios. Moreover, I introduced a novel modeling paradigm in public announcement logic, whereby as many knowledge operators as necessary (and no more) are sequentially removed from public announcements that the agent could not have understood otherwise. I showed that this modeling paradigm captures systematic patterns in the behaviour of human participants that deviate from a model of perfect recursive ToM reasoning and a model of informed random behaviour. It is my hope that the proposed experimental design and modeling paradigm will incite a wide variety of follow-up research at the intersection of cognitive modeling and dynamic epistemic logic.

REFERENCES

- J.R. Anderson. *How can the human mind occur in the physical universe?* New York, NY: Oxford University Press, 2007. 10.1093/acprof:oso/9780195324259.001.0001.
- J.R. Anderson and L.J. Schooler. The adaptive nature of memory. page 557–570. Oxford University Press, 2000.
- B. Arslan, A. Hohenberger, and R. Verbrugge. Syntactic recursion facilitates and working memory predicts recursive theory of mind. *PLoS ONE*, 12(1):e0169510, 2017a. 10.1371/journal.pone.0169510.
- B. Arslan, N.A. Taatgen, and R. Verbrugge. 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, 2017b. 10.3389/fpsyg.2017.00275.
- F. Arthaud and M. Rinard. Depth-bounded epistemic logic. In L.C. Verbrugge, editor, *Proceedings of the 19th conference on Theoretical Aspects of Rationality and Knowledge (TARK 23)*, pages 46–65, 2023. 10.4204/EPTCS.379.7.
- J.P. Borst. *The problem state bottleneck: Modeling the behavioral and neural signatures of a cognitive bottleneck in human multitasking*. PhD thesis, 2012.
- J.P. Borst, N.A. Taatgen, and H. van Rijn. The problem state: A cognitive bottleneck in multitasking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(2):363–382, 2010. 10.1037/a0018106.
- C.F. Camerer, T.H. Ho, and J.K. Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004. 10.1162/0033553041502225.
- Z. Cedegao, H. Ham, and W.H. Holliday. Does Amy know Ben knows you know your cards? A computational model of higher-order epistemic reasoning. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, volume 43, pages 2588–2594, 2021. URL <https://escholarship.org/uc/item/2kk1h4b2>.
- H. de Weerd, R. Verbrugge, and B. Verheij. Negotiating with other minds: The role of recursive theory of mind in negotiation with incomplete information. *Autonomous Agents and Multi-Agent Systems*, 31:250–287, 2017. 10.1007/s10458-015-9317-1.
- H. de Weerd, D. Diepgrond, and R. Verbrugge. Estimating the use of higher-order theory of mind using computational agents. *The B.E. Journal of Theoretical Economics*, 18(2), 2018. 10.1515/bejte-2016-0184.
- D.C. Dennett. Intentional systems. *The Journal of Philosophy*, 68(4):87–106, 1971. 10.2307/2025382.
- M. Devaine, G. Hollard, and J. Daunizeau. The social Bayesian brain: Does mentalizing make a difference when we learn? *PLoS Computational Biology*, 10(12):e1003992, 2014. 10.1371/journal.pcbi.1003992.
- L. Flobbe, R. Verbrugge, P. Hendriks, and I. Krämer. Children’s application of theory of mind in reasoning and language. *Journal of Logic, Language and Information*, 17:417–442, 2008. 10.1007/s10849-008-9064-7.
- H. Hayashi. Possibility of solving complex problems by recursive thinking. *The Japanese Journal of Psychology*, 73(2):179–185, 2002. 10.4992/jjpsy.73.179.
- S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL <http://www.jstor.org/stable/4615733>.

- C.M. Jonker and J. Treur. Modelling the dynamics of reasoning processes: Reasoning by assumption. *Cognitive Systems Research*, 4(2):119–136, 2003. 10.1016/S1389-0417(02)00102-X.
- M. Kaneko and N.Y. Suzuki. Epistemic logic of shallow depths and game theoretical applications. In *Advances In Modal Logic: Volume 3*, pages 279–298. World Scientific, 2002. 10.1142/9789812776471.0015.
- P. Kinderman, R.I. Dunbar, and R.P. Bentall. Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, 89(2):191–204, 1998. 10.1111/j.2044-8295.1998.tb02680.x.
- R. Laillier, A. Viard, M. Caillaud, H. Duclos, A. Bejanin, V. de La Sayette, F. Eustache, B. Desgranges, and M. Laisney. Neurocognitive determinants of theory of mind across the adult lifespan. *Brain and Cognition*, 136:103588, 2019. 10.1016/j.bandc.2019.103588.
- S. Lin, B. Keysar, and N Epley. Reflexively mind-blind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, 46(3):551–556, 2010. 10.1016/j.jesp.2009.12.019.
- John McCarthy. Formalization of two puzzles involving knowledge. *Formalizing Common Sense: Papers by John McCarthy*, pages 158–166, 1990.
- B. Meijering, H. van Rijn, N.A Taatgen, and R. Verbrugge. What eye movements can tell about theory of mind in a strategic game. *PLOS ONE*, 7(9):1–8, 2012. 10.1371/journal.pone.0045961.
- G.A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2): 81–97, 1956. 10.1037/h0043158.
- B. Mutter, M.B. Alcorn, and M. Welsh. Theory of mind and executive function: Working-memory capacity and inhibitory control as predictors of false-belief task performance. *Perceptual and Motor Skills*, 102(3):819–835, 2006. 10.2466/pms.102.3.819-835.
- R. Nagel. Unraveling in guessing games: An experimental study. *The American Economic Review*, 85(5):1313–1326, 1995. URL <http://www.jstor.org/stable/2950991>.
- R. Parikh. Knowledge and the problem of logical omniscience. In *ISMIS*, volume 87, pages 432–439. Citeseer, 1987.
- R. Parikh. Logical omniscience. In *International Workshop on Logic and Computational Complexity*, pages 22–29. Springer, 1994.
- J. Perner. Higher-order beliefs and intentions in children’s understanding of social interaction. In J. W. Astington, P. L. Harris, and D. R. Olson, editors, *Developing Theories of Mind*, pages 271–294, 1988.
- J. Plaza. Logics of public communications. *Synthese*, 158:165–179, 2007. 10.1007/s11229-007-9168-7.
- J.A. Plaza. Logics of public announcements. In M.L. Emrich, M.S. Pfeifer, M. Hadzikadic, and Z.W. Ras, editors, *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems: Poster Session Program*, pages 201–216, Oak Ridge National Laboratory, 1989.
- D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978. 10.1017/S0140525X00076512.
- G. Priest. *An Introduction to Non-Classical Logic: From If to Is*. Cambridge University Press, 2008.
- K.E. Stephan, W.D. Penny, J. Daunizeau, R.J. Moran, and K.J. Friston. Bayesian model selection for group studies. *Neuroimage*, 46(4):1004–1017, 2009. 10.1016/j.neuroimage.2009.03.025.
- J. Stiller and R.I. Dunbar. Perspective-taking and memory capacity predict social network size. *Social Networks*, 29(1):93–104, 2007. 10.1016/j.socnet.2006.04.001.
- J.W. Strachan, D. Albergo, G. Borghini, O. Pansardi, E. Scaliti, S. Gupta, K. Saxena, A. Rufo, S. Panzeri, Graziano M.S.A. Manzi, G., and Becchio C. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pages 1–11, 2024. 10.1038/s41562-024-01882-z.

