An Integrated Modal Approach to Rational Agents *

W. van der Hoek B. van Linder† J.-J. Ch. Meyer
Utrecht University – Department of Computer Science
P.O. Box 80.089 – 3508 TB Utrecht – The Netherlands

Abstract

In this paper we give an overview of work we have done to provide a framework in which many aspects of rational agency are integrated. The various attitudes of a rational agent, viz. the informational as well as the motivational ones, are modelled in the framework by means of a variety of modal operators that are interpreted by means of possible worlds, as usual in modal logic. A main point here is that we incorporate all these modal operators into one model, so that in principle the various modal operators can be mixed to describe an agent's complex attitudes.

1 Introduction

We present a single formal framework for describing crucial aspects of rational or intelligent agents. Before proposing such a formal framework, let us try and distinguish some of its necessary features in a more informal way. An essential feature of being an *agent* is that it has access to a variety of *actions*; by using the adjective *rational*, we mean that the agent is able to *reason*. Obviously there is an interesting interaction between these two features: in our view, a vital feature that links the two mentioned above is that of *information*.

Information is a key issue when describing an agent's existence and behaviour in its environment, the world. On the one hand, information about this world is received by the agent, and it can use this, together with its specific reasoning mechanism, to draw certain conclusions (about the world). Based on these conclusions, it may decide to undertake some specific actions. These actions, on the other hand, may change the world itself, and thereby the information that can be obtained about it.

We may distinguish two aspects or levels regarding information: the *declarative* or *static* level of information, at which the agent may state that he knows,

*This work was partially supported by ESPRIT III BRWG No. 8319 (ModelAge)
†Currently at Philips Research Laboratories, Eindhoven

believes or doubts a certain piece of information. But there is also an *operational* or *dynamic* aspect to information: an agent may perform actions that provide it with (new) information about the world: it may, for instance, perform observations, or communicate with other agents.

Thus, by performing particular actions, the agent may acquire more information on which it can base its reasoning. Then, how does this reasoning lead it to the performance of new actions? We think (and many with us) that a rational agent is *directed* towards certain actions more than other ones. And, by using the idea that an action is aimed at changing the world, we even adhere to the principle that the agent considers certain states of the world more preferable than other states. In order to approximate such states, the agent must be able to reason about complex actions, and *choose* some actions as candidates for a plan. Again, here we can distinguish a declarative or static level, at which it can for instance be expressed that the agent is *committed* (in the sense of being in a state of 'committedness') to some action, and an operational or dynamic level at which the agent can perform an action like committing itself to do something.

But a rational agent should also be aware of the *limits* and *constraints* of certain actions. These limits may stem from the world itself: an agent does not always have the *opportunity* to perform an action. But other constraints are the consequence of the agent's very own design: it may simply not be equipped with certain *abilities*.

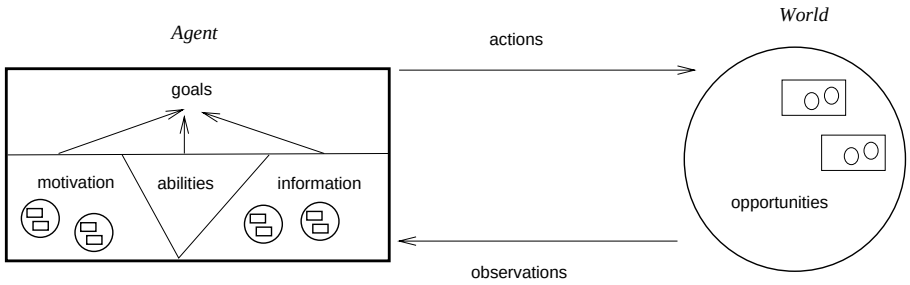

Figure 1: The agent and the world

Figure 1 depicts our framework of agents in a nutshell. Depending on the agent's information, abilities and motivation, it can formulate certain goals. This generally leads to performing certain actions, which have their impact on the world. Since the agent itself is part of the world, some of its actions may also change its own state, like the act of 'committing' that takes it to a state of 'being committed'. Thus, one of the agents in the world of agent a is a itself. The figure also shows samples of the world in the agent's informational

and motivational state: in the former state they represent the possible ways the world looks like on the basis of the agent's information, in the latter state they denote possible desired states of the world.

To express all these aspects of agency we will employ modal logic with Kripke-style possible world semantics. In fact, we will use a blend of dynamic, epistemic and doxastic logic, which we extend with certain elements (operators and actions) to suit our purpose, viz. giving a proper description of the informational and motivational attitudes of agents.

This chapter is organized as follows. In Section 2 we introduce the basic framework. Here we shall discuss the language that we will use as well as general considerations about its semantics. In Section 3 we will concentrate on the instantiation of the agent's informational attitudes, while in Section 4 we treat the motivational attitudes of an agent. We conclude the paper with some remarks about related work and an indication of further topics in Section 5.

2 The Framework

In this section we present a core language that is rich enough to reason about some of the agent's attitudes mentioned above–other operators will be defined on top of this– and indicate formal models for this language. Doing so, we will try not to lose ourselves in technical details (these can be found elsewhere, [13, 24, 23, 25]), but rather provide the reader with an intuitive grasp for the ideas underlying our formal definitions.

2.1 Language

The language $\mathcal{L}$ that we use to formalise these notions is based on a fixed set of propositional atoms, and the connectives $\wedge, \vee, \rightarrow, \neg$ to build formulas $\varphi, \psi, \ldots$ with their usual meaning; the canonical tautology $\top$ is defined to be $p \vee \neg p$ for p some arbitrary propositional atom, and the canonical contradiction $\bot$ is defined to be $\neg\top$. We denote the pure propositional language with $\mathcal{L}_0$. We extend this core language to deal with actions and epistemic and motivational attitudes.

Let i be a variable over a set of agents $\{1, \ldots, n\}$. Actions in the set Ac are either atomic actions (At $= \{a, b, \ldots\}$) or composed ($\alpha, \beta, \ldots$) by means of confirmation of formulas (`confirm` φ), sequencing ($\alpha; \beta$), conditioning (`if` φ `then` α `else` β) and repetition (`while` φ `do` α). (Actions that are either atomic or `confirm` actions will be called *semi-atomic* in this paper.) These actions α can then be used to build new formulas to express the possible *result* of the execution of α by agent i (the formula $\langle do_i(\alpha)\rangle\varphi$ denotes that φ is a result of i's execution of α), the *opportunity* for i to perform α ($\langle do_i(\alpha)\rangle\top$) and i's *capability* of performing the action α ($\mathbf{A}_i\alpha$). The formula $[do_i(\alpha)]\varphi$ is shorthand for $\neg\langle do_i(\alpha)\rangle\neg\varphi$, thus expressing that all possible results of performance of α by i imply φ, thereby being non-committal about the agent's opportunity to

perform α. We shall refer to the language built out of the constructs mentioned in this paragraph as the *core language*, denoted as $\mathcal{L}_c$. So $\mathcal{L}_c \subseteq \mathcal{L}$.

In order to successfully complete an action, both the opportunity and the ability to perform the action are necessary (we call this combination the *practical possibility* to do the action). 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. [20]). To formalise the knowledge of agents on their practical (im)possibilities, we introduce 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.

Definition 2.1 The Can-predicate and the Cannot-predicate are, for all agents i, actions α and formulae φ, defined as follows in terms of practical possibilities.

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

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.

In this paper we only consider *deterministic* actions, i.e. actions that have *at most one* successor state. In this case, the diamond-formula $\langle do_i(\alpha)\rangle\varphi$ is stronger than the box-formula$[do_i(\alpha)]\varphi$. To be more precise, we have the equivalence expressed by $\langle do_i(\alpha)\rangle\varphi \leftrightarrow ([do_i(\alpha)]\varphi \wedge \langle do_i(\alpha)\rangle\top)$. The diamond-formula expresses that agent i has the opportunity to perform α, and that φ is one of its effects, whereas the box-formula only asserts that agent i would end up in a situation in which φ holds if i would perform α, and as such it says nothing about the opportunity for i to do α.

Roughly speaking, the language $\mathcal{L}$ allows for *objective* formulas, i.e., formulas about an actual state of affairs, and we have *operators* to make assertions about the agents' (informational and motivational) attitudes, capacities and dynamics. Such operators can be *practitional*, when their argument is an action (like in $\mathbf{A}_i\alpha$ saying that i is able of doing α, or $\mathbf{Com}_i\alpha$ which expresses that i is committed to α). But operators can also be *assertional*, when their argument is a formula (like in $\mathbf{Goal}_i\varphi$—agent i has φ as a goal). In such a case, one could also speak about just a *modal* operator. In this dichotomy, $\langle do_i(\alpha)\rangle\varphi$ (and, for that matter, α, $[do_i(\alpha)]\varphi$) lives in both worlds: it expresses the practitional fact that i has the opportunity to do α, and the assertional fact that φ is one of its possible effects.

However, it is also important to note that the informational and motivational attitudes of the agents not only can be expressed at a *declarative* level (agent i believes φ, is committed to α), but are also considered at an *operational* level (agent i updates its beliefs with φ, or commits itself to α).

Whereas the compositional behaviours of results, abilities and opportunities are a main issue in our research, we feel that in order to dress up agents with cognitive features, one also has to come up with specific instances of atomic actions. In Section 3 we study some typical 'informational' actions. To be more precise, for every propositional formula φ and agent j, we consider the actions $\mathtt{obs}\,\varphi$, (agents can do observations), $\mathtt{inform}(\varphi, j, \mathbf{t})$ (an agent may inform another agent j about φ, 'tagged' with an additional $\mathbf{t}$, see Section 3) and $\mathtt{try\_jump}\,\varphi$ (an agent may apply one of its defaults in order to conclude φ). The three actions just mentioned provide the agent with mechanisms that trigger it to change its mind (he can observe or hear something, or have its own 'rules of thumb') but they don't necessarily specify *how* the agent exactly changes its mind. Under some general conditions, informative actions typically have the effect of changing the *epistemic state* of an agent. In the literature, a popular triple of such actions is *expansions, contractions* and *revisions*. In Section 3 we will study three corresponding informative actions (we use the generic term 'updates' for them) for each agent.

The operational aspects of the informational attitudes of agents are addressed by the atomic actions mentioned above. For the declarative aspects of information, we use modal operators $\mathbf{B}_i^k, \mathbf{B}_i^o, \mathbf{B}_i^c$ and $\mathbf{B}_i^d$. $\mathbf{B}_i^k\varphi$ means that agent i *knows* φ. Here, knowledge must be understood as the information that the agent is born or designed with, like propositional tautologies or domain-specific facts. As such, it comprises to the agent available, true, but fixed information: it determines the agent's basic perspective throughout its existence. A characteristic property of the other $\mathbf{B}_i$ operators is that they model information that is believed to be true by the agent and that is subject to change. In particular in planning it may be necessary for agents to acquire additional information about their world from whatever source possible. The $\mathbf{B}_i^o$ and $\mathbf{B}_i^c$ operator correspond to external sources the agent can use to gain information: the agent can *observe* a state of affairs, or he may *communicate* about it, respectively. The operator $\mathbf{B}_i^d$ represents those beliefs that the agent adopts *by default*. We assume that observational beliefs are always true, whereas this is not a prerequisite for the communicational and default beliefs. In Section 3 we will link these attitudes to the corresponding informative actions mentioned above.

Section 4 adds 'motivational' actions to the atomic kernel of actions. For any action α defined so far, we add the action $\mathtt{commit\_to}\,\alpha$ to the set of actions. In some situations, agents cannot fulfil their promises. Therefore, we also consider $\mathtt{uncommit}\,\alpha$ as a basic action. Our language is then completed with operators that deal with motivational attitudes. At the *assertion* level, we consider *wishes* and *goals*, represented as $\mathbf{W}_i$ and $\mathbf{Goal}_i$, respectively. At the *practition* level we define *commitments*, represented as $\mathbf{Com}_i$.

2.2 Semantics

Recalling that the full language described above is denoted with $\mathcal{L}$, we will now briefly indicate how to formally interpret formulas of $\mathcal{L}$. We use the following kind of Kripke models $M = \langle W, \pi, \mathtt{D}, \mathtt{I}, \mathtt{M} \rangle$ where W is a non-empty set of worlds. In order to determine whether a formula $\varphi \in \mathcal{L}$ is true in w (if so, we write $(M, w) \models \varphi$, where (M, w) is called a *state*), π, $\mathtt{P}$, $\mathtt{I}$ and $\mathtt{M}$ encode how to interpret the propositional atoms and the Dynamic, Informational and Motivational operators, respectively. For the propositional part, we stipulate $\mathcal{M}, s \models p$ iff $\pi(s)(p) = \mathit{true}$, and the logical connectives are interpreted in $\mathcal{M}, s$ as expected. We are going to extend the definition of $\mathcal{M}, s \models \varphi$ to arbitrary φ; if $\mathcal{M}$ is clear from context, we abbreviate $\{s \mid \mathcal{M}, s \models \varphi\}$ to $[\![\varphi]\!]$.

