A Logical Theory of Static Reasoning for Rational Agents

Gianpaolo Cugola
cugola@elet.polimi.it
Dipartimento di Elettronica e Informazione
Politecnico di Milano
P.zza Leonardo da Vinci 32
20133 Milano (Italy).
Tel.: +39-2-23993638
Fax: +39-2-23993411

February 5, 1996

Abstract

A formal language to describe mental states of rational agents is useful to specify the behavior of artificial agents that are supposed to act intelligently and to help human beings in evaluating their reasoning about the external environment.

This paper presents a multi-agent, propositional, temporal logic to describe the mental state of rational agents. The formal properties of beliefs and intentions of rational agents and the relationships between such mental attitudes are first informally sketched. Then, a logical language whose semantics captures these properties and relationships is presented. In particular, this logic provides two modal operators to describe rational agents' beliefs and intentions at each time instant. The above-mentioned properties of beliefs and intentions are formalized as logical axioms.

1 Introduction

One of the most exciting challenges for psychologists and AI researchers is to understand and formally describe the behavior of rational agents. In particular, they were interested in describing how the mental state of a rational agent is influenced by the external environment and how this mental state guides rational agent's actions. Therefore, during the last years, a great effort has been made to define logical languages to describe the mental state of rational agents. Once such a language had been defined, it would be possible to specify formally the behavior of artificial agents that are supposed to act intelligently, to specify how they should interact rationally with other artificial agents and human beings, and, finally, to help human beings in evaluating their reasoning about the external environment.

This paper gives a contribution to this research, pointing out the properties that a formal model of rational agents is expected to capture and defining a logical language whose semantics captures these properties.

The first step to obtain a language to describe the mental state of rational agents is to give a model of an agent situated in a changing environment. Figure 1 describes how a rational agent interacts with the external environment. Firstly, the rational agent *perceives* the external environment. Secondly, based on these perceptions, it develops a set of *beliefs* on the actual state of the external environment. Finally, from these beliefs, and from its *desires*, the agent develops a set of *intentions* that move its actual and future actions and influence its beliefs about the future. Perceptions, beliefs, desires and intentions are the agent's *mental attitudes*.

As shown in Figure 1, the different mental attitudes of a rational agent influence each others and interact to form its global *mental state*. The interactions between different mental attitudes can be classified into *static interactions* and *dynamic interactions* [12]. Static interactions describe the relationship between different mental attitudes at the same instant, whereas dynamic interactions describe how these mental attitudes change over time.

The dynamics of mental states is clearly related to perception. An agent totally isolated from the external environment does not have any reason to change its beliefs or its plans. According to this idea, a possible formalization of mental state dynamics can be obtained by giving a set of default rules which assert that no mental attitude changes over time, unless explicitly specified.

This paper does not formalize the rules which guide mental states dynamics, instead, it focuses on the static relationships between different mental attitudes. Future work will examine how mental state dynamics could be formalized and will extend the language developed in this paper to include this formalization.

Focusing on static reasoning, this paper does not take into account perception. Perception, in fact, is critical only when beliefs change over time. Moreover, in order to simplify the language, this work focuses on beliefs and intentions, leaving desires out. The reason of this choice is that, although high level desires certainly play an important role in determining agents' in-

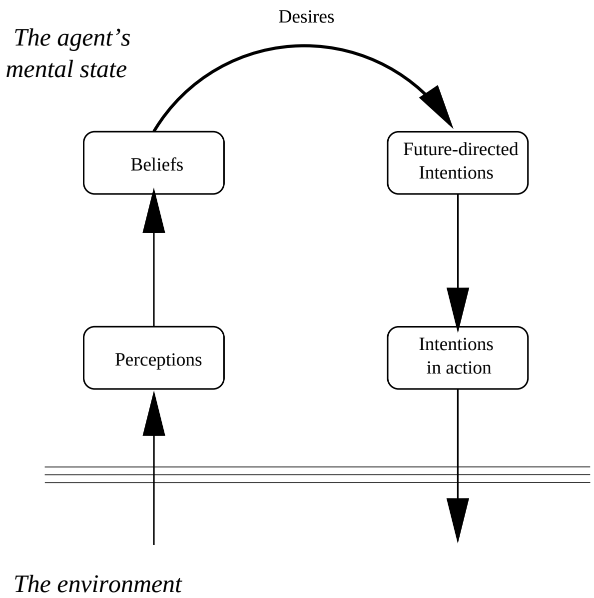

Figure 1: The model of a rational agent.

