Synthese (2007) 155:265–290
Knowledge, Rationality & Action 103–128
DOI 10.1007/s11229-006-9145-6

ORIGINAL PAPER

Towards a theory of intention revision

Wiebe van der Hoek · Wojciech Jamroga · Michael Wooldridge

Received: 2 November 2006 / Accepted: 18 December 2006 / Published online: 14 February 2007

Abstract Although the change of beliefs in the face of new information has been widely studied with some success, the revision of other mental states has received little attention from the theoretical perspective. In particular, *intentions* are widely recognised as being a key attitude for rational agents, and while several formal theories of intention have been proposed in the literature, the logic of intention *revision* has been hardly considered. There are several reasons for this: perhaps most importantly, intentions are very closely connected with other mental states—in particular, beliefs about the future and the abilities of the agent. So, we cannot study them in isolation. We must consider the interplay between intention revision and the revision of other mental states, which complicates the picture considerably. In this paper, we present some first steps towards a theory of intention revision. We develop a simple model of an agent's mental states, and define intention revision operators. Using this model, we develop a logic of intention dynamics, and then investigate some of its properties.

Keywords Intention revision · BDI agents · Belief change · Practical reasoning

1 Introduction

Given the success of theories of belief change over the last 20 years—whether we consider the Alchourrón–Gärdenfors–Makinson tradition (Alchourrón, Gärdenfors, & Makinson, 1985) of *belief revision*, or the Katsuno–Mendelzon paradigm (Katsuno & Mendelzon, 1991) of

W. van der Hoek (✉) · M. Wooldridge
Department of Computer Science, University of Liverpool,
Liverpool, UK
e-mail: wiebe@csc.liv.ac.uk

M. Wooldridge
e-mail: mjw@csc.liv.ac.uk

W. Jamroga
Department of Computer Science, Clausthal University of Technology,
Clausthal, Germany
e-mail: wjamroga@in.tu-clausthal.de

belief update[1]—it seems natural to ask whether these approaches can be adapted or extended to other, related mental states. Belief is, after all, just one of the attitudes of a rational agent: there are others, equally prominent in our cognitive landscape. If one is to consider extending such theories, then *pro-attitudes* of some kind seem very natural candidates (Bratman, Israel, & Pollack, 1988). A pro-attitude is one that influences an agent's actions. Pro-attitudes include desires, intentions and the like. To decide what to do, an agent needs both beliefs (which represent what the agent thinks about how the world looks), and pro-attitudes (which represent how he wants the world to look). For example, *practical syllogisms* have been suggested as providing a link between beliefs, pro-attitudes and actions. A practical syllogism is a reasoning pattern of the form:

if i desires φ and i believes that α leads to φ then i should intend to do α.[2]

There is no universal agreement on the exact nature and properties of belief (and its close relative, knowledge); and there is considerably less agreement on the nature of pro-attitudes such as desires and intentions. They seem altogether more complex beasts. There are several reasons for this. Perhaps most importantly, while it is possible to give an account of beliefs that is independent of other mental states, it does not really seem possible to do the same for pro-attitudes. This is because pro-attitudes are closely interwoven with other mental attitudes—beliefs in particular. If you say you intend to bring about φ, this implies something about your beliefs—that you believe φ is possible, for example. A rational agent would not intend φ and simultaneously believe that φ was impossible. So, we can expect the road to pro-attitude revision to be a good deal more rocky than the road to belief change. Nevertheless, the aim of the present paper is to begin to traverse this landscape—or at least, to catch a glimpse of some of the main obstacles that must be crossed.

We focus on *intention* revision. Our reasons for focusing on intention revision are threefold. First, intentions play a particularly important role in artificial intelligence: they correspond to what artificial intelligence researchers usually refer to as "goals" (Bratman et al., 1988; Pollack, 1992). Second, it has been argued that intentions play a very important role in communication and coordination, and are thus of particular interest to the multi-agent systems community (Cohen & Levesque, 1990; Jennings, 1993). And third, several significant attempts have been made to formalise a theory of intention (Cohen & Levesque, 1990; Rao & Georgeff, 1998; Wooldridge, 2000; Wooldridge & Jennings, 1999), and these provide useful guidance in formalising a theory of intention revision.

2 Intentions and intention revision

One of the best known, and most sophisticated attempts to show how the various components of an agent's cognitive structure could be combined to form a logic of rational agency is due to Cohen and Levesque (1990). Cohen and Levesque's formalism was originally used to develop a theory of intention (as in "I intended to. . ."), which the authors considered as

[1] We will return to the difference in Sect. 4.2; until then we use the terms *belief change*, *belief revision* and *belief update* interchangeably.

[2] It was Bratman (1987) who added the notion of intention to the Aristotelian model (Aristotle, 1998) of practical reasoning. It should be mentioned that this pattern is not what Bratman or Searle would see as a practical syllogism: Searle (2001), for instance, believes that it is incorrect to assume that a person's set of beliefs and desires are *causally sufficient* to determine an action. We refer the reader to Bratman (1987) and Searle (2001) and the overview (Atkinson, 2005) for a more detailed discussion on practical reasoning, practical syllogism and rational agency.

a pre-requisite for a theory of speech acts (see Cohen & Levesque, 1990 for full details). However, the logic has subsequently proved to be so useful for specifying and reasoning about the properties of agents that it has been used in an analysis of conflict and cooperation in multi-agent dialogue (Galliers, 1988), as well as several studies on theoretical foundations of cooperative problem solving (Jennings, 1992; Levesque, Cohen, & Nunes, 1990). This section will focus on the use of the logic in developing a theory of intention. The first step is to lay out the criteria that a theory of intention must satisfy.