Let us agree to call a modal operator X to be a necessity operator for a relation R (or a function f) if the truth of $X\varphi$ at state M, w is defined as the truth of φ in all states s for which Rws (or $s \in f(w)$, respectively). Then, the $\mathtt{D}$-part of a model consists of two functions: a result function $\mathtt{r}$ such that, for each atomic action a and agent i, $\mathtt{r}(i, a)(\mathcal{M}, w)$ yields all resulting states (M', w') of i doing a in w. We will explain below why we generalize the function $\mathtt{r}$ from a world transformation (as it is usually defined in dynamic logic, cf. [10]) to a transformation on states (M, w). $\mathtt{D}$ also contains a capability function $\mathtt{c}$ where $\mathtt{c}(i, a)(w)$ is true exactly for those atomic actions a that i is capable of in w. The function $\mathtt{r}$ is extended to arbitrary actions in a way standard for dynamic logic (cf. [10]) and then $[do_i(\alpha)]$ is interpreted as the necessity operator for this extended $\mathtt{r}^*$. In a similar fashion, $\mathtt{c}$ is extended for arbitrary actions. Then we have:
$\mathcal{M}, s \models [do_i(\alpha)]\varphi$ iff $\mathcal{M}, t \models \varphi$ for all $\mathcal{M}, t \in \mathtt{r}^*(\mathcal{M}, s)$
$\mathcal{M}, s \models \mathbf{A}_i\varphi$ iff $\mathtt{c}^*(\mathcal{M}, s) = \mathit{true}$

Definition 2.2 We now give the extensions of $\mathtt{r}$ and $\mathtt{c}$, respectively. Recall that we assume that all actions are *deterministic*. For convenience, we introduce the special state $\mathcal{E} = (\mathcal{M}, \epsilon)$, where ϵ is a world where all impossible actions lead to: from there, no other actions can be performed anymore.

$$
\begin{array}{ll}
\mathtt{r}^*(i, a)(\mathcal{M}, s) & = \mathtt{r}(i, a)(\mathcal{M}, s) \\
\mathtt{r}^*(i, \mathtt{confirm}\ \varphi)(\mathcal{M}, s) & = (\mathcal{M}, s) \text{ if } \mathcal{M}, s \models \varphi \text{ and } \mathcal{E} \text{ otherwise} \\
\mathtt{r}^*(i, \alpha_1; \alpha_2)(\mathcal{M}, s) & = \mathtt{r}^*(i, \alpha_2)(\mathtt{r}^*(i, \alpha_1)(\mathcal{M}, s)) \\
\mathtt{r}^*(i, \mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1 & = \mathtt{r}^*(i, \alpha_1)(\mathcal{M}, s) \text{ if } \mathcal{M}, s \models \varphi \text{ and} \\
\quad \mathtt{else}\ \alpha_2\ \mathtt{fi})(\mathcal{M}, s) & \quad \mathtt{r}^*(i, \alpha_2)(\mathcal{M}, s) \text{ otherwise} \\
\mathtt{r}^*(i, \mathtt{while}\ \varphi & = (\mathcal{M}', s') \text{ iff } \exists k \in \mathbb{N} \exists \mathcal{M}_0, s_0 \ldots \exists \mathcal{M}_k, s_k \\
\quad \mathtt{do}\ \alpha_1\ \mathtt{od})(\mathcal{M}, s) & \quad [\mathcal{M}_0, s_0 = \mathcal{M}, s \,\&\, \mathcal{M}_k, s_k = \mathcal{M}', s' \,\&\, \forall j < k \\
 & \qquad [\mathcal{M}_{j+1}, s_{j+1} = \mathtt{r}^*(i, \mathtt{confirm}\ \varphi; \alpha_1)(\mathcal{M}_j, s_j)] \\
 & \qquad \,\&\, \mathcal{M}', s' \models \neg\varphi] \\
\text{where } \mathtt{r}^*(i, \alpha)(\mathcal{E}) & = \mathcal{E} \\
\text{and} &
\end{array}
$$

$$
\begin{array}{ll}
\mathtt{c}^*(i,a)(\mathcal{M},s) & = \mathtt{c}(i,a)(s) \\
\mathtt{c}^*(i,\mathtt{confirm}\ \varphi)(\mathcal{M},s) & = \textit{true} \text{ if } \mathcal{M},s \models \varphi \text{ and } \textit{false} \text{ otherwise} \\
\mathtt{c}^*(i,\alpha_1;\alpha_2)(\mathcal{M},s) & = \mathtt{c}^*(i,\alpha_1)(\mathcal{M},s)\ \&\ \mathtt{c}^*(i,\alpha_2)(\mathtt{r}^*(i,\alpha_1)(\mathcal{M},s)) \\
\mathtt{c}^*(i,\mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1 & = \mathtt{c}^*(i,\mathtt{confirm}\ \varphi;\alpha_1)(\mathcal{M},s) \text{ or} \\
\quad \mathtt{else}\ \alpha_2\ \mathtt{fi})(\mathcal{M},s) & \quad \mathtt{c}^*(i,\mathtt{confirm}\ \neg\varphi;\alpha_2)(\mathcal{M},s) \\
\mathtt{c}^*(i,\mathtt{while}\ \varphi & = \textit{true} \text{ if } \exists k \in \mathbb{N}[\mathtt{c}^*(i,(\mathtt{confirm}\ \varphi;\alpha_1)^k; \\
\quad \mathtt{do}\ \alpha_1\ \mathtt{od})(\mathcal{M},s) & \qquad\qquad \mathtt{confirm}\ \neg\varphi)(\mathcal{M},s) = \textit{true}] \\
 & \quad \text{and } \textit{false} \text{ otherwise} \\
\text{where } \mathtt{c}^*(i,\alpha)(\mathcal{E}) & = \textit{true}
\end{array}
$$

Note that the clauses above only indicate how composed actions are interpreted. One main feature of our approach is the possibility to define actions that may effect the semantic units: actions may change the set of worlds that an agent considers epistemically possible, desirable, and also the sequence of actions it is planning to perform next. This will be exploited further in the following sections. With regard to the abilities of agents, the motivation for the choices made in Definition 2.2 is the following. The definition of $\mathtt{c}(i,\mathtt{confirm}\ \varphi)$ 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 get confirmation for 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 repetitive composition $\mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}$ iff it is able to perform the action $(\mathtt{confirm}\,\varphi;\alpha_1)^k;\mathtt{confirm}\,\neg\varphi$ for some natural number k, i.e. it is able to perform the k-th unwinding of the while-loop. Also note that in the ';-clause' for $\mathtt{c}^*$ we have chosen to model *optimistic* agents: as a consequence, we have that any agent finds itself capable of performing $\alpha;\beta$, whenever it lacks the opportunity to perform α (for a discussion on this choice, cf. [22]).

To explain the content of the informational layer $\mathtt{I}$ of the model, recall that, concerning the declarative part, we had four informational operators $\mathbf{B}_i^k, \mathbf{B}_i^o, \mathbf{B}_i^c$ and $\mathbf{B}_i^d$. Let x be a variable over $\{k,o,c,d\}$, then for each x, $\mathtt{I}$ contains a function B^x such that, for each state $w \in W$, and agent i, $\mathrm{B}^x(i,w)) \subseteq W$ denotes all the worlds that are x-believed to be possible by agent i, in world w. The interpretation of the $\mathbf{B}_i^x$ now becomes:
$\mathcal{M},s \models \mathbf{B}_i^x\varphi$ iff $\forall t \in \mathrm{B}^x(i,s), \mathcal{M},t \models \varphi$
This truth condition for the operators $\mathbf{B}_i^x$ ensures that the notions of belief satisfy at least the properties of the logic **K45** (cf. [7]). To ensure that both $\mathbf{B}_i^k$ and $\mathbf{B}_i^o$ behave knowledge-like (or **S5**-like), and to relate the different kinds of belief, we require some additional properties on our models, which will be made explicit in Section 3.

Now, since we allow for actions that change the beliefs of an agent it is clear why actions must be considered to transform states into states, since an

informative action, like an update of an agent i's beliefs at level x in a state (M, w) typically does not effect objective formulas in w, nor any of the functions $\mathrm{B}^x(j, \cdot)(i \neq j)$, but it will effect the function $\mathrm{B}^x(i, \cdot)$: as a result of an update, agent i may consider other states as epistemically indistinguishable. There is both a mathematical and a conceptual difference between the result function `r` and the information functions B^x. Firstly, `r` may pick states from a model $\mathcal{M}'$ different from the model $\mathcal{M}$ it is applied to. Secondly, the function `r` models a state *transition* caused by certain actions, whereas the informational functions yield states that are *compatible* with agent i's knowledge, observations, communication or default beliefs, respectively. In Section 3 we will see the need for further ingredients of the `I`-part of our models as we'll go along.

Dealing with motivational attitudes of agents is the part where we depart significally from standard modal approaches. We leave the instantiation of the `M`-part of the model to Section 4.

Without going into technical details at this point, it will be clear right now that our framework definitely embodies a modal-logical flavour. However, it is well-known that normal modal operators have certain properties that are occasionally considered undesirable for the common-sense 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, in some restricted situations). While using modal operators for epistemic notions may lead to *over-idealisation* of such a notion, when formalising motivational attitudes the undesired properties induced by closure under logical consequence become even more problematic. 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. To facilitate discussion of some of those properties, let us collect them in the following seven 'Logical Omniscience properties', which are all valid properties for normal modal operators.

Definition 2.3 Let $\varphi, \psi \in \mathcal{L}$ be formulae, and let $\mathbf{X}$ be some operator.

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

3 Gathering Information

In this section, we first describe various notions of belief in our framework, and then we discuss the notion of belief change. In particular, we distinguish ways that describe how to perform such a belief change (Section 3.2) from the effects of a belief change as a result of a particular action (Section 3.3).

3.1 Notions of Belief

The logics of knowledge and belief are by now well established and understood ([7, 27]). To cater for the desired properties, we dress up our models with the following requirements. For the function B^k we require that $w \in \mathrm{B}^k(i,w)$ and, for all $u, v \in W, u \in \mathrm{B}^k(i,v) \Leftrightarrow \mathrm{B}^k(i,u) = \mathrm{B}^k(i,v)$. This ensures that the relation $\mathtt{R}_i^k xy$ defined as $y \in \mathrm{B}^k(i,x)$ is an equivalence relation, and hence, that knowledge satisfies the **S5** properties. We put the same constraints on B^o, such that observational beliefs have the same properties as knowledge. Finally, we require that the B^x-functions are such that for all i and $s, s' \in W$:

- $\mathrm{B}^d(i,s) \neq \emptyset$
- $\mathrm{B}^d(i,s) \subseteq \mathrm{B}^c(i,s) \subseteq \mathrm{B}^o(i,s) \subseteq \mathrm{B}^k(i,s)$
- if $s' \in \mathrm{B}^o(i,s)$ then $\mathrm{B}^c(i,s') = \mathrm{B}^c(i,s)$ and $\mathrm{B}^d(i,s') = \mathrm{B}^d(i,s)$

The properties above ensure the following theorem.

Theorem 3.1 Any belief operator $\mathbf{X} \in \{\mathbf{B}_i^k, \mathbf{B}_i^o, \mathbf{B}_i^c, \mathbf{B}_i^d\}$ satisfies all the logical omniscience properties $LO1 \Leftrightarrow LO2$ of definition 2.3. Moreover, for all $\varphi \in \mathcal{L}$ we have:

1. $\models (\mathbf{B}_i^o\varphi \rightarrow \varphi) \wedge (\mathbf{B}_i^k\varphi \rightarrow \varphi)$ T
2. $\models \neg(\mathbf{X}\varphi \wedge \mathbf{X}\neg\varphi)$ D
3. $\models \mathbf{X}\varphi \rightarrow \mathbf{X}\mathbf{X}\varphi$ 4
4. $\models \neg\mathbf{X}\varphi \rightarrow \mathbf{X}\neg\mathbf{X}\varphi$ 5
5. $\models (\mathbf{B}_i^k\varphi \rightarrow \mathbf{B}_i^o\varphi) \wedge (\mathbf{B}_i^o\psi \rightarrow \mathbf{B}_i^c\psi) \wedge (\mathbf{B}_i^c\chi \rightarrow \mathbf{B}_i^d\chi)$

Here, we will not address the problem of logical omniscience for belief operators any further (for a discussion, see [35, 15]). Item 1 of theorem 3.1 says that knowledge and observational beliefs are veridical: they must be true. Item 2 states a weaker property, saying that one cannot have false beliefs. Items 3 and 4 are known as positive and negative introspection, respectively. Finally, item 5 relates the various notions of beliefs. It may seem a bit counter-intuitive to say, for instance, that knowing φ implies your 'observational-believing' φ, but this only indicates relative strength. There *is* a way to express that i o-believes φ purely on the basis of its observations. Let, for $x = k, o, c, d$, $\mathbf{Ignorant}_i^x\varphi \stackrel{\triangle}{=} \neg\mathbf{B}_i^x\varphi \wedge \neg\mathbf{B}_i^x\neg\varphi$. Then we can also define the following:

Definition 3.2 *For each $\varphi \in \mathcal{L}$ we define:*

- $\mathbf{Saw}_i\varphi \triangleq \mathbf{Ignorant}_i^k\varphi \wedge \mathbf{B}_i^o\varphi$
- $\mathbf{Heard}_i\varphi \triangleq \mathbf{Ignorant}_i^o\varphi \wedge \mathbf{B}_i^c\varphi$
- $\mathbf{Jumped}_i\varphi \triangleq \mathbf{Ignorant}_i^c\varphi \wedge \mathbf{B}_i^d\varphi$

Note that, for instance, the $\mathbf{Heard}_i$ operator does not formalise hearing *per se*, but rather *believing on the basis of being told*. One easily verifies that the operators defined above satisfy logical omniscience property $LO1$. However, they generally don't satisfy all properties of Logical Omniscience. For instance, $LO2$ is not valid: for no tautology ψ, one can have $\mathbf{Saw}_i\psi$. A last property we mention (for more, see also [22, Proposition 5.12]) is the following weakening of $LO3$: if $\models \varphi \to \psi$ then $\models \mathbf{B}_i^c\psi$, and hence $\models \varphi \to \psi$ implies $\models \mathbf{Heard}_i\varphi \to (\mathbf{Heard}_i\psi \vee \mathbf{Saw}_i\psi \vee \mathbf{B}_i^k\psi)$ The latter implication expresses that if an agent believes a formula on the basis of its communication, it must believe all its consequences on at least a reliable level: it must have heard or seen them, or else even know them.

3.2 Belief Changes: How?

Concerning the agent's attitude towards *new* information, we like to address two main issues within our framework. The first is: what makes an agent change its beliefs? When is it prepared to give up (part of) its ideas about how the world looks like? The other question is: how do agents update their beliefs? Can we guarantee that they don't change it unnecessarily, for instance, and if so, how do they decide which beliefs to hold on to, and which to give up? We will first tackle the second question, since it eases, technically speaking, our way to answer the first one.

