Incorrect Responses in First-Order False-Belief Tests

A Hybrid-Logical Formalization

Braüner, Torben

Published in:
Logic and Logical Philosophy

DOI:
10.12775/LLP.2020.003

Publication date:
2020

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

Citation for published version (APA):
Braüner, T. (2020). Incorrect Responses in First-Order False-Belief Tests: A Hybrid-Logical Formalization. *Logic and Logical Philosophy*, *29*(3), 415-445. https://doi.org/10.12775/LLP.2020.003

Logic and **L**ogical **P**hilosophy
Volume 29 (2020), 415–445
DOI: 10.12775/LLP.2020.003

Torben Braüner

Incorrect Responses in First-Order False-Belief Tests. A Hybrid-Logical Formalization

Abstract. In the paper (Braüner, 2014) we were concerned with logical formalizations of the reasoning involved in giving correct responses to the psychological tests called the Sally-Anne test and the Smarties test, which test children's ability to ascribe false beliefs to others. A key feature of the formal proofs given in that paper is that they explicitly formalize the perspective shift to another person that is required for figuring out the correct answers—you have to put yourself in another person's shoes, so to speak, to give the correct answer. We shall in the present paper be concerned with what happens when answers are given that are *not* correct. The typical incorrect answers indicate that children failing false-belief tests have problems shifting to a perspective different from their own, to be more precise, they simply reason from their own perspective. Based on this hypothesis, we in the present paper give logical formalizations that in a systematic way model the typical incorrect answers. The remarkable fact that the incorrect answers can be derived using logically correct rules indicates that the origin of the mistakes does not lie in the children's logical reasoning, but rather in a wrong interpretation of the task.

Keywords: logic in cognitive science; hybrid logic; natural deduction; false-belief tests; perspective shift

1. Introduction

In cognitive psychology there is a reasoning test called the *Sally-Anne test*. One formulation of this reasoning test is described below.

> *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*

Received October 9, 2018. *Revised* October 10, 2019. *Published online* February 17, 2020

> *her absence, the marble is moved by Anne and hidden in her box. Then Sally returns, and the child is asked: "Where will Sally look for her marble?"*

As reported in many experimental studies (see Wellman et al., 2001) typically developing children above the age of four correctly respond where Sally must falsely believe the marble to be (in the basket). Younger children, on the other hand, respond where they know the marble to be (in the box), but this information is not available to Sally, and hence this response is incorrect. For autistic children the cutoff age is higher than four years, which was first observed in (Baron-Cohen et al., 1985). Note that passing the Sally-Anne test involves taking the perspective to another person, namely Sally, and reasoning about what she believes. You have to put yourself in Sally's shoes to get the answer right.

The Sally-Anne test is one of a family of reasoning tests called *false-belief tests* showing the same pattern: Typically developing children above four answer correctly, but autistic children have to be older. Starting with the authors of (Baron-Cohen et al., 1985), many researchers in cognitive psychology have argued that there is a link between autism and a lack of what is called *theory of mind*, which is an ability to ascribe mental states, for example beliefs, to others. To be more precise, since the ability to take a different perspective is a precondition for figuring out the correct answer to false-belief tests, the fact that autistic children have a higher cutoff age is taken to support the claim that autists have a limited or delayed theory of mind. For a very general formulation of the theory of mind deficit hypothesis of autism, see (Baron-Cohen, 1995). A critical overview of these arguments can be found in the book (Stenning and van Lambalgen, 2008).

The results of false-belief tests are robust under many variations, for example across various countries and various task manipulations, as shown in the meta-analysis (Wellman et al., 2001) involving 178 individual false-belief studies with typically developing children.

In a range of works Michiel van Lambalgen et al. have given a detailed logical analysis (but not a full formalization, that is, a fully formal proof in a specified formal system) of the reasoning taking place in the Sally-Anne test and other false-belief tests in terms of non-monotonic closed world reasoning as used in logic programming; see in particular (Stenning and van Lambalgen, 2008). In (Arkoudas and Bringsjord, 2008) (and the extended version (Arkoudas and Bringsjord, 2009)) it is described how the reasoning in the Sally-Anne test has been implemented in an interactive theorem prover using axioms and proof-rules formulated in a many-sorted first-order modal logic. The proof-rules employed in (Stenning and van Lambalgen, 2008) and (Arkoudas and Bringsjord, 2008) do not explicitly formalize the perspective shift required to pass the Sally-Anne test.

In (Braüner, 2013, 2014) we gave a logical analysis of the perspective shift required to give correct answers to the Sally-Anne test and another false-belief test called the Smarties test, and it was demonstrated that the correct reasoning in these tests can be fully formalized in a hybrid-logical natural deduction system originally introduced by Jerry Seligman in the 1990s; see in particular (Seligman, 1997). Hybrid logic is an appropriate tool to analyse the reasoning in the Sally-Anne and Smarties tests since it can explicitly represent perspectives (perspectives can be named). Moreover, Seligmans's natural deduction system can explicitly represent shifts between different perspectives (it is dealt with by a specific proof-rule), which is a key feature which distinguishes our approach from other formalizations of false-belief tests.

Using Seligman's natural deduction system for hybrid logic, in the present paper we consider what goes wrong when incorrect answers are given. It turns out that a child either answers correctly, or tends to give a specific incorrect response in accordance with the child's own knowledge, in particular, in the case of the Sally-Anne test, the child reports the real location of the marble, namely the box. This indicates that children failing false-belief tests have problems shifting to a perspective different from their own. Based on the hypothesis that these children simply reason from their own perspective, we in the present paper give formalizations of the incorrect answers in terms of the hybrid-logical natural deduction rules. It is remarkable that the incorrect answers can be derived using logically correct rules, that is, rules living up to a normative standard of logical correctness. In other words, young children and autists give normatively appropriate responses to problems different from the problems that the experimenters intends them to solve.

The formalizations of the incorrect answers give rise to an analysis in terms of the two stages in reasoning emphasized in (Stenning and van Lambalgen, 2008), namely reasoning *to* and reasoning *from* an interpretation. This analysis indicates that the origin of the mistakes lies in the first of these stages (a wrong interpretation of the task), and not in the second stage (application of logically correct rules). This type of incorrect response, stemming from a wrong task interpretation by the subject, is described in a number of different places in the literature, see in particular in the book (Stanovich, 1999).

We proceed as follows: In Section 2 we introduce hybrid modal logic, the tool we use to represent perspectives, and in Section 3 we introduce the natural deduction system for hybrid modal logic which can represent shifts between perspectives. In Sections 4 and 5 we formalize the correct reasoning in the Smarties and Sally-Anne tests. In Section 6 we consider what goes wrong when incorrect answers are given, and we point out a "pattern of failure", and show that our logical formalizations in a systematic way can model the typical incorrect answers. In Section 7 this pattern of failure is analyzed in terms of van Lambalgen and Stenning's two stages in reasoning, reasoning to and reasoning from an interpretation. In Section 8 we discuss what is called realist bias; in Section 9 we make some brief remarks on related work, and finally, in Section 10 we describe further work.

The present paper is a follow-up to the paper (Braüner, 2014). To make the present paper self-contained we give a brief recapitulation of the formalizations in that paper.

1.1. Simulation-theory

A remark on terminology is appropriate at this stage: In the literature on false-belief tests it is common to talk about "perspectives" and "shift of perspective" in an intuitive and pre-theoretic sense. We follow this way to use the terminology in the present paper[1], but we also remark that the terminology can be underpinned by an established psychological theory on theory of mind, namely what is called simulation theory.

According to simulation-theory, theory of mind should be viewed as an ability to simulate other person's mental states, that is, an ability to represent other person's mental states by adopting their perspective.[2] Simulation-theory goes back to (Gordon, 1986), where as an example it is described how chess players playing against an opponent report that

[1] Note that the terminology is related to what is called egocentric logic, see the remark at the end of the next section.

