Computational Theory of Mind with Abstractions for Effective Human-Agent Collaboration

Extended Abstract

Emre Erdogan
Utrecht University
Utrecht, Netherlands
e.erdogan1@uu.nl

Rineke Verbrugge
University of Groningen
Groningen, Netherlands
l.c.verbrugge@rug.nl

Pınar Yolum
Utrecht University
Utrecht, Netherlands
p.yolum@uu.nl

ABSTRACT

Empowering artificially intelligent agents with capabilities that humans use regularly is crucial to enable effective human-agent collaboration. One of these crucial capabilities is the modeling of Theory of Mind (ToM) reasoning: the human ability to reason about the mental content of others such as their beliefs, desires, and goals. However, it is generally impractical to track all individual mental attitudes of all other individuals and for many practical situations not even necessary. Hence, what is important is to capture enough information to create an approximate model that is effective and flexible. Accordingly, this paper proposes a computational ToM mechanism based on abstracting beliefs and knowledge into higher-level human concepts, called *abstractions*, similar to the ones that guide humans to effectively interact with each other (e.g., trust). We develop an agent architecture based on epistemic logic to formalize the computational dynamics of ToM reasoning. We identify important challenges regarding effective maintenance of abstractions and accurate use of ToM reasoning and demonstrate how our approach addresses these challenges over multiagent simulations.

KEYWORDS

Theory of Mind; Abstraction; Human-AI Collaboration

ACM Reference Format:
Emre Erdogan, Rineke Verbrugge, and Pınar Yolum. 2024. Computational Theory of Mind with Abstractions for Effective Human-Agent Collaboration: Extended Abstract. 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

Hybrid Intelligence (HI) [2] refers to the combination of human and machine intelligence for the purpose of enhancing human intellect instead of replacing it. Success of HI relies on the effectiveness of collaborations between humans and computational agents where both parties complement each other in tasks that they perform together and engage in effective interactions to foster productive collaborations. To realize effective human-agent collaboration, agents need to be empowered with capabilities that humans use on a daily basis. One of these crucial capabilities is called Theory of Mind (ToM) reasoning [14]. This ability enables humans to reason about the mental contents of others such as their knowledge, beliefs, desires, and goals, making it possible to understand and predict their behaviour. It is even possible for humans to use higher-order ToM reasoning to infer how others employ ToM, which helps humans interact efficiently (e.g., "I believe that Alice knows that I trust her greatly, so she should assume that I will not check her work.").

Recently, many computational ToM models have been developed to understand its effectiveness in a variety of settings [3–6, 12, 16]. Although the results are mostly encouraging and indicate that employing ToM yields improved performance for the studied tasks, the current models have not seen widespread use as a computational tool in many real-life situations, indicating that developing a practical computational ToM is rather difficult. Many existing ToM-using agent models begin by representing individual beliefs about others and constructing a ToM model from these. However, in complex real-life settings, computationally tracking all such individual beliefs about others can be a costly approach. To continue being effective in their interactions with humans over time, agents should be efficient in keeping, maintaining, and utilizing these beliefs.

One candidate solution to this problem comes from human behaviour, called *abstracting* [9]. As a problem-solving technique, abstracting enables us humans to form a broad understanding of the problem and helps us approximate what we should look for in the social interaction to reach our goals [13]. Consider *trust* as an abstraction, which serves as a backbone in collaboration and generally captures people's confidence in one another's abilities, reliability, and commitment [8]. By using the abstracting technique, one can efficiently leverage pertinent information about one's partner to make trust-based decisions. Combined with ToM, one can gain insights into whether the partner reciprocates trust, enabling informed choices about the actions to take in one's interactions.

