Non-monotonic Logical Reasoning and Theory of Mind for Transparency in HRI

Michalina Jakubczak, Mohan Sridharan, and Masoumeh Mansouri

Intelligent Robotics Lab
School of Computer Science, University of Birmingham, UK
mjj783@student.bham.ac.uk,{m.sridharan,m.mansouri}@bham.ac.uk

Abstract. Our architecture seeks to enable transparency in a robot's decision making, which is crucial in Human-robot Interaction. For any given goal, the robot performs non-monotonic logical reasoning with incomplete commonsense domain knowledge to compute a plan to achieve the goal. In addition, it reasons with domain knowledge and Theory of Mind (ToM) models of specific human users to provide user-specific contextual descriptions of the relevant decisions and beliefs in response to the user's questions. These capabilities are illustrated with a motivating scenario of a simulated robot assisting a human in an indoor domain.

Keywords: Theory of Mind · Transparent decision making · Human-robot interaction.

1 Introduction

Consider a mobile robot assisting Mary, a recently retiree. While largely independent, she benefits from the robot's help with some tasks in her house, e.g., fetching or putting away objects. Mary's trust in the robot will depend on the robot's ability to consider Mary's preferences and prior knowledge to identify and communicate the information relevant to the task or query posed by Mary.

A robot's ability to "*explain*" its behavior by describing its decisions and beliefs to the human user is a key requirement for effective human-robot collaboration [1,13]. This transparency in decision making is related to the rich body of work on *explainable AI* [2,15]. Our work is directed towards transparency in integrated robot systems that sense, reason, interact with, and learn from complex dynamic domains. In this paper, we describe an architecture that enables a robot to reason with prior domain knowledge (e.g., some domain/robot attributes, action effects) for planning and providing on-demand relational descriptions of its decisions and beliefs. Specifically, our architecture:

- Performs non-monotonic logical reasoning with commonsense domain knowledge (i.e., robot and domain attributes, axioms governing change) at different resolutions to compute and execute plans to achieve the desired goal.
- Reasons with domain knowledge, a Theory of Mind model of the human interacting with it, and formal definitions of relevance, to automatically construct a response to the human's query about its decisions and beliefs.

We use Answer Set Prolog (ASP) to represent and reason with prior domain knowledge and ToM models in order to construct the desired relational descriptions. We abstract away the architecture's components for sensing, actuation, and learning; other work in our group has explored these components [22]. We use execution traces to illustrate the architecture's capabilities in the context of a simulated robot assisting humans in an indoor domain.

2 Related work

Work in *explainable AI* can be broadly classified into two groups [2,15,16]. Methods in the first group modify or map learned models or reasoning systems to make their decisions interpretable, e.g., frameworks that approximate learned "black box" models by equivalent interpretable models [18,19], or bias a planning system towards making decisions easier for humans to understand [23]. The second group consists of methods that seek to make a reasoning or learning system's decisions more transparent, e.g., by mapping a model's decisions to input features [10], describing planning decisions [5], or justifying decisions based on non-monotonic logical reasoning [6].

Theory of Mind (ToM), defined as "the cognitive capacity to attribute mental states to self and others" [4], has been used extensively to model beliefs and generate explanations that promote human understanding. In methods that use ToM models, the accuracy of the model determines the accuracy (or even the need) for explanation [11]. Examples of such work include the use of *model reconciliation* to address knowledge discrepancies between agents in the context of plan explanation [21], and making human-generated plans understandable [9]. Recent work has used epistemic logic planners for decision making, and extends the ToM model to include nested beliefs and epistemic goals, supporting the generation of user-specific descriptions of current state [20].

Our work is directed toward transparent, reliable, and efficient reasoning and learning in integrated robot systems in dynamic domains. It is inspired by work on explainable agency [12] in the context of robot architectures that reason and learn with relevant knowledge at different resolutions using some commonsense domain knowledge and observations [17,22].

3 Proposed architecture

Our proposed architecture (see Figure 1) performs non-monotonic logical reasoning with commonsense domain knowledge for planning, diagnostics, and inference. The architecture builds a ToM model to dynamically represent and reason about the user's beliefs. Guided by the user interaction, the robot extracts relevant information from the ToM model and its history of observations to construct suitable explanations of its decisions and beliefs.