The work of Alchourrón, Gärdenfors and Makinson (AGM, [1]) has become a standard reference in the area of belief revision. The basic building blocks here are a theory T, which is closed under classical consequences, some propositional formula ϕ representing novel information to be incorporated into T, and an operation that performs the change and yields a theory T'. AGM distinguishes the operations *expansion* (where ϕ is added to T), *contraction* (the resulting T' should not imply ϕ anymore) and *revision* (which can be defined as a contraction with $\neg\phi$, followed by an expansion with ϕ). Here, we will use the term *updates* as a generic term.

We will now define our updates, and, having done that, we will relate them to the AGM framework. In fact, we consider knowledge to be static, and thus not subject to any updates. Thus, our notion of knowledge comprises the properties the agent is designed or born with. In fact, since the *implementation* of the basic update operations are the same for o, c and d beliefs (not the *condition* under which they are performed, see below), in the following we will omit the superscript x, and just study updates of beliefs. Thus, we assume, for each

formula $\varphi \in \mathcal{L}_0$—we will explain this restriction to $\mathcal{L}_0$ shortly—to have the atomic actions $\texttt{expand}\ \varphi, \texttt{contract}\ \varphi$ and $\texttt{revise}\ \varphi$. We think the following properties must be shared by all of them:

Definition 3.3 We distinguish the following properties of actions α, where $\chi \in \mathcal{L}$.

- $\models \langle \mathrm{do}_i(\alpha) \rangle \top$ *realizability*
- $\models \langle \mathrm{do}_i(\alpha) \rangle \chi \rightarrow [\mathrm{do}_i(\alpha)]\chi$ *determinism*
- $\models \langle \mathrm{do}_i(\alpha; \alpha) \rangle \chi \leftrightarrow \langle \mathrm{do}_i(\alpha) \rangle \chi$ *idempotence*

Realizability of an action implies that agents have the opportunity to perform the action regardless of circumstances; determinism of an action means that performing the action results in a unique state of affairs, and idempotence of an action implies that performing the action an arbitrary number of times has the same effect as performing the action just once. We say that $\Delta \in \{\texttt{expand}, \texttt{contract}, \texttt{revise}\}$ satisfies any of the properties of Definition 3.3 if the action $\Delta\varphi$ satisfies that property for all $\varphi \in \mathcal{L}_0$.

Let us now see how we can model expansions. Informally, a belief expansion is an action that leads to a state of affairs in which some formula is included in the set of beliefs of an agent. In our framework uncertainties of agents are formalized through the different doxastic alternatives that the agent has: if an agent believes neither φ nor $\neg\varphi$ then it considers both doxastic alternatives supporting φ and doxastic alternatives supporting $\neg\varphi$ possible. Expanding the beliefs of the agent with φ may then be implemented by declaring all alternatives supporting $\neg\varphi$ to be 'doxastically impossible', i.e., on the ground of its beliefs the agent no longer considers these alternatives to be possible. Hence the *expansion* of the belief set of an agent can be modelled through a *restriction* of its set of doxastic alternatives.

Definition 3.4 Recall that we only consider two doxastic accessibility functions B^k,and B respectively, in the informational layer $\mathtt{I}$ now. Let some model $\mathcal{M} = \langle W, \pi, \mathtt{D}, \mathtt{I}, \mathtt{M} \rangle$ with $s \in W$, and $\varphi \in \mathcal{L}_0$ be given. We define:

$\mathtt{r}(i, \texttt{expand}\ \varphi)(\mathcal{M}, s) = \mathcal{M}', s$ where
$\mathcal{M}' = \langle W, \pi, \mathtt{D}, \mathtt{I}', \mathtt{M} \rangle$ with
$\mathrm{B}'(i', s') = \mathrm{B}(i', s')$ if $i' \neq i$ or $s' \notin \mathrm{B}^k(i, s)$
$\mathrm{B}'(i, s') = \mathrm{B}(i, s') \cap [\![\varphi]\!]$ if $s' \in \mathrm{B}^k(i, s)$

Definition 3.4 provides for an intuitively acceptable formalization of belief expansions as can be seen in the following proposition.

Proposition 3.5 *For all $\varphi, \psi \in \mathcal{L}_0$ we have:*

- $\models [\mathrm{do}_i(\texttt{expand}\ \varphi)]\mathbf{B}_i\varphi$
- $\models \mathbf{B}_i\psi \rightarrow [\mathrm{do}_i(\texttt{expand}\ \varphi)]\mathbf{B}_i\psi$
- $\models \mathbf{B}_i\varphi \rightarrow (\mathbf{B}_i\psi \leftrightarrow [\mathrm{do}_i(\texttt{expand}\ \varphi)]\mathbf{B}_i\psi)$

The first clause of Proposition 3.5 states that an expansion with some formula results in the formula being believed. The second clause states that beliefs are persistent under expansions. In this clause the restriction to *propositional* formulae ψ is in general necessary. For consider a situation in which an agent does not believe φ and by negative introspection believes that it does not believe φ. After expanding its beliefs with φ, the agent believes φ and, assuming that the resulting belief set is not the absurd one, it no longer believes that it does not believe φ. Hence not all beliefs of the agent persist in situations like these. Note that the first two clauses combined indicate that our definition of belief, in the context of allowing absurd belief sets, is a good one when dealing with expansions. For an expansion with some formula φ in a situation in which $\neg\varphi$ is already believed, results in the agent believing both φ and $\neg\varphi$ and hence having inconsistent beliefs. The third clause states that in situations where some formula is already believed, nothing is changed as the result of an expansion with that formula. This latter property is suggested by the *criterion of informational economy* [8], which states that since information is in general not gratuitous, unnecessary losses of information are to be avoided.

A belief contraction is the change of belief through which in general some formula that is believed beforehand is no longer believed afterwards. As such, apparent beliefs that an agent has are turned into doubts as the result of a contraction. In terms of our framework, this change of belief may be implemented by *extending* the set of doxastic alternatives of an agent in order to encompass at least one state not satisfying the formula that is to be contracted. Consider for example the situation of an agent i that believes p, i.e., p holds in all its doxastic alternatives. When contracting p from the belief set of the agent, some $\neg p$-worlds are added to the set of doxastic alternatives of the agent. In order to end up with well-defined Kripke models, these worlds that are to be added, need to be in the set of epistemic alternatives of s. For in our Kripke models, the set of doxastic alternatives for a given agent in a given state is contained in its set of epistemic alternatives in that state. Thus the worlds that are to be added to the set of doxastic alternatives of the agent are elements of the set of epistemic alternatives not supporting p.

The problem with defining contractions in this way, is that it is not straightforward to decide which worlds need to be added. From the basic idea that knowledge — acting as the principles of agents — provides some sort of lower bound of the belief set of an agent, it is clear that in the case of a contraction with φ some states need to be added that are elements of the set of epistemic alternatives of the agent and do not support φ, but it is not clear exactly *which* elements of this set need to be chosen. The approach that we propose to solve this problem is based on the use of so called *selection functions*. These are functions $\sigma : \mathcal{A} \times W \times \mathcal{L}_0 \to \wp(W)$ that (whenever possible) select a subset of the set of epistemic alternatives in such a way that the resulting `contract` action behaves rationally. Without giving the full details here (see [25], let us here mention the two requirements that $\sigma(i, s, \varphi) \subseteq \mathrm{B}^k(i, s)$ and that $\sigma(i, s, \varphi) = \sigma(i, s'\varphi)$, when-

ever $s' \in \mathrm{B}^k(i,s)$)), we claim that such a function can be defined in such a way that the following definition guarantees Proposition 3.7, stating such rationality properties. The reader may verify that the two requirements mentioned above guarantee the items 3 and 9 of Proposition 3.7, respectively).

The definition of $\mathtt{r}^*$ for the `contract` action is based on the use of selection functions: a contraction is performed by adding to the set of doxastic alternatives of the agent exactly those worlds that are picked out by the selection function.

Definition 3.6 Let some model $\mathcal{M} = \langle W, \pi, \mathtt{D}, \mathtt{I}, \mathtt{M}\rangle$ with $s \in W$ and $\varphi \in \mathcal{L}_0$ be given. Furthermore, let σ be an arbitrary but fixed selection function for $\mathcal{M}$. We define:

$\mathtt{r}^*(i, \mathtt{contract}\ \varphi)(\mathcal{M}, s) = \mathcal{M}', s$ where
$\mathcal{M}' = \langle W, \pi, \mathtt{D}, \mathtt{I}', \mathtt{M}\rangle$ with
$\mathrm{B}'(i', s') = \mathrm{B}(i', s')$ if $i' \neq i$ or $s' \notin \mathrm{B}^k(i,s)$
$\mathrm{B}'(i, s') = \mathrm{B}(i, s') \cup \sigma(i, s, \varphi)$ for all $s' \in \mathrm{B}^k(i,s)$

Proposition 3.7 *For all $\varphi, \psi, \vartheta \in \mathcal{L}_0, \chi \in \mathcal{L}$ we have:*

- $\models [\mathrm{do}_i(\mathtt{contract}\ \varphi)]\mathbf{B}_i\psi \rightarrow \mathbf{B}_i\psi$
- $\models \neg\mathbf{B}_i\varphi \rightarrow ([\mathrm{do}_i(\mathtt{contract}\ \varphi)]\mathbf{B}_i\psi \leftrightarrow \mathbf{B}_i\psi)$
- $\models \neg\mathbf{K}_i\varphi \rightarrow [\mathrm{do}_i(\mathtt{contract}\ \varphi)]\neg\mathbf{B}_i\varphi$
- $\models \mathbf{B}_i\varphi \rightarrow (\mathbf{B}_i\psi \rightarrow [\mathrm{do}_i(\mathtt{contract}\ \varphi; \mathtt{expand}\ \varphi)]\mathbf{B}_i\psi)$
- $\models ([\mathrm{do}_i(\mathtt{contract}\ \varphi)]\mathbf{B}_i\vartheta \wedge [\mathrm{do}_i(\mathtt{contract}\ \psi)]\mathbf{B}_i\vartheta) \rightarrow$
  $[\mathrm{do}_i(\mathtt{contract}\ \varphi \wedge \psi)]\mathbf{B}_i\vartheta$
- $\models [\mathrm{do}_i(\mathtt{contract}\ \varphi \wedge \psi)]\neg\mathbf{B}_i\varphi \rightarrow$
  $([\mathrm{do}_i(\mathtt{contract}\ \varphi \wedge \psi)]\mathbf{B}_i\vartheta \rightarrow [\mathrm{do}_i(\mathtt{contract}\ \varphi)]\mathbf{B}_i\vartheta)$
- $\models \neg\mathbf{B}_i\varphi \rightarrow ([\mathrm{do}_i(\mathtt{contract}\ \varphi)]\chi \leftrightarrow \chi)$
- $\models \mathbf{B}_i\varphi \rightarrow (\chi \leftrightarrow [\mathrm{do}_i(\mathtt{contract}\ \varphi; \mathtt{expand}\ \varphi)]\chi)$
- $\models \mathbf{K}_i(\varphi \leftrightarrow \psi) \rightarrow ([\mathrm{do}_i(\mathtt{contract}\ \varphi)]\chi \leftrightarrow [\mathrm{do}_i(\mathtt{contract}\ \psi)]\chi)$
- `contract` *satisfies realizability, determinism and idempotence*

The first clause of Proposition 3.7 states that after a contraction an agent believes at most the formulae that it believed before the contraction. The second clause states that in situations in which an agent does not believe φ, nothing changes as the result of contracting φ. Again this property reflects the criterion of informational economy. The third clause states that a contraction with a contractable formula, i.e., a formula not belonging to the agent's knowledge, results in the agent not believing the contracted formula. The fourth clause states that whenever an agent believes a formula, all beliefs in its original belief set are recovered after a contraction with that formula followed by an expansion with the same formula. The fifth clause formalizes the idea that all formulae that are believed both after a contraction with φ and after a contraction with ψ, are believed after a contraction with $\varphi \wedge \psi$. Clause 6 states that if a contraction with $\varphi \wedge \psi$ results in φ not being believed, then in order to contract φ no more

formulae need to be removed than those that were removed in order to contract $\varphi \wedge \psi$. This clause is related to the property of *minimal change* for contractions. The seventh clause states that contractions with disbelieved formulae cause no change at all. The next clause states that whenever an agent believes a formula, a contraction with that formula followed by an expansion with the same formula reduces to the void action and therefore causes no change. Clause 9 states that for formulae that an agent knows to be equivalent, a contraction with one formula causes exactly the same change as a contraction with the other formula. By the last clause, contractions obey the properties given in Definition 3.3.

Having defined actions that model expansions and contractions, we now turn to defining actions that model revisions. A revision is a change of belief through which some formula is added to the beliefs of an agent, while preserving consistency. Our definition of actions that model revisions is based on the *Levi identity* [21]. Levi suggested that revisions can be defined in terms of contractions and expansions: a revision with φ can be defined as a contraction with $\neg\varphi$ followed by an expansion with φ. Given the definitions of contractions and expansions of the previous paragraphs and the fact that the class of actions Ac that we consider is closed under sequential composition, the Levi identity provides for a means to define revisions as the sequential composition of a contraction and an expansion action.

Definition 3.8 Let some model $\mathcal{M} = \langle W, \pi, \mathtt{D}, \mathtt{I}, \mathtt{M} \rangle$ with $s \in W$ and $\varphi \in \mathcal{L}_0$ be given. We define:

- $\mathtt{r}^*(i, \mathtt{revise}\ \varphi)(\mathcal{M}, s) = \mathtt{r}^*(i, \mathtt{contract}\ \neg\varphi; \mathtt{expand}\ \varphi)(\mathcal{M}, s)$

Proposition 3.9 *For all $\varphi, \psi, \vartheta \in \mathcal{L}_0$ we have:*

- $\models [\mathrm{do}_i(\mathtt{revise}\ \varphi)]\mathbf{B}_i\varphi$
- $\models [\mathrm{do}_i(\mathtt{revise}\ \varphi)]\mathbf{B}_i\vartheta \rightarrow [\mathrm{do}_i(\mathtt{expand}\ \varphi)]\mathbf{B}_i\vartheta$
- $\models \neg\mathbf{B}_i\neg\varphi \rightarrow ([\mathrm{do}_i(\mathtt{expand}\ \varphi)]\mathbf{B}_i\vartheta \leftrightarrow [\mathrm{do}_i(\mathtt{revise}\ \varphi)]\mathbf{B}_i\vartheta)$
- $\models \mathbf{K}_i\neg\varphi \leftrightarrow [\mathrm{do}_i(\mathtt{revise}\ \varphi)]\mathbf{B}_i\bot$
- $\models \mathbf{K}_i(\varphi \leftrightarrow \psi) \rightarrow ([\mathrm{do}_i(\mathtt{revise}\ \varphi)]\mathbf{B}_i\vartheta \leftrightarrow [\mathrm{do}_i(\mathtt{revise}\ \psi)]\mathbf{B}_i\vartheta)$
- $\models [\mathrm{do}_i(\mathtt{revise}\ \varphi \wedge \psi)]\mathbf{B}_i\vartheta \rightarrow [\mathrm{do}_i(\mathtt{revise}\ \varphi; \mathtt{expand}\ \psi)]\mathbf{B}_i\vartheta$
- $\models \neg[\mathrm{do}_i(\mathtt{revise}\ \varphi)]\mathbf{B}_i\neg\psi \rightarrow$
  $([\mathrm{do}_i(\mathtt{revise}\ \varphi; \mathtt{expand}\ \psi)]\mathbf{B}_i\vartheta \rightarrow [\mathrm{do}_i(\mathtt{revise}\ \varphi \wedge \psi)]\mathbf{B}_i\vartheta)$

The first clause of Proposition 3.9 states that agents believe φ as the result of revising their beliefs with φ. The second clause states that a revision with φ results in the agent believing at most the formulae that it would believe after expanding its beliefs with φ, i.e., changing the belief set to incorporate φ consistently (if possible) — this is a revision with φ — results in a subset of the set of beliefs that results from straightforward inserting φ in the belief set — an expansion with φ. The third clause formalizes the idea that expansion is a special kind of revision: in cases where $\neg\varphi$ is not believed, expanding with φ

and revising with φ amount to the same action. The left-to-right implication of the fourth clause states that if $\neg\varphi$ is known, i.e., $\neg\varphi$ is among the agent's knowledge, then the revision with φ results in the agent believing $\bot$, i.e., the revision results in the absurd belief set. The right-to-left implication of the fourth clause states that the agent will believe $\bot$ only after it performs a revision with a negation of something that it knows. The fifth clause states that revisions with formulae that are known to be equivalent have identical results. The sixth clause formalizes the idea that the revision with the conjunction $\varphi \wedge \psi$ results in the agent believing at most the formulae that it would believe after a revision with φ followed by an expansion with ψ. The seventh clause states that if a revision with φ does not result in $\neg\psi$ being believed, then after revising with $\varphi \wedge \psi$ the agent believes at least the formulae that it would believe as the result of performing a revision with φ followed by an expansion with ψ. As Gärdenfors remarks, clauses 6 and 7 provide for some sort of *minimal change* condition on revisions.

Let us finally explain how our approach can be related to the AGM postulates for expansion, contraction and revision. In order to do so, let the belief set of agent i at state $\mathcal{M}, s$ be defined as $\mathcal{B}(i, \mathcal{M}, s) = \{\varphi \in \mathcal{L}_0 \mid \mathcal{M}, s \models \mathbf{B}_i\varphi\}$. Moreover, the *contraction* of $\mathcal{B}(i, \mathcal{M}, s)$ with $\varphi \in \mathcal{L}_0$, notation $\mathcal{B}^-_\varphi(i, \mathcal{M}, s)$ is defined as: $\mathcal{B}^-_\varphi(i, \mathcal{M}, s) = \{\psi \in \mathcal{L}_0 \mid \mathcal{M}, s \models [\mathrm{do}_i(\texttt{contract}\ \varphi)]\mathbf{B}_i\psi\}$ Similar sets $\mathcal{B}^+_\varphi(i, \mathcal{M}, s)$ and $\mathcal{B}^*_\varphi(i, \mathcal{M}, s)$ can be defined as the expansion and the revision at s with φ, respectively. Then, in [25] we show that these sets exactly satisfy the AGM postulates for AGM contraction, expansion and revision, respectively. There is one difference: whereas the AGM postulates assume classical logic as a backup system, in our setup this is provided by the agent's knowledge. For instance, whereas the AGM Success-postulate for contraction states that a formula φ that is not a classical tautology does not follow from a theory T contracted with φ, our variant of that postulate reads

$$\text{If } \mathcal{M}, s \models \neg\mathbf{B}^k_i\varphi \text{ then } \varphi \notin \mathcal{B}^-_\varphi(i, \mathcal{M}, s)$$

3.2.1 The ability to change one's mind

In the previous (sub)sections, we dealt with the formalization of the opportunity for, and the result of, the actions that model the belief changes of agents. Here we look at the *ability* of agents to change their beliefs.

For 'mental' actions, like testing (observing) and communicating, the abilities of agents are closely related to their (lack of) information. This observation seems to hold *a fortiori* for the abstract actions that cause agents to change their beliefs. For when testing and communicating, at least some interaction takes place, either with the real world in case of testing, or with other agents when communicating, whereas the changing of beliefs is a strictly mental, agent-internal, activity. Therefore, it seems natural to let the ability of an agent to change its beliefs be determined by its informational state only.

The intuitive idea behind the definitions as we present them, is that the ability to change beliefs can be used to guide the changes that the beliefs of an agent are subjected to. In particular, if an agent is able to change its beliefs in a certain way, then this change of belief should work out as desired, i.e., it should neither result in an absurd belief set nor cause no change at all. Another point of attention is given by the observation that the Levi identity should also be respected for abilities, i.e., an agent is capable of revising its beliefs with a formula φ if and only if it is able to contract its beliefs with $\neg\varphi$ and thereafter perform an expansion with φ.

Definition 3.10 Let $\mathcal{M}$ be some Kripke model with state s, and let $\varphi \in \mathcal{L}_0$ be arbitrary. We define the capability function $\mathtt{c}$ for the `expand`, `contract` and `revise` actions in the following manner:

$$\begin{array}{lll} \mathtt{c}(i, \mathtt{expand}\ \varphi)(\mathcal{M}, s) = \mathbf{1} & \Leftrightarrow & \mathcal{M}, s \models \neg\mathbf{B}_i\neg\varphi \\ \mathtt{c}(i, \mathtt{contract}\ \varphi)(\mathcal{M}, s) = \mathbf{1} & \Leftrightarrow & \mathcal{M}, s \models \neg\mathbf{K}_i\varphi \\ \mathtt{c}(i, \mathtt{revise}\ \varphi)(\mathcal{M}, s) & = & \mathtt{c}(i, \mathtt{contract}\ \neg\varphi; \mathtt{expand}\ \varphi)(\mathcal{M}, s) \end{array}$$

The first clause of Definition 3.10 states that an agent is able to expand its set of beliefs with a formula if and only if it does not already believe the negation of the formula. The second clause formalizes the idea that an agent is able to remove some formula from its set of beliefs if and only if it does not consider the formula to be one of its principles. The ability for the `revise` action is defined through the Levi identity.

Proposition 3.11 *For all $\varphi \in \mathcal{L}_0$ we have:*

- $\models \mathbf{A}_i\mathtt{expand}\ \varphi \leftrightarrow \mathbf{K}_i\mathbf{A}_i\mathtt{expand}\ \varphi$
- $\models \mathbf{A}_i\mathtt{expand}\ \varphi \rightarrow \langle \mathrm{do}_i(\mathtt{expand}\ \varphi)\rangle\neg\mathbf{B}_i\bot$
- $\models \mathbf{A}_i\mathtt{contract}\ \varphi \leftrightarrow \mathbf{K}_i\mathbf{A}_i\mathtt{contract}\ \varphi$
- $\models \mathbf{A}_i\mathtt{contract}\ \varphi \rightarrow \langle \mathrm{do}_i(\mathtt{contract}\ \varphi)\rangle\neg\mathbf{B}_i\varphi$
- $\models \mathbf{A}_i\mathtt{revise}\ \varphi \leftrightarrow \mathbf{A}_i\mathtt{contract}\ \neg\varphi$
- $\models \mathbf{A}_i\mathtt{revise}\ \varphi \rightarrow \langle \mathrm{do}_i(\mathtt{revise}\ \varphi)\rangle(\mathbf{B}_i\varphi \wedge \neg\mathbf{B}_i\bot)$

The first and third clause of Proposition 3.11 state that agents know of their ability to expand and contract their beliefs; a consequence of the fifth clause is that agents also know of their ability to revise their beliefs. The second, fourth and sixth clause formalize the idea that belief changes of which the agent is capable, behave as desired, i.e., an expansion does not result in absurd belief sets, a contraction leads to disbelief in the contracted formula, and a revision results in a combination of these.

3.3 Belief Changes: When?

3.3.1 Observations: seeing is believing

Through observations an agent *learns whether* some proposition is true of the state in which it is residing. For artificial agents it seems to be a reasonable as-

sumption to demand that observations are *truthful*. That is, if some observation yields information that φ, then it should indeed be the case that φ. Observations form the most trustworthy way of acquiring information: utterances like 'I've seen it with my own eyes' or 'Seeing is believing' support this claim. The formalisation that we propose is therefore such that observations overrule any beliefs acquired by other means. In situations where an agent does not observationally believe whether φ, its beliefs are revised by applying the $\mathtt{revise}^o$ function to the model and the state under consideration. If the agent already observationally believed whether φ no revision takes place. Note that observations will never conflict with an agent's observational beliefs or its knowledge: since both these notions are veridical, and observations are truthful, it is not possible that an observation that some formula φ holds contradicts observational belief or knowledge that $\neg\varphi$ holds, since this would force both φ and $\neg\varphi$ to be true in one and the same state.

Let, for $x = k, o, c, d$, $\mathbf{Bwh}_i^x\varphi \stackrel{\triangle}{=} \neg\mathbf{Ignorant}_i^x\varphi$.

Definition 3.12 *For all $\mathcal{M} \in \mathbf{M}^{\mathrm{I}}$ with state s and all $\varphi \in \mathcal{L}_0$ we define:*

$$\mathtt{r}(i, \mathtt{obs}\,\varphi)(\mathcal{M}, s) = \begin{cases} \mathcal{M}, s & \text{if } \mathcal{M}, s \models \mathbf{Bwh}_i^o\varphi \\ \mathtt{revise}^o(i,\varphi)(\mathcal{M}, s), s & \text{otherwise} \end{cases}$$

Now we can define $\mathtt{revise}^o$ in the same fashion as the semantic revision procedures of Section 3.2, but we have to do some additional work in order to keep the relations between $\mathbf{B}_i^o, \mathbf{B}_i^c$ and $\mathbf{B}_i^d$ in tact. To do so, let us agree on the following notation. Let $\mathcal{M}$ be given. For any formula φ and B^x, let

$$\kappa(\llbracket\varphi\rrbracket) \cdot \mathrm{B}^x(i,w) = \begin{cases} \mathrm{B}^x(i,w) \cap \llbracket\varphi\rrbracket & \text{if } \mathcal{M}, w \models \varphi \\ \mathrm{B}^x(i,w) \cap \llbracket\neg\varphi\rrbracket & \text{else} \end{cases}$$

Definition 3.13 *Let $\mathcal{M} = \langle W, \pi, \mathtt{D}, \mathtt{I}, \mathtt{M}\rangle$ be given, $\mathtt{I} = \{\mathrm{B}^k, \mathrm{B}^o, \mathrm{B}^c, \mathrm{B}^d\}$. Then $\mathtt{revise}^o(i,\varphi)(\mathcal{M}, s) = \langle W, \pi, \mathtt{D}, \{\mathrm{B}^k, \mathrm{B}^{o\prime}, \mathrm{B}^{c\prime}, \mathrm{B}^{d\prime}\}, \mathtt{M}, \rangle, s$ where ($y \in \{c, d\}$, $x \in \{o, c, d\}$):*

$$\begin{array}{ll}
\mathrm{B}^{x\prime}(j,t) = \mathrm{B}^x(j,t) & \text{if } \ j \neq i \ \text{ or } \ t \notin \mathrm{B}^x(i,s) \\
\mathrm{B}^{o\prime}(i,t) = \kappa(\llbracket\varphi\rrbracket) \cdot \mathrm{B}^o(i,t) & \text{if } t \in \mathrm{B}^o(i,s) \\
\mathrm{B}^{y\prime}(i,t) = \mathrm{B}^y(i,t) \cap \mathrm{B}^{o\prime}(i,t) & \text{if } \ \mathrm{B}^y(i,t) \cap \mathrm{B}^{o\prime}(i,t) \neq \emptyset,\ t \in \mathrm{B}^o(i,s) \\
\mathrm{B}^{c\prime}(i,t) = \mathrm{B}^{o\prime}(i,t) & \text{if } \ \mathrm{B}^c(i,t) \cap \mathrm{B}^{o\prime}(i,t) = \emptyset,\ t \in \mathrm{B}^o(i,s) \\
\mathrm{B}^{d\prime}(i,t) = \mathrm{B}^{c\prime}(i,t) & \text{if } \ \mathrm{B}^d(i,t) \cap \mathrm{B}^{o\prime}(i,t) = \emptyset,\ t \in \mathrm{B}^o(i,s)
\end{array}$$

Intuitively, Definition 3.13 expresses the following strategy, if agent i tries to revise its observational beliefs, when it is in state s of the model $\mathcal{M}$. First of all, there is no effect whatsoever for other agents (see item 3 of the next Proposition), or for states that are not observational equivalent to s, according to i. The next clause of the definition ensures that an agent has a correct observational belief after revising: if it revises with φ, and φ is true in s, then it only considers

as new observational alternatives those in which φ holds; if $\neg\varphi$ is true in s, however, it will only consider the $\neg\varphi$-alternatives, afterwards. The last three clauses of the definition deal with communicational and default alternatives. They stay within the observational cluster, so to speak, but if there are no communicational alternatives left in the new observational cluster, the agent performs a kind of reset: it then equalizes the communicational alternatives with the new observational ones. A similar property holds for the new default-compatible alternatives. The intuitive acceptability of this definition can be seen from the following proposition.

