Observation, Action, and Social Cognition

Jan van Eijck

(joint work – in progress – with Thomas Bolander)

Réunion DynRes, Nancy, April 14, 2015

Abstract

We present a very simple model of how false beliefs can be created by the interplay of action and observation or lack of observation. This is enough to model several false-belief tasks that are being used to attribute theory of mind to social agents in interaction, like the Sally-Anne task. The modelling is intended to shed new light on the interpretation of the results of action on belief in the light of changes in observational opportunities for the agents.

Key words: Social cognition, Sally-Anne tasks, theory of mind, belief revision, false beliefs.

Theory of Mind

One of the key questions in attributing social intelligence to animals or small children is whether they have *theory of mind* — the ability to attribute mental states to others, and to understand that someone else may be in a different mental state than their own. See [Ver09].

Einstein the Pig

For an earlier attempt to formalize false belief tasks in dynamic epistemic logic see [Bol14]. This new approach represents a considerable simplification, but still satisfies two criteria:

Robustness The formalism should be able to deal with a large range of false belief tasks, with no strict limit to the order of belief attribution.

Faithfulness Each action in the false belief task should correspond to an action *or a sequence of actions* in the formalism, and the connection should be straightforward.

Language

We are going to keep this as simple as possible, by focussing on the bare minimum that is needed to model the kind of social interaction that makes false belief tasks interesting.

Let P be a finite set of basic propositions and let A be finite set of agents. In the following BNF definition, assume p ranges over P, and a over A.

$$\varphi \ ::= \ \top \mid \bot \mid p \mid c_a p \mid \neg\varphi \mid \varphi \wedge \varphi \mid B_a\varphi \mid [\alpha]\varphi$$

$$\alpha \ ::= \ (a, p := \varphi) \mid (a, ?p) \mid (a, +p) \mid (a, -p) \mid \alpha; \alpha$$

Call this language $\mathcal{L}$, and call α the *actions* of $\mathcal{L}$. We will call composite actions *plans* or *scenarios*.

We assume the usual abbreviations for $\vee$, $\rightarrow$, $\leftrightarrow$. We use $\check{B}_a\varphi$ for $\neg B_a \neg\varphi$, and $\langle\alpha\rangle\varphi$ for $\neg[\alpha]\neg\varphi$.

The intention is that $c_a p$ expresses that p is within the compass (range of observation and action) of agent a, that B_a is interpreted as belief, while $[\alpha]\varphi$ asserts that if plan α succeeds, then formula φ will hold in the result state, and $\langle\alpha\rangle\varphi$ asserts that the plan α succeeds, with φ as a result.

The plans represent attempts by agents to change or inspect the world, or to change one's abilities to change or inspect the world.

Attempts at Changing the World

The action $p := \varphi$ is an attempt at change.

If agent a performs this, it depends on whether p is within the agent's *compass* what will happen.

If p is within the agent's compass, the value of p will be reset, otherwise the action aborts. If the action succeeds, its result will be visible to agents that have p within their compass, invisible to other agents.

If an action takes place that is invisible to an agent, a false belief may be created.

The Compass of an Agent

For the notion of *compass*, think of

> Love's not Time's fool, though rosy lips and cheeks
> *within his bending sickle's compass* come.
>
> Shakespeare, Sonnet 116

So the compass of an agent is his or her range of observation and command.

Sonnet 116

Let me not to the marriage of true minds
Admit impediments. Love is not love
Which alters when it alteration finds,
Or bends with the remover to remove:
O no; it is an ever-fixed mark,
That looks on tempests, and is never shaken;
It is the star to every wandering bark,
Whose worth's unknown, although his height be taken.
Love's not Time's fool, though rosy lips and cheeks
Within his bending sickle's compass come;
Love alters not with his brief hours and weeks,
But bears it out even to the edge of doom.
If this be error and upon me proved,
I never writ, nor no man ever loved.

Attempts at Inspecting the World

The action $?p$ is an attempt to inspect the value of p.

If agent a performs this, it again depends on whether p is within the agent's *compass* what will happen.

If p is within the agent's compass, a will learn the value of p, otherwise the action will abort. If the action succeeds, the other agents that have p within their compass learn that a has learned the value of p, while the other agents mistakenly believe that nothing has happened.

An inspect action $?p$ may correct a false belief about p, that is, inspection may result in *belief revision*.

Actions that Change the Compass

The action $+p$, when performed by agent a, always succeeds, and has the effect of bringing p within a's compass.

