Implementing Theory of Mind on a Robot Using Dynamic Epistemic Logic

Lasse Dissing and **Thomas Bolander**
DTU Compute, Denmark
{ldiha, tobo}@dtu.dk

Abstract

Previous research has claimed dynamic epistemic logic (DEL) to be a suitable formalism for representing essential aspects of a Theory of Mind (ToM) for an autonomous agent. This includes the ability of the formalism to represent the reasoning involved in false-belief tasks of arbitrary order, and hence for autonomous agents based on the formalism to become able to pass such tests. This paper provides evidence for the claims by documenting the implementation of a DEL-based reasoning system on a humanoid robot. Our implementation allows the robot to perform cognitive perspective-taking, in particular to reason about the first- and higher-order beliefs of other agents. We demonstrate how this allows the robot to pass a quite general class of false-belief tasks involving human agents. Additionally, as is briefly illustrated, it allows the robot to proactively provide human agents with relevant information in situations where a system without ToM-abilities would fail. The symbolic grounding problem of turning robotic sensor input into logical action descriptions in DEL is achieved via a perception system based on deep neural networks.

1 Introduction

With the coming of sophisticated service robots, human-robot interaction is becoming an increasingly important challenge. Robots are no longer confined to closed and static environments like factory floors, but instead, have to operate in dynamic multi-agent environments where robust interaction with non-technical human users is essential. For these robots to be accepted by the users, they will need to possess basic social skills and behave in a socially acceptable manner [Dautenhahn, 2007]. Consider a service robot handling small logistic tasks in an office space. If the robot sees an employee searching for something the robot knows has been moved, it would be rude for the robot not to notify her. Conversely, it will be a great annoyance if the robot repeatedly informs employees about the location of things they already correctly know where are—or are not even looking for.

To solve these problems in a *human-centric* way, the robot must be able to take the perspective of the employees and reason about their world views. This issue is already hampering real-world deployment of robots. Studies of logistics robots in hospitals [Barras, 2009] and office spaces [Mitsunaga *et al.*, 2008] include multiple examples of robots misbehaving due to a lack of understanding of the social context—resulting in frustration among the users.

To reason about other agents and their world views, one needs a *Theory of Mind* (ToM). Having a ToM means having the ability to understand and reason about the mental state of other agents, e.g. their beliefs, intentions, desires, and emotions [Premack and Woodruff, 1978]. In this paper, we will almost exclusively restrict attention to beliefs. Endowing a robot with the ability to reason about the beliefs of other agents is a prerequisite for acting socially acceptable. To decide whom to inform about what and when in the previous office space example requires the robot to be able to reason about what the employees already know, what they might falsely believe, and what they need to be informed about.

In developmental psychology, one of the standard methods to test the strength of a human child's ToM is *false-belief tasks* (Section 3). In these tests, the child is presented with a story involving multiple characters, where one or more of the characters end up with a false belief. The child is then asked a series of questions revealing whether she has correctly modelled the mental states (beliefs) of the characters.

False-belief tasks can be categorized by the *depth of reasoning* needed to solve the task, that is, how many recursive perspective shifts are needed. *First-order (false) beliefs* are (false) beliefs about the physical state of the world, while nth-order (false) beliefs are (false) beliefs about another agent's $(n-1)$th-order beliefs. An *nth-order false-belief task* is then one that tests the ability to attribute nth-order false beliefs to others (testing the nth-order ToM).

False-belief tasks can also be tested on robots, similarly to how they are tested on children. While these tests could in principle be passed with just cameras and a microphone, embodying the observing agent in a humanoid robot facilitates better human-robot interaction for the human participants [Złotowski *et al.*, 2015]. Lemaignan *et al.* [2015] finds that while reasoning about first-order beliefs is fairly well studied in the field of Human-Robot Interaction (HRI), higher-order reasoning remains an important next step. In this paper, we consider both first- and higher-order reasoning, allowing our robot to pass both first- and higher-order false-belief tasks.

Bolander [2018] argues that *Dynamic Epistemic Logic* (DEL) is a suitable formalism for representing a ToM and for potentially allowing an autonomous agent to pass false-belief tasks of arbitrary order. The argument is mainly that DEL allows the representation of both 1) first- and higher-order beliefs via *epistemic models* and 2) modelling of dynamic consequences of actions via *action models*. In this paper, we follow the approach suggested by Bolander and implement a ToM in a humanoid robot using DEL. The DEL model keeps track of all first- and higher-order beliefs of all observed agents, and how these beliefs change dynamically when events occur in the environment. Using such models, we can ensure that the robot can robustly pass a general class of false-belief tasks. The only possible source of failure to pass a false-belief task is the perception system (for instance due to the robot misclassifying an object or missing an event).

The contribution of this paper is to: 1) describe a robotic implementation capable of higher-order ToM reasoning (Section 4 and 5); 2) demonstrate its capability to pass first- and higher-order false-belief tasks (Section 6); and 3) briefly illustrate how ToM reasoning can play an important role in proactively helping human agents (also Section 6).