tentions, they can be omitted without limiting the possibility of modeling rational agents' behavior. In fact, once an agent has sorted out its desires, deciding those it wishes to achieve, the worlds it will be striving for are intended worlds.

In conclusion, this paper proposes a multi-agent, propositional, temporal logic, capable of expressing beliefs and intentions of rational agents. The formal semantics of this logic captures the properties of rational agents' intentions and beliefs and the static interactions between these mental attitudes as a set of constraints.

The paper is organized as follows: the next section describes the desirable properties of a theory of intentions and beliefs. Section 3 presents syntax and semantics of the language developed to describe rational agents.The set of constraints that the semantic relations describing mental attitudes have to satisfy in order to capture the desired properties of the theory, and the set of axioms of the proposed theory are introduced in Section 4. Subsequently, Section 5 compares this work with other related ones, highlighting the original contributions of this paper and, finally, Section 6 gives a conclusive summary of the paper and proposes some guidelines for further researches.

2 Expected properties of a theory of belief and intention

This section provides a list of properties that a formal model of a rational agent is expected to capture. As already mentioned, the present work mainly concerns the static interaction between intentions and beliefs of rational agents, therefore, this section firstly describes the expected properties of beliefs and intentions separately, then it describes the relationships between these mental attitudes at the same instant.

The properties of belief have been studied by philosophers and artificial intelligence researchers for a long time [6, 7, 10]. This paper adopts the well accepted model of belief due to Hintikka [7] (see also Halpern and Moses [6] for a detailed discussion of the properties of beliefs). According to this model, rational agents can not believe contradictory facts; if they believe a fact ϕ and believe that $\phi \Rightarrow \psi$, then they believe the fact ψ (they are "ideal believers", who know all logical consequences of their knowledge). Moreover, they are positive and negative introspective, that is, if they believe ϕ, then they believe that they believe ϕ, and if they do not believe ϕ, then they believe that they do not believe ϕ, respectively.

The theory of intention proposed in this paper is heavily influenced by Bratman's philosophical theory of plans and planning [2, 3] and by the theory of intention due to Searle [13].

First of all, Bratman (following Searle) distinguishes between "doing something intentionally" and "having the intention to do something". This work is mainly concerned with the second form of intention, that Searle calls "future-direct intention" (while the former is usually called "intention in action").

Bratman moves from the premise that the reasoning resources of rational agents are limited. Having such a constraint, agents form future-direct intentions to commit to certain courses of action, avoiding to take into account all possible alternatives, thus saving reasoning resources. In other words, agents form intentions to simplify the reasoning process, committing to certain courses of actions, and performing further reasoning on the assumption that they will actually carry out their plans. Moreover, future-direct intentions help in coordinating the activities of different agents: each agent may act on the assumption that other agents will achieve their plans.

Intentions of rational agents have to satisfy some constraints. Firstly, there are *consistency constraints*: the intentions of a rational agent have to be consistent among them and they have to be consistent with the beliefs of the agent. In other words, each agent has to believe that all its plans can be successfully executed in a world in which its beliefs are true.

Secondly, intentions must be reasonably stable. They can not be constantly subjected to reconsideration, otherwise there is no reason to adopt such intentions as a means to reduce the complexity of the reasoning process.

Thirdly, if an agent intends to do an action a and it believes that a implies b, then it does not have to intend necessarily b. This is the well-known "side-effect" problem: a rational agent does not need to intend all the believed consequences of its actions.

