Hybrid-Logical Reasoning in the Smarties and Sally-Anne Tasks

Braüner, Torben

Hybrid-Logical Reasoning in
the Smarties and Sally-Anne Tasks[*]

Torben Braüner

Programming, Logic and Intelligent Systems Research Group

Roskilde University

P.O. Box 260, DK-4000 Roskilde, Denmark

torben@ruc.dk

Abstract

The main aim of the present paper is to use a proof system for hybrid modal logic to formalize what are called false-belief tasks in cognitive psychology, thereby investigating the interplay between cognition and logical reasoning about belief. We consider two different versions of the Smarties task, involving respectively a shift of perspective to another person and to another time. Our formalizations disclose that despite this difference, the two versions of the Smarties task have exactly the same underlying logical structure. We also consider the Sally-Anne task, having a more complicated logical structure, presupposing a "principle of inertia" saying that a belief is preserved over time, unless there is belief to the contrary.

Keywords: Modal logic, hybrid logic, false-belief tasks, Theory of Mind

1   Introduction

In the area of cognitive psychology there is a reasoning task called the *Smarties task*. The following is one version of this task.

> A child is shown a Smarties tube where unbeknownst to the child the Smarties have been replaced by pencils.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: "If your mother comes into the room and we show this tube to her, what will she think is inside?"

It is well-known from experiments that most children above the age of four correctly say "Smarties" (thereby attributing a false belief to the mother) whereas younger children say "Pencils" (what they know is inside the tube). For autistic[1] children the cutoff age is higher than four years, which is one reason to the interest in the Smarties task.

The Smarties task is one out of a family of reasoning tasks called *false-belief tasks* showing the same pattern, that most children above four answer correctly, but autistic children have to be older. This was first observed in the paper [6] in connection with another false-belief task called the *Sally-Anne task*. Starting with the authors of that paper, many researchers in cognitive psychology

---

[1]Autism is a psychiatric disorder with the following three diagnostic criteria: 1. Impairment in social interaction. 2. Impairment in communication. 3. Restricted repetitive and stereotyped patterns of behavior, interests, and activities. For details, see *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV)*, published by the American Psychiatric Association.

have argued that there is a link between autism and a lack of what is called *theory of mind*, which is a person's capacity to ascribe mental states to oneself and to others, for example beliefs. For a very general formulation of the theory of mind deficit hypothesis of autism, see the book [5]. The results of false-belief tasks are robust under many different variations, for example across various countries and various task manipulations, as shown in the meta-analysis [37] involving 178 individual false-belief studies and more than 4000 children. Beside the research considering theory of mind at a cognitive level, such as in connection with false-belief tasks, there is also an extensive research from the point of view of neuropsychology, for example the paper [16], that suggests an explanation of theory of mind in terms of 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.

Giving a correct answer to the Smarties task involves a shift of perspective to another person, namely the mother. You have to put yourself in another person's shoes, so to speak. Since the capacity to take another perspective is a precondition for figuring out the correct answer to the Smarties task and other false-belief tasks, 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 critical overview of these arguments, see the book [33] by Keith Stenning and Michiel van Lambalgen. The books [33] and [5] not only consider theory of mind at a cognitive level, but they also discuss it from a biological point of view.

In a range of works van Lambalgen and co-authors have given a detailed logical analysis (but not a full formalization) of the reasoning taking place in the Smarties task and other false-belief tasks in terms of non-monotonic closed world reasoning as used in logic programming, see in particular [33]. The analysis of the Sally-Anne and Smarties tasks of [33], subsections 9.4.1–9.4.4, makes use of a modality[2] B for belief satisfying two standard modal principles. The first principle is $B(\phi \to \psi) \to (B\phi \to B\psi)$ (principle (9.5) at page 251 in [33]), which usually is called axiom K. The second principle is the rule called necessitation, that is, from ϕ derive $B\phi$ (this rule is not mentioned explicitly in [33], but in the Sally-Anne case, the necessitation rule is implicitly applied to clauses (9.12) and (9.13) at page 253, and in the Smarties case, it is implicitly applied to clause (9.22) at page 256). Axiom K together with the necessitation rule imply that belief is closed under logical consequence, that is, $B\psi$ can be derived from $\phi \to \psi$ and $B\phi$, which at least for human agents is implausible (when the modal operator stands for knowledge, this is called logical omniscience).[3]

The papers [2] and [3] by Arkoudas and Bringsjord give a formalization of the Sally-Anne task, but no other false-belief tasks. The papers have several aims, one of them is described as follows.

> One intended contribution of our present work is ... to provide a formal model of false-belief attributions, and, in particular, a description of the logical competence of an agent capable of passing a false-belief task. It addresses questions such as the following: What sort of principles is it plausible to assume an agent has to deploy in order to be able to succeed on a false-belief task? What is the depth and complexity of the required reasoning? Can such reasoning be automated, and if so, how? ([2], p. 18)

The papers specifiy a number of axioms and proof-rules formulated in a many-sorted first-order modal logic, and it is briefly described how the reasoning in the Sally-Anne task has been implemented in an interactive theorem prover using this machinery (but the papers do not explicitly give a full formalization). The paper [3] is an extended version of [2], containing a section discussing

---

[2]Strictly speaking, the modality B in [33] is not formalized in terms of modal logic, but in terms of what is called event calculus, where B is a predicate that can take formulas as arguments.

[3]The observation that [33] applies axiom K and (implicitly) the necessitation rule, raises the following question: How *could* the logical analyses of the Smarties and Sally-Anne tasks in [33] be turned into full formalizations, that is, fully formal proofs in some well-defined proof-system? The book [33] puts much emphasis on applying the closed world reasoning mechanism of logic programming, that is, the standard procedural evaluation mechanism of Horn clauses, extended with the metalinguistic predicate negation as failure (classical negation is not expressible using Horn clauses), and this would in a principled way have to be combined with machinery like axiom K and the necessitation rule, stemming from Hilbert-style axiom systems, which is a very different type of reasoning.

how to encode the system in (non-modal) many-sorted first-order logic. The papers describe how their axioms and proof-rules are taylormade to avoid logical omniscience, cf. page 22 in [2]. The proof-rules employed in the papers do not explicitly formalize the perspective shift required to pass the Sally-Anne task.

In the present paper we give a logical analysis of the perspective shift required to give a correct answer to the Smarties and Sally-Anne tasks, and we demonstrate that these tasks can be fully formalized in a hybrid-logical proof system not assuming principles implying logical omniscience, namely the natural deduction system described in Chapter 4 of the book [11], and the paper [10] as well. Beside not suffering from logical omniscience, why is a *natural deduction* system for *hybrid modal logic* appropriate to this end?

- The subject of proof-theory is the notion of proof and formal, that is, symbolic, systems for representing proofs. Formal proofs built according to the rules of proof systems can be used to represent—describe the structure of—mathematical arguments as well as arguments in everyday human practice. Beside giving a way to distinguish logically correct arguments from incorrect ones, proof systems also give a number of ways to characterize the structure of arguments. Natural deduction style proofs are meant to formalize the way human beings actually reason, so natural deduction is an obvious candidate when looking for a proof system to formalize the Smarties task in.