2 Related Work

Previous work on solving false-belief tasks with autonomous agents appear e.g. in the work Arkoudas and Bringsjord [2008], who use the event calculus to develop virtual characters capable of passing first-order false-belief tasks. Breazeal *et al.* [2009] develop a physical robot solving first-order tasks using a simulation-theoretic approach. The robot maintains a distinct belief base for every agent in its environment, representing the first-order beliefs of that agent. The robot filters all percepts such that an agent's belief base is only updated if the robot believes that the percept was visible to the agent. Sindlar *et al.* [2009] solve a first-order task in a virtual setting by adding a set of belief-tracking rules to the BDI model based on the 2APL agent programming language. Milliez *et al.* [2014] use a spatio-temporal belief base to maintain a complex 3D model of the world, allowing to represent the visual perspective of the different agents. Similarly to Brezeal *et al.*, they can solve a first-order false-belief task by maintaining a distinct belief base for each agent.

In essence, all of these prior works are based upon a simulation-theoretic approach where the beliefs of other agents are tracked by maintaining a distinct representation simulating the agent's view, i.e., "How would my belief base look like if I had received the percepts of the other agent". While higher-order reasoning up to some arbitrary bounded depth could in principle be achieved in this approach by using recursive belief bases, all of the surveyed implementations only support first-order belief attribution. Our ToM system based on DEL differs by supporting higher-order belief attribution to unbounded depth, one of the main attractions of models based on epistemic logic and DEL.

Recently, ToM has also been approached with machine learning techniques [Rabinowitz *et al.*, 2018; Nematzadeh *et al.*, 2018]. This is a very interesting alternative to building symbolic models. However, a crucial advantage of our approach is that by building symbolic DEL models that provably correctly keep track of all first- and higher-order beliefs (up to perception failures), our approach leads to a robot that will provably pass any first- or higher-order false-belief task within the bounds of the supported action types (and again up to perception failures), even previously unseen tasks, as we are going to illustrate in Section 6.2.

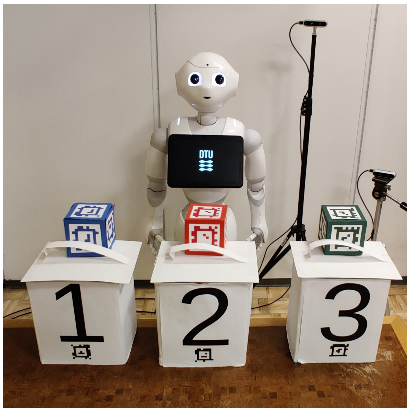

Figure 1: The Pepper robot and the two external cameras

3 False-Belief Tasks

A common false belief task is the *Sally-Anne task* [Wimmer and Perner, 1983] in which the child is shown a story about two girls, Sally and Anne, who are in a room with a basket and a box. Sally puts the marble into the basket, leaves the room, and then Anne moves the marble to the box in her absence. The child is then asked: "where does Sally believe the marble to be?". To pass the test, the child must answer "in the basket", since Sally did not see Anne moving the marble, and therefore Sally has the false belief that the marble is still in the basket. It is a first-order false belief task since the child taking the test has to be able to attribute a first-order false belief to Sally. Consider the following variant of the test: While Anne moves the marble, Sally secretly observes her through a window. Then Sally knows the marble has been moved, but Anne is not aware of that. Thus Sally does not get a false belief, but Anne does: Anne falsely believes that Sally believes the marble is in the basket. This is a second-order false belief task since the child needs to be able to attribute a second-order false belief to Anne [Flobbe, 2006].

In this paper, we consider a generalized version of the Sally-Anne task domain, consisting of two human participants A and B ("Sally" and "Anne") standing behind a table with three objects (coloured cubes) which can be put into three different containers (numbered boxes with lids), see Figure 1. There can be multiple objects in a container, and the content is not visible without lifting the lid. The human participants can interact with the environment by moving objects in and out of containers, and by entering and leaving the room. The robot keeps track of all interactions with the environment, including who observes what. This allows it to keep track of exactly who believes what, including higher-order beliefs.

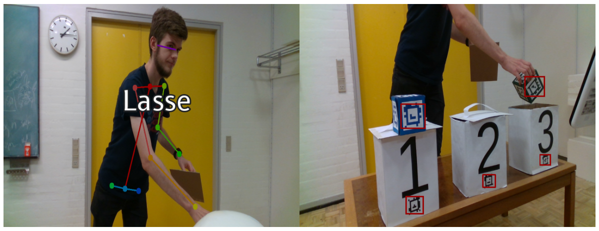

Figure 2: View from the two cameras, with object tracking overlay.

4 Implementation

We have used a Softbank Robotics Pepper robot. Two Intel RealSense D435i RGB+D cameras are used to provide high-resolution depth imagery. Both cameras are mounted on tripods next to the robot and oriented such that one has full coverage of the table while the other provides coverage for the person area behind the table, see Figure 2.

In order to successfully pass false-belief tasks, the robot must be able to reliably perceive and track humans, cubes, and boxes. Images from the cameras are passed to a battery of *detectors*, each of which is designed to detect a specific kind of feature such as faces, markers, and body poses. For each detected feature, a *percept* is created containing identification and spatial data. Person percepts are extracted using OpenPose realtime pose detection [Cao *et al.*, 2018] and dlib CNN face recognition [King, 2009]. The cubes and boxes are marked with unique AprilTag fiducial markers [Olson, 2011] in order to ensure robust detection.

Percepts produced by the perception system are used to maintain a low-level spatial *world model* representing the physical entities and their current position. Physical entities are split into a set of (names of) *objects* $\mathcal{O}$ and a set of (names of) *agents* $\mathcal{A}$. Each $o \in \mathcal{O}$ represents a uniquely identified object (a cube or box in our case), and each $i \in \mathcal{A}$ represents a unique person. In the world model, at any point in time, each $c \in \mathcal{O} \cup \mathcal{A}$ is assigned a position in a Cartesian coordinate system $\mathbb{R}^3$. The world model informs other components in the system using *events*. Two basic events produced by the world model are: $\texttt{Appear}(c)$, which is produced when the world model tracking locks onto a previously untracked entity c appearing in the robot's field of view; and $\texttt{Disappear}(c)$ which is produced when the world model is no longer able to track that entity.

More complex events are detected using *triggers*. An n-ary trigger is an *if-then* rule of the form 'for each $\omega \in (\mathcal{O} \cup \mathcal{A})^n$, if *condition*$(\omega)$ then produce *event*(ω)'. A trigger could e.g. be detecting whether a cube is put into a box. Each trigger is checked at a regular interval.

Triggers are used to monitor the spatial relations between the tracked cubes, boxes and persons. Each cube c has a small bounding sphere, and if the hand of agent i enters the bounding volume, the $\texttt{pickup}\ (i, c)$ event is produced. Similarly, each box b has a bounding box centered at its opening. If cube c was currently 'picked-up' by agent i when it disappeared inside the bounding box of b, the system produces a $\texttt{put}\ (i, c, b)$ event. It should be noted that these events only concern the robot's view of the physical state of the world, not its representation of the mental states of other agents, that we will now turn to.

5 Epistemic modelling on the robot

The DEL framework which our implementation is based on is essentially the version of DEL with *postconditions*, *edge-conditioned action models* and *observability propositions* introduced by Bolander [2018]. To keep the exposition simple and accessible to non-experts in DEL, we will not define general action models, but instead define the dynamics directly via the relevant model transformers. First we need to define our static epistemic language and its semantics.

Definition 1. *Let $\mathcal{O}$ and $\mathcal{A}$ be as above, and let Ψ be a set of predicates of first-order logic. The* epistemic language *$\mathcal{L}(\Psi, \mathcal{O}, \mathcal{A})$ is:*

$$\phi ::= P(\omega) \mid i \triangleleft j \mid \neg\phi \mid \phi \wedge \phi \mid B_i\phi$$

where $i, j \in \mathcal{A}$, $P \in \Psi$ is a predicate of arity $ar(P) \in \mathbb{N}$, and $\omega \in (\mathcal{O} \cup \mathcal{A})^{ar(P)}$. Formulas $P(\omega)$ and $i \triangleleft j$ are atoms*, and the set of these is denoted Atm.*

Note that the atoms of our logic is a combination of ground atoms of first-order logic and the special atoms $i \triangleleft j$. Atoms $i \triangleleft j$ are *observability atoms* read as "agent i sees agent j". The intended semantics of $i \triangleleft j$ is that agent i is currently looking at agent j and hence observing all actions performed by j. Formulas $B_i\phi$ are read: "agent i believes ϕ".

Example 1. *Consider the Sally-Anne task of Section 3. It can be modelled in $\mathcal{L}(\Psi, \mathcal{O}, \mathcal{A})$ with $\Psi = \{In\}$, $ar(In) = 2$, $\mathcal{O} = \{basket, box, marble\}$ and $\mathcal{A} = \{Sally, Anne\}$. The formula $B_{Anne}B_{Sally}In(marble, basket)$ then reads: "Anne believes that Sally believes the marble is in the basket."*

Definition 2. *An* epistemic model *of $\mathcal{L}(\Psi, \mathcal{O}, \mathcal{A})$ is $\mathcal{M} = (W, R, L)$ where W is a set of* (possible) worlds*; $R : \mathcal{A} \to \mathscr{P}(W \times W)$ maps each agent $i \in \mathcal{A}$ to an* accessibility relation *R_i; and $L : W \to \mathscr{P}(Atm)$ maps each world to its* labelling. *An* epistemic state *is a pair $(\mathcal{M}, w_0)$ where $w_0 \in W$ denotes the* actual world. *Formulas are evaluated in epistemic states as follows, with standard clauses for the propositional connectives '$\wedge$' and '$\neg$':*

- $(\mathcal{M}, w) \models p$ *iff* $p \in L(w)$ *for all* $p \in Atm$
- $(\mathcal{M}, w) \models B_i\phi$ *iff* $\forall v \in W, (w, v) \in R_i \Rightarrow (\mathcal{M}, v) \models \phi$

For all models considered, $i \triangleleft i$ is universally true for all agents i, but this will be left implicit.

Example 2. *The model s_4 of Figure 3 is an example of an epistemic state $(\mathcal{M}, w_0) = ((W, R, L), w_0)$ with worlds $W = \{w_l, w_r\}$ (represented by vertices) and with actual world $w_0 = w_l$ (represented by the black dot inside the vertex). Each world w is labelled by $L(w)$, so $In(cube_{red}, box_2)$ is true in w_l. The edges represent the accessibility relations, so e.g. $(w_l, w_r) \in R_A$. This edge has the consequence that despite $In(cube_{red}, box_2)$ being true in the actual world, agent A believes $In(cube_{red}, box_1)$. It corresponds to the situation at the end of the Sally-Anne task with Sally and Anne replaced by A and B, the marble replaced by a red cube, $cube_{red}$, and the basket and box replaced by box_1 and box_2, respectively.*

An *assignment* is an expression of the form $+p$ or $-p$ where $p \in Atm$. The assignment $+p$ is used to denote the atomic event of making p true. Similarly, $-p$ is the event

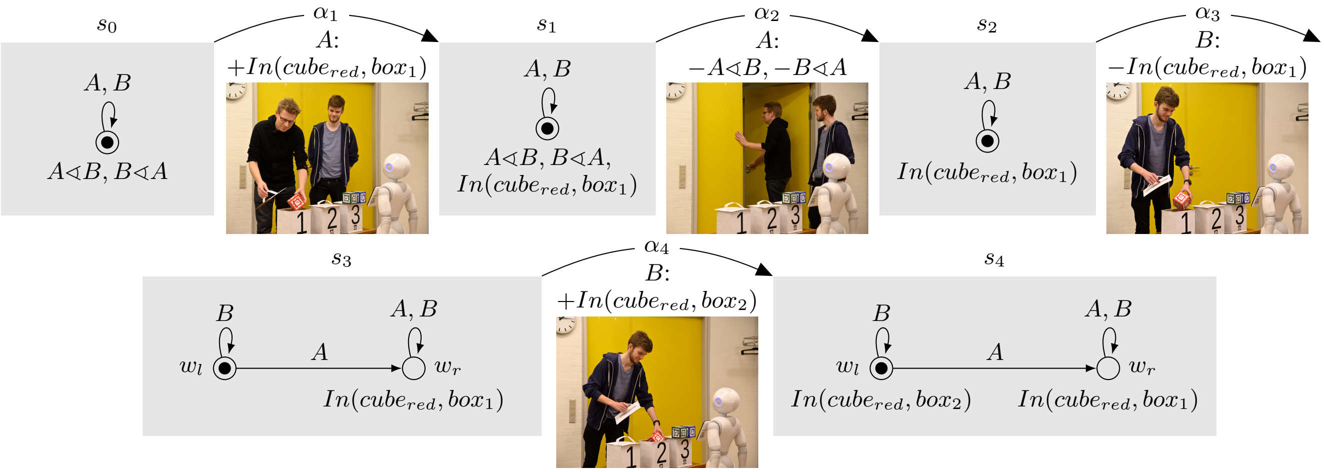

Figure 3: Evolution of states in a version of the Sally-Anne task. For all $n > 0$, we have $s_n = s_{n-1} \otimes \alpha_n$, where α_n is the action shown below the edge labelled α_n.

of making p false. An *action* is an expression $i{:}X$, where $i \in \mathcal{A}$ and X is a set of assignments. It denotes the action in which agent i brings about the events in X. For instance, $Anne{:}\{-In(marble, basket), +In(marble, box)\}$ denotes the action of Anne moving the marble from the basket to the box, and $Sally{:}\{-Sally \triangleleft Anne, -Anne \triangleleft Sally\}$ denotes the action of Sally leaving the room such that Sally and Anne no longer observe each other. We will often omit the set parentheses when writing actions.

