Formal semantics for the Sally-Anne tasks

Braüner, Torben; Blackburn, Patrick Rowan

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Braüner, T., & Blackburn, P. R. (2018). *Formal semantics for the Sally-Anne tasks*. 16-27. Paper presented at 4th Asian Workshop on Philosophical Logic, Beijing, China.

Formal semantics for the Sally-Anne tasks

Torben Braüner and Patrick Blackburn
Roskilde University
Denmark

October 1, 2018

Abstract

In the earlier papers [BBP16a, BBP16b] we gave hybrid-logical formalizations of the first-order and second-order Sally-Anne tasks. In these papers we argued that the reasoning used in the first-order case was essentially a simple form of propositional logic, whereas in the second-order case, a genuinely modal logic was needed. In this paper we give a more detailed justification of this claim by first discussing the simple propositional semantics for the logic used to analyse the first-order case, and then presenting a Kripke-style modal semantics for the logic underlying the second-order case.

1 Introduction

First-order false-belief tasks are a widely studied family of reasoning tasks used in cognitive and developmental psychology. A well known example is the *Sally-Anne task*:

> *A child is shown a scene with two doll protagonists, Sally and Anne, having respectively a basket and a box. Sally first places a marble into her basket. Then Sally leaves the scene, and in her absence, Anne moves the marble and puts it in her box. Then Sally returns, and the child is asked: "Where will Sally look for her marble?"*

Children above the age of four typically handle this task correctly: they say that Sally will look in the basket, which is where Sally (falsely) believes the marble to be. Younger children, on the other hand, say that Sally will look in the box: this is indeed where the marble is, but this information is not available to Sally and hence the response is incorrect. For children with *Autism Spectrum Disorder (ASD)*[1], the shift to correct responses usually occurs at a later age.

The attainment of first-order false-belief mastery is a milestone in the acquisition of *Theory of Mind (ToM)*, the capacity to ascribe mental states such as beliefs to oneself and others, and some researchers account for ASD using some form of a *ToM deficit hypothesis*; see [BC95].

Many first-order false-belief tasks have been devised, but second-order false-belief tasks are less well studied. Consider the following version of the *second-order Sally-Anne task* (the bold font highlights the new text added to the first-order version just given; the bracketed [Sally will] marks the shift in word order from 'will Sally'):

> *A child is shown a scene with two doll protagonists, Sally and Anne, having respectively a basket and a box. Sally first places a marble into her basket. Then Sally leaves the scene, and in her absence, Anne moves the marble and puts it in her box.* ***However, although Anne does not realise this, Sally is peeking through the window and sees what Anne is doing****. Then Sally returns, and the child is asked: "Where* ***does Anne think that*** *[Sally will] look for her marble?"*

[1] Autism Spectrum Disorder is a psychiatric disorder with the following diagnostic criteria: 1. Persistent deficits in social communication and social interaction. 2. Restricted, repetitive patterns of behavior, interests, or activities. For details, see *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-V)*, published by the American Psychiatric Association.

Again there is a transition age. Children above the age of six typically handle this task correctly: they respond that Anne thinks that Sally will look in the basket, which is where Anne (falsely) believes that Sally believes the marble to be. Younger children respond that Anne thinks that Sally will look in the box: this is where Sally knows the marble to be, but Anne does not know that Sally knows this and hence the response is incorrect. For children with ASD, results are mixed: Most studies indicate that performance on second-order tasks does not differ from that of typically developing children, matched on a number of variables.

The pioneering work on logical modeling of the first-order Sally-Anne task and other first-order false-belief tasks was carried out by Stenning and van Lambalgen, [Sv08], using non-monotonic closed world reasoning. The first-order Sally-Anne task has also been formalized using an interactive theorem prover for a many-sorted first-order modal logic [AB08]. Applications of logical models to second-order false-belief tasks are rare; one of the few examples is the paper [Bol18], which uses a version of dynamic epistemic logic.

As mentioned in the abstract, we presented hybrid-logical formalizations of both the first-order and second-order Sally-Anne tasks in the papers [BBP16a, BBP16b]; these were based on the earlier work on first-order false-belief tasks found in [Bra14]. The approach in [Bra14, BBP16a, BBP16b] differs from the other logical formalizations we have mentioned in a key respect: it takes the notion of *perspective shift* as fundamental. The key idea underlying these papers is that correctly handling the Sally-Anne tasks involves taking the perspective of another agent — in the first-order case Sally and in the second-order case Anne — and reasoning about what they believe. So to speak, you have to put yourself in another person's shoes.

Our previous works [Bra14, BBP16a, BBP16b] had a syntactic flavour; they were based on proof-theory, and in particular, a hybrid-logical natural deduction system. In the present paper we take a more semantic view, and develop a claim made in [BBP16a, BBP16b], namely that the reasoning in the first-order Sally-Anne task is essentially simple propositional logic, whereas in the second-order Sally-Anne task, a genuinely modal logic is needed. We substantiate this claim by first discussing a simple propositional semantics for the first-order case: essentially we are making use of distributed propositional logic: hybrid-logical machinery allows us to jump between different information locations (or perspective). We then move to the second-order case, and show the need for a full-blown modal semantics.