- J.D. Top, R. Verbrugge, and S. Ghosh. An automated method for building cognitive models for turn-based games from a strategy logic. *Games*, 9(3):44, 2018a. 10.3390/g9030044.
- J.D. Top, R. Verbrugge, and S. Ghosh. Automatically translating logical strategy formulas into cognitive models. In *16th International Conference on Cognitive Modelling*, pages 182–187, 2018b.
- J.D. Top, C. Jonker, R. Verbrugge, and H. de Weerd. Predictive theory of mind models based on public announcement logic. In Nina Gierasimczuk and Fernando R. Velázquez-Quesada, editors, *Dynamic Logic. New Trends and Applications: 5th International Workshop, DaLi 2023*, volume 14401, pages 85–103. Springer, 2023. 10.1007/978-3-031-51777-8_6.
- W. van der Hoek and R. Verbrugge. Epistemic logic: A survey. In *Game Theory and Applications*, volume 8, pages 53–94. 2002.
- H. van Ditmarsch, W. van Der Hoek, and B. Kooi. *Dynamic Epistemic Logic*, volume 337. Springer Science & Business Media, 2007.
- H. van Ditmarsch, M.I. Hartley, B. Kooi, J. Welton, and J.B.W. Yeo. Cheryl’s birthday. *Electronic Proceedings in Theoretical Computer Science*, 251:1—9, 2017. ISSN 2075-2180. 10.4204/eptcs.251.1. URL <http://dx.doi.org/10.4204/EPTCS.251.1>.
- H.P. van Ditmarsch, J. Ruan, and R. Verbrugge. Sum and product in dynamic epistemic logic. *Journal of Logic and Computation*, 18(4):563–588, 2008. 10.1093/logcom/exm081.
- K. Veltman, H. de Weerd, and R. Verbrugge. Training the use of theory of mind using artificial agents. *Journal on Multimodal User Interfaces*, 13(1):3–18, 2019. 10.1007/s12193-018-0287-x.
- R. Verbrugge. Logic and social cognition: The facts matter, and so do computational models. *Journal of Philosophical Logic*, 38(6):649–680, 2009. 10.1007/s10992-009-9115-9.
- R. Verbrugge, B. Meijering, S. Wierda, H. van Rijn, and N. Taatgen. Stepwise training supports strategic second-order theory of mind in turn-taking games. *Judgment and Decision Making*, 13(1):79–98, 2018.
- M. Verma, S. Bhambri, and S. Kambhampati. Theory of mind abilities of large language models in human-robot interaction: An illusion? In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 36–45, 2024. 10.1145/3610978.364076.

APPENDIX **A** _____
_____ EXPERIMENT MATERIALS

A.1 Consent Form

Participants were first asked to read and sign an informed consent form. Participants were informed of the general layout of the experiment, the broad purpose of the research, and privacy regulations.

Form for informed consent concerning human subject research

INFORMED CONSENT

I
(name participant)
hereby consent to be a participant in the current research performed by
(name researcher)
Andreea Minculescu

I have agreed to take part in the study entitled
Cheryl's Puzzle
and I understand that my participation is entirely voluntary. I understand that my responses will be kept strictly confidential and anonymous. I have the option to withdraw from this study at any time, without penalty, and I also have the right to request that my responses not be used.

The following points have been explained to me:

1. The goal of this study is
to identify an upper-limit to the recursive reasoning about other people's knowledge.
Participation in this study should help advance our understanding of
theory of mind.
2. I shall be asked to
solve a series of logic puzzles on a computer.
3. The current study will last approximately 60 minutes. At the end of the study, the researcher will explain to me in more detail what the research was about.
4. My responses will be treated confidentially and my anonymity will be ensured. Hence, my responses cannot be identifiable and linked back to me as an individual.
5. The researcher will answer any questions I might have regarding this research, now or later in the course of the study.

Date: _____ Signature researcher: _____

Date: _____ Signature participant: _____

Figure A.1.1: Informed consent form

A.2 Interface

Participants were asked to enter a participant number (see Figure A.2.1). Each participant number was linked to a pre-determined series of puzzles, for later reference.

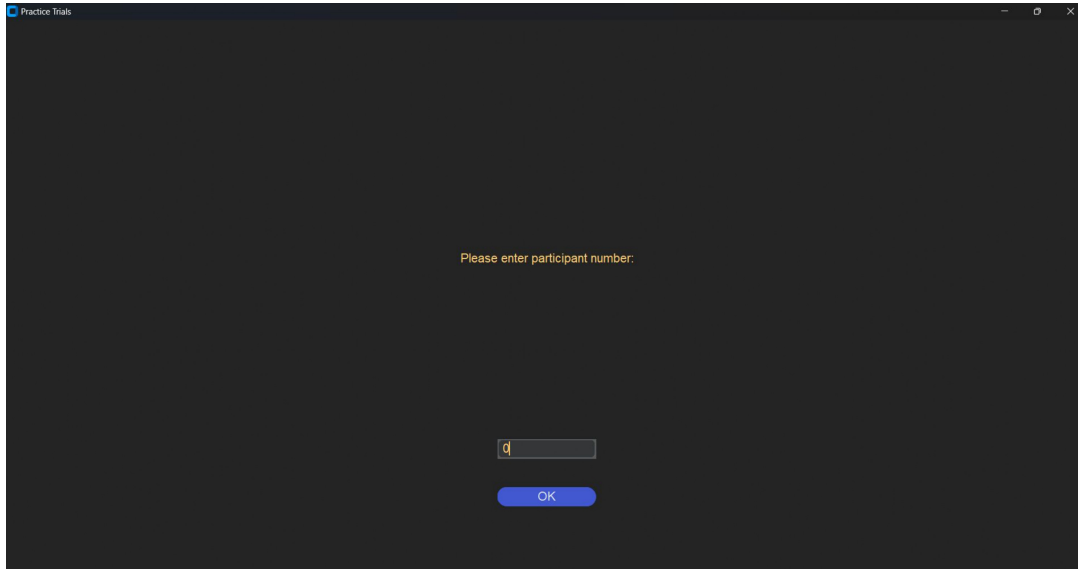


Figure A.2.1: Participants were asked to enter a participant number, ranging from 0 to 49 (each number unique for each of the 50 participants).

Next, participants were welcomed to the experiment and provided with additional instructions.

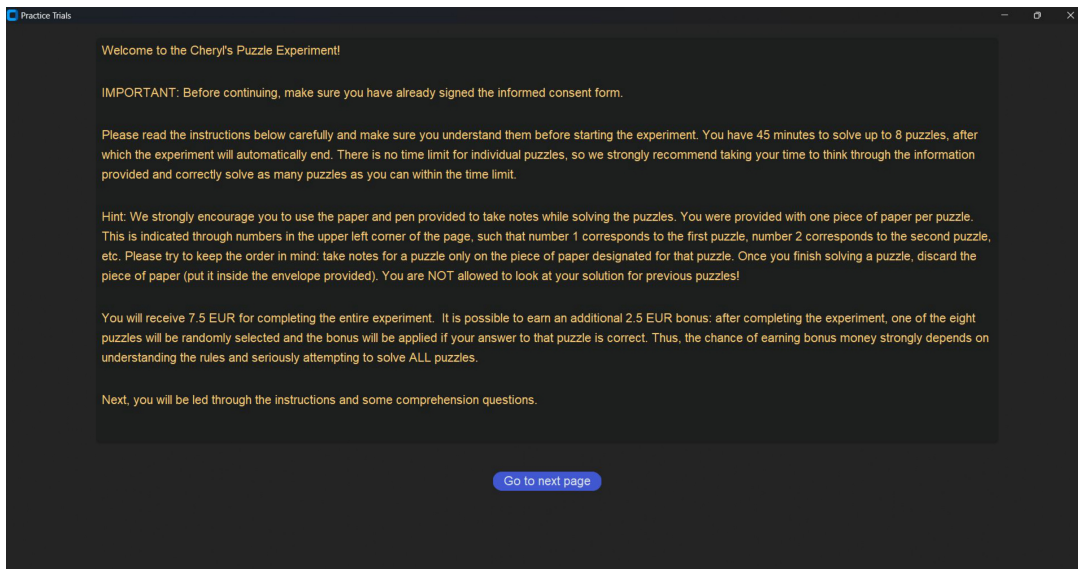


Figure A.2.2: Participants are welcomed to the experiment and provided with extra instructions on materials and monetary reward. Each paragraph is shown upon button click.

The experimental procedure comprised of three stages. Firstly, participants were tested on their (reading) comprehension of important concepts relevant to Cheryl's puzzle. See Comprehension Questions for more details.