[2] Simulation-theory is one of several views on theory of mind, another one being theory-theory, according to which theory of mind should be viewed as an explicit theory of the mental realm of another person, like the theories of the physical world usually going under the heading "naive physics".

they visualize the board from the other side, taking the opposing pieces for their own and vice versa, and further, pretending that their reasons for action have shifted accordingly. In other words, such a chess player switches to the opponent's perspective, makes a decision of what to do in the opponent's situation, and after having switched back again, predicts that the opponent will make the decision in question. Of course, the player has to make adjustments for relevant differences when taking the opponent's perspective. In (Gordon, 1986) it is even described how Sherlock Holmes makes use of this sort of hypothetical reasoning (quotation from Conan Doyle, 1994).

> You know my methods in such cases, Watson. I put myself in the man's place, and, having first gauged his intelligence, I try to imagine how I should myself have proceeded under the same circumstances.
>
> (Gordon, 1986, p. 162)

For an overview of simulation theory, see (Gordon, 2009). There is extensive research in giving a neuropsychological explanation of the simulation process in terms of what are called mirror neurons, which are neurons that fire not only when an individual performs a particular action, but also when the individual observes someone else performing the same action, see for example the paper (Gallese and Goldman, 1998).

2. Hybrid modal logic

In the standard Kripke semantics for modal logic, the truth-value of a formula is evaluated "locally" at a point, where points represent times, persons, locations, or something else. Hybrid logics are modal logics that have been extended such that the object language allows direct reference to such points. This in particular means that one can formulate statements about what is the case from the perspective of a specific person. This is the central idea in the hybrid-logical approach to false-belief tests.

The most fundamental hybrid logic is obtained by extending ordinary modal logic with *nominals*, which are new propositional symbols, interpreted such that a nominal is true at exactly one point (not an arbitrary set as with ordinary propositional symbols). Most often hybrid logics involve additional machinery; in the present paper we shall consider an operator called the *satisfaction operator*. The motivation for adding such operators is to be able to formalize a statement being true relative to a specific time, person, or something else. If a is a nominal and ϕ is an arbitrary formula, then a new formula $@_a\phi$ can be formula, where $@_a$ is a satisfaction operator. Such a formula is called a *satisfaction statement.* The formula $@_a\phi$ says that the formula ϕ is true at one specific point, namely the point that the nominal a refers to.

We now give the formal syntax and semantics of the hybrid logic outlined above. We assume that a set of ordinary propositional symbols and a countably infinite set of nominals are given. We assume that the two sets are disjoint. The metavariables p, q, r, ... range over ordinary propositional symbols and a, b, c, ... range over nominals.

Definition 2.1. Formulas are defined by the grammar:

$$S \ ::= \ p \mid a \mid S \wedge S \mid S \rightarrow S \mid \bot \mid @_a S \mid \Box S$$

We adopt the conventions that $\neg\phi$ and $\Diamond\phi$ are abbreviations for respectively $\phi \rightarrow \bot$ and $\neg\Box\neg\phi$.

Definition 2.2. A *model* for hybrid logic is a triple $(W, R, \{V_w\}_{w \in W})$ where

1. W is a non-empty set,
2. R is a binary relation on W, and
3. for each w, V_w is a function that maps ordinary propositional symbol to elements of $\{0, 1\}$.

Notice that a model for hybrid logic is the same as a model for ordinary modal logic. Given a model $\mathfrak{M} = (W, R, \{V_w\}_{w \in W})$, an *assignment* is a function g that maps nominals to elements of W. The relation $\mathfrak{M}, g, w \models \phi$ is defined by induction, where g is an assignment, w is an element of W, and ϕ is a formula.

$$\begin{array}{rcl}
\mathfrak{M}, g, w \models p & \text{iff} & V_w(p) = 1 \\
\mathfrak{M}, g, w \models a & \text{iff} & w = g(a) \\
\mathfrak{M}, g, w \models \phi \wedge \psi & \text{iff} & \mathfrak{M}, g, w \models \phi \text{ and } \mathfrak{M}, g, w \models \psi \\
\mathfrak{M}, g, w \models \phi \rightarrow \psi & \text{iff} & \mathfrak{M}, g, w \models \phi \text{ implies } \mathfrak{M}, g, w \models \psi \\
\mathfrak{M}, g, w \models \bot & \text{iff} & \text{falsum} \\
\mathfrak{M}, g, w \models @_a\phi & \text{iff} & \mathfrak{M}, g, g(a) \models \phi \\
\mathfrak{M}, g, w \models \Box\phi & \text{iff} & \text{for any } v \in W \text{ such that } wRv,\ \mathfrak{M}, g, v \models \phi
\end{array}$$

The above definitions of the syntax and semantics of hybrid logic are standard and can be found many different places. See (Areces and ten

Cate, 2007) for a detailed overview of hybrid logic and see (Braüner, 2011) on hybrid logic and its proof-theory.

When points in a model are taken to stand for local perspectives, in particular times or persons, hybrid logic can represent the different perspectives in the Sally-Anne and Smarties test. The history of what now is called hybrid logic goes back to the philosopher Arthur N. Prior's work in the 1960s, and letting points in a model represent persons is exactly Prior does in his *egocentric logic* (see Section 1.3 of Braüner, 2011, in particular pp. 15–16).

3. Seligman's natural deduction system

Natural deduction style proof systems are meant to formalize actual human reasoning (on natural deduction, see Prawitz, 1965, 2005) and some psychologists have found experimental support for the claim that formal rules in natural deduction style underlies human deductive reasoning:

> [...] a person faced with a task involving deduction attempts to carry it out through a series of steps that takes him or her from an initial description of the problem to its solution. These intermediate steps are licensed by mental inference rules, such as modus ponens, whose output people find intuitively obvious. (Rips, 1994, p. x)

Remark: The logical rule modus ponens is a rule in the standard natural deduction system for propositional logic. See also (Rips, 2008) which is a reproduction of some chapters from (Rips, 1994).

Seligman's natural deduction system is obtained by adding the rules in Figure 1 to the standard natural deduction system for propositional logic (modal operators are ignored as they are irrelevant in the present paper). The system, which is a modified version of a system originally introduced in (Seligman, 1997), is taken from Chapter 4 of the book (Braüner, 2011). Recently tableau systems have been developed along similar lines (see Blackburn et al., 2017; Jørgensen et al., 2016). In (Braüner, 2011) it is proved that Seligman's natural deduction system is sound and complete.

Theorem 3.1. *Let ψ be a formula and Γ a set of formulas. The first statement below implies the second statement (soundness) and vice versa (completeness).*

1. *The formula ψ is derivable from the formulas Γ in Seligman's natural deduction system.*
2. *For any model $\mathfrak{M}$, any world $w \in W$, and any assignment g, if, for each formula $\theta \in \Gamma$, it is the case that $\mathfrak{M}, g, w \models \theta$, then it is also the case that $\mathfrak{M}, g, w \models \psi$.*

Natural deduction systems usually have two sorts of rules for each connective; rules that introduce a connective and rules that eliminate a connective. The rules $(@I)$ and $(@E)$ displayed in Figure 1 are the introduction and elimination rules for the satisfaction operator.

A decisive feature of natural deduction systems is that such systems allow to make and discharge assumptions; discharge of assumptions is indicated by bracketing $[\dots]$ the assumptions in question. This is what is going on in the rule (*Term*) in Figure 1, which enables hypothetical reasoning about what is the case at a specific possible world (time or person), usually not the same as the actual world.