The logical investigations presented in this paper are carried out as part of a correlation and training study of second-order social reasoning competency in high-functioning children with ASD, the hypothesis being that training in linguistic recursion (in particular, handling of sentential complements) will improve their social cognition skills, as measured by second-order false-belief tests. More precisely, we measure the second-order reasoning capacity using a composite score involving all four reasoning patterns singled out by our logical analysis; see [BBP16b]. Our study involves 62 Danish-speaking children with ASD. See [PBB18] for results from the correlation study. See also [BPB18]. In the paper [BBP] we present some empirical results (for both children with ASD and typically developing children) on the effects of the four second-order false-belief reasoning patterns.

2 An indexed propositional logic (for the first-order task)

First we define the syntax and semantics of the fragment of hybrid logic we use for the formalization of the first-order Sally-Anne task, namely a version of Jerry Seligman's [Sel97] *Logic of Correct Description (LCD)*. The language we use is very simple: Atomic formulas are expressions like $Bl(i,t)$ and $Bm(t)$ as well as $B\neg m(t)$, which are treated as propositional symbols; we also have similar expression where the belief-modality B has been replaced by the seeing-modality S. In other words, our atomic formulas are modalized literals. We also have *nominals* s, a, b, c, ... which stand for people. The connectives are the usual Boolean connectives of propositional logic together with a *satisfaction operator* $@_s$ for each nominal s. A formula on the form $@_s\phi$ is called a *satisfaction statement*.

Definition 2.1 *A* model *for LCD is a tuple* $(P, \{V_i\}_{i \in P})$ *where*

1. *P is a non-empty set ("people") and*
2. *For each i, V_i is a function that to each propositional symbol (that is: modalized literal) assigns an element of* $\{0, 1\}$.

Thus, a model embodies propositional information indexed by points in a set. In the present approach to false-belief tasks, these points are taken to be people (in our running example Sally and Anne) and the information indexed by them is their beliefs, what they see, and so on.

Given a model $\mathfrak{M} = (P, \{V_i\}_{i \in P})$, an *assignment* is a function g that to each nominal assigns an element of P. The relation $\mathfrak{M}, g, i \models \phi$ is defined by induction, where g is an assignment, i is an element of P, and ϕ is a formula (and where p stands for an atomic formula as described above: that is, modalized literal).

$$\begin{array}{rcl}
\mathfrak{M}, g, i \models p & \text{iff} & V_i(p) = 1 \\
\mathfrak{M}, g, i \models a & \text{iff} & i = g(a) \\
\mathfrak{M}, g, i \models \phi \wedge \psi & \text{iff} & \mathfrak{M}, g, i \models \phi \text{ and } \mathfrak{M}, g, i \models \psi \\
\mathfrak{M}, g, i \models \phi \rightarrow \psi & \text{iff} & \mathfrak{M}, g, i \models \phi \text{ implies } \mathfrak{M}, g, i \models \psi \\
\mathfrak{M}, g, i \models \bot & \text{iff} & \text{falsum} \\
\mathfrak{M}, g, i \models @_a\phi & \text{iff} & \mathfrak{M}, g, g(a) \models \phi
\end{array}$$

Nominals here should be thought of as naming the unique person they are true at, for example, we shall use s as a nominal true at Sally; in effect s is a 'name' that picks out Sally. Satisfaction operators enable us to shift perspective between different people. Incidentally, letting nominals name people is exactly what is done in Arthur Prior's *egocentric logic*; see [Bla06] and also Section 1.3 in [Bra11], in particular pp. 15–16. Now, in Prior's original egocentric logic there was also a modality for talking about the taller-than relation between people. At the moment, the only modalities in our language are B and S, and they only occur in modalized literals, as our immediate aim is to show that (for the *first-order* Sally-Anne task) we are essentially working in propositional logic. But when we later consider the *second-order* Sally-Anne task, both B and S will become full-fledged modalities (rather like Prior's taller-than modality) and thus the resulting language could be viewed as a version of egocentric logic.

For more on nominals, satisfaction operators and hybrid logic more generally, see [Bla00] and [Bra11].

3 A natural deduction system for LCD

As far as the analysis of *first-order* false-belief tasks is concerned, expressions of the form $S\phi$ and $B\phi$ are essentially complicated-looking propositional symbols: they are only used in simple propositional reasoning and then fed into a perspective-shifting natural deduction rule called *Term*.