Example Domain 1 *[Assistive Robot (AR) Domain]*
Consider the robot assisting Mary in her house. The robot has some prior commonsense domain knowledge, i.e., some attributes of robot and domain (e.g.,

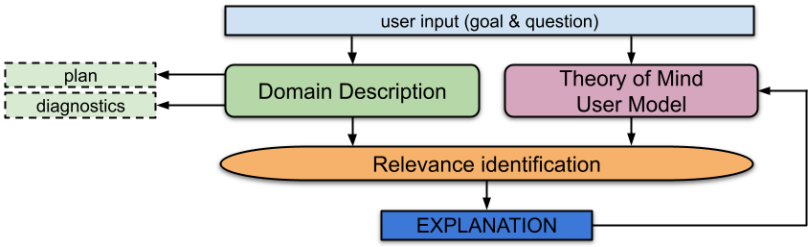

Fig. 1. Architecture overview: non-monotonic logical reasoning with commonsense knowledge supports planning and diagnostics; reasoning with ToM model provides on-demand descriptions of relevant decisions and beliefs adapted to user queries.

map of house, size and location of objects), relations between objects, and axioms governing change. It is able to plan and execute actions, and has some prior understanding of Mary's domain knowledge, preferences, and beliefs.

Mary asks the robot to bring her the cookbook that is usually in a designated box in the kitchen. However, instead of the cookbook, the robot finds a crossword puzzle inside the box; it eventually finds the cookbook in the living room. Mary asks the robot why it went to the living room. To answer this question, the robot must consider Mary's current beliefs and communicate the relevant information, e.g., that the cookbook was not found in its expected location.

3.1 Knowledge Representation and Reasoning

We first describe the basic knowledge representation and reasoning component.

Action Language In our architecture, the domain's transition diagrams are described in action language $\mathcal{AL}$ [8], a formal model of parts of natural language used for describing the behavior of dynamic systems. $\mathcal{AL}$ has a sorted signature with *statics*, *fluents* and *actions*. It allows three types of statements; causal law, state constraint, and executability condition.

Domain Representation Domain representation consists of system description $\mathcal{D}$, a collection of statements of $\mathcal{AL}$, and history $\mathcal{H}$. $\mathcal{D}$ has a signature Σ and axioms. Σ has basic sorts, e.g., *loc*, *agent*, *object*, *container*; domain attributes, e.g., static $next\_to(loc, loc)$ and fluent $in\_hand(agent, object)$; robot actions, e.g., $pick\_up(agent, object)$); and exogenous actions, e.g., $exo\_remove(object, container)$. Also, relation $holds(fluent, step)$ implies that a particular fluent is true at a particular timestep. $\mathcal{H}$ records the robot's observations, i.e., $obs(fluent, Bool, step)$, and action executions, i.e., $hpd(action, step)$, at specific time steps. Action execution in the physical world may require a finer-granularity representation, e.g., to pick up objects from specific locations. To support this ability, our architecture has transition diagrams at two resolutions, with the finer-granularity representation (grids in rooms, parts of objects) defined formally as a *refinement* of the coarse-granularity representation [22]; the robot automatically *zooms* to the part of the fine-granularity representation relevant to tasks at hand.

Theory of Mind For each human user i that it interacts with, the robot maintains a Theory of Mind model Ψ_i that has a format similar to the robot's domain description, i.e., a system description $\mathcal{D}_{\psi_i}$ and history $\mathcal{H}_{\psi_i}$. Axioms in $\mathcal{D}_{\Psi_i}$ capture the robot's understanding of how user i's beliefs change over time, including axioms governing the domain dynamics, default knowledge (e.g. "books are usually in the study"), initial beliefs, and axioms for identifying and resolving cognitive dissonance. Other work has explored how domain knowledge can be acquired incrementally [17]; here, we focus on reasoning with given ToM models.

Reasoning with Knowledge To perform the reasoning tasks, i.e., planning, diagnostics, and inference, the robot automatically constructs a program $\Pi(\mathcal{D}, \mathcal{H})$ in CR-Prolog [3], which is an extension of ASP[1]. Π includes generic helper axioms (e.g., inertia axioms, goal specification). Reasoning is reduced to computing *answer sets* of Π that represent the set of inferred beliefs of an agent associated with the program. When the robot has to reason with the ToM model Ψ_i (e.g., for generating explanations below), it constructs and solves $\Pi(\mathcal{D}_{\Psi_i}, \mathcal{H}_{\Psi_i})$; the resulting answer set is the robot's model of the user's beliefs.