Given a set of assignments X, we use $obs(X) = \{i \in \mathcal{A} \mid +i \triangleleft j \in X \text{ or } -i \triangleleft j \in X \text{ for some } j\}$ to denote the set of agents whose observational abilities are affected by X. Given an epistemic state s and action α, we use $s \otimes \alpha$ to denote the epistemic state resulting from executing α in s. The intuition underlying the semantics of $s \otimes \alpha$ defined below is this: In $s \otimes \alpha$, we make two copies of the set of worlds of s. The submodel induced by the first copy (the worlds of the form (w, c_1) below) represents the world view of those agents who don't observe α taking place, so this submodel is simply a copy of s. The submodel induced by the second copy (the worlds of the form (w, c_2) below) represents the world view of those agents who *do* observe α taking place. The agents observing an action $\alpha = i{:}X$ taking place are: 1) all agents currently observing agent i; 2) all agents whose observational abilities are affected by the assignments in X.[1]

Definition 3. *Let $s = ((W, R, L), w_0)$ be an epistemic state and $\alpha = i{:}X$ an action. Then the result of applying α in s is $s \otimes \alpha = ((W \times \{c_1, c_2\}, R', L'), (w_0, c_2))$ where:*

- *$((w, c_n), (v, c_m)) \in R'_j$ iff $(w, v) \in R_j$ and:*
  - *$n = m = 1$, or*
  - *$n = m = 2$ and $j \in \{k \mid s \models k \triangleleft i\} \cup obs(X)$, or*
  - *$n = 2, m = 1$ and $j \notin \{k \mid s \models k \triangleleft i\} \cup obs(X)$*
- *$L'((w, c_n)) =$*

$$\begin{cases} L(w) & n = 1 \\ (L(w) \cup \{p \mid +p \in X\}) - \{p \mid -p \in X\} & n = 2 \end{cases}$$

An epistemic state $s \otimes \alpha$ can end up containing worlds not reachable from the actual world. Such worlds do not affect what is true in the epistemic state, and will be omitted.

Example 3. *Consider again Figure 3. The initial state s_0 evolves into s_4 via the sequence of actions $\alpha_1, \dots, \alpha_4$, i.e. we have $s_4 = s_0 \otimes \alpha_1 \otimes \cdots \otimes \alpha_4$. The sequence $\alpha_1, \dots, \alpha_4$ represents the actions of a Sally-Anne task with the names introduced in Example 2. Initially, in s_0, the two agents see each other. Then agent A puts the red cube into box 1, which leads to s_1, i.e. $s_1 = s_0 \otimes \alpha_1 = s_0 \otimes A{:}{+In(cube_{red}, box_1)}$ (unreachable worlds omitted). Next, agent A leaves the room, which is the action α_2, leading to s_2. The final two actions represent agent B moving the cube from box 1 to box 2. Since the last two actions are performed in the absence of A, by the semantics of '$\otimes$' above this leads to the final situation s_4, where A has the false belief that the red cube is in box 1.*

5.1 From Perception Events to Actions

Our robot implements the semantics above, always keeping track of the beliefs of agents through an epistemic state s that is updated whenever an action α takes place. The robot infers which actions take place from the events described in Section 4. An event of the form `put` (i, c, b) is directly inferred to be the action $i{:}{+In(c, b)}$. All other actions can also be directly inferred from the corresponding events, except actions involving observability atoms. In our examples, we only need to keep track of which agents are co-present with the robot, and we can furthermore assume observability change to be public, significantly simplifying the handling of observability change. It is not strictly necessary to make these simplifying assumptions, but it simplifies our handling of higher-order false belief tasks in the following.

The tracking system maintains a set of currently tracked persons Φ through the `Appear` and `Disappear` events described in Section 4. When i enters the room, the `Appear` (i)

[1] See Bolander [2018] for further discussion. Bolander splits actions into *observability changing actions* and *ontic actions*, whereas we here generalise and simplify by considering those under one.

event is produced, and i is added to Φ; and when i leaves, Disappear(i) is produced and i is removed from Φ. Whenever Φ has been updated, the following action for updating the observability between agents is produced: $i{:}\{+i \triangleleft j \mid i,j \in \Phi\} \cup \{-j \triangleleft k \mid (j,k) \in (\Phi \times (\mathcal{A} - \Phi)) \cup ((\mathcal{A} - \Phi) \times \Phi)\}$. This action makes all agents in Φ become co-present, i.e., observe each other, and observability between agents in Φ and agents outside Φ is terminated.

5.2 Model Queries

By keeping track of the beliefs of agents through an epistemic state s, if the robot is asked a question such as "Does A believe that ϕ?", it can check whether $s \vDash B_A \phi$ holds, and answer correspondingly. To pass a Sally-Anne task, one also needs to be able to answer questions such as "Where does A believe the object c to be?". To handle such questions, we introduce *model queries*, borrowing notation from Calvanese *et al.* [2000].

Definition 4. *A* query *is a formula of $\mathcal{L}(\Psi, \mathcal{O}, \mathcal{A})$ where one or more constant symbols have been replaced by variables. We use standard notation $\phi(x_1, \ldots, x_n)$ for such formulas, where $\phi(c_1, \ldots, c_n)$ is the result of substituting c_i for x_i everywhere. The* answer *to a query $\phi(x_1, \ldots, x_n)$ in an epistemic state s is the formula $\phi(x_1, \ldots, x_n)^s := \{(c_1, \ldots, c_n) \in (\mathcal{O} \cup \mathcal{A})^n \mid s \vDash \phi(c_1, \ldots, c_n)\}$.*

On the robot, speech input is first transcribed using the DanSpeech neural automated speech recognition library by Nielsen and Jensen [2019]. The textual output is then parsed as a context-free language and transformed into an answer using a model query.