Let's make this precise. Here is the natural deduction system for LCD we shall use to analyse the first-order Sally-Anne task. We use the system obtained by extending the standard natural deduction system for classical propositional logic with the rules in Figure 1. This system is a modified version of a natural deduction system for LCD originally introduced by Seligman in [Sel97]. The system of [Sel97] was modified in [Bra04] and [Bra11] with the aim of obtaining a desirable property called closure under substitution, see Subsection 4.1.1 of [Bra11] for further information. Soundness and completeness proofs for this system can also be found in [Bra11].

Natural deduction style reasoning is based on two main ideas: The first is that there are two different kinds of rule for each logical connective: one to introduce it, the other to eliminate it. The second is that *conditional reasoning* is hardwired into natural deduction systems: we can make an assumption, work out its consequences, and then discharge it.[2] Natural deduction was originally developed to model mathematical argumentation, see [Pra65, Pra05], but there is now

[2] The discharge mechanism is a bit technical; it works as follows. All assumptions in a proof-tree are annotated with numbers. An assumption being discharged means that it is discharged by one particular rule-instance, and this

Figure 1: Natural deduction rules for LCD

$$\frac{a \qquad \phi}{@_a\phi}\,(@I) \qquad\qquad \frac{a \qquad @_a\phi}{\phi}\,(@E)$$

$$\frac{\phi_1 \quad \dots \quad \phi_n \qquad \begin{array}{c}[\phi_1]\dots[\phi_n][a]\\ \vdots \\ \psi\end{array}}{\psi}\,(\mathit{Term})^* \qquad\qquad \frac{\begin{array}{c}[a]\\ \vdots \\ \psi\end{array}}{\psi}\,(\mathit{Name})^\dagger$$

$*$ $\phi_1, \dots, \phi_n$ and ψ are satisfaction statements, and there are no undischarged assumptions in the derivation of ψ besides the specified occurrences of $\phi_1, \dots, \phi_n$ and the nominal a.

† The nominal a does not occur in the formula ψ or in any undischarged assumptions other than the specified occurrences of a.

some experimental backing for the claim that it is a mechanism underlying human deductive reasoning more generally; see [Rip08].

The rules $@I$ and $@E$ in Figure 1 are the introduction and elimination rules for satisfaction operators, in line with the first main idea behind natural deduction mentioned above. Both rules are natural and straightforward. For example, the introduction rule $@I$ can be viewed as capturing the following informal argument involving spatial location (taken from [Sel97]):

> This is Bloomington;
> the sun is shining;
> so the sun is shining in Bloomington.

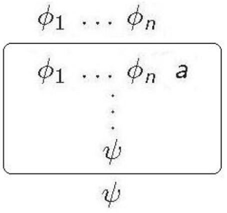
Similarly, the elimination rule $@E$ can be viewed as formalizing this argument (again from [Sel97]):

> This is Tokyo;
> people drive on the left in Tokyo;
> so people drive on the left.

But it is the *Term* rule that is central to the Sally-Anne formalization. This rule lets us switch to another perspective, do some hypothetical reasoning, and then switch back. Syntactically, the *Term* rule delimits a subderivation. Mathematically, the shift to a hypothetical person/world effected by the rule *Term* can be seen by inspecting the proof that the rule is sound (Theorem 4.1 in Chapter 4 of [Bra11]): The world of evaluation is shifted from the actual world to the hypothetical world where the nominal a is true (that is, the perspective of the person a is taken) then some reasoning is performed involving the delimited subderivation which by mathematical induction is assumed to be sound, and finally the world of evaluation is shifted back to the actual world.

The way *Term* delimits a subderivation is similar to the way a subderivation is delimited by the introduction rule for the modal operator $\Box$ in the natural deduction system for $\mathsf{S4}$ given in [BdP00], making use of explicit substitutions in derivations; more specifically, it is similar to the way subderivations are delimited by so-called boxes in linear logic. Using boxes in the style of linear logic, the *Term* rule could alternatively be formulated as follows (compare to our formulation in Figure 1).

is indicated by annotating the assumption and the rule-instance with identical numbers. A rule-instance annotated with some number discharges all undischarged assumptions that are above it and are annotated with the number in question, and moreover, are occurrences of a formula determined by the rule-instance. Two assumptions in a proof-tree belong to the same *parcel* if and only if they are annotated with the same number and are occurrences of the same formula, and moreover, either are both undischarged or have both been discharged by the same rule-instance. Thus, in this terminology, rules discharge parcels. We shall make use of the standard notation $[\phi^r]$ to denote a parcel where r is the number annotating the formulas in the parcel. We shall often omit the numbers when no confusion can occur.



We refer the reader to [Sel97] and Chapter 4 of [Bra11] for further discussion of the hybrid-logical rules. Here we'll simply add that while the rule *Name* is needed for completeness, it is not needed in our analysis of the Sally-Anne task.

4 Formalizing the first-order Sally-Anne task

To figure out where Sally will look for her marble, the subject views matters from Sally's perspective and reasons as follows. At the time t_0, Sally believed the marble to be in the basket. She saw no action to move it, so she still believed this at t_1. When she returned at t_2, she still believed the marble to be in the basket (after all, she was out of the room when Anne moved it at time t_1). So the subject concludes that Sally believes that the marble is still in the basket.

To formalize this we use the nominal s to name Sally, and the modal operators S (*sees that*) and B (*believes that*). The predicate $l(i,t)$ means that *the marble is at location i at time t*. Predicate $m(t)$ means that *the marble is moved at time t*. We take time to be discrete, and use $t+1$ as the successor of t. Using this vocabulary we can express the four belief formation principles we need:[3]

(D) $B\neg\phi \rightarrow \neg B\phi$
$(P1)$ $S\phi \rightarrow B\phi$
$(P2)$ $Bl(i,t) \wedge \neg Bm(t) \rightarrow Bl(i,t+1)$
$(P3)$ $Bm(t) \rightarrow Sm(t)$

With the help of these principles, the perspectival reasoning involved in the Sally-Anne task can be formalized as the derivation in Figure 3 (in the Appendix). Note that when applying the belief formation principles, we simply use them as rules.

This formalization was given in [Bra14], and later explicated in [BBP16a, BBP16b]. For the purpose of the present paper, we point out that all applications of the belief formation principles in Figure 3 stay within the syntax specified in the previous section, in fact, modal operators in Figure 3 only occur in front of literals. So what is really going on is indexed propositional reasoning with modalized literals, where the rule *Term* is employed to effect a shift between indexes (note that the belief formation principles are the only rules involving modal operators). In particular, in accordance with the syntax, the applications of belief formation principles in Figure 3 only involve pure Boolean combinations of propositional symbols.

[3] The "belief formation" (and "belief manipulation") terminology is borrowed from [Sv08], see [BBP16a, BBP16b] for detailed explanations. As for the belief formations principles themselves, Principle (D) is a common modal axiom and it says that beliefs are consistent, that is, if something is believed, then its negation is not also believed. Principle $(P1)$ states that a belief in ϕ may be formed as a result of seeing ϕ; this is principle (9.2) in [Sv08], page 251. Principle $(P2)$ is a principle of inertia: a belief that the predicate l is true is preserved from a time t to its successor $t+1$, unless it is believed that the marble moved at t. This is essentially Principle (9.11) from [Sv08], page 253, and axiom $[A_5]$ in [AB08], page 20. Principle $(P3)$ encodes the information that *seeing* the marble being moved is the only way a belief that the marble is being moved can be acquired (this is an arguable assumption in the Sally-Anne scenario, but it of course depends on the scenario under consideration, and other scenarios might call for other ways to acquire belief).

5 A hybrid modal logic (for the second-order task)

The language we now consider is the following: Atomic formulas are predicates like $l(i,t)$ and $m(t)$, which are treated as propositional symbols. Again we also have nominals $a, b, c, ...$ standing for people. The connectives are the usual Boolean connectives of propositional logic as well as satisfaction operators, the belief modality B, and the see modality S. So now modal operators are treated as genuine operators that can occur anywhere in a formula (not just in front of literals, as part of expressions treated as propositional symbols).

Now, the formal semantics. With the aim of interpreting the S modality, we equip the set of people with a binary relation, and with the aim of interpreting the B modality, we add a set of doxastic states equipped with a binary relation for each person. In our notion of a model, people and doxastic states are kept separate; as we note below, this keeps the logic straighforward in two respects.

Definition 5.1 *A* model *is a tuple* $(P, Q, W, \{R_i\}_{i \in P}, \{V_i\}_{i \in P})$ *where*

1. *P is a non-empty set ("people");*
2. *Q is a binary relation on P; and*
3. *W is a non-empty set ("doxastic states");*
4. *for each i, R_i is a binary relation on W; and*
5. *For each i, V_i is a function that to each propositional symbol (in the above sense) assigns a subset of W.*

Note that for any person i, the triple (W, R_i, V_i) constitutes a model for doxastic logic in the usual sense (except that no conditions have been imposed on the accessibility relation R_i).

Let a model $\mathfrak{M} = (P, Q, W, \{R_i\}_{i \in P}, \{V_i\}_{i \in P})$ be given. The relation $\mathfrak{M}, g, i, w \models \phi$ is defined by induction, where g is an assignment, i is an element of P, w is an element of W, and ϕ is a formula (and where p stands for an atomic formula as described above).