Secondly, a series of eight puzzles were shown in succession to the participants. One could advance to the next puzzle only by selecting an answer from the drop-down menu. After the first four puzzles, a p-beauty contest was inserted, in order to conduct a post-hoc analysis of the level of theory of mind showcased, on average, by participants. The analysis of the answers recorded at this stage was the main interest of the study. See Cheryl's Puzzle for the puzzles associated with participant 0.

Lastly, participants were asked to fill in a series of personal information, such as contact information, study background, and overall experience with the experimental procedure. All information that could be used to trace back the data to the participant (i.e. name and email address) was stored separately from the rest of the answers, in order to ensure anonymity. See Background Form for more details.

A.2.1 Comprehension Questions

All text shown in this section was identical for all participants.

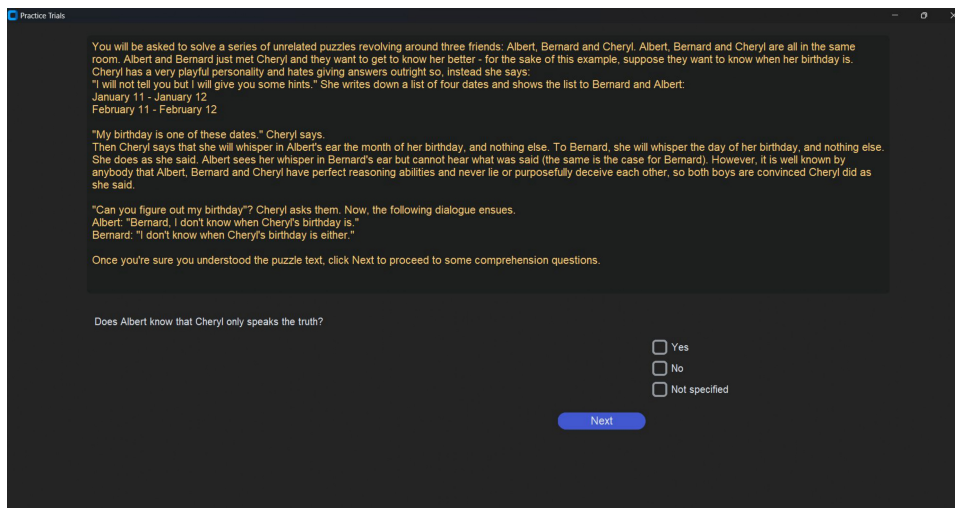


Figure A.2.3: Participants are shown a detailed instruction text, similar to the standard puzzle texts. Each paragraph is shown upon button click. Comprehension question 1 tests the participants' understanding of common knowledge and perfect logicians. Correct answer: "Yes", because *it is well known by anybody that [...] Cheryl [...] never lies*.

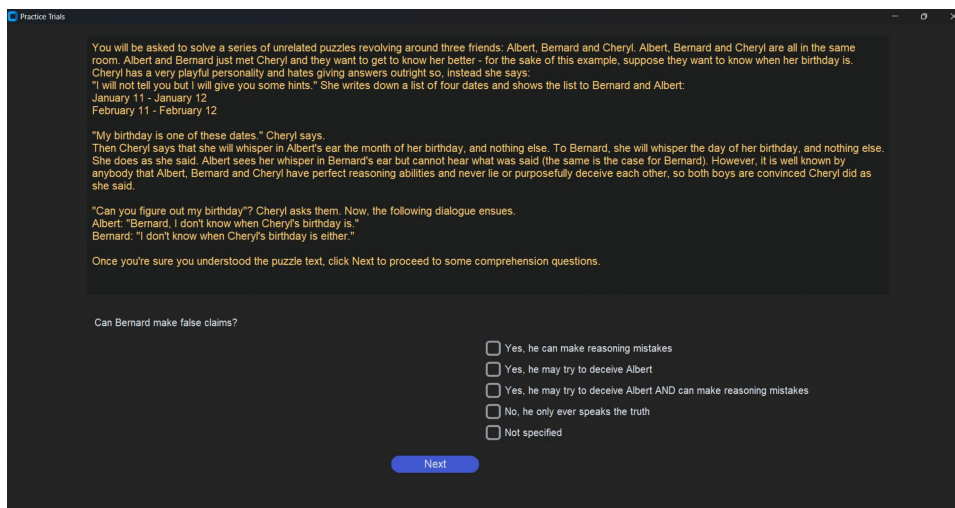


Figure A.2.4: Same text as in Figure A.2.3. Comprehension question 2 tests the participants' understanding of common knowledge and perfect logicians. Correct answer: "No, he only ever speaks the truth", because *it is well known by anybody that [...] Bernard [...] has perfect reasoning, never lies or purposefully deceives*.

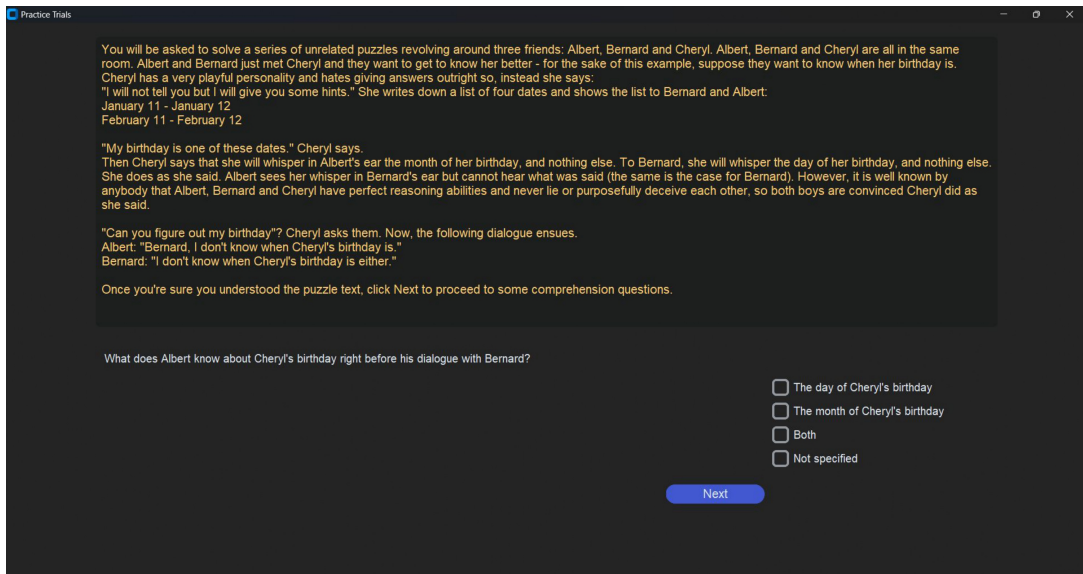


Figure A.2.5: Same text as in Figure A.2.3. Comprehension question 3 tests the participants' reading comprehension. Correct answer: "The month of Cheryl's birthday", because *Cheryl says that she will whisper in Albert's ear the month of her birthday, and nothing else.*

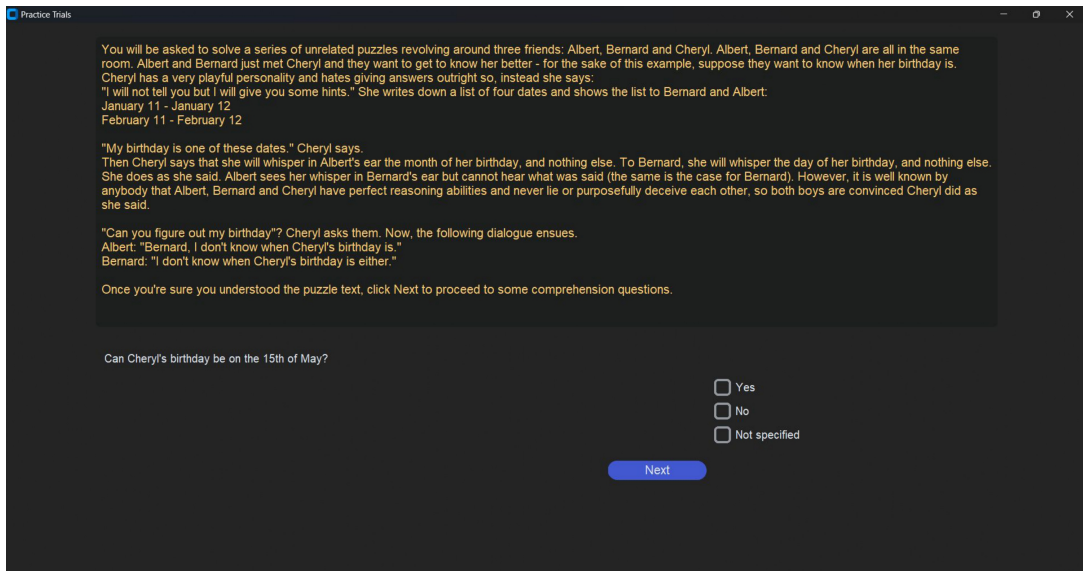


Figure A.2.6: Same text as in Figure A.2.3. Comprehension question 4 tests the participants' reading comprehension. Correct answer: "No", because Cheryl says that her month is on the list and May, 15 is not.