The action $-p$, when performed by agent a, always succeeds, and has the effect of removing p from a's compass.

Compass Models

A compass model $\mathcal{M}$ is a quadruple (W, R, V, C) where

- W is a finite, non-empty set of worlds,
- R is a function that assigns to each $a \in A$ a binary relation $R(a)$ on W that is transitive, serial and euclidean. We will write $R(a)$ as R_a, and we will use $R_a(w)$ for the set of worlds
$$\{v \in W \mid R_a(w, v)\}.$$
- V is a valuation on W, i.e. V is a function from W to $\mathcal{P}(P)$. The members of $V(w)$ are the basic propositions that are true at w.
- C is a compass function on A, i.e., C is a function from A to $\mathcal{P}(P)$ that assigns to any $a \in A$ thet set of basic propositions that are within a's compass.

Truth Definition

Let $\mathcal{M} = (W, R, V, C)$ be a compass model. Let $w \in W$.

- $\mathcal{M}, w \models \top$ always.
- $\mathcal{M}, w \models \bot$ never.
- $\mathcal{M}, w \models p$ iff $p \in V(w)$.
- $\mathcal{M}, w \models c_a p$ iff $p \in C(a)$.
- $\mathcal{M}, w \models \neg\varphi$ iff it is not the case that $\mathcal{M}, w \models \varphi$.
- $\mathcal{M}, w \models \varphi_1 \wedge \varphi_2$ iff both $\mathcal{M}, w \models \varphi_1$ and $\mathcal{M}, w \models \varphi_2$.
- $\mathcal{M}, w \models B_a\varphi$ iff it holds for all $v \in R_a(w)$ that $\mathcal{M}, v \models \varphi$.
- $\mathcal{M}, w \models [\alpha]\varphi$ iff either $[\![\alpha]\!](\mathcal{M}, w) = \uparrow$ (undefined), or $[\![\alpha]\!](\mathcal{M}, w) \models \varphi$, where $[\![\cdot]\!]$ is defined as below.

Action Updates

$[\![\alpha]\!]$ is a partial function from pairs of compass models and states to pairs of compass models and states.

Assume $\mathcal{M} = (W, R, V, C)$ and $w \in W$. We will model the change operator $p := \varphi$ by creating new sets of worlds $W \times \{1, 2\}$, i.e., the members of W' are pairs $(v, 1)$ or $(v, 2)$, with $v \in W$. We will write $(v, 1)$ as $v1$ and $(v, 2))$ as $v2$.

Change

- $[\![(a, p := \varphi)]\!](\mathcal{M}, w)$ is undefined in case $p \notin C(a)$.
  Explanation: if p is not in the compass of a, the attempt at change aborts.

- $[\![(a, p := \varphi)]\!](\mathcal{M}, w)$ equals $(\mathcal{M}', w1)$ in case $p \in C(a)$.
  The new actual world is the old actual world, with index 1.
  $\mathcal{M}' = (W', R', V', C)$ is given by:
  - $W' = W \times \{1, 2\}$.
  - R' is given by

$$
\begin{aligned}
R'_b(v1) &= \begin{cases} \{z1 \mid z \in R_b(v)\} & \text{if } p \in C(b) \\ \{z2 \mid z \in R_b(v)\} & \text{otherwise} \end{cases} \\
R'_b(v2) &= \{z2 \mid z \in R_b(v)\}
\end{aligned}
$$

    Explanation: the worlds with index 1 are visible for the agents that have p in their compass, while the other agents see the corresponding worlds with index 2 instead.

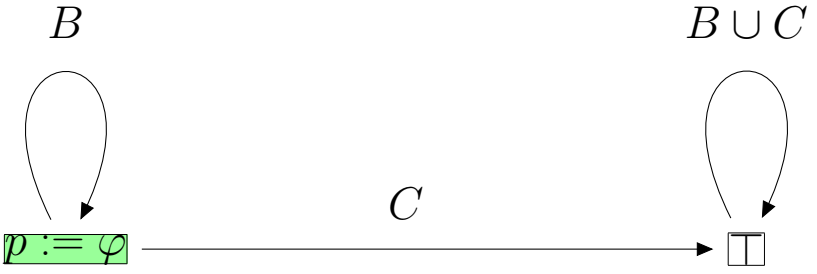
– V' is given by

$$\begin{aligned} V'(v1) &= \begin{cases} V(v) \cup \{p\} & \text{if } \mathcal{M}, v \models \varphi \\ V(v) - \{p\} & \text{if } \mathcal{M}, v \not\models \varphi \end{cases} \\ V'(v2) &= V(v). \end{aligned}$$

Explanation: if p is in the compass of a, p gets the value of φ (in the worlds with index 1) but this is invisible to agents that do not have p within their compass (the worlds with index 2).