$$\begin{array}{rll}
\mathfrak{M}, g, i, w \models p & \text{iff} & w \in V_i(p) \\
\mathfrak{M}, g, i, w \models a & \text{iff} & i = g(a) \\
\mathfrak{M}, g, i, w \models \phi \wedge \psi & \text{iff} & \mathfrak{M}, g, i, w \models \phi \text{ and } \mathfrak{M}, g, i, w \models \psi \\
\mathfrak{M}, g, i, w \models \phi \rightarrow \psi & \text{iff} & \mathfrak{M}, g, i, w \models \phi \text{ implies } \mathfrak{M}, g, i, w \models \psi \\
\mathfrak{M}, g, i, w \models \bot & \text{iff} & \text{falsum} \\
\mathfrak{M}, g, i, w \models @_a \phi & \text{iff} & \mathfrak{M}, g, g(a), w \models \phi \\
\mathfrak{M}, g, i, w \models S\phi & \text{iff} & \text{for some } j \in P,\ iQj \text{ and } \mathfrak{M}, g, j, w \models \phi \\
\mathfrak{M}, g, i, w \models B\phi & \text{iff} & \text{for all } v \in W,\ wR_i v \text{ implies } \mathfrak{M}, g, i, v \models \phi
\end{array}$$

Observe that the following two properties hold (these are a straightforward consequence of our decision to keep people and doxastic states separate):

- If the sublanguage consisting of propositional symbols (in the above sense), Boolean connectives, nominals, satisfaction operators, and the see modality S is considered, then the relation $\models$ behaves like in standard hybrid modal logic.
- If the sublanguage consisting of propositional symbols (in the above sense), Boolean connectives, and the belief modality B is considered, then the relation $\models$ behaves like in standard doxastic logic.

The natural deduction system for the second-order Sally-Anne task is obtained by extending the standard natural deduction system for classical propositional logic with the hybrid-logical rules in Figure 1 as well as the modal BM rule in Figure 2, which we will motivate in the next section.

Theorem 5.2 *Let ψ be a formula and Γ a set of formulas. The first statement below implies the second statement.*

Figure 2: Belief manipulation rule for the B operator

$$\dfrac{B\phi_1 \quad \ldots \quad B\phi_n \qquad \begin{matrix}[\phi_1]\ldots[\phi_n]\\ \vdots \\ \psi\end{matrix}}{B\psi}\ (\mathsf{BM})^{*}$$

$*$ There are no undischarged assumptions in the derivation of ψ besides the specified occurrences of ϕ_1, …, ϕ_n.

1. *The formula ψ is derivable from Γ.*

2. *For any model $\mathfrak{M}$, any assignment g, any person i, and any dowastic state w, if, for any formula $\theta \in \Gamma$, it is the case that $\mathfrak{M}, g, i, w \models \theta$, then $\mathfrak{M}, g, i, w \models \psi$.*

Proof The soundness proof is by induction on the structure of the derivation of ψ. There is a case for each rule.

Soundness of the standard natural deduction rules for classical propositional logic and the hybrid-logical rules in Figure 1 follows from the soundness proof, Theorem 4.1 given in [Bra11], this being the case since the additional parameter w to the $\models$ relation above does not affect the proof (for intuition, see the first observation made above).

We now consider the case with the BM rule in Figure 2 (for intuition, see the second observation made above). Let $\mathfrak{M}$ be a model, g an assignment, i a person, w a doxastic state, such that for any formula $\theta \in \Gamma$, $\mathfrak{M}, g, i, w \models \theta$. We have to prove that $\mathfrak{M}, g, i, w \models B\psi$, that is, for any doxastic state v such that wR_iv, $\mathfrak{M}, g, i, v \models \psi$. So let a v such that wR_iv be given. It follows by induction that $\mathfrak{M}, g, i, w \models B\phi_i$, where $i \in \{1, \ldots, n\}$, cf. Figure 2, and hence $\mathfrak{M}, g, i, v \models \phi_i$. According to the side-condition on the BM rule, there are no undischarged assumptions in the derivation of ψ besides the specified occurrences of ϕ_i, so it follows by induction that $\mathfrak{M}, g, i, v \models \psi$, which is what we wanted to prove. Q.E.D.

6 Formalizing the second-order Sally-Anne task

Our formalization of the second-order Sally-Anne task is based on the observation that the experimental subject has the same beliefs about Sally in the *first-order* Sally-Anne task as Anne has about Sally in the *second-order* task, to be more precise, in the first-order task the subject believes that Sally does not know the ball has moved whereas in the second-order task Anne analogously believes that Sally does not know the ball has moved. This suggests that we should take the proof-tree we have just given (formalizing the subject's reasoning about Sally), view it as formalizing Anne's reasoning about Sally, and nest it as appropriate inside a formalization of the subject's reasoning about the second-order task. That is, we should add another level of nesting to the perspectival analysis.

As pointed out in [BBP16a, BBP16b], the additional level of nesting requires that we introduce a recursive belief manipulation rule for the B operator. There are several ways this could be done; see for example [Wan94]. We have chosen the rule for the minimal modal logic K given in Figure 2. We call it the BM rule. It is a version of a rule from [Fit07] that fits naturally our tree-style natural deduction proofs. Thus, we treat B as a full-fledged modal operator.