Example 4. *Consider the robot being in an epistemic state s and receiving the question "Where is c?". The question will be translated into the query $In(c, x)$, and its answer $In(c, x)^s$ will be computed. If $In(c, x)^s \neq \emptyset$, the robot will say "The x is in $In(c, x)^s$", otherwise the robot will say "The x is not in any box." Hence, letting s be the epistemic state s_4 of Figure 3, if we ask the robot "where is the red cube?", the robot will compute $In(cube_{red}, x)^s = box_2$ and answer "The red cube is in box 2". The robot can also answer questions about beliefs such as "Where does i believe that c is?", producing the query $B_i In(c, x)$. Asking "Where does A believe the red cube is?" in s_4, the robot will compute the query answer $(B_A In(cube_{red}, x))^s = box_1$ and respond with "A believes the red cube is in box 1". By this, it passes the Sally-Anne task. Higher-order beliefs are supported by simply iterating the belief operator, e.g. asking "Where does A believe that B believe c is?" produces the query $B_A B_B In(c, x)$.*

6 Results

Bolander [2018] claimed that the presented DEL-formalism would be appropriate and sufficient for building an artificial agent that can pass the Sally-Anne task—and more generally, false-belief tasks of arbitrary order. It is claimed that the approach is *robust* (not tied to a specific task or limited by a maximal depth of reasoning) and *faithful* (making it simple to translate the physical execution of the test into its formal representation). The robotic system described in this paper is indeed able to pass false-belief tasks of arbitrary order (robustness), and the translation from observations into actions is fairly straightforward (faithfulness). We now illustrate this through a series of scenarios including a first-order Sally-Anne task and a second-order false-belief task.

6.1 First-Order False-Belief Task

We already described the individual components involved in making the robot pass a Sally-Anne task, but let us now put all the pieces together. Our scenario starts with all cubes placed directly on the table and no human agents present. The robot represents this as an initial epistemic state s_{-1} consisting of a single world with no labelling. Now the robot waits for events to occur. The first events are when the two human agents, A (Sally) and B (Anne), enter the room.[2] These events are translated into the appropriate actions as described in Section 5.1, and the robot produces the updated epistemic state using the $\otimes$ operator, leading to s_0 of Figure 3. Now the two agents execute the action sequence illustrated by the pictures in Figure 3, where the robot percepts are first translated into events and then into actions, and the epistemic state is updated accordingly. At the end of executing the action sequence, the epistemic state of the robot has become s_4 of Figure 3. Now the robot can be asked questions about s_4, including who believes what, as described in Section 5.2.

The human agents are of course free to execute any action sequence involving moving the cubes and leaving/entering the room, and at any point of time during the action execution, the robot will be able to answer questions about the beliefs of agents. There is no limit on the number of boxes, cubes, agents, and actions the system can handle. For the scenarios described in this paper, all computations (including model updates and queries) are performed in real-time (50 Hz).

For our experiments, we used 50 humans unfamiliar with the robot to play the role of Sally and Anne. The experiments were carried out in different physical settings under different lighting conditions. If the robot failed to recognize an action, e.g. missed a cube being moved, we asked the human participant to repeat the action. In all 25 experiments, the robot correctly determined the false beliefs of participants, although, in the majority of them, one or more actions had to be repeated due to a perception failure. The most common perception failure was due to moving the cubes too fast, resulting in blurred images preventing fiducial marker detection.

6.2 Second-Order False-Belief Task

Our second scenario consists of a second-order false-belief task novel to this paper. The task involves three agents, A, B and C. The initial state s_0 contains a single world where all observability atoms as well as $In(cube_{red}, box_1)$ are true.[3] Then A leaves the room. While A is outside, B moves the cube to box 2. Now B leaves the room, and while B is away, C moves the cube back to box 1. Now A re-enters, also moves the cube to box 2, and leaves. This leads to a final situation in

[2] Agent names are learned at an earlier stage where the robot on seeing an unknown face asks for the name and learns it, using a combination of the earlier mentioned face and speech recognition.

[3] We don't have to assume this initial state, but could start from an empty epistemic state and then let the agents enter the room one by one, and afterwards agent C could put the red cube into box 1.

