A Logical Approach to the Dynamics of Commitments

J.-J. Ch. Meyer W. van der Hoek B. van Linder*
Utrecht University
Department of Computer Science
P.O. Box 80.089
3508 TB Utrecht
The Netherlands
Email: jj@cs.uu.nl

Abstract

In this paper we present a formalisation of motivational attitudes, the attitudes that are the driving forces behind the actions of agents. We consider the statics of these attitudes both at the assertion level, i.e. ranging over propositions, and at the practition[1] level, i.e. ranging over actions, as well as the dynamics of these attitudes, i.e. how they change over time. Starting from an agent's wishes, which form the primitive, most fundamental motivational attitude, we define its goals as induced by those wishes that do not yet hold, i.e. are unfulfilled, but are within the agent's practical possibility to bring about, i.e. are implementable for the agent. Among these unfulfilled, implementable wishes the agent selects those that qualify as its goals. Based on its knowledge on its goals and practical possibilities, an agent may make certain commitments. In particular, an agent may commit itself to actions that it knows to be correct and feasible to bring about some of its known goals. As soon as it no longer knows its commitments to be useful, i.e. leading to fulfilment of some goal, and practically possible, an agent is able to undo these commitments. Both the act of committing as well as that of undoing commitments is modelled as a special model-transforming action in our framework, which extends the usual state-transition paradigm of Propositional Dynamic Logic. In between making and undoing commitments, an agent is committed to all the actions that are known to be identical for all practical purposes to the ones in its agenda. By modifying the agent's agenda during the execution of actions in a straightforward way, it is ensured that commitments display an intuitively acceptable behaviour with regard to composite actions.

1. Introduction

The formalisation of rational agents is a topic of continuing interest in Artificial Intelligence. Research on this subject has held the limelight ever since the pioneering work of Moore [40] and Morgenstern [41, 42] in which knowledge and actions are considered. Over the years important contributions have been made on both *informational* attitudes like knowledge and belief [20], and *motivational* attitudes like commitments and obligations [9]. Recent

*Currently at ABN AMRO Bank N.V.

[1] The term 'practition' is due to Castañeda [6].

developments include the Belief-Desire-Intention architecture [47], logics for the specification and verification of multi-agent systems [62], and cognitive robotics [28].

In a series of papers [37, 24, 32, 25, 34, 33, 35] we defined a *theorist* logic for rational agents, i.e. a formal system that may be used to *specify*, *analyse* and *reason about* the behaviour of rational agents. In the basic framework [37, 24], the *knowledge*, *belief* and *abilities* of agents, as well as the *opportunities* for and the *results* of their actions are formalised. In this so-called KARO framework it can for instance be modelled that an agent knows that some action is *correct* to bring about some state of affairs since it knows that performing the action will lead to the state of affairs, and that it knows that an action is *feasible* in the sense that the agent knows of its ability to perform the action.

Having dealt with both informational attitudes and various aspects of action in previous work, this paper is aimed at providing a formalisation of the *motivational* attitudes of rational agents. In the last decade various formalisations of different kinds of motivational attitudes have been proposed [9, 46, 55]. The approach presented in this paper makes three main contributions to the theory of formalising motivational attitudes. Firstly, we consider a fairly wide scope of motivational attitudes, situated at two different levels. At the *assertion* level, this is the level where operators deal with assertions, we consider *wishes* and *goals*. At the *practition* level, where operators range over actions, we define *commitments*. With respect to these commitments we introduce both an operator modelling the commitments that an agent has made, and an action which models the act of committing. The notions that we formalise avoid (most of) the well-known problems that plague formalisations of motivational attitudes. Secondly, our formalisation of the various notions is strictly *bottom up*. That is, after defining the fundamental notion of wishes, goals are defined in terms of wishes, and commitments are introduced using the notion of goals. In this way, we provide a formalisation of motivational attitudes that does not have to resort to 'tricks' like (circularly) defining the intention to do an action in terms of the goal to have done it. Lastly, in our formalisation we will also try to connect to some relevant insights on motivational attitudes as they have been gained in the philosophical research on practical reasoning, in particular some ideas of the philosopher Von Wright [63].

We end this introduction with a disclaimer. In recent years it has become apparent that in order to give a full treatment of intelligent agents (and especially their motivational attitudes), one needs – besides a logical framework – elements from decision theory and/or game theory, where utilities and preferences of goals, plans and courses of action are studied. (The seminal work on this area is Von Neumann & Morgenstern [60]. References to application to AI (and agent theory in particular) can be found in e.g. [3, 61, 13].) Although we appreciate this fully[2], we do not include these elements in our present account, but indicate the places in our approach where they may play a role. Thus we present a logical framework to describe motivational attitudes of agents, leaving 'extra-logical' issues aside.[3] For instance, for selecting a wish (as a candidate goal) we will just use a $\texttt{select}\,\varphi$ operator that has as a result that φ has been selected. We do not give criteria on the basis of which a rational agent will make this selection. Here, of course, decision-theoretic approaches using utilities and preferences,

[2]However,the role of decision theory in specifying the behaviour of intelligent agents should also not be overestimated: as stated in e.g. [5, 18], it might be that sometimes the system in which the agent has to operate is so dynamic and subject to change that elaborate decision-theoretic deliberation might be outdated when it is finished and ready to be used.

[3]We note that for instance Boutilier [3] attempts to give a logical account of this type of reasoning so that his theory may be a good candidate for integrating these decision-theoretic aspects into our theory.

may come in. We maintain that if one would wish to incorporate these approaches in our framework, one might do so by refining the model.

The rest of the paper is organised as follows. To sketch the context and the area of application of this research, we start in Section 2 with the (re)introduction of some of our ideas on knowledge, belief, abilities, opportunities, and results; furthermore the definition of our formal language is given. In Section 3 we discuss the motivational attitudes that we will treat in this paper, and extend the basic language with extra operators to express these attitudes. Next we turn to the semantics of our language: in Section 4 we present the models of our basic framework, in Section 5 we consider the semantics of the operator for expressing wishes and in Section 6 and Section 7 the notions of a goal and a commitment, respectively, are treated formally. We will be able to formalise goals in such a way that problems of 'logical omniscience' that plague most formalisations of motivational attitudes are avoided. Finally, in Section 8 we summarise and conclude. Selected proofs are provided in an appendix.

2. Knowledge, abilities, opportunities, and results

The main informational attitude that we consider is that of *knowledge*. In representing knowledge we follow the approach common in epistemic logic [20, 23]: the formula $\mathbf{K}_i\varphi$ denotes the fact that agent i knows φ, and is interpreted in a Kripke-style possible worlds semantics. In our approach it is left open whether the knowledge of an agent results from the agent's explicitly having an encoding of this knowledge at its disposal (like e.g. in Konolige's deduction model of belief [27]), or that knowledge is ascribed implicitly (or even metaphorically) to the agent in terms of a relationship between the agent and its environment (as in Rosenschein's situated automata [50, 51], in which an agent knows φ in a situation where its internal state is σ if φ holds in all possible situations in which the agent is in internal state σ).

At the action level we consider *results*, *abilities* and *opportunities*. Slightly simplifying ideas of Von Wright [63], we consider any aspect of the state of affairs brought about by the execution of an action by an agent in some state to be among the results of the event consisting of the execution of that particular action by that particular agent, in that particular state. An important aspect of any investigation of action is the relation that exists between ability and opportunity. In order to successfully complete an action, both the opportunity and the ability to perform the action are necessary. Although these notions are interconnected, they are surely not identical: the abilities of agents comprise mental and physical powers, moral capacities, and physical possibility, whereas the opportunity to perform actions is best described by the notion of circumstantial possibility (cf. [26]). The abilities of agents are formalised via the $\mathbf{A}_i$ operator; the formula $\mathbf{A}_i\alpha$ denotes the fact that agent i has the ability to do α. When using the descriptions of opportunities and results as given above, the framework of (propositional) dynamic logic provides an excellent means to formalise these notions. Using events $\mathrm{do}_i(\alpha)$ to refer to the performance of the action α by the agent i, we consider the formulae $\langle \mathrm{do}_i(\alpha)\rangle\varphi$ and $[\mathrm{do}_i(\alpha)]\varphi$. As we shall only encounter deterministic actions in this paper, $\langle \mathrm{do}_i(\alpha)\rangle\varphi$ is the stronger of these formulae; it represents the fact that agent i has the opportunity to do α and that doing α leads to φ. The formula $[\mathrm{do}_i(\alpha)]\varphi$ is noncommittal about the opportunity of the agent to do α but states that if the opportunity to do α is indeed present, doing α results in φ.

2.1. Definition. Let denumerable sets $\mathrm{A} = \{1, \ldots, n\}$ of agents, Π of propositional symbols and At of atomic actions be given. The language L is the smallest superset of Π such that:

- if $\varphi, \psi \in \mathrm{L}, i \in \mathrm{A}, \alpha \in \mathrm{Ac}$ then $\neg\varphi, \varphi \vee \psi, \mathbf{K}_i\varphi, \langle \mathrm{do}_i(\alpha)\rangle\varphi, \mathbf{A}_i\alpha \in \mathrm{L}$

where Ac is the smallest superset of At such that if $\varphi \in \mathrm{L}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}$ then

- $\varphi? \in \mathrm{Ac}$ *tests*
- $\alpha_1; \alpha_2 \in \mathrm{Ac}$ *sequential composition*
- $\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi} \in \mathrm{Ac}$ *conditional composition*
- $\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od} \in \mathrm{Ac}$ *repetitive composition*

The intuitive interpretation of the $\varphi?$ action is the verification whether φ holds: if it does, execution can be continued with the next action; if it doesn't, execution fails (remains pending without yielding a next state). Actions that are either atomic or a test, are called *semi-atomic*. The class of semi-atomic actions is denoted by At^+. The sequential composition $\alpha_1; \alpha_2$ is interpreted as α_1 followed by α_2. The conditional composition $\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}$ means the execution of α_1 if φ holds and that of α_2 otherwise. The repetitive composition $\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$ is interpreted as executing α as long as φ holds. The formal semantics of these actions will be treated in the next section. The purely propositional fragment of L is denoted by L_0. Constructs $\wedge, \rightarrow, \leftrightarrow, \top$ and $\bot$ are defined in the usual way, and $[\mathrm{do}_i(\alpha)]\varphi$ is used as an abbreviation of $\neg\langle \mathrm{do}_i(\alpha)\rangle\neg\varphi$. Finally, we define the class Acseq as the smallest superset of At^+ closed under sequential composition. Thus Acseq consists of sequences of semi-atomic actions.

2.1. The Can-predicate and the Cannot-predicate

To formalise the knowledge of agents on their practical (im)possibilities, we introduced the so-called Can-predicate and Cannot-predicate. These are binary predicates, pertaining to a pair consisting of an action and a proposition, and denoting that an agent knows that performing the action constitutes a practical (im)possibility to bring about the proposition. We consider practical possibility to consist of two parts, viz. correctness and feasibility: action α is *correct* with respect to φ iff $\langle \mathrm{do}_i(\alpha)\rangle\varphi$ holds and α is *feasible* iff $\mathbf{A}_i\alpha$ holds.

2.2. Definition. The Can-predicate and the Cannot-predicate are, for all agents i, actions α and formulae φ, defined as follows.

- $\mathbf{PracPoss}_i(\alpha, \varphi) =^{\mathrm{def}} \langle \mathrm{do}_i(\alpha)\rangle\varphi \wedge \mathbf{A}_i\alpha$
- $\mathbf{Can}_i(\alpha, \varphi) =^{\mathrm{def}} \mathbf{K}_i\mathbf{PracPoss}_i(\alpha, \varphi)$
- $\mathbf{Cannot}_i(\alpha, \varphi) =^{\mathrm{def}} \mathbf{K}_i\neg\mathbf{PracPoss}_i(\alpha, \varphi)$

Thus the Can-predicate and the Cannot-predicate express the agent's knowledge about its practical possibilities and impossibilities, respectively. Therefore these predicates are important for the agent's planning of actions.[4]

3. Motivational attitudes: wishes, goals and commitments

Motivational attitudes constitute what probably is the most fundamental, primitive and essential characteristic of agency. These attitudes provide the motive for any act on behalf of

[4] This Can-predicate is very closely related to similar notions in [40] and [42]. Both these notions also employ the notion of a (physical) precondition, which in our framework is expressed by means of the result operator in $\langle \mathrm{do}_i(\alpha)\rangle\varphi$, while we have generalised the ability / know-how aspect in these notions by means of the ability operator $\mathbf{A}_i$.

the agents, i.e. the acting of agents is driven by their motivational attitudes. Typical examples of motivational attitudes are amongst others wishes, desires, concerns, ambitions, goals, intentions and commitments. The meaning of most of these terms is intuitively much less clear than that of the informational attitudes of knowledge and belief, or of the aspects of action (result, opportunity, ability) that we considered in the previous section. It is therefore also not clear which of the aforementioned motivational attitudes are relevant, and worth formalising, when modelling rational agents. In their BDI-architecture, Rao & Georgeff [47] consider desires and intentions to be primitive, and define a notion of commitment in terms of these, Cohen & Levesque [9] consider goals to be primitive and define intentions using goals, and Shoham [54] restricts himself to formalising commitments.

Psychologically, it is rather doubtful whether notions such as goals, intentions and commitments are indeed primitive. Contemporary psychology maintains that motivational processes are stemming from (the interaction between) internal drives, such as hunger, and external incentives, such as the availability of food, and are experienced by humans subjectively as *conscious desires* (see e.g. [2]). Thus, abstracting from the way how these come about, wishes or desires seem to be a fundamental notion in explaining human motivated behaviour. Of course, one may question the adequacy of psychological insights for non-human, artificial agents. It is very dubious indeed whether a robot (and, *a fortiori* a softbot) can be ascribed human motivational attitudes. However, since it is to be expected that agent systems will consist of both artificial and human agents, and agent theories must therefore be applicable to both agents, human and artificial, it seems reasonable to stay as close in line with psychology as possible. In our approach we will therefore take *wishes* (or *desires*) as primitive, and define *goals* by means of these, where we also make use of our KARO framework to express the practical possibility of actions. Commitments are then defined as the result of certain commit actions applied to actions that are known to fulfil goals (and that *can* indeed be performed).

We claim that by proceeding in the above manner, in addition to being more in line with ideas from contemporary psychology, we obtain a theory that is also close to insights gained in analytical philosophy, especially those pertaining to the modelling of *practical reasoning*. The term 'practical reasoning' dates back to Aristotle, and refers to the process through which (human) agents conclude that they should perform certain actions in order to bring about some of the things that they like to be the case. It seems very likely that for autonomous agents in AI applications, which have to act (autonomously) to achieve some of their goals, practical reasoning accounts for the most essential and most frequently used kind of information processing. Hence an adequate formalisation of motivational attitudes should pay at least some attention to this kind of reasoning.

A further important feature of our framework is that within it – being founded on a theory of action – it is easy to incorporate the *act* of *selecting*: agents have to make choices among the things they desire to be the case, thereby deciding which of these they will try to achieve next. For it might be impossible to satisfy all of an agent's wishes simultaneously, since these wishes become easily either analytically inconsistent or incompatible given the agent's resources. More generally, we believe that basing motivational attitudes of agents on a theory of action (like the KARO framework) provides a framework in which dependencies between motivational notions can be expressed naturally. We also claim that the use of an action-based theory yields direct connections with 'programs', so that we can on the one hand profit from programming / computing theory while developing our framework, and on the other obtain a theory that is closer to programming practice. Interestingly, even if one does not share our views of the desirability of keeping our theory as close to psychology and

philosophy as possible, the approach thus leads to a very pragmatic point of view. We will return to this issue later.

Thus the notions that will play an essential role in our formalisation are *wishes*, *goals* and *commitments*. Of these, wishes constitute the primitive motivational attitude that models the things that an agent likes to be the case. As such, wishes naturally range over propositions, corresponding to the idea that agents wish for certain aspects of the world. We formalise wishes through a normal modal operator, i.e. an operator validating just the so-called K-axiom and the necessitation rule. Agents set their goals by selecting among their wishes. However, agents are not allowed to select arbitrary wishes as their goals, but instead may only select wishes that are unfulfilled yet implementable. Whenever an agent knows that it has some goal, it may commit itself to any action that it knows to be correct and feasible with respect to the goal. This act of committing to an action is itself formalised as a special kind of action. Commitments to actions are in general to persist until all of the goals for which the commitment was made are fulfilled. Having said so, agents should not be forced to remain committed to actions that have either become useless in that they do not lead to fulfilment of any goal any more, or impossible in that the agent no longer knows that it has the opportunity and ability to perform the action. Phrased differently, an agent should be allowed to uncommit itself whenever an action is no longer known to be correct and feasible with respect to one of the agent's goals.

To formalise wishes, goals and commitments and their associated concepts, we introduce a modal operator modelling wishes, operators modelling implementability, (made) choices and (made) commitments, and action constructors modelling the acts of selecting, committing and uncommitting.

3.1. Definition. To define the language L^{C}, the alphabet is extended with the wish operator $\mathbf{W}\_\_$, the implementability operator $\Diamond\_\_$, the selected operator $\mathbf{C}\_\_$, the commitment operator $\mathbf{Committed}\_\_$ and the action constructors $\mathtt{select}\_$, $\mathtt{commit\_to}\_$ and $\mathtt{uncommit}\_$.

The acts of committing and uncommitting are of an essentially different nature than the regular actions, execution of which changes the state of the world. Through the former actions agents (un)commit themselves to actions of the latter kind. Intuitively it does not make much sense to allow agents to commit themselves to making commitments: it is not at all clear how a statement like 'i is committed to commit itself to do α' is to be interpreted. Also statements like 'it is implementable for agent i to become committed' seem to be of a rather questionable nature. To avoid these kinds of counterintuitive situations, we define the language L^{C} on top of the language L that we defined previously. That is, the operators modelling wishes, implementability and selections are defined in such a way that they range over formulae from L rather than those from L^{C}. The operator modelling the commitments that an agent has made is defined to range over the actions from Ac, the class of actions associated with L, and not over Ac^{C} (to be defined below). Analogously, the special actions in Ac^{C}, as there are the action modelling the act of selecting and those modelling the making and undoing of commitments, take as arguments elements from L and Ac rather than L^{C} and Ac^{C}.