With this machinery in place, the reasoning in the second-order Sally-Anne task can be formalized by the proof-tree[4] in Figure 4 in the Appendix, where we use the nominal a as a name

[4] This proof-tree differs slightly from the proof-tree given in [BBP16a, BBP16b]; it is slightly simpler, as here we have omitted a reasoning step involving a third modal operator D (*deduces that*). We return to this point in the papers conclusion.

for Anne. Note that the first-order proof-tree is nested inside: the dots in the upper-right corner of Figure 4 indicate where.

The proof-tree's conclusion, $@_a B @_s Bl(basket, t_2)$, says that Anne believes that Sally believes that the marble is in the basket at the time t_2, and this is indeed the correct response to the second-order task. Note that the embedded proof-tree (which reasons from Sally's perspective) yields the conclusion $@_s Bl(basket, t_2)$, which is the correct response to the first-order task. The essential step in the proof-tree is the way the belief manipulation rule BM glues together the two levels of perspectival reasoning.

7 From first-order to second-order mastery

The first-order formalization has a simple structure: Beside the application of the *Term* rule, the proof-tree consists of sequencing applications of belief formation principles until the crucial formula $@_s Bl(basket, t_2)$ is deduced. In short, the analysis consists of *Perspectival Reasoning + Belief Formation* correctly combined. On the other hand, the second-order formalization requires the belief manipulation rule BM, which allows unrestricted *Belief Manipulation* as well.

We believe that this analysis sheds some light on the status of the shift from first-order to second-order mastery, which there is no consensus about. Starting with [SZTF94], some researchers have viewed second-order mastery as a reasonably straightforward addition to first-order mastery: the acquisition of second-order mastery occurs when the child has sufficiently strengthened his or her information processing capacities, such as working memory and sequencing; following [Mil09, Mil12] we call this the *complexity only* position. Other researchers, starting with [PW85], have argued that the transition marks a more fundamental cognitive shift; again we follow Miller and call this the *conceptual change* position.

Our analysis suggests that the transition to second-order competence marks a more significant development than is suggested by the complexity only position: Second-order competence marks the stage where beliefs become objects in their own right that can be manipulated. This shift is mirrored in our analysis: we jumped from a logic that permitted only *Belief Formation + Perspectival Reasoning* to one that also allowed unrestricted *Belief Manipulation*, as embodied by the BM rule.

8 Conclusion

In this paper we clarified the claim (made in our earlier work on the Sally-Anne tasks) that the reasoning involved in the first-order task was essentially propositional while the reasoning in the second-order task was essentially modal. The first-order analysis is carried out in the LCD fragment of hybrid logic. This is a very simple fragment of hybrid logic (it is easy to see that it has an NP-complete satisfiability problem) and moreover, the first-order proof has a very simple structure: the outputs of simply propositional reasoning involving the belief formation principles are fed into a single application of the *Term* rule. Moreover, the semantics we gave for the richer language in which we analysed the second-order problem was genuinely modal: it was a simple and fairly standard Kripke semantics and we used it to prove the soundness of the crucial BM rule. It follows that we are carrying out the second-order analysis in a standard (hybrid) modal logic (that is, a PSPACE hard logic). There indeed seems to be a jump involved in the move from first-order to second-order competency.

At a superficial level, the above considerations seems to be contradicted by recent results in the paper [vdPvRS18] which analyzes the computational complexity of theory of mind reasoning, as formalized in terms of dynamic epistemic logic. This paper shows that theory of mind reasoning is intractable (PSPACE-complete), and more surprisingly, that so formalized, this is independent of the order of the reasoning.[5]

[5] According to the paper [vdPvRS18], "Higher-order thinking could still be a source of difficulty for performing theory of mind for other reasons than computational complexity (e.g., due to limited working-memory)." In this

However, we note that what we do in the present paper is to model reasoning from the perspective of the subject doing the reasoning, that is, the reasoning of the subject is represented by a formal proof, built according to the rules of a precisely defined proof-system, thus, the subject's reasoning process is represented step-by-step by instances of proof-rules[6]. This is different from the paper [vdPvRS18] where the modeling tool is dynamic epistemic logic. See Section 5 of [Ver09] for a general discussion of using epistemic logic as a model for human social cognition. Summing up, there are two different sorts of modelling, via proof-theory and via dynamic epistemic logic–we shall leave a more direct comparison to future work, and for now only mention that there are recent attempts at modelling human reasoning involving both perspectives at the same time, see [SS18].

