Towards a logical formalisation of Theory of Mind: a study on False Belief Tasks

Anthia Solaki and Fernando R. Velázquez-Quesada

Institute for Logic, Language and Computation, Universiteit van Amsterdam.
`a.solaki2@uva.nl`, `F.R.VelazquezQuesada@uva.nl`

Abstract. Theory of Mind, the cognitive capacity to attribute internal mental states to oneself and others, is a crucial component of social skills. Its formal study has become important, witness recent research on reasoning and information update by intelligent agents, and some proposals for its formal modelling have put forward settings based on Epistemic Logic (*EL*). Still, due to intrinsic idealisations, it is questionable whether *EL* can be used to model the high-order cognition of 'real' agents. This manuscript proposes a mental attribution modelling logical framework that is more in-line with findings in cognitive science. We introduce the setting and some of its technical features, and argue why it does justice to empirical observations, using it for modelling well-known False-Belief Tasks.

Keywords: theory of mind, mental state attribution, false belief tasks, temporal model, dynamic epistemic logic

1 Introduction

An important feature of how people function in social scenarios is that of *Theory of Mind* (ToM), the cognitive capacity to attribute internal mental states, such as knowledge and beliefs, to oneself and others [1].[1] Theory of Mind is a crucial component of social skills: someone who understands that others might have mental states different from hers, and can reason about those states, is much better suited to understand their behaviour, and thus act and react appropriately.

Theory of Mind is slowly developed in the course of our lives [3,4] (and at different speed for different types of persons [5,6]), starting with the ability to make *first-order* attributions (e.g., someone knowing/believing that *"Mary believes that the ball is in the bag"*) and progressing through attributions of *second-order* mental states (e.g., someone knowing/believing that *"Mary believes that John believes that the ball is in the closet"*). When testing one's ToM, an extensively used experiment is the *Sally-Anne* False-Belief Task.

Example 1 (The *Sally-Anne* (*SA*) task) The following is adapted from [3].

---

[1] There has been a debate on how this understanding of others' mental states is achieved (see, e.g., [2]). Some argue that it is by acquiring a *theory* of commonsense psychology (*theory theory*); some others argue that it comes from a direct *simulation* of others' mental states (*simulation theory*). We will use the term ToM without endorsing any of these views, as such discussion falls outside the scope of this proposal.

> *Sally and Anne are in a room in which there are a basket and a box. Sally is holding a marble. Then, after putting the marble into the basket, Sally leaves the room. While Sally is away, Anne transfers the marble to the box. Then Sally comes back.*

To pass the test, the subject should answer correctly the question *"where does Sally believe the marble to be?"*. This requires for the subject to distinguish between her own true belief (*"the marble is in the box"*) and Sally's *false* belief (*"the marble is in the basket"*). Experiments have shown that, while children older than 4 years old tend to answer correctly, younger children (or children on the autism spectrum) tend to fail the test, reporting their own belief [3]. (But see [7].) ◀

In the enterprise of studying and understanding ToM, there has been a growing interest on the use of formal frameworks. A seemingly natural choice is Epistemic Logic (*EL*) [8,9], as it provides tools for representing not only the knowledge/beliefs agents have about ontic facts, but also the knowledge/beliefs they have about their own and others' knowledge/beliefs. However, using *EL* has some drawbacks. First, within *EL*'s standard relational 'Kripke' semantics, knowledge/beliefs are closed under logical consequence (the *logical omniscience* problem; [10]). Moreover, the extra relational requirements for 'faithful' representations of knowledge and beliefs turn them into S5 and KD45 modal logics, respectively, thus yielding fully (positive and negative) introspective agents.

There is an even more fundamental reason why *EL* might not be well-suited for representing realistic high-order attributions. Semantically, both knowledge and beliefs correspond to a universal quantification (ϕ is known/believed iff it is the case in *all* the alternatives the agent considers possible); still, for real agents, these notions involve more elaborate considerations (e.g., observation, communication, reasoning). This 'simple' universal quantification works because *EL* uses a loaded model, which contains not only the (maximally consistent) alternatives the agent considers possible, but also every other alternative *every other agent* considers possible.[2] In a few words, the semantic interpretation of (high-order) knowledge/beliefs formulas is simple because the model is complex. Real agents might not be able to have such a loaded structure 'in their mind', and thus it is questionable whether the use of traditional *EL* can provide a proper picture of the way real agents deal with mental attribution scenarios.

In light of these issues, one could even wonder whether it makes sense to use logical tools for dealing with results of empirical research. Indeed, it has been argued that psychological experiments and logic are essentially different[3], understanding the former as the study of empirical findings on the behaviour of real 'fallible' agents, and the latter as a normative discipline studying what 'rational' agents *should* do. However, other authors (e.g., [14,15]) have justified why bridging these two views is a worthwhile endeavour that also has promising applications (especially on reasoning and information update by intelligent

---

[2] Frameworks for representing acts of private communication [11] make this clear. Their additional structures, *action models*, have one 'event' for each different perspective the agents might have about the communication, and the model after the communication contains roughly one copy of the original model for each one of these perspectives.

[3] *Anti-Psychologism* (e.g., [12]) has long been against attempts to reconcile the two [13].

agents). Indeed, empirical research benefits from using formal tools to explain their discoveries and understand their consequences, and logical frameworks become richer and more 'useful' when they capture human limitations and prescribe behaviour attainable by real agents.

This work seeks a ToM's logical setting that is more in-line with the findings in cognitive science, with non-trivial and competent agents whose underlying reasoning is reflected in the syntax and semantics.[4] To that end, we aim at the converse direction to that of *EL*. Our structures are simple, encoding only basic facts, and thus resembling the 'frugal' way real agents keep information stored. However, interpretations of mental state attributions show that agents engage in the, oftentimes strenuous, process of recalling these facts and deriving further information on their basis.

Outline The text is organised as follows. Section 2 introduces the *temporal visibility* framework, presenting its model and formal language, and also discussing some of its technical aspects. Then, Section 3 relates the features of the setting with findings in the cognitive science literature, using it to model well-known mental attribution tasks in detail, and comparing it with other related formal settings. Section 4 closes, recapitulating the highlights, discussing ways in which the framework can be extended, and suggesting lines for further research.

2 Visibility in a temporal setting

In most mental attribution tasks, beliefs[5] are, at their lower (ontic) order, about the location of certain objects (e.g., the marble's location in the Sally-Anne Task). We do take objects as the main entities about which agents have mental attitudes; still, for simplicity, we will work with these objects' *colours*. Let $A \neq \varnothing$ be the set of agents $(a, b, \ldots)$, and $O \neq \varnothing$ be the set of objects $(o, p, q, \ldots)$. For each $o \in O$, the set R_o contains the colours the object might have; define $R_O := \bigcup_{o \in O} R_o$. The model is a temporal structure, with each stage (*state*) fully described by both the colour of each object and the objects and agents each agent sees.

DEFINITION 2.1 (TEMPORAL VISIBILITY MODEL) A *temporal visibility (TV) model* is a tuple $\langle n, S, \tau, \kappa, \nu \rangle$ with ***(i)*** $n \in \mathbb{N}$ the index of the 'most recent' (current) stage; ***(ii)*** S a finite set of states with $|S| = n$; ***(iii)*** $\tau : S \to \{1..n\}$ the temporal index (bijective) function, indicating the temporal index $\tau(s) \in \{1..n\}$ of each state $s \in S$; ***(iv)*** $\kappa : S \to (O \to R_O)$ the *colouring* function, with $\kappa(s, o)$ (abbreviated as $\kappa_s(o)$) the colour object o has at state s;[6] ***(v)*** $\nu : S \to (A \to \wp(A \cup O))$ the *visibility* function, with $\nu(s, a)$ (abbreviated as $\nu_s(a)$) the *entities* (agents and objects) agent a sees at state s.[7] Given a *TV* model, let $s_{last} \in S$ be its (unique) state satisfying $\tau(s_{last}) = n$. ◀

---

[4] In particular, one goal is to find a system that provides a plausible answer on why people find mental attribution tasks increasingly difficult as their order increases.

[5] Following the common parlance in the literature describing the tasks we later model, the term *belief* will be used for referring to an agent's mental state.

[6] Each object has a proper colour: $\kappa_s(o) \in R_o$ holds for all $s \in S$ and $o \in O$.

[7] Every agent can see herself in every state: $a \in \nu_s(a)$ holds for all $s \in S$ and all $a \in A$.

Example 2 Take the Sally-Anne Task, with Sally (Sa), Anne (An) and the marble (mar). Consider a two-state model M with ***(i)*** s_1 the initial state, where both agents see all agents and objects ($\nu_{s_1}(Sa) = \nu_{s_1}(An) = \{Sa, An, mar\}$) and the object is black ($\kappa_{s_1}(mar) = black$, read as 'the marble is in Sally's hands'), and ***(ii)*** s_2 the 'next' state, where both agents still see everything, but now the object is white ($\kappa_{s_2}(mar) = white$, read as 'the marble is in the basket'). The model is depicted as

◄

Representing actions A TV model contains not only a state representing the current situation (the state $\tau^{-1}(n)$) but also states indicating how the situation was in the past (up to the initial $\tau^{-1}(1)$). One can provide operations that *extend* the current model with a state depicting the outcome of a certain activity (the way the situation *will* be). In the Sally-Anne Task, some acts modify the colour of objects (Sally puts the marble into the basket) and some others modify the agents' visibility (Sally leaves the room). Here are operations for them.

Definition 2.2 (Colour change) Let $M = \langle n, S, \tau, \kappa, \nu \rangle$ be a TV model, with $s_{new} \notin S$; take a set of objects $\{p_1, \dots, p_k\} \subseteq O$, with $c_i \in R_{p_i}$ a proper colour for each p_i. The colour assignment $[p_1{:=}c_1, \dots, p_k{:=}c_k]$ produces the TV model

$$M_{[p_1:=c_1,\dots,p_k:=c_k]} = \langle n+1, S \cup \{s_{new}\}, \tau', \kappa', \nu' \rangle$$

in which ***(i)*** τ' preserves the temporal position of states in S, making s_{new} the most recent (so $\tau'(s) := \tau(s)$ for $s \in S$, and $\tau'(s_{new}) := n+1$); ***(ii)*** κ' is exactly as κ for states in S, with the new s_{new} taking the colouring of s_{last} for objects not mentioned by the assignment, and following the assignment for the colour of the objects it mentions (so, for any $o \in O$, define $\kappa'_s(o) := \kappa_s(o)$ for $s \in S$, with $\kappa'_{s_{new}}(o) := \kappa_{s_{last}}(o)$ when $o \notin \{p_1, \dots, p_k\}$, and $\kappa'_{s_{new}}(p_j) := c_j$ when $o = p_j$); ***(iii)*** ν' preserves the visibility assignment for states in S, with visibility in s_{new} exactly as in s_{last} (so, for any $a \in A$, define $\nu'_s(a) := \nu_s(a)$ for $s \in S$, and $\nu'_{s_{new}}(a) := \nu_{s_{last}}(a)$).◄

Definition 2.3 (Visibility change) Let $M = \langle n, S, \tau, \kappa, \nu \rangle$ be a TV model, with $s_{new} \notin S$; take a set of agents $\{b_1, \dots, b_k\} \subseteq A$, and let $X_i \subseteq A \cup O$ be a set of agents and objects for every b_i, satisfying $b_i \in X_i$. The visibility assignment $[b_1{\leftarrow}X_1, \dots, b_k{\leftarrow}X_k]$ produces the TV model

$$M_{[b_1\leftarrow X_1,\dots,b_k\leftarrow X_k]} = \langle n+1, S \cup \{s_{new}\}, \tau', \kappa', \nu' \rangle$$

in which ***(i)*** τ' preserves the temporal position of states in S, making s_{new} the most recent (so $\tau'(s) := \tau(s)$ for $s \in S$, and $\tau'(s_{new}) := n+1$); ***(ii)*** κ' preserves the colouring assignment for states in S, with the colouring in s_{new} exactly as in s_{last} (so, for any $o \in O$, define $\kappa'_s(o) := \kappa_s(o)$ for $s \in S$, and $\kappa'_{s_{new}}(o) := \kappa_{s_{last}}(o)$); ***(iii)*** ν' is exactly as ν for states in S, with the new s_{new} taking the visibility of s_{last} for agents not mentioned by the assignment, and following the assignment for those agents it mentions (so, for any $a \in A$, define $\nu'_s(a) := \nu_s(a)$ for $s \in S$, with $\nu'_{s_{new}}(a) := \nu_{s_{last}}(a)$ when $a \notin \{b_1, \dots, b_k\}$, and $\nu'_{s_{new}}(b_j) := X_j$ when $a = b_j$). ◄