The hypothetical reasoning in the (*Term*) rule is the reasoning represented by the subderivation indicated by vertical dots and having discharged assumptions $[\phi_1]$, …, $[\phi_n]$, $[a]$ and conclusion ψ. The world where the hypothetical reasoning takes place is the world referred to by the nominal discharged by the rule—indicated by $[a]$. This nominal can be called the point-of-view nominal. It is important to note that the side-condition on the (*Term*) rule, that the assumptions ϕ_1, …, ϕ_n and the conclusion ψ all have to be satisfaction statements, ensures that their truth-values are not affected when perspective is shifted from the actual world to the hypothetical world.

The way the (*Term*) rule delimits a subderivation is similar to the way subderivations are delimited by what are called proof boxes in linear logic. Formulated using such proof boxes, the (*Term*) rule looks as follows (compare to our formulation in Figure 1).

$$\begin{array}{c}\phi_1 \;\cdots\; \phi_n\\ \boxed{\begin{array}{c}\phi_1 \;\cdots\; \phi_n \quad a\\ \vdots\\ \psi\end{array}}\\ \psi\end{array}$$

The perspective shift required to give a correct answer to the Sally-Anne and Smarties tests is captured particularly well by the (*Term*) rule, see (Braüner, 2014) for more information.

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

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

* The formulas $\phi_1, \dots, \phi_n$, and ψ are all satisfaction statements and there are no undischarged assumptions in the derivation of ψ besides the specified occurrences of $\phi_1, \dots, \phi_n$, and a.
** The nominal a does not occur in ψ or in any undischarged assumptions other than the specified occurrences of a.

Figure 1. Hybrid-logical rules

4. Correct response in the Smarties test

First a brief description of how the correct reasoning in the Smarties test is formalized in (Braüner, 2014). The Smarties test comes in two versions, namely a version where the experimental subject shifts perspective to a second person, and a version where there is a shift of perspective to an earlier time, see (Gopnik and Astington, 1988). Here is a formulation of the temporal version.

> *A child is shown a Smarties tube where unbeknownst to the child the Smarties have been replaced by pencils. The child is asked: "What do you think is inside the tube?" The child answers "Smarties!" The tube is then shown to contain pencils only. The child is then asked: "Before this tube was opened, what did you think was inside?"*

We start with an informal analysis. Let us call the child Peter. Let the nominal a denote the time when Peter answers the first question, and let t be the time where he answers the second question. To answer the second question, Peter imagines himself being at the earlier time a where he was asked the first question. At that time he deduced that there were Smarties inside the tube from the fact that it is a Smarties tube. Imagining being at the time a, Peter reasons that since he at that time deduced that the tube contained Smarties, he must also have come

$$\dfrac{@_a Dp \qquad \dfrac{[a] \qquad \dfrac{\dfrac{[a] \qquad [@_a Dp]}{Dp}\,(@E)}{Bp}\,(\text{P0})}{@_a Bp}\,(@I)}{@_a Bp}\,(\mathit{Term})$$

Figure 2. Formalization of an experimental subject's correct reasoning in the Smarties test (both temporal and person shift versions)

to believe that the tube contained Smarties. Therefore, at t he concludes that at the earlier time a he believed that the tube contained Smarties.

With the aim for formalizing the Smarties task, we extend the language of hybrid logic with two modal operators, D and B. We use the following symbolizations

p	The tube contains Smarties
D	Peter deduces that ...
B	Peter believes that ...
a	The time where Peter answers the first question

and we take the principle

$$D\phi \to B\phi \tag{P0}$$

as an axiom. This is principle (9.4) in the book (Stenning and van Lambalgen, 2008, p. 251). Given this machinery, the shift of temporal perspective in the Smarties test can be formalized directly as the derivation in Figure 2, where a is the point-of-view nominal and where principle (P0) has been formulated as a rule (more compact and in line with the natural deduction reasoning style). In this derivation, the premise $@_a Dp$ expresses that the experimental subject Peter at the earlier time a deduced that the tube contained Smarties, which he remembers at t.

Besides the temporal version, we also consider the version of the Smarties test where there is a shift of perspective to another person. The only difference between the two versions of the test is the second question where

> *"Before this tube was opened, what did you think was inside?"*

is replaced by

> *"If your mother comes into the room and we show this tube to her, what will she think is inside?"*

To give a correct answer to the last of these two questions, the child Peter imagines being the mother coming into the room. Imagining being the mother, Peter reasons that the mother must deduce that the tube contains Smarties from the fact that it is a Smarties tube, and from that, she must also come to believe that the tube contains Smarties. Therefore, Peter concludes that the mother would believe that the tube contains Smarties.

The derivation formalizing this line of reasoning is exactly the same as in the temporal version, Figure 2, but some symbols are interpreted in a different way, namely

D	Deduces that ...
B	Believes that ...
a	The imagined mother

So now the nominal a refers to a person rather than a time. Thus, the premise $@_a Dp$ in the derivation in Figure 2 expresses that the imagined mother deduces that the tube contains Smarties, which the child doing the reasoning takes to be the case since the mother is imagined to be present in the room.

5. Correct response in the Sally-Anne test

In the present section we give a brief description of how the correct reasoning in the Sally-Anne test is formalized in (Braüner, 2014) (see that paper for a detailed description as well as a detailed comparison to the formalizations of the Sally-Anne test given in (Arkoudas and Bringsjord, 2008; Stenning and van Lambalgen, 2008)).

We start with an informal analysis: Let us call the child Peter again. We shall consider three successive times t_0, t_1, t_2, where t_0 is the time at which Sally leaves the scene, t_1 is the time at which the marble is moved to the box, and t_2 is the time after Sally has returned when Peter answers the question. To answer the question, Peter imagines himself being Sally, and he reasons as follows: At the time t_0 when Sally leaves, she believes that the marble is in the basket since she sees it, and she sees no action to move it, so when she is away at t_1, she still believes the marble is in the basket. She does not see that the marble is moved at t_1, so she does not believe that this is the case, and hence, at t_2 after she has returned, she still believes that the marble is in the basket. Therefore, Peter concludes that Sally at t_2 believes that the marble is in the basket.

In our formalization we use the modal operators S and B as well as the predicates $l(i,t)$ and $m(t)$. The argument i in the predicate $l(i,t)$ denotes a location, and the argument t in $l(i,t)$ and $m(t)$ denotes a timepoint. We take time to be discrete, and the successor of a time t is denoted $t+1$.

$l(i,t)$	The marble is at location i at time t
$m(t)$	The marble is moved at time t
S	Sees that ...
B	Believes that ...
a	The person Sally

We also use the following four principles:

$$B\phi \rightarrow \neg B\neg\phi \tag{D}$$

$$S\phi \rightarrow B\phi \tag{P1}$$

$$Bl(i,t) \wedge \neg Bm(t) \rightarrow Bl(i,t+1) \tag{P2}$$

$$Bm(t) \rightarrow Sm(t) \tag{P3}$$

Principle (D) is a common modal axiom which says that beliefs are consistent, that is, if something is believed, then its negation is not also believed. We will use $B\neg\phi \rightarrow \neg B\phi$ which is equivalent to (D).

Principle (P1) formalizes how a belief in something may be formed, namely by seeing it being the case. This is principle (9.2) in (Stenning and van Lambalgen, 2008, p. 251).

Principle (P2) is reminiscent of principle (9.11) in (Stenning and van Lambalgen, 2008, p. 253) and axiom $[A_5]$ in (Arkoudas and Bringsjord, 2008, p. 20). Principle (P2) formalizes a “principle of inertia” saying that a belief in the predicate l being true is preserved over time, unless it is believed that an action has taken place causing the predicate to be false.

Principle (P3) encodes the information that *seeing* the marble being moved is the only way to acquire a belief that the marble is being moved.