- In the standard Kripke semantics for modal logic, the truth-value of a formula is relative to points in a set, that is, a formula is evaluated "locally" at a point, where points usually are taken to represent possible worlds, times, locations, epistemic states, persons, states in a computer, or something else. Hybrid logics are extended modal logics where it is possible to directly refer to such points in the logical object language, whereby locality can be handled explicitly, for example, when reasoning about time one can formulate a series of statements about what happens at specific times, which is not possible in ordinary modal logic. Thus, when points in the Kripke semantics represent local perspectives, hybrid-logical machinery can handle explicitly the different perspectives in the Smarties task.

For the above reasons, we have been able to turn our informal logical analysis of the Smarties and Sally-Anne tasks into fully formal hybrid-logical natural deduction proofs closely reflecting the shift between different perspectives.

The natural deduction system we use for our formalizations is a modified version of a natural deduction system for a logic of situations similar to hybrid logic, originally introduced in the paper [32] by Jerry Seligman. The modified system was introduced in the paper [10], and later on considered in Chapter 4 of the book [11], both by the present author. In what follows we shall simply refer to the modified system as Seligman's system. Very recently a tableau system has been developed along similar lines, see [8].

Now, Seligman's natural deduction system allows any formula to occur in it, which is different from the most common proof systems for hybrid logic that only allow formulas of a certain form called satisfaction statements. This is related to a different way of reasoning in Seligman's system, which captures particularly well the reasoning in the Smarties and Sally-Anne tasks. We prove a completeness result which also says that Seligman's system is analytic, that is, we prove that any valid formula has a derivation satisfying the subformula property. Analyticity guarentees that any valid argument can be formalized using only subformulas of the premises and the conclusion. The notion of analyticity goes back to G.W. Leibniz (1646–1716) who called a proof analytic if and only if the proof is based on concepts contained in the proven statement, the main aim being to be able to construct a proof by an analysis of the result, cf. [4].

The present paper is structured as follows. In the second section we recapitulate the basics of hybrid logic, readers well-versed in hybrid logic can safely skip this section. In the third section we introduce Seligman's natural deduction system for hybrid logic and in the fourth section we give a first example of reasoning in this system. In the fifth and sixth sections we formalize two versions of the Smarties task using this system, and in the seventh section we formalize the Sally-Anne task. In the eight section there are some brief remarks on other work, and in the final section

some remarks on further work. In the appendix we prove the above mentioned completeness result, which also demonstrates analyticity.

2   Hybrid logic

The term "hybrid logic" covers a number of logics obtained by adding further expressive power to ordinary modal logic. The history of what now is known as hybrid logic goes back to the philosopher Arthur Prior's work in the 1960s. See the handbook chapter [1] for a detailed overview of hybrid logic. See the book [11] on hybrid logic and its proof-theory.

The most basic hybrid logic is obtained by extending ordinary modal logic with *nominals*, which are propositional symbols of a new sort. In the Kripke semantics a nominal is interpreted in a restricted way such that it is true at exactly one point. If the points are given a temporal reading, this enables the formalization of natural language statements that are true at exactly one time, for example

> it is five o'clock May 10th 2007

which is true at the time five o'clock May 10th 2007, but false at all other times. Such statements cannot be formalized in ordinary modal logic, the reason being that there is only one sort of propositional symbol available, namely ordinary propositional symbols, which are not restricted to being true at exactly one point.

Most hybrid logics involve further additional machinery than nominals. There is a number of options for adding further machinery; here we shall consider a kind of operator called *satisfaction operators*. The motivation for adding satisfaction operators is to be able to formalize a statement being true at a particular time, possible world, or something else. For example, we want to be able to formalize that the statement "it is raining" is true at the time five o'clock May 10th 2007, that is, that

> at five o'clock May 10th 2007 it is raining.

This is formalized by the formula $@_a r$ where the nominal a stands for "it is five o'clock May 10th 2007" as above and where r is an ordinary propositional symbol that stands for "it is raining". It is the part $@_a$ of the formula $@_a r$ that is called a satisfaction operator. In general, if a is a nominal and ϕ is an arbitrary formula, then a new formula $@_a \phi$ can be built (in some literature the notation $a : \phi$ is used instead of $@_a \phi$). A formula of this form is called a *satisfaction statement*. The formula $@_a \phi$ expresses that the formula ϕ is true at one particular point, namely the point to which the nominal a refers. Nominals and satisfaction operators are the most common pieces of hybrid-logical machinery, and are what we need for the purpose of the present paper.

In what follows we give the formal syntax and semantics of hybrid logic. It is assumed that a set of ordinary propositional symbols and a countably infinite set of nominals are given. The sets are assumed to be disjoint. The metavariables p, q, r, ... range over ordinary propositional symbols and a, b, c, ... range over nominals. Formulas are defined by the following grammar.

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

The metavariables ϕ, ψ, θ, ... range over formulas. Negation is defined by the convention that $\neg \phi$ is an abbreviation for $\phi \to \bot$. Similarly, $\Diamond \phi$ is an abbreviation for $\neg \Box \neg \phi$.

Definition 2.1 *A* model *for hybrid logic is a tuple* $(W, R, \{V_w\}_{w \in W})$ *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 to each ordinary propositional symbol assigns an element of $\{0, 1\}$.*

The pair (W, R) is called a *frame*. Note 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 to each nominal assigns an element 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.

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

By convention $\mathfrak{M}, g \models \phi$ means $\mathfrak{M}, g, w \models \phi$ for every element w of W and $\mathfrak{M} \models \phi$ means $\mathfrak{M}, g \models \phi$ for every assignment g. A formula ϕ is *valid* if and only if $\mathfrak{M} \models \phi$ for any model $\mathfrak{M}$.

3   Seligman's system

In this section we introduce Seligman's natural deduction systems for hybrid logic. Before defining the system, we shall sketch the basics of natural deduction. Natural deduction style derivation rules for ordinary classical first-order logic were originally introduced by Gerhard Gentzen in [18] and later on developed much further by Dag Prawitz in [26, 27]. See [35] for a general introduction to natural deduction systems. With reference to Gentzen's work, Prawitz made the following remarks on the significance of natural deduction.

> ...the essential logical content of intuitive logical operations that can be formulated in the languages considered can be understood as composed of the atomic inferences isolated by Gentzen. It is in this sense that we may understand the terminology *natural* deduction.
>
> Nevertheless, Gentzen's systems are also natural in the more superficial sense of corresponding rather well to informal practices; in other words, the structure of informal proofs are often preserved rather well when formalized within the systems of natural deduction. ([27], p. 245)

Similar views on natural deduction are expressed many places, for example in a textbook by Warren Goldfarb.

> What we shall present is a system for *deductions*, sometimes called a system of *natural deduction*, because to a certain extent it mimics certain natural ways we reason informally. In particular, at any stage in a deduction we may introduce a new premise (that is, a new supposition); we may then infer things from this premise and eventually eliminate the premise (*discharge* it). ([20], p. 181)

Basically, what is said by the second part of the quotation by Prawitz, and the quotation by Goldfarb as well, is that the structure of informal human arguments can be described by natural deduction derivations.