The operations describe a change in the current situation; in this sense, they are analogous to model operations in *Dynamic Epistemic Logic* (*DEL*; [16,17]). Still, there is an important difference. Typically, *DEL* models describe only the current situation, so model operations return a structure representing also a single situation (the 'next' one). In contrast, while a *TV* model describes how the situation is at the current stage (the state $\tau^{-1}(n)$), it might also describe how the situation was in the past (the other states). Thus, while the operations add a state describing the situation the action produces, they also retain the states of the original model, hence keeping track of the past. In this sense, the *TV* setting can be understood as a 'dynamic temporal': an underlying temporal structure that can be *extended* by dynamic 'model change' operations. Other proposals using similar ideas include [18] (cf. [19,20]), which redefines the operation representing acts of (public and) private communication [11] to preserve previous stages, and [21], whose models 'remember' the initial epistemic situation.

A formal language The language $\mathcal{L}$, for describing *TV* models, contains basic formulas expressing the (high-order) beliefs agents have about the colour of an object, and it is closed under both the standard Boolean operators as well as modalities for describing what will be the case after an action takes place.

Definition 2.4 (Language $\mathcal{L}$) Given A, O and $\{R_o\}_{o\in O}$, formulas ϕ of the language $\mathcal{L}$ are given by

$$\begin{array}{ll} \phi ::= B_{a_1}\cdots B_{a_k}(o{\lhd}c) \mid \neg\phi \mid \phi\wedge\phi \mid [\alpha]\,\phi & \text{for } k\geqslant 1, \{a_1,\ldots,a_k\}\subseteq A, o\in O, c\in R_o \\ \alpha ::= p_1{:=}c_1,\ldots,p_i{:=}c_i \mid b_1{\leftarrow}X_1,\ldots,b_j{\leftarrow}X_j & \text{for } i\geqslant 1, \{p_1,\ldots,p_i\}\subseteq O, c_i\in R_{p_i}, \\ & j\geqslant 1, \{b_1,\ldots,b_j\}\subseteq A, X_i\subseteq A\cup P \text{ with } b_i\in X_i \end{array}$$

Formulas of the form $B_{a_1}\cdots B_{a_k}(o{\lhd}c)$, called *mental attribution formulas*, are read as *"agent a_1 believes that ... that agent a_k believes that o has colour c"*. Other Boolean connectives ($\vee, \rightarrow, \leftrightarrow$) are defined in the standard way. ◀

Formulas in $\mathcal{L}$ are evaluated in a *TV* model with respect its last state s_{last}, the fullest representation of the scenario available up that point. Nevertheless, as the definition shows, the truth-value of formulas is influenced by earlier states.

Definition 2.5 (Semantic interpretation) Let $M = \langle n, S, \tau, \kappa, \nu\rangle$ be a temporal visibility model. The following definitions will be useful.

- Take $\chi := B_{a_1}\cdots B_{a_k}(o{\lhd}c)$. Its *visibility condition* on $s\in S$, denoted by $\text{vis}_\chi(s)$, and listing the requirements for χ to be evaluated at s (agent a_1 can see agent a_2, ..., agent a_{k-1} can see agent a_k, agent a_k can see object o), is given by

$$\text{vis}_\chi(s) \quad \textit{iff}_{\,def} \quad a_2\in\nu_s(a_1)\ \&\ \ldots\ \&\ a_k\in\nu_s(a_{k-1})\ \&\ o\in\nu_s(a_k).$$

- Take $s\in S$ and $t\leqslant\tau(s)$. The *t-predecessor* of s, denoted by $[s]_{-t}$, is the (unique) state appearing exactly t stages before s,[8] and it is formally defined as

$$[s]_{-t} := \tau^{-1}(\tau(s)-t)$$

---

[8] In particular, $[s]_{-0} = s$. Note also how $[s]_{-t}$ is undefined for $t > \tau(s)$.

For evaluating $\chi := B_{a_1} \cdots B_{a_k}(o{\lhd}c)$, the process starts from s_{last}, going 'back in time' one step at the time, looking for a state satisfying χ's visibility condition. If such s' is reached, χ's truth-value depends only on whether o has colour c at s'; otherwise, χ is false. Formally, and by using "⅋" for a natural-language disjunction (just as "&" stands for a natural-language conjunction), the satisfaction relation $\Vdash$ between a *TV* model and a mental attribution formula is given by

$$M \Vdash B_{a_1} \cdots B_{a_k}(o{\lhd}c) \qquad \mathit{iff}_{\,\mathit{def}} \qquad \underset{i=0}{\overset{\tau(s_{last})-1}{\text{⅋}}} \left(\begin{array}{c} \overbrace{\underset{j=1}{\overset{i}{\text{\&}}} \text{not}\ \mathrm{vis}_{B_{a_1} \cdots B_{a_k}(o\lhd c)}([s_{last}]_{-(j-1)})}^{\textsf{no}-\textsf{latter}-\textsf{vis}} \\ \text{\&} \\ \underbrace{\mathrm{vis}_{B_{a_1} \cdots B_{a_k}(o\lhd c)}([s_{last}]_{-i})}_{\textsf{vis}} \ \ \text{\&} \ \ \underbrace{\kappa_{[s_{last}]_{-i}}(o) = c}_{\textsf{col}} \end{array} \right)$$

Thus, $B_{a_1} \cdots B_{a_k}(o{\lhd}c)$ holds at M when there is a state (the quantification indicated by the main disjunction) in which the visibility condition is satisfied (the vis part), the object has the indicated colour (the col part), and there is no 'more recent' state satisfying the visibility condition (the no−latter−vis part).