Proposition 3.14 *For all $i, j \in \mathcal{A}$, $\varphi, \psi \in \mathcal{L}_0$ and $\chi \in \mathcal{L}$ we have:*

1. $\mathtt{obs}\,\varphi$ *is deterministic, idempotent and realisable*
2. $\models [\mathrm{do}_i(\mathtt{obs}\,\varphi)]\mathbf{Bwh}_i^o\varphi \wedge ((\varphi \to [\mathrm{do}_i(\mathtt{obs}\,\varphi)]) \wedge (\neg\varphi \to [\mathrm{do}_i(\mathtt{obs}\,\neg\varphi)]$
3. $\models \mathbf{B}_j^k\psi \leftrightarrow \langle \mathrm{do}_i(\mathtt{obs}\,\varphi)\rangle\mathbf{B}_j^k\psi$
4. $\models \mathbf{B}_j^o\psi \to \langle \mathrm{do}_i(\mathtt{obs}\,\varphi)\rangle\mathbf{B}_j^o\psi$
5. $\models (\varphi \wedge \langle \mathrm{do}_i(\mathtt{obs}\,\varphi)\rangle\mathbf{B}_i^o\psi) \leftrightarrow (\varphi \wedge \mathbf{B}_i^o(\varphi \to \psi))$
6. $\models (\neg\varphi \wedge \langle \mathrm{do}_i(\mathtt{obs}\,\varphi)\rangle\mathbf{B}_i^o\psi) \leftrightarrow (\neg\varphi \wedge \mathbf{B}_i^o(\neg\varphi \to \psi))$
7. $\models \langle \mathrm{do}_i(\mathtt{obs}\,\varphi)\rangle\chi \leftrightarrow \langle \mathrm{do}_i(\mathtt{obs}\,\neg\varphi)\rangle\chi$
8. $\models \varphi \wedge \mathbf{Ignorant}_i^k\varphi \to \langle \mathrm{do}_i(\mathtt{obs}\,\varphi)\rangle\mathbf{Saw}_i\varphi$
9. $\models \neg\varphi \wedge \mathbf{Ignorant}_i^k\varphi \to \langle \mathrm{do}_i(\mathtt{obs}\,\varphi)\rangle\mathbf{Saw}_i\neg\varphi$
10. $\models \varphi \wedge (\mathbf{Heard}_i\neg\varphi \vee \mathbf{Jumped}_i\neg\varphi) \to \langle \mathrm{do}_i(\mathtt{obs}\,\varphi)\rangle\mathbf{Saw}_i\varphi$
11. $\models \varphi \wedge \mathbf{B}_i^c\neg\varphi \to \langle \mathrm{do}_i(\mathtt{obs}\,\varphi)\rangle((\mathbf{B}_i^c\chi \leftrightarrow \mathbf{B}_i^o\chi) \wedge (\mathbf{B}_i^d\chi \leftrightarrow \mathbf{B}_i^o\chi))$

The properties given in the first item are not uncommon for informative actions: determinism and idempotence are strongly related to the AGM postulates, and are also encountered for the other informative actions. Realisability is typical for observations; it models the idea that our agents are perfect observers which always have the opportunity to make an observation. The second item of Proposition 3.14 formalises two essential properties of observations, viz. their informativeness and truthfulness. Item 3 states that the knowledge fluents — the propositional formulae known to be true — of all agents remain unaffected under execution of an `observe` action by one of them, and item 4 states an analogous, but slightly weaker property, for their observational beliefs. The fifth and sixth item express the fact that an agent adapts only new beliefs ψ after doing an observation, if it already believed on beforehand that this observation would entail this new belief ψ. In item 7 it is formalised that the $\mathtt{obs}\,\varphi$ action models 'observing whether φ': observing whether φ is in all aspects equivalent to observing whether $\neg\varphi$. Items 8 and 9 state that for knowledge-ignorant agents observations actually lead to *learning by seeing*. Item 10 — a special case of item 8 — is intuitively a very nice one: it states that observations are the most credible source of information. Observations overrule other beliefs acquired through communication or adopted by default, i.e. incorrect communicational or default beliefs are *revised* in favour of observational beliefs. The last item of Proposition 3.14 sheds some more light on the (rigorous) way in which beliefs

are revised: observing something that contradicts communicational beliefs leads to a *reset* of both the latter and the default beliefs of the agent, i.e. after such a revision all the beliefs of the agent are at least grounded in its observations. Phrased differently, there no longer is any formula that the agent believes due to it being told or assuming it by default, i.e. after such a revision there is no formula ψ for which either $\mathbf{Heard}_i\psi$ or $\mathbf{Jumped}_i\psi$ holds.

3.3.2 Communication: hearing is believing

The second source of information available to an agent consists of the information communicated by other agents. As we present it here, communication is reduced to its barest form, viz. the transfer of information. That is, we are not dealing with concepts like communication protocols, synchronisation and the like, but instead consider communication as consisting of an agent transferring some of its information to another agent. In our implementation of revising by communication, we made some choices that are in some sense rather arbitrary. On the other hand, we believe our framework demonstrates to be both expressive in that it makes those choices explicit, and flexible in that it can easily be adapted to other choices. In general, agents have the opportunity to send all of their beliefs, and nothing else. However, the sending agent should also provide his information with his source, in order to let the receiving agent rationally decide on how to deal with the new incoming information. For these sources, we take the tags $\mathbf{k}, \mathbf{s}, \mathbf{h}$ and $\mathbf{j}$, denoting that the sending agent's source for the message he is sending is that he knows it, observed it, heard it or concluded it by using some default, respectively. This assumes that each agent knows, or has some awareness about how he comes to his beliefs, for instance whether $\neg\mathbf{B}_i^c \rightarrow \mathbf{B}_i^k\neg\mathbf{B}_i^c\varphi$ hold. Such a systematic investigation is beyond the scope of this chapter, but one may read [12]. Thus, we introduce, for each $\varphi \in \mathcal{L}_0$ and $\mathbf{t} \in \{\mathbf{k}, \mathbf{s}, \mathbf{h}, \mathbf{j}\}$, an action $\texttt{inform}(\varphi, i, \mathbf{t})$, expressing that agent i is informed about φ, and that the status of this fact at the sender's perspective is $\mathbf{t}$.

Depending on the credibility of both the sending agent and the information that it sends, the receiving agent may use this information to revise its beliefs. For reasons of simplicity, we define the credibility of the sending agent as a binary notion, i.e. the agent is either credible or it is not credible, without distinguishing degrees of credibility. The notion of credibility is modelled through the so-called dependence operator, originally proposed by Huang [17]. This ternary operator, pertaining to a pair of agents and a formula, models that there is a relation of trust, dependence or credibility between the agents with respect to the formula: $\mathbf{D}_{i,j}\varphi$ indicates that agent i accepts agent j as an authority on φ, or that j is a teacher of i on the subject φ. The $\mathbf{D}_{i,j}$ operator is interpreted by incorporating a function $\mathrm{D} : \mathcal{A} \times \mathcal{A} \rightarrow W \rightarrow \wp(\mathcal{L})$ in the $\mathtt{I}$-part of our models: $\mathbf{D}_{i,j}\varphi$ is true in a state s of some model iff φ is in $\mathrm{D}(i, j)(s)$. The credibility of the transferred information is determined by both the credibility that the sending agent attaches to this information and the credibility that

the receiving agent attaches to any of i beliefs that contradict this information. That is, if the sending agent itself observationally believes the information that it is sending, then this information overrules any contradicting beliefs that the receiving agent may have. If the sending agent itself was told the information that it is now transferring, the receiving agent will accept this information only if it does not have any contradicting information that it believes at least with the credibility attached to communicational beliefs. Information that the sending agent adopted by default is considered to be too weak to ever justify a revision of the receiving agent's beliefs. These intuitive ideas are formalised in Definition 3.15. The definition of $\texttt{revise}^c$ below is in the same spirit of Definition 3.13 for $\texttt{revise}^o$, we omit it here for reasons of space (see [22] for details).

Definition 3.15 *For all $\mathcal{M} \in \mathbf{M}^{\mathrm{I}}$ with state s, $i, j \in \mathcal{A}$, $\mathbf{t} \in \{\mathbf{k}, \mathbf{s}, \mathbf{h}, \mathbf{j}\}$, $\psi \in \mathcal{L}$ and $\varphi \in \mathcal{L}_0$ we define:*

$$\mathcal{M}, s \models \mathbf{D}_{i,j}\psi \Leftrightarrow \psi \in \mathrm{D}(i,j)(s)$$

$$\mathtt{r}(j, \mathtt{inform}(\varphi, i, \mathbf{t}))(\mathcal{M}, s) = \begin{cases} \emptyset & \text{if } \mathbf{t} = \mathbf{k} \text{ and } \mathcal{M}, s \models \neg \mathbf{B}^k_j \varphi, \text{ or} \\ & \text{if } \mathbf{t} = \mathbf{s} \text{ and } \mathcal{M}, s \models \neg \mathbf{Saw}_j \varphi, \text{ or} \\ & \text{if } \mathbf{t} = \mathbf{h} \text{ and } \mathcal{M}, s \models \neg \mathbf{Heard}_j \varphi, \text{ or} \\ & \text{if } \mathbf{t} = \mathbf{j} \text{ and } \mathcal{M}, s \models \neg \mathbf{Jumped}_j \varphi \\ \mathtt{revise}^c(i, \varphi)(\mathcal{M}, s), s & \text{if } \mathbf{t} = \mathbf{j} \text{ and } \mathcal{M}, s \models \mathbf{D}_{i,j}\varphi \wedge \mathbf{Ignorant}^c_i \varphi) \\ & \text{if } \mathbf{t} \in \{\mathbf{k}, \mathbf{s}\} \text{ and } \mathcal{M}, s \models \mathbf{D}_{i,j}\varphi \wedge \mathbf{Ignorant}^o_i \varphi \\ \mathcal{M}, s & \text{otherwise} \end{cases}$$

In the following proposition we summarise some validities describing the behaviour of the communication actions. These validities show that our formalisation indeed corresponds to the intuitive ideas unfolded above.

Proposition 3.16 *For all $i, i', j \in \mathcal{A}$, $\varphi, \psi \in \mathcal{L}_0$ and $\chi \in \mathcal{L}$, $\mathbf{t} \in \{\mathbf{k}, \mathbf{s}, \mathbf{h}, \mathbf{j}\}$, $\mathbf{t}' \in \{\mathbf{k}, \mathbf{s}, \mathbf{h}\}$ we have:*

1. $\mathtt{inform}(\varphi, i, \mathbf{t})$ *is deterministic and idempotent*
2. $\models \mathbf{B}^x_{i'}\psi \rightarrow [\mathrm{do}_j(\mathtt{inform}(\varphi, i, \mathbf{t}))]\mathbf{B}^x_{i'}\psi$ *for* $x \in \{k, o\}$
3. $\models \mathbf{Jumped}_j\varphi \leftrightarrow \langle \mathrm{do}_j(\mathtt{inform}(\varphi, i, \mathbf{j}))\rangle \top$
4. $\models \mathbf{B}^d_j\varphi \wedge \neg \mathbf{D}_{i,j}\varphi \rightarrow (\langle \mathrm{do}_j(\mathtt{inform}(\varphi, i, \mathbf{t}))\rangle\chi \leftrightarrow \chi)$
5. $\models \mathbf{D}_{i,j}\varphi \wedge \mathbf{B}^c_j\varphi \rightarrow (\langle \mathrm{do}_j(\mathtt{inform}(\varphi, i, \mathbf{k}))\rangle \mathbf{Bwh}^c_i\varphi \vee$
   $\langle \mathrm{do}_j(\mathtt{inform}(\varphi, i, \mathbf{s}))\rangle \mathbf{Bwh}^c_i\varphi \vee \langle \mathrm{do}_j(\mathtt{inform}(\varphi, i, \mathbf{h}))\rangle \mathbf{Bwh}^c_i\varphi)$
6. $\models \mathbf{D}_{i,j}\varphi \wedge \mathbf{Heard}_j\varphi \wedge \mathbf{Ignorant}^c_i\varphi \rightarrow \langle \mathrm{do}_j(\mathtt{inform}(\varphi, i, \mathbf{t}'))\rangle \mathbf{Heard}_i\varphi$
7. $\models \mathbf{D}_{i,j}\varphi \wedge \mathbf{Heard}_j\varphi \wedge \mathbf{Bwh}^c_i\varphi \rightarrow (\langle \mathrm{do}_j(\mathtt{inform}(\varphi, i, \mathbf{h}))\rangle\chi \leftrightarrow \chi)$
8. $\models \mathbf{D}_{i,j}\varphi \wedge \mathbf{Heard}_j\varphi \wedge \mathbf{Ignorant}^c_i\varphi \rightarrow$
   $(\langle \mathrm{do}_j(\mathtt{inform}(\varphi, i, \mathbf{t}'))\rangle \mathbf{B}^c_i\psi \leftrightarrow \mathbf{B}^c_i(\varphi \rightarrow \psi))$
9. $\models \mathbf{D}_{i,j}\varphi \wedge \mathbf{Saw}_j\varphi \wedge \mathbf{Ignorant}^o_i\varphi \wedge \mathbf{B}^c_i\neg\varphi$
   $\rightarrow (\langle \mathrm{do}_j(\mathtt{inform}(\varphi, i, \mathbf{s}))\rangle \mathbf{B}^c_i\psi \leftrightarrow \mathbf{B}^o_i(\varphi \rightarrow \psi))$
10. $\models \mathbf{D}_{i,j}\varphi \wedge \mathbf{Jumped}_j\varphi \rightarrow (\langle \mathrm{do}_j(\mathtt{inform}(\varphi, i, \mathbf{j}))\rangle\chi \leftrightarrow \chi)$