3.2. Definition. The language L and the class Ac of actions are as in Definition 2.1, i.e. L is the smallest superset of Π closed under the core clauses of Definition 2.1 and Ac is the smallest superset of At satisfying the core clauses of Definition 2.1 again.

The language L^{C} is the smallest superset of Π such that the core clauses of Definition 2.1 are validated and furthermore

- if $\varphi \in \mathrm{L}$ and $i \in \mathrm{A}$ then $\mathbf{W}_i\varphi \in \mathrm{L}^{\mathrm{C}}$
- if $\varphi \in \mathrm{L}$ and $i \in \mathrm{A}$ then $\Diamond_i\varphi \in \mathrm{L}^{\mathrm{C}}$
- if $\varphi \in \mathrm{L}$ and $i \in \mathrm{A}$ then $\mathbf{C}_i\varphi \in \mathrm{L}^{\mathrm{C}}$
- if $\alpha \in \mathrm{Ac}$ and $i \in \mathrm{A}$ then $\mathbf{Committed}_i\alpha \in \mathrm{L}^{\mathrm{C}}$

The class Ac^{C} is the smallest superset of At closed under the core clauses of Definition 2.1 and such that

- if $\varphi \in \mathrm{L}$ then $\mathtt{select}\,\varphi \in \mathrm{Ac}^{\mathrm{C}}$
- if $\alpha \in \mathrm{Ac}$ then $\mathtt{commit\_to}\,\alpha \in \mathrm{Ac}^{\mathrm{C}}$
- if $\alpha \in \mathrm{Ac}$ then $\mathtt{uncommit}\,\alpha \in \mathrm{Ac}^{\mathrm{C}}$

3.3. Definition. For $i \in \mathrm{A}$, $\alpha \in \mathrm{Ac}^{\mathrm{C}}$ and $\varphi \in \mathrm{L}^{\mathrm{C}}$, the abbreviations $\mathbf{PracPoss}_i(\alpha, \varphi)$ and $\mathbf{Can}_i(\alpha, \varphi)$ are defined as in the language L.

4. Semantics

As is usual for a modal language we use Kripke-style semantics. Essentially, our models consist of a set of possible worlds, a truth assignment function for the propositional symbols, and accessibility relations for the modal operators that we consider. For the core language defined in Definition 2.1, this means that we need accessibility relations for the knowledge operator and the operator $\langle \mathrm{do}_i(\alpha)\rangle$. For technical convenience, we will use for the latter a function rather than a relation. Finally, we need a function to indicate the agent's capabilities at each possible world.

4.1. Definition. The class $\mathbf{M}$ of models contains all $\mathrm{M} = \langle \mathrm{S}, \pi, \mathrm{R}, \mathtt{r}_0, \mathtt{c}_0\rangle$ where

- S is a set of possible worlds, or states.
- $\pi : \Pi \times \mathrm{S} \to \{\mathbf{0}, \mathbf{1}\}$ assigns a truth value to propositional symbols in states.
- $\mathrm{R} : \mathrm{A} \to \wp(\mathrm{S} \times \mathrm{S})$ is a function that yields the epistemic accessibility relations for a given agent. It is demanded that $\mathrm{R}(i)$ is an equivalence relation for all i. We define $[s]_{\mathrm{R}(i)}$ to be $\{s' \in \mathrm{S} \mid (s, s') \in \mathrm{R}(i)\}$, the $\mathrm{R}(i)$-equivalence class of s.
- $\mathtt{r}_0 : \mathrm{A} \times \mathrm{At} \to (\mathrm{S} \cup \{\emptyset\}) \to (\mathrm{S} \cup \{\emptyset\})$ is such that $\mathtt{r}_0(i, a)(s)$ yields the (possibly empty) state transition in s caused by the event $\mathrm{do}_i(a)$. For technical reasons we stipulate that always $s \notin \mathtt{r}_0(i, a)(s)$. This means that a successful performance of an atomic action always results in a state transition to *another* state in the model. Of course, it may happen that this state satisfies the same formulas as the original one.[5]
- $\mathtt{c}_0 : \mathrm{A} \times \mathrm{At} \to (\mathrm{S} \cup \{\emptyset\}) \to \{\mathbf{0}, \mathbf{1}\}$ is the capability function such that $\mathtt{c}_0(i, a)(s)$ indicates whether the agent i is capable of performing the action a in s.

The notion of a *state* in our sense refers to the state in which the (objective) world can be, i.e. (via the function π) a complete description of the world in terms of truth and falsity of the propositional atoms. Thus it refers to a state of the *world* rather than that of an *agent*.[6]

[5] This might be understood in the following sense: even an action that has no effect on the truth of the formulas such as a 'skip' action, still results in a state where the *history* of performed actions is changed. This idea is very similar to the notion of a (Herbrand) situation in the situation calculus ([49, 31, 53, 28]), and can be made formal by considering propositions (or variables) recording the history. In order to not complicate our models further we will mostly leave this implicit in this paper, apart from some places where it is crucial.

[6] In a 'situated automata' approach to knowledge as mentioned in Section 2, the state may be thought of to include the 'internal' or 'local' state of an agent, which is determined completely by the propositional atoms 'local to the agent'. In this case states (here often referred to as global states) are typically represented as vectors of 'local' states of the agents in the system at hand. See e.g. [15, 38].

It should not be confused with the epistemic, or more generally, mental state of an agent. The latter can be viewed as being represented in our framework by the relation R and the function c_0 in the model (w.r.t. the epistemic and capability aspects, respectively).

The models for the language L^{C} are further equipped with elements used to interpret the agents' wishes, selections and commitments. Wishes are interpreted through an accessibility relation on worlds that denotes worlds that are (more) desirable from the agent's point of view. Selections are straightforwardly interpreted through a set of formulae that denotes the choices that an agent has made. From a formal point of view, this set acts as a kind of *awareness* on the wishes of an agent, thereby ensuring an intuitively acceptable behaviour of goals. Originally, Fagin & Halpern [14] introduced the idea of awareness sets as a means to solve the so-called problems of logical omniscience. As we will see in Section 6, the effect of the selection sets on the behaviour of goals is similar to that of the awareness sets on the properties of knowledge. The agents' commitments are interpreted by means of an agenda function, which yields for each agent in every state the commitments that it has made and is up to. Detailed accounts of the respective interpretations are given in the following sections.

4.2. Definition. A model M for the language L^{C} is a tuple containing the core elements of Definition 4.1, the functions $\mathrm{W} : \mathrm{A} \to \wp(\mathrm{S} \times \mathrm{S})$, which determines the desirability relation of an agent in a state, and $\mathrm{C} : \mathrm{A} \times \mathrm{S} \to \wp(\mathrm{L})$ denoting the choices made by an agent in a state, and a function $\mathrm{Agenda} : \mathrm{A} \times \mathrm{S} \to \wp(\mathrm{Ac})$, which records the commitments of agents.

Note that the agenda records *syntactical* entities (as opposed to our earlier preliminary account in [36], where semantical entities, viz. computation sequences, are recorded). In order to get strong results about the agent's knowledge of commitments (Prop. 7.17, in particular part 2) we need a syntactical representation, which may yield different (semantical) computation sequences in different possible worlds (e.g. when if-statements are recorded in the agenda). We will return to this issue in Section 7.

Note furthermore that in general it is allowed that the agenda contains multiple actions. This means that the agent has multiple commitments, and should be in need of some strategy to select and perform these. In this paper we will not pursue this. On the contrary, to keep matters simple we will stipulate in Definition 7.9 below that an agent will only be able to commit to a new action if its agenda is empty. We interpret the acts of selecting, committing and uncommitting as model-transformations. The act of selecting changes a model by affecting the set of choices, and the act of (un)committing transforms the agent's agenda. To account for these modifications, we introduce the set of possible result models of a given model for L^{C}.

4.3. Definition. Let $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ be some model for L^{C}. The class $\mathrm{M}^{\mathrm{C}}_{\sim} \subseteq \mathbf{M}^{\mathrm{C}}$ contains all models that (possibly) differ from M only in the C or the Agenda functions.

In our interpretation of actions as given in Definition 4.4 below, we generalise the standard paradigm of actions as state-transitions [21] by interpreting actions as transitions between pairs (Model, State) rather than transitions between states *per se*. Using this more general interpretation we can both account for regular actions that cause a transition between states upon execution, and special actions that transform models. Among the special actions that we considered elsewhere [35] were those modelling informational attitudes such as observations and communication; in Section 7 we will formalise the acts of committing and uncommitting in a similar way.

Furthermore, to keep in line with the idea expressed in Definition 4.1 that the performance of actions, when successful, always leads to different states, we sometimes need to provide 'copies' of states. More formally, this is done by assuming in the model for every state s a special variable h such that $s(h)$ yields the list of semi-atomic actions performed successively so far (the 'history').[7] The variable h doesn't appear in the language L_0. For a state s, and a semi-atomic action a, we denote by s_a the state that is as s (w.r.t. all properties regarding $\pi, \mathtt{r}_0, \mathtt{c}_0$ and all accessibility relations such as R and W), but in which $s_a(h)$ is equal to $s(h)$ appended with a. Since h doesn't occur in L_0, we have that for all formulae $\varphi \in \mathrm{L}_0$, $\mathrm{M}, s \models \varphi \Leftrightarrow \mathrm{M}, s_a \models \varphi$.

4.4. DEFINITION. The binary relation $\models$ between a formula from L and a pair M, s consisting of a model $\mathrm{M} \in \mathbf{M}$ and a state s in M, is for φ a propositional symbol, a negation, or a disjunction inductively defined as usual. For the other cases $\mathrm{M}, s \models \varphi$ is defined by:

$$
\begin{array}{ll}
\mathrm{M}, s \models \mathbf{K}_i\varphi & \Leftrightarrow \forall s' \in \mathrm{S}((s,s') \in \mathrm{R}(i) \Rightarrow \mathrm{M}, s' \models \varphi) \\
\mathrm{M}, s \models \langle \mathrm{do}_i(\alpha)\rangle\varphi & \Leftrightarrow \exists \mathrm{M}', s'(\mathrm{M}', s' \in \mathtt{r}(i,\alpha)(\mathrm{M}, s)\,\&\,\mathrm{M}', s' \models \varphi) \\
\mathrm{M}, s \models \mathbf{A}_i\alpha & \Leftrightarrow \mathtt{c}(i,\alpha)(\mathrm{M}, s) = \mathbf{1}
\end{array}
$$

where $\mathtt{r}$ and $\mathtt{c}$ are defined as follows:

$$
\begin{array}{ll}
\mathtt{r}(i,a)(\mathrm{M}, s) & = \mathrm{M}, \mathtt{r}_0(i,a)(s) \\
\mathtt{r}(i,\varphi?)(\mathrm{M}, s) & = \mathrm{M}, s_{\varphi?} \text{ if } \mathrm{M}, s \models \varphi \text{ and } \emptyset \text{ otherwise} \\
\mathtt{r}(i,\alpha_1;\alpha_2)(\mathrm{M}, s) & = \mathtt{r}(i,\alpha_2)(\mathtt{r}(i,\alpha_1)(\mathrm{M}, s)) \\
\mathtt{r}(i, \mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1 & = \mathtt{r}(i,\alpha_1)(\mathrm{M}, s_{\varphi?}) \text{ if } \mathrm{M}, s \models \varphi \text{ and} \\
\quad \mathtt{else}\ \alpha_2\ \mathtt{fi})(\mathrm{M}, s) & \quad \mathtt{r}(i,\alpha_2)(\mathrm{M}, s_{\neg\varphi?}) \text{ otherwise} \\
\mathtt{r}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(\mathrm{M}, s) & = \mathrm{M}', s' \text{ such that } \exists k \in \mathbb{N} \exists \mathrm{M}_0, s_0 \ldots \exists \mathrm{M}_k, s_k \\
 & \quad (\mathrm{M}_0, s_0 = \mathrm{M}, s\,\&\,\mathrm{M}_k, s_k = \mathrm{M}', s'\,\&\,\forall j < k \\
 & \quad (\mathrm{M}_{j+1}, s_{j+1} = \mathtt{r}(i,\varphi?;\alpha)(\mathrm{M}_j, s_j))\,\&\,\mathrm{M}', s' \models \neg\varphi) \\
\text{where } \mathtt{r}(i,\alpha)(\emptyset) & = \emptyset \\
\text{and} & \\
\mathtt{c}(i,a)(\mathrm{M}, s) & = \mathtt{c}_0(i,a)(s) \\
\mathtt{c}(i,\varphi?)(\mathrm{M}, s) & = \mathbf{1} \text{ if } \mathrm{M}, s \models \varphi \text{ and } \mathbf{0} \text{ otherwise} \\
\mathtt{c}(i,\alpha_1;\alpha_2)(\mathrm{M}, s) & = \mathtt{c}(i,\alpha_1)(\mathrm{M}, s)\,\&\,\mathtt{c}(i,\alpha_2)(\mathtt{r}(i,\alpha_1)(\mathrm{M}, s)) \\
\mathtt{c}(i, \mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1 & = \mathtt{c}(i,\varphi?;\alpha_1)(\mathrm{M}, s) \text{ or} \\
\quad \mathtt{else}\ \alpha_2\ \mathtt{fi})(\mathrm{M}, s) & \quad \mathtt{c}(i,\neg\varphi?;\alpha_2)(\mathrm{M}, s) \\
\mathtt{c}(i, \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})(\mathrm{M}, s) & = \mathbf{1} \text{ if } \mathtt{c}(i,(\varphi?;\alpha)^k;\neg\varphi?)(\mathrm{M}, s) = \mathbf{1} \\
 & \qquad\qquad \text{for some } k \in \mathbb{N} \text{ and } \mathbf{0} \text{ otherwise} \\
\text{where } \mathtt{c}(i,\alpha)(\emptyset) & = \mathbf{0}
\end{array}
$$

Validity in a model and in a class of models is defined as usual.

With regard to the abilities of agents, the motivation for the choices made in Definition 4.4 is the following. The definition of $\mathtt{c}(i,\varphi?)$ expresses that an agent is able to get confirmation for a formula φ iff φ holds. An agent is capable of performing a sequential composition $\alpha_1;\alpha_2$ iff it is capable of performing α_1 (now), and it is capable of executing α_2 after it has performed α_1. An agent is capable of performing a conditional composition, if either it is able to confirm the condition and thereafter perform the then-part, or it is able to confirm the negation of the condition and perform the else-part afterwards. An agent is capable of performing a

[7] Strictly speaking this goes slightly beyond the propositional set-up, where variables only get the values 'true' or 'false'.

repetitive composition $\texttt{while}\,\varphi\,\texttt{do}\,\alpha\,\texttt{od}$ iff it is able to perform the action $(\varphi?;\alpha_1)^k;\neg\varphi?$ for some natural number k, i.e. it is able to perform the kth unwinding of the while-loop. More about capabilities and their properties can be found in [37].

5. Formalising wishes

Wishes are here taken to be the most primitive motivational attitudes, i.e. *in ultimo* agents are motivated to fulfil their wishes. As mentioned in Section 3, we formalise wishes through a plain normal modal operator, i.e. wishes are straightforwardly interpreted as a necessity operator over the accessibility relation W.

5.1. Definition. The binary relation $\models^{\mathrm{C}}$ between a formula in L^{C} and a pair M, s consisting of a model M for L^{C} and a state s in M is for wishes defined as follows:

$\mathrm{M}, s \models^{\mathrm{C}} \mathbf{W}_i\varphi \Leftrightarrow \forall s' \in \mathrm{S}((s,s') \in \mathrm{W}(i) \Rightarrow \mathrm{M}, s' \models^{\mathrm{C}} \varphi)$

In order to allow for maximal flexibility regarding wishes (which may be quite irrational), we do not impose any restrictions on the accessibility relation W. In fact, one may even argue that representing wishes by a (normal) modal operator in itself already imposes too much rationality. It is well-known that normal modal operators have certain properties that are occasionally considered undesirable for the commonsense notions that they are intended to formalise. For example, although the formal notions of knowledge and belief are closed under logical consequence, this property will in general not hold for human knowledge and belief (although it will for instance hold for the information that is recorded in a database, and it can also be defended for the knowledge and belief of an artificial agent). When formalising motivational attitudes the undesired properties induced by closure under logical consequence become even more pregnant. For agents do in general not desire all the logical consequences of their wishes, nor do they consider the logically inevitable to be among their goals. For example, an agent that wants its teeth to be restored will in general not want or wish for the pain that inevitably accompanies such a restoration. And although the sun rises in the east there will hardly be an agent that desires this to be the case. The problem embodied by the former example is known as the *side-effect* problem; the problem that all logical tautologies are wishes (goals) of an agent is known as the *transference* problem. Both in syntactical shape as in meaning, these problems are closely related to the problems of logical omniscience that have plagued formalisations of informational attitudes for many years. In terms of our framework, seven of the most (in)famous problems of logical omniscience can be formulated as follows.

5.2. Definition. Let $\varphi, \psi \in \mathrm{L}$ be formulae, and let $\mathbf{X}$ be some operator.

- $\models \mathbf{X}\varphi \wedge \mathbf{X}(\varphi \rightarrow \psi) \rightarrow \mathbf{X}\psi$ LO1
- $\models \varphi \Rightarrow \models \mathbf{X}\varphi$ LO2
- $\models \varphi \rightarrow \psi \Rightarrow \models \mathbf{X}\varphi \rightarrow \mathbf{X}\psi$ LO3
- $\models \varphi \leftrightarrow \psi \Rightarrow \models \mathbf{X}\varphi \leftrightarrow \mathbf{X}\psi$ LO4
- $\models (\mathbf{X}\varphi \wedge \mathbf{X}\psi) \rightarrow \mathbf{X}(\varphi \wedge \psi)$ LO5
- $\models \mathbf{X}\varphi \rightarrow \mathbf{X}(\varphi \vee \psi)$ LO6
- $\models \neg(\mathbf{X}\varphi \wedge \mathbf{X}\neg\varphi)$ LO7

Properties LO1 and LO3 as given in Definition 5.2 capture the side-effect problem, and property LO2 captures the transference problem. Of the other properties given not all are equally problematic when formalising wishes. In our opinion, property LO4 is not that problematic, and could even be considered desirable, dependent on the demands for rationality that one is willing to make. Property LO5, which we like to think of as representing 'the problem of *unrestricted combining*', is in general undesirable when formalising motivational attitudes. This is for instance shown by the example of a (human) agent that likes both a sandwich with peanut butter and one with Camembert cheese, but not together. Property LO6, for which we coin the term 'the problem of *unrestricted weakening*', is a special instantiation of the side-effect problem. That this property is undesirable is shown by the example of an agent desiring itself to be painted green, without desiring being green or being crushed under a steam roller[8]. Property LO7 is unacceptable for certain kinds of motivational attitudes but a necessity for others. It is for instance perfectly possible for agents to have contradicting wishes[9], but it seems hardly rational to allow agents to try and fulfil these conflicting wishes simultaneously. Thus, whereas the absence of LO7 is essential when formalising wishes, the presence is when formalising goals.

It turns out that our formalisation of wishes validates all but one of the properties of logical omniscience.

5.3. Proposition. *All of the properties of logical omniscience formalised in Definition 5.2, with the exception of LO7, are valid for the* $\mathbf{W}_i$ *operator.*

Although we argued against the properties of logical omniscience when formalising motivational attitudes, we do not consider it a serious problem that our formalisation of wishes validates (almost all of) these properties. For these wishes are both *implicit* in the terminology of Levesque [29] and *passive* in the sense of Castelfranchi *et al.* [7]. Being implicit, it will not be the case that agents *explicitly* desire all of their wishes[10]. Being passive, wishes in themselves do not *actively* influence the course of action that an agent is going to take. Through the act of *selecting*, agents turn some of their implicit, passive wishes into explicit, active goals. Hence even though an agent implicitly and passively desires all logical consequences of one of its wishes, it will not do so explicitly and actively. Therefore Proposition 5.3 is not taken to represent a severe problem for a formalisation of (implicit and passive) wishes, whereas it would for a formalisation of (explicit and active) goals. In the following section it will be shown how the properties of logical omniscience are avoided for goals.

6. Setting goals

As remarked previously, an agent's goals are not primitive but induced by its wishes. Basically, an agent selects among its (implicit and passive) wishes those that it (explicitly and actively) aims to fulfil. Given the rationality of agents, these selected wishes should be both

[8] The problem of unrestricted weakening is intuitively related to Ross's paradox [52], well-known in deontic logic[1, 39]. The standard counterexample towards the desirability of LO6 in a deontic context, where the operator $\mathbf{X}$ is interpreted as 'being obliged to', is that of an agent that is obliged to mail a letter while (intuitively) not being obliged to either mail the letter or burn it.

[9] Even stronger, human agents will almost always suffer from conflicts between their wishes.

[10] For the implicit belief that, in combination with awareness, constitutes explicit belief in the approach of Fagin & Halpern [14], it is also considered unproblematic that the properties of logical omniscience are validated.

unfulfilled and implementable: it does not make sense for an agent to try and fulfil a wish that either already has been fulfilled or for which fulfilment is not a practical possibility. We do not take the latter constraint too stringently, i.e. we only demand wishes to be individually implementable without requiring a simultaneous implementability of all chosen wishes. However, if desired, constraints like simultaneous implementability are easily formulated. The act of selecting is treated as a fully-fledged action by defining the opportunity, ability and result of selecting. Informally, an agent has the *opportunity* to select any of its wishes, corresponding to the idea that choices are only restricted by the elements among which is to be chosen. However, an agent is *capable* of selecting only those formulae that are unfulfilled and implementable, which can be thought of as it having a built-in aversion against selecting fulfilled or practically impossible formulae. The *result* of a selection will consist of the selected formula being marked chosen.

As stated in the introduction the selection mechanism as proposed here is perhaps too liberal to be realistic for a rational agent. One might consider replacing the selection condition of 'unfulfilled and implementable' by something stronger based on decision-theoretic utilities and preferences.[11] Of course, the model should then be enriched accordingly to accommodate for this. Here we have chosen to abstract from this and concentrate on the overall logical framework.

The notion of unfulfilledness is straightforwardly formalised as 'not holding', i.e. a formula φ is unfulfilled in a state s of some model M if and only if $\mathrm{M}, s \not\models^{\mathrm{C}} \varphi$. Defining implementability is a little more elaborate. Roughly speaking, we define a formula φ to be implementable for an agent i, denoted by $\Diamond_i\varphi$, if i has the practical possibility to fulfil φ by performing an appropriate sequence of atomic actions[12].

6.1. Definition. The binary relation $\models^{\mathrm{C}}$ between a formula in L^{C} and a pair M, s consisting of a model M for L^{C} and a state s in M is for implementability formulae defined by:

$$\mathrm{M}, s \models^{\mathrm{C}} \Diamond_i\varphi \Leftrightarrow \exists k \in \mathbb{N} \exists a_1, \ldots, a_k \in \mathrm{At}(\mathrm{M}, s \models^{\mathrm{C}} \mathbf{PracPoss}_i(a_1; \ldots; a_k, \varphi))$$

Having defined unfulfilledness and implementability, we can now formally introduce the `select` action.

6.2. Definition. For $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ with state s, $i \in \mathrm{A}$ and $\varphi \in \mathrm{L}$ we define:

$$\mathtt{r}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s) = \begin{cases} \emptyset & \text{if } \mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{W}_i\varphi \\ \mathtt{choose}(i, \varphi)(\mathrm{M}, s), s & \text{if } \mathrm{M}, s \models^{\mathrm{C}} \mathbf{W}_i\varphi \end{cases}$$

where for $\mathrm{M} = \langle \mathrm{S}, \pi, \mathrm{R}, \mathtt{r}_0, \mathtt{c}_0, \mathrm{W}, \mathrm{C}, \mathrm{Agenda} \rangle$ we define

$$\begin{aligned} &\mathtt{choose}(i, \varphi)(\mathrm{M}, s) = \langle \mathrm{S}, \pi, \mathrm{R}, \mathtt{r}_0, \mathtt{c}_0, \mathrm{W}, \mathrm{C}', \mathrm{Agenda} \rangle \text{ with} \\ &\quad \mathrm{C}'(i', s') = \mathrm{C}(i', s') \text{ if } i \neq i' \text{ or } s \neq s' \\ &\quad \mathrm{C}'(i, s) \;\; = \mathrm{C}(i, s) \cup \{\varphi\} \end{aligned}$$

[11] A very interesting semantic approach along these lines is proposed by [61, 13].

[12] As was pointed out by Maarten de Rijke, defining the implementability operator in this way makes it a kind of 'dual master modality' (cf. [17, 56]). A formula consisting of a formula φ prefixed by the master modality is true in some state s of a model iff φ holds at all states that are reachable by any finite sequence of transitions from s. Such a formula is false iff there is some state s', reachable by some finite sequence of transitions from s, at which φ does not hold. This indeed makes our implementability modality to be a dual master modality.

$\mathtt{c}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s) = \mathbf{1} \Leftrightarrow \mathrm{M}, s \models^{\mathrm{C}} \neg\varphi \wedge \Diamond_i\varphi$

The binary relation $\models^{\mathrm{C}}$ between a formula in L^{C} and a pair M, s consisting of a model M for L^{C} and a state s in M is for choices defined by:

$\mathrm{M}, s \models^{\mathrm{C}} \mathbf{C}_i\varphi \Leftrightarrow \varphi \in \mathrm{C}(i, s)$

The definition of $\mathtt{r}^{\mathrm{C}}$ for the selection actions indeed provides for a correct model transformation.

6.3. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $\varphi \in \mathrm{L}$*, if* $\mathrm{M}', s = \mathtt{r}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s)$ *then* $\mathrm{M}' \in \mathrm{M}^{\mathrm{C}}_{\sim}$.

Besides being correct in that well-defined models are transformed into well-defined models, our formalisation of the act of selecting is also correct with respect to minimal change. That is, the change caused by selecting some formula is minimal given that the formula is to be marked chosen, which implies that our formalisation of selections does not suffer from the frame problem. The following proposition provides a (partial) formalisation of this property.

6.4. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $\varphi \in \mathrm{L}$*, if* $\mathrm{M}', s = \mathtt{r}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s)$ *then for all states* s' *in* M, $\mathrm{M}, s' \models^{\mathrm{C}} \psi$ *iff* $\mathrm{M}', s' \models^{\mathrm{C}} \psi$*, for all* $\psi \in \mathrm{L}$.

Proposition 6.4 states that all formulae from L are interpreted identically in a model M and in the one resulting from selecting some formula in an arbitrary state of M. As a direct consequence of this proposition we have the following corollary, which states that the interpretation of wishes and implementability formulae persists under selecting some formula.

6.5. Corollary. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $\varphi \in \mathrm{L}$*, if* $\mathrm{M}', s = \mathtt{r}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s)$ *then for all states* s' *in* M *and all* $\psi \in \mathrm{L}$*:*

- $\mathrm{M}, s' \models^{\mathrm{C}} \mathbf{W}_i\psi \Leftrightarrow \mathrm{M}', s' \models^{\mathrm{C}} \mathbf{W}_i\psi$
- $\mathrm{M}, s' \models^{\mathrm{C}} \Diamond_i\psi \Leftrightarrow \mathrm{M}', s' \models^{\mathrm{C}} \Diamond_i\psi$

Having defined wishes and selections one might be tempted to straightforwardly define goals to be selected wishes, i.e. $\mathbf{Goal}_i\varphi =^{\mathrm{def}} \mathbf{W}_i\varphi \wedge \mathbf{C}_i\varphi$. This definition is however not adequate to formalise the idea of goals being selected, *unfulfilled*, *implementable* wishes. The reason for this is that in well-defined models from $\mathbf{M}^{\mathrm{C}}$ no relation is imposed between 'being selected' and 'being unfulfilled and implementable', i.e. one is not prevented by Definition 4.2 to come up with a well-defined model M in which for certain i and s the set $\mathrm{C}(i, s)$ contains formulae φ that are either fulfilled or not implementable. We see basically two ways of solving this problem, a semantical and a syntactical one. Semantically one could restrict the set of well-defined models for L^{C} to those in which the set $\mathrm{C}(i, s)$ contains for all agents i and states s only unfulfilled and implementable formulae, thereby ensuring beforehand that goals are unfulfilled and implementable when using the definition suggested above. Syntactically one could define goals to be only those selected wishes that are indeed unfulfilled and implementable. Hence instead of (semantically) restricting the set of well-defined models for L^{C} one (syntactically) expands the definition of goals. Although both the semantic and the syntactic approach are equally well applicable, we will restrict ourselves here to pursuing the syntactic one. Therefore, goals are defined to be those wishes that are unfulfilled, implementable and selected.

6.6. DEFINITION. The $\mathbf{Goal}_i$ operator is for $i \in \mathrm{A}$ and $\varphi \in \mathrm{L}$ defined by:
$\mathbf{Goal}_i\varphi =^{\mathrm{def}} \mathbf{W}_i\varphi \wedge \neg\varphi \wedge \Diamond_i\varphi \wedge \mathbf{C}_i\varphi$

As mentioned above, the goals of agents, being the explicit and active notions that they are, are not to validate the properties of logical omniscience as formalised in Definition 5.2. Fortunately, though not surprisingly, this indeed turns out to be the case when defining goals as in Definition 6.6.

6.7. PROPOSITION. *None of the properties of logical omniscience formalised in Definition 5.2, with the exception of LO7, is valid for the* $\mathbf{Goal}_i$ *operator.*

The only property of logical omniscience satisfied by the goal operator, viz. LO7, formalises the idea that an agent's goals are consistent. This is a highly desirable property for rational creatures. For although it is quite possible for a rational agent to have contradictory wishes, it is rather irrational to try and fulfil these simultaneously.

Besides invalidating the undesired ones among the properties of logical omniscience, particularly those embodying the side-effect and transference problem, our definition of goals and selections has some other pleasant and desirable features. The following proposition formalises some of these features together with some properties characterising the act of selecting.

6.8. PROPOSITION. *For all* $i \in \mathrm{A}$ *and* $\varphi \in \mathrm{L}$ *we have:*

1. $\models^{\mathrm{C}} \mathbf{W}_i\varphi \leftrightarrow \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\top$
2. $\models^{\mathrm{C}} \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\top \leftrightarrow \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\mathbf{C}_i\varphi$
3. $\models^{\mathrm{C}} \neg\mathbf{A}_i\mathtt{select}\,\varphi \rightarrow [\mathrm{do}_i(\mathtt{select}\,\varphi)]\neg\mathbf{Goal}_i\varphi$
4. $\models^{\mathrm{C}} \mathbf{PracPoss}_i(\mathtt{select}\,\varphi, \top) \leftrightarrow \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\mathbf{Goal}_i\varphi$
5. $\models^{\mathrm{C}} \varphi \Rightarrow \models^{\mathrm{C}} \neg\mathbf{Goal}_i\varphi$
6. $(\varphi \rightarrow \psi) \rightarrow (\mathbf{Goal}_i\varphi \rightarrow \mathbf{Goal}_i\psi)$ *is not for all* $\varphi, \psi \in \mathrm{L}$ *valid*
7. $\mathbf{K}_i(\varphi \rightarrow \psi) \rightarrow (\mathbf{Goal}_i\varphi \rightarrow \mathbf{Goal}_i\psi)$ *is not for all* $\varphi, \psi \in \mathrm{L}$ *valid*

The first item of Proposition 6.8 states that agents have the opportunity to select all, and nothing but, their wishes. The second item formalises the idea that every choice for which an agent has the opportunity results in the selected wish being marked chosen. In the third item it is stated that whenever an agent is unable to select some formula, then selecting this formula will not result in it becoming one of its goals. The related item 4 states that all, and nothing but, practically possible selections result in the chosen formula being a goal. The fifth item provides a strengthening of the invalidation of the second property of logical omniscience, which embodies the transference problem. It states that no logically inevitable formula qualifies as a goal. Hence whenever a formula is valid this does not only not necessarily imply that it is a goal but it even necessarily implies that it is not. The last two items of Proposition 6.8 are related to the avoidance of the transference problem, and state that goals are neither closed under implications nor under known implications.

7. Formalising commitments

The last part of our formalisation of motivational attitudes concerns the agents' commitments. Commitments to actions represent promises to perform these actions, i.e. an agent that is committed to an action has promised itself to perform the action. As mentioned above,

commitments may be made to plans for goals, i.e. whenever an agent is committed it should be to an action that is correct and feasible to bring about at least one of its goals.

Not only do we formalise this static aspect of made commitments, but we also consider the dynamic aspect of making and undoing commitments. The act of committing is related to, and can be seen as, an elementary implementation of practical reasoning, the process through which agents decide that they should perform certain actions (their ought-to-do's) on the basis of their wishes, desires or goals (their ought-to-be's). Ever since Aristotle, the study of practical reasoning has formed a major constituent of the research in analytical philosophy [48]. According to Von Wright [64], the essence of practical reasoning is best captured by the following syllogism:

> i intends to make it true that φ
>
> i thinks that, unless it does α, it will not achieve this
>
> Therefore i intends to do α.

The simplified version of practical reasoning that we aim to formalise through the act of committing can be described by the following syllogism,

> i *knows* that φ is one of its *goals*
>
> i *knows* that α is *correct* and *feasible* with respect to φ
>
> Therefore i has the *opportunity* to *commit* itself to α

which corresponds to the idea that commitments may be made to actions that are known to be correct and feasible to achieve some of the agent's known goals.[13]

Commitments are formalised through the **Committed**_ _ operator: $\mathbf{Committed}_i\alpha$ denotes that agent i is committed to the action α. The act of committing is modelled by the (special) action `commit_to` _: $\texttt{commit\_to}\,\alpha$ represents the act of committing to the (regular) action α. As mentioned in Section 3, commitments, though in general persistent, should not be maintained when having become useless or impossible, i.e. agents should have the possibility to undo useless or impossible commitments. This act of uncommitting is formalised by the `uncommit` _ action: $\texttt{uncommit}\,\alpha$ denotes the act of undoing the commitment to the action α. In the sequel we successively formalise the act of committing, the commitments that have been made, and the act of uncommitting.

7.1. Getting committed

The act of committing, though of a special nature compared to other actions, is treated as a fully-fledged action, i.e. we define what it means to have the ability or opportunity to commit, and what the result of committing is. To start with the latter notion, given the relation between the infinitive 'to commit' and the past participle 'committed', it seems rather obvious that the act of committing should result in the agent being committed. Determining

[13] We abstract here from the way how exactly an agent comes to know that α is *correct* and *feasible* with respect to φ. This may be by means of a classical planner like STRIPS [16], or by consulting / using a kind of pre-compiled plan base as in certain agent architectures. Also, in case there are more possibilities for actions α to get to the goal φ, we do not consider a particular mechanism to choose such action (in some optimal way). We only say that when φ is a known goal and α is known to be correct and feasible for obtaining φ, the agent may (has the opportunity) to commit to that action α, allowing maximal flexibility on the agent's part whether it will actually perform such a commitment.

when an agent has the opportunity to perform a $\texttt{commit\_to}\,\alpha$ action is equally obvious, for it is inspired by the syllogism describing our version of practical reasoning given above. Hence agent i has the opportunity to perform the action $\texttt{commit\_to}\,\alpha$ if and only if it knows that α is correct and feasible to bring about one of its goals. This leaves to determine the constituents of the ability of an agent to commit itself. Our definition of this ability is inspired by the observation that situations where agents are committed to two (or more) essentially different actions are highly problematic. Since 'being committed to α' intuitively corresponds to 'having promised (to oneself) to perform α next (or at least a.s.a.p.)', it is unclear how to interpret the case where an agent is committed to two different actions. Should both actions be performed simultaneously? But what does it mean that actions are performed simultaneously? Are they performed concurrent, interleaved or in parallel? Or should the actions be performed sequentially? If so, in which order? And what then if the commitment to perform one action does not persist under execution of the other action? It is clear that these questions have no unique answers. Here we shall assume simply that agents do not have multiple commitments. One way to ensure this is to let an agent have the ability to commit itself only if it is not up to any previously made commitments, i.e. an agent is capable to commit only if it is not already committed. (However, it should also be clear that in principle our framework allows other, more complicated situations, if one would be interested to model that. However, since some of the results in the sequel depend on the present choice, these should be reconsidered then.)