Comparison With Action Model Update

The update result can be compared to the action model update [BM04], with added substitutions for modelling factual change [BvEK06], with the following action model, where it is assumed that B are the agents that have p within their compass, and C are the agents that have not. The coloured frame indicates the actual action.



Test

- $[\![(a, ?p)]\!](\mathcal{M}, w)$ is undefined in case $p \notin C(a)$.
- $[\![(a, ?p)]\!](\mathcal{M}, w)$ equals $(\mathcal{M}', w)$ in case $p \in C(a)$.
  Here $\mathcal{M}' = (W', R', V', C)$, with

$$\begin{aligned} W' &= \{v1 \mid v \in W \text{ and } \mathcal{M}, v \models p\} \\ &\cup \{v2 \mid v \in W \text{ and } \mathcal{M}, v \not\models p\} \\ &\cup \{v3 \mid v \in W\} \end{aligned}$$

  R' given by:

$$\begin{aligned} R'_a(v1) &= \begin{cases} \{z1 \mid z \in R_a(v)\} & \text{if this set } \neq \emptyset, \\ \{v1\} & \text{otherwise} \end{cases} \\ R'_a(v2) &= \begin{cases} \{z2 \mid z \in R_a(v)\} & \text{if this set } \neq \emptyset, \\ \{v2\} & \text{otherwise} \end{cases} \\ R'_a(v3) &= \{z3 \mid z \in R_a(v)\}. \end{aligned}$$

If $b \neq a, p \in C(b)$ then

$$R'_b(v1) = R'_b(v2) = \{z1 \mid z \in R_b(v)\} \cup \{z2 \mid z \in R_b(v)\},$$

$$R'_b(v3) = \{z3 \mid z \in R_b(v)\},$$

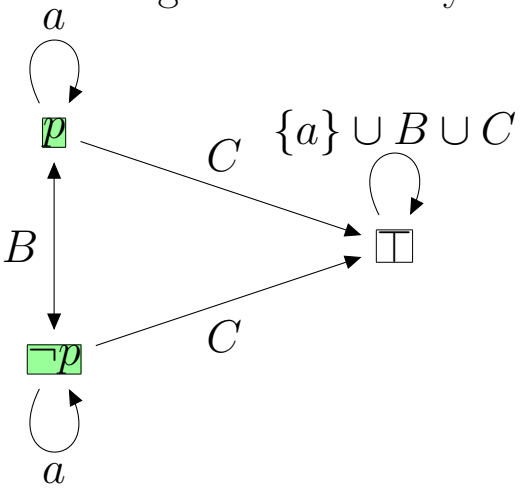
if $p \notin C(b)$ then

$$R'_b(v1) = R'_b(v2) = R'_b(v3) = \{z3 \mid z \in R_b(v)\}.$$

V' is given by $V'(vx) = V(v)$.

Comparison With Action Model Update

To explain the semantics of the update with $(a, ?p)$, consider the action model, where a learns the value of p, the B agents are aware of this, and the C agents mistakenly believe that nothing happens.



The actions with the tests for p and for $\neg p$ are both actual. What actually happens depends on the value of p in the input model.

Exception

- Consider the case where $p \in C(a)$ but we are in a world v with the truth value of p at v different from the truth value of p at *any* of the R_a-successors of v.
- In this case it is impossible for a to learn the value of p by 'arrow elimination'.
- We remove the inconsistency by cutting the link between v and its R_a successors while adding an a-self loop at v. Note that such 'belief revision' is beyond the power of action model update, where the update actions can only *remove* arrows.

Compass Change

- $[\![(a, +p)]\!](\mathcal{M}, w)$ equals $(\mathcal{M}', w)$, where $\mathcal{M}' = (W, R, V, C')$, with $C'(a) = C(a) \cup \{p\}$, and $C'(b) = C(b)$ for $b \neq a$.
- $[\![(a, -p)]\!](\mathcal{M}, w)$ equals $(\mathcal{M}', w)$, where $\mathcal{M}' = (W, R, V, C')$, with $C'(a) = C(a) - \{p\}$, and $C'(b) = C(b)$ for $b \neq a$.