In this paper, we propose a computational ToM design based on aggregating individual beliefs and knowledge into higher-level abstractions that serve as practical approximations for agents to use in human-agent collaboration. The utilization of abstractions in ToM reasoning, along with the computational mechanisms it requires, necessitates formalization, for which we provide an agent architecture based on epistemic logic [10]. Specifically, we provide a modular structure for storage and maintenance of individual beliefs, knowledge, and abstractions that the agent creates and updates over time. For this agent to effectively collaborate with humans, we highlight important challenges regarding maintenance of abstractions and accurate use of ToM reasoning, propose different mechanisms to study them in the context of a medical scenario, and provide an experimental evaluation over agent simulations.

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).

2 AGENT ARCHITECTURE

Our proposed agent architecture, which is based on epistemic logic [10], consists of three modules: The knowledge and belief module, the abstraction module, and the deliberation module.

Knowledge and Belief Module: Whenever the agent interacts with others, the information that reaches the agent is stored in this module. It is a dynamic set such that new knowledge and beliefs can be added to it in a conflict-free manner.

Abstraction Module: This module is for the storage and maintenance of the agent's abstractions: Human-inspired, abstract decision-making heuristics (e.g., trust, respect, affinity, etc.) which can guide agents in their interaction decisions. The agent creates and updates them with the help of logical derivation rules, called abstraction rules, which are also kept in the abstraction module. They can be different for each agent.

Deliberation Module: The deliberation module is for making decisions on how to interact with other agents. By way of deliberation rules, it defines how the agent will deliberate with others based on its abstractions (it does not use the knowledge and belief module). These rules can evolve over time based on interactions with others.

3 SCENARIO

To illustrate our ideas, we consider a human-agent collaboration scenario in which a computational agent doctor X and a human doctor Y collectively decide on the diagnosis of a patient Z (see [7] for the details of the scenario). We are particularly interested in situations featuring a conflict between X and Y (e.g., X and Y have different opinions on the diagnosis of Z) in which effective application of social skills, like trust, are needed for resolution [11, 15]. By aggregating its beliefs and knowledge that are contextually relevant, X can determine whether it should trust Y or not. More interestingly, X can also reason about how Y abstracts her knowledge and beliefs to decide whether to trust X or not (i.e., how Y does her own approximation for trust) with the help of its computational ToM of Y. These abstractions can help X in choosing the best response to go with when a dispute occurs regarding the diagnostic decision. Table 1 illustrates all possible options of action that X can take depending on the trust dynamics between itself and Y. Since both partners are trying to collaborate, their main *goal* would be to achieve the **converse** action in case of a conflict.

4 CHALLENGES

For our agent architecture to be useful in achieving effective human-agent collaboration, we need to address three important challenges. We briefly explain them below in the context of the conflict scenario:

Abstraction Consistency: X's knowledge and beliefs can change over time. Ideally, the abstractions that X holds should consistently and efficiently change with the knowledge and beliefs that pertain to the abstractions (e.g., X's trust in Y). At the same time, X should not constantly update its abstractions after every change in the world. What would be an efficient mechanism to realize monitoring and updating abstractions effectively?

Theory of Mind (ToM) Consistency: Y's knowledge and beliefs can also change over time which may change Y's own abstractions (e.g., Y's trust in X). It is up to X to understand such changes and react accordingly to update its ToM of Y. This requires not only monitoring one's own beliefs but also cross-checking expectations with actual behavior of others to stay consistent with the actual situation. What would be a robust mechanism to identify and resolve ToM inconsistencies effectively?

Goal Consistency: Y may lose her trust in X due to a change in Y's perception about X. If X's main goal is to **converse** with Y in case of a conflict, then X should take initiative to sustain Y's trust for itself, which requires doing more than passively monitoring to update others' ToM models. How can an agent decide to proactively take actions that are appropriate in different situations?

Table 1: Strategies that the agent doctor X can use for resolving conflicts with the human doctor Y.