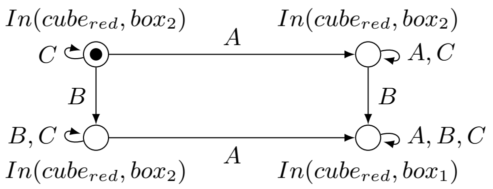

Figure 4: The end state of the second-order false-belief task

which both A and B have correct first-order beliefs about the cube: it is in box 2. However, there is now a second-order false belief: A falsely believes B to believe the cube is in box 1, since A didn't see B moving it, and A knows that B didn't see A moving it. Symmetrically, B also comes to falsely believe A to believe the cube is in box 1. This action sequence is formalised as follows by the robot:

1. A leaves the room:

   $\alpha_1 = A{:}{-}A \lhd B, -B \lhd A, -A \lhd C, -C \lhd A, +B \lhd C, +C \lhd B$

2. B moves the cube to box 2 and leaves the room:

   $\alpha_2 = B{:}{-}In(cube_{red}, box_1)$
   $\alpha_3 = B{:}{+}In(cube_{red}, box_2)$
   $\alpha_4 = B{:}{-}A \lhd B, -B \lhd A, -A \lhd C, -C \lhd A, -B \lhd C, -C \lhd B$

3. C moves the cube back into box 1:

   $\alpha_5 = C{:}{-}In(cube_{red}, box_2)$
   $\alpha_6 = C{:}{+}In(cube_{red}, box_1)$

4. A re-enters, moves the cube to box 2, and leaves again:

   $\alpha_7 = A{:}{-}A \lhd B, -B \lhd A, +A \lhd C, +C \lhd A, -B \lhd C, -C \lhd B$
   $\alpha_8 = A{:}{-}In(cube_{red}, box_1)$
   $\alpha_9 = A{:}{+}In(cube_{red}, box_2)$
   $\alpha_{10} = A{:}{-}A \lhd B, -B \lhd A, -A \lhd C, -C \lhd A, -B \lhd C, -C \lhd B$

Applying this sequence of actions to the initial state s_0 generates the state $s = s_0 \otimes \alpha_1 \otimes \cdots \otimes \alpha_{10}$ shown in Figure 4. Note that the model correctly represents the second-order false beliefs of A and B. Performing the actions and asking the robot "Where does A believe the red cube is?" will make it answer "in $(B_A In(cube_{red}, x))^s$", i.e., "in box 2". If asked "Where does A believe that B believes the red cube is?", it will answer "in $(B_A B_B In(cube_{red}, x))^s$", i.e. "in box 1", hence passing the second-order task.

Note that we didn't program the robot specifically to pass this second-order task. The general framework we implemented on the robot is sufficient to pass any false-belief task involving the type of actions our system supports (moving cubes in and out of boxes and moving agents in and out of the room). We first designed this second-order false belief task on paper, and later simply enacted it in front of the robot, and then it managed to pass it.

6.3 Application examples

Passing a false-belief task is not necessarily in itself important for applications of human-robot interaction. However, the belief tracking that our framework supports has a number of potentially interesting applications, a couple of which we will now discuss. Since the robot can track true and false beliefs in human agents, it can also assess what is relevant to announce to those humans. Currently, we have only implemented a very simple version of this. A human agent, i, can say "I have come to pick up object c", after which the robot will run the query $\phi(x) := In(c, x) \wedge \neg B_i In(c, x)$. This query checks whether agent i has a false belief concerning the location of object c. If $\phi(x)^s = \emptyset$, where s is the current epistemic state of the robot, agent i does not have a false belief concerning the location of c. In this case the robot will not say anything. It would be annoying if the robot said "c is in box $In(c, x)^s$", a fact agent i already knows in this case. This relates back to our service robot example from the introduction: We will only rarely want a robot to tell us things we already know. However, if $\phi(x)^s \neq \emptyset$, the robot knows that i has a false belief concerning the location of c, and will say "c is in $In(c, x)^s$."

The point is that the robot should only proactively make announcements to the human if she has false beliefs concerning objects relevant to her intentions, and otherwise keep quiet. This kind of behaviour would of course be impossible to achieve if the robot didn't have a ToM in which to represent the mental states of other agents.

In the example of the introduction, we also mentioned a case where an employee is searching for something the robot knows to have been moved, and the robot ought to inform her. This requires an additional intention recognition layer. To be able to test just a simple version of this, we made the robot interpret the action of an agent i reaching out for a box b as an intent to interact with box b, i.e., to pick up one of the cubes believed to be in b. This is implemented using an additional trigger for when the hand of agent i is near box b. The trigger will lead to the following query being tested: $\psi(x) := \neg In(x, b) \wedge B_i In(x, b)$. This query tests whether something is falsely believed by i to be in b. As before, if $\psi(x)^s = \emptyset$, the robot stays quiet. Otherwise, the robot informs i about her false beliefs by, for each $c \in \psi(x)^s$, announcing: "If you are looking for cube c, it is now in box $In(c, x)^s$". Without the ability to represent the beliefs of agent i, the robot would not have had any idea about what to announce to i, even if having correctly recognized her intention.