Of course, the observation that natural deduction derivations often can formalize, or mimic, informal reasoning does not itself prove that natural deduction is the mechanism underlying human deductive reasoning, that is, that formal rules in natural deduction style are somehow built into the human cognitive architecture. However, this view is held by a number of psychologists, for example Lance Rips in the book [28], where he provides experimental support for the claim.

> ...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. ([28], p. x)

This is the main claim of the "mental logic" school in the psychology of reasoning. See also [29] which is a reproduction of some chapters from the book [28]. The major competitor of the "mental logic" school is the "mental models" school, claiming that the mechanism underlying human reasoning is the construction of models, rather than the application of topic-neutral formal rules, see [22].

We have now given a brief motivation for natural deduction and proceed to a formal definition. A *derivation* in a natural deduction system has the form of a finite tree where the nodes are labelled with formulas such that for any formula occurrence ϕ in the derivation, either ϕ is a leaf of the derivation or the immediate successors of ϕ in the derivation are the premises of a rule-instance which has ϕ as the conclusion. In what follows, the metavariables $\pi, \tau, \ldots$ range over derivations. A formula occurrence that is a leaf is called an *assumption* of the derivation. The root of a derivation is called the *end-formula* of the derivation. All assumptions are annotated with numbers. An assumption is either *undischarged* or *discharged*. If an assumption is discharged, then it is discharged at one particular rule-instance and this is indicated by annotating the assumption and the rule-instance with identical numbers. We shall often omit this information when no confusion can occur. 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 derivation belong to the same *parcel* 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 at the same rule-instance. Thus, in this terminology rules discharge parcels. We shall make use of the standard notation

$$\begin{array}{c} [\phi^r] \\ \vdots\, \pi \\ \psi \end{array}$$

which means a derivation π where ψ is the end-formula and $[\phi^r]$ is the parcel consisting of all undischarged assumptions that have the form ϕ^r.

We shall make use of the following conventions. The metavariables $\Gamma, \Delta, \ldots$ range over sets of formulas. A derivation π is called a *derivation of* ϕ if the end-formula of π is an occurrence of ϕ, and moreover, π is called a *derivation from* Γ if each undischarged assumption in π is an occurrence of a formula in Γ (note that numbers annotating undischarged assumptions are ignored). If there exists a derivation of ϕ from $\emptyset$, then we shall simply say that ϕ is *derivable*.

A typical feature of natural deduction is that there are two different kinds of rules for each connective; there are rules called introduction rules which introduce a connective (that is, the connective occurs in the conclusion of the rule, but not in the premises) and there are rules called elimination rules which eliminate a connective (the connective occurs in a premiss of the rule, but not in the conclusion). Introduction rules have names in the form $(\ldots I \ldots)$, and similarly, elimination rules have names in the form $(\ldots E \ldots)$.

Now, Seligman's natural deduction system is obtained from the rules given in Figure 1 and Figure 2. We let $\mathbf{N}'_{\mathcal{H}}$ denote the system thus obtained. The system $\mathbf{N}'_{\mathcal{H}}$ is taken from [10] and Chapter 4 of [11] where it is shown to be sound and complete wrt. the formal semantics given in the previous section. As mentioned earlier, this system is a modified version of a system originally introduced in [32]. The system of [32] was modified in [10] and [11] with the aim of obtaining a desirable property called closure under substitution, see Subsection 4.1.1 of [11] for further explanation.

The way of reasoning in Seligman's system is different from the way of reasoning in most other proof systems for hybrid logic[4]. This in particular applies to rule (*Term*), which is very different from other rules in proof systems for hybrid logic, roughly, this rule replaces rules for equational

---

[4]We here have in mind natural deduction, Gentzen, and tableau systems for hybrid logic, not Hilbert-style axiom systems. Proof systems of the first three types are suitable for actual reasoning, carried out by a human, a computer, or in some other medium. Axiom systems are usually not meant for actual reasoning, but are of a more foundational interest.

Figure 1: Rules for connectives

$$\frac{\phi \quad \psi}{\phi \wedge \psi}\ (\wedge I) \qquad\qquad \frac{\phi \wedge \psi}{\phi}\ (\wedge E1) \quad \frac{\phi \wedge \psi}{\psi}\ (\wedge E2)$$

$$\begin{array}{c}[\phi]\\ \vdots\\ \dfrac{\psi}{\phi \to \psi}\ (\to I)\end{array} \qquad\qquad \frac{\phi \to \psi \quad \phi}{\psi}\ (\to E)$$

$$\begin{array}{c}[\neg\phi]\\ \vdots\\ \dfrac{\bot}{\phi}\ (\bot)^{*}\end{array}$$

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

$$\begin{array}{c}[\Diamond c]\\ \vdots\\ \dfrac{@_c\phi}{\Box\phi}\ (\Box I)^{\dagger}\end{array} \qquad\qquad \frac{\Box\phi \quad \Diamond e}{@_e\phi}\ (\Box E)$$

$*$ ϕ is a propositional letter.
$\dagger$ c does not occur free in $\Box\phi$ or in any undischarged assumptions other than the specified occurrences of $\Diamond c$.
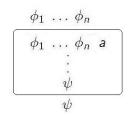
Figure 2: Rules for nominals

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

$*$ ϕ_1, ..., ϕ_n, and ψ are all satisfaction statements and there are no undischarged assumptions in the derivation of ψ besides the specified occurrences of ϕ_1, ..., ϕ_n, and a.
$\dagger$ a does not occur in ψ or in any undischarged assumptions other than the specified occurrences of a.

reasoning in other systems, see for example the rules in the natural deduction system given in Section 2.2 of the book [11].

Syntactically, the (*Term*) rule delimits a subderivation. This is similar to the way a subderivation is delimited by the introduction rule for the modal operator $\Box$ in the natural deduction system for S4 given in [7], making use of explicit substitutions in derivations, and 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 2).

$$
\frac{\begin{array}{|c|}
\phi_1 \ \ldots \ \phi_n \\
\hline
\phi_1 \ \ldots \ \phi_n \ \ a \\
\vdots \\
\psi \\
\end{array}}{\psi}
$$

4   A first example

In this section we give the first example of reasoning using the (*Term*) rule, displayed in Figure 2. The example involves spatial locations, more concretely, cities. Beside the (*Term*) rule, the key rules in the example are the rules (@I) and (@E), displayed in Figure 1, which are the introduction and elimination rules for the satisfaction operator. The rules (@I) and (@E) formalizes the following two informal arguments, adapted from [32].

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

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

Now, the (*Term*) rule enables hypothetical reasoning where reasoning is about what is the case at a specific location, possibly different from the actual location. Consider the following informal argument, also adapted from [32].

> Alcohol is forbidden in Abu Dabi; if alcohol is forbidden then Sake is forbidden, so Sake is forbidden in Abu Dabi.

The reasoning in this example argument is about what is the case in the city of Abu Dabi. If this argument is made at a specific actual location, the location of evaluation is first shifted from the actual location to a hypothetical location, namely Abu Dabi, then some reasoning is performed involving the premise "if alcohol is forbidden then Sake is forbidden", and finally the location of evaluation is shifted back to the actual location. The reader is invited to verify this shift of spatial location by checking that the argument is correct, and note that the reader himself (or herself) imagines being at the location Abu Dabi. Note that the premise "if alcohol is forbidden then Sake is forbidden" represents a relation holding at all locations.

Now, in a spatial setting, the side-condition on the rule (*Term*) requiring that all the formulas $\phi_1, \ldots, \phi_n, \psi$ are satisfaction statements (see Figure 2) ensures that these formulas have the same truth-value at all locations, so the truth-value of these formulas are not affected by a shift of spatial perspective. The rule would not be sound if this was not the case, that is, if one or more of the formulas were *spatially-indexical* (in the terminology of [32]).

We now proceed to the formalization of the above argument about what is the case in the city of Abu Dabi. We make use of the following symbolizations

Figure 3: First example formalization

$$\cfrac{\cfrac{\cfrac{[a] \quad [@_a p]}{p}\,(@E) \qquad \cfrac{}{p \to q}\,(Axiom)}{\cfrac{q}{\cfrac{[a]}{@_a q}}\,(@I)}\,(\to E)}{@_a q}$$

$$@_a p \qquad \qquad (Term)$$

p	Alcohol is forbidden
q	Sake is forbidden
a	This is Abu Dabi

and we take the formula $p \to q$ as an axiom since it represents a relation between p and q holding everywhere (note that we use an axiom since the relation $p \to q$ holds between the particular propositions p and q, we do not use an axiom schema since the relation obviously does not hold between any pair of propositions).[5] Then the argument can be formalized as the derivation in Figure 3.

Note that the derivation in Figure 3 is obtained by applying the (*Term*) rule to the subderivation below.

$$\cfrac{\cfrac{\cfrac{a \quad @_a p}{p}\,(@E) \qquad \cfrac{}{p \to q}\,(Axiom)}{q}\,(\to E)}{\cfrac{a \qquad \qquad \qquad}{@_a q}\,(@I)}$$

Thus, the (*Term*) rule delimits a piece of reasoning taking place at the hypothetical location Abu Dabi, namely the piece of reasoning encompassed by the subderivation above.

Formally, the shift to a hypothetical point of evaluation effected by the rule (*Term*) can be seen by inspecting the proof that the rule (*Term*) is sound: The world of evaluation is shifted from the actual world to the hypothetical world where the nominal a is true (see Figure 2), then some reasoning is performed involving the delimited subderivation which by induction is assumed to be sound, and finally the world of evaluation is shifted back to the actual world. Soundness of the system $\mathbf{N'}_{\mathcal{H}}$, including soundness of the rule (*Term*), is proved in Theorem 4.1 in Section 4.3 of [11].

5 The Smarties task (temporal shift version)

In this section we will give a formalization which has exactly the same structure as the formalization in the previous section, but which in other respects is quite different. It turns out that a perspective shift like the one just described in the previous section also takes place in the following version of the Smarties task where there is a shift of perspective to another time.[6]

---

[5]One may ask why the premise "if alcohol is forbidden then Sake is forbidden" is formalized as $p \to q$ using classical implication, rather than a form of non-monotonic implication. Like in many cases when classical logic is used to formalize natural language statements, there is an idealization in our choice of classical implication. We think this idealization is justified since our main goal is to formalize the perspective shift involved in the example argument. We note in passing that classical implication is also used in [32] where this example stems from, or to be precise, machinery equivalent to classical implication. See also Footnote 7.

[6]The author thanks Michiel van Lambalgen for mentioning the Smarties task in an email exchange in 2011 where the author suggested that the shift of perspective in the hybrid-logical rule (*Term*) could be of relevance in

Figure 4: Formalization of the child's correct response in the Smarties task (both temporal and person shift versions)

$$
\cfrac{
  @_a Dp \qquad
  \cfrac{[a] \qquad
    \cfrac{
      \cfrac{[a] \qquad [@_a Dp]}{Dp}\,(@E) \qquad \cfrac{}{Dp \to Bp}\,(Axiom\ schema)
    }{Bp}\,(\to E)
  }{@_a Bp}\,(@I)
}{@_a Bp}\,(Term)
$$

> 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?"

Compare to the version in the introduction of the present paper where there is a shift of perspective to another person. See [21] for more on the temporal version of the Smarties task.

Below we shall formalize each step in the logical reasoning taking place when giving a correct answer to the temporal version of the task, but before that, we give an informal analysis. Let us call the child Peter. Let a be the time where Peter answers the first question, and let t be the time where he answers the second one. 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 there were Smarties inside, he must also have come to believe that there were Smarties inside. Therefore, at the time t he concludes that at the earlier time a he believed that there were Smarties inside.

We now proceed to the full formalization. We first extend the language of hybrid logic with two modal operators, D and B. We make use of the following symbolizations

D Peter deduces that ...
B Peter believes that ...
p There are Smarties inside the tube
a The time where the first question is answered

and we take the principle $D\phi \to B\phi$ as an axiom schema (it holds whatever proposition is substituted for the metavariable ϕ, hence an axiom schema). This is principle (9.4) in [33].[7] Then the shift of temporal perspective in the Smarties task can be formalized very directly in Seligman's system as the derivation in Figure 4. Recall that the derivation is meant to formalize each step in Peter's reasoning at the time t where the second question is answered. The premise $@_a Dp$ in the derivation says that Peter at the earlier time a deduced that there were Smarties inside the tube, which he remembers at t.

Note that the formalization in Figure 4 does not involve the $\Box$ operator, so this operator could have been omitted together with the associated rules $(\Box I)$ and $(\Box E)$ in Figure 1. Since this proof system is complete, the $\Box$ operator satisfies logical omniscience. The operators D and B are only taken to satisfy the principle $D\phi \to B\phi$, as mentioned above.

---

connection with the theory of mind view of autism.

[7] Analogous to the question in Footnote 5, it can be asked why we use classical implication in $D\phi \to B\phi$, rather than a form of non-monotonic implication. Again, the answer is that this is an idealization. In this connection we remark that principle (9.4) in [33] also uses classical implication (the non-monotonicity in the logical analysis of the Smarties task of [33] does not concern principle (9.4), but other principles).

Compare the derivation in Figure 4 to the derivation in Figure 3 in the previous section and note that the structure of the derivation is exactly the same. Note that what we have done is that we have formalized the logical reasoning taking place when giving the correct answer "Smarties". Note also that information about the actual content of the tube, namely pencils, is not needed to draw the correct conclusion, in fact, the actual content is not even mentioned in the formalization.

6   The Smarties task (person shift version)

As a stepping stone between the temporal version of the Smarties task we considered in the previous section, and the Sally-Anne task we shall consider in the next section, we in the present section take a look again at the version of the Smarties task described in the introduction. The only difference between the version in the introduction and the version in the previous section is the second question where

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

obviously gives rise to a temporal shift of perspective, whereas

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

gives rise to a shift of perspective to another person, namely the imagined mother.

To give a correct answer to the latter 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 there are Smarties inside the tube from the fact that it is a Smarties tube, and from that, she must also come to believe that there are Smarties inside. Therefore, Peter concludes that the mother would believe that there are Smarties inside.