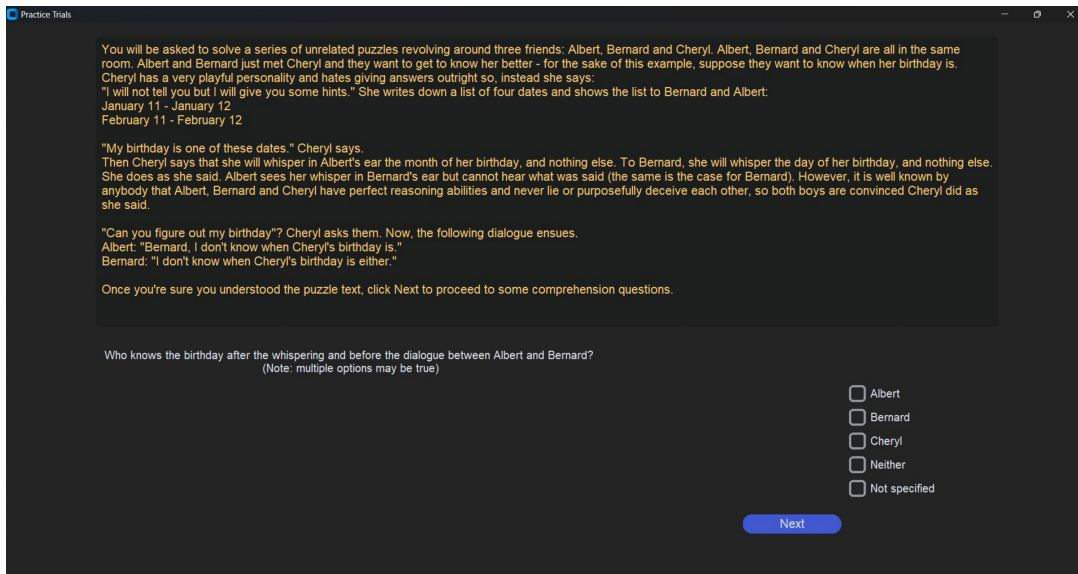


Figure A.2.7: Same text as in Figure A.2.3. Comprehension question 5 tests the participants' dialogue comprehension. Correct answer: "Cheryl", because *Albert and Bernard just met Cheryl* and they are only told the month and day, respectively. This questions relies on common sense: one cannot know another's birthday only from the month (which leaves 28-31 options) or day (which leaves 7, 11 or 12 options).

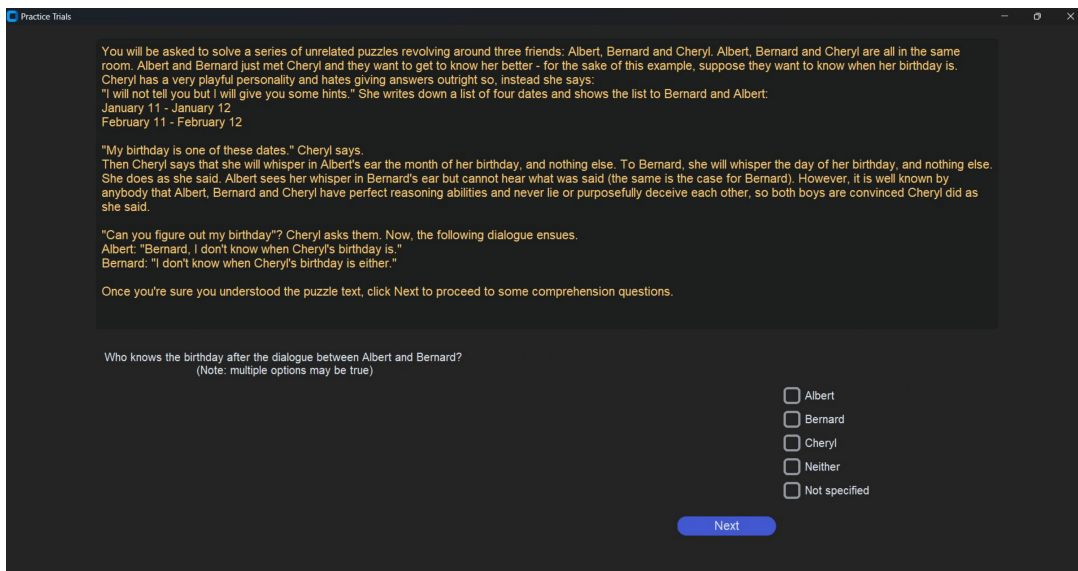


Figure A.2.8: Same text as in Figure A.2.3. Comprehension question 5 tests the participants' dialogue comprehension. Correct answer: "Cheryl", because Albert and Bernard both say that they do not know when Cheryl's birthday is.

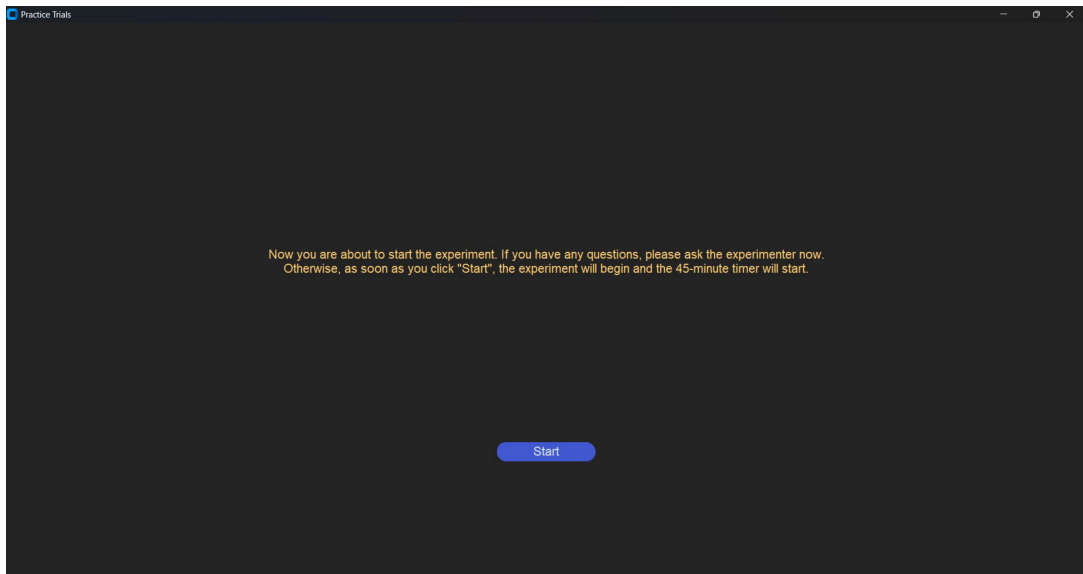


Figure A.2.9: Message shown after the participant has answered all comprehension questions correctly.

A.2.2 Cheryl's Puzzle

Each participant received a unique series of puzzles, drawn from the same pool of 64 puzzles. Below, the puzzles for participant 0 are shown.