Given the above machinery, the shift of person perspective in the Sally-Anne test can be formalized as the derivation in Figure 3, where a is the point-of-view nominal and where we have omitted names of the introduction and elimination rules for the @ operator to save space.

The first two premises $@_a Sl(\mathit{basket}, t_0)$ and $@_a S\neg m(t_0)$ in the derivation express that Sally at the earlier time t_0 saw that the marble was in the basket and that no action was taken to move it, which the child Peter remembers. The third premise, $@_a \neg Sm(t_1)$, expressess that Sally

$$\dfrac{@_a Sl(basket,t_0) \quad @_a S\neg m(t_0) \quad @_a \neg Sm(t_1) \quad \dfrac{[a] \qquad \dfrac{\dfrac{\dfrac{\dfrac{\dfrac{[a][@_a Sl(basket,t_0)]}{Sl(basket,t_0)}}{Bl(basket,t_0)}\,\text{(P1)} \quad \dfrac{\dfrac{\dfrac{[a][@_a S\neg m(t_0)]}{S\neg m(t_0)}}{B\neg m(t_0)}\,\text{(P1)}}{\neg Bm(t_0)}\,\text{(D)}}{Bl(basket,t_1)}\,\text{(P2)} \quad \dfrac{\dfrac{[a][@_a \neg Sm(t_1)]}{\neg Sm(t_1)}}{\neg Bm(t_1)}\,\text{(P3)}}{Bl(basket,t_2)}\,\text{(P2)}}{@_a Bl(basket,t_2)}}{@_a Bl(basket,t_2)}\,(\mathit{Term})$$

The temporal progression of the hypothetical reasoning is pointed out using colors: Red is what happens at t_0, blue at t_1, and magenta at t_2. The colors are obviously not a formal part of the derivation. See Subsection 7.2 for further analysis.

Figure 3. Formalization of an experimental subject's correct reasoning in the Sally-Anne test

Table 1	Correct response	Formula
Smarties (temporal version)	At the time of question one Peter believes that the tube contains Smarties	$@_a Bp$
Smarties (person version)	The imagined mother believes that the tube contains Smarties	$@_a Bp$
Sally-Anne	Sally believes that the marble is in the basket at the time t_2	$@_a Bl(basket, t_2)$

did not see the marble being moved at the time t_1, this being the case since she was absent, which Peter remembers.

6. What goes wrong when incorrect answers are given?

The derivations in Figures 2 and 3 are formalizations of the reasoning taking place in the cases where correct responses are given to the Smarties and the Sally-Anne tests. The correct answers are summed up below in Table 1. Note the nominal a (the point-of-view nominal) that is discharged by the instances of the (Term) in Figures 2 and 3, shows that the formulas in Table 1 are derived via a perspective shift to a, representing respectively the time where the first question is answered, the imagined mother, and the doll Sally.

Let b stand for the experimental subject's own perspective[3], that is, in the temporal version of the Smarties test, b stands for the time where the second question is answered and in the person version of the Smarties test b stands for the person Peter. Also in the Sally-Anne test, b stands for the person Peter. Thus, to derive the correct responses in Figures 2 and 3, the perspective is shifted from b to a, and then back from a to b.

Now, the derivations of the correct responses in Figures 2 and 3 do not tell what happens when incorrect responses are given. But it turns out that a subject either answers correctly, or tends to give a specific incorrect response in accordance with the subject's own knowledge, in particular, in the case of the Sally-Anne test, the subject reports the real location of the marble. This is clear from studies where the experimental design does not involve a forced choice between responses, for example (Baron-Cohen et al., 1985):

> The critical question was, "Where will Sally look?" after she returns. [...] normal preschool children answered by pointing to where the marble was put in the first place. [...] The autistic group, on the other hand, answered by pointing consistently to where the marble really was. They did not merely point to a 'wrong' location, but rather to the actual location of the marble. This becomes especially clear on trial 2 where the autistic children never pointed to the box (which had been the wrong location on trial 1), but instead to the experimenters pocket—that is, again to where the marble really was.
>
> (Baron-Cohen et al., 1985, pp. 42–43)

Thus, the autistic children have a systematic tendency to report their own beliefs, rather than that of different persons (we shall discuss this phenomenon more in Section 8). Based on the hypothesis that these children simply reason from their own perspective, we shall in what follows analyze this pattern in the incorrect responses. To this end we let

[3] Note that b does not occur in the formal derivations in Figures 2 and 3, like it is not mentioned in a formal mathematical proof that it has been carried out by a certain mathematician. The formal derivation itself does not care who carries out the reasoning (or for that matter whether the reasoning takes place in a computer, or in some other medium). Note also that b does in fact occur in Figures 4 and 5, but this is because the latter derivations are about what is the case from the perspective b, which happens to be the same perspective as the perspective of the experimental subject who carries out the reasoning. Thus, the nominal b in Figures 4 and 5 is true because the subject Peter is reasoning from his own perspective.

Table 2	Incorrect response	Formula
Smarties (temporal version)	At the time of question one Peter believes that the tube contains pencils	$@_a Bq$
Smarties (person version)	The imagined mother believes that the tube contains pencils	$@_a Bq$
Sally-Anne	Sally believes that the marble is in the box at the time t_2	$@_a Bl(box, t_2)$

Table 3	True proposition	Formula
Smarties (temporal version)	At the time of question two Peter believes that the tube contains pencils	$@_b Bq$
Smarties (person version)	Peter believes that the tube contains pencils	$@_b Bq$
Sally-Anne	Peter believes that the marble is in the box at the time t_2	$@_b Bl(box, t_2)$

the propositional symbol q stand for “The tube contains pencils”. Then the incorrect responses can be summed up as follows in Table 2. The formulas in Table 2 are false in the scenarios described by the reasoning tasks—the simple reason being that the formulas represent the incorrect answers—but observe the following: If the perspective a in the formulas above is replaced by the experimental subject’s own perspective b, then true formulas are obtained, namely the formulas in Table 3.

Indeed, the formulas $@_b Bq$ and $@_b Bl(box, t_2)$ in Table 3 are derivable in Seligman’s system, extended with the principles introduced in the previous section.

The formula $@_b Bq$ in Table 3 is derivable from the nominal b and the formula $@_b Sq$ by the simple derivation in Figure 4. The nominal b is true since it denotes the perspective of the subject Peter who happens to be the person doing the reasoning, that is, Peter is reasoning from his own perspective, and $@_b Sq$ is obviously true in the temporal as well as the person version of the test, in both cases since Peter when the second question is answered sees that the tube contains pencils.

The formula $@_b Bl(box, t_2)$ in Table 3 is derivable from b together with $@_b Sl(box, t_1)$ and $@_b S\neg m(t_1)$ by the derivation in Figure 5. Again, the nominal b is true since Peter is reasoning from his own perspective. The formulas $@_b Sl(box, t_1)$ and $@_b S\neg m(t_1)$ express that Peter at the earlier time t_1 saw that the marble was in the box and that no action was taken to move it, which he remembers. From these three formulas,

$$\dfrac{b \qquad \dfrac{\dfrac{b \qquad @_b Sq}{Sq}\,(@E)}{Bq}\,(\mathrm{P1})}{@_b Bq}\,(@I)$$

Figure 4. Formalization of incorrect reasoning in the Smarties test (what is the case from the subject's own perspective)

$$\dfrac{b \qquad \dfrac{\dfrac{\dfrac{b \qquad @_b Sl(box, t_1)}{Sl(box, t_1)}\,(@E)}{Bl(box, t_1)}\,(\mathrm{P1}) \qquad \dfrac{\dfrac{\dfrac{b \qquad @_b S\neg m(t_1)}{S\neg m(t_1)}\,(@E)}{B\neg m(t_1)}\,(\mathrm{P1})}{\neg Bm(t_1)}\,(\mathrm{D})}{Bl(box, t_2)}\,(\mathrm{P2})}{@_b Bl(box, t_2)}\,(@I)$$

The temporal progression is pointed out using colors: Blue is what happens at t_1 and magenta at t_2. See Subsection 7.2 for further analysis.

Figure 5. Formalization of incorrect reasoning in the Sally-Anne test (what is the case from the subject's own perspective)

the formula $@_b Bl(box, t_2)$ is derivable using the principles indicated in Figure 5, including the principle of inertia (P1). The principle of inertia is employed as Peter cannot see the content of the box at t_2, but at t_1 he came to believe that the marble was in the box, and this belief is preserved over time to t_2, since he does not believe that an action has been taken to move the marble.

Note that the (*Term*) rule is not employed in the derivations in Figures 4 and 5, and since b is the child Peter's own perspective, there is no shift to a different perspective in these derivations.

The formulas considered in the three tables above can be classified along the two dimensions in Table 4.

Note the pattern in Table 4: The child giving an incorrect answer (lower left quarter) reports what is believed to be the case from the child's own perspective (lower right quarter), and the child does not perform the shift of perspective required to be able to report what is believed to be the case from the second perspective (upper left quarter).

Table 4	The second perspective (the nominal a)	The child's own perspective (the nominal b)
Involves the false statements p and $l(basket, t_2)$	Correct responses cf. Table 1 $@_a Bp$ and $@_a Bl(basket, t_2)$	
Involves the true statements q and $l(box, t_2)$	Incorrect responses cf. Table 2 $@_a Bl(box, t_2)$	True statements cf. Table 3 $@_b Bq$ and $@_b Bl(box, t_2)$

The above "pattern of failure" shows that our logical formalizations in a systematic way model the typical incorrect answers.

7. Where is the origin of mistakes?

As indicated earlier, the child giving an incorrect answer does not perform the shift of perspective required to figure out the correct answer (upper left quarter in Table 4), but instead reports what is believed to be the case from the child's own perspective (lower right quarter in Table 4), namely the formulas $@_b Bq$ and $@_b Bl(box, t_2)$. For example, the formula $@_b Bl(box, t_2)$ says that "Peter believes that the marble is in the box at the time t_2" where t_2 is the time when Peter answers the question. But as shown in Figures 4 and 5, these two formulas are actually derivable using hybrid-logical rules, thus, the incorrect answers can be derived using normatively correct rules.

The fact that the incorrect answers can be derived using logically correct rules suggests that the origin of the mistakes does not lie in the subject's logical reasoning, but rather in a wrong interpretation of the task.[4] This type of error is extensively discussed in the book (Stanovich, 1999), and also in the paper (Stanovich and West, 2000). According to the abstract of (Stanovich and West, 2000), a distinction can be made between four different types of errors taking place when human responses deviate from the performance considered normative, that is, in accordance with a given normative model (logical, statistical, or otherwise):

(1) performance errors,
(2) computational limitations,

[4] Thanks to one of the anonymous reviewers of (Braüner, 2015) for pointing this out.

(3) the wrong norm being applied by the experimenter, and
(4) a different construal of the task by the subject.

In that paper "[...] performance errors represent algorithmic-level problems that are transitory in nature. Nontransitory problems at the algoritmic level that would be expected to recur on a readministration of the task are termed computational limitations", cf. p. 646, where"the algorithmic level" refers to the three levels of description put forward by David Marr and others; see Subsection 7.1. Thus, in the classification of that paper, the notion of a performance error exclusively encompass nonsystematic deviations from the norm (this contrary to some other works, in particular, the more inclusive notion of performance employed in Chomsky's competence/performance distinction, which we shall come back to in Subsection 7.1).

Summing up, the previous considerations indicate that the reasoning errors made by young children and autists belong to the fourth category given in (Stanovich and West, 2000): The subject gives a normatively appropriate response to a problem different from the problem that the experimenter intends the subject to solve.

This diagnosis that a subject's incorrect answer can be traced back to a wrong task interpretation, rather than logical errors in the subject's reasoning, can be further analyzed in terms of the two stages in human reasoning emphasized in (Stenning and van Lambalgen, 2008), namely reasoning *to* and reasoning *from* an interpretation: First a domain of discourse is fixed, together with an interpretation of logical and non-logical expressions, and only after this has been achieved, a set of normatively correct formal rules can be determined. Along similar lines, the book (Stanovich, 1999, p. 99) the chapter on different task construal by the subject, writes that "It is now widely recognized that the evaluation of the normative appropriateness of a response to a particular task is always relative to a particular interpretation of the task." In terms of the reasoning to/from distinction, the origin of the mistakes made by young children and autists seems to be located in the first stage, that is, in the process of interpreting the task, rather than in the second stage, that is, the logical reasoning in accordance with the interpretation obtained in the first stage.

According to Stenning and van Lambalgen (2008, pp. 23–24) what a subject concretely does can be described using a type-token distinction: "The domain mentally constructed while interpreting a discourse is a concrete instance – a token – of a general kind – the type – which determines the logical properties of the token." According to Stenning and van Lambalgen (2008, p. 25) the technical part of reasoning to an interpretation involves the following:[5]:

1. Fixing a formal language. In our case this is done by Definition 2.1.
2. Fixing a semantics for the formal language. This includes a notion of a mathematical representation of the domain, usually called a model, together with a definition of satisfaction, connecting the models to the formal language. Our models are defined by Definition 2.2 and our satisfaction relation $\models$ is defined immediately after Definition 2.2.
3. Fixing a definition of valid arguments in the language. Our definition of valid arguments is embodied in the formulation of Theorem 3.1 (soundness and completeness).

Reasoning from such a fixed interpretation then involves applying rules which are normatively correct according to the interpretation. If the above three steps are taken for granted, it seems most plausible that the origin of the mistakes made by young children and autists lies in the second step, namely in fixing a semantics, more specifically a Kripke model, presumably including only one perspective, namely the subject's own perspective.

7.1. Digging deeper: Competence versus performance

Above we described the two-stage process put forward by Stenning and van Lambalgen (2008), first an interpretation is fixed, and then formal logical rules determining normatively appropriate responses can be given. We now dig one step deeper in the reasoning process: According to

[5] These three items are in (Stenning and van Lambalgen, 2008) described as three successive stages in reasoning to an interpretation, but there is actually an alternative order: One could start by fixing a set of semantic objects, that is, a set of mathematical objects meant for interpretations of formulas, without specifying how the formulas should be built, and after that, one can consider the concrete logical connectives as well as the definition of valid arguments. For example, in the case of classical propositional logic, one might start by stipulating that a formula built using the propositional symbols p, q, r, ... is interpreted as an element of the set of functions $(\{p, q, r, \ldots\} \to \{0, 1\}) \to \{0, 1\}$ and given this initial stipulation, one can independently fix the remaining components of reasoning to an interpretation: A set of connectives equipped with truth-tables (which raises the question of functional completeness) and a definition of valid arguments (where truth-preservation is an obvious requirement).

Stenning and van Lambalgen (2008), the formal rules determined by this process provides the competence model, the ideal norm against which performance must be judged, and a possible performance is described by an algorithm corresponding to this competence model.

The competence/performance distinction goes back to Noam Chomsky: According to Chomsky, linguistic competence is the knowledge of language manifest in a speaker's idealized capacity to produce and to understand an infinite number of sentences, whereas the actual use of language in concrete situations is a matter of linguistic performance (see Chomsky, 1965).These two levels—the competence level and the performance level—are in (Stenning and van Lambalgen, 2008, p. 348) also described in terms of the first two levels of David Marr's three levels of analysis of cognitive systems:

1. The information-processing task as an input-output function.
2. An algorithm which computes that function.
3. The neural implementation of the algorithm. Analogous levels of analysis can be found in several other works of cognitive science[6], see the overview in (Stanovich, 1999, pp. 9–12).