The derivation formalizing this argument is exactly the same as in the temporal case dealt with in previous section, Figure 4, but the symbols are interpreted differently, namely as

D	Deduces that ...
B	Believes that ...
p	There are Smarties inside the tube
a	The imagined mother

So now nominals refer to persons rather than times. Accordingly, the modal operator B now symbolize the belief of the person represented by the point of evaluation, rather than Peter's belief at the time of evaluation, etc. Thus, the premise $@_a D p$ in the derivation in Figure 4 says that the imagined mother deduces that there are Smarties inside the tube, which the child doing the reasoning takes to be the case since the mother is imagined to be present in the room.

Incidentally, letting points in the Kripke model represent persons is exactly what is done in Arthur Prior's *egocentric logic*, see Section 1.3 in the book [11], in particular pp. 15–16. In egocentric logic the accessibility relation represents the taller-than relation, but this relation is obviously not relevant here.

7   The Sally-Anne task

In this section we will present a formalization of a more complicated reasoning task called the Sally-Anne task. In a number of places we shall compare to the detailed logical analysis of the Sally-Anne task given in the book [33], and we shall also make some remarks in relation to the formalization of the Sally-Anne task given in the papers [2] and [3]. The following is one version of this 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, 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?"

Most children above the age of four correctly responds where Sally must falsely believe the marble to be (in the basket) whereas younger children respond where they know the marble to be (in the box). Again, for autists, the cutoff is higher.

Below we shall formalize the correct response to the task, but before that, we give 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 also believe the marble is in the basket. At t_2, after she has returned, she still believe that the marble is in the basket since she has not seen Anne moving it at the time t_1. Therefore, Peter concludes that Sally believes that the marble is in the basket.

In our formalization we make use of a tiny fragment of first-order hybrid logic, involving the predicates $l(i,t)$ and $m(t)$ as well as the modal operators S and B, but no quantifiers. This should be compared to [33] and [2] which both uses the event calculus, encoded in first-order logic, involving quantification[8] over times, and in the case of [2], also quantification over events and fluents (a fluent is a condition that can change truth-value over time). However, we think that quantification over times is really not needed for formalizing the Sally-Anne task, the reason being that the scenario only involves a fixed finite number number of times (in our formalization three distinct times). Even though [33] and [2] both uses the event calculus, it should be remarked that they interpret logical constructs in a very different way: Clauses are in [33] evaluated as Horn clauses in logic programs, and negation is interpreted using negation as failure (classical negation is not expressible using Horn clauses, to be more precise, a Horn clause program cannot have negative consequences in the classical sense, cf. for example [31], page 151). On the other hand, logical constructs are in [2] interpreted using an interactive theorem prover having a classical logic basis.

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 the timepoint t is denoted $t+1$. This should be compared to [33] and [2] where time is taken to be continuous, since this is how time is represented in the event calculus. Now, we make use of the following symbolizations

$l(i,t)$	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 make use of the following three principles

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

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 (D) is the only purely modal principle we are going to make use of. Strictly speaking, we 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. This principle is identical to principle (9.2) in the book [33], page 251.

---

[8]Formally, there are no quantifiers in the object language used by [33] to formalize the Sally-Anne task, but quantification relies on the fact that uninstantiated variables in logic programs are automatically quantified, as described in Footnote 9.

Principle $(P2)$ is remiscent of principle (9.11)[9] in the book [33], page 253, and axiom $[A_5]$[10] in the paper [2], page 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. Of course, this requires taking into account all actions that might make the predicate l false, but only one such action is mentioned in the Sally-Anne scenario, namely the action of moving the marble, formalized as m. One might envisage other actions that could make i false, for example heating the marble so much that it evaporates, but no such action is mentioned in the scenario. If another such action h had been mentioned, the predicate $m(t)$ in Principle $(P2)$ would have to be replaced by $m(t) \vee h(t)$, and the formalization of the Sally-Anne task we give below, would have to be adjusted accordingly.

Note that Principle $(P2)$ is not schematic in the sense that it only holds for the particular predicates l and m, not predicates or formulas in general, the reason being that it encodes a specific interaction between the two predicates: If the action m takes place, then it causes the predicate l to be false.

The principle $l(i,t) \wedge \neg m(t) \rightarrow l(i, t+1)$ obtained by removing the occurrences of the belief modality B from $(P2)$ says that if a predicate is true, and no action takes place causing it to be false, then the predicate stays true. In Artificial Intelligence this principle is captured by so-called successor-state axioms, which is one standard way to solve the famous frame problem, see for example the textbook [30]. In accordance with the idea of successor-state axioms, the action of moving the marble is formalized as a predicate, the predicate $m(t)$, meaning that the same generic action can be performed at different times (so we consider the action as a *type*). This is contrary to the approach taken in the paper [2], where actions are taken as events, and hence are time-stamped (thus, that paper considers actions as *tokens*). This is related to the fact that [2] involve quantification over events. See [17] for a discussion on types versus tokens in temporal reasoning. The use of first-order machinery in our formalization of the Sally-Anne tasks, in comparison to the formalizations of [2] and [33], can be summed up as follows, cf. also footnotes 9 and 10.

---

[9]Slightly reformulated, principle (9.11) of [33] is the following

$$B_a l(i,t) \wedge t < u \wedge \neg B_a\, clipped(t,i,u) \rightarrow D_a l(i,u)$$

where a stands for an agent, and where $clipped(t,i,u)$ means that the marble ceases to be at location i at some time between t and u, that is, there exists a time r between t and u such thet the marble ceases to be at location i at the time r. This principle is in [33] interpreted as a clause in a logic program where the negation prefixing the second occurrence of the B_a operator is interpreted using negation as failure. The predicate $clipped(t,i,u)$ stems from the event calculus where it is defined as

$$\forall t \forall f \forall u (clipped(t,f,u) \leftrightarrow \exists e \exists r (happens(e,r) \wedge t < r < u \wedge terminates(e,f,r)))$$

and where f stands for fluents and e stands for events. Notice the existential quantifiers $\exists e$ and $\exists r$ ranging over events and times. The book [33] defines $clipped$ using the clauses (9.12) and (9.13) at page 253. These two clauses are similar, and slightly reformulated, clause (9.12) is the following

$$l(i,t) \wedge m(r) \wedge t < r < u \rightarrow clipped(t,i,u)$$

Note that there is no quantification in the object language of this clause, rather, quantification is taken care of by the evaluation mechanism in logic programs, where uninstantiated variables are automatically quantified over, cf. also Footnote 8, in particular, an uninstantiated variable in the body of a clause, not occurring in its head, is automatically existentially quantified, which is exactly what is going on with the variable r above. So the book [33] existentially quantifies over times like in the usual definition of $clipped$, displayed above, but contrary to the usual definition, the book does not quantify over events, rather, the action of moving the marble is formalized as a predicate, namely the predicate $m(r)$.

[10] Axiom $[A_5]$ of [2] is the following

$$C \forall a \forall f \forall t \forall u (B_a\, holds(f,t) \wedge B_a(t < u) \wedge \neg B_a\, clipped(t,f,u) \rightarrow B_a\, holds(f,u))$$

where C is the common knowledge operator. The paper [2] defines $clipped$ as usual in the event calculus, that is, as displayed in Footnote 9, except that the definition is prefixed by the common knowledge operator. So [2] makes use of quantification over times and events as well as fluents.