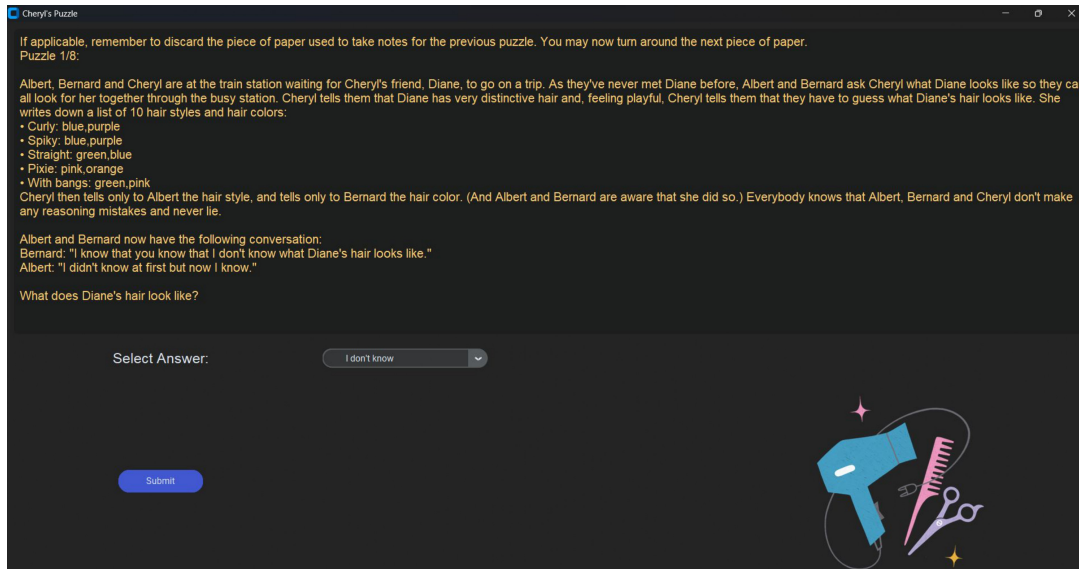


Figure A.2.10: Puzzle 1: scenario hair, level-4 ToM. Correct answer: *With bangs, green*

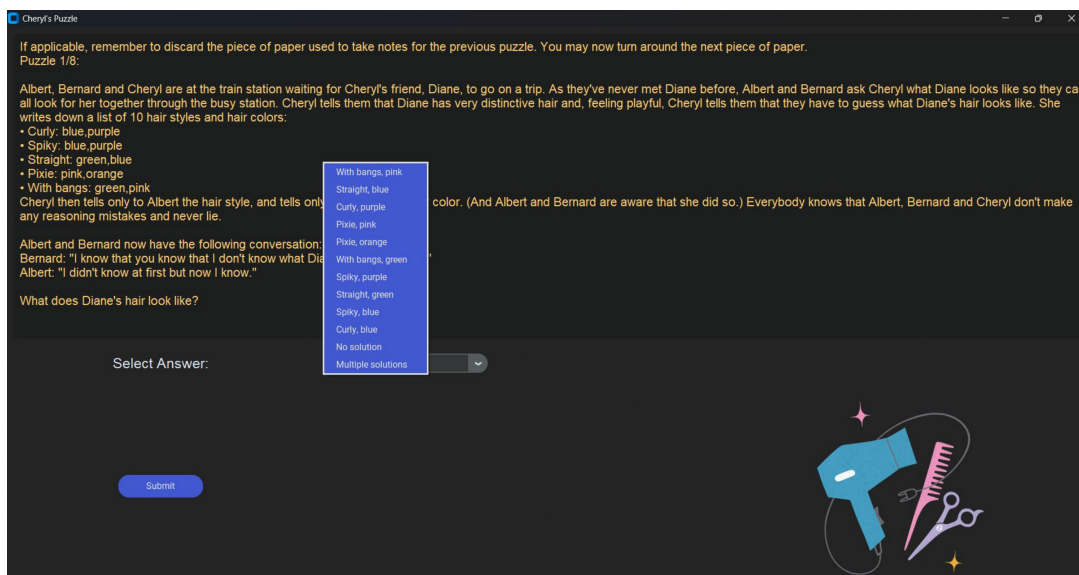


Figure A.2.11: Same puzzle text as in Figure A.2.10. The answer can be selected from the drop-down menu. Note that "Multiple solutions" and "No solution" are never the correct answer. The participants have to select an option and click on the "Submit" button to proceed to the next puzzle. For brevity, showing the drop-down menu options will be skipped for the remainder of the puzzles.

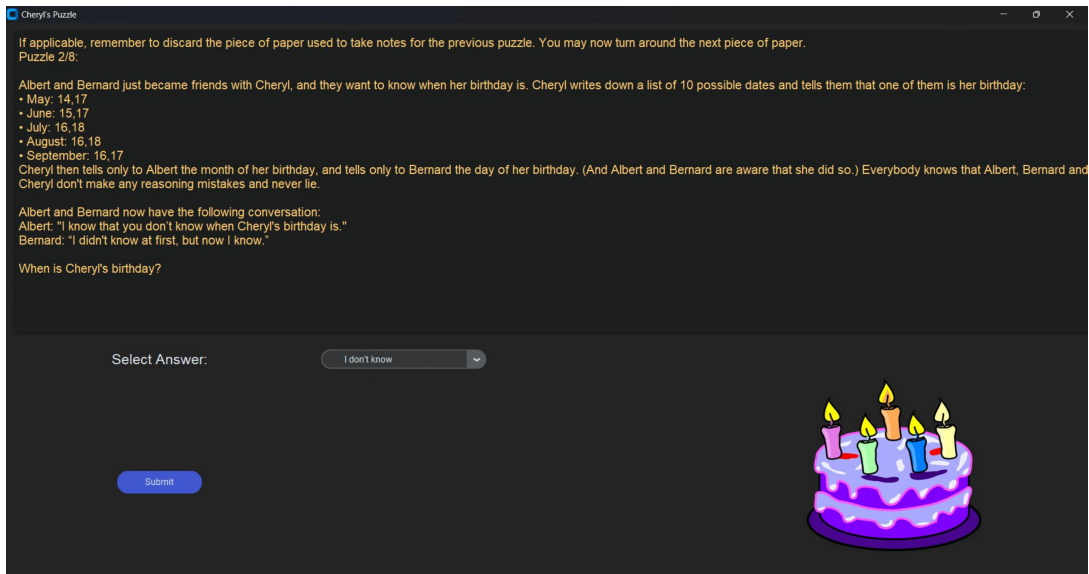


Figure A.2.12: Puzzle 2: scenario birthday, level-3 ToM. Correct answer: *September, 17*

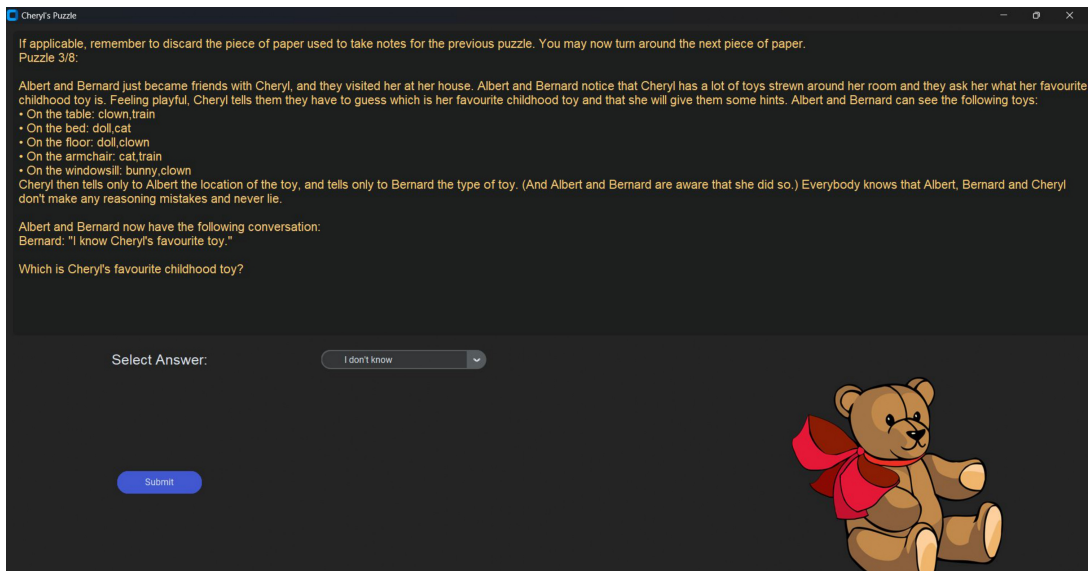


Figure A.2.13: Puzzle 3: scenario toy, level-1 ToM. Correct answer: *On the windowsill, bunny*