11. $\models \mathbf{B}_i^x\varphi \rightarrow (\langle \mathrm{do}_i(\mathtt{inform}\,\varphi, i, \mathbf{t})\rangle\chi \rightarrow \chi)$ *for* $x \in \{k, o, c, d\}$

The first item of Proposition 3.16 states that the `inform` action also obeys the properties of determinism and idempotence that are intuitively related to the AGM postulates for belief revision. The second item states that both the knowledge and the observational belief fluents of all agents persist under execution of an `inform` action by one of them. Thus communication does at most affect the regions of beliefs that it should affect, viz. the communicational and default belief clusters. Note that, in contrast with the corresponding items of Propositions 3.14 and 3.18, the communicational beliefs of the receiving agent do not necessarily persist. The reason for this is given in item 9, where communicational beliefs are genuinely revised (and not just expanded). Item 3 states that agents may transfer, using the tag $\mathbf{j}$, all, and nothing but, the beliefs they jumped to. Attempts to use a wrong tag are doomed to fail. Agents are therefore not even allowed to tell white lies; they are utterly honest. Note that item 3 also shows that the `inform` action is — in contrast with the `observe` action — generally not realisable. Corresponding properties hold for the other tags. Item 4 formalises that authority is a *conditio sine qua non* for effectual communication: if the receiving agent does not trust the sending agent on the transferred information, it lets the information pass without revising its beliefs (or changing anything else for that matter). Item 5 states that if some trustworthy agent j tells another agent i some formula φ that j either knows, observed or was told, this leads to a state of affairs in which the receiving agent believes whether φ at least with the credibility attached to communicational beliefs; whenever i is beforehand ignorant with regard to φ on the level of communicational beliefs, the receiving agent actually *learns* φ by being told (item 6). Item 7 states that if agent j tells i some formula that j itself was told while i already communicationally believes whether φ, nothing changes, really. Items 8 and 9 deal with the ways in which the receiving agent's beliefs are revised as the result of it acquiring information through communication. If the sending agent j heard the transferred formula φ and the receiving agent i is communicationally ignorant with respect to φ, then an expansion with φ of the communicational beliefs of i takes place (item 8). If j saw φ and i is observationally ignorant with respect to φ while communicationally believing $\neg\varphi$, then i's communicational beliefs consist henceforth of the observational beliefs that are implied by φ (item 9). Item 10 states that default beliefs are not transferable: the credibility of this kind of information is too low to have any effect upon being heard. The last item in particular shows that agents cannot increase the credibility they attach to information by talking to themselves, since this talking to themselves does not change anything. Thus our agents are not susceptible to this kind of auto-suggestion.

3.3.3 Default jumps: jumping is believing

The last possible source of information that we consider in this chapter is the only endogenous one, and consists of the possibility to adopt beliefs by default. In general in default reasoning, plausible yet fallible conclusions are derived on the basis of the presence of certain information and the absence of other information. In the formalisation of default reasoning as proposed by Reiter [33], defaults are formalised as special inference rules $\varphi_1 \ : \ \varphi_2/\varphi_3$, which should be interpreted as stating that if φ_1 holds and it is consistent to assume φ_2 then φ_3 may be derived. Here we consider the most basic form of default reasoning, in which no information is required to be present and the information that needs to be absent is strongly related to the conclusion that is to be derived. In terms of Reiter's framework, the kind of default reasoning that we consider here uses only *supernormal* defaults, i.e. defaults of the form $: \ \varphi/\varphi$; these defaults can be seen as *possible hypotheses* in Poole's system [31]. Here we shall introduce these supernormal defaults as syntactical constructs that are at the agent's disposal. Formally, we include in the $\mathtt{I}$-part of our models a function $\mathrm{N} : \mathcal{A} \to W \to \wp(\mathcal{L})$, such that $\mathrm{N}(i,s)$ yields the set of defaults available to agent i at state s. In the language we include an operator $\mathbf{N}_i$, such that $\mathbf{N}_i\varphi$ expresses that φ is a default of i; $\mathbf{N}_i\varphi$ is true in a state s of a model iff $\varphi \in \mathrm{N}(i,s)$.

Using the $\mathbf{N}_i$-operator to represent defaults, we now come to the formalisation of the attempted jumps to conclusion that constitute (supernormal) default reasoning. Since adopting a belief by default accounts for acquiring information of the lowest credibility, it is obvious that default jumps are effective only for agents that are completely ignorant with respect to the default that is jumped to.

Definition 3.17 *For all $\mathcal{M} \in \mathbf{M}^{\mathtt{I}}$ with state s, $i \in \mathcal{A}$, and $\varphi \in \mathcal{L}_0$ we define:*

$$\mathtt{r}(i, \mathtt{try\_jump}\,\varphi)(\mathcal{M}, s) = \begin{cases} \emptyset & \text{if } \mathcal{M}, s \models \neg\mathbf{N}_i\varphi \\ \mathtt{revise}^d(i,\varphi)(\mathcal{M}, s), s & \text{if } \mathcal{M}, s \models \mathbf{N}_i\varphi \wedge \mathbf{Ignorant}_i^d\varphi \\ \mathcal{M}, s & \text{otherwise} \end{cases}$$

Proposition 3.18 *For all $i, j \in \mathcal{A}$, $\varphi, \psi \in \mathcal{L}_0$ and $\chi \in \mathcal{L}$, we have:*

1. $\mathtt{try\_jump}\,\varphi$ *is deterministic, idempotent and d-informative with regard to* φ
2. $\models \mathbf{B}_j^x\psi \to [\mathrm{do}_i(\mathtt{try\_jump}\,\varphi)]\mathbf{B}_j^x\psi$ *for* $x \in \{d, c, o, k\}$
3. $\models \mathbf{N}_i\varphi \leftrightarrow \langle \mathrm{do}_i(\mathtt{try\_jump}\,\varphi)\rangle\top$
4. $\models \langle \mathrm{do}_i(\mathtt{try\_jump}\,\varphi)\rangle\top \leftrightarrow \langle \mathrm{do}_i(\mathtt{try\_jump}\,\varphi)\rangle\mathbf{Bwh}_i^d\varphi$
5. $\models \mathbf{N}_i\varphi \wedge \mathbf{Ignorant}_i^d\varphi \to \langle \mathrm{do}_i(\mathtt{try\_jump}\,\varphi)\rangle\mathbf{Jumped}_i\varphi$
6. $\models \mathbf{N}_i\varphi \wedge \mathbf{Ignorant}_i^d\varphi \to (\langle \mathrm{do}_i(\mathtt{try\_jump}\,\varphi)\rangle\mathbf{B}_i^d\psi \leftrightarrow \mathbf{B}_i^d(\varphi \to \psi))$
7. $\models \mathbf{N}_i\varphi \wedge \mathbf{Bwh}_i^d\varphi \to (\langle \mathrm{do}_i(\mathtt{try\_jump}\,\varphi)\rangle\chi \leftrightarrow \chi)$

The first item of Proposition 3.18 deals again with the properties more or less typical for informative actions. It is obvious that the property of realisability is

not validated in general since an agent may attempt to jump to a non-default, and in 3 it is formalised that such attempted jumps are doomed to fail. Item 2 states that all information of all agents persists under the attempted jump to a formula by one of them. Item 4 states that default jumps for which an agent has the opportunity always result in the agent believing by default whether the formula that is jumped to holds. The fifth item formalises the idea that agents that are completely ignorant with respect to some formula φ jump to new default beliefs by applying the $\texttt{try\_jump}\,\varphi$ action. The incorporation of these new beliefs is brought about by expanding the default beliefs of the agent with the default that is adopted (item 6). The last item states that attempted jumps to default conclusions yield information for totally ignorant agents only.

3.3.4 The ability to gather information

As we see it, in the ability of intelligent information agents to execute informative actions two different notions are combined. On the one hand, the abilities of an agent restrict its practical possibility to acquire information: only the actions that are within the agent's capacities can be used as means to extend its information. This corresponds to the idea that agents are not just able to acquire all the information they would like to acquire. On the other hand, we use the agents' abilities to steer the way in which information is acquired. That is, through its abilities an agent is forced to prefer more credible sources of information. We will, for instance, define an agent to be able to try to adopt some formula by default only if it cannot acquire this information through its — more credible — observations.

Since, by nature, observations provide the most credible source of information, the ability to observe is determined strictly by the fact that the agents' information gathering is limited, and does not depend on other means to acquire information. In our opinion it is reasonable to consider this limit to the agent's ability to perform observations as given by the construction of the agent. For instance, human agents are built in such a way that they cannot observe objects at great distances, or objects that are outside the spectrum of human observation, like for instance X-rays. In the case of artificial agents, one could think of two robots, working together to explore a strange, new world. Each robot is equipped with its own, personal set of sensors: one robot is for instance able to observe whether its environment is radioactive, the other whether the planet's atmosphere contains oxygen, but neither robot is able to observe both. Being a construction decision, we assume that the observational capacities of agents are determined beforehand.

In defining the capabilities of agents to inform other agents, we consider the *moral* component of ability to be most relevant. That is, we demand our agents to be sincere in that they are morally unable to lie or gossip. The things that an agent is capable of telling to other agents are exactly the things that it itself believes, be it with the lowest credibility. In this way the information acquisition

of an agent that is due to communication is restricted to those formulae that are believed by some authority.

The ability to attempt to jump to a default captures both aspects described above, i.e. both the aspect of restricted information acquisition as well as that of preferring the most credible source of information are visible in the definition of the ability to (attempt to) jump. Concerning the first aspect, an agent is able to jump to a default only if it *knows* it to be a default, i.e. agents have to know their defaults in order to be able to use them. With respect to the second aspect, the capability to jump is defined to depend on the observational capacities of the agent: an agent is able to attempt a jump to a formula only if it knows that it is not able to observe whether the formula holds. In this way it is ensured that agents resort to default jumps only if the possibility of acquiring the information through observations is excluded.

Definition 3.19 *Let $\mathcal{M}$ be some model. The function* $\mathtt{c}$ *is for the informative actions defined as follows, where s is some state in $\mathcal{M}$, i, j are agents and $\varphi \in \mathcal{L}_0$ is some propositional formula.*

$$\begin{aligned} \mathtt{c}(j, \mathtt{inform}(\varphi, i, \mathbf{t}))(\mathcal{M}, s) &= \mathtt{r}(j, \mathtt{inform}(\varphi, i, \mathbf{t}))(\mathcal{M}, s) \neq \emptyset \\ \mathtt{c}(i, \mathtt{try\_jump}\,\varphi)(\mathcal{M}, s) &= \mathtt{c}(i, \mathtt{confirm}\ \mathbf{K}_i(\neg \mathbf{A}_i \mathtt{obs}\,\varphi \wedge \mathbf{N}_i \varphi))(\mathcal{M}, s) \end{aligned}$$

where $\lambda\varphi \in \mathcal{L}_0.(\mathtt{c}(i, \mathtt{obs}\,\varphi)(\mathcal{M}, s))$ is a function in φ such that

$$\mathtt{c}(i, \mathtt{obs}\,\varphi)(\mathcal{M}, s) = \mathtt{c}(i, \mathtt{obs}\,\neg\varphi)(\mathcal{M}, s)$$

Proposition 3.20 *For all $i, j \in \mathcal{A}$ and $\varphi \in \mathcal{L}_0$ and any tag $\mathbf{t}$ we have:*

1. $\models \mathbf{A}_i \mathtt{obs}\,\varphi \leftrightarrow \mathbf{A}_i \mathtt{obs}\,\neg\varphi$
2. $\models \mathbf{A}_j \mathtt{inform}(\varphi, i, \mathbf{t}) \leftrightarrow \langle \mathrm{do}_j(\mathtt{inform}(\varphi, i, \mathbf{t}))\rangle \top$
3. $\models \mathbf{A}_i \mathtt{try\_jump}\,\varphi \rightarrow \langle \mathrm{do}_i(\mathtt{try\_jump}\,\varphi)\rangle \top$
4. $\models \mathbf{A}_i \mathtt{obs}\,\varphi \rightarrow \neg \mathbf{A}_i \mathtt{try\_jump}\,\varphi$

The first item of Proposition 3.20 again expresses that the `observe` actions formalise 'observing whether'. Item 2 and 3 state that both the actions modelling communication and those modelling default reasoning, though not realisable *per se*, are *A-realisable*, i.e. having the ability to perform these actions implies having the opportunity to do so. The property of A-realisability is considered unacceptable for mundane actions, but given our view on the notion of ability for the non-mundane, informative actions, this property seems much less controversial. For both the ability and the opportunity to perform informative actions are defined in terms of informational attitudes, and thus the distinction between these notions is less clear than it is for mundane actions. The last item of Proposition 3.20 expresses that agents are able to attempt to jump to a formula only if it is not within their capacities to observe whether the formula holds.

