Hintikka's World: Agents with Higher-order Knowledge

François Schwarzentruber
Univ Rennes, CNRS, IRISA
francois.schwarzentruber@ens-rennes.fr

Abstract

In this demonstration paper, we present a pedagogical tool called *Hintikka's world* for showing how artificial agents can reason about higher-order knowledge (an agent knows that another agent knows that...). The system provides famous AI examples such as Muddy children and Russian cards. The system also allows to implement user's own examples via the description of a Kripke model or via its generation by the generic tableau method prover MetTeL2.

1 Introduction

Higher-order knowledge of agents is relevant in many applications: game theory [Aumann, 1999], robotics ([Scassellati, 2002], [Devin and Alami, 2016]), specifications of distributed systems [Halpern and Fagin, 1989], etc. Dynamic epistemic logic (DEL) ([Baltag *et al.*, 1998], [van Ditmarsch *et al.*, 2008]) extends epistemic logic for describing and reasoning about epistemic properties and information change. The famous tool in the community is called DEMO [van Eijck, 2007] and is a model checker for DEL, that has been used in practice [van Ditmarsch *et al.*, 2012]. It also provide symbolic techniques [van Benthem *et al.*, 2015].

Nevertheless, there are no tools with an intuitive graphical user interface that may be used by roboticians, game theorists, psychologists, etc. In this paper, we present such a tool called *Hintikka's world*.

The idea of tool we propose, called *Hintikka's world* is simple: represent Kripke models by comic strips, as shown in Figure 1. The tool is available at the following address: `http://hintikkasworld.irisa.fr/`.

Hintikka's world is a proof of concept of a graphical user interface that shows artificial agents mental states. It could be used in debugging contexts and for explaining behaviors of the agents that takes their decisions with respect to their beliefs. In other words, it takes part in Explainable Artificial Intelligence. The artificial agent could be a humanoid robot that interact with humans ([Scassellati, 2002], [Devin and Alami, 2016]) or several autonomous agents that have imperfect information [Saffidine *et al.*, 2018].

Another application is to provide a tool for psychiatrists to test ability of children to reason about higher-order knowledge (see [Arslan *et al.*, 2015], [Wimmer and Perner, 1983]).

Finally, the tool has a pedagogical aim. It illustrates the concepts of Kripke models, modal formulas, model checking and satisfiability problem in a modal logic course. It also enables to explain easily how to model higher-order knowledge to other scientists.

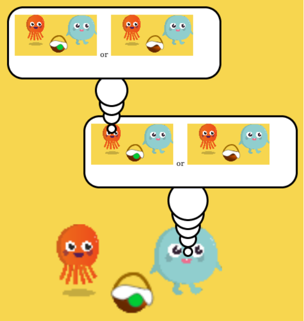

Figure 1: Graphical user interface of *Hintikka's world*

2 Demonstration Outline

2.1 Already Implemented Examples

First, the user can run AI examples that illustrate important concepts:

- Agents can learn information from messages of the form 'an agent does not know...': Muddy children puzzle, Consecutive numbers [van Ditmarsch and Kooi, 2015]
- False beliefs of agents: Sally and Anne [Wimmer and Perner, 1983]
- Public announcements that are secure in the sense that intruders of a system do not not learn relevant information: Russian cards [van Ditmarsch, 2003]
- Evolution of knowledge in asynchronous systems [Knight *et al.*, 2017]
- Evolution of knowledge in agents that run knowledge-based programs over a QdecPOMDP [Saffidine *et al.*, 2018].
- Simulation of cellular automata, for proving undecidability of epistemic planning [Sébastien Lê Cong, 2018].

2.2 User Interaction

The tool adopts the point of view of Halpern and Vardi for modeling an epistemic situation: model checking is more suitable than theorem proving [Halpern and Vardi, 1991]. In other words, the current situation is modeled as a pointed Kripke model. By clicking on a given agent a, the interface opens a thought bubble that displays the possible worlds for agent a. Actually, the comic strips shows the unfolding of the current pointed Kripke model that represents the current situation.

On the left, the software shows buttons for possible actions (public announcement, public actions, private actions, etc.). Actions are modeled by pointed event models of Dynamic epistemic logic [Baltag *et al.*, 1998]. By clicking on a button, the corresponding action is executed: the product of the pointed Kripke model and the pointed event model becomes the current pointed Kripke model.