Now, according to Stenning and van Lambalgen (2008, p. 350) an algorithm corresponding to a competence model computes an information-processing task where information is extracted from given data: The input to the algorithm is a set of premises ψ_1, ..., ψ_m encoding the given data, and the output is a conclusion ϕ derivable from the premises using a set of predetermined formal logical rules. We use a turnstile $\vdash$ to denote the derivability relation generated by a set of formal rules, then we have $\psi_1, \ldots, \psi_m \vdash \phi$ and the information-processing algorithm determines *how* the formula ϕ is derived from the formulas $\psi_1, \ldots, \psi_m$.

[6] In fact, such a layering can be found many different places, in particular in computers, for example in the area of programming languages: The denotational semantics of a computer program determines the output of running the program, that is, *what* the program computes, for example a natural number. Usually the denotational semantics of a program is a structure-preserving function between appropriate mathematical structures. On the other hand, an operational semantics specifies *how* the output is computed, that is, it specifies an algorithm that computes the output. An operational semantics is said to be *sound* if it coincides with the denotational semantics in this way, that is, if for any input, it calculates the same output as the denotational semantics. Note that different sound operational semantics compute the same output, but employ different algorithms. See (Winskel, 1993) for more on formal semantics of programming languages.

Note that the derivability relation is—as the name suggests—a relation, not a function.[7]

As mentioned earlier, Stenning and van Lambalgen (2008) gives a logical analysis of the Sally-Anne test, and based on this analysis, the book gives an informal description of an algorithm that carries out the information processing when a correct answer to the test is given.

> In the false-belief task the preprocessing of the data involves understanding of the task instruction, and recruiting information concerning the relation between perception and belief, and of general causal knowledge. Information extraction is performed by an algorithm *simulating* the temporal evolution of the initial model up to the time that Maxi [in the present paper: Sally] returns to the room and starts looking for the chocolate [in the present paper: the marble], using the principle of inertia applied to both beliefs and the world
>
> (Stenning and van Lambalgen, 2008, p. 354, italics as in original)

We remark that the relation between perception and belief referred to in this quotation is obtained using principle (9.2) in (Stenning and van Lambalgen, 2008), which is identical to principle (P1), namely $S\phi \rightarrow B\phi$, considered in Section 5 of the present paper, and the principle of inertia referred to in the quotation is principle (9.11) in (Stenning and van Lambalgen, 2008), which is reminiscent of principle (P2), namely $Bl(i,t) \wedge \neg Bm(t) \rightarrow Bl(i,t+1)$, considered in Section 5.

7.2. Competence versus performance in our formalizations

We shall in what follows give an informal algorithmic analysis of the information processing in the Sally-Anne test like the above quoted from (Stenning and van Lambalgen, 2008), but instead based on our analysis and formalization as the derivation in Figure 3, where the perspective shift is explicitly represented by an instance of the (Term) rule. Thus, we take the competence/performance distinction for granted, where the competence model is provided by our hybrid-logical proof-rules, and where a possible performance is described by an (informal) algorithm that applies the hybrid-logical proof-rules to derive an answer.

Now, we actually already gave a very informal algorithmic analysis of the Sally-Anne task in the second paragraph of Section 5, but we shall

[7] Contrary to the case outlined in footnote 6 with the formal semantics of deterministic computer programs.

now be more precise, and explain the correspondence to the derivation in Figure 3, in particular, how the derivation is actually derived. If $\vdash$ denotes the derivability relation generated by our hybrid-logical rules, then the derivation in Figure 3 is a witness of the following relationship

$$@_a Sl(basket, t_0), @_a S\neg m(t_0), @_a \neg Sm(t_1) \vdash @_a Bl(basket, t_2)$$

where the nominal a stands for Sally. So the antecedent formulas are the undischarged assumptions of the derivation in Figure 3 and the succedent formula is the conclusion of the derivation.

The method of reasoning in natural deduction systems is called "forward" reasoning: You start with assumptions, and using the rules, you step by step build derivations of new formulas.[8] Thus, to give an algorithmic analysis of the derivation in Figure 3, we need to specify how the derivation was built, in particular, we need to specify how the subderivation delimited by the (*Term*) rule was built. This is the subderivation that formalizes the experimental subject's hypothetical reasoning from Sally's perspective. It seems plausible that the subject, whom we called Peter, step by step, in the temporal order of the narrative of the task, derives more and more information, that is, the formulas in the derivation are calculated in temporal order: First what is the case at t_0, then what is the case at t_1, and finally what is the case at t_2.

Building the derivation in temporal order fits with our initial informal analysis in the second paragraph of Section 5, which we recapitulate below, slightly reformulated, and with formulas inserted, corresponding to formulas in Figure 3. The temporal progression in Peter's hypothetical reasoning is indicated using colors: Red is what happens from Sally's perspective at t_0, blue what happens at t_1, and magenta what happens at t_2 (the same color codes are used in Figure 3). Peter's reasoning before he switches to Sally's perspective, and after he has switched back again, is black.

> Peter imagines himself being Sally, and he reasons as follows: At the time t_0 when Sally leaves, she sees that the marble is in the basket (formalization $Sl(basket, t_0)$) so she believes the marble is in the basket (formalization $Bl(basket, t_0)$), and she sees no action to move the marble (formalization $S\neg m(t_0)$), so when she

[8] This is contrary to for example tableau systems which are backward reasoning systems since you explicitly start with a formula you want to prove, and then you try to build a proof of it using tableau rules.

> is away at t_1, she still believes that the marble is in the basket (formalization $Bl(\mathit{basket}, t_1)$). She does not see that the marble is moved at t_1 (formalization $\neg Sm(t_1)$), so she does not believe that it is moved (formalization $\neg Bm(t_1)$), and hence, at t_2 after she has returned, she still believes that the marble is in the basket (formalization $Bl(\mathit{basket}, t_2)$). Therefore, Peter concludes that Sally at t_2 believes that the marble is in the basket (formalization $@_a Bl(\mathit{basket}, t_2)$).

Note that the word "still" occurs two times above—these are the two places where the principle of inertia, principle (P2), are applied, effecting one step forward in time. To see the connection to the derivation in Figure 3, let us zoom in on the second application of the principle of inertia, which happens in the following excerpt.

> [...] at t_1, she [...] believes that the marble is in the basket (formalization $Bl(\mathit{basket}, t_1)$). [...] she does not believe that it is moved (formalization $\neg Bm(t_1)$), and hence, at t_2 [...] she still believes that the marble is in the basket (formalization $Bl(\mathit{basket}, t_2)$).

Formally, this reasoning step is carried out by the rightmost instance of the (P2) rule in Figure 3, that is, the following instance:

$$\frac{Bl(\mathit{basket}, t_1) \qquad \neg Bm(t_1)}{Bl(\mathit{basket}, t_2)}\,(\mathrm{P2})$$

Of course, it is a choice to derive the answer formula in Figure 3 in a particular way, but in our concrete case it seems cognitively plausible that formulas are derived in the temporal order of the narrative, for example, it only seems relevant to derive the blue formula $\neg Bm(t_1)$, formalizing that Sally does not believe that the marble is moved at t_1, *after* having derived the red formulas $Bl(\mathit{basket}, t_0)$ and $\neg Bm(t_0)$, implying that the marble is also in the basket at t_1, formalized by the blue formula $Bl(\mathit{basket}, t_1)$.