When building intelligent agents—particularly agents that must interact with humans—it is important that a *rational balance* is achieved between the beliefs, desires and intentions of the agents. Quoting Cohen and Levesque:

> For example, the following are desirable properties of intention: an autonomous agent should act on its intentions, not in spite of them; adopt intentions it believes are feasible and forego those believed to be infeasible; keep (or commit to) intentions, but not forever; discharge those intentions believed to have been satisfied; alter intentions when relevant beliefs change; and adopt subsidiary intentions during plan formation (Cohen & Levesque, 1990, p. 214)

Following Bratman (1987, 1990), Cohen and Levesque identify seven specific properties that must be satisfied by a reasonable theory of intention:

1. Intentions pose problems for agents, who need to determine ways of achieving them.
2. Intentions provide a "filter" for adopting other intentions, which must not conflict.
3. Agents track the success of their intentions, and are inclined to try again if their attempts fail.
4. Agents believe their intentions are possible.
5. Agents do not believe they will not bring about their intentions.
6. Under certain circumstances, agents believe they will bring about their intentions.
7. Agents need not intend all the expected side effects of their intentions.

Given these criteria, Cohen and Levesque adopt a two tiered approach to the problem of formalising a theory of intention. First, they construct the logic of rational agency, "being careful to sort out the relationships among the basic modal operators" (Cohen & Levesque, 1990, p. 221). On top of this framework, they introduce a number of derived constructs, which constitute a "partial theory of rational action" (Cohen & Levesque, 1990, p. 221). Intention is one of these constructs.

The first major derived construct is that of a *persistent* goal. An agent has a persistent goal of p if:

1. He has a goal that p eventually becomes true, and believes that p is not currently true.
2. Before he drops the goal, one of the following conditions must hold:
   (a) the agent believes the goal has been satisfied;
   (b) the agent believes the goal will never be satisfied.

It is a small step from persistent goals to a first definition of intention, as in "intending to act". Note that "intending that something becomes true" is similar, but requires a slightly different definition; see (Cohen & Levesque, 1990). An agent i intends to perform action α if he has a persistent goal to have brought about a state where he had just believed he was about to perform α, and then did α. Cohen and Levesque go on to show how such a definition meets many of Bratman's criteria for a theory of intention (outlined above). In particular, by basing the definition of intention on the notion of a *persistent goal*, Cohen and Levesque are

able to avoid overcommitment or undercommitment. An agent will only drop an intention if he believes that the intention has either been achieved, or is unachievable.[3]

The intention theory of Cohen and Levesque has been enormously influential, and has prompted a number of other studies in related areas, e.g. (Cohen & Levesque, 1990; Wooldridge & Jennings, 1999). However, their theory does not address the *dynamics* of intentions: how an agent should rationally choose them or drop them. This issue has been most studied in the context of the *belief-desire-intention* (BDI) agent model (Bratman et al., 1988; Georgeff & Lansky, 1987; Georgeff & Rao, 1995), a popular approach to the design of autonomous agents that emerged in the late 1980s.

The BDI model gets its name from the fact that it recognises the primacy of beliefs, desires and intentions in rational action. Intuitively, an agent's *beliefs* correspond to information the agent has about the world. These beliefs may be incomplete or incorrect. An agent's *desires* are states of affairs that the agent would, in an ideal world, wish to bring about. Finally, an agent's *intentions* represent desires that he has *committed* to achieving. The idea is that an agent will not be able to deliberate indefinitely over which states of affairs to bring about; ultimately, he must fix upon some subset of his desires and commit to achieving them. These chosen desires are *intentions*.

A more formal description of the control loop of a BDI agent is shown in Fig. 1, which is based on the BDI agent control loop presented in Rao & Georgeff (1992) and (Wooldridge, 2000, p. 38). The idea is that an agent has beliefs $\mathcal{B}$ about the world, intentions $\mathcal{I}$ to achieve and a plan π that can be used to bring about the intended state of affairs. In lines 2–4, the beliefs, intentions and the plan are initialised. The main control loop is then in lines 5–20. In lines 6–7, the agent perceives and updates his beliefs; in line 8, he decides whether to reconsider his pro-attitudes or not; in lines 9–13 the agent deliberates, by generating new options (i.e. desires $\mathcal{D}$) and deliberating over these; in line 12, the agent generates a plan for achieving his intentions; and in lines 15–18 an action of the current plan is executed. Because the purpose of the functions used in this loop can be easily derived from their names, we omit the actual formalisations here for reasons of space, but direct the reader to Wooldridge (2000, ch. 2).

It is necessary for a BDI agent to *reconsider* his intentions from time-to-time (Bratman et al., 1988; Kinny & Georgeff, 1991; Wooldridge & Parsons, 1999). One of the key properties of intentions is that they enable the agent to be goal-driven rather than event-driven, i.e. by committing to intentions the agent can pursue long-term goals. But when circumstances have changed and, for example, an intention cannot be achieved anymore, the agent will be better off if he drops that intention. Similarly, when opportunities arise that enable bringing about desires that have been so far impossible to achieve, the agent should reconsider his intentions. On the other hand, reconsideration is itself a potentially costly computational process; therefore one would not want the agent to reconsider his intentions at every possible moment. The purpose of the *reconsider*(...) function as shown in Fig. 1 is precisely this: to deliberate when it pays to deliberate (i.e. when deliberation is likely to lead to a change in intentions), and otherwise not to deliberate, but to act.

The goal within the remainder of this paper is to develop (the beginnings of) a theory that sits somewhere between the Cohen–Levesque theory and the control loop in Fig. 1. That is, we want to develop a theory which characterises mental states of a rational agent, in much the same way as the Cohen–Levesque theory, but which incorporates the dynamics of intentions in the manner implemented in BDI agents *à la* Fig. 1.

[3] A critique of Cohen and Levesque's theory of intention is presented in Singh (1992); space restrictions prevent a discussion here.

```
Algorithm: BDI Agent Control Loop
1.
2.   B := B0;
3.   I := I0;
4.   π := null;
5.   while (true) do
6.        get next observation ω;
7.        revise B on the basis of ω;
8.        if (reconsider(B,I)) then
9.             D := options(B,I);
10.            I := filter(B,D,I);
11.            if (not sound(π,I,B)) then
12.                 π := plan(B,I);
13.            end-if
14.       end-if
15.       if (not empty(π)) then
16.            α := hd(π);
17.            execute(α);
18.            π := tail(π);
19.       end-if
20.  end-while
```

Fig. 1 The abstract BDI agent control loop. The loop consists of continuous observation, deliberation, planning and execution. To perform optimally, the *reconsider*(...) function decides whether deliberation and planning is necessary

3 Mental states

We noted above that, in order to say anything meaningful about an agent's intentions and how he revises them, we need to say how they relate to the agent's other mental states—beliefs and the like. This, in turn implies that we must commit to some particular model of the agent's mental activities. We represent the mental state Γ of an agent as a tuple

$$\Gamma = \langle \mathcal{B}, \mathcal{D}, \langle \mathcal{I}, \mathcal{A} \rangle, \mathcal{P} \rangle,$$

where $\mathcal{B}$ is a *belief set*, $\mathcal{D}$ a set of *desires*, $\langle \mathcal{I}, \mathcal{A} \rangle$ a pair of *intentions* and *active plans*, and $\mathcal{P}$ a set of *practical reasoning rules*. The beliefs of an agent will represent the information that the agent has about his environment; desires represent those states of affairs that the agent might in principle like to achieve; intentions represent those states of affairs that the agent has committed to achieving, and for which there is a current plan in place that seems feasible to achieve it; and practical reasoning rules represent the agent's "know how"—his knowledge about how to achieve possible intentions.

3.1 Propositions, options and beliefs

We start with beliefs. As will become very clear later, we view beliefs as being primarily about the future, and so we want a language for belief that allows us to express temporal properties. We use the language of Linear Temporal Logic (LTL); we pre-suppose some familiarity with LTL, and do not give a complete definition of this language or its proof system: see, e.g. (Emerson, 1990) for an introduction. We use "$\bigcirc$" as the "next" or "tomorrow" operator, "$\Diamond$" as the "eventually" operator and "$\Box$" as the "always" operator.[4] We denote the LTL proof relation by $\vdash_{\text{LTL}}$.

Now, a (strong) *belief set* $\mathcal{B}$ is a set of sentences of LTL, closed under LTL consequence. A non-trivial belief set is LTL-consistent; the only non-consistent belief set is denoted by $\mathcal{B}_\perp$.

[4] For the moment, we do not consider intentions that use the until operator $\mathcal{U}$.

Unless explicitly mentioned, a belief set is assumed to be non-trivial. The idea behind strong beliefs is that they represent the agent's ideas about the current situation—and, more importantly, the future—within which he can plan his actions. More precisely, they represent the agent's ideas about what is inevitable, no matter how he would act in the world.

We note here that a strong belief set can give rise to two different notions of belief. For instance, suppose that $\mathcal{B}$ does not entail (via LTL entailment) that atomic property *rich* will ever be true. An agent with such a belief set can adopt the intention to make *rich* true some time in the future, in which case he would believe that he would eventually become rich, although it is not one of his *strong beliefs*. In other words, after the agent has deliberated about his possible intentions, and committed to some of them, it might be that he has beliefs about the future that are not entailed by $\mathcal{B}$ alone. To extend this example a little further, if $\mathcal{B}$ entailed "eventaully *rich*" then it would not make sense for the agent to say that he is committed to making *rich* true sometime: he already (strongly) believes *rich* will eventually become true, regardless of what he does (see Sect. 3.5 for details).

States of affairs are described with propositional (*objective*) formulas α from a basic language $\mathcal{L}_0$ over a set of atoms P, with typical element p[5]

$$\alpha \ := \ p \mid \neg p \mid \alpha \wedge \alpha \mid \alpha \vee \alpha.$$

We assume that beliefs, desires and intentions refer to temporal properties of the world; we call these properties *options*. Let T be a temporal operator from $\{\bigcirc, \Diamond, \Box\}$. We define our temporal language of options as follows:

$$\varphi := T\alpha \mid T\varphi \mid \varphi \wedge \varphi \mid \varphi \vee \varphi.$$

For example, having eggs on the next day ($\bigcirc egg$), some day in the future ($\Diamond egg$), or at least every other day ($\Box(egg \vee \bigcirc egg)$) are sample options that an agent might consider when planning his breakfast for the coming week. And, the agent may believe that the second option is inevitable, while the others are not ($\Diamond egg \in \mathcal{B}$, $\bigcirc egg \notin \mathcal{B}$, $\Box(egg \vee \bigcirc egg) \notin \mathcal{B}$). Unlike for desires and intentions (which we assume to be future-oriented), the agent can also have beliefs about the current situation, expressed by propositional formulae α (e.g. the agent may believe that he is having eggs right now: $egg \in \mathcal{B}$).

To sum up more formally, a *belief set* $\mathcal{B}$ is a set of options and/or propositional formulae that satisfies the following constraints:

1. $\mathcal{B}$ is LTL-consistent (i.e. $\mathcal{B} \nvdash_{\text{LTL}} \bot$), and
2. $\mathcal{B}$ is closed under LTL consequence (i.e. if $\mathcal{B} \vdash_{\text{LTL}} \varphi$ then $\varphi \in \mathcal{B}$).

Note that we do not allow for iteration of belief operators. For the one-agent case, this is not a severe restriction, at least for the most popular modal logic of belief KD45, since an agent's

[5] Note that our propositional and temporal languages include negation only on the level of atomic propositions. Indeed, we do not need negation elsewhere: as $\neg\Box\alpha$ is equivalent to $\Diamond\neg\alpha$, $\neg\Diamond\alpha$ is equivalent to $\Box\neg\alpha$, and $\neg \bigcirc\alpha$ is equivalent to $\bigcirc\neg\alpha$; and thanks to De Morgan laws, we can always transform a nested formula with negation in front of it to an equivalent formula with negations on the level of atoms only. This restriction is perhaps not necessary in the context of beliefs, but it becomes important when we discuss which options should be *intended* by the agent. In particular, we propose that more complex intentions should be decomposed into more primitive ones using so called *practical reasoning rules* (cf. Sect. 3.3). While such practical reasoning looks relatively straightforward for positive (i.e. non-negated) intentions—e.g. intending to $\bigcirc(eggs \wedge nicecoffee)$ should force the agent to intend both $\bigcirc eggs$ and $\bigcirc nicecoffee$ if the original intention is to be treated seriously—a direct decomposition of negative intentions would be rather unnatural.

We will sometimes abuse notation in the remainder of the paper by writing temporal operators in the scope of negations, especially when describing general properties of agents' mental states and processes. In such cases, the presumed meaning of $\neg\varphi$ is "the formula which is LTL-equivalent to $\neg\varphi$, and which includes negation only on the level of atoms".

beliefs can be characterised by a set of non-modal formulas [Meyer & van der Hoek, 1995, Proposition 3.1.2], and every belief sentence in KD45 is equivalent to one without nestings of belief operators (see for instance the proof of [Meyer & van der Hoek, 1995, Theorem 1.7.6.4], which is for S5, but does not in fact use the reflexivity axiom). Disallowing nestings keeps us closer to the AGM belief revision paradigm (Alchourrón et al., 1985), although we *do* allow for beliefs about temporal properties. In the multi-agent setting, iterated belief operators are not reducible any longer: we will come back to this issue in Sect. 5.

3.2 Desires

In the theory of BDI agents, desires have received far less attention than beliefs or intentions, and their exact status is often confused. From our point of view, an agent's desires represent different ways that he may want the world to look. That is, if φ is a desire, then the agent would like φ to be satisfied. Crucially, we do *not* require that sets of desires are consistent: one can desire to finish writing a paper and desire to go to a party while these desires cannot *jointly* be satisfied. Similarly, we do not require that an agent's desires are consistent with his beliefs. One can desire to be a famous rock star while believing that this will never happen. Of course, the same is not true for intentions, and this is in fact one of the key concerns of this paper.

Example 1 We introduce the running example that we will use throughout the remainder of the paper. Our agent is thinking about his breakfast for tomorrow. He desires a nice breakfast, and both a continental and an English one, and, should the breakfast tomorrow not provide cereals, our agent would have yogurt. Finally, he also desires to be energetic tomorrow and happy for the coming week. Formally:

$$\mathcal{D} = \{\bigcirc nicebf, \bigcirc contbf, \bigcirc Englishbf, \bigcirc energetic, \bigcirc(\neg cereals \rightarrow yogurt), \bigwedge_{i=1\ldots7} \bigcirc^i happy\},$$

where T^i is a shorthand for the temporal operator T repeated i times.

Suppose furthermore that the agent believes: that every breakfast is either continental or English, but never both; that he will never be able to have any tea; that he will manage to have an egg at most every even day starting from today; and that he will be able to fetch a newspaper at most every second day starting tomorrow. Thus, the set $\mathcal{B}$ must include the following:

$$\Box(contbf \leftrightarrow \neg Englishbf), \Box\neg tea, \bigcirc^{2n+1}\neg egg, \bigcirc^{2n}\neg newspaper, \bigcirc nicecoffee.$$

Note that the agent's belief set must be closed under logical consequence, so it should also include $\bigcirc(contbf \leftrightarrow \neg Englishbf)$, $\Diamond\neg tea$, etc. Note also that the set $\mathcal{D}$ is not consistent with the agent's beliefs $\mathcal{B}$ in this example.

3.3 Practical reasoning rules

How does an agent go from beliefs about what the world is like, and desires/intentions about how he wishes the world to be, to actions that shape and modify it? In the introduction, we discussed the idea of a practical syllogism—a reasoning pattern which says that if you believe that a certain action will bring about a certain desired effect then this action should be performed. Here, we use a closely related idea. We assume the agent to have a set of *practical reasoning rules*. These rules simply decompose an intention into a group of sub-intentions.

They have the general form:

$$\text{to achieve } \alpha, \text{ one should strive to achieve } \alpha_1, \ldots, \alpha_n.$$

Practical reasoning rules are thus a type of *goal reduction rule* (Kaelbling & Rosenschein, 1990, p. 39). Note that for any given α on the left hand side of a rule, an agent may have multiple practical reasoning rules, corresponding to the different ways that a given intention may be achieved.

More formally, the format of a practical reasoning rule ρ is as follows:

$$\rho : (\alpha \leftarrow \alpha_1, \alpha_2, \ldots, \alpha_n).$$

The practical reasoning rule $\rho : (\alpha \leftarrow \alpha_1, \alpha_2, \ldots, \alpha_n)$ expresses that if $T\alpha$ is an intention, then $T(\alpha_1 \wedge \alpha_2 \wedge \cdots \wedge \alpha_n)$ is a possible fulfilment of it. In our approach, the intention $T\alpha$ can be thus divided into sub-intentions $T\alpha_1, \ldots, T\alpha_n$, with the intended meaning that their joint presence is a good way of achieving $T\alpha$.

For a rule ρ as displayed above, we call $\alpha = hd(\rho)$ the *head* of the rule, and each $\alpha_i \in bd(\rho)$ a member of the *body* of ρ. Moreover, for a set of rules $\mathcal{P}$, we define $hd(\mathcal{P}) = \{hd(\rho) \mid \rho \in \mathcal{P}\}$ and $bd(\mathcal{P}) = \cup_{\rho \in \mathcal{P}} bd(\rho)$. A set of practical reasoning rules $\mathcal{P}$ is *acyclic* if there is no sequence of rules $\rho_1, \ldots, \rho_n \in \mathcal{P}$ and propositional formulae $\alpha_1, \ldots, \alpha_n$ such that $\alpha_i \in hd(\rho_i)$ for $i = 1, \ldots, n$, and $\alpha_{i+1} \in bd(\rho_i)$ for $i = 1, \ldots, n-1$, and $\alpha_1 \in bd(\rho_n)$.

Example 2 With respect to the "breakfast" scenario discussed above, here are the agents' practical reasoning rules:

$$\begin{array}{ll}
\rho_1 : \mathit{nicebf} & \leftarrow \mathit{milk}, \mathit{cereals}, \mathit{nicecoffee}, \\
\rho_2 : \mathit{nicebf} & \leftarrow \mathit{milk}, \mathit{cereals}, \mathit{tea}, \\
\rho_3 : \mathit{happy} & \leftarrow \mathit{newspaper}, \\
\rho_4 : \mathit{happy} & \leftarrow \mathit{egg}, \\
\rho_5 : \mathit{energetic} & \leftarrow \mathit{nicecoffee}, \\
\rho_6 : \neg\mathit{happy} & \leftarrow \mathit{bill}, \mathit{nicebf}, \\
\rho_7 : \mathit{nicecoffee} & \leftarrow \mathit{espresso}.
\end{array}$$

These rules express that a nice breakfast comprises of a combination of milk and cereals with either nice coffee or tea. Moreover, there are two ways to make the agent happy: either by providing him an egg or a newspaper. Finally, getting a bill during a nice breakfast makes the agent unhappy, and an espresso counts as a nice coffee.

We consider formulas that occur in the body of some rule ρ but not in the head of any rule, to be *atomic means* of some kind: they serve to satisfy subgoals of the agent, but cannot be decomposed themselves into further means. For instance, there may be several ways to make the agent *energetic*: it may occur on its own (i.e. without substituting *energetic* with the body of rule ρ_5), or by satisfying the "more primitive" property *nicecoffee* (which, in turn, can occur in its own or be decomposed to *espresso*). On the other hand, the only way to make *espresso* true is to have some espresso, since no practical reasoning rule applies to this proposition. We remark here that the presence of practical reasoning rules with the same head (like ρ_3 and ρ_4 in our example), will facilitate the agent to revise his intentions due to the "instrumentality of his beliefs":[6] our framework makes it possible, building upon Example 2, that the agent trades an intention $\bigcirc\mathit{newspaper}$ for $\bigcirc\mathit{egg}$, given the intention $\bigcirc\mathit{happy}$ and the belief $\Box\neg\mathit{newspaper}$ (that is, if the agent has the intention to be happy tomorrow, and

[6] This term was suggested by one of the referees.

finds out that one instrumental intention to achieve this is impossible to achieve—e.g. to read a newspaper—he can switch to the instrumental intention of having an egg).

Remark 1 One possible reading of rule $\rho : (\alpha \leftarrow \alpha_1, \ldots, \alpha_n)$ is that achieving $\alpha_1, \ldots, \alpha_n$ *guarantees* making α true. Note that $\bigcirc(\alpha_1 \wedge \cdots \wedge \alpha_n) \equiv \bigcirc\alpha_1 \wedge \cdots \wedge \bigcirc\alpha_n$, and $\Box(\alpha_1 \wedge \cdots \wedge \alpha_n) \equiv \Box\alpha_1 \wedge \cdots \wedge \Box\alpha_n$. So, under this reading, and for a temporal operator $T = \bigcirc, \Box$, bringing about $T\alpha_1, \ldots, T\alpha_n$ is *sufficient* to bring about $T\alpha$.

The situation is not so clear for "achievement" intentions, i.e. ones that define that a property should be achieved *sometime in the future* (e.g. $\Diamond$*nicebf*). Note that $\Diamond(\alpha_1 \wedge \cdots \wedge \alpha_n)$ implies $\Diamond\alpha_1 \wedge \cdots \wedge \Diamond\alpha_n$, but not the other way around. Thus, adding sub-intentions $\Diamond\alpha_1 \wedge \cdots \wedge \Diamond\alpha_n$ when decomposing intention $\Diamond\alpha$ with practical rule $\rho = (\alpha \leftarrow \alpha_1, \ldots, \alpha_n)$, we only add *necessary* (but not sufficient!) conditions to *bring about* $\Diamond\alpha$ *in the way suggested by* ρ. Therefore, in the case of "achievement" intentions, practical reasoning rules should be treated as *guidelines* that force the agent to accept necessary sub-intentions but do not guarantee success. For example achieving $\Diamond$*milk*, $\Diamond$*cereals* and $\Diamond$*nicecoffee* does not guarantee $\Diamond$*nicebf*, but the former is necessary to achieve the latter if the agent uses rule ρ_1 to decompose intention $\Diamond$*nicebf*.

Again, note that our language for practical reasoning rules is arguably weak: for instance, we cannot impose any temporal order within a plan: like in $\rho_1' :$ *nicebf* $\leftarrow$ *milk*, *cereals*, $\bigcirc$*nicecoffee*, expressing that a nice breakfast requires some nice coffee *after* having enjoyed milk with cereals. For the moment, we abstract from such temporal dependencies between sub-goals, and refer to the huge literature on planning in AI (Allen, Hendler, & Tate, 1990; Georgeff & Lansky, 1986).

What is the aim of practical reasoning rules? On one hand, they represent the agent's "know-how" about how particular properties can be achieved. On the other hand, we want to provide the agent's intentions with a kind of closure, in the sense that an intention to bring about a complex property φ forces the agent to accept more primitive intentions that would (or at least may) indeed bring about φ.

Notice that practical reasoning rules are *not* implications, and in particular, they are not implications that are part of the agents' *beliefs*. First of all, a believed implication $p \rightarrow q$ does not necessarily give a recipe for achieving q, for which taking $p =$ "set fire to the house" and $q =$"get warmer" are witnesses. Moreover, we do not want beliefs and practical reasoning rules to interact as implications do: ρ: *kill* $\leftarrow$ *shoot* seems a good recipe for action, although the corresponding implication *shoot* $\rightarrow$ *kill* seems rather far-fetched, so, e.g. inferring that $\neg$*shoot* from the fact that nobody was killed does not seem plausible. This emphasises the distinguishing feature of practical reasoning rules: they are used by the agent to plan his future actions, whereas beliefs, as we will see in Sect. 3.5, are typically about what will become true regardless of the agent's decisions.

However, just to complicate matters, some practical rules may be read as "achieving $\alpha_1, \ldots, \alpha_n$ guarantees making α true", in which case they should obviously have their corresponding implications $(\alpha_1 \wedge \cdots \wedge \alpha_n \rightarrow \alpha)$ in the agent's belief set $\mathcal{B}$. So, (*espresso* $\rightarrow$ *nicecoffee*) $\in \mathcal{B}$ means that *nicecoffee* $\leftarrow$ *espresso* $\in \mathcal{P}$ is not only a suggestion on how to choose one's action; it is actually a guarantee that if we achieve *espresso*, then *nicecoffee* will also be achieved.

3.4 Intentions

While the definition of a belief set as one, that is closed under consequences is commonly accepted in the belief revision literature (Alchourrón et al., 1985), the question of what properties a set of intentions should have enjoys far less agreement (Cohen & Levesque,

1990; Rao & Georgeff, 1998). We will proceed to build a model of intentions in steps, each step capturing one important aspect.

First, in our view, intentions are usually *relativized* (Cohen & Levesque, 1990). That is, an intention usually forms a small part of a larger canvas of intentions. For example, you may intend to work hard to write a paper, but there is a *reason* for this: you want the paper published in order to get promotion, for example. Thus, first of all, we consider intentions as structures consisting of some end that we want to achieve ("write papers"), and some reason, or motivation, for them ("get promoted").[7] By keeping track of the reasons for adopting our intentions, we can rationally update them as the future unfolds: once we believe we have been promoted, we can rationally drop our intention to write papers.[8]

Suppose the agent from Example 2 adopts the desire to have a nice breakfast and be energetic tomorrow. This should mean, among other things, that he has the intention to have milk, cereals and nice coffee tomorrow, and also espresso. Now suppose that the agent learns (comes to believe) that in fact it is impossible to have a nice breakfast tomorrow. We would expect him to give up the intention of having milk and cereals for the next day, but *not* that of having nice coffee, neither of having espresso. If, alternatively, he learned that there would be no nice coffee tomorrow, we would expect him to give up all the intentions concerning being energetic, whereas he might keep the intention of having a nice breakfast—although now with the sub-intentions of having milk, cereals and tea.

We need several concepts before defining what an intention is. Let $\mathcal{B}, \mathcal{D}$, and $\mathcal{P}$ be a set of beliefs, desires and practical reasoning rules, respectively. First, we define *plans*. Given an LTL formula $T\alpha$, π is a plan for $T\alpha$ if:

1. if there is a rule $\rho\colon (\alpha \leftarrow \alpha_1, \ldots, \alpha_n) \in \mathcal{P}$, then π is a tree with root $T\alpha$, and branches from the root to sub-trees $\pi_1, \ldots, \pi_n$, which are plans for $T\alpha_1, \ldots, T\alpha_n$, respectively;
2. else, π is a tree consisting solely of one node $T\alpha$.

The set of all plans for φ is denoted by $\Pi(\varphi)$. We assume the notions of leaf of a tree, parents, children and sub-trees to be familiar. We write $\varphi \rightsquigarrow_\pi \psi$ if in the tree π, node ψ occurs in the sub-tree generated by φ: in such a case we also say that φ is an *ancestor* of ψ. Given a tree π, $Nodes(\pi)$ collects all the nodes in it. Similarly, if Pl is a set of plans (trees), $Nodes(\mathsf{Pl}) = \cup_{\pi \in \mathsf{Pl}} Nodes(\pi)$. Note: in a particular plan π all the children of a node are generated by one practical reasoning rule. Also, π is assumed to be *full* in the sense that no rule is applicable to any of the leafs.

Theorem 1 *For each φ:*

1. $\Pi(\varphi)$ *is non-empty.*
2. *If $\mathcal{P}$ is acyclic and finite, then every $\pi \in \Pi(\varphi)$ is finite, and $\Pi(\varphi)$ is finite.*

Proof

1. The construction of such a $\pi \in \Pi(\varphi)$ is given in the above definition of a plan.
2. Consider a single branch from $\pi \in \Pi(\varphi)$. $\mathcal{P}$ is acyclic, so each rule from $\mathcal{P}$ could have been applied at most once along this branch. Thus, the length of branches in π is bounded by the number of rules in $\mathcal{P}$, and this is finite. Moreover, $\Pi(\varphi)$ contains plans of bounded length with nodes labeled with formulas from a finite set (all the propositional formulas that occur in any rule from $\mathcal{P}$). Therefore $\Pi(\varphi)$ can include only a finite number of plans.

□

[7] For more on such 'instrumental goals', see e.g. (Dignum, Dunin-Keplicz, & Verbrugge, 2001; Dignum & Conte, 1998).

[8] But please do not tell our managers we said this.

Definition 1 Let $\mathcal{T}$ be a set of trees, i.e. plans. Given the belief set $\mathcal{B}$, we define a labelling $\lambda(\varphi, \pi, \mathcal{B}) \in \{+, -, \bot\}$ for every node φ of π, given the beliefs $\mathcal{B}$. The label "+" means that φ is not believed to be fulfilled yet (nor has it ceased to be instrumental to some higher-level intention), and hence it is a proper intention. The label "−" for φ means that either φ is already believed to be true, or it is needed in π for a node that turns out to be believed true. Formally:

$$\lambda(\varphi, \pi, \mathcal{B}) = \begin{cases} \bot \text{ if } \varphi \notin \mathit{Nodes}(\pi), \\ - \text{ if } \mathcal{B} \vdash_{\text{LTL}} \varphi, \\ - \text{ if } \exists \psi \in \mathit{Nodes}(\pi) \text{ s.t. } \psi \rightsquigarrow_{\pi} \varphi \,\&\, \lambda(\psi, \pi, \mathcal{B}) = -, \\ + \text{ otherwise.} \end{cases}$$

Moreover, given a plan π, $\mathit{Nodes}^{+}(\pi, \mathcal{B})$ is the set of nodes $\varphi \in \pi$ for which $\lambda(\varphi, \pi, \mathcal{B}) = +$. For a *set* of plans PI, $\mathit{Nodes}^{+}(\mathsf{PI}, \mathcal{B})$ is the set of nodes for which there is a $\pi \in \mathsf{PI}$ such that $\lambda(\varphi, \pi, \mathcal{B}) = +$.

Intentions will be the nodes labelled with a + in a set of plans, but not every set of plans can be kept by a rational agent:

Definition 2 Given $\mathcal{B}, \mathcal{D}$ and $\mathcal{P}$, a set of Active Plans $\mathcal{A}$ is a set of trees such that:

1. $\mathcal{B} \cup \mathit{Nodes}^{+}(\mathcal{A}, \mathcal{B})$ is consistent;
2. For every $\pi \in \mathcal{A}$, $\text{root}(\pi) \in \mathcal{D}$ and $\lambda(\text{root}(\pi), \pi, \mathcal{B}) = +$;
3. There is no $\varphi \in \mathcal{D}$ with a plan π for φ such that $\mathcal{A} \cup \{\pi\}$ satisfies items 1 and 2.

Definition 3 Given a belief set $\mathcal{B}$ and a set of active plans $\mathcal{A}$, we define $\mathit{Int}(\mathcal{A}, \mathcal{B}) = \mathit{Nodes}^{+}(\mathcal{A}, \mathcal{B})$, and call this the *intentions* of the agent, generated by $\mathcal{B}$ and $\mathcal{A}$. The agent's current active desires are $\mathit{Des}(\mathcal{A}) = \{\varphi \mid \exists \pi \in \mathcal{A} : \text{root}(\pi) = \varphi\}$.

Coming back to our example, let the agent have the desires $\bigcirc\mathit{nicebf}$ and $\bigcirc\mathit{energetic}$, and consider the plan π_1 for $\bigcirc\mathit{nicebf}$ that uses ρ_1 and ρ_7, and the plan π_2 for $\bigcirc\mathit{energetic}$ that uses ρ_5 and ρ_7 (see Fig. 2). In principle, all the nodes of both plans π_1 and π_2 are candidates to be intentions for the agent. But not so if the agent believes them to already be guaranteed: if $\bigcirc\mathit{cereals} \in \mathcal{B}$, the agent does not have $\bigcirc\mathit{cereals}$ as an intention, and, that is why $\lambda(\bigcirc\mathit{cereals}, \pi_1, \mathcal{B}) = -$. We do not remove this leaf from π_1, however, since the agent might give up his belief that cereals are guaranteed, and may want to re-instantiate this intention (we come back to this later). Now here is the advantage of representing the intention $\bigcirc\mathit{nicecoffee}$ twice: suppose the agent would learn that $\bigcirc\mathit{nicebf}$, and hence include this formula in $\mathcal{B}$. It would mean that all the nodes of π_1 become labelled "−", but $\bigcirc\mathit{nicecoffee}$ and $\bigcirc\mathit{espresso}$ should stay intentions: since they are needed for an intention different from $\bigcirc\mathit{nicebf}$, they cannot be given up totally.

Item 1 of Definition 2 requires that the current active intentions have to be consistent with what the agent believes: it does not make sense to intend $\bigcirc\mathit{espresso}$ if you already believe

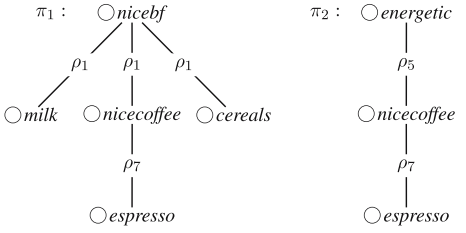

Fig. 2 Two plans

$\Box\neg$*espresso*. However, the plan π_1 still may make sense even if $\Box\neg$*espresso* is believed: if the agent also believes $\bigcirc$*nicecoffee*, he might still intend $\bigcirc$*nicebf*, it just would mean that $\bigcirc$*nicecofee* is guaranteed somehow, and the agent does not have to bother about specific ways to achieve it (technically, $\bigcirc$*espresso* would be labelled "$-$" in π_1, since it is the child of the "$-$" node $\bigcirc$*nicecoffee* in π_1). Item 2 guarantees that an agent eventually only chases up desires, but never those that are already believed to be fulfilled. Finally, regarding item 3 of Definition 2, a rational agent tries to pursue as many desires as possible, respecting the previous two conditions.

We add an additional requirement on agents which has a repercussion on their active plans. Let $\rho : \alpha \leftarrow \alpha_1, \ldots \alpha_n$. An agent is said to *believe ρ to be correct with respect to T* if, whenever $T\alpha_1, \ldots, T\alpha_n$ are in $\mathcal{B}$, then so is $T\alpha$. This corresponds closely to the notion of practical reasoning rules (discussed in Sect. 3.3) as ones that would *guarantee* $T\alpha$, had $T\alpha_1, \ldots, T\alpha_n$ been achieved. Consider an agent with plan π_1 (from Fig. 2) in his set of active plans $\mathcal{A}$, who comes to believe $\bigcirc$*milk* $\wedge \bigcirc$*nicecoffee* $\wedge \bigcirc$*cereals*. If the agent believes rule *nicebf* $\leftarrow$ *milk*, *cereals*, *nicecoffee* to be correct with respect to $\bigcirc$, he should be expected to stop working on plan π_1, and instead use his resources elsewhere. An agent believes his plans are correct if this is the case for all rules ρ used in Pl. It is not difficult to see that the labelling λ for any agent that believes his plans are correct would be "upward closed" with respect to "$-$": that is: for such an agent, for any $\mathcal{A}$, if λ assigns "$-$" to all the children of a node $\varphi \in \pi$, then it also assigns "$-$" to φ itself. Unless stated otherwise, from now on, we will assume that agents believe their plans are correct with respect to all temporal operators $T = \bigcirc, \Diamond, \Box$.

Remark 2 We have chosen to define intentions $Int(\mathcal{A}, \mathcal{B})$ given a maximal set of current plans $\mathcal{A}$ that are in focus by the agent. Alternatively, we could have defined intentions $\mathcal{I}$ as a primitive and then expect the agent to generate a set of plans $\mathcal{A}$ satisfying $\mathcal{I} = Int(\mathcal{A}, \mathcal{B})$; this would correspond to the "conventional" AI planning approach, of generating a plan to achieve some end given beliefs about how the world currently is. Indeed, these two approaches seem interdependent: an agent choosing a set of intentions should guarantee that they are mutually consistent with his future beliefs and that one intention possibly triggers another; he should implicitly have some set of active plans for them. It is not hard to see that active plans, given the agent's desires $\mathcal{D}$, beliefs $\mathcal{B}$ and practical reasoning rules $\mathcal{P}$, give a unique set of intentions, where, conversely, given sets $\mathcal{I}, \mathcal{D}, \mathcal{B}$ and $\mathcal{P}$, there may be in general several $\mathcal{A}$'s for which $\mathcal{I} = Int(\mathcal{A}, \mathcal{B})$. This again, we claim, makes sense: two agents with different current plans can have the same actual intentions.

Given a set of plans $\mathcal{A}'$, and sets $\mathcal{B}, \mathcal{D}, \mathcal{P}$, the agent can build his active plan set $\mathcal{A}$ (and hence, his intentions $\mathcal{I} = Int(\mathcal{A})$) from $\mathcal{A}'$ in two ways: he can either chose as many different desires φ from $\mathcal{D}$ and add a plan for φ to $\mathcal{A}'$, or, alternatively, he may decide to maximize the number of alternative plans for some specific desires. An abstract non-deterministic algorithm to build such a set of plans $\mathcal{A}$, given some initial plans $\mathcal{A}'$, is given in Fig. 3.

The process of plan formation can be also seen as follows. Let $choices(\mathcal{A}', \mathcal{B}, \mathcal{D}, \mathcal{P})$ return all $\mathcal{A} \supseteq \mathcal{A}'$ such that $\mathcal{A}$ is an active plan set with respect to $\mathcal{B}$, $\mathcal{D}$ and $\mathcal{P}$, and let $commit(\cdot)$ select one of them. Now, $activePlans(\mathcal{A}', \mathcal{B}, \mathcal{D}, \mathcal{P})) = commit(choices(\mathcal{A}', \mathcal{B}, \mathcal{D}, \mathcal{P}))$. The selection of one of the candidate $\mathcal{A}$'s uses some kind of *commitment function* by the agent (cf. the selection functions of AgentSpeak Rao, 1996). A commitment function will thus characterise an agent's *focus of attention* in a particular domain. We will not be concerned with the properties of commitment functions in this paper, and simply assume that one is given.

Fig. 3 Generating an active plan set $\mathcal{A}$ given $\mathcal{A}', \mathcal{B}, \mathcal{D}, \mathcal{P}$

```
Algorithm: activePlans(·,·,·,·)
Input: A', B, D and P
Output: a set of active plans A

1.  A := A';
2.  while   ∃φ ