Combining Theory of Mind and Abductive Reasoning in Agent-Oriented Programming

(JAAMAS track)

Nieves Montes
Artificial Intelligence Research Institute (IIIA-CSIC)
Bellaterra, Barcelona, Spain
nmontes@iiia.csic.es

Michael Luck
University of Sussex
Brighton, United Kingdom
Michael.Luck@sussex.ac.uk

Nardine Osman
Artificial Intelligence Research Institute (IIIA-CSIC)
Bellaterra, Barcelona, Spain
nardine@iiia.csic.es

Odinaldo Rodrigues
King's College London
London, United Kingdom
odinaldo.rodrigues@kcl.ac.uk

Carles Sierra
Artificial Intelligence Research Institute (IIIA-CSIC)
Bellaterra, Barcelona, Spain
sierra@iiia.csic.es

ABSTRACT

In this paper we present TomAbd, a novel agent model extending the BDI architecture with Theory of Mind capabilities, i.e. the capacity to adopt and reason from the perspective of others. By combining the Theory of Mind of TomAbd agents with abductive reasoning, agents can infer explanations for the behaviour of others, which they can incorporate into their own decision-making. We have implemented the TomAbd agent model and successfully tested its performance in the cooperative board game Hanabi.

KEYWORDS

Theory of mind, Abductive reasoning, Agent-oriented programming, Social AI, Hanabi

ACM Reference Format:
Nieves Montes, Michael Luck, Nardine Osman, Odinaldo Rodrigues, and Carles Sierra. 2024. Combining Theory of Mind and Abductive Reasoning in Agent-Oriented Programming: (JAAMAS track). In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Theory of Mind (ToM) is the human cognitive ability to perceive, interpret and reason about others in terms of their mental states, such as their beliefs, goals and intentions [2]. It is an essential requirement for effective participation in social life, and is strongly linked to the feeling of empathy [4] and moral judgement [3]. The emergent field of social AI acknowledges the need of ToM-like capabilities for software agents to successfully interact with other agents as well as humans [1, 6].

In this work, we present an overview of our novel TomAbd agent model [5], a Theory of Mind extension of the Belief-Desire-Intention (BDI) agent architecture. A TomAbd agent i is designed to operate in the following generic scenario. A different (not necessarily TomAbd or even BDI) agent l, denoted as the *actor*, executes some action a_l observable by i. Upon observing the action, i uses ToM and substitutes its belief base with the beliefs it estimates l has. Once this change of perspective is complete, i engages in *abductive reasoning* to derive the possible beliefs that might have led l to decide on a_l. After some post-processing, i incorporates the information thus derived into its own belief base, to inform its posterior decision-making.

Besides ToM, the second main component of the TomAbd agent model is abductive reasoning. Abduction is a logical inference paradigm differing from traditional deductive reasoning [7]. Classical deduction follows the *modus ponens* rule: from knowledge of ϕ and of the implication $\phi \rightarrow \psi$, ψ is inferred. In contrast, abduction makes inferences in the opposite direction: from knowledge of the implication $\phi \rightarrow \psi$ and the *observation* of ψ, ϕ is inferred as a possible *explanation* for ψ. The explanation ϕ may be further constrained by the need to be consistent with prior knowledge. In the TomAbd agent model, observations refer to actions executed by other agents (`action(`l, a_l`)`), while explanations refer to the beliefs that might have led l towards a_l.

In the remainder of this paper, we present the key features of the TomAbd agent model (Section 2) as well as its performance in the cooperative board game Hanabi and the main takeaways from this work (Section 3).

2 THE TOMABD AGENT MODEL

Figure 1 presents the architecture of the TomAbd agent model.[1] The core ToM functionality is provided by the AdoptViewpoint function. Agent i operates according to the logic program T_i contained in its belief base. T_i is composed of a set of ground literals l and Horn clauses $h \leftarrow b$. Among the clauses are domain-dependent *ToM clauses*. ToM clauses have literals `believes(Ag,F)` as their head, to express the fact that agent `Ag` believes some fact `F` to be true.

[1] A full implementation of TomAbs is available at: https://github.com/nmontesg/tomabd

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

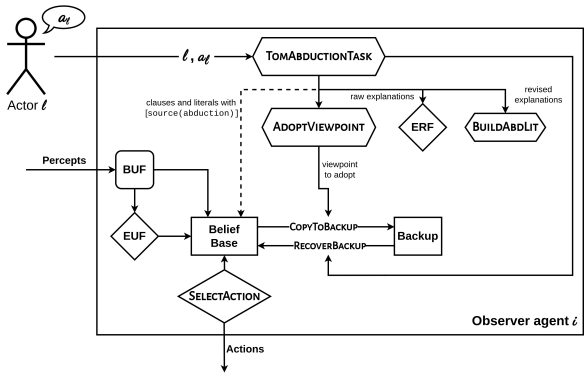

Figure 1: Architecture of the TomAbd agent model.

The ToM capabilities of TomAbd agents comes from their ability to substitute their logic program T_i with the logic program they estimate that others have, and reason from that perspective. Hence, the program that agent i estimates that agent j has is denoted by $T_{i,j}$ and given by:

$$T_{i,j} = \{\phi \mid T_i \models \texttt{believes}(j, \phi)\} \tag{1}$$

The switch from T_i to $T_{i,j}$ constitutes *first-order* ToM, as agent i adopts the beliefs that it estimates agent j has. However, eq. (1) can be extended down to an arbitrary level recursion:

$$T_{i,j,\ldots,k,l} = \{\phi \mid T_{i,j,\ldots,k} \models \texttt{believes}(l, \phi)\} \tag{2}$$

Hence, eq. (2) constitutes n^{th}*-order* ToM. We denote the sequence $[j, \ldots, k, l]$ of agents whose beliefs i recursively incorporates as the *perspective* that i adopts. The AdoptViewpoint function essentially takes as input a perspective P and adopts it by applying eq. (2). First, however, it saves a copy of the agent i's original program T_i in a back-up belief base, so that it can return to it later and continue the reasoning from i's perspective.