	Our work	The book [33]	The paper [2]
Terms referring to times	Yes	Yes	Yes
Quantification over times	No	Yes	Yes
Quantification over events	No	No	Yes
Quantification over fluents	No	No	Yes

Beside the principles (D), $(P1)$ and $(P2)$, we shall also encode the information[11] 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). We encode this information as

$$(P3) \qquad Bm(t) \to Sm(t)$$

Strictly speaking, we will use the contrapositive formulation $\neg Sm(t) \to \neg Bm(t)$. When Peter figures out that Sally does not believe that the marble has been moved at t_1, he uses that she did not see this happening, as she was not present in the room. This step in Peter's reasoning is exactly what Principle $(P3)$ enables. The information encoded in $(P3)$ is of course dependent on the concrete scenario, wherefore one cannot expect that it can be formalized as an axiom schema.

Principle $(P3)$ is not explicitly taken as a principle in the book [33], but it is implicit, the reason being that the $(P1)$ principle $S\phi \to B\phi$ is interpreted as a clause in a logic program, cf. Subsection 9.4.2 in the book, and in the absence of other clauses whose heads can be instantiated to $Bm(t)$, the logic programming query $?Bm(t)$ is successful only if $?Sm(t)$ is successful[12]. If negation is interpreted using negation as failure, this means by contraposition that if the query $?Sm(t)$ is not successful, that is, if $?\neg Sm(t)$ is successful, then $?\neg Bm(t)$ is successful, which is parallel to our Principle $(P3)$. Since our formalization rely on classical logic, not the closed world reasoning mechanism in logic programming, we need explicit encoding of the information that a belief that the marble is being moved can only be acquired by seeing it.

In order to make the formalization more compact, and also more in the spirit of natural deduction style, we do not take the four principles above as axioms or axiom schemas, but instead we turn them into the following proof-rules.

$$\frac{B\neg\phi}{\neg B\phi} \text{ (D)} \qquad \frac{S\phi}{B\phi} \, (P1) \qquad \frac{Bl(i,t) \quad \neg Bm(t)}{Bl(i,t+1)} \, (P2) \qquad \frac{\neg Sm(t)}{\neg Bm(t)} \, (P3)$$
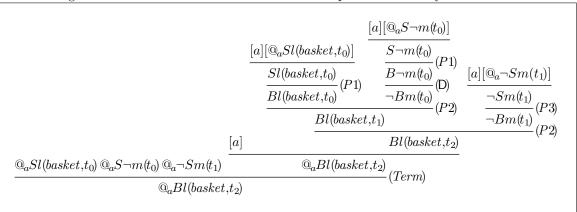
With this machinery in place, the shift of person perspective in the Sally-Anne task can be formalized as the derivation in Figure 5 where to save space, we have omitted names of the introduction and elimination rules for the @ operator. Recall that the derivation in Figure 5 is meant to formalize the child Peter's reasoning at the time t_2 where the question is answered. The first two premises $@_a Sl(basket, t_0)$ and $@_a S\neg m(t_0)$ in the derivation say that Sally (the reference the nominal a) 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, $@_a \neg Sm(t_1)$, says that Sally did not see the marble being moved at the time t_1, this being the case since she was absent, which Peter remembers.

Note that the actual position of the marble at the time t_2 is irrelevant to figure out the correct response. Note also that in the Sally-Anne task there is a shift of person perspective which we deal with in a modal-logical fashion letting points of evaluation stand for persons, like in the person version of the Smarties task in the previous section, but there is also a temporal shift in the Sally-Anne task, which we deal with using first-order machinery.

---

[11]Thanks to one of the reviewers for a comment prompting the author to include this information in the formalization, thereby making it more cognitively faithful.

[12]In the interest of comparison, we here disregard that [33] allows a couple of other ways to acquire belief than *seeing* something, but this this can also be incorporated in our formalization at the expence of making it slightly more complicated.

Figure 5: Formalization of the child's correct response in the Sally-Anne task

$$
\dfrac{
\begin{array}{c}
\dfrac{[a]\,[@_a Sl(basket,t_0)]}{\dfrac{Sl(basket,t_0)}{Bl(basket,t_0)}(P1)}\,(P1)
\qquad
\dfrac{\dfrac{[a]\,[@_a S\neg m(t_0)]}{\dfrac{S\neg m(t_0)}{\dfrac{B\neg m(t_0)}{\neg Bm(t_0)}(D)}(P1)}}{Bl(basket,t_1)}(P2)
\qquad
\dfrac{[a]\,[@_a \neg Sm(t_1)]}{\dfrac{\neg Sm(t_1)}{\neg Bm(t_1)}(P3)}
\\[6pt]
Bl(basket,t_2)
\end{array}
}{\,}(P2)
$$

$$
\dfrac{@_a Sl(basket,t_0) \quad @_a S\neg m(t_0) \quad @_a \neg Sm(t_1) \qquad \dfrac{[a] \qquad Bl(basket,t_2)}{@_a Bl(basket,t_2)}}{@_a Bl(basket,t_2)}(Term)
$$

8 Some remarks on other work

Beside analysing the reasoning taking place when giving a correct answer to a reasoning task, the works by van Lambalgen and co-authors also analyse what goes wrong when an incorrect answer is given. We note that Stenning and van Lambalgen in [33] warn against simply characterizing autism as a lack of theory of mind. Rather than being an explanation of autism, Stenning and van Lambalgen see the theory of mind deficit hypothesis as "an important label for a problem that needs a label", cf. [33], page 243. They argue that another psychological theory of autism is more fundamental, namely what is called the *executive function deficit theory*. Very briefly, executive function is an ability to plan and control a sequence of actions with the aim of obtaining a goal in different circumstances. In comparison with the work of the present paper, a decisive difference is which psychological theory is taken as the basis of the logical analysis. If the executive function deficit theory is taken as the basis, then it appears natural to try to formalize a false-belief task in some sort of non-monotonic logic. This is what van Lambalgen and co-authors do. On the other hand, if the theory of mind deficits theory is taken as the basis, then we find that it is natural to use a classical version of hybrid logic and hybrid-logical proof-theory.

The paper [25] reports empirical investigations of closed world reasoning in adults with autism. Incidentally, according to the opening sentence of that paper, published in 2009, "While autism is one of the most intensively researched psychiatric disorders, little is known about reasoning skills of people with autism."

With motivations from the theory of mind literature, the paper [36] models examples of beliefs that agents may have about other agents' beliefs, one example is an autistic agent that always believes that other agents have the same beliefs as the agent's own. This is modelled by different agents preference relations between states, where an agent prefers one state over another if the agent considers it more likely. The beliefs in question turn out to be frame-characterizable by formulas of epistemic logic.

The paper [15] reports empirical investigations of what is called *second-order* theory of mind, which is a person's capacity to take into account other people's beliefs about other people's beliefs, for example the person's own beliefs (in comparison to *first-order* theory of mind, which is the capacty to take into account other peoples beliefs about simple world facts, in line with what we previously in the present paper just have called theory of mind). The investigations in [15] make use of a second-order false-belief task, as well as other tasks.