While Bratman's considerations about the roles played by agents' intentions in the reasoning process and the first and second constraint about intentions are reasonable, in our opinion the side-effect problem has to be better investigated. To assert that an operator that formalizes rational agents' intention does not have to distribute over believed implications, does not seem to be enough. Instead, in order to better formalize the side-effect problem, we need to distinguish between *logical implication* and *causal connection*. An example will better explain this concept. Suppose you intend to buy a red car, and suppose that, being a "rational" agent, you know that if you buy a red car then you buy a car. Having these premises, you must have the intention to buy a car. In fact, it is not possible to buy a red car without buying a car. Now, suppose you intend to get a tooth filled and suppose that, being uninformed about modern anesthetics, you believe that having your tooth filled will cause you much pain. Having these premises, it is correct to say that you choose to suffer but it is not correct to deduce that you intend to suffer. Now the questions are: what are the differences between the first and the second example? Why in the first example it is correct to deduce that you intend b, from the premises that you intend a and you believe that a implies b, while in the second one you could intend $\neg b$ even if the premises are the same?

To answer these questions it has to be observed that in the first example, action a (to buy a red car), *logically implies* action b (to buy a car). This means that, in every possible situation, if you perform action a then you perform action b too. A rational agent that intend to do a have to intend to do b, because it is not possible to do the first action without doing the second. Conversely, in the second example, action a (to have a tooth filled), *causes* action b (to feel pain). This means that there could be situations in which an agent performs the first action without causing the second. Therefore, an agent can intend to do a without intending b and this behavior is "rational" because it knows that a does not necessarily implies b.

In conclusion, a theory of intention is *side-effect-free* if the following principle holds:

The side-effect-free principle: A rational agent does not need to intend all the believed causal consequences of its actions, but it intends all the logically implied consequences of its actions.

Consequently, a language to specify the behavior of rational agents has to allow for the distinction between causal implication and logical implication. Unfortunately, modeling causality by using logic is a really difficult task. Many authors [14, 9, 15] tried to find a logical operator with the right properties to model causality, but nothing really convincing has been found yet. This work does not introduce a new logical operator to model causality but it uses the standard material implication connective '$\Rightarrow$' to approximate such an operator. As it will be formally explained in the next section, the approach to the side-effects problem chosen in this paper is to model intentions of rational agents in such a way that if the agent intends a and believes that $a \Rightarrow b$ then it does not necessarily have to intend b, while if a logically implies b ($M \models a \Rightarrow b$), then every agent that intends a must intend b too.

In addition to the characteristics of intention pointed out by Bratman, a formal model of rational agents could capture the fact that intentions are conscious. Accepting this assumption means to accept the fact that if a rational agent intends ϕ then it believes that it intends ϕ, and if it does not intend ϕ then it believes that it does not intend ϕ. In other words, such rational agents, besides being introspective with respect to beliefs, could be introspective with respect to intentions (*intentions introspective agents*). The correctness of this assumption is very controversial and this paper does not intend to take sides in this argument. Instead, Section 4 shows a constraint on the models of the logical language developed that can be adopted to model rational agents that are intentions introspective.

3 The language

The logical language defined in this section is a multiple agent, linear time, temporal language with two parametric modal operators, I_x and B_x, that allow the formalization of intentions and beliefs of a rational agent x, respectively.

3.1 Syntax

The alphabet of the language is composed of four disjoint, nonempty, denumerable sets:

- a set P of propositional symbols,
- a set A of action symbols,
- a set Agt of agent symbols,
- a set T of time point symbols,

and it comprises the following symbols:

- the symbol $\top$,
- the propositional connectives $\neg$ and $\Rightarrow$,
- the symbol $Doing$,
- the modal symbols: $B, I, Alw, Alw_f, Alw_p, Next, Prev$

Definition 1 *The well formed formulas (wffs) of the language are defined as follows:*

- *a propositional symbol p is a wff,*
- *if a is an action symbol and x is an agent symbol then: $(Doing_x\, a)$ is a wff,*
- *if ϕ and ψ are wffs, x is an agent symbol, and t is a time point symbol then:*

$$(\neg\phi), (\phi \Rightarrow \psi), \phi^t, B_x\,\phi, I_x\,\phi, Alw\,\phi, Alw_f\,\phi, Alw_p\,\phi, Next\,\phi, Prev\,\phi$$

are wffs. □

In the rest of this paper redundant parentheses will be omitted.

