Towards a Verification Framework for Communicating Rational Agents

Nils Bulling[1] and Koen V. Hindriks[2]

[1]Department of Informatics
Clausthal University of Technology, Germany
bulling@in.tu-clausthal.de

[2] Faculty Electrical Engineering, Mathematics and Computer Science
Delft University of Technology, The Netherlands
k.v.hindriks@tudelft.nl

Abstract. We present an abstract framework for verifying communicative actions for rational agent programming languages. Firstly, a multi-agent verification logic based on the computational semantics is introduced; and subsequently, this multi-agent logic is embedded into a more expressive modal logic over a standard run-based semantics. We formally relate both logics, prove expressivity results, and argue why it is useful to have a (more expressive) *standard* modal logic and semantics at hand.

1 Introduction

In the literature the gap between agent programming languages and agent logics has frequently been discussed and first steps for bridging and analysing it have been done [5, 6]. Just recently, a computational semantics for the GOAL agent programming language for communicative actions based on mental models was proposed in [4, 1].

In this paper we relate such agent programming languages offering communication abilities to a "standardized" agent logic. For this purpose we propose an abstract setting for communicating agents, based on the message-passing system introduced in [3], and extend the verification logic from [5] to be applicable to the communicative setting. We continue to show that the computational semantics can be embedded into a run-based modal semantics. This result can be seen as a conservative extension of the single-agent result presented in [5] to the multi-agent setting introduced here; however, in this paper we leave out some operators whose addition is straightforward.

Due to the space limitations we will often refer to [1] for more details and the proofs of the main theorems.

2 Preliminaries

In the following we present the basic multi-agent model, sketch the essentials of an agent programming language and its extension by communicative actions.

The Multi-Agent Model. For our multi-agent model we reuse the well-established theory on distributed systems from [3]. Our model assumes a fixed number of agents with associated *agent names* $\mathcal{A}gt = \{a_1, \ldots, a_n\}$. A *global state* g of a multi-agent system (MAS) is a tuple $\langle l_{a_1}, \ldots, l_{a_n}, l_e \rangle$ where l_{a_i} is the local state of agent a_i. We use g_a to denote the local state of agent a. The non-empty set $G = L_{a_1} \times \ldots \times L_{a_n}$ represents all *(global) states*.

In each state an agent a may perform an action, drawn from a set of *actions* Act_a where $Act_a \cap Act_b = \emptyset$ when $a \neq b$. We use α_i to denote actions and Act denotes the union of the action sets of all agents.

The effects of performing an action are represented by a *transition function* $\tau : G \times Act \rightarrow G$. Actions are assumed to update only the local state of the agent performing it. That is, $\tau(g, \alpha)_a = g_a$ whenever $\alpha \notin Act_a$. An exception to this rule will be made below for communicative actions. The behavior of a multi-agent system is given by a *run* r which is a mapping $\mathbb{N} \rightarrow G \times Act$. $r_1(i)$ (resp. $r_2(i)$) is used to denote the projection of $r(i)$ onto the first (resp. second) component of $r(i)$. We thus use an interleaving semantics to model the execution of a MAS, i.e. one action is executed per time step. A *multi-agent system model* $\mathcal{R}$, *system* for short, is defined as a set of runs.

Programming with Mental Models. Rational agents are programs that derive their choice of action from their beliefs and goals. An agent programming language provides a framework for programming with *mental models* that consist of an agent's beliefs and goals. Whereas in the single agent setting a mental model consists of the agent's own beliefs and goals only, in the multi-agent setting, that we consider here, we use the notion of a mental state that consists of mental models of other agents as well (cf. [4, 1]). The idea is that these mental models are used to (partially) reconstruct the beliefs and goals of another agent.

The beliefs and goals of an agent are declarative sentences which are represented in standard propositional logic $\mathcal{L}_{PL}$ built over a set of propositional atoms *Atom* and the usual Boolean connectives. $\models_{\text{PL}}$ denotes the usual consequence relation associated with $\mathcal{L}_{PL}$.

Formally, a *mental model* is a pair $\langle \Sigma, \Gamma \rangle$ with $\Sigma \subseteq \mathcal{L}_{PL}$ a *belief* and $\Gamma \subseteq \mathcal{L}_{PL}$ a *goal base* which satisfy the usual rationality constraints: (i),(ii) Consistency of beliefs and goals; and (iii) goals are not believed to be achieved (cf. [1, 4]).

