

Model Checking Epistemic Properties

Wiebe van der Hoek

*Department of Computer Science
University of Liverpool
wiebe@csc.liv.ac.uk*

Introduction

I will report on work that has been done in the Agent ART group in Liverpool for the last two years on model checking epistemic properties. In this abstract, I will do no justice to other work in this area, and I will only refer to such work as far as it relates to ours.

Although model checking is a dominant verification technique for Multi Agent Systems, the emphasis has been on temporal, rather than for instance epistemic properties. To the best of our knowledge, there are no model checkers available to which one can directly feed epistemic properties for verification.

Knowledge and Linear Time: Local Propositions

In [?] we use the concept of *local propositions*, as introduced in [?], to propose a way to model check knowledge properties, with an underlying *interpreted systems semantics* ([?]) in Linear Time systems. Let us use LTL for the language of Linear Time Logic, and CKL for the propositional temporal logic with operators K_i (i an agent) and C for Common Knowledge.

Roughly, the idea is as follows. Call a propositional formula φ to be *i -local* if agent i knows its value, i.e., if in all states u and v with $u \sim_i v$, formula φ has the same truth-value. For any CKL formula ψ , and an interpreted system $\mathcal{I} = \langle \mathcal{R}, \pi \rangle$, we are now interested whether ψ holds initially in every run r in \mathcal{I} , or more formally, whether $(\langle \mathcal{R}, \pi \rangle, (r, 0)) \models_{\text{CKL}} \psi$, for all $r \in \mathcal{R}$. Let us abbreviate this to $mc_{\text{CKL}}(\mathcal{I}, \psi)$. Our aim is now to reduce this problem to an LTL model checking problem $mc_{\text{LTL}}(\mathcal{I}, \psi')$.

The idea to do this is simple: rather than the agent's knowledge, we use his (propositional) evidence: Suppose ψ is an epistemic formula $K_i\beta$, and that there is an i -local property φ . Then, verify that $mc_{\text{LTL}}(\mathcal{I}, \diamond\varphi \wedge \square(\varphi \rightarrow \beta')$, where β' is obtained from β using the same procedure. We show that this procedure is 'correct', and will discuss an example. Also, we will argue that in an *adversarial* setting, i.e. one in which absence of knowledge is to be proven, this technique is rather cumbersome.

Alternating-time Temporal Epistemic Logic

The framework ATL extends CTL in that it generalises the path quantifiers A ("on all paths") and its dual E to *coalition modalities* $\langle\langle \Sigma \rangle\rangle$ for every set of agents Σ , with intended meaning of $\langle\langle \Sigma \rangle\rangle\varphi$, that the coalition Σ has a strategy, so that, no matter what the agents outside Σ do, φ will be true. In ATL, φ 's main operator must be a temporal one.

In [?] we proposed to enrich ATL with an epistemic component, giving rise to a language ATEL. In this language a whole range of properties are expressible, referring to (secret) communication, like $K_a\phi \wedge \neg K_b\phi \wedge \neg K_c\phi \wedge \langle\langle a, b \rangle\rangle \diamond (K_b\phi \wedge \neg K_c\phi)$, or to knowledge pre-conditions, as in $(\neg \langle\langle a \rangle\rangle \circ \phi) \mathcal{U} K_a\psi$ or as in the ATL* formula $K_b(c = s) \rightarrow \langle\langle b \rangle\rangle (\langle\langle b \rangle\rangle \circ o) \mathcal{U} \neg(c = s)$ ("If Bob knows the code c of the safe s , he is able to open it until the code changes"). In the simplest setting (no interaction between

knowledge and abilities assumed), the setting that is the subject of [?], the model checking problem for ATEL is PTIME-complete.

Agents that Know how to Play

Once a basic framework for ATL with epistemics is in place, it becomes interesting to look at properties one can impose on agents with incomplete information. Some first steps in this direction are taken in [?], although perhaps the main message of that paper is that matters can become pretty complicated. One main source of complication in ATL is the way a *strategy* is defined, which assumes perfect information and perfect memory by the agents. Analogous to an insight by Moore ([?]), it is noted in [?] that the distinction between *de dicto* and *de re* when referring to the knowledge of having a strategy is helpful. [?] then proposes various ways to restrict the set of allowed strategies, but especially for the multi-agent abilities, it remains unclear to how exactly interpret a formula like $C_\Sigma \langle\langle \Sigma \rangle\rangle \diamond win_\Sigma$; even if it is common knowledge within Σ that they have a strategy for winning, this does not immediately imply that this strategy will be played (e.g., it may not be unique!).

There are some other unexpected logical problems with expressing intuitively simple ATEL properties. As an example, take the scheme $\langle\langle i \rangle\rangle \circ \varphi \rightarrow K_i \langle\langle i \rangle\rangle \circ \varphi$: agents are aware of what they can achieve. The most obvious requirement to impose this seems to be that agents have the same capabilities in indistinguishable states (formally: $q \sim_i q' \Rightarrow \delta(q, i) = \delta(q', i)$, where $\delta(q, i)$ represents the set of possible sets of next states that i can bring about, in q). However, as recently demonstrated by Ågotnes ([?]) things don't work out that way. The key problem here is that, even if the choices of agent i are 'the same' in two states q and q' , it is possible that he can achieve $\circ p$ in state q because of the other agents' limited abilities there, while the other agents' abilities are not so much restricted in q' .

Acknowledgements

The work reported here is based on joint work with Wojciech Jamroga, Sieuwert van Otterloo and Michael Wooldridge. During the past years, our progress in this benefited a lot from work and discussions with Valentin Goranko, Alessio Lomuscio, Carsten Lutz, Ron van der Meyden, Wojciech Penczek, Dirk Walther, and Frank Wolter.