Now that we understand how TomAbd agents implement ToM, we present how they combine it with abductive reasoning to derive the beliefs motivating the action of other agents. The integration of ToM and abductive reasoning into a single functionality is provided by TomAbductionTask, which constitutes the core function of the TomAbd agent model.

Function TomAbductionTask takes as input: (i) an observer perspective *Pobs*; (ii) an actor agent l; and (iii) the action a_l executed by l. It starts by adopting the *perspective of the actor Pact*, generated by appending l to the observer perspective *Pobs*. Once the agent i is operating from the logic program it estimates the actor l to have (with as many intermediate perspective switches as indicated by *Pobs*), i proceeds to generate explanations for the observation $\texttt{action}(l,a_l)$. This step relies on an abductive meta-interpreter specific to the TomAbd agent model. This meta-interpreter is based on classical SLD clause resolution with a small extension to handle abducible literals. We use Φ to denote the set of explanations generated by the abductive meta-interpreter. Φ is revised for consistency against the logic program that i estimated the actor to have. Next, Φ is also checked for consistency against the logic program that i estimates the *observer* to have. Again, the observer's perspective is obtained by calling the AdoptViewpoint function. From these two consistency checks, agent i obtains two (possibly different) sets of abductive explanations: Φ_{act} and Φ_{obs} (revised from the actor's (observer's) perspectives). Φ_{act} and Φ_{obs} are transformed into a suitable format to be incorporated into agent i's own belief base, to be queried during its own decision-making process.

Next we provide an overview of the other components from Figure 1 that have not been mentioned so far. The *explanation revision function*, ERF, is called by TomAbductionTask to ensure the consistency of explanations both from the actor's and the observer's perspectives. BuildAbdLit is an auxiliary function to transform abductive explanations in a format suitable to be added to the agent i's belief base. The *explanation update function*, EUF, is called from the standard BDI *belief update function* BUF. EUF is triggered at every perception step of the BDI cycle, and is in charge of removing explanations from agent i's belief base if they are no longer informative. Finally, the SelectAction function is in charge of selecting agent i's next action, taking into account the abductive explanations currently present in its belief base.

3 RESULTS AND CONCLUSIONS

We have tested the TomAbd agent model on the Hanabi benchmark, a cooperative board game where agents can see each other's cards, but not their own, and can provide hints to other agents about their cards. The objective is to maximize a single team score by playing the cards. A full description of the game rules is available elsewhere.[2]

We have compared the performance in Hanabi of teams of TomAbd agents that use 1st-order ToM versus teams where agents do not use ToM capabilities. Our results show that teams where players use ToM score significantly higher, with acceptably low execution overhead. Besides the score, players with ToM exchange information more efficiently, and we also observed that a large percentage of the gain in score could be attributed to the information acquired by the agents through ToM reasoning.

In summary, our work presents a novel model for agents with Theory of Mind. It provides the cognitive machinery for agents to adopt and reason at multiple levels of perspective of their peers, and abduce the potential reasons leading their peers to act the way they do, and hence increasing the agent's own understanding of its environment. Our model endows autonomous agents with essential social abilities, which are becoming increasingly important in the current AI landscape.

Our model opens the door to several avenues for future work. Namely, the trade-offs between higher level of ToM reasoning, the increased uncertainty about the information abduced and the computational cost associated with recursive changes in perspective should be explored. This can potentially inform psychological research on the limits that humans have for adopting high levels of ToM.

ACKNOWLEDGMENTS

Authors acknowledge funding from the VAE project (#TED2021-131295B-C31), the VALAWAI project (HORIZON #101070930), and the TAILOR project (H2020 #952215).

[2] https://www.ultraboardgames.com/hanabi/game-rules.php

REFERENCES

[1] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. Cooperative AI: machines must learn to find common ground. *Nature* 593, 7857 (may 2021), 33–36. https://doi.org/10.1038/d41586-021-01170-0

[2] Chris Frith and Uta Frith. 2005. Theory of mind. *Current Biology* 15, 17 (sep 2005), R644–R645. https://doi.org/10.1016/j.cub.2005.08.041

[3] Joshua Knobe. 2005. Theory of mind and moral cognition: exploring the connections. *Trends in Cognitive Sciences* 9, 8 (aug 2005), 357–359. https://doi.org/10.1016/j.tics.2005.06.011

[4] Bertram Malle. 2022. *Theory of Mind.* DEF publishers, Champagne, IL.

[5] Nieves Montes, Michael Luck, Nardine Osman, Odinaldo Rodrigues, and Carles Sierra. 2023. Combining theory of mind and abductive reasoning in agent-oriented programming. *Autonomous Agents and Multi-Agent Systems* 37, 2 (aug 2023), 41. https://doi.org/10.1007/s10458-023-09613-w

[6] Ana Paiva et al. 2020. WP6 – Social AI: Learning and Reasoning in Social Contexts. https://www.tailor-social-ai.eu/home. ICT-48 TAILOR: Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization.

[7] Douglas Walton. 2014. *Abductive Reasoning.* University of Alabama Press, Tuscaloosa, AL. 320 pages.