2.3 Building New Examples

The tool also allows the final user to building their own examples. They are two ways to specify a new epistemic situation. First, the user can describe the pointed Kripke models in JavaScript, by giving the list of worlds, their valuations and the epistemic relations. Second, the user can specify the situation by a formula ϕ epistemic logic. The BNF is:

$$\begin{aligned}\phi := \quad & p \mid (\texttt{not}\ \phi) \mid (\phi\ \texttt{and}\ \phi) \mid (\phi\ \texttt{or}\ \phi) \\ & \mid (\texttt{K a}\ \phi) \mid (\texttt{Kpos a}\ \phi) \mid (\texttt{CK G}\ \phi) \mid (\texttt{CKpos G}\ \phi)\end{aligned}$$

where p is an atomic proposition, a is an agent and G is a group of agents. E.g. 'p does not holds but agent a imagines that it is possible that p holds' (`((Kpos a p) and (not p))`), agent a and b commonly know that agent c does not know the value of p (`(CK (a b) ((not (K b p)) and (not (K b (not p)))))`), etc. The user writes a set of formulas, one formula per line.

Then the system solves the satisfiability problem and generates a pointed epistemic model.

3 System Description

3.1 Class Architecture

Figure 2 shows the main part of the architecture of *Hintikka's world*. The interesting part is the fact that the graphical user interface (GUI) is independent from the current example that is running (muddy children, Sally and Anne, etc.). In particular, adding a new example only requires to add a new class that inherits from `World` and to implement the method for drawing the scene from data (valuations, numbers, etc.) that are members of the class.

3.2 Model Checking

The tool highly rely on model checking. Indeed, for instance, performing the public announcement of ϕ requires to compute the subset of worlds in which ϕ holds and to prune the current Kripke model. We chose to write the model checking procedure in Javascript. Since model checking is in PTIME – thus is an easy task – and is used intensively, it suitable to run run it on the client-side for performance reasons.

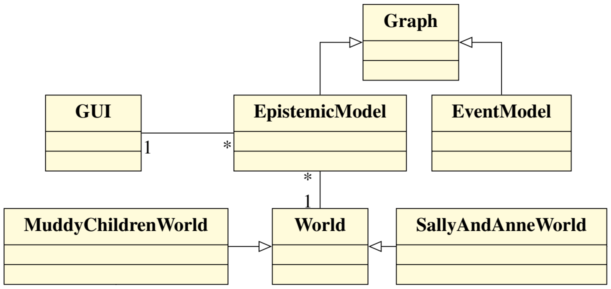

Figure 2: Architecture of *Hintikka's world*

3.3 Satisfiability Problem

By default, the system runs a tableau method for modal logic $KD45_n$ (logic for beliefs). Other logics, as K_n or $S5_n$ are available. Termination of the procedure is granted by a blocking rule that merges two possible worlds that contain the same subformulas (see p. 208 of [Horrocks *et al.*, 2007]). The tableau method is implemented in the generic tableau prover MetTeL2 [Tishkovsky *et al.*, 2012]. As the satisfiability problem of epistemic modal logic with common knowledge is EXPTIME-complete [Halpern and Moses, 1992], we require the use of an efficient tool although generic, which justify the use of MetTeL2. MetTeL2 runs on the server side for efficiency reasons.

4 Future Work

Reasoning tool. The model checking is in P but actions correspond to a product update in DEL [Baltag *et al.*, 1998] and the size of the Kripke model is exponential in the number of performed actions in worst case. That is why we plan to extend the tool with succinct Kripke models ([van Benthem *et al.*, 2015], [Charrier and Schwarzentruber, 2015]). We also want to extend the tool by implementing algorithms for epistemic planning (even bounded epistemic planning because epistemic planning is undecidable in the general case ([Bolander and Andersen, 2011], [Aucher and Bolander, 2013], [Sébastien Lê Cong, 2018])) and arbitrary public announcements ([Charrier and Schwarzentruber, 2015]).

Graphical user interface. By clicking on agents, the interface shows some posible worlds. We want to implement heuristics for displaying the most relevant epistemic worlds when there are too many possible worlds for a given agent.

Acknowledgments

The author thanks Tristan Charrier, Sébastien Lê Cong, Sophie Pinchinat, and Noël Plouzeau for their helpful comments. We thank Renate Schmidt and Dmitry Tishkovsky for their help to use MetTeL2. The author also thanks the art school LISAA Rennes, and especially François Camard, Florent Chamiot-Poncet, Anne-Laure Delacroix, Oriane Desfontaine and Clémentine Lucas for *Hintikka's world* artwork. We also thank the anonymous reviewers for their valuable comments.

References

[Arslan *et al.*, 2015] Burcu Arslan, Rineke Verbrugge, Niels Taatgen, and Bart Hollebrandse. Teaching children to attribute second-order false beliefs: A training study with feedback. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015, Pasadena, California, USA, July 22-25, 2015*, 2015.

[Aucher and Bolander, 2013] Guillaume Aucher and Thomas Bolander. Undecidability in epistemic planning. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, pages 27–33, 2013.

[Aumann, 1999] Robert J. Aumann. Interactive epistemology I: knowledge. *Int. J. Game Theory*, 28(3):263–300, 1999.