The paper [19] does not deal with false-belief tasks or theory of mind, but it is nevertheless relevant to mention since it uses formal proofs to compare the cognitive difficulty of deductive tasks. To be more precise, the paper associates the difficulty of a deductive task in a version of the Mastermind game with the minimal size of a corresponding tableau tree, and it uses this measure of difficulty to predict the empirical difficulty of game-plays, for example the number of

steps actually needed for solving a task.

The method of reasoning in tableau systems can be seen as attempts to construct a model of a formula: A tableau tree is built step by step using rules, whereby more and more information about models for the formula is obtained, and either at some stage a model can be read off from the tableau tree, or it can be concluded that there cannot be such a model (in fact, in the case of [19], any formula under consideration has exactly one model, so in that case it is a matter of building a tableau tree that generates this model). Hence, if the building of tableau trees is taken to be the underlying mechanism when a human is solving Mastermind tasks, then the investigations in [19] can be seen to be in line with the mental models school (see the third section of the present paper).

A remark from a more formal point of view: The tableau system described in [19] does not include the cut-rule[13]. Much has been written on the size of proofs in cut-free proof systems, in particular, the paper [9] gives examples of first-order formulas whose derivations in cut-free systems are much larger than their derivations in natural deduction systems, which implicitly allow unrestricted cuts (in one case more than 10^{38} characters compared to less than 3280 characters). Similarly, the paper [14] points out that ordinary cut-free tableau systems have a number of anomalies, one of them being that for some classes of propositional formulas, decision procedures based on cut-free systems are much slower than the truth-table method (in the technical sense that there is no polynomial time computable function that maps truth-table proofs of such formulas to proofs of the same formulas in cut-free tableau systems). Instead of prohibiting cuts completely, the paper [14] advocates allowing a restricted version of the cut-rule, called the analytic cut-rule.

9   Future work

We would like to extend the work of the present paper to further false-belief tasks, perhaps using different hybrid-logical machinery (and moreover, use hybrid-logical proof-theory to analyse what goes wrong when incorrect answers are given). Not only will formalization of further reasoning tasks be of interest on their own, but we also expect that such investigations can be feed back into logical research, either as corroboration of the applicability of existing logical constructs, or in the form of new logical constructs, for example new proof-rules or new ways to add expressive power to a logic.

We are also interested in further investigations in when two seemingly dissimilar reasoning tasks have the same underlying logical structure, like we in the present paper have disclosed that two different versions of the Smarties task have exactly the same underlying logical structure[14]. Such investigations might be assisted by a notion of identity on proofs (exploiting the longstanding effort in proof-theory to give a notion of identity between proofs, that is, a way to determine if two arguments have common logical structure, despite superficial dissimilarity). If two experiments make use of seemingly dissimilar reasoning tasks, but which have the same underlying logical structure, then we would expect similar empirical results (for example in terms of number of correct answers and/or reaction time). In this case the identity of logical structure can be seen as an explanation of the similarity of the results. On the other hand, if the experiments give differing empirical results, despite having the same logical structure, then it calls for an explanation: One such explanation could be differing levels of abstraction, in the extreme case a purely symbolic reasoning task in comparison to a reasoning task dealing with a familiar everyday situation[15].

---

[13]The cut-rule says that the end of any branch in a tableau tree can extended with two branches with ϕ on the one branch and $\neg\phi$ on the other (expressing the bivalence of classical logic).

[14]Other examples of dissimilar, but logically equivalent, reasoning tasks are the two-player games called Marble Drop and the Matrix Game, used to investigate higher-order social reasoning. They are equivalent in the sense of being game-theoretically equivalent. See the papers [23] and [34].

[15]It is well-known that such differences can give rise to differences in performance, see for example the extensive literature on the *Wason selection task*, as surveyed in [33] and [24]. Another example is the above mentioned games Marble Drop and the Matrix Game. In [23] it is demonstrated that subjects perform better when a game is embedded in a concrete physical context (Marble Drop) than when it is given a more abstract formulation (the Matrix Game).

Acknowledgements

A   Proof of analyticity

Usually, when considering a natural deduction system, one wants to equip it with a normalizing set of reduction rules such that normal derivations satisfy the subformula property. Normalization says that any derivation by repeated applications of reduction rules can be rewritten to a derivation which is normal, that is, no reduction rules apply. From this it follows that the system under consideration is analytic.

Now, the works [10] and [11], Section 4.3, by the present author devise a set of reduction rules for $\mathbf{N}'_{\mathcal{H}}$ obtained by translation of a set of reduction rules for a more common natural deduction system for hybrid logic. This more common system, which we denote $\mathbf{N}_{\mathcal{H}}$, can be found in [10] and in [11], Section 2.2. All formulas in the system $\mathbf{N}_{\mathcal{H}}$ are satisfaction statements. Despite other desirable features, it is not known whether the reduction rules for $\mathbf{N}'_{\mathcal{H}}$ are normalizing, and normal derivations do not always satisfy the subformula property. In fact, Chapter 4 of the book [11] ends somewhat pessimistically by exhibiting a normal derivation without the subformula property. It is remarked that a remedy would be to find a more complete set of reduction rules, but the counter-example does not give a clue how such a set of reduction rules should look.

In what follows we shall take another route. We prove a completeness result saying that any valid formula has a derivation in $\mathbf{N}'_{\mathcal{H}}$ satisfying a version of the subformula property. This is a sharpened version of a completeness result for $\mathbf{N}'_{\mathcal{H}}$ originally given in [10] and in Section 4.3 of [11] (Theorem 4.1 in [11]). Thus, we prove that $\mathbf{N}'_{\mathcal{H}}$ is analytic without going via a normalization result. So the proof of the completeness result does not involve reduction rules. The result is mathematically weaker than normalization together with the subformula property for normal derivations, but it nevertheless demonstrates analyticity. Analyticity is a major success criteria in proof-theory, one reason being that analytic provability is a step towards automated theorem proving (which obviously is related to Leibniz' aim mentioned in the intoduction of the present paper).

In the proof below we shall refer to $\mathbf{N}_{\mathcal{H}}$ as well as a translation $(\cdot)^{\circ}$ from $\mathbf{N}_{\mathcal{H}}$ to $\mathbf{N}'_{\mathcal{H}}$ given in [10] and Section 4.3 of [11]. This translates a derivation π in $\mathbf{N}_{\mathcal{H}}$ to a derivation π° in $\mathbf{N}'_{\mathcal{H}}$ having the same end-formula and parcels of undischarged assumptions. The reader wanting to follow the details of our proof is advised to obtain a copy of the paper [10] or the book [11]. The translation $(\cdot)^{\circ}$ satisfies the following.

Lemma A.1 *Let π be a derivation in $\mathbf{N}_{\mathcal{H}}$. Any formula θ occuring in π° has at least one of the following properties.*

1. *θ occurs in π.*

2. *$@_a\theta$ occurs in π for some satisfaction operator $@_a$.*

3. *θ is a nominal a such that some formula $@_a\psi$ occurs in π.*

Proof Induction on the structure of the derivation of π. Each case in the translation $(\cdot)^{\circ}$ is checked. Q.E.D.