Finally, a *mental state* is a mapping m from $\mathcal{A}gt$ to mental models, i.e. $m(a) = \langle \Sigma, \Gamma \rangle$ is a mental model for each $a \in \mathcal{A}gt$. The set of all mental states is denoted by $MS(\mathcal{A}gt)$. The intuition is that a mental state m_a encodes a's beliefs about b's beliefs and goals by mapping agent name b to a mental model $m_a(b) = \langle \Sigma, \Gamma \rangle$ where Σ encodes b's beliefs and Γ encodes b's goals.

Agents need to be able to inspect their mental state and the different mental models part of it. Thus, the language of *mental state conditions* over $\mathcal{A}gt$, $\mathcal{L}_{MS}(\mathcal{A}gt)$, is defined by: $\psi ::= \mathbf{B}^a\phi \mid \mathbf{G}^a\phi \mid \neg\psi \mid \psi \wedge \psi$ where $a \in \mathcal{A}gt$ and $\phi \in \mathcal{L}_{PL}$. The semantics of mental state conditions is defined relative to a mental state m where $m(a) = \langle \Sigma_a, \Gamma_a \rangle$. So, we have, for instance, $m \models_{\text{MS}} \mathbf{B}^a\phi$ iff $\Sigma_a \models_{\text{PL}} \phi$; and $m \models_{\text{MS}} \mathbf{G}^a\phi$ iff $\exists\gamma \in \Gamma_a$ such that $\gamma \models_{\text{PL}} \phi$. The semantics for negation and conjunction is given in the usual way.

Communication Among Agents. The communicative actions that we introduce here affect the mental state of the *receiving* agent. Following [4, 1], a communicative action is of the form $send(a, b, msg) \in Act_a$ where msg denotes a message that is being sent by agent a to b. Three indicators are introduced that intuitively correspond with the sentence types most often used in natural language: $\bullet$ for *declarative*, **?** for *interrogative*, and **!** for *imperative* sentences. Hence, a *message* is of the form $\bullet\phi$, $?\phi$, or $!\phi$ where $\phi \in \mathcal{L}_{PL}$.

3 The Formal Model for Communicating Agents

In this section we present a formal and abstract model for communicating rational agents based on the concepts introduced in Section 2 (again, we try to be as brief as possible and refer to [1] for a more detailed presentation). Mental states of an agent can be seen as concrete instantiations of the local states of Section 2; thus, we get $G = MS_{a_1} \times \ldots \times MS_{a_n}$ where MS_{a_i} denotes the set of mental states for agent a_i. The behavior of an agent is determined by its mental state. Here, the transition functions τ of Section 2 are therefore named *mental state transformers*. A corresponding run is called a(n) *(agent) trace*. $\tau(g, \alpha)_a(b)$ must satisfy the three rationality constraints of mental models for any $a, b \in \mathcal{A}gt$. Here, $\tau(g, \alpha)$ denotes a global state, $\tau(g, \alpha)_a$ denotes the mental state of agent a, and $\tau(g, \alpha)_a(b)$ denotes the mental model agent a associates with agent b, a notation we will often use below.

We extend a mental state transformer $\tau : G \times Act \to G$ to *message-passing mental state transformer* such that it can be applied to $send(a, b, msg)$ by imposing the following constraints:

1. If $b \neq a$, $\alpha \in Act_a$, $\alpha \neq send(a, b, \mathsf{m})$, then $\tau(g, \alpha)_b = g_b$
2. If $\alpha = send(a, b, \mathsf{m})$, then (i) $\tau(g, \alpha)_i = g_i \ \forall i \in \mathcal{A}gt \setminus \{b\}$,
   (ii) $\tau(g, \alpha)_b(i) = g_b(i) \ \forall i \in \mathcal{A}gt \setminus \{a\}$, and

$$\tau(g, \alpha)_b(a) := \begin{cases} \langle \Sigma_a \oplus \phi, \{\gamma \in \Gamma_a \mid \Sigma_a \oplus \phi \not\models \gamma\}\rangle & \text{if } \mathsf{m} = \bullet\phi \\ \langle \Sigma_a \ominus \phi, \Gamma_a\rangle & \text{if } \mathsf{m} = ?\phi \\ \langle \Sigma_a \ominus \phi, \Gamma_a \cup \{\phi\}\rangle & \text{if } \mathsf{m} = !\phi \end{cases}$$

Following [4, 1], communicating a message m thus modifies the mental model $\langle \Sigma_a, \Gamma_a\rangle$ of the sender a maintained by receiver b. $\oplus$ and $\ominus$ are understood as update and revision operators. For details we refer to [4, 1].