[Baltag *et al.*, 1998] Alexandru Baltag, Lawrence S Moss, and Slawomir Solecki. The logic of public announcements, common knowledge, and private suspicions. In *Proceedings of the 7th conference on Theoretical aspects of rationality and knowledge*, pages 43–56. Morgan Kaufmann Publishers Inc., 1998.

[Bolander and Andersen, 2011] Thomas Bolander and Mikkel Birkegaard Andersen. Epistemic planning for single and multi-agent systems. *Journal of Applied Non-Classical Logics*, 21(1):9–34, 2011.

[Charrier and Schwarzentruber, 2015] Tristan Charrier and François Schwarzentruber. Arbitrary public announcement logic with mental programs. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2015, Istanbul, Turkey, May 4-8, 2015*, pages 1471–1479, 2015.

[Devin and Alami, 2016] Sandra Devin and Rachid Alami. An implemented theory of mind to improve human-robot shared plans execution. In *The Eleventh ACM/IEEE International Conference on Human Robot Interation, HRI 2016, Christchurch, New Zealand, March 7-10, 2016*, pages 319–326, 2016.

[Halpern and Fagin, 1989] Joseph Y. Halpern and Ronald Fagin. Modelling knowledge and action in distributed systems. *Distributed Computing*, 3(4):159–177, 1989.

[Halpern and Moses, 1992] Joseph Y. Halpern and Yoram Moses. A guide to completeness and complexity for modal logics of knowledge and belief. *Artif. Intell.*, 54(2):319–379, 1992.

[Halpern and Vardi, 1991] Joseph Y. Halpern and Moshe Y. Vardi. Model checking vs. theorem proving: A manifesto. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR'91). Cambridge, MA, USA, April 22-25, 1991.*, 1991.

[Horrocks *et al.*, 2007] Ian Horrocks, Ullrich Hustadt, Ulrike Sattler, and Renate Schmidt. 4 computational modal logic. In Patrick Blackburn, Johan Van Benthem, and Frank Wolter, editors, *Handbook of Modal Logic*, volume 3 of *Studies in Logic and Practical Reasoning*, pages 181 – 245. Elsevier, 2007.

[Knight *et al.*, 2017] Sophia Knight, Bastien Maubert, and François Schwarzentruber. Reasoning about knowledge and messages in asynchronous multi-agent systems. *Mathematical Structures in Computer Science*, page 1–42, 2017.

[Saffidine *et al.*, 2018] Abdallah Saffidine, François Schwarzentruber, and Bruno Zanuttini. Knowledge-based policies for qualitative decentralized pomdps. In *In Sheila McIlraith and Kilian Weinberger, editors, Proc. 32nd AAAI Conference on Artificial Intelligence (AAAI 2018)*, 2018.

[Scassellati, 2002] Brian Scassellati. Theory of mind for a humanoid robot. *Auton. Robots*, 12(1):13–24, 2002.

[Sébastien Lê Cong, 2018] François Schwarzentruber Sébastien Lê Cong, Sophie Pinchinat. Small undecidable problems in epistemic planning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI) and the 23rd European Conference on Artificial Intelligence (ECAI), Stockholm, 13-19 July 2018*, 2018.

[Tishkovsky *et al.*, 2012] Dmitry Tishkovsky, Renate A. Schmidt, and Mohammad Khodadadi. The tableau prover generator mettel2. In *Logics in Artificial Intelligence - 13th European Conference, JELIA 2012, Toulouse, France, September 26-28, 2012. Proceedings*, pages 492–495, 2012.

[van Benthem *et al.*, 2015] Johan van Benthem, Jan van Eijck, Malvin Gattinger, and Kaile Su. Symbolic model checking for dynamic epistemic logic. In *Logic, Rationality, and Interaction - 5th International Workshop, LORI 2015 Taipei, Taiwan*, pages 366–378, 2015.

[van Ditmarsch and Kooi, 2015] Hans van Ditmarsch and Barteld Kooi. *One Hundred Prisoners and a Light Bulb*. Springer, 2015.

[van Ditmarsch *et al.*, 2008] Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi. *Dynamic Epistemic Logic*. Springer, Dordecht, 2008.

[van Ditmarsch *et al.*, 2012] Hans van Ditmarsch, Jan van Eijck, Ignacio Hernández-Antón, Floor Sietsma, Sunil Simon, and Fernando Soler-Toscano. Modelling cryptographic keys in dynamic epistemic logic with DEMO. In *Highlights on Practical Applications of Agents and Multi-Agent Systems, PAAMS 2012 Special Sessions, Salamanca, Spain, 28-30 March, 2012*, pages 155–162, 2012.

[van Ditmarsch, 2003] Hans P. van Ditmarsch. The russian cards problem. *Studia Logica*, 75(1):31–62, 2003.

[van Eijck, 2007] Jan van Eijck. Demo—a demo of epistemic modelling. In *Interactive Logic. Selected Papers from the 7th Augustus de Morgan Workshop, London*, volume 1, pages 303–362, 2007.

[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.