3.1.1 Rewritings

If ϕ and ψ are two wffs, then formulas $\phi \wedge \psi$ and $\phi \vee \psi$ stand for $\neg(\phi \Rightarrow \neg\psi)$ and $(\neg\phi) \Rightarrow \psi$, respectively.

If ϕ is a wff, a is an action symbol and t is a time point symbol then:

$\bot$	stands for	$\neg\top$
$I_x\, a$	stands for	$I_x(Doing_x\, a)$
$I_x\, a^t$	stands for	$I_x(Doing_x\, a)^t$
$Som\, \phi$	stands for	$\neg Alw\, \neg\phi$
$Som_f\, \phi$	stands for	$\neg Alw_f\, \neg\phi$
$Som_p\, \phi$	stands for	$\neg Alw_p\, \neg\phi$

3.2 Semantics

To simplify the language, a discrete time structure (the set of integers) is used. Moreover, every action is supposed to have a duration of one 'tick'.

3.2.1 Models

A model for a set of well formed formulas is a structure:

$$M =< W, \mathcal{AGT}, \mathbb{Z}, \mathcal{A}, \Pi, \tau, \alpha, \lambda, \mathcal{AGT}, \mathcal{B}, \mathcal{I} >$$

where:

- W is a set of possible worlds representing possible courses of events. Each possible world is a complete history specifying the state of the world at all instants of time.
- $\mathcal{AGT}$ is a set of agents.
- $\mathbb{Z}$ (the set of integers) represents instants of time.
- $\mathcal{A}$ is a set of concrete actions.
- $\Pi : W \times \mathbb{Z} \times P \rightarrow \{True, False\}$ is a function that assign truth values to propositional symbols at each time instant of every possible world.
- $\tau : T \rightarrow \mathbb{Z}$ is the interpretation function of time symbols.
- $\alpha : Agt \rightarrow \mathcal{AGT}$ is the interpretation function of symbols indicating agents.
- $\lambda : A \rightarrow \mathcal{A}$ is the interpretation function of action symbols.
- $\mathcal{ACT} : \mathcal{AGT} \times W \times \mathbb{Z} \rightarrow \mathrm{P}\, \mathcal{A}$ is a function that matches each agent with the set of actions that it is supposed to do in a certain world at a certain time.
- $\mathcal{B}, \mathcal{I} \subseteq (\mathcal{AGT} \times W \times \mathbb{Z}) \times W$ are the accessibility relations of the modal operators B and I, respectively.

3.2.2 Truth, satisfaction and validity of a wff

The notion of truth of a well formed formula ϕ at time n of a possible world w in a model M (written $M, w, n \models \phi$) is defined as follows:

$$
\begin{array}{lll}
M, w, n \models \top & & \\
M, w, n \models p & \text{iff} & \Pi(w, n, p) = True \\
M, w, n \models Doing_x\, a & \text{iff} & \lambda(a) \in \mathcal{ACT}(\alpha(x), w, n) \\
M, w, n \models \neg\phi & \text{iff} & M, w, n \not\models \phi \\
M, w, n \models \phi \Rightarrow \psi & \text{iff} & M, w, n \not\models \phi \text{ or } M, w, n \models \psi \\
M, w, n \models \phi^t & \text{iff} & M, w, n + \tau(t) \models \phi \\
M, w, n \models B_x\phi & \text{iff} & M, \overline{w}, n \models \phi \text{ for each } \overline{w} \in W \text{ such that} \\
 & & (\alpha(x), w, n)\mathcal{B}\,\overline{w} \\
M, w, n \models I_x\phi & \text{iff} & M, \overline{w}, n \models \phi \text{ for each } \overline{w} \in W \text{ such that} \\
 & & (\alpha(x), w, n)\mathcal{I}\,\overline{w} \\
M, w, n \models Alw\, \phi & \text{iff} & M, w, \overline{n} \models \phi \text{ for each } \overline{n} \in \mathbb{Z} \\
M, w, n \models Alw_f\, \phi & \text{iff} & M, w, \overline{n} \models \phi \text{ for each } \overline{n} \in \mathbb{Z} \text{ such that } \overline{n} > n \\
M, w, n \models Alw_p\, \phi & \text{iff} & M, w, \overline{n} \models \phi \text{ for each } \overline{n} \in \mathbb{Z} \text{ such that } \overline{n} < n \\
M, w, n \models Next\, \phi & \text{iff} & M, w, n + 1 \models \phi \\
M, w, n \models Prev\, \phi & \text{iff} & M, w, n - 1 \models \phi
\end{array}
$$