The Verification Language $\mathcal{L}_V$. The temporal language $\mathcal{L}_V$ to reason about communicating agents is an extension of the verification logic introduced in [5]. We enrich the logic by $\mathbf{B}_a^b\phi$ (a believes that b believes ϕ), and by $\mathbf{G}_a^b\phi$ (a believes that b has goal ϕ). The *verification language* $\mathcal{L}_V$ is given by the set of formulae χ defined by the following grammar:

$$\chi ::= \mathbf{B}_a^b\phi \mid \mathbf{G}_a^b\phi \mid \neg\chi \mid \chi \wedge \chi \mid \chi\mathbf{U}\chi \mid \mathbf{X}\chi \mid done_a(\alpha)$$

where $\phi \in \mathcal{L}_{PL}$, $\alpha \in Act_a$ and $a, b \in \mathcal{A}gt$. We also write $\mathbf{B}_a$ for $\mathbf{B}_a^a$ and $\mathbf{G}_a$ for $\mathbf{G}_a^a$. A trace generated by several agents and a message passing mental state

transformer serves as a model for $\mathcal{L}_V$. Given such a trace t and a time point $i \in \mathbb{N}$ the semantics of $\mathcal{L}_V$-formulae is defined in a straightforward way:

$$
\begin{array}{ll}
t,i \models_{\mathrm{V}} \mathbf{B}_a^b \phi & \text{iff } g_a \models_{\mathrm{MS}} \mathbf{B}^b \phi \text{ where } g = t_1(i) \\
t,i \models_{\mathrm{V}} \mathbf{G}_a^b \phi & \text{iff } g_a \models_{\mathrm{MS}} \mathbf{G}^b \phi \text{ where } g = t_1(i) \\
t,i \models_{\mathrm{V}} \mathbf{X}\chi & \text{iff } t,i+1 \models_{\mathrm{V}} \chi \\
t,i \models_{\mathrm{V}} \chi \mathbf{U} \chi' & \text{iff } \exists j \geq i : t,j \models_{\mathrm{V}} \chi' \text{ and } \forall k : i \leq k < j \Rightarrow t,k \models_{\mathrm{V}} \chi \\
t,i \models_{\mathrm{V}} \mathit{done}_a(\alpha) & \text{iff } i > 0 \text{ and } t_2(i-1) = \alpha
\end{array}
$$

and in the usual way for negation and conjunction. This logic allows to verify basic properties of a multi-agent system.

4 Embedding $\mathcal{L}_V$ in the Modal Logic $\mathcal{L}_M$

One disadvantage of $\mathcal{L}_V$ is that it is *non-standard* and not very expressive. In this section we introduce the modal logic $\mathcal{L}_M$ which is used to reason about runs. Then, we relate the verification logic $\mathcal{L}_V$ and its semantics to the modal logic $\mathcal{L}_M$ and present expressiveness results.

$\mathcal{L}_M$: Syntax and Semantics. The language $\mathcal{L}_M$ given by the grammar:

$$\varphi ::= \mathsf{p} \mid \neg\varphi \mid \varphi \wedge \varphi \mid B_a\varphi \mid G_a\varphi \mid \bigcirc\varphi \mid \varphi \mathcal{U} \psi \mid \mathit{Done}_a(\alpha)$$

is built over atoms $\mathsf{p} \in \mathit{Atom}$ and the temporal constructs $\bigcirc\varphi$ for φ holds in the next state, $\varphi\mathcal{U}\psi$ for φ holds until ψ holds, belief operators $B_a\varphi$ for $a \in \mathcal{A}gt$ believes φ, goal operators $G_a\varphi$ for a has goal φ, and $\mathit{Done}_a(\alpha)$ for a has performed $\alpha \in \mathit{Act}$.

The behavior of a MAS is modelled by a set of runs (cf. Section 2); thus, an $\mathcal{L}_M$-model $\mathfrak{M}$ is a tuple $\langle \mathcal{R}, \{\mathcal{B}_a \mid a \in \mathcal{A}gt\}, \{\mathcal{G}_a \mid a \in \mathcal{A}gt\}, V\rangle$ consisting of a set $\mathcal{R}$ of runs, serial belief and goal accessibility relations, one for each agent $\mathcal{B}_a, \mathcal{G}_a \subseteq \mathcal{R} \times \mathbb{N} \times \mathcal{R} \times \mathbb{N}$, and a valuation function $V : \mathcal{R} \times \mathbb{N} \to \mathcal{P}(\mathit{Atom})$which labels states with the facts true in it.