3.2 Constructing relevant explanations

When a human user poses a question $\mathcal{Q}$, the robot provides an explanation $\mathcal{E}(\mathcal{Q}, \sigma_g)$ in the form of a relational description of relevant decisions and beliefs. This response is tailored to $\mathcal{Q}$, current goal σ_g, and the user's beliefs, by reasoning with $\mathcal{D}$, $\mathcal{H}$, and Ψ_i (i.e., $\mathcal{D}_{\psi_i}$ and $\mathcal{H}_{\psi_i}$). The robot executes the following steps:

1. Identify relevant object constants in Σ, i.e., $relCon(\mathcal{Q}, \sigma_g)$, using σ_g and $\mathcal{Q}$.
2. Reason with Ψ_i to compute updated beliefs of user i. Use any user beliefs that violate the robot's beliefs to identify relevant object constants.
3. Identify relevant signature $\Sigma(\mathcal{Q},\sigma_g)$ using $relCon(\mathcal{Q}, \sigma_g)$.
4. Restrict $\mathcal{D}$ and $\mathcal{H}$ to signature $\Sigma(\mathcal{Q},\sigma_g)$ to obtain $\mathcal{D}(\mathcal{Q}, \sigma_g)$ and $\mathcal{H}(\mathcal{Q}, \sigma_g)$.
5. Construct relevant explanation $\mathcal{E}(\mathcal{Q}, \sigma_g)$.

We briefly describe definitions used by the robot to construct suitable explanations for a given user question $\mathcal{Q}$ and goal σ_g.

Definition 1 *[Relevant object constants]*
Let user's question $\mathcal{Q}$ be a set of ground literals $\{q\}$. Let σ_g be the robot's goal, and $relCon(\mathcal{Q}, \sigma_g)$ be the set of object constants identified as follows:

1. If$f(x_1, ..., x_n, y)$ is a literal formed of a domain attribute and occurs in $\mathcal{Q}$, all object constants of the sorts of arguments $x_1, ..., x_n, y$ are in $relCon(\mathcal{Q}, \sigma_g)$.
2. If$f(x_1, ..., x_n, y)$ is a literal formed of a domain property, and belongs to σ_g then object constants $x_1, ..., x_n, y$ are in $relCon(\mathcal{Q}, \sigma_g)$.
3. If $\mathcal{B}_{i,v}$ is a non-empty set of violated beliefs obtained by restricting the user i's beliefs ($\mathcal{B}_i$) to object constants identified above, and identifying fluent literals whose range is not the robot's current beliefs, only object

[1] "ASP" and "CR-Prolog" are used interchangeably in this paper.

constants that are arguments of beliefs in $\mathcal{B}_{i,v}$ and the arguments of the corresponding violating (robot's) beliefs remain in $\mathcal{B}_{i,v}$ and the arguments of the corresponding violating (robot's) beliefs.

First two conditions help identify relevant constants for different types of questions; the third one constrains this set whenever the user's beliefs (based on Ψ for user) are violated by robot's own beliefs. Then $relCon(\mathcal{Q}, \sigma_g)$ is the set of object constants relevant to the explanation $\mathcal{E}(\mathcal{Q}, \sigma_g)$ for query $\mathcal{Q}$ and goal σ_g.

Definition 2 *[Relevant signature]*
Let Σ be the signature of $\mathcal{D}$. Let $\Sigma(\mathcal{Q}, \sigma_g)$ be the signature computed as follows:

1. Sorts of Σ with a non-empty intersection with $relCon(\mathcal{Q}, \sigma_g)$ are in $\Sigma(\mathcal{Q},\sigma_g)$.
2. For each basic sort of Σ corresponding to the range of a static attribute, all constants of the sort are in $\Sigma(\mathcal{Q},\sigma_g)$.
3. For each basic sort of Σ corresponding to the range of a fluent, or domain of a fluent or static, constants of the sort in $relCon(\mathcal{Q}, \sigma_g)$ are in $\Sigma(\mathcal{Q},\sigma_g)$.
4. Domain properties restricted to basic sorts of $\Sigma(\mathcal{Q},\sigma_g)$ are in $\Sigma(\mathcal{Q},\sigma_g)$.

Then $\Sigma(\mathcal{Q}, \sigma_g)$ is the signature relevant to the explanation $\mathcal{E}(\mathcal{Q},\sigma_g)$.