Note that in item 1 of the lemma above, the formula θ must be a satisfaction statement since only satisfaction statements occur in π. In what follows $@_d\Gamma$ denotes the set of formulas $\{@_d\xi \mid \xi \in \Gamma\}$.

Theorem A.2 *Let π be a normal derivation of $@_d\phi$ from $@_d\Gamma$ in $\mathbf{N}_{\mathcal{H}}$. Any formula θ occuring in π° has at least one of the following properties.*

1. *θ is of the form $@_a\psi$ such that ψ is a subformula of ϕ, some formula in Γ, or some formula of the form c or $\Diamond c$.*

2. *θ is a subformula of ϕ, some formula in Γ, or some formula of the form c or $\Diamond c$.*

3. *θ is a nominal.*

4. *θ is of the form $@_a(p \to \bot)$ or $p \to \bot$ where p is a subformula of ϕ or some formula in Γ.*

5. *θ is of the form $@_a\bot$ or $\bot$.*

Proof Follows from Lemma A.1 above together with Theorem 2.4 (called the quasi-subformula property) in Subsection 2.2.5 of [11]. Q.E.D.

We are now ready to give our main result, which is a sharpened version of the completeness result given in Theorem 4.1 in Section 4.3 of [11].

Theorem A.3 *Let ϕ be a formula and Γ a set of formulas. The first statement below implies the second statement.*

1. *For any model $\mathcal{M}$, any world w, and any assignment g, if, for any formula $\xi \in \Gamma$, $\mathcal{M}, g, w \models \xi$, then $\mathcal{M}, g, w \models \phi$.*

2. *There exists of derivation of ϕ from Γ in $\mathbf{N}'_{\mathcal{H}}$ such that any formula θ occuring in the derivation has at least one of the five properties listed in Theorem A.2.*

Proof Let d be a new nominal. It follows that for any model $\mathcal{M}$ and any assignment g, if, for any formula $@_d\xi \in @_d\Gamma$, $\mathcal{M}, g \models @_d\xi$, then $\mathcal{M}, g \models @_d\phi$. By completeness of the system $\mathbf{N}_{\mathcal{H}}$, Theorem 2.2 in Subsection 2.2.3 of the book [11], there exists a derivation π of $@_d\phi$ from $@_d\Gamma$ in $\mathbf{N}_{\mathcal{H}}$. By normalization, Theorem 2.3 in Subsection 2.2.5 of the book, we can assume that π is normal. We now apply the rules $(@I)$, $(@E)$, and $(Name)$ to π° obtaining a derivation of ϕ from Γ in $\mathbf{N}'_{\mathcal{H}}$ satisfying at least one of the properties mentioned in Theorem A.2. Q.E.D.

Remark: If the formula occurrence θ mentioned in the theorem above is not of one of the forms covered by item 4 in Theorem A.2, and does not have one of a finite number of very simple forms not involving propositional symbols, then either θ is a subformula of ϕ or some formula in Γ, or θ is of the form $@_a\psi$ such that ψ is a subformula of ϕ or some formula in Γ. This is the version of the subformula property we intended to prove.

References

[1] C. Areces and B. ten Cate. Hybrid logics. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*, pages 821–868. Elsevier, 2007.

[2] 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.

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

[4] M. Baaz and A. Leitsch. *Methods of Cut-Elimination*, volume 34 of *Trends in Logic Series*. Springer, 2011.

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

[6] S. Baron-Cohen, A.M. Leslie, and U. Frith. Does the autistic child have a 'theory of mind'? *Cognition*, 21:37–46, 1985.

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

[8] P. Blackburn, T. Braüner, T. Bolander, and K.F. Jørgensen. A Seligman-style tableau system. In K. McMillan, A. Middeldorp, and A. Voronkov, editors, *Logic for Programming, Artificial Intelligence, and Reasoning*, volume 8312 of *Lecture Notes in Computer Science*, pages 147–163. Springer Publishing Company, 2013.

[9] G. Boolos. Don't eliminate cut. *Journal of Philosophical Logic*, 13:373–378, 1984.

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

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

[12] T. Braüner. Hybrid-logical reasoning in false-belief tasks. In B.C. Schipper, editor, *Proceedings of Fourteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pages 186–195, 2013. ISBN 978-0-615-74716-3, available at http://tark.org.

[13] T. Braüner. Hybrid-logical reasoning in the Smarties and Sally-Anne tasks. *Journal of Logic, Language and Information*, 23:415–439, 2014. Revised and extended version of [12].

[14] M. D'Agostino and M. Mondadori. The taming of the cut. Classical refutations with analytical cut. *Journal of Logic and Computation*, 4:285–319, 1994.

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

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

[17] A. Galton. Operators vs. arguments: The ins and outs of reification. *Synthese*, 150:415–441, 2006. Special issue edited by T. Braüner, P. Hasle, and P. Øhrstrøm.

[18] G. Gentzen. Investigations into logical deduction. In M.E. Szabo, editor, *The Collected Papers of Gerhard Gentzen*, pages 68–131. North-Holland Publishing Company, 1969.

[19] N. Gierasimczuk, H. van der Maas, and M. Raijmakers. An analytic tableaux model for Deductive Mastermind empirically tested with a massively used online learning system. *Journal of Logic, Language and Information*, 22:297–314, 2013.

[20] W. Goldfarb. *Deductive Logic*. Hackett Pub. Co., 2003.

[21] A. Gopnik and J.W. Astington. 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, 1988.

[22] P.N. Johnson-Laird. Mental models and deductive reasoning. In J.E. Adler and L.J. Rips, editors, *Reasoning: Studies of Human Inference and Its Foundations*, pages 206–222. Cambridge University Press, 2008.

[23] B. Meijering, L. van Maanen, H. van Rijn, and R. Verbrugge. The facilitative effect of context on second-order social reasoning. In R. Catrambone and S. Ohlsson, editors, *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 1423–1428. Cognitive Science Society, 2010.

[24] C. Dutilh Novaes. *Formal Languages in Logic: A Philosophical and Cognitive Analysis.* Cambridge University Press, 2012.

[25] J. Pijnacker, B. Geurts, M. van Lambalgen, C.C. Kan, J.K. Buitelaar, and P. Hagoort. Defeasible reasoning in high-functioning adults with autism: Evidence for impaired exception-handling. *Neuropsychologia*, 47:644–651, 2009.

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

[27] D. Prawitz. Ideas and results in proof theory. In J. E. Fenstad, editor, *Proceedings of the Second Scandinavian Logic Symposium*, volume 63 of *Studies in Logic and The Foundations of Mathematics*, pages 235–307. North-Holland, 1971.

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

[29] 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.

[30] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach, 3rd Edition.* Prentice Hall, 2009.

[31] U. Schöning. *Logic for Computer Scientists.* Birkhäuser Verlag, 1989.

[32] 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.

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

[34] J. Szymanik, B. Meijering, and R. Verbrugge. Using intrinsic complexity of turn-taking games to predict participants' reaction times. In M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth, editors, *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pages 1426–1432. Cognitive Science Society, 2013.

[35] A.S. Troelstra and H. Schwichtenberg. *Basic Proof Theory*, volume 43 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, 1996.

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

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