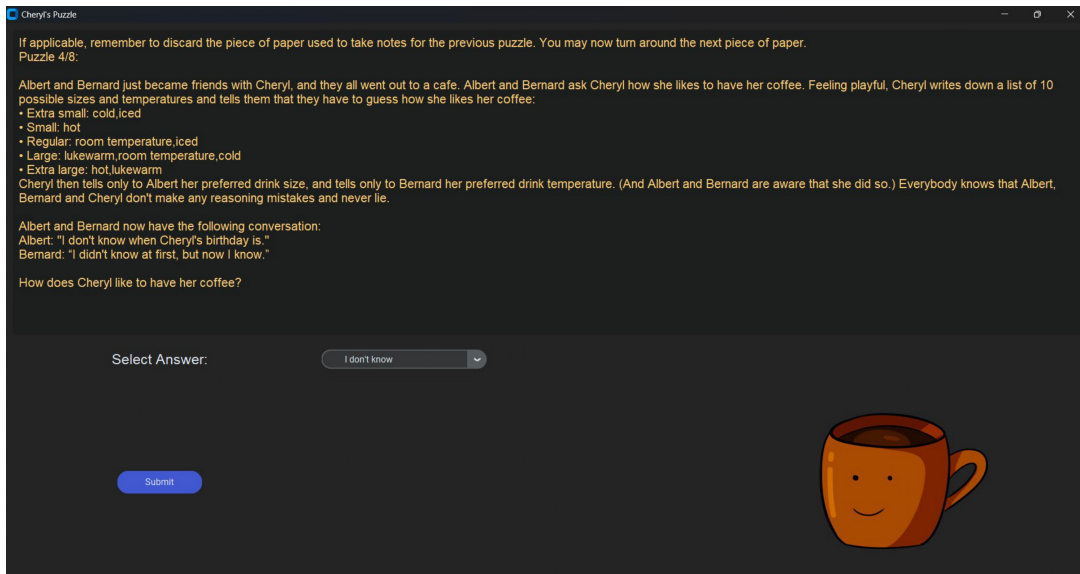


Figure A.2.14: Puzzle 4: scenario drink, level-2 ToM. Correct answer: *Extra large, hot*

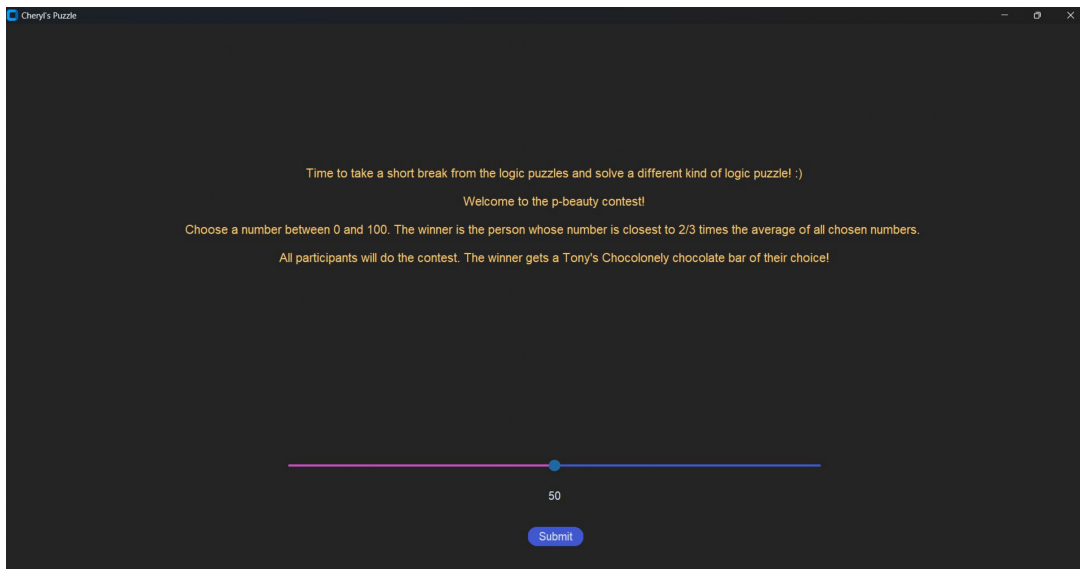


Figure A.2.15: P-Beauty Contest

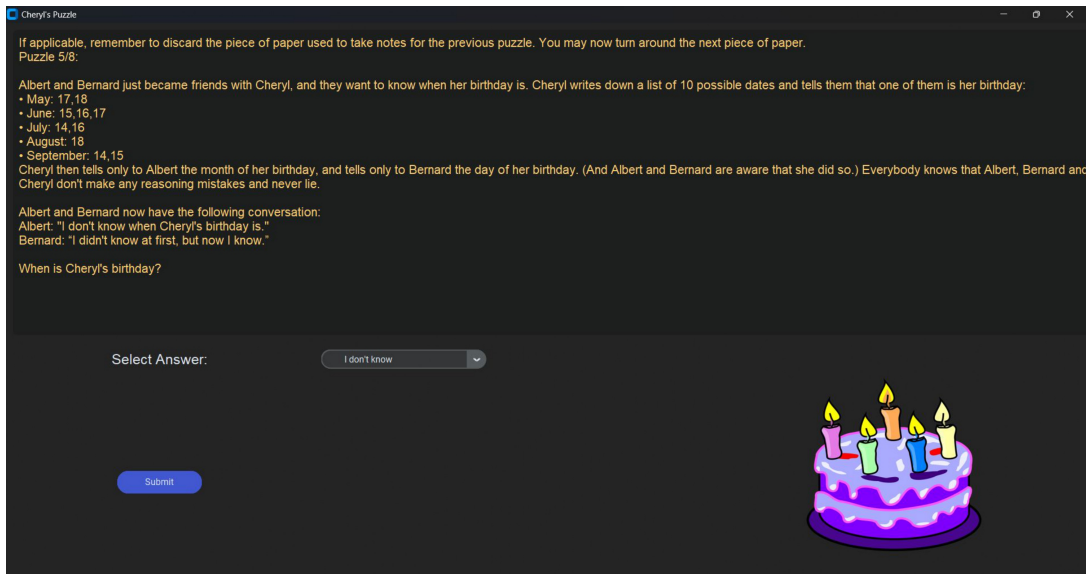


Figure A.2.16: Puzzle 5: scenario birthday, level-2 ToM. Correct answer: *May, 18*

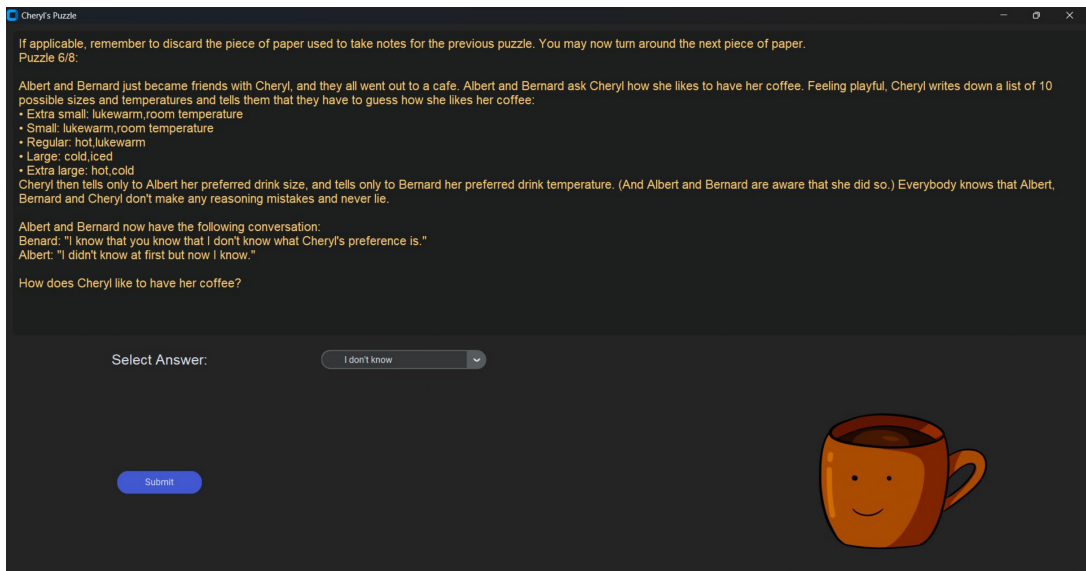


Figure A.2.17: Puzzle 6: scenario drink, level-4 ToM. Correct answer: *Extra large, hot*