Definition 3 *[Relevant system description]*
Let $\mathcal{D}$ be the robot's system description such that $\mathcal{D}_\Psi \in \mathcal{D}$. Then, $\mathcal{D}(\mathcal{Q},\sigma_g)$, the system description relevant to explanation $\mathcal{E}(\mathcal{Q},\sigma_g)$ comprises signature $\Sigma(\mathcal{Q}, \sigma_g)$ and axioms of $\mathcal{D}$ restricted to $\Sigma(\mathcal{Q},\sigma_g)$. We make the reasonable assumption that the robot's knowledge includes the ToM models of specific users.

Definition 4 *[Relevant history]*
If history $\mathcal{H}$ is the set of the robot's observations of domain attributes and the occurrences of actions. $\mathcal{H}(\mathcal{Q}, \sigma_g)$ is the restriction of $\mathcal{H}$ to $relCon(\mathcal{Q},\sigma_g)$.

4 Execution traces and Discussion

We present two execution traces to demonstrate our architecture's capabilities.

Execution Example 1 *[Question of the form "Why X?"]*
Recall the situation in **Example Domain** 1, with Mary asking the robot ($robot_1$) to find her *cookbook*.

1. Robot's goal can be stated as: $\sigma_g = in\_hand(robot_1, cookbook)$, while the subsequent question is: $\mathcal{Q} = in\_room(robot_1, living\_room)$. The $relCon(\mathcal{Q}, \sigma_g)$ (without belief violations) includes *cookbook*, $robot_1$ and all rooms of sort *loc*, i.e., *living_room*, *kitchen*, and *bedroom*.
2. Reasoning with relevant parts of its history $\mathcal{H}(\mathcal{Q}, \sigma_g)$ and that of the ToM model $\mathcal{H}_\Psi(\mathcal{Q}, \sigma_g)$, the robot identifies $\mathcal{B}_v : in\_room(cookbook, kitchen)$ and the corresponding $in\_room(cookbook, living\_room)$ as belief violations. Not all incorrect beliefs are included (e.g., $inside(cookbook, box)$), as they are not fully defined by the object constants in $relCon(\mathcal{Q}, \sigma_g)$. The revised set of relevant object constants is: *cookbook*, *living_room* and *kitchen*.

3. Next, the robot reasons with its relevant history and the beliefs in the user's ToM to find the literals fully defined by the revised $relCon(\mathcal{Q}, \sigma_g)$. The following literals are identified to form explanation $\mathcal{E}(\mathcal{Q}, \sigma_g)$ to Mary's question:

$$exo\_move\_room(cookbook, living\_room), 0).$$
$$obs(in\_room(cookbook, living\_room), false, 2),$$
$$obs(in\_room(cookbook, kitchen), true, 3).$$

   The robot explains that it found the cookbook in the living room, potentially because it had been exogenously moved there from the kitchen.

Execution Example 2 *[Question of the form "What X?"]*
Having received the explanation from **Execution Example 1.**, Mary asks the robot "What was in the box instead of the cookbook?".

1. This new question can be expressed as $inside(\#object, box)$, while the goal remains the same. The relevant object constants $relCon(\mathcal{Q}, \sigma_g)$ are thus *box* and all constants of sort *object*, e.g., *cookbook*, $crossword\_puzzle$, $crime\_novel$.
2. Based on the relevant histories $\mathcal{H}(\mathcal{Q}, \sigma_g)$ and $\mathcal{H}_\Psi(\mathcal{Q}, \sigma_g)$, the robot identified the violated beliefs $\mathcal{B}_v$: $inside(cookbook, box)$ and $inside(crossword\_puzzle, box)$. The $relCon(\mathcal{Q}, \sigma_g)$ is limited to *box*, *cookbook* and $crossword\_puzzle$.
3. The explanation $\mathcal{E}(\mathcal{Q}, \sigma_g)$ consists of:

$$hpd(exo\_remove(cookbook, box), 0),$$
$$hpd(exo\_place(crossword\_puzzle, box), 1),$$
$$obs(inside(cookbook, box), false, 2),$$
$$obs(inside(crossword\_puzzle, box), true, 2).$$

   The robot explains that it found the $crossword\_puzzle$ in the *box* instead of the *cookbook*, potentially due to an unobserved exogenous action.