Nonetheless, the semantics we have given should not be considered a final analysis. For a start, the semantics we have given is probably too simple: keeping people and doxastic states separate results in a logically straightforward system, but a more fine-grained approach that modelled interactions between seeing and believing might teach us more. Moreover, as we mentioned in Footnote 6, in this paper we discussed a slightly simpler versions of our original proofs as we did not make use of the D (deduces) operator. Now, this operator is only used in a very limited way in our original analysis: it occurs only in one belief formation principle, namely $D\varphi \rightarrow B\varphi$ (if you can deduce something then you believe it) and it was easy to simplify the proof so that it was not required. Indeed, one of the reasons we used this principle was because Stenning and van Lambalgen [Sv08] make use of a similar principle in their closed world reasoning analysis; we wanted to be able to compare our approach with theirs. But the ease with which we could simplify the proof points to a deeper issue: how should we go about picking suitable belief formation rules in the first place? What predicates/modalities should they be stated in? Perhaps using 'seeing' can be justified (this verb is used in the statement of the tests) but do we really need 'deduces'? More generally: what exactly should we be able to expect in the way of belief formation principles.

We don't have answers to such questions at this stage. We'll simply repeat what we said earlier: the semantics given in this paper was introduced to make certain technical claims more precise, but it is hardly the end of the story.

9 Acknowledgements

The authors acknowledge the funding received from the VELUX FOUNDATION for the project *Hybrid-Logical Proofs at Work in Cognitive Psychology* (VELUX 33305).

References

[AB08] K. Arkoudas and S. Bringsjord. Toward formalizing common-sense psychology: An analysis of the false-belief task. In T.-B. Ho and Z.-H. Zhou, editors, *PRICAI 2008: Trends in Artificial Intelligence*, volume 5351 of *Lecture Notes in Computer Science*, pages 17–29. Springer-Verlag, 2008.

[AHV17] B. Arslan, A. Hohenberger, and R. Verbrugge. Syntactic recursion facilitates and working memory predicts recursive theory of mind. *PLOS ONE*, 12(1):e0169510, 2017. Available at https://doi.org/10.1371/journal.pone.0169510.

[BBP] T. Braüner, P. Blackburn, and I. Polyanskaya. Being deceived: Information asymmetry in second-order false belief tasks. *Topics in Cognitive Science*. In press.

connection we remark that working memory has been shown to play a significant role in higher-order false belief reasoning, cf. [AHV17] and also our own work [PBB18].

[6] Like when a formalized mathematical proof represents–describes the structure of–a proof carried out by a human mathematican.

[BBP16a] T. Braüner, P. Blackburn, and I. Polyanskaya. Recursive belief manipulation and second-order false-beliefs. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, pages 2579–2584. Philadelphia, Pennsylvania, USA: Cognitive Science Society, 2016.

[BBP16b] T. Braüner, P. Blackburn, and I. Polyanskaya. Second-order false-belief tasks: Analysis and formalization. In *Proceedings of Workshop on Logic, Language, Information and Computation (WoLLIC 2016)*, volume 9803 of *Lecture Notes in Computer Science*, pages 125–144. Springer-Verlag, 2016.

[BC95] S. Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind.* MIT Press, 1995.

[BdP00] G.M. Bierman and V. de Paiva. On an intuitionistic modal logic. *Studia Logica*, 65:383–416, 2000.

[Bla00] P. Blackburn. Representation, reasoning, and relational structures: a hybrid logic manifesto. *Logic Journal of the IGPL*, 8:339–365, 2000.

[Bla06] P. Blackburn. Arthur Prior and hybrid logic. *Synthese*, 150:329–372, 2006. Special issue edited by T. Braüner, P. Hasle, and P. Øhrstrøm.

[Bol18] T. Bolander. *Seeing Is Believing: Formalising False-Belief Tasks in Dynamic Epistemic Logic*, pages 207–236. Springer International Publishing, 2018.

[BPB18] T. Braüner, I. Polyanskaya, and P. Blackburn. A logical investigation of false-belief tasks. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society.* Madison, Wisconsin, USA: Cognitive Science Society, 2018. In press.

[Bra04] T. Braüner. Two natural deduction systems for hybrid logic: A comparison. *Journal of Logic, Language and Information*, 13:1–23, 2004.

[Bra11] T. Braüner. *Hybrid Logic and its Proof-Theory*, volume 37 of *Applied Logic Series.* Springer, 2011.

[Bra14] T. Braüner. Hybrid-logical reasoning in the Smarties and Sally-Anne tasks. *Journal of Logic, Language and Information*, 23:415–439, 2014.

[Fit07] M. Fitting. Modal proof theory. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*, pages 85–138. Elsevier, 2007.

[Mil09] S.A. Miller. Children's understanding of second-order mental states. *Psychological Bulletin*, 135:749–773, 2009.

[Mil12] Scott Miller. *Theory of mind: Beyond the preschool years.* Psychology Press, 2012.

[PBB18] I. Polyanskaya, T. Braüner, and P. Blackburn. Second-order false beliefs and recursive complements in children with Autism Spectrum Disorder. In *BUCLD 42: Proceedings of the 42nd annual Boston University Conference on Language Development*, pages 632–643. Cascadilla Press, 2018.

[Pra65] D. Prawitz. *Natural Deduction. A Proof-Theoretical Study.* Almqvist and Wiksell, Stockholm, 1965.