Boolean operators are interpreted as usual. For 'action' modalities,

$$M \Vdash [\alpha]\,\phi \qquad \mathit{iff}_{\,\mathit{def}} \qquad M_{[\alpha]} \Vdash \phi \qquad\qquad \blacktriangleleft$$

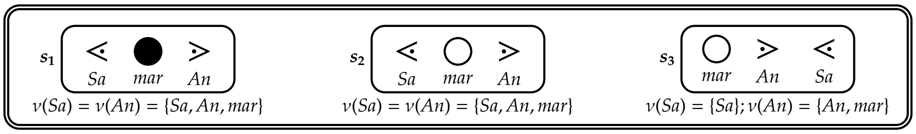
Before an example of the framework at work, there are four points worthwhile to emphasise. ***(i)*** The semantic interpretation of $\chi := B_{a_1} \cdots B_{a_k}(o{\lhd}c)$ captures the discussed intuitive idea. On the one hand, if the visibility condition fails at every state, the formula is false (every disjunct fails in its vis part). On the other hand, if some states satisfy the visibility condition, let s' be the time-wise latest (i.e., $s' := \tau^{-1}(\max\{\tau(s) \mid \mathrm{vis}_{\chi}(s)\})$); then, $M \Vdash \chi$ iff $\kappa_{s'}(o) = c$. ***(ii)*** For the sake of simplicity, we assume that, when an agent a sees an agent b, and b sees an object o, then a in fact sees b *seeing* o, as it should be intuitively the case in order for a formula like $B_a\, B_b(o{\lhd}c)$ to be evaluated.[9] ***(iii)*** The term 'belief' here does not have the strong *EL* reading; it is rather understood as *"truth according to the agent's current information about what has happened so far"* (a form of *default reasoning* [24,25]: the agent assumes that things remain the way she saw them last). ***(iv)*** Attributions to oneself boil down to the col part of the interpretation, given the properties of ν, thus giving any agent full positive introspection.

EXAMPLE 3 Recall the Sally-Anne Task, with its first two stages represented by the model M in Example 2. The story continues with Sally leaving the room, after which she can see neither Anne nor the marble anymore, and Anne can only

---

[9] Notice that visibility of each agent is not 'common knowledge': knowledge relies on visibility, and an agent can see without being seen (Subsection 3.1). Additionally, our simplifying assumption might be a problem for attributions under (semi-)private actions. Work of [22,23] can be especially relevant in that respect.

see the marble. This is represented by an operation extending the model with a new state (s_3) in which both *Sa*'s and *An*'s visibility have changed, yielding the model $M_{[Sa \leftarrow \{Sa\}, An \leftarrow \{An, mar\}]} = M'$ below.



- Does Anne believe that the marble is white? Intuitively, the answer should be *"yes"*, and the system agrees: $M' \Vdash B_{An}(mar \triangleleft white)$ holds, as at s_{last} Anne sees the marble ($mar \in v_{s_3}(An)$), and the marble is indeed white ($\kappa_{s_3}(mar) = white$).
- Does Sally believe that the marble is white? The answer is *"yes"*, but for a different reason: $M' \Vdash B_{Sa}(mar \triangleleft white)$ holds because ***(i)*** although *Sa* cannot see *mar* now (at s_3), ***(ii)*** the last time she saw it (s_2), *mar* was white.
- Does Anne believe that Sally believes that the marble is white? The relevant state is the last time Anne saw Sally looking at the marble, i.e., s_2. Since *mar* is white at s_2, indeed $M' \Vdash B_{An} B_{Sa}(mar \triangleleft white)$.
- Finally, does Sally believe that Anne believes that the marble is white? As before, we can verify that $M' \Vdash B_{Sa} B_{An}(mar \triangleleft white)$. ◀

***TV* models from a modal perspective** Readers familiar with modal logic [26] will have noticed that a *TV* model is just a domain with a predecessor relation (more precisely, a finite linear temporal structure); thus, it can also be described by more standard modal languages. This will be made precise now, in order to make explicit what the semantic evaluation of mental attribution formulas boils down to. For simplicity, the focus will be $\mathcal{L}'$: the fragment of $\mathcal{L}$ that does not include the dynamic modalities $[p_1{:=}c_1, \ldots, p_i{:=}c_i]$ and $[b_1 \leftarrow X_1, \ldots, b_j \leftarrow X_j]$.

A modal language for describing a *TV* model requires special atoms for agents' visibility and objects' colour. For the modalities, evaluating mental attribution formulas might require visiting previous states, so temporal operators are needed. A suitable one for expressing what mental attribution formulas encode is the *since* operator $\mathsf{S}(\phi, \psi)$ [27] (more precisely, its *strict* version, found also in, e.g., [28]), read as *"since ϕ was true, ψ has been the case"*.[10] Given a linear structure $\mathsf{M} = \langle W, \prec, V \rangle$ and $w \in W$, the formula is interpreted as follows.

$$(\mathsf{M}, w) \Vdash \mathsf{S}(\phi, \psi) \quad \textit{iff}_{\,def} \quad \text{there is } u \in W \text{ with } \textbf{\textit{(i)}}\ u \prec w,\ \textbf{\textit{(ii)}}\ (\mathsf{M}, u) \Vdash \phi, \text{ and } \textbf{\textit{(iii)}}\ (\mathsf{M}, v) \Vdash \psi \text{ for every } v \in W \text{ such that } u \prec v \prec w.$$[11]

---

[10] Note: a single 'predecessor' modality is insufficient, as the number of back steps the recursive exploration requires is *a priori* unknown. A modality for its reflexive and transitive closure is still not enough: it takes care of the recursive search for a state satisfying the visibility condition, but on its own cannot indicate that every state up to that point should *not* satisfy it. More on the adequacy of *since* can be found in [27].

[11] Within propositional dynamic logic [29], and in the presence of the converse $\succ$, the *since* modality can be defined as $\mathsf{S}(\phi, \psi) := \langle (\succ; (?\phi \cup ?(\neg\phi \wedge \psi)))^{+} \rangle \phi$, with "?" indicating relational test, ";" indicating sequential composition, "$\cup$" indicating non-deterministic choice, and "$^{+}$" indicating one or more iterations.

Thus, let $\mathcal{L}_{\mathrm{S}}$ be the modal language whose formulas are given by

$$\phi ::= \lessdot_a b \mid \lessdot_a o \mid o{\lhd}c \mid \neg\phi \mid \phi \wedge \phi \mid \mathrm{S}(\phi, \phi)$$

for $a, b \in A$, $o \in O$ and $c \in R_o$. The semantic interpretation of the atoms $\lessdot_a b$, $\lessdot_a o$ and $o{\lhd}c$ over a *TV* 'pointed' model (M, s) is the natural one (look at s's contents, given by ν_s and κ_s); the semantic interpretation of $\mathrm{S}(\phi, \psi)$ is as above, with $\prec$ taken to be the *"strictly earlier than"* relation over states in S, defined as $s \prec s'$ *iff*$_{def}$ $\tau(s) < \tau(s')$. Then, by using the abbreviation $\mathrm{vis}_{a_1 \cdots a_n o} := \lessdot_{a_1} a_2 \wedge \cdots \wedge \lessdot_{a_{k-1}} a_k \wedge \lessdot_{a_k} o$ (so $\mathrm{vis}_{a_1 \cdots a_n o} \in \mathcal{L}_{\mathrm{S}}$ expresses the visibility condition of the formula $B_{a_1} \cdots B_{a_k}(o{\lhd}c)$), the translation $tr : \mathcal{L}' \to \mathcal{L}_{\mathrm{S}}$ is defined as

$$\begin{gathered} tr(B_{a_1} \cdots B_{a_k}(o{\lhd}c)) := (\mathrm{vis}_{a_1 \cdots a_n o} \wedge o{\lhd}c) \vee (\neg\, \mathrm{vis}_{a_1 \cdots a_n o} \wedge \mathrm{S}(\mathrm{vis}_{a_1 \cdots a_n o} \wedge o{\lhd}c, \neg\, \mathrm{vis}_{a_1 \cdots a_n o})), \\ tr(\neg\phi) := \neg tr(\phi), \qquad\qquad tr(\phi \wedge \psi) := tr(\phi) \wedge tr(\psi). \end{gathered}$$

Then, $M \Vdash \phi$ iff $(M, s_{last}) \Vdash tr(\phi)$ holds for any *TV* model M and any $\phi \in \mathcal{L}'$. The crucial case, for mental attribution formulas, is apparent: $tr(B_{a_1} \cdots B_{a_k}(o{\lhd}c))$ holds at s_{last} in M if and only if either the visibility condition holds and the object has the indicated colour ($\mathrm{vis}_{a_1 \cdots a_n o} \wedge o{\lhd}c$), or else the visibility condition fails ($\neg\, \mathrm{vis}_{a_1 \cdots a_n o}$) and there is a state in the past where both visibility and colour were satisfied, and since then visibility has failed ($\mathrm{S}(\mathrm{vis}_{a_1 \cdots a_n o} \wedge o{\lhd}c, \neg\, \mathrm{vis}_{a_1 \cdots a_n o})$). This is exactly what the semantic interpretation of $B_{a_1} \cdots B_{a_k}(o{\lhd}c)$ in M requires.

Bisimulation The translation *tr* provides an insight on the semantic clause for mental attribution formulas. Equally illuminating is a bisimulation for $\mathcal{L}'$.

DEFINITION 2.6 (*TV*-BISIMULATION) Two *TV* models $M = \langle n, S, \tau, \kappa, \nu \rangle$ and $M' = \langle m, S', \tau', \kappa', \nu' \rangle$ (with s_{last} and s'_{last} their respective 'last' states) are said to be *TV-bisimilar* (notation: $M \leftrightarrows M'$) if and only if, for any mental attribution formula $\chi := B_{a_1} \cdots B_{a_k}(o{\lhd}c)$, **(I) Forth:** if there is $t \in S$ such that ***(i)*** $\mathrm{vis}_\chi(t)$ holds, ***(ii)*** $\mathrm{vis}_\chi(r)$ fails for every $r \in S$ with $\tau(t) < \tau(r) \leqslant \tau(s_{last})$, and ***(iii)*** $\kappa_t(o) = c$, then there is $t' \in S'$ such that ***(i)*** $\mathrm{vis}_\chi(t')$ holds, ***(ii)*** $\mathrm{vis}_\chi(r')$ fails for every $r' \in S$ with $\tau'(t') < \tau'(r') \leqslant \tau'(s'_{last})$, and ***(iii)*** $\kappa_{t'}(o) = c$. **(II) Back:** vice versa. ◀

It can be proved that, whenever M and M' are *TV*-bisimilar, both models satisfy the same $\mathcal{L}'$-formulas.[12] The colour of an object is relevant only if some agent can see it (so, no 'atom' clause is needed). Note also how two *TV* models satisfying the same $\mathcal{L}'$-formulas might differ in their cardinality, and also make the same formula true in different ways (e.g., $\neg\, B_a(o{\lhd}c)$ holds in M because, at s_{last}, agent a sees o having a colour other than c, but it holds in M' because, as far as M' is concerned, agent a has never seen o). Finally, notice how, although *TV*-bisimulation implies $\mathcal{L}'$-equivalence, it does not imply $\mathcal{L}$-equivalence. Take $A = \{a\}$ and $O = \{o\}$, with s_1 a state in which a sees o being white, and s_2 one in which a does not see o. Take M to be the model with only s_1, and M' to be the model with both s_1 and s_2. The models are *TV*-bisimilar, hence $\mathcal{L}'$-equivalent.

---

[12] Since $\mathcal{L}'$-formulas are evaluated with respect to a *TV* model's last state, it is enough for a bisimulation to establish a connection between those states, as the definition does.

Yet, they can be distinguished by the formula $[o{:=}black]\, B_a(o \triangleleft black)$ (true in M, false in M'): the different reasons why $\mathcal{L}'$-formulas are made true in bisimilar models become salient when actions enter the picture. For a bisimulation for $\mathcal{L}_S$, it is enough to consider the mutual satisfaction of atoms in bisimilar points, and suitable *Since* conditions, as the ones discussed in [30, p.413].

3 On modelling mental attribution scenarios

The *TV* framework aims to model belief attributions in a more cognitively plausible way (compared with *EL*), revealing features thought of as crucial ingredients of social cognition. Let's justify these claims.

Informational economy On the one hand, a state in a *TV* model contains a bare informational 'minimum': only basic facts regarding objects and agents' visibility. The operations on the model also induce 'minimal' changes, in accordance to the criterion of informational economy in belief revision [31]. On the other hand, the non-standard semantic clause for belief is complex, as the state representing the current situation might not have all information necessary to evaluate a complex belief attribution, and thus the information at other (previous) stages might be needed. A 'backtracking' process might be difficult and time-consuming, depending on how many different states an agent needs to 'remember', and our clause is sensitive to this observation, unlike the usual modal interpretations. The level of complexity that one finds on the *TV* framework for both representing a situation (low) and evaluating mental attributions (high) can be contrasted with what *EL* does, as discussed in Section 1.

Perspective shifting Another important feature, identified in analyses of ToM and formalisations of False-Belief Tasks (*FBT*s), is *perspective shifting* [32]. Successful performance in the tasks (i.e., making correct attributions) requires a perspective shift: stepping into the shoes of another agent. [13] Asking for the visibility condition ensures precisely that agents change perspectives, even if that means having to recall earlier stages. Making multiple shifts, e.g. in complex high-order attributions, may be difficult compared to plainly attributing one's own belief to others, capturing why agents might fail in the tasks.

Principle of inertia A further crucial notion is the *principle of inertia* [6,33,34]: an agent's beliefs are preserved unless there is reason to the contrary. In our case, reason to the contrary amounts to the satisfaction of visibility; if this is not satisfied in the state of evaluation, then, essentially, the agent maintains beliefs formed in earlier stages, where necessary information was available.

Dual process theories of reasoning Besides ToM, the *TV* setting is in agreement with the literature supporting the *dual process theories of reasoning* [35,36,37]. According to them, there are two systems underlying human reasoning. System 1 (the *fast* mode) is quick, unconscious and automatic, often governed by habit, biases and heuristics developed in the course of evolution. System 2 (the *slow* mode) is gradual, deliberate and rule-based, and requires cognitive effort.

---

[13] In fact, unsuccessful performance, e.g. of autistic children, is often connected with a failure in perspective shifting, resulting in the subject reporting her own beliefs [6,33].

System 1 is at play most of the time, constructing our idea of the world with elementary cues and avoiding cognitive overload. When rule-based calculations become necessary, e.g. in face of a demanding task, System 2 takes over, building on inputs of System 1 to slowly produce an output in a step-wise fashion.

We argue that agents' higher order reasoning roughly follows this pattern. System 1 keeps track only of a bare-minimum of information (basic facts), without overloading memory with information that can be later inferred. Whenever a task requires more than what is stored (as higher-order attributions), System 2 takes over, using the inputs of System 1. This is precisely the pattern of our semantics, with our models and updates encoding only basic facts. Whenever a demanding task appears, such as the evaluation of a mental attribution, our agents follow the cognitively hard calculations of our semantic clause.[14] On the basis of elementary facts regarding whom/what they observed, they test certain conditions and trace back earlier states. It is only after this slow and effortful process that they can determine whether a higher-order attribution holds.

3.1 Detailed examples

False-Belief Tasks use stories to test the ability to attribute mental states to others. In what follows, we provide formal representations of some of these storylines, to the level of abstraction allowed by our framework's constructions.

EXAMPLE 4 (FIRST-ORDER *FBT*: THE *Sally-Anne* (*SA*) TASK) The full storyline (Example 1) can be represented within the *TV* framework, modulo minor changes, as already hinted at. ***(1)*** Sally and Anne are in a room, with Sally holding the marble (the model with only state s_1 in Example 2). ***(2)*** Sally puts the marble into the basket (the full model in Example 2). ***(3)*** Sally leaves the room (the model in Example 3). ***(4)*** Anne transfers the marble to the box (the model in Figure 1). The task's last step, Sally coming back to the room, prepares the audience for the crucial question: *"where does Sally believe the marble is?"*. The action changes Sally's visibility (she can see Anne now), but it does not change the crucial fact that she cannot see the marble. Thus, it is not relevant for our purposes.

So, which are Anne's and Sally's final high-order beliefs? According to the framework, with M the model in Figure 1 (top): $M \Vdash B_{Sa}(mar \lhd white) \wedge B_{An}(mar \lhd green)$, and $M \Vdash B_{Sa}\, B_{An}(mar \lhd white) \wedge B_{An}\, B_{Sa}(mar \lhd white)$. ◀

EXAMPLE 5 (SECOND-ORDER *FBT*: THE *chocolate* (*C*) TASK) Adapted from [39], the task is as follows. ***(1)*** Mary and John are in a room, with a chocolate bar in the room's table. ***(2)*** John puts the chocolate into the drawer, then ***(3)*** leaving the room. ***(4)*** Mary transfers the chocolate to the box. ***(5)*** John peeks into the room, without Mary noticing, and sees the chocolate in the box.

The *TV* modelling works stepwise, with the initial situation represented by s_1 (*black* indicates the chocolate is on the table), and each subsequent action adding

---

[14] Although it is always possible to evaluate attributions of any length (like in possible-worlds semantics), our semantic clause offers a mechanism to account for human reasoning limitations, indicated by empirical research, e.g. on working memory [38]. It allows us to trace how many states need to be held in working memory, and therefore explain why attribution-making might fail from some point on.

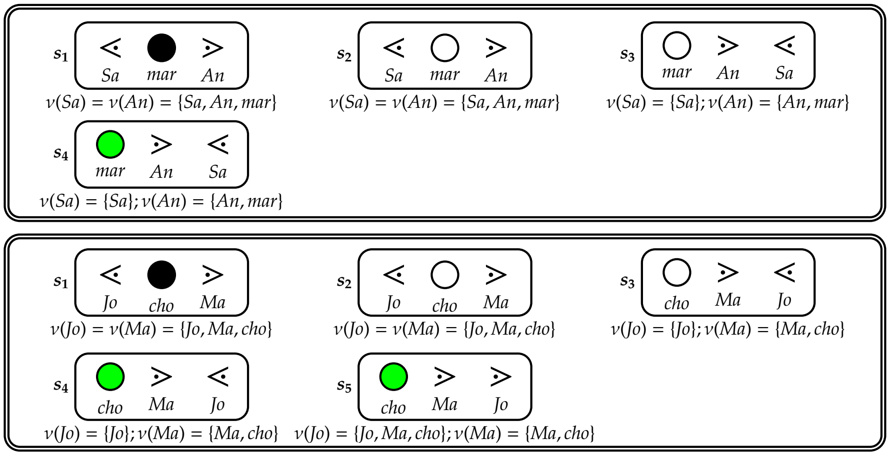


Fig. 1: *TV* representations of Sally-Anne Task (top) and Chocolate Task (bottom).