3.4 Possible extensions

One can think of several obvious yet interesting extensions of the framework presented in this Section, most of which are in fact not too hard to implement. A possible extension was suggested by Castelfranchi [3], and consists of refining the communication part of information acquisition. In this more refined form of communication it should be possible to relate the credibility that an agent attaches to some formula that it has been told, to the number of agents that have been communicating this formula, i.e. the more agents that tell that some formula is true, the more fiercely it is believed. It is possible to formalise this intuitive idea by splitting the communicational belief cluster of an agent into different parts, one for each of the other agents. Whenever an agent j informs an agent i of the truth of φ, i's communicational belief cluster that is associated with j is revised with φ. Over the different communicational belief clusters a kind of *graded* belief modality could be defined (in the spirit of [16]), which formalises the credibility attached to the agent's communicational beliefs. The formula $\mathbf{B}_i^{c,0.5}\varphi$ would then be taken to represent that agent i communicationally believes φ with credibility 0.5, i.e. φ holds in at least half of the communicational belief clusters of i. It is clear that in this way it can indeed be modelled that the credibility attached by i to one of its communicational beliefs depends on the number of agents that have informed i of the truth of this formula. Another possibility to refine communication could be to allow a form of 'demand-driven' communication in addition to the 'supply-driven' one considered here. In this extended form of communication an agent may request information on the truth or falsity of certain proposition from another, trusted, agent. One possible way of formalising this kind of extended communication is given in [23], where an additional modal operator is used to record the agents' requests.

Another very interesting extension of the framework presented in this chapter concerns the incorporation of actions associated with belief *updates* (cf. [19]) in addition to the ones associated with belief revision that we considered here. Whereas belief revisions are changes in information on an unchanging world, updates are information changes that are associated with a change in the state of the world. The interaction between these different information changes might well be worth looking at.

The framework could also be extended to make it a suitable formalisation tool for special agents, like *intelligent information retrieval agent*. These agents assist a user with finding relevant information, particularly in cyberspace, that satisfies one of its information needs. To this end they communicate, either with their user or with other (intelligent information retrieval) agents, go to World Wide Web sites to seek for relevant information, or make assumptions by default. A preliminary formalisation of these intelligent information retrieval agents based on the framework proposed in this chapter was presented by Huibers & Van Linder [18].

4 Setting Goals

As explained in the introduction a rational agent does not only possess information processing capabilities; it should also be endowed with motivations and should be able to modify these. In this section we will discuss how these 'motivational attitudes' can be incorporated into our modal framework.

Informally speaking, we will equip an agent with an *agenda*, which contains the action(s) that the agent has committed itself to and is supposed to perform. Formally, this agenda will take the form of a function that yields for a given agent in a given possible world the set of actions that the agent is committed to in that world. The main technical difficulty that has to be solved is the proper representation in the model of the way the agent maintains its agenda while performing actions. Intuitively speaking, the agent will have to drop from its agenda the actions that it has performed already. Although this sounds rather commonsense and clear, the formal implementation of this idea turns out to be rather involved, as we shall see, mainly due to the need for considering the right mix of syntax and semantics of the actions in the agenda.

Formally, we go about as follows. We start with the formalisation of the concept of a *wish* or *desire*. We then continue with that of the notions of selecting wishes which may then be committed to (and next possibly also be uncommitted to). We will again interpret the acts of selecting, committing and uncommitting as model transformations rather than mere world 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.

We extend the language as follows. We build the motivational part on top of the core language $\mathcal{L}_c$. (At the moment we have no need for more complicated formulas expressing, for example, that one wishes to believe something, or one is committed to a revision operator, but in principle this could be done if one likes.) The language $\mathcal{L}^{\mathrm{M}}$ is obtained by adding the following clauses:

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

We also consider an extended class Ac^{M} of actions that is the smallest superset of At closed under the clauses of the core language and such that

- if $\varphi \in \mathcal{L}_c$ then $\texttt{select}\,\varphi \in \mathrm{Ac}^{\mathrm{M}}$
- if $\alpha \in \mathrm{Ac}$ then $\texttt{commit\_to}\,\alpha \in \mathrm{Ac}^{\mathrm{M}}$
- if $\alpha \in \mathrm{Ac}$ then $\texttt{uncommit}\,\alpha \in \mathrm{Ac}^{\mathrm{M}}$

4.1 Formalising wishes

In our approach we consider wishes to be the most primitive, fundamental motivational attitudes, that is to say, *in ultimo* agents are motivated to fulfil their wishes. Wishes are represented by means of a plain normal modal operator, i.e.

wishes are straightforwardly interpreted as a necessity operator over an accessibility relation W.

Thus we instantiate the M-part of our models to cater for wishes as follows:

The M-part of a model $\mathcal{M} = \langle W, \pi, \mathtt{D}, \mathtt{I}, \mathtt{M}\rangle$ contains the functions $\mathrm{W} : \mathcal{A} \to \wp(W \times W)$, which determines the desirability relation of an agent in a state, and $\mathrm{C} : \mathcal{A} \times W \to \wp(\mathcal{L}_c)$ denoting the choices made by an agent in a state, and a function Agenda : $\mathcal{A} \times W \to \wp(\mathrm{AcSeq})$, which records the commitments of agents, per state. Here AcSeq stands for the set of sequences of semi-atomic actions.

We can now interpret the $\mathbf{W}_i$ operator as usual:
$\mathcal{M}, s \models \mathbf{W}_i\varphi \Leftrightarrow \forall s' \in W((s, s') \in \mathrm{W}(i) \Rightarrow \mathcal{M}, s' \models \varphi)$

4.2 Selecting wishes

In order to transform wishes to goals, an agent has to first select candidate goals from its set of wishes on the basis of the criteria of unfulfilledness and implementability. In more ordinary language this means that an agent can choose a wish if it is (as yet) unfulfilled and implementable, i.e. the agent is able and has the opportunity to fulfil the wish by means of a finite sequence of atomic actions. Unfulfilledness of a formula φ is easily expressed in the language by means of the classical negation $\neg\varphi$. The notion of implementability is somewhat more involved. For this purpose we introduce an implementability operator $\Diamond_i$, which we interpret as follows:
$\mathcal{M}, s \models \Diamond_i\varphi \Leftrightarrow \exists k \in \mathbb{N}\exists a_1, \ldots, a_k \in \mathrm{At}(\mathcal{M}, s \models \mathbf{PracPoss}_i(a_1; \ldots; a_k, \varphi))$,
that is to say, φ is implementable by i if i has the practical possibility to perform a finite sequence of atomic actions yielding φ.

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

Definition 4.1 *For model $\mathcal{M}$ with state s, $i \in \mathcal{A}$ and $\varphi \in \mathcal{L}_c$ we define:*

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

where for $\mathcal{M} = \langle W, \pi, \mathtt{D}, \mathtt{I}, \mathtt{M}\rangle$ with $\mathtt{M} = \langle \mathrm{W}, \mathrm{C}, \mathrm{Agenda}\rangle$ we define

$\mathtt{choose}(i, \varphi)(\mathcal{M}, s) = \langle W, \pi, \mathtt{D}, \mathtt{I}, \mathtt{M}'\rangle$ *with* $\mathtt{M}' = \langle \mathrm{W}, \mathrm{C}', \mathrm{Agenda}\rangle$ *such that*
$\mathrm{C}'(i', s') = \mathrm{C}(i', s')$ *if* $i \neq i'$ *or* $s \neq s'$
$\mathrm{C}'(i, s) = \mathrm{C}(i, s) \cup \{\varphi\}$

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

Finally we define the interpretation of the $\mathbf{C}_i$ operator:
$\mathcal{M}, s \models \mathbf{C}_i\varphi \Leftrightarrow \varphi \in \mathrm{C}(i, s)$

It can be shown ([26]) the act of selecting a formula φ causes minimal change in the sense that the formula φ is marked to be chosen, and nothing else of the model is changed. This has as a corollary that, for example, wishes and implementability formulas remain true after selecting if they were so before.

4.3 Goals

Having defined wishes and selections, one might be tempted to straightforwardly define goals to be selected wishes, i.e. $\mathbf{Goal}_i\varphi \stackrel{\triangle}{=} \mathbf{W}_i\varphi \wedge \mathbf{C}_i\varphi$. This definition is however not adequate to formalise the idea of goals being selected, *unfulfilled*, *implementable* wishes. In the selection operator the criteria of unfulfilledness and implementability have not been incorporated yet. So, an easy way to do this is just to add them. Therefore, goals are defined to be those wishes that are unfulfilled, implementable and selected.

Definition 4.2 *The* $\mathbf{Goal}_i$ *operator is for* $i \in \mathcal{A}$ *and* $\varphi \in \mathcal{L}_c$ *defined by:*

$$\mathbf{Goal}_i\varphi \stackrel{\triangle}{=} \mathbf{W}_i\varphi \wedge \neg\varphi \wedge \Diamond_i\varphi \wedge \mathbf{C}_i\varphi$$

Below we state a few properties of wishes, selections and goals.

Proposition 4.3 *For all* $i \in \mathcal{A}$ *and* $\varphi \in \mathcal{L}_c$ *we have:*

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

The first item of Proposition 4.3 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 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 4.3 state that goals are neither closed under implications nor under known implications.

4.4 Commitments

After having defined goals in our framework, we go on with a formal description of how an agent can commit itself to a goal. To this end we introduce a `commit_to` α action that, when successful, will have as a result that the agent is committed to the action α. In order to accommodate our model to this action, we need some extra formal machinery.

In order to prepare grounds for the formal semantics of the `commit_to` action we introduce the notion of a transition relation in the spirit of Structural Operational Semantics of Plotkin ([30]). This method is widely used in computer science and we can employ it here fruitfully to describe below what happens with the agent's agenda when it performs actions (from that agenda, so to speak).

In our set-up we consider transitions of the form $\langle \alpha, s\rangle \rightharpoonup_{i,a} \langle \alpha', s'\rangle$, where $\alpha, \alpha' \in \mathrm{Ac}, i \in \mathcal{A}, a$ semi-atomic and $s, s' \in W$. Such a transition expresses that if in state s, i has to perform the action α, after performing the (semi-)atomic action a, this leads to a state s' in which a remaining action α' has still to be performed. 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:

Definition 4.4 *The transition system T is given by the following axioms:*

- $\langle \alpha, s\rangle \rightharpoonup_{i,\alpha} \langle \Lambda, s'\rangle$ *with* $s' = \pi_2(\mathtt{r}(i,\alpha)(\mathcal{M}, s))$ *if* α *is semi-atomic.*
- $\langle \mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1\ \mathtt{else}\ \alpha_2\ \mathtt{fi}, s\rangle \rightharpoonup_{i,\mathtt{confirm}\ \varphi} \langle \alpha_1, s\rangle$ *if* $s \models \varphi$
- $\langle \mathtt{if}\ \varphi\ \mathtt{then}\ \alpha_1\ \mathtt{else}\ \alpha_2\ \mathtt{fi}, s\rangle \rightharpoonup_{i,\mathtt{confirm}\ \neg\varphi} \langle \alpha_2, s\rangle$ *if* $s \not\models \varphi$
- $\langle \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}, s\rangle \rightharpoonup_{i,\mathtt{confirm}\ \varphi} \langle \alpha; \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}, s\rangle$ *if* $s \models \varphi$
- $\langle \mathtt{while}\,\varphi\,\mathtt{do}\,\alpha\,\mathtt{od}, s\rangle \rightharpoonup_{i,\mathtt{confirm}\ \neg\varphi} \langle \Lambda, s\rangle$ *if* $s \not\models \varphi$

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

Next, we introduce for convenience's sake an Intend predicate, analogously to the Can predicate. This predicate expresses that the agent intends to do an action with some result if it can perform the action with this result and, moreover, he knows that this result is a goal.

Definition 4.5 *For $\alpha \in \mathrm{Ac}^{\mathrm{M}}, i \in \mathcal{A}$ and $\varphi \in \mathcal{L}_c$ we define:*

$\mathbf{Intend}_i(\alpha, \varphi) \stackrel{\triangle}{=} \mathbf{Can}_i(\alpha, \varphi) \wedge \mathbf{K}_i\mathbf{Goal}_i\varphi$

Having established the formal prerequisites, we can now present the formal semantics of the `commit_to` action. Informally, a $\texttt{commit\_to}\,\alpha$ is only successful if the agent intends to do α with some result φ. If it is successful, the agent's agenda is updated in the worlds it is in. Moreover, also in all other possible worlds that are related to this world, either by performing actions or by considering epistemic alternatives, the agenda is updated. The general idea is that if the agent performs (part of) an action that is in its agenda, the agenda in the resulting state will contain the remainder of the action that was in its agenda in the original state. Moreover, since we like to model that an agent knows what it is committed to, we also stipulate that epistemic alternatives of states contain the same agenda.

Finally, it is stipulated that an agent is able to commit to an action iff its agenda is empty, and is ready to receive another commitment, so to speak. This models a simple kind of agent which could be called a *single-minded* one. By varying this definition one may model other agents as well. However, for the sake of simplicity—it is very convenient to have to consider only at most one action sequence in the agenda—we have chosen this definition here.

Definition 4.6 [1] *For all models* $\mathcal{M} = \langle W, \pi, \mathtt{D}, \mathtt{I}, \mathtt{M}\rangle$ *with state* s*, for all* $i \in \mathcal{A}$ *and* $\alpha \in \mathrm{Ac}$ *we define:*

$\mathtt{r}(i, \texttt{commit\_to}\,\alpha)(\mathcal{M}, s) = \emptyset$ *if* $\mathcal{M}, s \models \neg\mathbf{Intend}_i(\alpha, \varphi)$ *for all* $\varphi \in \mathrm{C}(i, s)$
$\mathtt{r}(i, \texttt{commit\_to}\,\alpha)(\mathcal{M}, s) = \mathcal{M}', s$
with $\mathcal{M}' = \langle W, \pi, \mathtt{D}, \mathtt{I}, \mathtt{M}'\rangle$ *and* $\mathtt{M}' = \langle \mathrm{W},\ \mathrm{C},\ \mathrm{Agenda}'\rangle$
where Agenda' *is minimal such that it is closed under the following:*
for all $s' \in \mathrm{B}^k(i, s)$, $\mathrm{Agenda}'(i, s') = \mathrm{Agenda}(i, s') \cup \{\alpha\}$
and for all $s', s'', s''' \in W, \alpha' \in \mathrm{Agenda}'(i, s')$
such that, for some semi-atomic a*,*
$\langle \alpha', s'\rangle \rightharpoonup_{i,a} \langle \alpha'', s''\rangle$ *and* $s''' \in \mathrm{B}^k(i, s'')$*:*
$\mathrm{Agenda}'(i, s''') = \mathrm{Agenda}(i, s''') \cup \{\alpha''\}$
otherwise

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

Again one can show ([26]) that the `commit_to` action is minimal in the sense that only the agenda of agent i is updated and that only the agenda in the states that are affected in the sense described above are updated.