[Pra05] D. Prawitz. Logical consequence from a constructivist point of view. In S. Shapiro, editor, *The Oxford Handbook of Philosophy of Mathematics and Logic*, pages 671–695. Oxford University Press, 2005.

[PW85] J. Perner and H. Wimmer. “John thinks that Mary thinks that...” attribution of second-order beliefs by 5-to 10-year-old children. *Journal of Experimental Child Psychology*, 39:437–471, 1985.

[Rip08] L.J. Rips. Logical approaches to human deductive reasoning. In J.E. Adler and L.J. Rips, editors, *Reasoning: Studies of Human Inference and Its Foundations*, pages 187–205. Cambridge University Press, 2008.

[Sel97] J. Seligman. The logic of correct description. In M. de Rijke, editor, *Advances in Intensional Logic*, volume 7 of *Applied Logic Series*, pages 107 – 135. Kluwer, 1997.

[SS18] S. Smets and A. Solaki. The effort of reasoning: Modelling the inference steps of boundedly rational agents. In *Logic, Language, Information, and Computation - 25th International Workshop, WoLLIC 2018, Bogota, Colombia, July 24-27, 2018, Proceedings*, pages 307–324, 2018.

[Sv08] K. Stenning and M. van Lambalgen. *Human Reasoning and Cognitive Science*. MIT Press, 2008.

[SZTF94] K. Sullivan, D. Zaitchik, and H. Tager-Flusberg. Preschoolers can attribute second-order beliefs. *Developmental Psychology*, 30:395–402, 1994.

[vdPvRS18] I. van de Pol, I. van Rooij, and J. Szymanik. Parameterized complexity of theory of mind reasoning in dynamic epistemic logic. *Journal of Logic, Language and Information*, 27:255–294, 2018.

[Ver09] R. Verbrugge. Logic and social cognition - the facts matter, and so do computational models. *Journal of Philosophical Logic*, 38:649–680, 2009.

[Wan94] H. Wansing. Sequent calculi for normal modal propositional logics. *Journal of Logic and Computation*, 4:125–142, 1994.

Appendix

The appendix contains the formalization of the first-order and second-order Sally-Anne tasks (Figures 3 and 4).

Figure 3: Formalization of the child's correct response in the first-order Sally-Anne task

$$\dfrac{@_s Sl(basket,t_0)\quad @_s S\neg m(t_0)\quad @_s\neg Sm(t_1)\qquad \dfrac{[s]\qquad \dfrac{\dfrac{\dfrac{\dfrac{[s]\;[@_s Sl(basket,t_0)]}{Sl(basket,t_0)}\,(@E)}{Bl(basket,t_0)}\,(P1)\qquad \dfrac{\dfrac{\dfrac{[s]\;[@_s S\neg m(t_0)]}{S\neg m(t_0)}\,(@E)}{B\neg m(t_0)}\,(P1)}{\neg Bm(t_0)}\,(\mathsf{D})}{Bl(basket,t_1)}\,(P2)\qquad \dfrac{\dfrac{[s]\;[@_s\neg Sm(t_1)]}{\neg Sm(t_1)}\,(@E)}{\neg Bm(t_1)}\,(P3)}{Bl(basket,t_2)}\,(P2)}{@_s Bl(basket,t_2)}\,(@I)}{@_s Bl(basket,t_2)}\,(Term)$$

Figure 4: Formalization of the child's correct response in the second-order Sally-Anne task

$$\dfrac{@_a S@_s Sl(basket,t_0)\; @_a S@_s S\neg m(t_0)\; @_a B@_s\neg Sm(t_1)\qquad \dfrac{\dfrac{\dfrac{\dfrac{[a][@_a S@_s Sl(basket,t_0)]}{S@_s Sl(basket,t_0)}}{B@_s Sl(basket,t_0)}\,(P1)\qquad \dfrac{\dfrac{[a][@_a S@_s S\neg m(t_0)]}{S@_s S\neg m(t_0)}}{B@_s S\neg m(t_0)}\,(P1)\qquad \dfrac{[a][@_a B@_s\neg Sm(t_1)]}{B@_s\neg Sm(t_1)}\qquad \begin{matrix}[@_s Sl(basket,t_0)][@_s S\neg m(t_0)][@_s\neg Sm(t_1)]\\ \vdots\\ @_s Bl(basket,t_2)\end{matrix}}{B@_s Bl(basket,t_2)}\,(\mathsf{BM})\qquad [a]}{@_a B@_s Bl(basket,t_2)}}{@_a B@_s Bl(basket,t_2)}\,(Term)$$

The vertical dots in the upper-right corner represent the derivation in Figure 3. So this proof-tree contains two applications of *Term*: the concluding application, which is shown, and the one inside the earlier proof-tree, which is not. To save space, we have omitted names of the introduction and elimination rules for the @ operator.