As mentioned previously, an agent's commitments are interpreted by means of the so-called agenda function. The idea is that this function yields, for a given agent and a given state, the actions that the agent is committed to. Whenever an agent successfully commits itself to an action the agent's agenda is updated accordingly. The actual formal definition capturing this fairly unsophisticated idea is itself rather complicated. The reason for this lies in various desiderata that commitments and the act of committing should meet.

The first of these desiderata is that commitments should be known, i.e. agents should be aware of the commitments that they have made. To bring about this knowledge of commitments, epistemic equivalence classes rather than states are considered in an agenda update. Thus whenever agent i commits itself to action α in some state s of a model, the agenda of all states s' that are epistemically equivalent with s is updated appropriately.

The second and very important desideratum imposed on commitments is that they behave compositionally correct, i.e. the commitment to a composite action is linked in a rational way to commitments to its constituents. It is for example desirable that an agent that is committed to an action $\texttt{if}\,\varphi\,\texttt{then}\,\alpha_1\,\texttt{else}\,\alpha_2\,\texttt{fi}$ is also committed to α_1 whenever it knows that φ holds, and that an agent committed to the action $\alpha_1;\alpha_2$ is (currently) committed to α_1 and committed to α_2 in the state of affairs that results from executing α_1.

To bring about rational behaviour of commitments with respect to sequentially composed actions the actual update does not just concern the epistemic equivalence class of the current state, but also that of all the states that lay alongside the execution trajectory of the action. For example, if an agent i commits itself to $\alpha_1;\alpha_2$ in the state s of some model, then the epistemic equivalence class of s is updated with the commitment to α_1, and the epistemic equivalence class of the state s'' that results from executing α_1 in some s' that is an element of the epistemic equivalence of s is updated with the commitment to α_2.

Although we will put in the agent's agenda (syntactical) actions that represent actions it is committed to, we will, of course, also need semantical entities to link these actions to what actually happens when an agent performs (part of) the action in its agenda. To this end we

will use some notions that are well-known from the area of the semantics of programming languages, viz. computation sequences, computation runs and transition relations. Let us start with the introduction of computation sequences and runs.

Since the actions from Ac are deterministic, for each event built out of these actions there is at most one finite computation sequence which consists of the (semi-)atomic actions that occur in the halting executing of the event. Or phrased differently, the set of finite computation runs of a given event $\mathrm{do}_i(\alpha)$ is either empty or a singleton set. This property of deterministic actions facilitates the definition of finite computation runs to a considerable extent: we simply define it to be the unique finite computation sequence for which execution terminates.

Recall the definition of Acseq from Definition 2.1.

7.1. Definition. The function CS, yielding the *finite computation sequences* of a given action, is inductively defined as follows.

$$
\begin{array}{ll}
\mathrm{CS} & : \ \mathrm{Ac} \to \wp(\mathrm{Acseq}) \\
\mathrm{CS}(\alpha) & = \{\alpha\} \text{ if } \alpha \text{ is semi-atomic} \\
\mathrm{CS}(\alpha_1;\alpha_2) & = \{\alpha_1';\alpha_2' \mid \alpha_1' \in \mathrm{CS}(\alpha_1), \alpha_2' \in \mathrm{CS}(\alpha_2)\} \\
\mathrm{CS}(\texttt{if } \varphi \texttt{ then } \alpha_1 & \\
\qquad \texttt{else } \alpha_2 \texttt{ fi}) & = \mathrm{CS}(\varphi?;\alpha_1) \cup \mathrm{CS}(\neg\varphi?;\alpha_2) \\
\mathrm{CS}(\texttt{while } \varphi \texttt{ do } \alpha \texttt{ od}) & = \cup_{k=1}^{\infty} \mathit{Seq}_k(\texttt{while } \varphi \texttt{ do } \alpha \texttt{ od}) \cup \{\neg\varphi?\}
\end{array}
$$

where for $k \geq 1$

$$
\begin{array}{ll}
\mathit{Seq}_k(\texttt{while } \varphi \texttt{ do } \alpha \texttt{ od}) & = \{(\varphi?;\alpha_1'); \ldots ;(\varphi?;\alpha_k');\neg\varphi? \\
& \qquad \mid \alpha_j' \in \mathrm{CS}(\alpha_1) \text{ for } j = 1,\ldots,k\}
\end{array}
$$

7.2. Definition. Since Ac is closed under the core clauses of Definition 2.1 only, the function $\mathrm{CS} : \mathrm{Ac} \to \wp(\mathrm{Acseq})$ is defined as usual. For $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ the function $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}} : \mathrm{A} \times \mathrm{Ac} \times \mathrm{S} \to \wp(\mathrm{Acseq})$ is defined by:

$$\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha,s) = \{\alpha' \in \mathrm{CS}(\alpha) \mid \mathtt{r}^{\mathrm{C}}(i,\alpha')(\mathrm{M},s) \neq \emptyset\}$$

Note that the actions that we consider are deterministic, i.e. the set $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha,s)$ consists of at most one element, for any i,α,s. This must be kept in mind below, when we consider properties of computation runs. Moreover, since actions are deterministic we will often write (sloppily) $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha,s) = \alpha'$ instead of $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha,s) = \{\alpha'\}$.

Let, for sets $A_1, A_2 \in \wp(\mathrm{Acseq})$, $A_1;A_2$ stand for the set $\{\alpha_1;\alpha_2 \mid \alpha_1 \in A_1, \alpha_2 \in A_2\}$. Then we can state the following.

7.3. Proposition. *If* $\mathtt{r}^{\mathrm{C}}(i,\alpha_1;\alpha_2)(\mathrm{M},s) \neq \emptyset$ *then* $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha_1;\alpha_2,s) = \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha_1,s);\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha_2,s')$ *for (the unique)* $s' \in \mathtt{r}^{\mathrm{C}}(i,\alpha_1)(\mathrm{M},s)$.

Next we introduce the notion of a transition relation in the spirit of Structural Operational Semantics of Plotkin ([43]), which is a very neat and elegant way to describe computation by means of (single) transition steps. This method is widely used in computer science and we can use it here fruitfully to describe what happens with the agent's agenda when it performs actions step by step (from that agenda, so to speak).

In our set-up we (first) consider transitions of the form $\langle \alpha, s\rangle \rightarrow^{\mathrm{M}}_{i,a} \langle \alpha', s'\rangle$, where $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}, \alpha, \alpha' \in \mathrm{Ac}, i \in \mathrm{A}, a \in \mathrm{At}^{+}$ and $s, s' \in \mathrm{S}$. (To ease notation in the sequel we will drop the superscript M if the set of states S in the model M is understood.) We use the symbol Λ for

the empty action, with as property that $\Lambda;\alpha = \alpha;\Lambda = \alpha$. Furthermore, we use the projection function π_2, which is assumed to yield the second element of a pair.

Transitions are given by the following deductive system, often called a transition system:

7.4. DEFINITION. Let the model $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ be given. The transition system T_{M} is given by the following axioms:

- $\langle \alpha, s\rangle \rightharpoonup_{i,\alpha} \langle \Lambda, s'\rangle$ with $s' = \pi_2(\mathbf{r}^{\mathrm{C}}(i,\alpha)(\mathrm{M},s))$ if α is semi-atomic and $\mathbf{r}^{\mathrm{C}}(i,\alpha)(\mathrm{M},s) \neq \emptyset$
- $\langle \mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1\ \mathtt{else}\ \alpha_2\ \mathtt{fi}, s\rangle \rightharpoonup_{i,\varphi?} \langle \alpha_1, s_{\varphi?}\rangle$ if $s \models \varphi$
- $\langle \mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1\ \mathtt{else}\ \alpha_2\ \mathtt{fi}, s\rangle \rightharpoonup_{i,\neg\varphi?} \langle \alpha_2, s_{\neg\varphi?}\rangle$ if $s \not\models \varphi$
- $\langle \mathtt{while}\ \varphi\ \mathtt{do}\ \alpha\ \mathtt{od}, s\rangle \rightharpoonup_{i,\varphi?} \langle \alpha; \mathtt{while}\ \varphi\ \mathtt{do}\ \alpha\ \mathtt{od}, s_{\varphi?}\rangle$ if $s \models \varphi$
- $\langle \mathtt{while}\ \varphi\ \mathtt{do}\ \alpha\ \mathtt{od}, s\rangle \rightharpoonup_{i,\neg\varphi?} \langle \Lambda, s_{\neg\varphi?}\rangle$ if $s \not\models \varphi$

  and the following rule:
- $\frac{\langle \alpha_1, s\rangle \rightharpoonup_{i,a} \langle \alpha_1', s'\rangle}{\langle \alpha_1;\alpha_2, s\rangle \rightharpoonup_{i,a} \langle \alpha_1';\alpha_2, s'\rangle}$

Obviously, there is a relation between transitions and the computation runs we introduced earlier. This relation is given by the following proposition. (Here we use as a convention that $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\Lambda,s)$ is the empty sequence of actions.)

7.5. PROPOSITION. $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha,s) = a;\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha',s')$ *iff* $T_{\mathrm{M}} \vdash \langle \alpha, s\rangle \rightharpoonup_{i,a} \langle \beta, s'\rangle$ *for some* β *with* $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\beta,s') = \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha',s')$.

This proposition gives us as a corollary how computation runs can be viewed as being generated by transitions.

7.6. COROLLARY. $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha,s) = \{a_1;a_2;\ldots;a_n\}$ *iff* $T_{\mathrm{M}} \vdash \langle \alpha, s\rangle \rightharpoonup_{i,a_1} \langle \alpha_1, s_1\rangle \rightharpoonup_{i,a_2} \langle \alpha_2, s_2\rangle \rightharpoonup_{i,a_3} \ldots \rightharpoonup_{i,a_n} \langle \alpha_n, s_n\rangle$ *for some* $\alpha_1,\ldots,\alpha_n \in \mathrm{Ac}, s_1,\ldots,s_n \in \mathrm{S}$, *such that* $\alpha_n = \Lambda$.

Finally, we extend the transition relation $\langle \alpha, s\rangle \rightharpoonup^{\mathrm{M}}_{i,a} \langle \alpha', s'\rangle$ to the relation $\langle \alpha, s\rangle \rightarrow^{\mathrm{M}}_{i,a} \langle \alpha', s'\rangle$, which is like the former, but now also the performance of a semi-atomic action a *not* matching with the start of the action α is covered. As we will employ this transition system to describe how the agent's agenda is transformed under performance / execution of actions, in this case the agenda (containing α) should remain the same (for nothing from it is being done). So we stipulate:

7.7. DEFINITION. For $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}, \alpha, \alpha' \in \mathrm{Ac}, i \in \mathrm{A}, a \in \mathrm{At}^+$ and $s, s' \in \mathrm{S}$ we define the transition relation $\langle \alpha, s\rangle \rightarrow^{\mathrm{M}}_{i,a} \langle \alpha', s'\rangle$ as the transition relation $\langle \alpha, s\rangle \rightharpoonup^{\mathrm{M}}_{i,a} \langle \alpha', s'\rangle$ extended with the clause

$\langle \alpha, s\rangle \rightarrow^{\mathrm{M}}_{i,a} \langle \alpha, s'\rangle$

where $s' = \pi_2(\mathbf{r}^{\mathrm{C}}(i,a)(\mathrm{M},s))$ and a is not a prefix of $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i,\alpha,s)$. (In the sequel we will drop the superscript M again.)

Intuitively, the extra clause for $\rightarrow_{i,a}$ means that the execution of a in this case does not help resolving the agenda, so that this remains to be done in its totality.[14]

For reasons of convenience we introduce, analogously to the Can-predicate, a so-called Intend-predicate, which is meant to formalise the (possible) intentions of agents. The definition of this predicate is based on the idea that agents (possibly) intend to do all the actions that are correct and feasible with respect to some of their goals. As such, possible intention provides the precondition for successful commitment[15].

7.8. DEFINITION. For $\alpha \in \mathrm{Ac}^{\mathrm{C}}, i \in \mathrm{A}$ and $\varphi \in \mathrm{L}$ we define:
$\mathbf{PossIntend}_i(\alpha, \varphi) =^{\mathrm{def}} \mathbf{Can}_i(\alpha, \varphi) \wedge \mathbf{K}_i\mathbf{Goal}_i\varphi$

Having established the formal prerequisites, we can now present the definitions formalising the intuitive description of the act of committing as presented above.

7.9. DEFINITION. For all $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ with state s, for all $i \in \mathrm{A}$ and $\alpha \in \mathrm{Ac}$ we define:

$\mathtt{r}^{\mathrm{C}}(i, \mathtt{commit\_to}\,\alpha)(\mathrm{M}, s) = \emptyset$ if $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{PossIntend}_i(\alpha, \varphi)$ for all $\varphi \in \mathrm{C}(i, s)$
$\mathtt{r}^{\mathrm{C}}(i, \mathtt{commit\_to}\,\alpha)(\mathrm{M}, s) = \mathrm{M}', s$ with $\mathrm{M}' = \langle \mathrm{S}, \pi, \mathrm{R}, \mathtt{r}_0, \mathtt{c}_0, \mathrm{W}, \mathrm{C}, \mathrm{Agenda}' \rangle$
where Agenda' is minimal such that it is closed under the following conditions:
for all $s' \in [s]_{\mathrm{R}(i)}$, $\mathrm{Agenda}'(i, s') = \mathrm{Agenda}(i, s') \cup \{\alpha\}$
and for all $s', s'', s''' \in \mathrm{S}, \alpha' \in \mathrm{Agenda}'(i, s')$ such that, for some semi-atomic a,
$T_{\mathrm{M}} \vdash \langle \alpha', s' \rangle \rightarrow_{i,a} \langle \alpha'', s'' \rangle$ and $s''' \in [s'']_{\mathrm{R}(i)}$:
$\mathrm{Agenda}'(i, s''') = \mathrm{Agenda}(i, s''') \cup \{\alpha''\}$
otherwise

$\mathtt{c}^{\mathrm{C}}(i, \mathtt{commit\_to}\,\alpha)(\mathrm{M}, s) = \mathbf{1}$ iff $\mathrm{Agenda}(i, s) = \emptyset$

This definition makes sure that in the state where the commitment to an action α is made, the agent's agenda is updated with α, as well as in any state that is epistemically equivalent with s, i.e. in the epistemic equivalence class of s. The latter will have as an effect that the commitment to α is known to the agent, since every epistemic alternative to s will contain the same commitment in the agenda. Moreover, the rest of the definition ensures that the agenda is also updated in all states reachable from s (or elements from its equivalence class) by performing actions β that are according to the fulfilment of executing the action α by putting in the agenda the 'remainder' of α after its 'partial execution' β.

The clause in the definition regarding the capability function is such that an agent is only able to commit to some α if its agenda is empty. This is, of course, a very strong restriction, to which we will return at the end of our paper.

Note, by the way, that the definition above is in line with the restrictions we've put on the models in Definition 4.1 in the sense that the execution of a commit action preserves the property of having at most one action in the agent's agenda (which is the case on which we

[14] Our model may be (again) too liberal in the sense that nothing in our framework prevents an agent to perform, at any time, something else than is in its agenda (which then remains being recorded). In practice it may be imperative to e.g. start right away with the execution of (the actions recorded in) its agenda or not interrupt the execution of its agenda when having started this. We believe that in general it is difficult to give conditions for this; we feel it heavily depends on the application. Here we have again chosen for a simple model as a first approximation, which may be modified easily if one would like to.

[15] Our paraphrase of Cohen & Levesque's motto 'intention is choice plus commitment' [9] could therefore be stated as 'commitments are chosen possible intentions'.

have focused). Without the restriction mentioned there in the fourth bullet the successful performance of an atomic action might result in the same state, so that this state would get two values of the agent's agenda (viz. that associated with the situation before the execution of the atomic action, and one associated with that after the execution), which would violate the property mentioned. Moreover, intuitively, in this case the agent cannot determine whether the atomic action of concern has yet to be performed or not, which seems a rather undesirable property that we have excluded here by the restriction. For example, suppose $a; b \in \mathrm{Agenda}'(i, s)$. If it now were the case that $s \in \mathtt{r}_0(i, a)(s)$, the closure condition in our definition above would yield that also $b \in \mathrm{Agenda}'(i, s)$. So, since then both a and $a; b$ are in the agent's agenda in this situation, it cannot distinguish whether a has been performed already or not. In Definition 4.4, an analogous problem is circumvented for test actions, so that this problem does not occur for any semi-atomic action.

A further important point to notice is that only the agenda of i is modified, and that only in those states that are somehow, i.e. by a combination of state-transitions and epistemic accessibility relations, connected to the state in which the commitment is being made. All other elements of the model remain unchanged.

The latter aspect mentioned above, i.e. the minimality of the change caused by performing a commitment, is partly formalised in Proposition 7.11 given below. Proposition 7.10 states the correctness of the definition of $\mathtt{r}^{\mathrm{C}}$ as presented above in the sense that it yields a (unique) well-defined model when applied to a well-defined model.

7.10. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $\alpha \in \mathrm{Ac}$*, if* $\mathrm{M}', s = \mathtt{r}^{\mathrm{C}}(i, \mathtt{commit\_to}\,\alpha)(\mathrm{M}, s)$ *then* $\mathrm{M}' \in \mathrm{M}^{\mathrm{C}}_{\sim}$.

7.11. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $\alpha \in \mathrm{Ac}$*, if* $\mathrm{M}', s = \mathtt{r}^{\mathrm{C}}(i, \mathtt{commit\_to}\,\alpha)(\mathrm{M}, s)$ *then for all states* s' *in* M, $\mathrm{M}, s' \models^{\mathrm{C}} \varphi$ *iff* $\mathrm{M}', s' \models^{\mathrm{C}} \varphi$*, for all* $\varphi \in \mathrm{L}$.

Additional properties related to the `commit` actions are given in 7.4.

We remark that due to the fact that the agenda contains *syntactical* elements (viz. actions), we have the *same* element in the agenda within an epistemic equivalence class (so that this is known to be committed to by the agent), but the interpretation (= execution) of this agenda element may be quite different in the different worlds of the equivalence class. This means that the agent may not know where the execution of its commitment amounts to (and results in). This is quite a realistic feature, and is also in a sense reminiscent of the Moore - Morgenstern discussion [40, 41, 42] concerning the need for a *rigid* designator for an action in order to enable an agent to know how to do the action. As far as syntax is concerned, the description is rigid (the same in every epistemic alternative world); as to semantics / interpretation it is not. For example, an action consisting of a statement "$\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}$" may amount to the execution of "α_1" in some epistemic alternatives (viz. where φ holds), while it may amount to "α_2" in other alternatives (where φ doesn't hold), so that the agent doesn't (*a priori*) know whether α_1 or α_2 will be executed. But still he *does* know that he is committed to "$\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi}$"!

7.2. Being committed

After the rather elaborate and fairly complicated definition formalising the act of committing, defining what it means to be committed is a relatively straightforward and easy job. Basically,

agents are committed to the actions in its agenda.[16] The only additional aspect that has to be taken into account when defining the semantics of the **Committed_** operator is that agents should start at the very beginning (a very good place to start), and should therefore be also committed to initial executions of the actions in its agenda. As we can represent executions by computation runs, we can thus say that an agent committed to an action α should also be committed to actions that have computation runs that are initial fragments of the computation run of α. Formally we ensure this behaviour by using the prefix relation on basic actions. The notation $\mathrm{Prefix}(\alpha, \beta)$ for $\alpha, \beta \in \mathrm{Acseq}$ expresses that action sequence α is a prefix of action sequence β. The definition of $\models^{\mathrm{C}}$ for the $\mathbf{Committed}_i$ operator could then be informally interpreted as 'an agent is committed to those actions of which the computation run is a prefix of one of the actions in its agenda'.

7.12. DEFINITION. The binary relation $\models^{\mathrm{C}}$ between a formula in L^{C} and a pair M, s consisting of a model M for L^{C} and a state s in M is for commitments defined by:

$\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Committed}_i \alpha \Leftrightarrow$
$\forall s' \in [s]_{\mathrm{R}(i)} \exists \alpha_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha, s') \exists \alpha_2 \in \mathrm{Agenda}(i, s') \exists \alpha_2' \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha_2, s')(\mathrm{Prefix}(\alpha_1, \alpha_2'))$

An investigation of the properties of the commitment operator is postponed to 7.4.

7.3. Getting uncommitted

By performing an uncommit action, agents may undo previously made commitments that turned out to be either useless or impossible. That is, as soon as an agent no longer knows some commitment to be correct and feasible for any of its goals it may undo this commitment. Just as we did for the commit action, we have to decide upon the constituents of the result, opportunity and ability for the actions formalising the act of uncommitting. The result of such an action is obvious: agents should no longer be committed to α after a successful performance of an $\mathtt{uncommit}\,\alpha$ action[17]. Defining what it means to have the opportunity and ability to uncommit represents a somewhat more arbitrary choice. We have decided to let an agent have the opportunity to undo any of its commitments, i.e. there is nothing in its circumstances that may prevent an agent to undo a commitment. Our loyal, diligent agents are however only (morally) capable of undoing commitments that have become redundant. The actual definition of the functions $\mathtt{r}^{\mathrm{C}}$ and $\mathtt{c}^{\mathrm{C}}$ consists of nothing but a formalisation of these intuitive ideas.

7.13. DEFINITION. For all $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ with state s, for all $i \in \mathrm{A}$ and $\alpha \in \mathrm{Ac}$ we define:
(For technical convenience, in the following we define, simultaneously with the new agenda, the set $\mathrm{Reachable}_{\mathrm{M},s}$ of states that are reachable from s by alternatively considering epistemic alternatives and performing actions.)

[16] In our framework being committed to some action corresponds to having the intention to do this action in the sense of Cohen & Levesque [9].

[17] As was pointed out to us by John Fox, this description of the result of undoing a commitment comprises a major simplification. For in real life, undoing commitments may involve more than just abandoning future commitments: it may also be necessary to (try to) undo all the effects that followed from initially pursuing the commitment. For example, if an agent that is committed to $\alpha_1; \alpha_2$ finds out after having done α_1 that its commitment to α_2 should be undone, then it is very plausible that it should not only remove α_2 from its agenda but also try to undo as many of the effects of α_1 as possible.

$\mathtt{r}^{\mathrm{C}}(i, \mathtt{uncommit}\,\alpha)(\mathrm{M}, s) = \emptyset$ if $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{Committed}_i\alpha$
$\mathtt{r}^{\mathrm{C}}(i, \mathtt{uncommit}\,\alpha)(\mathrm{M}, s) = \mathrm{M}', s$ with $\mathrm{M}' = \langle \mathrm{S}, \pi, \mathrm{R}, \mathtt{r}_0, \mathtt{c}_0, \mathrm{W}, \mathrm{C}, \mathrm{Agenda}' \rangle$
where Agenda' and $\mathrm{Reachable}_{\mathrm{M},s}$ are minimal such that they are closed under the conditions:
for all $s' \in [s]_{\mathrm{R}(i)}$:
$s' \in \mathrm{Reachable}_{\mathrm{M},s}$ and
$\mathrm{Agenda}'(i, s') = \mathrm{Agenda}(i, s') \setminus \{\beta \mid \mathrm{Prefix}(\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha, s'), \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \beta, s'))\}$
and for all $s' \in \mathrm{Reachable}_{\mathrm{M},s}$, $s'', s''' \in \mathrm{S}$ with $\alpha' \in \mathrm{Agenda}'(i, s')$ and such that,
for some semi-atomic a, $T_{\mathrm{M}} \vdash \langle \alpha', s' \rangle \rightarrow_{i,a} \langle \alpha'', s'' \rangle$ and $s''' \in [s'']_{\mathrm{R}(i)}$:
$s''' \in \mathrm{Reachable}_{\mathrm{M},s}$ and
$\mathrm{Agenda}'(i, s''') = \mathrm{Agenda}(i, s''') \setminus \{\beta \mid \mathrm{Prefix}(\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha'', s'''), \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \beta, s'''))\}$
otherwise

$\mathtt{c}^{\mathrm{C}}(i, \mathtt{uncommit}\,\alpha)(\mathrm{M}, s) = \mathbf{1}$ iff $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{PossIntend}_i(\alpha, \varphi)$ for all $\varphi \in \mathrm{C}(i, s)$

Note the slight complication in this definition regarding prefixes: due to our definition of being committed to an action if it constitutes the partial execution of some action that is recorded in the agent's agenda, we must also take care that when we uncommit to some action α that we remove from the agent's agenda all actions β that have α as their partial execution. Formally this means that we remove all β that have the computation run of α as a prefix of their computation run. This will ensure the desirable property below (Proposition 7.16, fifth item) stating that when the agent uncommits to an action to which it was committed, it will indeed be not committed to this action afterwards.

The clause regarding $\mathtt{c}^{\mathrm{C}}$ of the uncommit action states that an agent is (cap)able to $\mathtt{uncommit}\,\alpha$ ('drop the intention to do α') iff it doesn't consider α a possible intention anymore, viz. if it doesn't know α to be correct and feasible for any known goal φ anymore.

Our definition of $\mathtt{r}^{\mathrm{C}}$ for the `uncommit` actions is also twofold correct: not only does performing an `uncommit` action provide for a correct model-transformation, but also does it do so while causing minimal change.

7.14. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $\alpha \in \mathrm{Ac}$*, if* $\mathrm{M}', s = \mathtt{r}^{\mathrm{C}}(i, \mathtt{uncommit}\,\alpha)(\mathrm{M}, s)$ *then* $\mathrm{M}' \in \mathrm{M}^{\mathrm{C}}_{\sim}$.

7.15. Proposition. *For all* $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ *with state* s*, for all* $i \in \mathrm{A}$ *and* $\alpha \in \mathrm{Ac}$*, if* $\mathrm{M}', s = \mathtt{r}^{\mathrm{C}}(i, \mathtt{uncommit}\,\alpha)(\mathrm{M}, s)$ *then for all states* s' *in* M*,* $\mathrm{M}, s' \models^{\mathrm{C}} \varphi$ *iff* $\mathrm{M}', s' \models^{\mathrm{C}} \varphi$*, for all* $\varphi \in \mathrm{L}$.

Additional validities characterising the `uncommit` action are given below.

7.4. The statics and dynamics of commitments

Here we characterise the statics and dynamics of commitments by presenting some validities for $\models^{\mathrm{C}}$. For a start we consider a number of validities characterising the dynamics of commitments.

7.16. Proposition. *For all* $i \in \mathrm{A}$*,* $\alpha, \beta \in \mathrm{Ac}$ *and* $\varphi \in \mathrm{L}$ *we have:*

1. $\models^{\mathrm{C}} \mathbf{PossIntend}_i(\alpha, \varphi) \rightarrow \langle \mathrm{do}_i(\mathtt{commit\_to}\,\alpha) \rangle \top$
2. $\models^{\mathrm{C}} \langle \mathrm{do}_i(\mathtt{commit\_to}\,\alpha) \rangle \top \leftrightarrow \langle \mathrm{do}_i(\mathtt{commit\_to}\,\alpha) \rangle \mathbf{Committed}_i\alpha$
3. $\models^{\mathrm{C}} \mathbf{Committed}_i\alpha \rightarrow \neg\mathbf{A}_i\mathtt{commit\_to}\,\beta$
4. $\models^{\mathrm{C}} [\mathrm{do}_i(\mathtt{commit\_to}\,\alpha)] \neg\mathbf{A}_i\mathtt{commit\_to}\,\beta$

5. $\models^{\mathrm{C}} \mathbf{Committed}_i\alpha \leftrightarrow \langle \mathrm{do}_i(\mathtt{uncommit}\,\alpha)\rangle\neg\mathbf{Committed}_i\alpha$
6. $\models^{\mathrm{C}} \mathbf{PossIntend}_i(\alpha,\varphi) \rightarrow \neg\mathbf{A}_i\mathtt{uncommit}\,\alpha$
7. $\models^{\mathrm{C}} (\mathbf{C}_i\varphi \leftrightarrow \mathbf{K}_i\mathbf{C}_i\varphi) \rightarrow (\mathbf{A}_i\mathtt{uncommit}\,\alpha \leftrightarrow \mathbf{K}_i\mathbf{A}_i\mathtt{uncommit}\,\alpha)$
8. $\models^{\mathrm{C}} \mathbf{Committed}_i\alpha \wedge \neg\mathbf{Can}_i(\alpha,\top) \rightarrow \mathbf{Can}_i(\mathtt{uncommit}\,\alpha, \neg\mathbf{Committed}_i\alpha)$

The first two items of Proposition 7.16 jointly formalise our version of the syllogism of practical reasoning as described above. In the third item it is stated that being committed prevents an agent from having the ability to (re)commit. The fourth item states that the act of committing is ability-destructive with respect to future `commit` actions, i.e. by performing a commitment an agent loses its ability to make any other commitments. Item 5 states that being committed is a necessary and sufficient condition for having the opportunity to uncommit; as mentioned above, agents have the opportunity to undo all of their commitments. In item 6 it is stated that agents are (morally) unable to undo commitments to actions that are still known to be correct and feasible to achieve some goal. In item 7 it is formalised that agents know of their abilities to uncommit to some action. The last item states that whenever an agent is committed to an action that is no longer known to be practically possible, it knows that it can undo this impossible commitment.

The following proposition formalises some of the desiderata for the statics of commitments that turn out to be valid in the class $\mathbf{M}^{\mathrm{C}}$ of models for L^{C}.

7.17. Proposition. *For all* $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}$ *and all* $\varphi \in \mathrm{L}$ *we have:*

1. $\models^{\mathrm{C}} \mathbf{Committed}_i\alpha \rightarrow \mathbf{K}_i\mathbf{Committed}_i\alpha$
2. $\models^{\mathrm{C}} \mathbf{Committed}_i(\alpha_1;\alpha_2) \rightarrow \mathbf{Committed}_i\alpha_1 \wedge \mathbf{K}_i[\mathrm{do}_i(\alpha_1)]\mathbf{Committed}_i\alpha_2$
3. $\models^{\mathrm{C}} \mathbf{Committed}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi} \wedge \mathbf{K}_i\varphi \rightarrow \mathbf{Committed}_i(\varphi?;\alpha_1)$
4. $\models^{\mathrm{C}} \mathbf{Committed}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi} \wedge \mathbf{K}_i\neg\varphi \rightarrow \mathbf{Committed}_i(\neg\varphi?;\alpha_2)$
5. $\models^{\mathrm{C}}\mathbf{Committed}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od} \wedge \mathbf{K}_i\varphi \rightarrow \mathbf{Committed}_i((\varphi?;\alpha);\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$

The first item of Proposition 7.17 states that commitments are known. The second item states that a commitment to a sequential composition $\alpha_1;\alpha_2$ of actions implies a commitment to the initial part α_1, and that the agent knows that after execution of this initial part α_1 it will be committed to the remainder α_2. The third and fourth item formalise the rationality of agents with regard to their commitments to conditionally composed actions. The last item concerns the unfolding of a while-loop: if an agent is committed to a while-loop while knowing the condition of the loop to be true, then the agent is also committed to the then-part of the while-loop.

Let us give a simple example in which many of our treated terms re-occur.

7.18. Example. Consider an agent Eve living in a blocks world. Blocks may have different sizes. Here, we assume to have only three formats for real blocks: those of type A are bigger than those of B, which are in turn bigger that those of type C. Block Y can be on top of block X ($\mathtt{is\_on}(X,Y)$). If there is no such block Y on top of X, we write $\mathtt{is\_clear}(X)$. We distinguish the following atomic actions. Eve can do a $\mathtt{drop}(X)$ action, meaning that X is put on the floor, and she can perform $\mathtt{put}(X,Y)$, which has as an effect that Y is placed on X. We assume that Eve only has the opportunity of performing a $\mathtt{put}(X,Y)$-action, if X is clear. She is not able to put heavy blocks: $\neg\mathbf{A}_e\mathtt{put}(X,Y)$, for any block Y of type A. There is no opportunity to put a block X on Y if $X > Y$, even if Eve is able to do so. Concerning dropping, she is only able to perform $\mathtt{drop}(X)$ if X is clear. .

Let us summarise these assumptions in a first-order like language: in fact, this could also be done in propositional logic. The formula $\texttt{type}(X, A)$ denotes that block X is of type A. In the following, free variables must be understood to be universally quantified. The constraints $A_1, A_2, \ldots$ concern Eve's abilities, $O_1, O_2 \ldots$ are about opportunities, and the constraints $E_1, E_2 \ldots$ describe the effects of certain actions, whereas the constraints $N_1, N_2, \ldots$ are concerned with the *non-effects* of putting and dropping. For instance, N_1 says clear blocks on which no block is put, remain clear. Let us write $\mathbf{O}_e(\alpha)$ for the statement that Eve has the opportunity to do α, i.e., $\mathbf{O}_e(\alpha) \equiv \langle \mathrm{do}_e(\alpha)\rangle\top$. Finally, the properties $C_1, C_2, \ldots$ are used to specify additional constraints for our example. In particular, the axioms C_2, C_3 and C_4 denote that Eve is aware of her goals, abilities and the results of her actions, respectively. Let $i \in \{e, \ldots\}$.

C_1 $(\texttt{type}(X, A) \wedge \texttt{type}(Y, B) \wedge \texttt{type}(Z, C)) \rightarrow ((X > Y) \wedge (Y > Z) \wedge (X > Z))$

A_1 $\texttt{type}(Y, A) \leftrightarrow \neg\mathbf{A}_i\texttt{put}(X, Y)$

A_2 $(\texttt{clear}(X) \wedge \neg\texttt{type}(X, A)) \leftrightarrow \mathbf{A}_i\texttt{drop}(X)$

O_1 $\texttt{clear}(X) \wedge X \neq Y \wedge \neg(X < Y) \leftrightarrow \mathbf{O}_i(\texttt{put}(X, Y))$

O_2 $\mathbf{O}_i\texttt{drop}(X)$

E_1 $[\mathrm{do}_i(\texttt{put}(X, Y))](\texttt{is\_on}(X, Y) \wedge \neg\texttt{is\_clear}(X))$

E_2 $\texttt{is\_on}(X, Y) \rightarrow [\mathrm{do}_i(\texttt{drop}(Y))](\texttt{is\_clear}(X) \wedge \texttt{Floor}(Y))$

N_1 $(\texttt{is\_clear}(Z) \wedge Z \neq X) \rightarrow [\mathrm{do}_i(\texttt{put}(X, Y))](\texttt{is\_clear}(Z)$

N_2 $(\texttt{is\_on}(V, Z) \wedge Z \neq Y) \rightarrow [\mathrm{do}_i(\texttt{put}(X, Y))]\texttt{is\_on}(V, Z)$

N_3 $\texttt{is\_on}(X, Y) \rightarrow [\mathrm{do}_i(\texttt{drop}(Y))](\texttt{is\_clear}(X) \wedge \texttt{Floor}(Y))$

N_4 $(\texttt{is\_on}(X, Y) \wedge \texttt{is\_on}(U, V) \wedge X \neq U) \rightarrow [\mathrm{do}_i(\texttt{drop}(Y))]\texttt{is\_on}(U, V)$

N_5 $\texttt{is\_clear}(Z) \rightarrow [\mathrm{do}_i(\texttt{drop}(Y))]\texttt{is\_clear}(Z)$

N_6 $(X = Y \wedge U \neq V) \rightarrow [\mathrm{do}_i(\alpha)](X = Y \wedge U \neq V)$

C_2 $\mathbf{Goal}_e\varphi \rightarrow \mathbf{K}_e\mathbf{Goal}_e\varphi$

C_3 $\mathbf{A}_i\alpha \rightarrow \mathbf{K}_e\mathbf{A}_i\alpha$

C_4 $\langle \mathrm{do}_e(\alpha)\rangle\varphi \rightarrow \mathbf{K}_e\langle \mathrm{do}_e(\alpha)\rangle\varphi$

On top of these global properties, which remain invariant in our example, we also have the following minimality principles. To formulate them properly, let us say that the formulas $\texttt{type}(X, A), \texttt{is\_on}(X, Y), \texttt{is\_clear}(X), \texttt{on\_floor}(X)$ are all *objective* formulas, of which the truth value is determined by some valuation π (respecting the dependencies between them, like expressed by $(\texttt{is\_on}(X, Y) \rightarrow \neg\texttt{is\_clear}(X))$).

M_1 suppose $(M', s') = \langle S', \pi', R', \mathtt{r}'_0, \mathtt{c}'_0, W', C'Agenda'\rangle \in \mathtt{r}^c(i, a)(M, s)$.Then:
if $\pi \neq \pi'$, then $a \in \{\texttt{put}, \texttt{drop}\}$
if for some φ, ψ, $(M, s) \models (\mathbf{W}_i\varphi \wedge \neg\mathbf{W}_i\psi)$, $(M', s') \models (\neg\mathbf{W}_i\varphi \vee \mathbf{W}_i\psi)$, then $a \notin \{\texttt{put}, \texttt{drop}\}$
if $C(i', s') \neq C(i, s)$, then $i' \neq i$ or $a \in \{\texttt{select}\ \varphi\}$
if $\mathrm{Agenda}'(i', s') \neq \mathrm{Agenda}(i, s)$, then $i' \neq i$ or $a \in \{\texttt{commit\_to}\ \alpha, \texttt{uncommit}\ \alpha\}$

M_1 expresses that the truth of objective formulas is only affected by putting or dropping blocks, that putting or dropping blocks does not affect an agent's wishes, and that the set of selected wishes or the agenda of agent i can only be altered by agent i itself, and only by making a choice, or by doing an (un-)commitment, respectively.

Initial state s_0. In our initial situation, we have blocks that are situated as in Figure

1, determining completely $\pi(s_0)$. We also have the following initial constraints about Eve's selected wishes and agenda:

M_2 $C(e, s_0) = \emptyset$

M_3 $\text{Agenda}(e, s_0) = \emptyset$

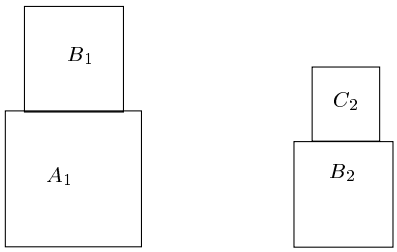

Figure 1: The initial block world

Since we only have four blocks here, we can define, for any blocks $X_1 \ldots X_4$ a predicate for towers as follows:

C_5 $\texttt{tower}\langle X_1, X_2, X_3, X_4\rangle \equiv$ $\bigwedge_{i\neq j}(X_i \neq X_j) \wedge \texttt{Floor}(X_1)$
$\wedge \texttt{is\_on}(X_1, X_2) \wedge \texttt{is\_on}(X_2, X_3) \wedge \texttt{is\_on}(X_3, X_4)$
$\wedge \texttt{is\_clear}(X_4))$

Additionally, let θ denote that $\texttt{tower}(X_1, X_2, X_3, X_4)$ is true, for some choice of the arguments from the blocks A_1, B_1, B_2 and C_2. Eve has two wishes: one is to have a tower ($\mathbf{W}_e\theta$) and the other is to have at least four blocks around ($\mathbf{W}_e\rho$). Since there are in principle 24 ways to stack the blocks into a tower, these comprise the worlds that are W-accessible from the initial state s. Determining the goals of Eve, recall that

$$\mathbf{Goal}_i\varphi =^{\text{def}} \mathbf{W}_i\varphi \wedge \neg\varphi \wedge \Diamond_i\varphi \wedge \mathbf{C}_i\varphi$$

Now, in the initial state s_0, we obviously have $\mathbf{W}_e\theta \wedge \neg\theta$. Also, ρ cannot be Eve's goal, since it is already fulfilled. This is also the reason that Eve is not able to make ρ a selected wish. Is she able to select θ? Which of the 24 a priori possibilities of building towers are implementable for Eve? Due to her (dis-)abilities recorded in the constraints A_1 and A_2, and since there is no opportunity to put larger blocks on smaller ones (O_1), there only remain two options: one is $\texttt{tower}(A_1, B_1, B_2, C_2)$, and the other is $\texttt{tower}(A_1, B_2, B_1, C_2)$. To see why θ is implementable by Eve, let us consider the plan

$$\alpha = \texttt{drop}(C_2); \texttt{put}(B_1, B_2); \texttt{put}(B_2, C_2)$$

We will also write α as $a_1; a_2; a_3$. We have to check whether in s_0, for our particular α and θ, $\mathbf{PracPoss}_e(\alpha, \theta)$ holds. This boils down to checking both $\langle \text{do}_e(\alpha)\rangle\theta$ and $\mathbf{A}_e\alpha$ in the initial state, which is an immediate consequence from the properties A_1, A_2, O_1, O_2, E_1 and E_2, applied to the initial configuration. Hence, we have $\Diamond_e\theta$. For θ to become a goal, Eve has to select it from her wishes. By using C_3 and C_4, we derive that $\mathbf{K}_e\mathbf{PracPoss}_e(\alpha, \theta)$ holds in s_0, i.e., $\mathbf{Can}_e(\alpha, \theta)$. In particular, this implies that Eve is able to choose θ. Let us assume Eve makes this choice, i.e, that she selects θ.

The next state s_1. Now, assuming that Eve made the choice θ, by the definition of the result of selecting (Definition 6.2), we know that s_1 only differs from s_0 in the set of

selected wishes $C(e, s_1)$ ascribed to s_1. Thus, in s_1, like in s_0, we have $\mathbf{Goal}_e\theta$, and, by property C_2, $\mathbf{K}_e\mathbf{Goal}_e\theta$. Thus, we have $\mathbf{Can}_e(\alpha, \theta) \wedge \mathbf{K}_e\mathbf{Goal}_e\theta$, which, by definition, entails $\mathbf{PossIntend}_e(\alpha, \theta)$. This, in its turn, implies that Eve has the opportunity to commit to the action α to fulfil her goal. Since we know from the constraints M_3 and M_1 that Eve's agenda is empty in s_1, Eve is also able to commit to α. Let us suppose that Eve decides to commit to the action α resulting in a state s_2.

The next state s_2. Now, $\text{Agenda}(e, s_2) = \{\alpha\}$. Let s_{21}, s_{22} and s_{23} be as follows: $\langle \alpha, s_2\rangle \rightharpoonup_{e,a_1} \langle (a_2; a_3), s_{21}\rangle \rightharpoonup_{e,a_2} \langle a_3, s_{22}\rangle \rightharpoonup_{e,a_3} \langle \Lambda, s_{23}\rangle$. Thus, we have $\text{Agenda}(e, s_{21}) = \{a_2; a_3\}$ and $\text{Agenda}(e, s_{22}) = \{a_3\}$ (under the assumption M_3).

As an effect, in s_2, Eve is committed to the actions a_1, $a_1; a_2$ and $a_1; a_2; a_3$, yielding, among others, $\mathbf{Committed}_e(a_1; a_2; a_3)$. Similar commitments hold for the states s_{2i}. Since the act of committing only influences the contents of the agenda, we have in s_2, as in s_1, that $\mathbf{K}_e\mathbf{Goal}_e\theta \wedge \mathbf{PossIntend}_e(\alpha, \theta)$. Let us now suppose that Eve executes a_1, i.e., she drops C_2 on the floor.

The state s_3. Here, by constraint E_2, $\texttt{is\_clear}(B_2)$ holds. Also, by M_1, Eve is still committed to a_2 and $a_2; a_3$. Hence, in principle, she can decide to perform a_2, possibly followed by a_3. Is θ still a goal? First of all, since the block B_2 is clear now, we have $\neg\theta$. Moreover, by constraint M_1, we still have $\mathbf{C}_e\theta \wedge \mathbf{W}_e\theta$, in s_3. Since the constraint N_5 on non-effects guarantees that dropping block C_2 does not affect $\texttt{is\_clear}(B_2)$, one readily checks that we also have $\mathbf{PracPoss}_e(a_2; a_3, \theta)$ so that, by using M_1, we indeed conclude $\mathbf{Goal}_i\theta$, in s_3. We can use constraints C_3 and C_4 to conclude $\mathbf{K}_e(\langle \text{do}_e(a_2; a_3)\rangle\theta \wedge \mathbf{A}_e(a_2; a_3))$, giving $\mathbf{Can}_e(a_2; a_3, \theta)$, so that, indeed, we have $\mathbf{PossIntend}_e(a_2; a_3, \theta)$. This implies that Eve has an opportunity to commit to doing $a_2; a_3$, but, since her agenda is already filled (with $a_2; a_3$), she is not able to commit again.

Of course, Eve can decide to execute a_2 now, leading her to state s_4. In a similar line of reasoning as above, she then can decide to perform a_3, leading her to a state s_5 in which her goal θ is fulfilled and, hence, will not be one of Eve's goals anymore.

Note how, to ensure that Eve's goal θ did persist while moving to state s_3, we had to use several persistence and non-effects assumptions. This seems reasonable: goals persist during execution of a plan to fulfil them, unless unexpected side-effects of actions occur. In our example, we ruled out the possible side-effects that may prevent θ to remain a goal while doing a_1, but in general, it may be too hard to sum up all these possible non-effects, and, moreover, one cannot always guarantee that they will not occur. In other words, we do *not* want to advocate a general property like

$$\mathbf{K}_i\mathbf{Goal}_i\varphi \rightarrow [\text{do}_i(\alpha)]\mathbf{K}_i\mathbf{Goal}_i\varphi \tag{1}$$

First of all, if (1) would be valid, then goals would never be achieved! Of course, to overcome this, we might weaken (1) to $\mathbf{K}_i\mathbf{Goal}_i\varphi \rightarrow [\text{do}_i(\alpha)](\mathbf{K}_i\mathbf{Goal}_i\varphi \vee \varphi)$, but another problem with (1) is that it does not link α to ways to achieve the goal: α may be a step in the 'wrong' direction for achieving φ, destroying the conditions constituting φ as a goal. Hence, a better way to guarantee persistence of goals might be the following:

$$(\mathbf{PossIntend}_i(\alpha; \beta, \varphi) \wedge \mathbf{Committed}_i\alpha; \beta) \rightarrow \langle \text{do}_i(\alpha)\rangle(\mathbf{PossIntend}_i(\beta, \varphi) \vee \varphi) \tag{2}$$

expressing that an intention, to which a commitment has been made, persists while executing the initial part of a plan to achieve it, unless it is achieved already. Property (2) does not hold in our framework: from $\mathbf{PossIntend}_i(\alpha;\beta,\varphi)$ one can derive $\mathbf{K}_i(\langle \mathrm{do}_i(\alpha;\beta)\rangle\varphi \wedge \mathbf{A}_i(\alpha;\beta)) \wedge \mathbf{K}_i\mathbf{Goal}_i\varphi$, which, in case α is deterministic, entails $\mathbf{K}_i\langle \mathrm{do}_i(\alpha)\rangle(\langle \mathrm{do}_i(\beta)\rangle\varphi \wedge \mathbf{A}_i\beta) \wedge \mathbf{K}_i\mathbf{Goal}_i\varphi$, but in our framework we do not guarantee that this knowledge of the effect of doing α remains there after α has been done. We believe that this is as it should be.

8. Summary and conclusions

In this paper we presented a formalisation of motivational attitudes, the attitudes that explain why agents act the way they do. This formalisation concerns operators both on the assertion level, where operators range over propositions, and on the practition level, where operators range over actions. An important feature of our formalisation is the attention paid to the acts associated with selecting between wishes and with (un)committing to actions. Starting from the primitive notion of wishes, we defined goals to be selected, unfulfilled, implementable wishes. Commitments may be made to actions that are known to be correct and feasible with respect to some goal and may be undone whenever the action to which an agent has committed itself has either become impossible or useless. Both the act of making, and the act of undoing commitments are formalised as model-transforming actions in our framework. The actions that an agent is committed to are recorded in its agenda in such a way that commitments are closed under prefix-taking and under practical identity, i.e. having identical computation runs. On the whole our formalisation is a rather expressive one, which tries to be faithful to a certain extent to both commonsense intuition and philosophical insights.

8.1. Future work

A first obvious extension would be to allow the agent (via the capability function) to put multiple actions (commitments) in its agenda. Of course, this would also call for a selection mechanism (sometimes called arbitration) for choosing committed actions to be executed in the line of the work of Bratman *et al.* [5] on the selection of plans.

A further, related extension to the framework presented here concerns a formalisation of the *actual execution* of actions. Although the conditional nature of a framework based on dynamic logic makes it perhaps less suitable for an adequate formalisation of 'doing', one could think of a practition operator indicating which action is actually performed next. Using this predicate would enhance expressiveness in that it would be possible to formulate relations between actions that agents are committed to, and actions that they actually perform.

Another way to extend the framework would be by establishing further relations with deontic notions like obligations and violations. A combination of the 'doing'-predicate with a deontic notion modelling violations or penalties would then allow one to model that agents should execute the actions that they are committed to if they want to avoid penalties. Research along these lines was initiated by Dignum & Van Linder [10, 11].

A very important extension that we will pursue is the extension to deal with multi-agent systems properly. In this case the agent's agenda should also contain some means to refer to (actions of) other agents, for instance requests for information or requests / commands for other agents to perform actions. At the moment we are looking at communicative aspects of agents expressed in agent (programming) languages [59, 58], in which primitives for requesting and sending information are incorporated. It will be interesting to see how this experience

can help us to give proper models (together with a logical specification language) for this crucial aspect.

Furthermore, we will investigate how the model as presented may serve as a rigorous semantic framework for agent-oriented languages such as AGENT0 [54], PLACA [57], and AgentSpeak(L) [45], so that, based on this model, specification methods for programs written in these languages can be obtained. A first step in this direction is the definition of the 'abstract' agent language 3APL [22], which on the one hand can be easily related to these other languages and, on the other hand, has programming constructs of the form $\pi \leftarrow \varphi \mid \pi'$, which, in our present terminology, express 'agenda maintenance': in a situation where φ holds, the agent may replace π by π' in its agenda of commitments. In case π' is empty, such a rule expresses the possibility of dropping a commitment (performance of an uncommitment) under certain circumstances (indicated by φ). In 3APL we have also considered rule (plan) selection and execution mechanisms, clearly related to the issues mentioned before. We plan to investigate the use of the logical formalism presented in this paper to specify agents written in the language 3APL.

8.2. Related work

The formalisation of motivational attitudes has received much attention within the agent research community. Probably the most influential account of motivational attitudes in AI is due to Cohen & Levesque [9], inspired by the conceptual groundwork of the philosopher M.E. Bratman [4]. Starting from the primitive notions of implicit goals and beliefs, Cohen & Levesque define so-called persistent goals, which are goals which agents give up only when they think they are either satisfied or will never be true, and intentions, both ranging over propositions and over actions. The idea underlying persistent goals is similar to that underlying our notion of goals. In the framework of Cohen & Levesque agents intend to bring about a proposition if they intend to do some action that brings about the proposition. An agent intends to do an action if it has the persistent goal to have done the action. This reduction of intentions to do actions for goals is a rather artificial and philosophically very questionable one: although intentions to actions should be related to goals, this relation should express that doing the action helps in bringing about some goal and not that doing the action in itself is a goal. Furthermore the coexistence of goals and intentions ranging over propositions seems to complicate matters unnecessarily.

As compared to the approach of Cohen & Levesque, we claim that our approach is more 'computational' in nature. At the basis of our theory we employ dynamic logic, a programming logic with explicit reference to actions (programs) within the language. Intentions are represented by commitments that have a very computational flavour, consisting of actions, and we employ concepts (like transition systems) from the realm of the semantics of programming.

Another important formalisation of motivational attitudes is proposed by Rao & Georgeff [47] by means of their BDI-logic(s). Treating desires and intentions as primitive, Rao & Georgeff focus on the process of intention revision rather than the 'commitment acquisition' which is essential to our formalisation. Another major difference is that BDI-logic rests on temporal logic rather than dynamic logic as in the case of our KARO-framework. Both desires and intentions in their framework suffer from the problems associated with logical omniscience. To avoid these problems, Cavedon *et al.* [8] propose the use of non-normal logics of intention and belief in the BDI-logic, and more in particular Rantala's 'impossible worlds' framework

[44]. This 'impossible worlds' approach was originally proposed as a way to solve the problems of logical omniscience for informational attitudes. Hence, whereas we more or less employ the awareness approach, Cavedon *et al.* propose yet another technique developed to solve the problems of logical omniscience. It therefore may come as no surprise that the properties that Cavedon *et al.* acquire for intentions are highly similar to the properties of goals given in Section 6.

Recently Rao & Georgeff [19] have also considered the *dynamics* of BDI-notions. Although they, of course, employ their BDI-logic to discuss the maintenance of beliefs, desires and intentions, this work is very close in spirit to ours. Focusing on *intention* maintenance we observe a very important difference, though. In [19] Rao & Georgeff mention the problem of logical omniscience again, which in the dynamic setting appears to have even graver consequences. For instance, if one intends φ, then if one uses a normal modal logic for intentions (such as Rao & Georgeff's BDI-logic), one obtains also that one intends $\varphi \vee \psi$, for arbitrary ψ. However, if intention φ is now dropped due to some reason (e.g. because it is not attainable for the agent), this will result (under some reasonable assumption of minimal change) in the intention of ψ, which is rather absurd! In order to solve this problem Rao & Georgeff introduce an "only intends" operator in a similar vein as Levesque's "only knows" operator [30]. However, in our framework we work with commitment (agenda) maintenance where *actions* are maintained or modified (by means of a $\texttt{commit\_to}\,\alpha$ operator) rather than *assertions*. In our set-up we thus do no have a problem with logical omniscience regarding the maintenance of commitments. Recording actions in one's agenda and reasoning about these has a very different logic than doing the same for assertions, or put more concisely, the logic of $\texttt{commit\_to}\,\alpha$ is completely different from that of intend as in BDI-logic, and does not suffer from the logical omniscience problem.[18]

We also like to mention here the work by Singh [55], which bears some resemblance to our approach. It also considers intended actions and provides a calculus to 'maintain the agent's agenda'. The main difference is again that his theory rests on temporal branching-time logic. (He also has some constructs *à la* dynamic logic in his language, but these, too, have an explicit temporal interpretation.) Furthermore, the emphasis in his work is put on exploring the intricacies of intentional notions as related to issues of nondeterminism, which we have not considered in this paper at all.

The last formalisation of motivational attitudes that we would like to mention is the one proposed by Dignum *et al.* [12]. In this formalisation, which is inspired by and based on research on deontic logic as carried out by Dignum *et al.*, notions like decisions, intentions and commitments are modelled. Of these, decisions and the act of committing are interpreted as so-called meta-actions, a notion similar to that of model-transformers. Despite its complexity, which is due to the incorporation of an algebraic semantics of actions and a trace semantics to model histories, some of the essential ideas underlying the formalisation of Dignum *et al.* are not unlike those underlying the formalisation presented here.

Acknowledgements. We like to thank Wieke de Vries for reading an earlier version of this paper and providing us with some very useful comments. We are also grateful to the at-

[18] In deontic logic, the logic of obligation and prohibition, one encounters a similar distinction with respect to e.g. the obligation operator: such an obligation operator with an action as argument ('ought-to-do') follows a entirely different logic than one with an assertion as argument ('ought-to-be'), cf. [39].

tendants of ATAL'95, Modelage'96 and the Workshop on Logic, Language and Computation 7 for their helpful response to preliminary versions of this work. Furthermore, the partial support of ESPRIT Working Group MODELAGE (BRWG 8319) is gratefully acknowledged. Finally we owe much to the anonymous referees who gave us useful suggestions for improvement, and, in particular, pointed us to the link with decision-theoretic aspects.

A. Selected proofs

5.3. PROPOSITION. *All of the properties of logical omniscience formalised in Definition 5.2, with the exception of LO7, are valid for the* $\mathbf{W}_i$ *operator.*

PROOF: Properties LO1 and LO2 state that $\mathbf{W}_i$ is a normal modal operator and are shown as for any necessity operator. Property LO3 follows directly by combining LO1 and LO2, and LO4 is a direct consequence of LO3. Properties LO5 and LO6 are typical for necessity operators in a normal modal logic: for whenever both φ and ψ hold at a set of designated worlds, $\varphi \wedge \psi$ also holds at all the worlds from that set (LO5), and if φ holds at all worlds from some set then $\varphi \vee \psi$ does also (LO6). That LO7 is not valid for the $\mathbf{W}_i$ operator is seen by considering a model M with state s such that no state s' exists with $(s, s') \in \mathrm{W}(i)$. Then it holds that $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{W}_i\varphi \wedge \mathbf{W}_i\neg\varphi$, for all $\varphi \in \mathrm{L}$.
⊠

6.7. PROPOSITION. *None of the properties of logical omniscience formalised in Definition 5.2, with the exception of LO7, is valid for the* $\mathbf{Goal}_i$ *operator.*

PROOF: Properties LO1, LO3, LO4, LO5 and LO6 are most easily seen not to hold for the goal operator by noting the absence of any closure properties on the set $\mathrm{C}(i, s)$, for $i \in \mathrm{A}$ and s some state. Due to this absence it is perfectly possible that φ and $\varphi \to \psi$ are both in $\mathrm{C}(i, s)$ while ψ is not (LO1), that $\varphi \in \mathrm{C}(i, s)$ and $\psi \notin \mathrm{C}(i, s)$ while $\models^{\mathrm{C}} \varphi \to \psi$ (LO3) or $\models^{\mathrm{C}} \varphi \leftrightarrow \psi$ (LO4), that $\{\varphi, \psi\} \subseteq \mathrm{C}(i, s)$ and $\varphi \wedge \psi \notin \mathrm{C}(i, s)$ (LO5), or that $\varphi \in \mathrm{C}(i, s)$ while $\varphi \vee \psi \notin \mathrm{C}(i, s)$ (LO6), for appropriate $i \in \mathrm{A}$ and s a state in some model. Property LO2 is seen not to hold by observing that $\models^{\mathrm{C}} \varphi$ implies that φ is fulfilled always and everywhere, which means that φ is not a goal. In fact, one can show that whenever φ is inevitable, i.e. $\models^{\mathrm{C}} \varphi$ holds, it is necessarily not a goal, i.e. $\models^{\mathrm{C}} \neg\mathbf{Goal}_i\varphi$ holds (cf. item 5 of Proposition 6.8). That LO7 holds for goals is a direct consequence of their unfulfilledness. For in any possible state s of any possible model M, either φ holds and thereby $\mathrm{M}, s \not\models^{\mathrm{C}} \mathbf{Goal}_i\varphi$, or $\neg\varphi$ holds and thereby $\mathrm{M}, s \not\models^{\mathrm{C}} \mathbf{Goal}_i\neg\varphi$. Hence LO7 is a valid property for goals.
⊠

6.8. PROPOSITION. *For all* $i \in \mathrm{A}$ *and* $\varphi \in \mathrm{L}$ *we have:*

1. $\models^{\mathrm{C}} \mathbf{W}_i\varphi \leftrightarrow \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\top$
2. $\models^{\mathrm{C}} \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\top \leftrightarrow \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\mathbf{C}_i\varphi$
3. $\models^{\mathrm{C}} \neg\mathbf{A}_i\mathtt{select}\,\varphi \to [\mathrm{do}_i(\mathtt{select}\,\varphi)]\neg\mathbf{Goal}_i\varphi$
4. $\models^{\mathrm{C}} \mathbf{PracPoss}_i(\mathtt{select}\,\varphi, \top) \leftrightarrow \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\mathbf{Goal}_i\varphi$
5. $\models^{\mathrm{C}} \varphi \Rightarrow \models^{\mathrm{C}} \neg\mathbf{Goal}_i\varphi$
6. $(\varphi \to \psi) \to (\mathbf{Goal}_i\varphi \to \mathbf{Goal}_i\psi)$ *is not for all* $\varphi, \psi \in \mathrm{L}$ *valid*
7. $\mathbf{K}_i(\varphi \to \psi) \to (\mathbf{Goal}_i\varphi \to \mathbf{Goal}_i\psi)$ *is not for all* $\varphi, \psi \in \mathrm{L}$ *valid*

PROOF: We successively show all items. Let $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ with state s and $\varphi \in \mathrm{L}$ be arbitrary.

1. An easy inspection of Definition 6.2 shows that $\mathbf{r}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s) = \emptyset$ iff $\mathrm{M}, s \not\models^{\mathrm{C}} \mathbf{W}_i\varphi$. Thus $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{W}_i\varphi \leftrightarrow \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\top$, which was to be shown.
2. If $\mathrm{M}', s = \mathbf{r}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s)$, then M' is such that $\mathrm{C}'(i, s)$ contains φ. Then by definition $\mathrm{M}', s \models^{\mathrm{C}} \mathbf{C}_i\varphi$, and thus $\mathrm{M}, s \models^{\mathrm{C}} \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\mathbf{C}_i\varphi$ if $\mathrm{M}, s \models^{\mathrm{C}} \langle \mathrm{do}_i(\mathtt{select}\,\varphi)\rangle\top$, which suffices to conclude item 2.
3. Suppose $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{A}_i\mathtt{select}\,\varphi$, i.e. $\mathrm{M}, s \models^{\mathrm{C}} \varphi \vee \neg\Diamond_i\varphi$. Now by definition, $\varphi \in \mathrm{L}$, and hence, by Proposition 6.4, $\mathrm{M}', s \models^{\mathrm{C}} \varphi$ if $\mathrm{M}, s \models^{\mathrm{C}} \varphi$ whenever $\mathrm{M}', s = \mathbf{r}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s)$. By Corollary 6.5 it follows that for M' as aforementioned holds that $\mathrm{M}', s \models^{\mathrm{C}} \neg\Diamond_i\varphi$ if $\mathrm{M}, s \models^{\mathrm{C}} \neg\Diamond_i\varphi$. Thus if $\mathrm{M}, s \models^{\mathrm{C}} \varphi \vee \neg\Diamond_i\varphi$ then it holds for $\mathrm{M}', s = \mathbf{r}^{\mathrm{C}}(i, \mathtt{select}\,\varphi)(\mathrm{M}, s)$ that $\mathrm{M}', s \models^{\mathrm{C}} \varphi \vee \neg\Diamond_i\varphi$. By definition it then directly follows that $\mathrm{M}', s \models^{\mathrm{C}} \neg\mathbf{Goal}_i\varphi$, and thus $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{A}_i\mathtt{select}\,\varphi \rightarrow [\mathrm{do}_i(\mathtt{select}\,\varphi)]\neg\mathbf{Goal}_i\varphi$, which was to be shown.
4. This item follows by combining item 2 of this proposition with Proposition 6.4 and Corollary 6.5.
5. If $\models^{\mathrm{C}} \varphi$ holds, then $\mathrm{M}, s \models^{\mathrm{C}} \varphi$ for all $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ with state s. Hence $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{Goal}_i\varphi$ for all $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ and their states s, and thus $\models^{\mathrm{C}} \neg\mathbf{Goal}_i\varphi$.
6. This item is easily shown by selecting an appropriate contingency φ and an arbitrary tautology ψ, such that for certain M and s holds that $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Goal}_i\varphi$. For then $\mathrm{M}, s \models^{\mathrm{C}} (\varphi \rightarrow \psi) \wedge \mathbf{Goal}_i\varphi$ while — by the previous item — $\mathrm{M}, s \not\models^{\mathrm{C}} \mathbf{Goal}_i\psi$.
7. Item 7 is proved similarly to item 6.

⊠

7.16. Proposition. *For all $i \in \mathrm{A}$, $\alpha, \beta \in \mathrm{Ac}$ and $\varphi \in \mathrm{L}$ we have:*

1. $\models^{\mathrm{C}} \mathbf{PossIntend}_i(\alpha, \varphi) \rightarrow \langle \mathrm{do}_i(\mathtt{commit\_to}\,\alpha)\rangle\top$
2. $\models^{\mathrm{C}} \langle \mathrm{do}_i(\mathtt{commit\_to}\,\alpha)\rangle\top \leftrightarrow \langle \mathrm{do}_i(\mathtt{commit\_to}\,\alpha)\rangle\mathbf{Committed}_i\alpha$
3. $\models^{\mathrm{C}} \mathbf{Committed}_i\alpha \rightarrow \neg\mathbf{A}_i\mathtt{commit\_to}\,\beta$
4. $\models^{\mathrm{C}} [\mathrm{do}_i(\mathtt{commit\_to}\,\alpha)]\neg\mathbf{A}_i\mathtt{commit\_to}\,\beta$
5. $\models^{\mathrm{C}} \mathbf{Committed}_i\alpha \leftrightarrow \langle \mathrm{do}_i(\mathtt{uncommit}\,\alpha)\rangle\neg\mathbf{Committed}_i\alpha$
6. $\models^{\mathrm{C}} \mathbf{PossIntend}_i(\alpha, \varphi) \rightarrow \neg\mathbf{A}_i\mathtt{uncommit}\,\alpha$
7. $\models^{\mathrm{C}} (\mathbf{C}_i\varphi \leftrightarrow \mathbf{K}_i\mathbf{C}_i\varphi) \rightarrow (\mathbf{A}_i\mathtt{uncommit}\,\alpha \leftrightarrow \mathbf{K}_i\mathbf{A}_i\mathtt{uncommit}\,\alpha)$
8. $\models^{\mathrm{C}} \mathbf{Committed}_i\alpha \wedge \neg\mathbf{Can}_i(\alpha, \top) \rightarrow \mathbf{Can}_i(\mathtt{uncommit}\,\alpha, \neg\mathbf{Committed}_i\alpha)$

Proof: We show the second, third, fourth, seventh and eight item; the other ones follow directly from the respective definitions. Let $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ with state s, and $i \in \mathrm{A}$, $\alpha, \beta \in \mathrm{Ac}$ be arbitrary.

2. Let $\mathrm{M}, s \models^{\mathrm{C}} \langle \mathrm{do}_i(\mathtt{commit\_to}\,\alpha)\rangle\top$ and let $\mathrm{M}', s = \mathbf{r}^{\mathrm{C}}(i, \mathtt{commit\_to}\,\alpha)(\mathrm{M}, s)$. We have to show that $\mathrm{M}', s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha$, i.e. we have to show that $\forall s' \in [s]_{\mathrm{R}'(i)} \exists\alpha_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}'}(i, \alpha, s') \exists\alpha_2 \in \mathrm{Agenda}'(i, s') \exists\alpha_2' \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha_2, s')(\mathrm{Prefix}(\alpha_1), \alpha_2'))$. An inspection of Definition 7.9 shows that for all $s' \in [s]_{\mathrm{R}(i)} = [s]_{\mathrm{R}'(i)}$ holds that $\mathrm{Agenda}'(i, s')$ contains α. This implies that $\mathrm{M}', s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha$. Thus $\mathrm{M}, s \models^{\mathrm{C}} \langle \mathrm{do}_i(\mathtt{commit\_to}\,\alpha)\rangle \mathbf{Committed}_i\alpha$, which suffices to conclude that item 2 holds.
3. If $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha$ then, by Definition 7.12, we have that $\mathrm{Agenda}(i, s) \neq \emptyset$. Hence, by Definition 7.9, $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{A}_i\mathtt{commit\_to}\,\beta$.
4. If $\mathbf{r}^{\mathrm{C}}(i, \mathtt{commit\_to}\,\alpha)(\mathrm{M}, s) = \emptyset$ then $\mathrm{M}, s \models^{\mathrm{C}} [\mathrm{do}_i(\mathtt{commit\_to}\,\alpha)]\neg\mathbf{A}_i\mathtt{commit\_to}\,\beta$ is trivially true. Else $\mathrm{M}, s \models^{\mathrm{C}} \langle \mathrm{do}_i(\mathtt{commit\_to}\,\alpha)\rangle\mathbf{Committed}_i\alpha$ by item 2 of this proposition, and, by item 3, this implies $\mathrm{M}, s \models^{\mathrm{C}} \langle \mathrm{do}_i(\mathtt{commit\_to}\,\alpha)\rangle\neg\mathbf{A}_i\mathtt{commit\_to}\,\beta$, which suffices to conclude item 4.

7. Suppose $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{C}_i\varphi \leftrightarrow \mathbf{K}_i\mathbf{C}_i\varphi$ and $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{A}_i\mathtt{uncommit}\,\alpha$. This implies that $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{PossIntend}_i(\alpha, \varphi)$, for all $\varphi \in \mathrm{C}(i, s)$. That is, $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{Can}_i(\alpha, \varphi) \vee \neg\mathbf{K}_i\mathbf{Goal}_i\varphi$ for all $\varphi \in \mathrm{C}(i, s)$. But by the introspective properties of knowledge the latter implies that $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{K}_i\neg\mathbf{Can}_i(\alpha, \varphi) \vee \mathbf{K}_i\neg\mathbf{K}_i\mathbf{Goal}_i\varphi$, for all $\varphi \in \mathrm{C}(i, s)$. Hence $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{K}_i(\neg\mathbf{Can}_i(\alpha, \varphi) \vee \neg\mathbf{K}_i\mathbf{Goal}_i\varphi)$, for all $\varphi \in \mathrm{C}(i, s)$, and thus for all $s' \in [s]_{\mathrm{R}(i)}$ it holds that $\mathrm{M}, s' \models^{\mathrm{C}} \neg\mathbf{PossIntend}_i(\alpha, \varphi)$ for all $\varphi \in \mathrm{C}(i, s)$, i.e. for all φ such that $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{C}_i\varphi$. By the assumption $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{C}_i\varphi \leftrightarrow \mathbf{K}_i\mathbf{C}_i\varphi$, we obtain that $\mathrm{M}, s' \models^{\mathrm{C}} \neg\mathbf{PossIntend}_i(\alpha, \varphi)$ for all φ such that $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{K}_i\mathbf{C}_i\varphi$, i.e. for all $\varphi \in \mathrm{C}(i, s')$ for all $s' \in [s]_{\mathrm{R}(i)}$. Consequently, $\mathrm{M}, s' \models^{\mathrm{C}} \mathbf{A}_i\mathtt{uncommit}\,\alpha$. Thus $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{K}_i\mathbf{A}_i\mathtt{uncommit}\,\alpha$, which suffices to conclude that item 7 indeed holds.
8. Suppose $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha \wedge \neg\mathbf{Can}_i(\alpha, \top)$. Then $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{Can}_i(\alpha, \varphi)$ for all $\varphi \in \mathrm{L}$, and thus $\mathrm{M}, s \models^{\mathrm{C}} \neg\mathbf{PossIntend}_i(\alpha, \varphi)$ for all $\varphi \in \mathrm{C}(i, s)$. Then, by definition of $\mathtt{c}^{\mathrm{C}}$, $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{A}_i\mathtt{uncommit}\,\alpha$, and, by the previous item, $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{K}_i\mathbf{A}_i\mathtt{uncommit}\,\alpha$. Also, $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha$ implies $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{K}_i\mathbf{Committed}_i\alpha$ by Proposition 7.17(1), and, by item 5 of this proposition, $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{K}_i\langle\mathrm{do}_i(\mathtt{uncommit}\,\alpha)\rangle\neg\mathbf{Committed}_i\alpha$. Thus $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{K}_i\langle\mathrm{do}_i(\mathtt{uncommit}\,\alpha)\rangle\neg\mathbf{Committed}_i\alpha \wedge \mathbf{K}_i\mathbf{A}_i\mathtt{uncommit}\,\alpha$. This implies that $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Can}_i(\mathtt{uncommit}\,\alpha, \neg\mathbf{Committed}_i\alpha)$, which suffices to conclude item 8.

⊠

7.17. Proposition. *For all* $i \in \mathrm{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}$ *and all* $\varphi \in \mathrm{L}$ *we have:*

1. $\models^{\mathrm{C}} \mathbf{Committed}_i\alpha \rightarrow \mathbf{K}_i\mathbf{Committed}_i\alpha$
2. $\models^{\mathrm{C}} \mathbf{Committed}_i(\alpha_1; \alpha_2) \rightarrow \mathbf{Committed}_i\alpha_1 \wedge \mathbf{K}_i[\mathrm{do}_i(\alpha_1)]\mathbf{Committed}_i\alpha_2$
3. $\models^{\mathrm{C}} \mathbf{Committed}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi} \wedge \mathbf{K}_i\varphi \rightarrow \mathbf{Committed}_i(\varphi?; \alpha_1)$
4. $\models^{\mathrm{C}} \mathbf{Committed}_i\mathtt{if}\,\varphi\,\mathtt{then}\,\alpha_1\,\mathtt{else}\,\alpha_2\,\mathtt{fi} \wedge \mathbf{K}_i\neg\varphi \rightarrow \mathbf{Committed}_i(\neg\varphi?; \alpha_2)$
5. $\models^{\mathrm{C}}\mathbf{Committed}_i\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od} \wedge \mathbf{K}_i\varphi \rightarrow \mathbf{Committed}_i((\varphi?; \alpha); \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od})$

Proof: We successively show all items. Let $\mathrm{M} \in \mathbf{M}^{\mathrm{C}}$ with state s and $\varphi \in \mathrm{L}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}$ be arbitrary.