where $p \in P$, $x \in Agt$, $t \in T$, $a \in A$, and ϕ is a wff.

A well formed formula ϕ is true in a model M, (written $M \models \phi$), if:

$$\forall w \in W, n \in \mathbb{Z} \quad M, w, n \models \phi$$

Similarly, a well formed formula ϕ is valid if it is satisfied in all models.

4 Model constraints and axioms

4.1 Model constraints

Section 2 provided a list of properties of beliefs and intentions that a formal model of rational agents has to capture. To fulfill these properties, this section provides a set of constraints on the semantic relationships that model different mental attitudes.

First of all, relation $\mathcal{B}$ has to be Euclidean, transitive and serial. This ensures that beliefs are consistent and that agents are introspective [6].

Similarly, relation $\mathcal{I}$ has to be serial to ensure that intentions are consistent (an agent can not intend a and $\neg a$).

If the language has to model intentions introspective agents, semantic relationships $\mathcal{B}$ and $\mathcal{I}$ have to satisfy the following constraint:

$$\mathcal{B} \circ \mathcal{I} \equiv \mathcal{I}$$

where:

$$
\begin{array}{rcl}
\mathcal{B} \circ \mathcal{I} & = & \{(x, w_1, n, w_2) \in \mathcal{AGT} \times W \times \mathbb{Z} \times W \mid \\
 & & \exists w \in W((x, w_1, n)\mathcal{B}w \wedge (x, w, n)\mathcal{I}w_2\}
\end{array}
$$

The rest of this paper assumes the adoption of this constraint.

Modeling the fact that intentions of an agent have to be consistent with its beliefs presents some difficulties. As correctly observed by Rao and Georgeff in [11] there are three possibilities:

Realism This constraint states that the intention-accessible worlds of an agent are a subset of its belief-accessible ones. This constraint was firstly stated by Cohen and Levesque [5] and it ensures that the worlds intended by an agent are not ruled out by its beliefs. This constraint is unsatisfactory because it does not avoid some forms of the side-effect problem as observed by Cohen and levesque themselves and by Allen in [1]. Moreover, assuming this constraint, any belief about the future becomes adopted as a goal. This is the problem of transference cited by Rao and Georgeff in [11].

Weak realism To overcome the problems raised up by the realism constraint, it is possible to adopt the weak realism constraint. It states that, for each agent, the intersection between its belief-accessible worlds and its intention-accessible worlds is not empty. This form of the realism constraint ensures that the agent's intentions are consistent with its beliefs. At the same time, it avoids the side-effect problem.

Strong realism Another possibility is to suppose that the belief-accessible worlds of an agent are a subset of its intention-accessible ones. This formulation of the realism axiom has the same properties of the previous one, but it also ensures that if an agent intends a then it believes that it will achieve a.

The logic presented in this paper adopts the strong realism constraint:

Axiom 1 (Strong realism) *:*

$$\mathcal{B} \subseteq \mathcal{I}$$

The reason of this choice is that, as already mentioned, it ensures that the agent's intentions are consistent with its beliefs. At the same time, it makes sure that if an agent intends a then it believes that it will achieve a. This is the best way to guarantee that the intentions of each agent play the expected role in helping it in coordinating its activities. An example will explain this claim. Suppose John intend to go to the cinema tonight, and suppose he believes that if he goes to the cinema, he does not have the chance to meet his friend Paul. If, from his intention to go to the cinema tonight, John were not able to deduce that he will actually go to the cinema tonight, then he could believe that he will meet his friend Paul tonight too. Formally, if from $I_{John}\, goToTheCinema^{tonight}$ John is not able to deduce $B_{John}(Doing\, goToTheCinema)^{tonight}$ then the following formula would be satisfiable:

$$I_{John}\, goToTheCinema^{tonight} \quad \wedge$$

$$
\begin{aligned}
B_{John}((Doing_{John}\, goToTheCinema)^{tonight} \Rightarrow \\
\neg(Doing_{John}\, meetPaul)^{tonight}) \quad \wedge \\
B_{John}(Doing_{John}\, meetPaul)^{tonight}
\end{aligned}
$$

Observe that from John's *present* intention to go to the cinema tonight, he deduces the *present* belief that he will go to the cinema tonight. This does not mean that John will actually go to the cinema tonight. He could change his mind later, rejecting his intention to go to the cinema and his belief that he will go to the cinema, but this has to do with the dynamics of mental states, a problem which is completely neglected in the present paper.

Given these constraints, the language presented in this paper meets all the expected properties of intentions and beliefs stated in Section 2. In particular it does not suffer of the side-effect problem. An agent can intend ϕ, believe that $\phi \Rightarrow \psi$ and it is not confined to intend ψ. At the same time if an agent intend ϕ and ϕ logically implies ψ then the agent has to intend ψ too. This agrees with the intuition given in Section 2.

4.2 Axioms

Given the model constraints described in the previous section, it is easy to prove that the following formulas are valid. They capture as theoretical axioms the properties of static reasoning of rational agents that were informally stated in Section 2.

1. $(B_x\phi \wedge B_x(\phi \Rightarrow \psi)) \Rightarrow B_x\psi$
2. $B_x\phi \Rightarrow B_xB_x\phi$
3. $\neg B_x\phi \Rightarrow B_x\neg B_x\phi$
4. $\neg B_x\bot$
5. $(I_x\phi \wedge I_x(\phi \Rightarrow \psi)) \Rightarrow I_x\psi$
6. $\neg I_x\bot$
7. $I_x\phi \Rightarrow B_xI_x\phi$
8. $\neg I_x\phi \Rightarrow B_x\neg I_x\phi$
9. $I_x\phi \Rightarrow B_x\phi$

Formulas 1 to 4 capture the expected properties of beliefs. In particular Formula 1 states that rational agents are logically omniscient, Formulas 2 and 3 state that they are positive and negative introspective and Formula 4 states that their beliefs are consistent. Furthermore, Formula 6 formalizes the fact that agents' intentions have to be consistent and Formula 5 settles the principle that a rational agent that intends ϕ and that intends that ϕ implies ψ, intends ψ too. This last principle is motivated by the fact that if the agent intends ϕ and, at the same time, it intends to act in such a way that if ϕ is true than ψ is also true, then it has to intend ψ too. Finally, Formulas 7 and 8 are valid for intentions introspective agents and Formula 9 express as an axiom of the language the strong realism constraint introduced in Section 4(an agent intending ϕ beliefs that it will actually achieve ϕ).

As far as the side-effect problem is concerned, it is easy to prove that the theory of rational agents presented in this paper satisfies the side-effect-free principle stated in Section 2. In fact, it is always possible to find a model M such that:

$$M \models I_x\,\phi \wedge B_x(\phi \Rightarrow \psi) \wedge \neg I_x\,\psi$$

On the other hand, if $M \models \phi \Rightarrow \psi$ then $M \models I_x\,\phi \Rightarrow I_x\,\psi$.

5 Related work

This section provides a comparison between the language to model rational agents developed in this paper and other related works.