Building the derivation in Figure 3 in temporal order is in line with what in the above quotation from (Stenning and van Lambalgen, 2008) is referred to as simulating "the temporal evolution of the initial model". In fact our algorithmic analyses above appears to be in accordance with the simulation-theory view of theory of mind, cf. the introductory section of the present paper, since the temporarily progressing simulation delimited by the (*Term*) rule, takes place from another perspective than the subject's own perspective.

Now, the main concern of the present paper is what goes wrong when incorrect answers are given. According to our analysis in Section 6, the child giving an incorrect answer to the Sally-Anne test reports what is believed to be the case from the child's own perspective, formalized by the derivation in Figure 5. The derivation in Figure 5 is a witness of the relationship

$$b, @_b Sl(box, t_1), @_b S\neg m(t_1) \vdash @_b Bl(box, t_2)$$

where the nominal b stands for Peter, the child doing the reasoning, thus, the antecedent formulas are the undischarged assumptions in Figure 5 and the succedent formula is the conclusion.

Building on the analysis of the incorrect answer in Section 6, we shall now give an informal algorithmic analysis of the information processing in the incorrect reasoning, corresponding to building the derivation in Figure 5 in temporal order. Blue is what happens at t_1 and magenta what happens at t_2 (same color codes as above and in Figure 5).

> Remembering the information that was available to him at the earlier stage t_1, Peter reasons that at the time t_1, he saw that the marble is in the box (formalization $Sl(box, t_1)$) so he believed that the marble is in the box (formalization $Bl(box, t_1)$), and he saw no action to move the marble (formalization $S\neg m(t_1)$), so at t_2, he still believes that the marble is in the box (formalization $Bl(box, t_2)$). Therefore, Peter at t_2 believes that the marble is in the box (formalization $@_b Bl(box, t_2)$)

Also, note that there is only one perspective involved here, namely the subject's own perspective, denoted by the nominal b, and hence no perspective shift.

8. Realist bias

In Table 2, and the lower left quarter of Table 4, the incorrect answers to the Smarties and Sally-Anne tests are summed up, where the experimental subjects report the belief of their own, rather than that of someone else, as required to give the correct response. This systematic tendency subjects have to report what they themselves believe of reality, rather than what others believe of reality, is similar to the bias in adults' mindreading ability which some authors call a *realist bias*, cf. (Mitchell et al., 1996), or *curse of knowledge*, cf. (Birch and Bloom, 2007). In the present context, this realist bias amounts to reporting what is the case from the subject's own perspective, rather than what can be deduced to be the case from another person's perspective.

In the paper (Birch and Bloom, 2007) a study is reported in which the Sally-Anne scenario is modified such there are four containers instead of two, and instead of judging where Sally will look, the experimental subjects rated the probability that she will look in the four containers, that is, for each of the four containers, the subjects rated the probability that she will look in the container in question. In some trials, the subjects knew in which container the marble really was, like in the original version of the Sally-Anne test where the subjects know that the marble had been moved to the box, but on other trials the subjects only knew that it had been moved to another container than where it was initially. The study reported that in the case where the subjects knew the real location of the marble, they judged it more probable that Sally would search in the location where the marble in fact was, compared to the case where they did not know the location of the marble.

The crucial point in the modified Sally-Anne scenario is that to determine the correct answer, it is not relevant whether or not the subject know where the marble really is, that is, the subject's own knowledge about the real location is not relevant—what *is* relevant is Sally's knowledge, which is the same in both cases. In particular, whether the subject knows the actual location of the marble or the subject does not know the actual location, this piece of information is obviously not included in Sally's knowledge. This is in line with the fact that the actual location of the marble in the Sally-Anne test—the box—is not even mentioned in the formalization of the correct answer in Figure 3. Similarly, in the Smarties test, information about the actual content of the tube—pencils—is not involved in figuring out the correct answer, to be more precise, the propositional symbol q standing for “The tube contains pencils” does not even occur in the formalization of the correct answer in Figure 2.

The experimental subjects in (Birch and Bloom, 2007) are adults, but it is suggested that the problems children under four have to pass false-belief tests to some extent should be accounted for in terms of an exaggerated curse-of-knowledge bias—not only in terms of a limited concept of belief, or more generally, a limited concept of mental state, the latter being a typical explanation in existing literature.

In their earlier paper (Birch and Bloom, 2003), the authors of (Birch and Bloom, 2007) reported experiments with three to five year old children, where the experiments showed that three to four year old children were particularly susceptible to the curse-of-knowledge bias, compared to older children of the age five. With reference to these experiments and other works, (Birch and Bloom, 2007) calls for more experiments, where different false-belief tests are used to clarify the role that the curse-of-knowledge bias play in the mental-state reasoning of children.

9. Related work

The approach taken in the present paper, based on our earlier papers (Braüner, 2013, 2014), is to model the reasoning in false-belief tests from the perspective of the subject doing the reasoning, to be more precise, the subject's reasoning is modeled syntactically. This is also the approach taken by the earlier mentioned works (Stenning and van Lambalgen, 2008) and (Arkoudas and Bringsjord, 2008), applying syntactic machinery different from ours, respectively the procedural evaluation mechanism of logic programming and a proof-system for a many-sorted first-order modal logic. Note that this is not the same thing as explicitly formalizing the perspective shift required to pass a false-belief test (we do that in the present paper, and we did it in (Braüner, 2013, 2014) as well, but neither (Stenning and van Lambalgen, 2008) nor (Arkoudas and Bringsjord, 2008) model the perspective shift explicitly).

As described earlier, the book (Stenning and van Lambalgen, 2008) analyze the reasoning taking place in a number of false-belief tests. The book analyze the reasoning taking place when giving the correct response, as well as what goes wrong when an incorrect response is given. In this connection the book discusses four main psychological theories of autism: The theory of mind deficit theory (described in the first section of the present paper), the affective foundation theory, the weak central coherence theory, and the executive function deficit theory. We note that the book argue that the executive function deficit theory is more fundamental than the theory of mind deficit theory. Rather than being an explanation of autism, the book sees the theory of mind deficit theory as "an important label for a problem that needs a label" (cf. Stenning and van Lambalgen, 2008, p. 243). Now, very briefly, executive function is the ability to plan and control a sequence of actions with the aim of obtaining a goal. If the executive function deficit theory is taken as the basis, then it appears appropriate to try to formalize the reasoning in a false-belief task in some sort of non-monotonic logic, which is what the book (Stenning and van Lambalgen, 2008) do. On the other hand, if the theory of mind deficits theory is taken as the basis, then we find that it is appropriate to use hybrid logic together with hybrid-logical proof-theory. Thus, a decisive difference between our work and (Stenning and van Lambalgen, 2008) is which psychological theory is taken as the basis of the logical analysis.

Another approach than modeling reasoning from the perspective of the subject doing the reasoning, is to model the reasoning from a global perspective, that is, from the perspective of the modeler. This approach has been taken in a number of works, for example (Bolander, 2018) which uses a version of dynamic epistemic logic to model the reasoning in the the Sally-Anne test and other false-belief tests. The main feature of epistemic logic is that reasoning is modeled with Kripke structures characterizing the uncertainty of agents: There is an accessibility relation for each agent, and two possible worlds, that is, doxastic states, are related if and only is the agent cannot distinguish between the states on the basis of the information available to the agent. Epistemic logic can model a static state of affairs, like at a specific time, Sally believes that the marble to be in the basket. In dynamic epistemic logic further machinery has been added that can update a model when an action has taken place, for example when Anne has moved the marble from the basket to the box. From a mathematical point of view, epistemic logic is very elegant, but one drawback of epistemic logic is that belief (or knowledge) is closed under logical consequence, that is, $B\psi$ can be derived from $\phi \rightarrow \psi$ and $B\phi$, which at least for human agents is implausible (when the modal operator stands for knowledge, this is called *logical omniscience*). See Section 5 of (Verbrugge, 2009) for a general discussion of the problems with epistemic logic as a model for human social cognition.