Formulae are interpreted over $\mathcal{L}_M$-models in the standard way (see e.g. [3]). We use $\mathfrak{M}, r, i \models \varphi$ to denote that φ is satisfied on r at time i in model $\mathfrak{M}$. Again, we skip the standard cases (see [1] for more details):

$$
\begin{array}{ll}
\mathfrak{M}, r, i \models B_a\varphi & \text{iff } \forall (r', i') \in \mathcal{B}_a(r,i) : \mathfrak{M}, r', i' \models \varphi \\
\mathfrak{M}, r, i \models G_a\varphi & \text{iff } \forall (r', i') \in \mathcal{G}_a(r,i) : \mathfrak{M}, r', i' \models \varphi \\
\mathfrak{M}, r, i \models \bigcirc\varphi & \text{iff } \mathfrak{M}, r, i+1 \models \varphi \\
\mathfrak{M}, r, i \models \varphi\mathcal{U}\psi & \text{iff } \exists j : j \geq i \text{ and } \mathfrak{M}, r, j \models \psi \text{ s.t. } \forall k : i \leq k < j \Rightarrow \mathfrak{M}, r, k \models \varphi \\
\mathfrak{M}, r, i \models \mathit{Done}_a(\alpha) & \text{iff } i > 0 \text{ and } r_2(i-1) = \alpha \in \mathit{Act}_a
\end{array}
$$

We define $X_a(r,i) = \{(r', i') \mid X_a(r, i, r', i')\}$ for $X \in \{\mathcal{B}, \mathcal{G}\}$ and, as usual, abbreviate $B_a\varphi \wedge \varphi$ as $K_a\varphi$.

Equivalence and Correspondence Results We formally relate the logics $\mathcal{L}_V$ and $\mathcal{L}_M$ by embedding $\mathcal{L}_V$ into $\mathcal{L}_M$. We do so by introducing a translation tr from $\mathcal{L}_V$-formulae to $\mathcal{L}_M$-formulae defined as stated below:

$$tr(\mathbf{B}_a^b\phi) = \begin{cases} B_a B_b \phi & \text{if } a \neq b \\ B_a \phi & \text{if } a = b \end{cases} \qquad tr(\neg\varphi) = \neg tr(\varphi)$$

$$tr(\varphi \wedge \psi) = tr(\varphi) \wedge tr(\psi)$$

$$tr(\mathbf{G}_a^b\phi) = \begin{cases} B_a G_b \Diamond \phi & \text{if } a \neq b \\ G_a \Diamond \phi & \text{if } a = b \end{cases} \qquad tr(\mathbf{X}\varphi) = \bigcirc tr(\varphi)$$

$$tr(\varphi \mathbf{U} \phi) = tr(\varphi) \mathcal{U} tr(\psi)$$

$$tr(done_a(\alpha)) = Done_a(\alpha)$$

We show that this translation preserves truth which shows that the logic $\mathcal{L}_M$ can be used to reason about communicating agents instead of the non-standard $\mathcal{L}_V$. Our first result shows that $\mathcal{L}_M$ and its models are at least as expressive as $\mathcal{L}_V$ over traces; i.e., the modal logic can be used to reason about traces.[1]

Theorem 1. *Let t be a trace. Then there is an $\mathcal{L}_M$-model $\mathfrak{M} = \langle \mathcal{R}, \{\mathcal{B}_a \mid a \in \mathcal{A}gt\}, \{\mathcal{G}_a \mid a \in \mathcal{A}gt\}, V\rangle$ and a run $r^t \in \mathcal{R}$ such that for all $\varphi \in \mathcal{L}_V$ and $i \in \mathbb{N}$ we have: $t, i \models_V \varphi$ iff $\mathfrak{M}, r^t, i \models tr(\varphi)$.*

To obtain a correspondence result in the other direction, it is clear we need to impose some constraints on $\mathcal{L}_M$-models to ensure they model mental states and meet the rationality constraints of mental states and message passing mental state transformers. The consistency requirements for beliefs and goals are satisfied due to the seriality of the belief and goal relations. To match the third condition (goals are not believed to be achieved), we introduce the following postulate:

(R1) $\forall a, b \in \mathcal{A}gt: \ \mathcal{G}_a^b(r,i) \subseteq [\![\Diamond\varphi]\!]_{\mathfrak{M}} \Rightarrow \mathcal{B}_a^b(r,i) \not\subseteq [\![\varphi]\!]_{\mathfrak{M}}$

where $[\![\varphi]\!]_{\mathfrak{M}} := \{(r,i) \mid \mathfrak{M}, r, i \models \varphi\}$, the *denotation of* φ, consists of the points that satisfy φ and $\mathcal{B}_a^b(r,i) := (\mathcal{B}_b \circ \mathcal{B}_a)(r,i) = \{(r',i') \mid \exists (r'',i'') \in \mathcal{B}_a(r,i) : (r',i') \in \mathcal{B}_b(r'',i'')\}$. $\mathcal{G}_a^b := \mathcal{G}_b \circ \mathcal{B}_a$ is defined analogously. The subscript $\mathfrak{M}$ is omitted if clear from context.