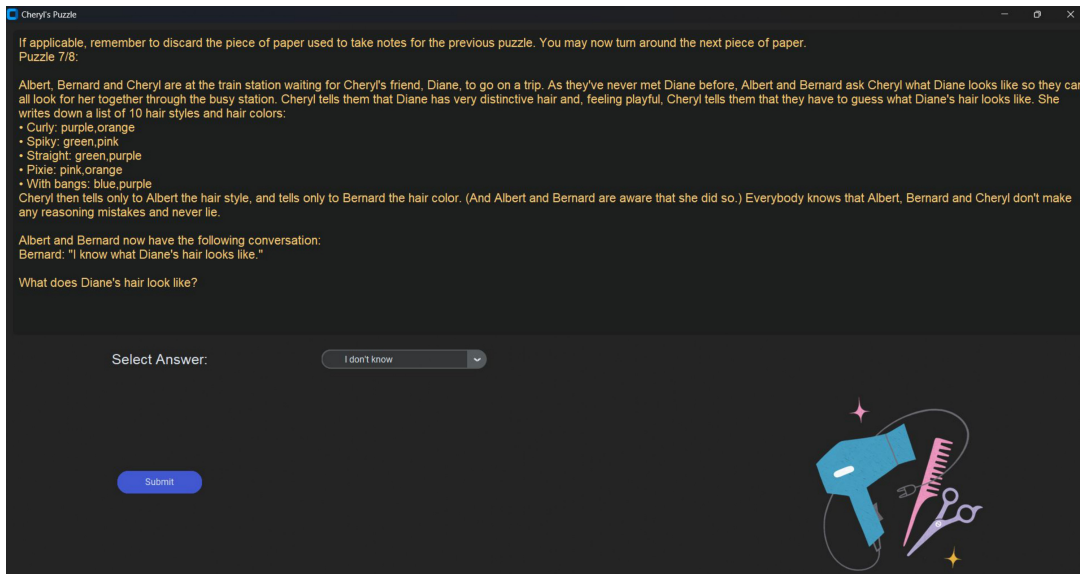


Figure A.2.18: Puzzle 7: scenario hair, level-1 ToM. Correct answer: *With bangs, blue*

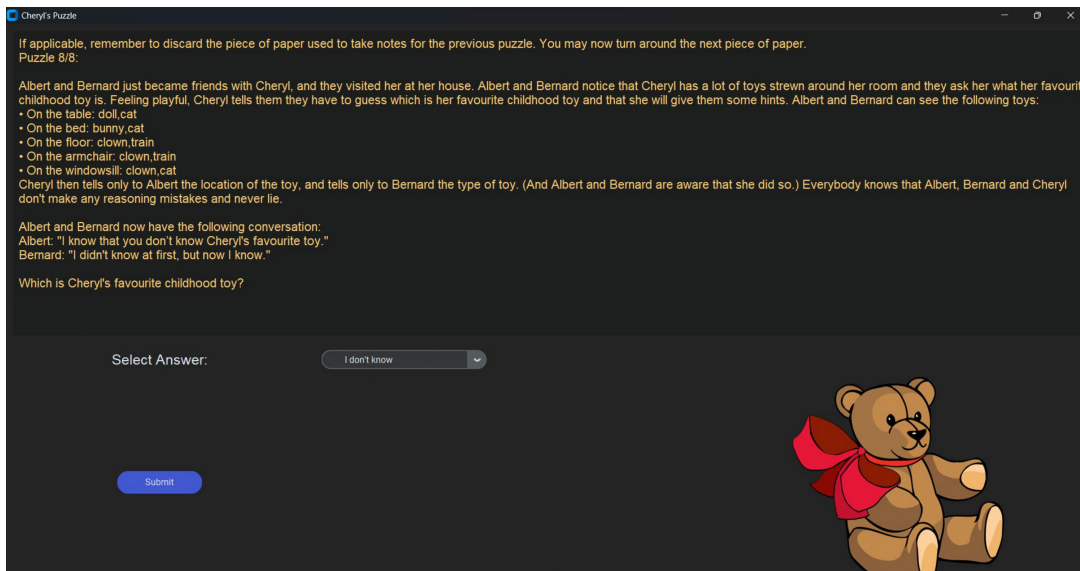


Figure A.2.19: Puzzle 8: scenario toy, level-3 ToM. Correct answer: *On the windowsill, cat*

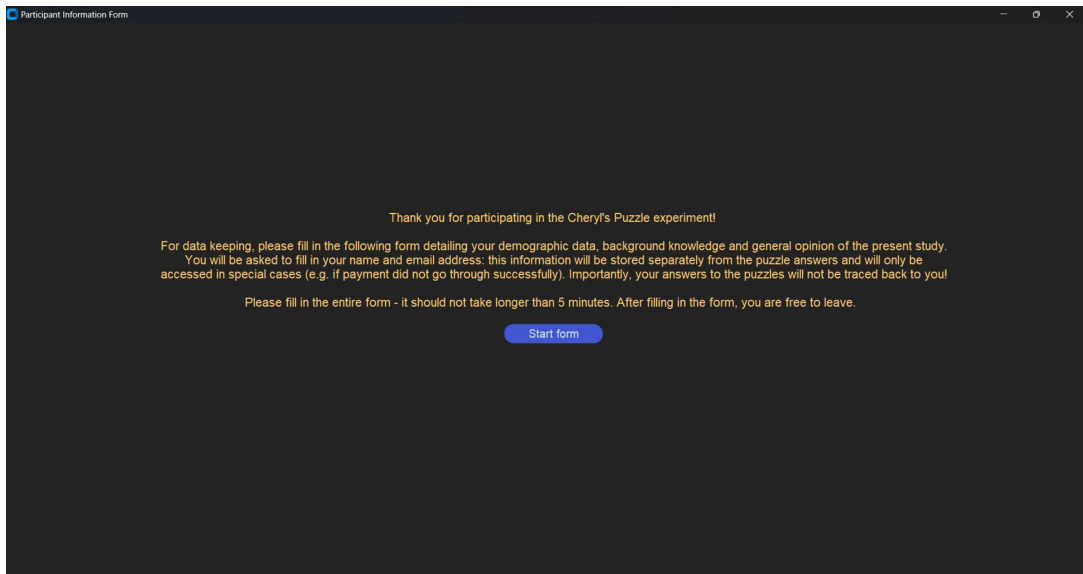
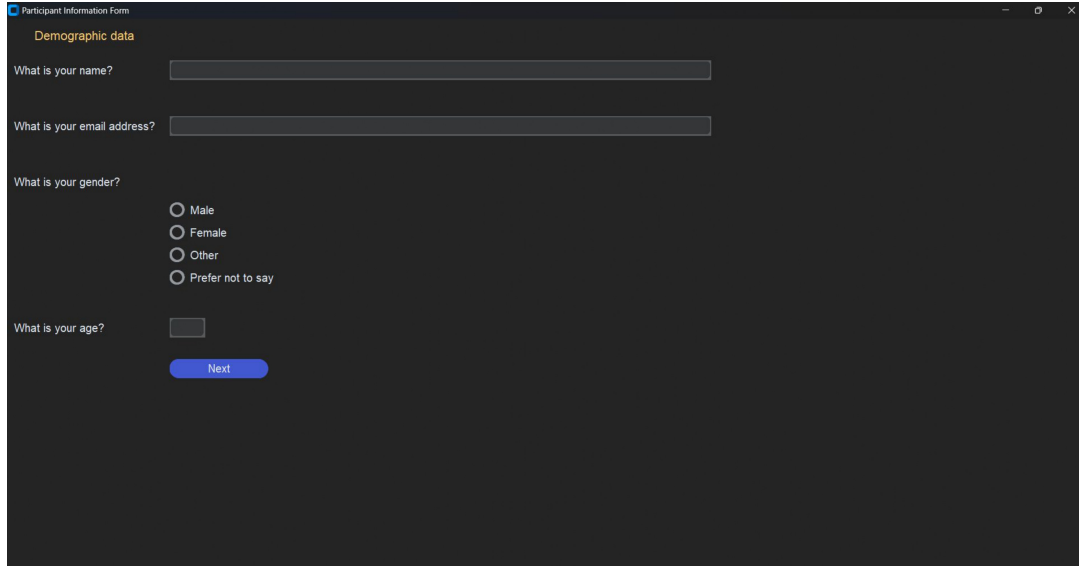