A global perspective is also taken in the paper (van Ditmarsch and Labuschagne, 2007), which uses a version of epistemic logic to model examples of beliefs that agents may have about other agents' beliefs. One example is what in the paper is called an autistic agent that always believes that other agents have the same beliefs as the agent's own beliefs. This is modelled by equipping each agent with a preference relations between states, where an agent prefers one state over another if the agent considers it more likely. The paper show that these beliefs are frame-characterizable by formulas of epistemic logic.

There are also computational cognitive models of false-belief tasks, for example (Arslan et al., 2013), which models the gradual development in false-belief reasoning using the ACT-R cognitive architecture.

10. Further work

We would like to lift the line of work presented here to what is called *second-order* false-belief tests, where the subject has to realize that someone can hold a false belief about someone's belief about a state of affair in the world, in comparison to the *first-order* case, where the child "just" has to realize that someone can hold a false belief about a state of affair in the world, which is what we have considered previously in the present paper. In (Braüner et al., 2016) we give a formalization of a second-order version of the Sally-Anne test, which we use to argue in favour of a view on second-order false-beliefs, which says that going from first-order to second-order false-belief understanding constitititutes a specific *conceptual change* (in contrast to the *complexity-only* view, which says that it is a matter of enhanced general cognitive capacities like working memory).

An interesting phenomenon crops up when going from first-order to second-order false-belief understanding: In the first-order case there is one obvious incorrect answer, namely the answer at the zero-order level where the subjects reports their own belief about world facts, but in some second-order tests there is more than one obvious incorrect answer, in particular, there is an incorrect answer at the zero-order level as well as an incorrect answer at the first-order level. We plan to investigate this with our logical tools. See (Braüner et al., 2020) for a comparison between four different second-order false-belief tasks. These four tasks play a crucial role in the empirical study of false-belief reasoning by autistic children reported in the PhD dissertation (Polyanskaya, 2019).

Acknowledgements. The present paper is a journal version of (Braüner, 2015). Thanks to Patrick Blackburn for many useful comments on a preliminary version of the present paper. Also, thanks to Irina Polyanskaya for many discussions on the issues of the present paper. The author acknowledges the funding received from the Velux Foundation for the project *Hybrid-Logical Proofs at Work in Cognitive Psychology* (VELUX 33305).

References

Areces, C., and B. ten Cate, 2007, "Hybrid logics", pages 821–868 in P. Blackburn, J. van Benthem and F. Wolter (eds.), *Handbook of Modal Logic*, Elsevier.

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

Arkoudas, K., and S. Bringsjord, 2009, "Propositional attitudes and causation", *International Journal of Software and Informatics* 3: 47–65.

Arslan, B., N. Taatgen, and R. Verbrugge, 2013, "Modeling developmental transitions in reasoning about false beliefs of others", pages 77–82 in *Proceedings of the 12th International Conference on Cognitive Modeling*, Ottawa: Carleton University.

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

Baron-Cohen, S., A. M. Leslie and U. Frith, "Does the autistic child have a 'theory of mind'?", *Cognition* 21: 37–46.

Birch, S. A. J., and P. Bloom, 2003, "Children are cursed: An asymmetric bias in mental state attribution", *Psychological Science* 14: 283–286.

Birch, S. A. J., and P. Bloom, 2007, "The curse of knowledge in reasoning about false beliefs", *Psychological Science* 18: 382–386.

Blackburn, P., T. Bolander, T. Braüner and K. F. Jørgensen, 2017, "Completeness and termination for a Seligman-style tableau system", *Journal of Logic and Computation* 27: 81–107.

Bolander, T., 2018, "Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic", pages 207–236 in *Jaakko Hintikka on Knowledge and Game-Theoretical Semantics*, Springer International Publishing.

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

Braüner, T., 2013, pages 186–195, "Hybrid-logical reasoning in false-belief tasks", in B. C. Schipper (ed.), *Proceedings of Fourteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*. Available at http://tark.org.

Braüner, T., 2014, "Hybrid-logical reasoning in the Smarties and Sally-Anne tasks", *Journal of Logic, Language and Information* 23: 415–439. Revised and extended version of (Braüner, 2013).

Braüner, T., 2015, "Hybrid-logical reasoning in the Smarties and Sally-Anne tasks: What goes wrong when incorrect responses are given?", pages 273–278 in *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, Pasadena, California: Cognitive Science Society.

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

Braüner, T., P. Blackburn and I. Polyanskaya, 2020, "Being deceived: Information asymmetry in second-order false belief tasks", *Topics in Cognitive Science* (in press).

Chomsky, N., *Aspects of the theory of syntax*, MIT Press, 1965.

Conan Doyle, A., "The Musgrave Ritual", in *The Memoirs of Sherlock Holmes*, Harper Bros., 1894.

Gallese, V., and A. Goldman, 1998, "Mirror neurons and the simulation theory of mind-reading", *Trends in Cognitive Sciences* 2: 493–501.

Gopnik, A., and J. W. Astington, 1988, "Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction", *Child Development* 59: 26–37.

Gordon, R. M., 1986, "Folk psychology as simulation", *Mind and Language* 1: 158–171.

Gordon, R. M., 2009, "Folk psychology as mental simulation", in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Stanford University. On-line encyclopedia article available at http://plato.stanford.edu/entries/folkpsych-simulation.

Jørgensen, K. F., P. Blackburn, T. Bolander and T. Braüner, 2016, "Synthetic completeness proofs for Seligman-style tableau systems", pages 302–321 in A. M. Lev Beklemishev, S. Demri (ed.), *Proceedings of Advances in Modal Logic 2016*, volume 11 of *Advances in Modal Logic*, College Publications.

Mitchell, P., E. J. Robinson, J. E. Isaacs and R. M. Nye, 1996, "Contamination in reasoning about false belief: an instance of realist bias in adults but not children", *Cognition* 59: 1–21.

Polyanskaya, I., 2019, "Second-order false belief reasoning by children with autism: A correlation and training study", PhD thesis, Department of People and Technology, Roskilde University, Denmark.

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

Prawitz, D., 2005, “Logical consequence from a constructivist point of view”, pages 671–695 in S. Shapiro (ed.), *The Oxford Handbook of Philosophy of Mathematics and Logic*, Oxford University Press.

Rips, L. J., 1994, *The Psychology of Proof: Deductive Reasoning in Human Thinking*, MIT Press.

Rips, L. J., 2008, “Logical approaches to human deductive reasoning”, , pages 187–205 in J. E. Adler and L. J. Rips (eds.), *Reasoning: Studies of Human Inference and Its Foundations*, Cambridge University Press.

Seligman, J., “The logic of correct description”, pages 107–135 in M. de Rijke (ed.), *Advances in Intensional Logic*, volume 7 of “Applied Logic Series”, Kluwer, 1997.

Stanovich, K. E., 1999, *Who is Rational? Studies of Individual Differences in Reasoning*, Lawrence Erlbaum.

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

Stanovich, K. E., and R. F. West, 2000, “Individual differences in reasoning: Implicationations for the rationality debate”, *Behavioral and Brain Sciences* 23: 645–726.

van Ditmarsch, H., and W. Labuschagne, 2007, “My beliefs about your beliefs – a case study in theory of mind and epistemic logic”, *Synthese* 155: 191–209.

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

Wellman, H. M., D. Cross and J. Watson, 2001, “Meta-analysis of theory-of-mind development: The truth about false-belief”, *Child Development*, 72: 655–684.

Winskel, G., 1993, *The Formal Semantics of Programming Languages: An Introduction*, Foundation of computing series, MIT Press.

Torben Braüner
Department of People and Technology
Roskilde University, Denmark
torben@ruc.dk