In order to be able to match the communication semantics of message-passing mental state transformers, two additional postulates are required. Let r be a run and $X \in \{B, G\}$. The second postulate says that only the beliefs and goals of an action executing agent may change provided it is not a send action; and the third that only the mental state of the agent who receives the message is allowed to change in the prescribed way.

(R2) If $send(\cdot,\cdot,msg) \neq r_2(i) \in Act_a$ then for all $c, d \in \mathcal{A}gt$, $c \neq a$: $X_c(r,i) = X_c(r,i+1)$ and $X_c^d(r,i) = X_c^d(r,i+1)$

(R3) If $r_2(i) = send(a,b,msg)$ then for all $c, d \in \mathcal{A}gt$: $X_c(r,i) = X_c(r,i+1)$ and $X_c^d(r,i) = X_c^d(r,i+1)$ except if:
$msg = \bullet\varphi$ and φ consistent then $\mathcal{B}_b^a(r,i+1) \subseteq [\![\varphi]\!]$;
$msg = ?\varphi$ and φ no tautology then $\mathcal{B}_b^a(r,i+1) \not\subseteq [\![\varphi]\!]$;
$msg = !\varphi$ and φ no tautology then $\mathcal{B}_b^a(r,i+1) \not\subseteq [\![\varphi]\!]$ and $\mathcal{G}_b^a(r,i+1) \subseteq \mathcal{G}_b^a(r,i) \cap [\![\Diamond\varphi]\!]$.

[1] Proofs can be found in [1].

A run is called *trace-consistent* if it satisfies **(R1-3)**; and an $\mathcal{L}_M$-model is said to be *trace-consistent* if it contains at least one trace-consistent run.

Theorem 2. *Let $\mathfrak{M}$ be a trace-consistent $\mathcal{L}_M$-model. For each trace-consistent run r, all $\varphi \in \mathcal{L}_V$, and $i \in \mathbb{N}$: $\mathfrak{M}, r, i \models tr(\varphi)$ iff $t, i \models \varphi$.*

Benefits of the Modal Logic Approach. Why do we need two logics ($\mathcal{L}_V$ and $\mathcal{L}_M$) for the same purpose? An advantage of $\mathcal{L}_M$ is that it is more standard and thus better comparable to other logics, it is more expressive and allows to reuse existing results and tools (e.g. wrt. model checking). Hence, $\mathcal{L}_M$ seems especially suitable for the specification and verification of communication in MAS.

5 Conclusion and Related Work

We proposed first steps towards a theoretical model for *communicating rational agents*. We extended the verification logic from [5] to be applicable to the new setting and introduced a more expressive modal logic over standard models, based on [3], *to reason about communicating agents*. Links between both logics were established allowing to use the benefits of the "standard" modal logic.

The expressiveness of the logic to reason about communicating agents is limited compared to other logics that have been proposed [7, 2], and remains an issue for future research, but an advantage of our approach is that it is based on the computational semantics introduced in [4]. Our work is very much related to [5]; actually, it can be seen as an extension of it. Here, however, we are even more general and relate agent programming languages to standard modal logic rather than Cohen and Levesque's Intention Logic [2].

References

1. N. Bulling and K. V. Hindriks. Communicating Rational Agents: Semantics and Verification. In *Technical Report*, Clausthal, Germany, 2009. Clausthal University of Technology.
2. P.R. Cohen and H.J. Levesque. Communicative actions for artificial agents. In *Proc. of the 1st Int. Conf. on Multi-agent Systems (ICMAS'95)*, 1995.
3. R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge*. MIT, 1995.
4. K. V. Hindriks and M. B. van Riemsdijk. A Computational Semantics for Communicating Rational Agents Based on Mental Models. In *The 7th International Workshop on Programming Multiagent Systems (ProMAS'09)*, 2009.
5. K.V. Hindriks and W. van der Hoek. GOAL agents instantiate intention logic. In *Proc. of the 11th European Conference on Logics in Artificial Intelligence (JELIA'08)*, pages 232–244, 2008.
6. John-Jules Ch. Meyer. Our quest for the holy grail of agent verification. In *TABLEAUX '07: Proceedings of the 16th international conference on Automated Reasoning with Analytic Tableaux and Related Methods*, pages 2–9, Berlin, Heidelberg, 2007. Springer-Verlag.
7. M.P. Singh. *Multiagent Systems: A Theoretical Framework for Intentions, Know-How, and Communications*. Springer-Verlag, 1994.