Cohen and Levesque, in [5], present a logical language to describe intentions of rational agents. They consider intentions as mental attitudes derived from beliefs and goals. According to Cohen and Levesque, intentions are a particular form of persistent goals. An agent intending to do an action a has the goal to do a and retracts this goal only when it believes that a has been already executed or that it will not be possible to execute a in the future. Cohen and Levesque take into account only atemporal intentions so that $(Intend\,x\,a)$ means that the agent x intends to do the action a sometimes in the future. Conversely, we consider intentions as a primitive mental attitude and we address timed intentions, that is, intentions in which the time of the intended action is explicitly mentioned. In our language, in fact, $I_x\,a^t$ means that the agent x intends to do the action a within t clock tick and $I_x\,a$ means that it intends to do the action a immediately (intention in action). Moreover, while the logic presented in this paper adopts the strong realism axiom, Cohen and Levesque language adopts the realism axiom. The main unsatisfactory consequence of this assumption is that an agent described using Cohen and Levesque logic adopts as a goal each fact that it believes (Cohen and Levesque logic suffers of the problem of transference).

Konolige and Pollack, in [8], claim that normal modal logics (NMLs), that is, logics in which the semantics of the modal operators is defined by accessibility relations over possible worlds, are ill-suited to a formal theory of intention. The problem related to NMLs is that they are closed under logical consequence and this does not make possible to avoid a particular form of the side-effect problem. In fact, using a language that adopts a normal modal semantics, it is not possible to avoid that if an agent intend ϕ and $\phi \models \psi$ then the agent has to intend ψ too. According to Konolige and Pollack this form of consequential closure can not be assumed for rational agents' intentions. They present an alternative model of intention based on the minimal model semantic of Chellas [4] that avoids this kind of problem. This paper (see Section 2) adopts the opposite point of view: rational agent's intentions does not have to be closed under causal consequence but have to be closed under logical consequence. This motivates our choice of using a normal modal operator to describe intentions of rational agents.

In Konolige and Pollack logic, intentions and beliefs are related by the weak-realism axiom. This leads to the consequence that, from the intention to do a, it is not possible to derive the belief that a will be actually carried out. This reduces the effectiveness of agent's intention as a mean to help it in coordinating its actions, as pointed out in Section 4.

Rao and Georgeff in [12] face the problem of studying autonomous agents embedded in a changing environment and, in particular, the so called agent-oriented systems. Such systems are viewed as rational agents having certain mental attitudes that determine their behavior. Rao and Georgeff are mainly interested in developing a logic to specify agent oriented systems and, in particular, in defining a methodology to verify that the design of such a system verifies its specification. Their logic adopts a possible world, branching-time structure to define the semantics of three modal operators that describe agent's beliefs, desires and intentions. Beliefs and intentions are related by the strong realism axiom but the implication of adopting other forms of realism is examined in [11]. Rao and Geogeff disregard to point out the axioms of their logic and focus on the problem of verifying that a particular model, derived from the agent's design, satisfies the agent's specification. This makes their logic ill-suited to prove agent's properties using the proof theoretic approach we are interested in. Moreover, Rao and Georgeff logic uses modal temporal operators (like next-time, until, every-time) as opposed to our explicit timed operators. Thus, using their logic, it is not possible to specify timed intentions and this partially reduces the expressiveness of the language.

In [16] Thomas presents a logic to represent agents' beliefs, their plans concerning future actions and goals, and their capabilities. Thomas work is similar to ours, but it focuses on agent capabilities and on the relationships between these capabilities and the agent's behavior. Thomas `PlanToDo` and `PlanToAch` operators are similar to the operator I presented in this paper, but are used to capture something slightly different. An agent plans to achieve ϕ if it not only has chosen to pursue ϕ, but also has constructed a plan for achieving ϕ and means to carry out that plan.

Thomas logic, like ours, is based on the strong realism axiom, but from the plan to do a it allows one to derive the belief that a will be successfully carried out and that its postconditions will be true. The logic presented in this paper, instead, from the agent's intention to do a at time t makes it possible only to derive the belief that the agent will be carrying out a at time t. Nothing is said about whether a will be successfully carried out or not. Formally, Thomas logic has the axiom: [1]

$$\mathtt{PlanToDo}^t(a) \Rightarrow B\,\mathtt{Did}^t(a)$$

[1]The semantics of Thomas operator `Did` is such that $\mathtt{Did}^t(a)$ is true in a possible world w if and only if the postconditions of action a are true in w.

while our logic has the axiom:

$$I_x\, a^t \Rightarrow B_x(Doing_x\, a)^t$$

Thomas axiom seems to be too strong. An agent can not be sure that it will successfully perform action a. It can only believe that he will be doing action a. This is the main reason why the weaker axiom we adopted better explain the reasoning of a rational agent.