1. Assume that $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha$. By Definition 7.12 it then follows that $\forall s' \in [s]_{\mathrm{R}(i)}\exists\alpha_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha, s')\exists\alpha_2 \in \mathrm{Agenda}(i, s')\exists\alpha_2' \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha_2, s')(\mathrm{Prefix}(\alpha_1, \alpha_2'))$. Since $[s]_{\mathrm{R}(i)}$ is an equivalence class we have that $\forall s'' \in [s]_{\mathrm{R}(i)}\forall s' \in [s'']_{\mathrm{R}(i)}\exists\alpha_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha, s')\exists\alpha_2 \in \mathrm{Agenda}(i, s')\exists\alpha_2' \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha_2, s')(\mathrm{Prefix}(\alpha_1, \alpha_2'))$, which implies $\mathrm{M}, s'' \models^{\mathrm{C}} \mathbf{Committed}_i\alpha$ for all $s'' \in [s]_{\mathrm{R}(i)}$, and thus $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{K}_i\mathbf{Committed}_i\alpha$.
2. Let $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha_1; \alpha_2$, i.e. for all $s' \in [s]_{\mathrm{R}(i)}$ some $\beta_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha_1; \alpha_2, s')$, $\beta_2 \in \mathrm{Agenda}(i, s')$ and $\beta_2' \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \beta_2, s')$ exist such that $\mathrm{Prefix}(\beta_1, \beta_2')$ holds. But then also for all $s' \in [s]_{\mathrm{R}(i)}$ some $\gamma_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha_1, s')$, $\gamma_2 \in \mathrm{Agenda}(i, s')$ and $\gamma_2' \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \gamma_2, s')$ exist such that $\mathrm{Prefix}(\gamma_1, \gamma_2')$ holds (viz. $\gamma_2 = \beta_2$ and $\gamma_2' = \beta_2'$ satisfy the requirement, since $\mathrm{Prefix}(\beta_1, \gamma_1)$). Thus $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Committed}_i\alpha_1$.

   Furthermore, consider some state $s' \in [s]_{\mathrm{R}(i)}$ and some state $s'' \in \mathtt{r}^{\mathrm{C}}(i, \alpha_1)(\mathrm{M}, s')$. By the above we obtain that there exist some $\beta_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha_1; \alpha_2, s')$, $\beta_2 \in \mathrm{Agenda}(i, s')$ and $\beta_2' \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \beta_2, s')$ exist such that $\mathrm{Prefix}(\beta_1, \beta_2')$ holds. By Proposition 7.3 we have that $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \beta_1, s') = \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha_1, s'); \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha_2, s'')$. Suppose $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha_1, s') = \{b_1; b_2; \ldots; b_m\}$. Since $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \beta_2, s') = \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha_1; \alpha_2, s'); \gamma = \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha_1, s'); \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha_2, s''); \gamma = b_1; b_2; \ldots; b_m; \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha_2, s''); \gamma$, for some $\gamma \in \mathrm{Acseq}$, we have by Proposition 7.5 that $\langle s', \beta_2\rangle \rightharpoonup_{b_1} \ldots \rightharpoonup_{b_m} \langle s'', \delta_2\rangle$, with δ_2 such that $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \delta_2, s'') = \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha_2, s''); \gamma$. This implies that $\delta_2 \in \mathrm{Agenda}(i, s''')$ for all $s''' \in [s'']_{\mathrm{R}(i)}$. Thus we have that, for all $s''' \in [s'']_{\mathrm{R}(i)}$, there exist $\delta_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \alpha_2, s''')$, $\delta_2 \in \mathrm{Agenda}(i, s''')$ and $\delta_2' \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \delta_2, s''')$ such that $\mathrm{Prefix}(\delta_1, \delta_2')$.

3. Assume that $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Committed}_i \mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1\ \mathtt{else}\ \alpha_2\ \mathtt{fi} \wedge \mathbf{K}_i\varphi$. By definition of $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}$ and CS we have $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1\ \mathtt{else}\ \alpha_2\ \mathtt{fi}, s') = \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \varphi?; \alpha_1, s')$ for all $s' \in [s]_{\mathrm{R}(i)}$. Hence it follows that $\exists\beta_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1\ \mathtt{else}\ \alpha_2\ \mathtt{fi}, s')\exists\beta_2 \in \mathrm{Agenda}(i, s')\exists\beta_2' \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \beta_2, s')(\mathrm{Prefix}(\beta_1, \beta_2'))$ implies $\exists\beta_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, (\varphi?; \alpha_1), s')$ $\exists\beta_2 \in \mathrm{Agenda}(i, s')\exists\beta_2' \in \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \beta_2, s')(\mathrm{Prefix}(\beta_1, \beta_2'))$ for all $s' \in [s]_{\mathrm{R}(i)}$. Then it is indeed the case that from $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Committed}_i \mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1\ \mathtt{else}\ \alpha_2\ \mathtt{fi}$ it follows that $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Committed}_i(\varphi?; \alpha_1)$, which suffices to conclude this item.
4. This item is completely analogous to the previous one.
5. From the definition of $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}$ and CS it follows that in the case that $\mathrm{M}, s \models^{\mathrm{C}} \varphi$, $\mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, \mathtt{while}\ \varphi\ \mathtt{do}\ \alpha\ \mathtt{od}, s) = \mathrm{CR}^{\mathrm{C}}_{\mathrm{M}}(i, (\varphi?; \alpha); \mathtt{while}\ \varphi\ \mathtt{do}\ \alpha\ \mathtt{od}, s)$. By a similar argument as the one given in the proof of item 3 one concludes that $\mathrm{M}, s \models^{\mathrm{C}} \mathbf{Committed}_i \mathtt{while}\ \varphi\ \mathtt{do}\ \alpha\ \mathtt{od} \wedge \mathbf{K}_i\varphi \rightarrow \mathbf{Committed}_i((\varphi?; \alpha); \mathtt{while}\ \varphi\ \mathtt{do}\ \alpha\ \mathtt{od})$, which concludes item 5.

⊠

References

[1] L. Åqvist. Deontic logic. In D.M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 2, chapter 11, pages 605–714. D. Reidel, Dordrecht, 1984.

[2] R.L. Atkinson, R.C. Atkinson, E.E. Smith, D.J. Bem, and S. Nolen-Hoeksema. *Hilgard's Introduction to Psychology, 12th edition.* Harcourt Brace College Publishers, Fort Worth, 1996.

[3] C. Boutilier. Toward a logic for qualitative decision theory. In P. Torasso J. Doyle, E Sandewall, editor, *Proceedings KR'94*, pages 75–86, 1994.

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

[5] M.E. Bratman, D. Israel, and M.E. Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4:349–355, 1988.

[6] H.-N. Castañeda. The paradoxes of deontic logic: the simplest solution to all of them in one fell swoop. In Risto Hilpinen, editor, *New Studies in Deontic Logic*, pages 37–85. Reidel, Dordrecht, 1981.

[7] C. Castelfranchi, D. D'Aloisi, and F. Giacomelli. A framework for dealing with belief-goal dynamics. In M. Gori and G. Soda, editors, *Topics in Artificial Intelligence*, volume 992 of *Lecture Notes in Computer Science (subseries LNAI)*, pages 237–242. Springer-Verlag, 1995.

[8] L. Cavedon, L. Padgham, A. Rao, and E. Sonenberg. Revisiting rationality for agents with intentions. In X. Yao, editor, *Bridging the Gap: Proceedings of the Eight Australian Joint Conference on Artificial Intelligence*, pages 131–138. World Scientific, 1995.

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

[10] F. Dignum and B. van Linder. Modelling rational agents in a dynamic environment: Putting Humpty Dumpty together again. In J.L. Fiadeiro and P.-Y. Schobbens, editors, *Proceedings of the 2nd Workshop of the ModelAge Project*, pages 81–91, 1996.

[11] F. Dignum and B. van Linder. Modelling social agents: Communication as action. In J.P. Müller, M.J. Wooldridge, and N.R. Jennings, editors, *Intelligent Agents Volume III – Agent Theories, Architectures, and Languages (ATAL'97)*, pages 205–218. Springer, LNCS 1193, Berlin, 1997.

[12] F. Dignum, J.-J.Ch. Meyer, R.J. Wieringa, and R. Kuiper. A modal approach to intentions, commitments and obligations: Intention plus commitment yields obligation. In M.A. Brown and J. Carmo, editors, *Deontic Logic, Agency and Normative Systems*, Springer Workshops in Computing, pages 80–97. Springer-Verlag, 1996.

[13] Jon Doyle, Yoav Shoham, and Michael P. Wellman. A logic for relative desire. In Z.W. Ras and M. Zemankova, editors, *Proceedings of the Sixth International Symposium on Methodologies for Intelligents Systems (ISMIS'91)*, pages 16–31. Springer Verlag, 1991.

[14] R. Fagin and J.Y. Halpern. Belief, awareness and limited reasoning. *Artificial Intelligence*, 34:39–76, 1988.

[15] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, MA, 1995.

[16] R. Fikes and N.J. Nilsson. Strips: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 3:189–208, 1971.

[17] G. Gazdar, G. Pullum, R. Carpenter, E. Klein, T. Hukari, and R. Levine. Category structures. *Computational Linguistics*, 14:1–19, 1988.

[18] M. Georgeff and A. Rao. Rational software agents: From theory to practice. In N.R. Jennings and M.J. Wooldridge, editors, *Agent Technology: Foundations, Applications, and Markets*, pages 139–160. Springer, Berlin, 1998.

[19] M.P. Georgeff and A.S. Rao. The semantics of intention maintenance for rational agents. In *Proceedings of 14th Int. Joint Conference on Artificial Intelligence (IJCAI–95)*, pages 704–710, Montreal, Canada, 1995.

[20] J.Y. Halpern and Y. Moses. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, 54:319–379, 1992.

[21] D. Harel. Dynamic logic. In D.M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 2, chapter 10, pages 497–604. D. Reidel, Dordrecht, 1984.

[22] K.V. Hindriks, F.S. de Boer, W. van der Hoek, and J.-J.Ch. Meyer. Formal semantics for an abstract agent programming language. In M.P. Singh, A. Rao, and M.J. Wooldridge, editors, *Intelligent Agents IV*, volume 1365 of *LNAI*, pages 215–229. Springer, Berlin, 1998.

[23] J. Hintikka. *Knowledge and Belief*. Cornell University Press, Ithaca, NY, 1962.

[24] W. van der Hoek, B. van Linder, and J.-J. Ch. Meyer. A logic of capabilities. In A. Nerode and Yu. V. Matiyasevich, editors, *Proceedings of the Third International Symposium on the Logical Foundations of Computer Science (LFCS'94)*, volume 813 of *Lecture Notes in Computer Science*, pages 366–378. Springer-Verlag, 1994.

[25] W. van der Hoek, B. van Linder, and J.-J. Ch. Meyer. Unravelling nondeterminism: On having the ability to choose (extended abstract). In P. Jorrand and V. Sgurev, editors, *Proceedings of the Sixth International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA'94)*, pages 163–172. World Scientific, 1994.

[26] A. Kenny. *Will, Freedom and Power*. Basil Blackwell, Oxford, 1975.

[27] K. Konolige. *A Deduction Model of Belief*. Pitman / Morgan Kaufmann, London / Los Altos, 1986.

[28] Y. Lespérance, H. Levesque, F. Lin, D. Marcu, R. Reiter, and R. Scherl. Foundations of a logical approach to agent programming. In M. Wooldridge, J.P. Müller, and M. Tambe, editors, *Intelligent Agents Volume II – Agent Theories, Architectures, and Languages*, volume 1037 of *Lecture Notes in Computer Science (subseries LNAI)*, pages 331–347. Springer-Verlag, 1996.

[29] H. Levesque. A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence (AAAI'84)*, pages 198–202. The AAAI Press/The MIT Press, 1984.

[30] H.J. Levesque. All I know: A study in autoepistemic logic. *Artificial Intelligence*, 42:263–309, 1990.

[31] F. Lin and R. Reiter. State constraints revisited. *Journal of Logic and Computation, Special Issue on Actions and Processes*, 1994.

[32] B. van Linder, W. van der Hoek, and J.-J. Ch. Meyer. Tests as epistemic updates. In A.G. Cohn, editor, *Proceedings of the 11th European Conference on Artificial Intelligence (ECAI'94)*, pages 331–335. John Wiley & Sons, 1994.

[33] B. van Linder, W. van der Hoek, and J.-J. Ch. Meyer. Actions that make you change your mind. In A. Laux and H. Wansing, editors, *Knowledge and Belief in Philosophy and Artificial Intelligence*, pages 103–146. Akademie Verlag, 1995.

[34] B. van Linder, W. van der Hoek, and J.-J. Ch. Meyer. The dynamics of default reasoning. *Data and Knowledge Engineering*, 21(3):317–346, 1997.

[35] B. van Linder, W. van der Hoek, and J.-J. Ch. Meyer. Seeing is believing (and so are hearing and jumping). *Journal of Logic, Language and Information*, 6(2):33–61, 1997.

[36] B. van Linder, W. van der Hoek, and J.-J.Ch. Meyer. Formalising motivational attitudes of agents: On preferences, goals and commitments. In M. Wooldridge, J.P. Mueller, and M. Tambe, editors, *Intelligent Agents Volume II – Agent Theories, Architectures, and Languages (ATAL'95)*, pages 17–32. Springer, LNCS 1037, Berlin, 1996.

[37] B. van Linder, W. van der Hoek, and J.-J.Ch. Meyer. Formalizing abilities and opportunities of agents. *Fundamenta Informaticae*, 34(1,2):53–101, 1998.

[38] J.-J. Ch. Meyer and W. van der Hoek. *Epistemic Logic for AI and Computer Science*. Cambridge University Press, Cambridge, UK, 1995.

[39] J.-J. Ch. Meyer and R.J. Wieringa. Deontic logic: A concise overview. In J.-J. Ch. Meyer and R.J. Wieringa, editors, *Deontic Logic in Computer Science*, chapter 1, pages 3–16. John Wiley & Sons, 1993.

[40] R.C. Moore. Reasoning about knowledge and action. Technical Report 191, SRI International, 1980.

[41] L. Morgenstern. A first order theory of planning, knowledge, and action. In *Proceedings of the 1st Conference on Theoretical Aspects of Reasoning about Knowledge (TARK86)*, pages 99–114, 1986.

[42] L. Morgenstern. Knowledge preconditions for actions and plans. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence (IJCAI-87)*, pages 867–874, Milan, Italy, 1987.

[43] G. Plotkin. A structural approach to operational semantics. Technical Report DAIME FN-19, Aarhus University, 1981.

[44] V. Rantala. Impossible worlds semantics and logical omniscience. *Acta Philosophica Fennica*, 35:106–115, 1982.

[45] A.S. Rao. AgentSpeak(L): BDI agents speak out in a logical computable language. In W. Van de Velde and J.W. Perram, editors, *Agents Breaking Away (Proc. MAAMAW'96)*, volume 1038 of *LNAI*, pages 42–55. Springer, Berlin, 1996.

[46] 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 (IJCAI'91)*, pages 498–504. Morgan Kaufmann, 1991.

[47] A.S. Rao and M.P. Georgeff. Modeling rational agents within a BDI-architecture. In J. Allen, R. Fikes, and E. Sandewall, editors, *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 473–484. Morgan Kaufmann, 1991.

[48] J. Raz, editor. *Practical Reasoning*. Oxford Readings in Philosophy. Oxford University Press, 1978.

[49] R. Reiter. Proving properties of states in the situation calculus. *Artificial Intelligence*, 64:337–351, 1993.

[50] S.J. Rosenschein. Formal theories of AI in knowledge and robotics. *New Generation Computing*, 3:345–357, 1985.

[51] S.J. Rosenschein and L.P. Kaelbling. The synthesis of digital machines with provable epistemic properties. In *Proceedings of Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 83–86, Los Altos, CA, 1986. Morgan Kaufmann.

[52] A. Ross. Imperatives and logic. *Theoria*, 7:53–71, 1941.

[53] E. Sandewall and Y. Shoham. Non-monotonic temporal reasoning. In C.J. Hogger D.M. Gabbay and J.A. Robinson, editors, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 4, pages 439–498. Clarendon Press, Oxford, 1995.

[54] Y. Shoham. Agent-oriented programming. *Artificial Intelligence*, 60(1):51–92, 1993.

[55] M.P. Singh. *Multiagent Systems*, volume 799 of *LNAI*. Springer-Verlag, Berlin-Heidelberg, 1994.

[56] E. Spaan. *Complexity of Modal Logics*. PhD thesis, Universiteit van Amsterdam, 1993.

[57] S.R. Thomas. *PLACA, an Agent-Oriented Programming Language*. PhD thesis, Stanford University, 1993.

[58] R.M. van Eijk, F.S. de Boer, W. van der Hoek, and J.-J.Ch. Meyer. Systems of communicating agents. In *Proceedings of 13th European Conference on Artificial Intelligence (ECAI–98)*, pages 293–297, Brighton, UK, 1998.

[59] R.M. van Eijk, F.S. de Boer, W. van der Hoek, and J.-J.Ch. Meyer. Information-passing and belief revision in multi-agent systems. In *Proceedings of Agent Theories, Architectures and Languages (ATAL'98)*, Berlin, 1999. Springer, to appear.

[60] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, 1944.

[61] Michael P. Wellman and Jon Doyle. Preferential semantics for goals. In *Proceedings of the Ninth National Conference on Artificial Intelligence, AAAI-91*, pages 698–703, 1991.

[62] M. Wooldridge and N. R. Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995.

[63] G.H. von Wright. *Norm and Action*. Routledge & Kegan Paul, London, 1963.

[64] G.H. von Wright. On so-called practical inference. In J. Raz, editor, *Practical Reasoning*, chapter III, pages 46–62. Oxford University Press, 1978.