a state. By putting the chocolate into the drawer (*white*), John produces s_2, and by leaving the room he produces s_3. Mary creates s_4 when she moves the chocolate to the box (*green*), and finally s_5 emerges when John peeks into the room. In the final model, displayed in Figure 1 (bottom), we have the following: ***(i)*** $M \Vdash B_{Ma}(cho \lhd green) \wedge B_{Jo}(cho \lhd green)$, ***(ii)*** $M \Vdash B_{Ma}\, B_{Jo}(cho \lhd white) \wedge B_{Jo}\, B_{Ma}(cho \lhd green)$, and ***(iii)*** $M \Vdash B_{Ma}\, B_{Jo}\, B_{Ma}(cho \lhd white) \wedge B_{Jo}\, B_{Ma}\, B_{Jo}(cho \lhd white)$. ◀

Other *FBT*s (the *Ice Cream Task* [40], the *Puppy Task* [41] and the *Bake-sale* task [42]) can be also represented in the *TV* framework, their crucial ToM features still preserved. Still, some sources of change in zero- or higher- order information in such dynamic scenarios might not be captured by our operations. While conceptually similar examples can fit into our setting, up to some level of abstraction, different operations might be required for other scenarios (Section 4).

3.2 Comparison with other proposals for mental attributions

Through a relational 'preference' framework for modelling different degrees of belief, [43] studies three kinds of agents (including agents on the autism spectrum), each endowed with specific "properties" as higher-order reasoners. Our attempt does not focus on agents with specific strategies when evaluating belief attributions, working instead on *any* agent's reasoning behind such process.

In [6], the authors provide a non-monotonic, closed-world reasoning formalization of first-order *FBT*s, implemented within logical programming. They use *event calculus*, with belief treated as a predicate, and rely on the principle of inertia. While we design a different formalism, we still account for these features without restricting ourselves to specific types of agents or orders of beliefs.

Another interesting logical formalization of *FBT*s is given in [32,33,34]. These papers use a proof-theoretic Hybrid Logic system for identifying perspective shifts, while using inertia. The straightforward difference is that our approach is rather semantic, with models keeping track of the actions involved, and in which the evaluation of mental attributions reflects their cognitive difficulty.

The framework of [44] uses *EL*-beliefs plus special atoms indicating the location of objects and the agents' visibility, then representing changes in the situation as action-model-based acts of (private) communication that rely on agents' visibility.[15] The differences between our proposal and [44] have been discussed: the contrast between complex models that simplify answering mental attribution questions (*EL*) and simple states that require a complex process for deciding high-order belief issues (here). The representation of actions also differs: while [44] uses (a variation of) the heavy action models machinery (for private communication), the actions of visibility and colour change presented here simply modify atomic information. Finally, [44] also proposes two criteria of success in formalizing *FBT*s: ***(i)*** *robustness* (being able to deal with as many *FBT*s as possible, with no strict limit on the order of belief attribution), and ***(ii)*** *faithfulness* (each action of the story should correspond to an action in the formalism in a natural way). The *TV* framework fulfils these requirements: it is robust enough to deal with different *FBT*s (see Subsection 3.1 and the discussion therein), and the actions in the stories have a straightforward representation.

4 Summary and ongoing/future work

This paper has introduced a temporal framework suitable for capturing 'real' agents' mental state attributions. Its most important feature is the contrast between a 'simple' semantic model (encoding only objects' colours and agents' visibility) and a 'complex' clause for interpreting mental state attributions (essentially a temporal *"since"* operator). We have argued for its adequacy towards representing important features of social cognition, as informational economy, perspective shifting, inertia, and connections with dual process theories, with these points exemplified through the modelling of common *FBT*s.

This project presents several lines for further research. On the technical side, there are still aspects of the logical setting to be investigated (e.g., axiomatisation). Equally interesting is the exploration of extensions for modelling more empirical findings. The main points made above on the adequacy of the framework make for a suitable basis for such extensions. Here are two possibilities.

A perspective function The setting can be fine-tuned to capture special types of high-order reasoning (see case-studies of [16]). For example, autistic children tend to fail the *FBT*s because they attribute their own beliefs to others [5]. This and other similar situations can be accommodated through the introduction of a *perspective* function $\pi : A \to (A \to A)$ (with $\pi_a(b) = c$ understood as *"agent a considers agent b to have the perspective of agent c"*), which then can be used to define an appropriate variation of the visibility condition. In this way, an autistic agent a would be one for which $\pi_a(x) = a$ for any $x \in A$, essentially relying only on her own information, and thus attributing her own belief to others.

Different states for different agents at the same stage Another extension is towards capturing scenarios involving communicative actions, including lying

---

[15] For example, the act through which, in the absence of Sally, Anne moves the marble from the basket to the box, is understood as a private announcement through which only Anne is informed about the marble's new location.

and spread of misinformation (e.g., the Puppy Task, the Bake Sale Task) and other manifestations of social cognition (e.g., negotiations, games). With them, it makes sense to include different states for different agents at the same stage, each one of them representing the (potentially different) information different agents might have about the situation at the same stage.