Next we define an operator $\mathbf{Com}_i$ that indicates that the agent i is committed to an action. We define the interpretation of this operator such that, in any state epistemically equivalent with the state it is in, the agent is committed to all actions that are (semantically equivalent to) 'initial parts' of the actions written

[1] In fact, in order for this definition to be well-defined regarding the agenda function some restrictions have to be put on the models regarding the function $\mathtt{r}_0$ and the interpretation of the `confirm` actions. For ease of presentation these are omitted here. Details can be found in [26].

in its agenda. In order to capture the notion of semantical equivalence of actions we use our transition systems again, and define the notion of a *computation run.*

Definition 4.7 $\mathrm{CR}^{\mathrm{C}}_{\mathcal{M}}(i,\alpha,s) \ni a_1;a_2;\ldots;a_n$ *iff* $\langle\alpha,s\rangle \rightharpoonup_{i,a_1} \langle\alpha_1,s_1\rangle \rightharpoonup_{i,a_2} \langle\alpha_2,s_2\rangle \rightharpoonup_{i,a_3} \ldots \rightharpoonup_{i,a_n} \langle\alpha_n,s_n\rangle$ *for some* $\alpha_1,\ldots,\alpha_n \in \mathrm{Ac}, s_1,\ldots,s_n \in W$, *such that* $\alpha_n = \Lambda$.

Note that due to the fact that our actions are deterministic, the set $\mathrm{CR}^{\mathrm{C}}_{\mathcal{M}}(i,\alpha,s)$ contains *at most one* element. This should be kept in mind while considering the definitions below. Actions that have the same computation run (with respect to a certain starting state) are considered to be semantically equivalent (in that state). So now we are able to give our interpretation of the $\mathbf{Com}_i$ operator.
$\mathcal{M},s \models \mathbf{Com}_i\alpha \Leftrightarrow$
$\forall s' \in \mathrm{B}^k(i,s)\exists\alpha_1 \in \mathrm{CR}^{\mathrm{C}}_{\mathcal{M}}(i,\alpha,s')\exists\alpha_2 \in \mathrm{Agenda}(i,s')\exists\alpha_2' \in \mathrm{CR}^{\mathrm{C}}_{\mathcal{M}}(i,\alpha_2,s')$ $(\mathrm{Prefix}(\alpha_1,\alpha_2'))$
where Prefix stands for the prefix relation on computation runs (which are sequences of actions).

Finally, we note that to let an agent be really rational, it should also be capable in certain situations to abandon its commitments, for instance, when the goal is achieved or is not implementable any more. In the definition below this is put as follows: an $\mathtt{uncommit}\alpha$ action is only successful if the agent i was committed to the action α, and the agent is able to uncommit iff it does no longer intend to do α for any purpose φ.

Definition 4.8 *For all models* $\mathcal{M} = \langle W,\pi,\mathtt{D},\mathtt{I},\mathtt{M}\rangle$ *with state* s, *for all* $i \in \mathcal{A}$ *and* $\alpha \in \mathrm{Ac}$ *we define*[2]*:*

$\mathtt{r}(i,\mathtt{uncommit}\alpha)(\mathcal{M},s) = \emptyset$ *if* $\mathcal{M},s \models \neg\mathbf{Com}_i\alpha$
$\mathtt{r}(i,\mathtt{uncommit}\alpha)(\mathcal{M},s) = \mathcal{M}',s$
with $\mathcal{M}' = \langle W,\ \pi,\ \mathtt{D},\ \mathtt{I},\mathtt{M}'\rangle$ *and* $\mathtt{M}' = \langle \mathrm{W},\ \mathrm{C},\ \mathrm{Agenda}'\rangle$
where for all $s' \in \mathrm{B}^k(i,s)$,
$\mathrm{Agenda}'(i,s') = \mathrm{Agenda}(i,s')\setminus$
$\{\beta \mid \mathrm{Prefix}(\mathrm{CR}^{\mathrm{C}}_{\mathcal{M}}(i,\alpha,s'),\mathrm{CR}^{\mathrm{C}}_{\mathcal{M}}(i,\beta,s'))\}$
and for all $s',s'',s''' \in W$ *with* $\alpha' \in \mathrm{Agenda}'(i,s')$ *and such that,*
for some semi-atomic a, $\langle\alpha',s'\rangle \rightharpoonup_{i,a} \langle\alpha'',s''\rangle$ *and* $s''' \in \mathrm{B}^k(i,s'')$,
$\mathrm{Agenda}'(i,s'') = \mathrm{Agenda}(i,s'')\setminus$
$\{\beta \mid \mathrm{Prefix}(\mathrm{CR}^{\mathrm{C}}_{\mathcal{M}}(i,\alpha'',s''),\mathrm{CR}^{\mathrm{C}}_{\mathcal{M}}(i,\beta,s''))\}$
otherwise

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

The complication in this definition, as compared to that of the `commit_to` operator, is due to the fact that 'committedness' is closed under taking prefixes

[2]Actually, in order to let this notion be well-defined, we need certain minimality conditions which we omit here for simplicity's sake (for a more rigorous treatment, see [26])

of (computation runs of) actions, so that in order to successfully uncommit to an action α also all actions that have α as a prefix (with respect to computation runs) should be removed from the agent's agenda.

Some properties of the operators treated in this section are given in the following proposition.

Proposition 4.9 *For all $i \in \mathcal{A}$, $\alpha, \beta \in \mathrm{Ac}$ and $\varphi \in \mathcal{L}_c$ we have:*

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

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.

Finally, the following proposition states some intuitive properties of the $\mathbf{Com}_i$ operator with respect to complex actions.

Proposition 4.10 *For all $i \in \mathcal{A}$, $\alpha, \alpha_1, \alpha_2 \in \mathrm{Ac}$ and all $\varphi \in \mathcal{L}_c$ we have:*

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

The first item of Proposition 4.10 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.

5 Conclusion

All the characterizing features of agents that emerge in our formal model have popped up in the literature before.

Dynamic logic has been studied extensively in computer science (see Goldblatt's [9]), *epistemic* and *doxastic* logic have been of interest among philosophers since Hintikka's [11], and have since then been formalized and proven useful for computers science and artificial intelligence (see the volumes [7, 27]). It was Moore ([28, 29]) who realized that dynamic and epistemic logic can be perfectly combined into one modal framework for actions and knowledge. Our contribution here is that we formalized and developed the ideas of Moore: we were able to come up with an axiomatization for a basic multi-agent system for actions and knowledge ([13]). We also introduces several epistemic attitudes([25]), thus taking into account the source of the information the knowledge (the variants are called 'beliefs') stems from.

We have also come across several accounts of *abilities* in the literature. The approach of Brown [2] and Elgesem [6] is essentially *endogenous* in the sense that actions are not referred to explicitly. They propose modal operators, ranging over formulae, to formalise ability: $\mathbf{A}_i\varphi$ is then to be read as 'agent i is able to bring about circumstances in which φ is true'. The only formal system that we know of in which it is hinted at an exogenous approach towards ability is the one proposed by Penther. And we agree with her that it is the *compositional behaviour* of ability that should be at the focus of attention, here.

There is also an extensive literature on *motivational attitudes.* Probably the most influential account of motivational attitudes is due to Cohen & Levesque [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. In our approach we do not use such a reduction technique.

Another important formalisation of motivational attitudes is proposed by Rao & Georgeff [32] in their BDI-architecture. Treating desires and intentions as primitive, Rao & Georgeff focus on the process of intention revision rather than the 'commitment acquisition' which is essential in our formalisation. Both

desires and intentions in their framework suffer from the problems associated with logical omniscience, which we have avoided in our formalisation of goals. (Not in our treatment of wishes, though, but these only play a subsidiary role to get to goals.)

Perhaps the closest in spirit to our approach is that of Singh ([34]). Singh also aims at an integrated framework for describing intelligent agents. His framework rests mainly on branching-time temporal logic extended with 'dynamic-logic-like' operators that nevertheless have an explicit temporal interpretation, which makes the elaboration of the theory quite distinct from ours, and somewhat more involved to our taste. His focus is slightly different, as well: less on the belief-revisional attitudes and more on the intricacies of intentional notions, as related to nondeterminism and know-how. Furthermore it considers communications as based on speech act theory, which is beyond the scope of our model (But cf. [5] for a first integrating attempt in this direction). In future work we will try to obtain a more compact theory and give a treatment of communication along the lines of what has been done in the realm of distibuted systems rather than linguistics, although we certainly see the relevance of this for the enterprise of describing intelligent agents. But we also believe that somewhere the theory should be constrained in order to keep it useful. If one is not careful one ends up with a Theory of Everything, which will be very hard to use indeed for the specification of a practical agent system! As has been mentioned before, we consider it one of our main contributions to have proposed a single framework in which all of these aspects can be dealt with, to some considerable extent. In our view modal logic has proven to be an excellent tool for this purpose.

Finally we like to say some words on the restriction to deterministic actions in this paper. It has been shown in [14] that also nondeterminism can be treated within our framework by extending the set of action constructors by choice operators. In fact, in [14] we have considered two such operators: the internal non-deterministic operator $\oplus$ models those choices that are to be made by the agent, and is therefore supposed to behave *angelic*, i.e. as helpful as possible for the agent. Then, $\neg[do_i(\alpha \oplus \beta)]\varphi$ is equivalent to $\neg[do_i(\alpha)]\varphi \vee \neg[do_i(\alpha)]\varphi$: if agent i wants to avoid φ as a result of doing $(\alpha \oplus \beta)$, it is sufficient to either avoid φ as a result of α or as a result of β. The external non-deterministic choice operator $+$, on the other hand, has a *demonic* behaviour from the agent's perspective: since it must be prepared for the worst, it has to guarantee both $[do_i(\alpha)]\varphi$ and $[do_i(\beta)]\varphi$ in order to ensure that $[do_i(\alpha + \beta)]\varphi$ holds.

References

[1] C.E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[2] M.A. Brown. On the logic of ability. *Journal of Philosophical Logic*, 17:1–26, 1988.

[3] C. Castelfranchi. Personal communication.

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

[5] F. Dignum and B. van Linder. Modelling social agents: Communication as action. In J. Mueller, M. Wooldridge, and N. Jennings, editors, *Intelligent Agents III –Proceedings of the 3rd International Workshop on Agent Theories, Architectures, and Languages (ATAL-96)*, pages 83–93, 1996.

[6] D. Elgesem. *Action Theory and Modal Logic*. PhD thesis, Institute for Philosophy, University of Oslo, Oslo, Norway, 1993.

[7] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge MA, 1994. To appear.

[8] P. Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. The MIT Press, Cambridge, Massachusetts and London, England, 1988.

[9] R. Goldblatt. *Axiomatising the Logic of Computer Programming*, volume 130 of *LNCS*. Springer-Verlag, 1982.

[10] R. Goldblatt. *Logics of Time and Computation*, volume 7 of *CSLI Lecture Notes*. CSLI, Stanford, 1992. Second edition.

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

[12] W. van der Hoek. Systems for knowledge and beliefs. *Journal of Logic and Computation*, 3(2):173–195, 1993.

[13] 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.

[14] 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.

[15] W. van der Hoek and J.-J. Ch. Meyer. Possible logics for belief. *Logique & Analyse*, 127–128:177–194, 1989.

[16] W. van der Hoek and J.-J.Ch. Meyer. Graded modalities for epistemic logic. *Logique et Analyse*, 34(133-134):251–270, 1991.

[17] Z. Huang. Logics for belief dependence. In E. Börger, H. Kleine Büning, M.M. Richter, and W. Schönfeld, editors, *Computer Science Logic, 4th Workshop CSL'90*, volume 533 of *Lecture Notes in Computer Science*, pages 274–288. Springer-Verlag, 1991.

[18] T.W.C. Huibers and B. van Linder. Formalising intelligent information retrieval agents. In F. Johnson, editor, *Proceedings of the 18th BCS IRSG Annual Colloquium on Information Retrieval Research*, pages 125–143, 1996.

[19] H. Katsuno and A.O. Mendelzon. On the difference between updating a knowledge base and revising it. In P. Gärdenfors, editor, *Belief revision*, pages 183–203. Cambridge University Press, 1992.

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

[21] I. Levi. Direct inference. *The Journal of Philosophy*, 74:5–29, 1977.

[22] B. van Linder. *Modal Logics for Rational Agents*. PhD thesis, Utrecht University, 1996.

[23] B. van Linder, W. van der Hoek, and J.-J. Ch. Meyer. Communicating rational agents. In B. Nebel and L. Dreschler-Fischer, editors, *KI-94: Advances in Artificial Intelligence*, volume 861 of *Lecture Notes in Computer Science (subseries LNAI)*, pages 202–213. Springer-Verlag, 1994.

[24] 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.

[25] 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.

[26] B. van Linder, J.-J. Ch. Meyer, and W. van der Hoek. Formalising motivational attitudes of agents using the karo framework. Technical Report UU-CS-1997-03, Utrecht University, 1997.

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

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

[29] R.C. Moore. A formal theory of knowledge and action. Technical Report 320, SRI International, 1984.

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

[31] D. Poole. A logical framework for default reasoning. *Artificial Intelligence*, 36:27–47, 1988.

[32] 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.

[33] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[34] M.P. Singh. *Multiagent Systems: A Theoretical Framework for Intentions, Know-How and Communications*, volume 799 of *Lecture Notes in Computer Science (subseries LNAI)*. Springer-Verlag, 1994.

[35] E. Thijsse. On total awareness logics (with special attention to monotonicity constraints and flexibility). In M. de Rijke, editor, *Diamonds and Defaults*, volume 229 of *Synthese Library*, pages 309–347. Kluwer Academic Publishers, 1993.