Sequencing

- $[\![\alpha_1; \alpha_2]\!](\mathcal{M}, w) = [\![\alpha_1]\!]([\![\alpha_2]\!](\mathcal{M}, w))$.

Example 1 (The Sally-Anne Scenario [WP83]) *Consider the following scenario involving the two agents s (Sally) and a (Anne). In the original story Sally and Anne are hiding a marble, but let's simplify the task from hiding a marble to making a basic proposition m true or false. Assume that initially both s and a have m in their compass: $C(s) = C(a) = \{m\}$. Then the Sally-Anne scenario looks like this:*

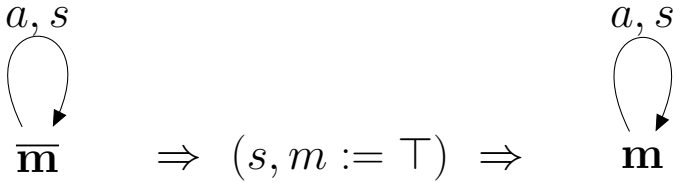
$$(s, m := \top); (s, -m); (a, m := \bot); (s, +m).$$

First Sally makes m true. Next, m gets removed from Sally's compass (Sally leaves the room). Then a makes m false. Finally Sally returns to the room, bringing m within her compass again.

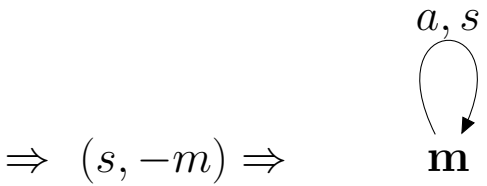
It will be useful to spell out the semantics. starting from a model where both Sally and Anne know that m is false (say).

Example 2 (Semantics of the Sally-Anne Scenario) *We start with a situation where m is false, and both agents know this, and then show the results of performing the action updates one by one.*

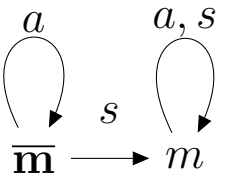
In the belief models, the actual world is in bold.



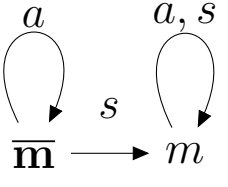
$C(a) = C(s) = \{m\}$ $\qquad$ $C(a) = C(s) = \{m\}$



$C(a) = \{m\}, C(s) = \emptyset$

$\Rightarrow\ (a, m := \bot) \Rightarrow$ 

$$C(a) = \{m\}, C(s) = \emptyset$$

$\Rightarrow\ (s, +m) \Rightarrow$ 

$$C(a) = C(s) = \{m\}$$

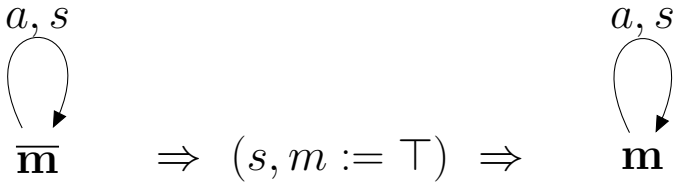
In the final model, Anne has a true belief about m, Sally has a false belief about m, and Anne is aware of Sally's false belief.

In the DEL formalization of the Sally-Anne false belief task in [Bol14] there is a problem with a shortened version of the protocol, where Sally does not leave the room. Let us see whether the present formalization fares better.

Example 3 (Shortened Sally-Anne Scenario) *Sally does not leave the room. This is rendered simply as:*

$$(s, m := \top); (a, m := \bot).$$

And the representation we get, again starting out from the situation where m is false and both s and a know this.



$$C(a) = C(s) = \{m\} \qquad\qquad C(a) = C(s) = \{m\}$$

$$\Rightarrow \ (a, m := \bot) \ \Rightarrow \quad \overset{a,\, s}{\circlearrowright}\ \overline{\mathbf{m}}$$

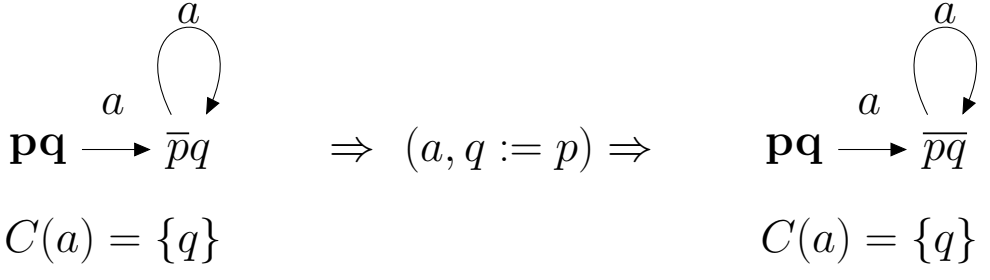
$$C(a) = C(s) = \{m\}$$

So we end up in the situation that we started with, with both Sally and Ann knowing that m is false. This is correct.

Example 4 (False Beliefs About Actions) *Consider an agent a that has a true belief about q and a false belief about p. Let us say that p and q are both true, but a believes that p is false.*

Suppose a has q within his compass. Then the action $(a, q := p)$ will result in a false belief about q.

In the actual world of the initial model, $q \wedge B_a q$ is true.



In the actual world of the resulting model, $q \wedge B_a \neg q$ is true.

Another task that has been used in tests for social cognitive ability is the so-called "second-order chocolate task."

Example 5 (Second Order Chocolate Scenario) *John and Mary are in a room, with a chocolate bar. John puts the chocolate in the drawer and leaves the room. Mary transfers the chocolate to the box. John secretly observes where Mary has put the bar.*

What does Mary believe about where John thinks the chocolate is?

To formalize this, we use the action of performing a test for the truth of a basic proposition. Again, we simplify the task of hiding the chocolate as a change between c and $\overline{c}$. Let j and m be the agents John and Mary. Let's assume c is initially true, and both John and Mary know this. Then the scenario that unfolds is:

$$(j, c := \bot); (j, -c); (m, c := \top); (m, -c); (j, +c); (j, ?c); (m, +c).$$

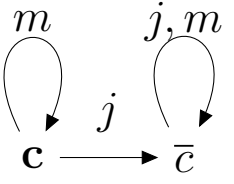
j makes c true, next removes c from his compass. Then m makes c false again, next removes c from her compass. Next, j brings c within his compass again, and observes its value. Finally, m brings c within her compass. The result should be that m has a correct belief about c, but attributes to j a false belief about c. I.e., in the final update result the formulas c, $B_m c$ and $B_m B_j \neg c$ are all true.

$\Rightarrow (j, c := \bot) \Rightarrow$

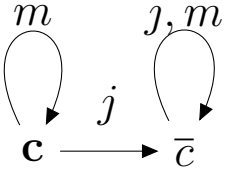
$C(j) = C(m) = \{c\}$ $\qquad$ $C(j) = C(m) = \{c\}$

$\Rightarrow (j, -c) \Rightarrow$

$C(j) = \emptyset, C(m) = \{c\}$

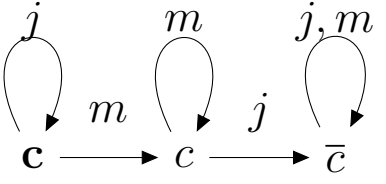
$\Rightarrow (m, c := \top) \Rightarrow$ 

$$C(j) = \emptyset, C(m) = \{c\}$$

$\Rightarrow (m, -c); (j, +c) \Rightarrow$ 

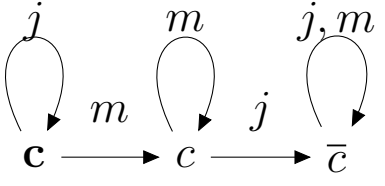
$$C(j) = \{c\}, C(m) = \emptyset$$

$\Rightarrow (j, ?c) \Rightarrow$



$C(j) = \{c\}, C(m) = \emptyset$

$\Rightarrow (m, +c) \Rightarrow$



$C(j) = C(m) = \{c\}$

It is clear from the picture that this update sequence does indeed create a second-order false belief.

Axiomatisation

The base logic is multimodal KD45, so we have rules Modus Ponens and B Necessitation:

$$\frac{\varphi \quad \varphi \to \psi}{\psi}$$

$$\frac{\varphi}{B_a \varphi}$$

Axioms are all propositional validities, plus the K,D, 4 and 5 axioms for B. We have to add axioms expressing that the compass is global, in the sense that the compass values are the same at every world.

$$
\begin{array}{ll}
K & B_a(\varphi \to \psi) \to B_a\varphi \to B_a\psi \\
D & B_a\top \\
4 & B_a\varphi \to B_aB_a\varphi \\
5 & \neg B_a\varphi \to B_a\neg B_a\varphi \\
C^+_{aa} & c_ap \to B_ac_ap \\
C^+_{ba} & c_ap \to B_bc_ap \\
C^-_{aa} & \neg c_ap \to B_a\neg c_ap \\
C^-_{ba} & \neg c_ap \to B_b\neg c_ap
\end{array}
$$

This takes care of the static part of the language. To axiomatize the dynamic part of the language, use the standard DEL approach by means of reduction axioms. This shows that the action modalities do not increase the expressive power of the base language.