Figure A.2.20: Message shown at the end of the puzzle series.

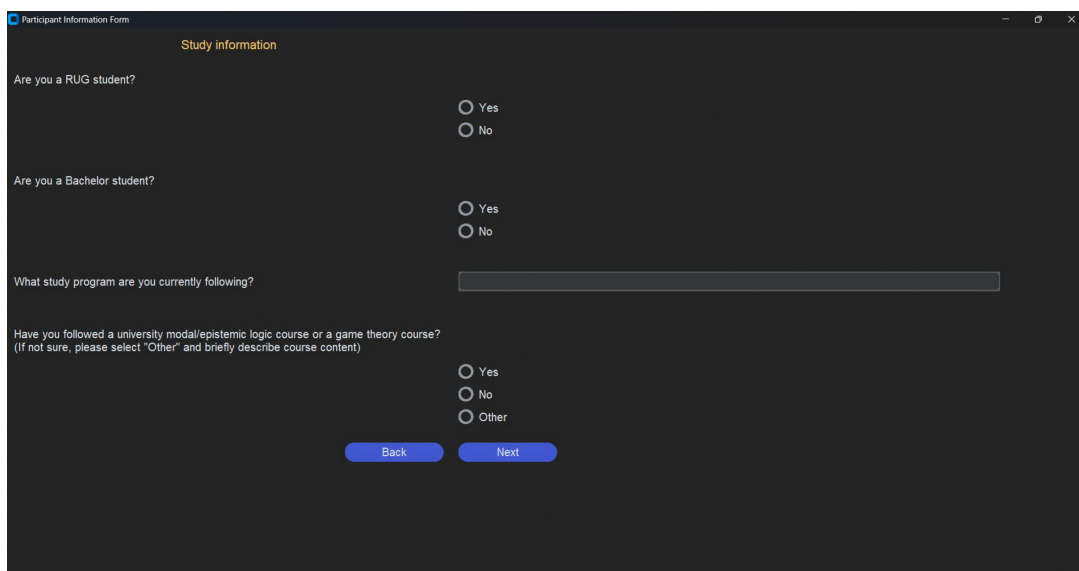
A.2.3 Background Form

All text shown in this section was identical for all participants.



The screenshot shows a window titled "Participant Information Form" with a sub-header "Demographic data". It contains four questions: "What is your name?" with a text input field; "What is your email address?" with a text input field; "What is your gender?" with four radio button options: "Male", "Female", "Other", and "Prefer not to say"; and "What is your age?" with a text input field. A blue "Next" button is located at the bottom of the form.

Figure A.2.21: First page of the form. Participants are asked to fill in their demographic data. Validity checkers were implemented for the age entry (must be a number between 15 and 50) and email address (must follow the X@Y.Z format). Additionally, no entry can be left empty.



The screenshot shows a window titled "Participant Information Form" with a sub-header "Study information". It contains four questions: "Are you a RUG student?" with "Yes" and "No" radio buttons; "Are you a Bachelor student?" with "Yes" and "No" radio buttons; "What study program are you currently following?" with a text input field; and "Have you followed a university modal/epistemic logic course or a game theory course? (If not sure, please select 'Other' and briefly describe course content)" with "Yes", "No", and "Other" radio buttons. Below the last question, a text input field is visible. At the bottom, there are blue "Back" and "Next" buttons.

Figure A.2.22: Second page of the form. Participants are asked to fill in their educational background. For the last question, if "Other" is selected, a text-box field is revealed. No entry can be left empty. Note: RUG is the abbreviation for the University of Groningen.

Participant Information Form

Experiment experience (1/2)

How easy was it to understand the instructions for Cheryl's Puzzle?

"I had great trouble understanding" 5 "I understood immediately"

Had you heard of Cheryl's (Birthday) Puzzle or a similar puzzle prior to the experiment? (If not sure, please select "Other" and briefly describe puzzle)

Yes
 No
 Other

On average, how much did you enjoy the puzzles?

"I didn't like them at all" 5 "I enjoyed them greatly"

On average, how difficult did you find the puzzles?

"I found it very easy" 5 "I found it very difficult"

Back Next

Figure A.2.23: Third page of the form. Participants are asked to recount their experience with the experimental procedure. For the second question, if “Other” is selected, a text-box field is revealed. No entry can be left empty.

Participant Information Form

Experiment experience (2/2)

Briefly, please describe the strategies you used to solve Cheryl's Puzzle (if any) and how these strategies evolved with practice.

Please rate your mood today.

"I feel very sad" 5 "I feel very happy"

Any final remarks? (optional)

Back Submit form

Figure A.2.24: Fourth page of the form. Participants are asked to recount their experience with the experimental procedure. Only the final entry can be left empty.

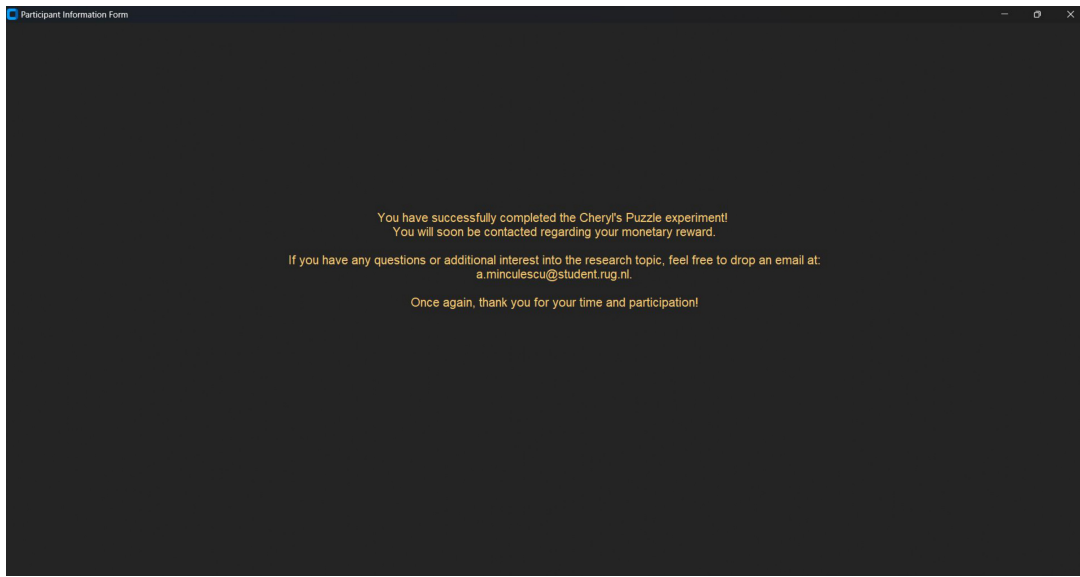


Figure A.2.25: Final message of the interface. Participants are thanked for their contribution and encouraged to contact the experimenter with any questions.