Discussion: Although these examples show an instance of two types of questions, our architecture can handle different types of questions, i.e., those based on *why, what, when, how, where* etc. These correspond to descriptive, contrastive, and counterfactual questions, which are known to be important in the context of interactions between robots and/or humans [7,14]. In addition, if the user requires information at a finer-granularity, the robot can automatically provide that using the refined domain descriptions. Furthermore, reasoning with commonsense knowledge enables the robot to guide learning and recover from any incorrect inferences drawn. Future work will describe examples corresponding to these capabilities, demonstrate the use of this architecture in conjunction with the modules for sensing, actuation, and learning (e.g., of ToM models), and explore the impact of the generated relational descriptions on effective human-robot interaction in complex domains.

References

1. Ethics guidelines for trustworthy ai (Apr 2019), https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
2. Anjomshoae, S., Najjar, A., Calvaresi, D., Framling, K.: Explainable agents and robots: Results from a systematic literature review. In: International Conference on Autonomous Agents and Multiagent Systems. Montreal, Canada (2019)
3. Balduccini, M., Gelfond, M.: Logic programs with consistency-restoring rules. In: International Symposium on Logical Formalization of Commonsense Reasoning, AAAI Spring Symposium Series (2003)
4. Blum, C., Winfield, A.F., Hafner, V.V.: Simulation-based internal models for safer robots. Frontiers in Robotics and AI **4**, 74 (2018)
5. Borgo, R., Cashmore, M., Magazzeni, D.: Towards Providing Explanations for AI Planner Decisions. In: IJCAI Workshop on Explainable Artificial Intelligence. pp. 11–17 (2018)
6. Fandinno, J., Schulz, C.: Answering the "Why" in Answer Set Programming: A Survey of Explanation Approaches. Theory and Practice of Logic Programming **19**(2), 114–203 (2019)
7. Fox, M., Long, D., Magazzeni, D.: Explainable Planning. In: IJCAI Workshop on Explainable AI (2017)
8. Gelfond, M., Kahl, Y.: Knowledge representation, reasoning, and the design of intelligent agents: The answer-set programming approach. Cambridge University Press (2014)
9. Grover, S., Sengupta, S., Chakraborti, T., Mishra, A.P., Kambhampati, S.: Radar: automated task planning for proactive decision support. Human–Computer Interaction **35**(5-6), 387–412 (2020)
10. Hohman, F., Kahng, M., Pienta, R., Chau, D.H.: Visual analytics in deep learning: An interrogative survey for the next frontiers. IEEE transactions on visualization and computer graphics **25**(8), 2674–2693 (2018)
11. Kiesler, S.: Fostering common ground in human-robot interaction. In: ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005. pp. 729–734. IEEE (2005)
12. Langley, P., Meadows, B., Sridharan, M., Choi, D.: Explainable Agency for Intelligent Autonomous Systems. In: Innovative Applications of Artificial Intelligence (2017)
13. Melkas, H., Hennala, L., Pekkarinen, S., Kyrki, V.: Impacts of robot implementation on care personnel and clients in elderly-care institutions. International Journal of Medical Informatics **134**, 104041 (2020)
14. Menzies, P., Beebee, H.: Counterfactual Theories of Causation. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, winter 2020 edn. (2020)
15. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)
16. Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N.: Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review **55**(5), 3503–3568 (2022)
17. Mota, T., Sridharan, M., Leonardis, A.: Integrated Commonsense Reasoning and Deep Learning for Transparent Decision Making in Robotics. Springer Nature Computer Science **2**(242), 1–18 (2021)
18. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)

19. Rodríguez-Pérez, R., Bajorath, J.: Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. Journal of computer-aided molecular design **34**(10), 1013–1026 (2020)
20. Shvo, M., Klassen, T.Q., McIlraith, S.A.: Explaining the plans of agents via theory of mind. In: Explainable AI Planning Workshop (2021)
21. Sreedharan, S., Chakraborti, T., Kambhampati, S.: Foundations of explanations as model reconciliation. Artificial Intelligence **301**, 103558 (2021)
22. Sridharan, M., Gelfond, M., Zhang, S., Wyatt, J.: Reba: A refinement-based architecture for knowledge representation and reasoning in robotics. Journal of Artificial Intelligence Research **65**, 87–180 (2019)
23. Zhang, Y., Sreedharan, S., Kulkarni, A., Chakraborti, T., Zhuo, H.H., Kambhampati, S.: Plan explicability and predictability for robot task planning. In: International Conference on Robotics and Automation (2017)