7 Future work

To the best of our knowledge, this is the first robot passing a higher-order false-belief task. However, many false-belief tasks rely on announcing and recovering from false beliefs, something that cannot be modelled with standard epistemic DEL models [Bolander, 2018], but require (at least) plausibility models [Baltag and Smets, 2008]. Moving to plausibility models would also simplify generalizing our work to a setting where the robot is not assumed omniscient (where it can have false beliefs on its own, and also recover from these). Furthermore, extending the models to incorporate other aspects of ToM such as intentions and desires remains an important next step. Finally, our examples concerning intention recognition and "helpful announcements" were quite restrictive. A more general approach for the robot to know what to announce to whom would be to use epistemic planning with implicit coordination [Engesser *et al.*, 2017], fitting naturally into the existing DEL-based approach.

Acknowledgement. We are indebted to Mathias Kaas-Olsen for implementing the DEL-based reasoning engine.

References

[Arkoudas and Bringsjord, 2008] Konstantine Arkoudas and Selmer Bringsjord. Toward formalizing common-sense psychology: An analysis of the false-belief task. In *Pacific Rim International Conference on Artificial Intelligence*, pages 17–29. Springer, 2008.

[Baltag and Smets, 2008] Alexandru Baltag and Sonja Smets. A qualitative theory of dynamic interactive belief revision. *Logic and the foundations of game and decision theory (LOFT 7)*, 3:9–58, 2008.

[Barras, 2009] Colin Barras. Useful, lovable and unbelievably annoying. *New Scientist*, 204(2738):22–23, 2009.

[Bolander, 2018] Thomas Bolander. Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic. In *Jaakko Hintikka on Knowledge and Game-Theoretical Semantics*, pages 207–236. Springer, 2018.

[Breazeal *et al.*, 2009] Cynthia Breazeal, Jesse Gray, and Matt Berlin. An embodied cognition approach to mindreading skills for socially intelligent robots. *The International Journal of Robotics Research*, 28(5):656–680, 2009.

[Calvanese *et al.*, 2000] Diego Calvanese, Giuseppe De Giacomo, and Maurizio Lenzerini. Answering queries using views over description logics knowledge bases. *AAAI/IAAI*, 2000:386–391, 2000.

[Cao *et al.*, 2018] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. *arXiv preprint arXiv:1812.08008*, 2018.

[Dautenhahn, 2007] Kerstin Dautenhahn. Socially intelligent robots: Dimensions of human–robot interaction. *Philosophical transactions of the royal society B: Biological sciences*, 362(1480):679–704, 2007.

[Engesser *et al.*, 2017] Thorsten Engesser, Thomas Bolander, Robert Mattmüller, and Bernhard Nebel. Cooperative epistemic multi-agent planning for implicit coordination. In *Proceedings of Methods for Modalities*, Electronic Proceedings in Theoretical Computer Science, 2017.

[Flobbe, 2006] Liesbeth Flobbe. Children's development of reasoning about other people's minds. *Unpublished Master's Thesis, University of Groningen*, 2006.

[King, 2009] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[Lemaignan and Dillenbourg, 2015] Séverin Lemaignan and Pierre Dillenbourg. Mutual modelling in robotics: Inspirations for the next steps. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 303–310. IEEE, 2015.

[Milliez *et al.*, 2014] Grégoire Milliez, Matthieu Warnier, Aurélie Clodic, and Rachid Alami. A framework for endowing an interactive robot with reasoning capabilities about perspective-taking and belief management. In *The 23rd IEEE international symposium on robot and human interactive communication*, pages 1103–1109. IEEE, 2014.

[Mitsunaga *et al.*, 2008] Noriaki Mitsunaga, Zenta Miyashita, Kazuhiko Shinozawa, Takahiro Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. What makes people accept a robot in a social environment - discussion from six-week study in an office. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3336–3343. IEEE, 2008.

[Nematzadeh *et al.*, 2018] Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Thomas L Griffiths. Evaluating theory of mind in question answering. *arXiv preprint arXiv:1808.09352*, 2018.

[Nielsen and Jensen, 2019] Martin Carsten Nielsen and Rasmus Arpe Fogh Jensen. Danspeech. https://github.com/danspeech/danspeech, 2019.

[Olson, 2011] Edwin Olson. AprilTag: A robust and flexible visual fiducial system. In *2011 IEEE International Conference on Robotics and Automation*, pages 3400–3407. IEEE, 2011.

[Premack and Woodruff, 1978] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[Rabinowitz *et al.*, 2018] Neil C Rabinowitz, Frank Perbet, H Francis Song, Chiyuan Zhang, SM Eslami, and Matthew Botvinick. Machine Theory of Mind. *arXiv preprint arXiv:1802.07740*, 2018.

[Sindlar *et al.*, 2009] Michal P Sindlar, Mehdi M Dastani, and John-Jules Ch Meyer. BDI-based development of virtual characters with a theory of mind. In *International Workshop on Intelligent Virtual Agents*, pages 34–41. Springer, 2009.

[Wimmer and Perner, 1983] 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.

[Złotowski *et al.*, 2015] Jakub Złotowski, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. Anthropomorphism: Opportunities and challenges in human–robot interaction. *International Journal of Social Robotics*, 7(3):347–360, 2015.