References

1. Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences **1**(4) (1978) 515–526
2. Carruthers, P., Smith, P.K.: Theories of Theories of Mind. CUP (1996)
3. Wimmer, H., Perner, J.: Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition **13**(1) (1983) 103 – 128
4. Wellman, H.M.: From desires to beliefs: Acquisition of a theory of mind. In: Natural theories of mind: Evolution, development and simulation of everyday mindreading. Basil Blackwell, Cambridge, MA, US (1991) 19–38
5. Baron-Cohen, S., Leslie, A.M., Frith, U.: Does the autistic child have a "theory of mind"? Cognition **21**(1) (1985) 37–46
6. Stenning, K., van Lambalgen, M.: Human Reasoning and Cognitive Science. MIT Press (2008)
7. Setoh, P., Scott, R.M., Baillargeon, R.: Two-and-a-half-year-olds succeed at a traditional false-belief task with reduced processing demands. Proceedings of the National Academy of Sciences **113**(47) (2016) 13360–13365
8. Hintikka, J.: Knowledge and Belief. Cornell University Press, Ithaca, N.Y. (1962)
9. Fagin, R., Halpern, J.Y., Moses, Y., Vardi, M.Y.: Reasoning about knowledge. The MIT Press, Cambridge, Mass. (1995)
10. Stalnaker, R.: The problem of logical omniscience, I. Synthese **89**(3) (1991) 425–440
11. Baltag, A., Moss, L.S., Solecki, S.: The logic of public announcements and common knowledge and private suspicions. In Gilboa, I., ed.: Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK-98), Evanston, IL, USA, July 22-24, 1998, Morgan Kaufmann (1998) 43–56
12. Frege, G.: Grundlagen der Arithmetik. Breslau: Wilhelm Koebner (1884)
13. Pelletier, F.J., Elio, R., Hanson, P.: Is logic all in our heads? From naturalism to psychologism. Studia Logica **88**(1) (2008) 3–66
14. Verbrugge, R.: Logic and social cognition. Journal of Philosophical Logic **38**(6) (2009) 649–680
15. van Benthem, J.: Logic and reasoning: Do the facts matter? Studia Logica **88**(1) (2008) 67–84
16. van Ditmarsch, H., van der Hoek, W., Kooi, B.: Dynamic Epistemic Logic. Volume 337 of Synthese Library Series. Springer, Dordrecht, The Netherlands (2008)
17. van Benthem, J.: Logical Dynamics of Information and Interaction. CUP (2011)
18. Yap, A.: Dynamic epistemic logic and temporal modality. In Girard, P., Roy, O., Marion, M., eds.: Dynamic Formal Epistemology. Springer (2011) 33–50
19. Sack, J.: Temporal languages for epistemic programs. Journal of Logic, Language and Information **17**(2) (2008) 183–216
20. Renne, B., Sack, J., Yap, A.: Logics of temporal-epistemic actions. Synthese **193**(3) (2016) 813–849
21. Baltag, A., Özgün, A., Sandoval, A.L.V.: APAL with memory is better. In Moss, L.S., de Queiroz, R.J.G.B., Martínez, M., eds.: 25th International Workshop WoLLIC 2018. Volume 10944 of Lecture Notes in Computer Science., Springer (2018) 106–129

22. Gasquet, O., Goranko, V., Schwarzentruber, F.: Big brother logic: visual-epistemic reasoning in stationary multi-agent systems. Autonomous Agents and Multi-Agent Systems **30**(5) (2016) 793–825
23. Charrier, T., Herzig, A., Lorini, E., Maffre, F., Schwarzentruber, F.: Building epistemic logic from observations and public announcements. In: Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning. KR'16, AAAI Press (2016) 268–277
24. Reiter, R.: A logic for default reasoning. Artificial Intelligence **13**(1-2) (1980) 81–132
25. Ben-David, S., Ben-Eliyahu-Zohary, R.: A modal logic for subjective default reasoning. Artificial Intelligence **116**(1-2) (2000) 217–236
26. Blackburn, P., de Rijke, M., Venema, Y.: Modal logic. CUP, Cambridge, UK (2001)
27. Kamp, H.: Tense Logic and the Theory of Linear Order. PhD thesis, University of California (1968)
28. Burgess, J.P.: Axioms for tense logic. I. "since" and "until". Notre Dame Journal of Formal Logic **23**(4) (1982) 367–374
29. Harel, D., Kozen, D., Tiuryn, J.: Dynamic Logic. MIT Press, Cambridge, USA (2000)
30. Kurtonina, N., De Rijke, M.: Bisimulations for temporal logic. Journal of Logic, Language and Information **6**(4) (1997) 403–425
31. Gärdenfors, P.: Knowledge in Flux. Modelling the Dynamics of Epistemic States. MIT Press (1988)
32. Braüner, T.: Hybrid-logical reasoning in the smarties and Sally-Anne tasks. Journal of Logic, Language and Information **23**(4) (2014) 415–439
33. Braüner, T.: Hybrid-logical reasoning in the smarties and Sally-Anne tasks: What goes wrong when incorrect responses are given? In: Proceedings of the 37th Annual Meeting of the Cognitive Science Society, Pasadena, California, USA, Cognitive Science Society (2015) 273–278
34. Braüner, T., Blackburn, P., Polyanskaya, I.: Second-order false-belief tasks: Analysis and formalization. In: 23rd International Workshop, WoLLIC 2016, Springer (2016) 125–144
35. Kahneman, D.: Thinking, fast and slow. Farrar, Straus and Giroux, New York (2011)
36. Evans, J.: Dual process theories. In Ball, L., Thompson, V., eds.: The Routledge International Handbook of Thinking and Reasoning. Routledge (2018) 151–64
37. Stanovich, K.E., West, R.F.: Individual differences in reasoning: Implications for the rationality debate? Behavioral and Brain Sciences **23**(5) (2000) 645–665
38. Cowan, N.: The magical number 4 in short-term memory: A reconsideration of mental storage capacity. The Behavioral and Brain Sciences **24** (2001) 87–114
39. Flobbe, L., Verbrugge, R., Hendriks, P., Krämer, I.: Children's application of theory of mind in reasoning and language. Journal of Logic, Language and Information **17**(4) (2008) 417–442
40. Perner, J., Wimmer, H.: "John *Thinks* that Mary *Thinks* that . . . " attribution of second-order beliefs by 5- to 10-year-old children. Journal of Experimental Child Psychology **39**(3) (1985) 437–471
41. Sullivan, K., Zaitchik, D., Tager-Flusberg, H.: Preschoolers can attribute second-order beliefs. Developmental Psychology **30** (1994) 395–402
42. Hollebrandse, B., van Hout, A., Hendriks, P.: Children's first and second-order false-belief reasoning in a verbal and a low-verbal task. Synthese **191**(3) (2014) 321–333
43. Ditmarsch, H.V., Labuschagne, W.: My beliefs about your beliefs: A case study in theory of mind and epistemic logic. Synthese **155**(2) (2007) 191–209
44. Bolander, T.: Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic. In van Ditmarsch, H., Sandu, G., eds.: Jaakko Hintikka. Springer (2018) 207–236