Finally, Thomas logic uses explicit timed operators similar to ours, but does not define any operator to quantify over time (like the operators Alw, Alw_f, and Alw_p) and this reduces the logic expressiveness.

To conclude this comparison, it is important to point out that all related works cited in this section, with the exception of the first one, take into account only the single agent case, while this paper describes a language to model multiple agents.

6 Conclusion

This paper presents a logical language to model rational agents. The expected properties of such a language has been pointed out, and a logic meeting these properties has been developed. The present work focuses on the static relationships between agents' intentions and beliefs, leaving dynamic reasoning out.

The need to distinguish between logical implication and causal connection, in order to better formalize the side-effects problem, has been highlighted. The logic presented in this paper does not try to introduce a new operator to model causality but uses the standard operator '$\Rightarrow$' to approximate it.

Different relationships between the intention to do a and the belief about a's achievement have been discussed. The strong realism approach has been finally adopted because it ensures that intentions can play the expected role of coordinating agent's activities.

Future work will deeply examine the effectiveness of using a logical operator to formalize better the causal consequences of agent's actions. Moreover, it will extend the language in order to formalize the dynamic relationships between different mental attitudes.

Acknowledgments

I would like to thank Prof. Marco Colombetti who was my advisor for the work presented in this paper. He was always able to find time to discuss with me even if his agenda was really full. I would also thank Elisabetta Di Nitto who reviewed the last version of the paper.

References

[1] J. Allen. Two views of intention: Comments on bratman and on cohen and levesque. In P. R. Cohen, J. M. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 71–75. MIT Press, Cambridge, MA, 1990.

[2] M. E. Bratman. *Intention, Plans and Practical Reason.* Harvard University Press, Cambridge, MA, 1987.

[3] M. E. Bratman. What is intention? In P. R. Cohen, J. M. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 15–31. MIT Press, Cambridge, MA, 1990.

[4] B. F. Chellas. *Modal Logic: An Introduction.* Cambridge University Press, 1980.

[5] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

[6] J. Y. Halpern and Y. O. Moses. A guide to the modal logics of knowledge and belief. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, Los Angeles, CA, 1985.

[7] J. Hintikka. *Knowledge and Belief.* Cornell University Press, 1962.

[8] K. Konolidge and M. E. Pollack. A representationalist theory of intention. In Bajcsy R.(ed.), editor, *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pages 390–395, San Mateo, CA, 1993. Morgan Kaufmann Publishers.

[9] J. L. Mackey. *The Cement of the Universe:a Study of Causation.* Oxford University Press, 1974.

[10] R. C. Moore. *Reasoning about Knowledge and Action.* PhD thesis, Massachusset Institute of Technology, Cambridge, MA, 1990.

[11] A. S. Rao and M. P. Georgeff. Asymmetry thesis and side-effect problems in linear time and branching time intention logics. In J Mylopoulos and R. Reiter, editors, *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, Sydney, Australia, 1991. Morgan Kaufmann Publishers.

[12] A. S. Rao and M. P. Georgeff. A model-theoretic approach to the verification of situated reasoning systems. In Bajcsy R.(ed.), editor, *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, San Mateo, CA, 1993. Morgan Kaufmann Publishers.

[13] J. R. Searle. *Intentionality, an essay in the philosophy of mind.* Cambridge University Press, 1983.

[14] Y. Shoham. *Reasoning About Change, Time and Causation from the STandpoint of Artificial Intelligence.* MIT Press, 1988.

[15] E. Sosa. *Causation and Conditionals.* Oxford University Press, 1975.

[16] B. Thomas. A logic for representing actions, beliefs, capabilities, and plans. In Bajcsy R.(ed.), editor, *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, San Mateo, CA, 1993. Morgan Kaufmann Publishers.