Situation	*Action*
Both X and Y trust each other	**Converse** with Y for joint resolution
X trusts Y but Y does not trust X	**Agree** with Y's opinion
X does not trust Y but Y trusts X	**Persuade** Y with own opinion
Neither X nor Y trusts each other	Advise Y to **Consult** another doctor

5 DISCUSSION

Our work provides a novel approach in ToM-based agent modeling with explicit use and maintenance of abstractions. For evaluation, we have designed two computational agents to simulate the behaviours of the agent doctor and the human doctor, interacting in a setting where they can create new beliefs, update already existing ones, and take actions to resolve conflicts. Using this setting, we propose different solutions for the challenges highlighted in Section 4 and run agent-based simulations to evaluate their effectiveness (see [1]). In addressing the challenge of maintaining abstraction consistency, we have devised various strategies for updating abstractions, taking into account three crucial factors: Frequency (i.e., updating on predefined frequencies), change (i.e., updating only if a particular belief of the agent changes), and engagement (i.e., updating when the agent engages in a deliberation). We measure *abstraction-effectiveness* of these strategies, which is equal to the number of consistent abstractions used in deliberations per abstraction update, to evaluate their performance. Our results show that for a strategy to be abstraction-effective, it should update abstractions only before deliberation and only if an abstraction-related change happens in the agent's system. Furthermore, we have designed additional strategies to study the remaining challenges and measure their performance in additional simulation experiments. Overall, our simulation results indicate that (i) truthful communication plays an important role in achieving ToM consistency and becomes more effective when done comprehensively, and (ii) agents need to be proactively interacting with others to create preferred abstractions which in turn results in high goal consistency.

ACKNOWLEDGMENTS

This research was funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research (https://www.hybrid-intelligence-centre.nl).

REFERENCES

[1] [n.d.]. Computational Theory of Mind with Abstractions - Project Page. https://git.science.uu.nl/e.erdogan1/ToM_Project. Accessed: 2024-02-08.

[2] Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, et al. 2020. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer* 53, 08 (2020), 18–28.

[3] Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1, 4 (2017), 1–10.

[4] Harmen de Weerd, Rineke Verbrugge, and Bart Verheij. 2015. Higher-order theory of mind in the tacit communication game. *Biologically Inspired Cognitive Architectures* 11 (2015), 10–21.

[5] Harmen De Weerd, Rineke Verbrugge, and Bart Verheij. 2022. Higher-order theory of mind is especially useful in unpredictable negotiations. *Autonomous Agents and Multi-Agent Systems* 36, 2 (2022), 30.

[6] Sandra Devin and Rachid Alami. 2016. An implemented theory of mind to improve human-robot shared plans execution. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 319–326.

[7] Emre Erdogan, Frank Dignum, Rineke Verbrugge, and Pınar Yolum. 2022. Abstracting Minds: Computational Theory of Mind for Human-Agent Collaboration. In *HHAI2022: Augmenting Human Intellect*. IOS Press, 199–211.

[8] Paul W Mattessich and Barbara R Monsey. 1992. *Collaboration: what makes it work. A review of research literature on factors influencing successful collaboration.* ERIC.

[9] Sally McBrearty and Alison S Brooks. 2000. The revolution that wasn't: a new interpretation of the origin of modern human behavior. *Journal of human evolution* 39, 5 (2000), 453–563.

[10] J-J Ch Meyer and Wiebe Van Der Hoek. 2004. *Epistemic logic for AI and computer science.* Number 41. Cambridge University Press.

[11] Jakki Mohr and Robert Spekman. 1994. Characteristics of partnership success: partnership attributes, communication behavior, and conflict resolution techniques. *Strategic management journal* 15, 2 (1994), 135–152.

[12] 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 (2023), 36.

[13] George Polya. 2004. *How to solve it: A new aspect of mathematical method.* Vol. 85. Princeton university press.

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

[15] Kenneth W Thomas. 1992. Conflict and conflict management: Reflections and update. *Journal of organizational behavior* (1992), 265–274.

[16] Alan F. T. Winfield. 2018. Experiments in Artificial Theory of Mind: From Safety to Story-Telling. *Frontiers in Robotics and AI* 5 (2018), 75.