Planning

A plan is a non-empty finite sequence of atomic actions α. Let $|\alpha|$ denote the size (length) of a plan α.

Question

Starting from a situation where two agents have true beliefs about a proposition p, what is the size of the smallest plan that will create a false first order belief about p?

Answer

Suppose that a has p in its compass and b has not. Then a single action $(a, p := \neg p)$ is enough. This is essentially the Sally-Ann false belief plan.

Question

Starting from a situation where two agents have true beliefs about a proposition p, what is the size of the smallest plan that will create a false second order belief about p?

Answer

We need size at least 4. For suppose that a has p in its compass and b has not. Then the shortest plan that will create a false second order belief is:

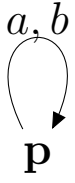
$$(a, p := \neg p); (a, -p); (b, +p); (b, ?p).$$

This is essentially the second order chocolate plan.

Note that the following plan will not succeed in creating a second-order false belief:

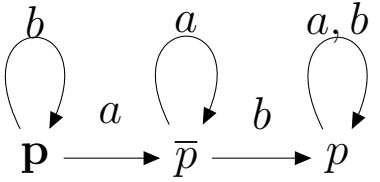
$$(a, p := \neg p); (a, -p); (b, +p); (b, p := \neg p).$$

For suppose that initially p is true. Then the result of carrying out the plan is:



$$\Rightarrow (a, p := \neg p); (a, -p); (b, +p); (b, p := \neg p) \Rightarrow$$

$$C(a) = \{p\}, C(b) = \emptyset$$



$$C(a) = \emptyset, C(b) = \{p\}$$

This creates a second order *true* belief about p.

Plan Length Operators

Extend the language with an infinite set of operators $[n]$, where $n \in \mathbb{N}^+$, and with the following semantics:

$$\mathcal{M}, w \models [n]\varphi \text{ iff for all } \alpha \text{ with } |\alpha| \leq n :$$

$$\text{if } [\![\alpha]\!](\mathcal{M}, w) \neq \uparrow \text{ then } [\![\alpha]\!](\mathcal{M}, w) \models \varphi.$$

Define $\langle n \rangle \varphi$ as $\neg[n]\neg\varphi$.

$\langle n \rangle \varphi$ expresses that some plan of length at most n makes φ true.

This does not increase the expressive power of the language $\mathcal{L}$.

It is possible to exhaustively enumerate all plans up to a given size (this uses the assumption that the sets of agents and propositions are finite), so $[n]\varphi$ can be viewed as an abbreviation of a (very long) formula.

Arbitrary Plans

Consider the operator $[*]$, with the following semantics:

$$\mathcal{M}, w \models [*]\varphi \text{ iff for all } \alpha \text{ it holds that}$$

$$\text{if } [\![\alpha]\!](\mathcal{M}, w) \neq \uparrow \text{ then } [\![\alpha]\!](\mathcal{M}, w) \models \varphi.$$

Define $\langle * \rangle \varphi$ as $\neg[*]\neg\varphi$.

The formula $[*]\varphi$ is not definable in $\mathcal{L}$.

Let $\mathcal{L}^*$ be $\mathcal{L}$ extended with $[*]$.

We believe model checking for $\mathcal{L}^*$ is still decidable.

Questions that Need to be Answered

Question What is the complexity of model checking for $\mathcal{L}^*$?

Question What does a complete axiomatisation of $\mathcal{L}^*$ look like?

Question Is satisfiability for $\mathcal{L}^*$ still decidable?

References

[BM04] A. Baltag and L.S. Moss. Logics for epistemic programs. *Synthese*, 139(2):165–224, 2004.

[Bol14] Thomas Bolander. Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic. In Andreas Herzig and Emiliano Lorini, editors, *Proceedings of the European Conference on Social Intelligence (ECSI-2014)*, volume 1283 of *CEUR Workshop Proceedings*, pages 87–107, 2014.

[BvEK06] J. van Benthem, J. van Eijck, and B. Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.

[Ver09] Rineke Verbrugge. Logic and social cognition. *Journal of Philosophical Logic*, 38(6):649–680, 2009.

[WP83] Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983.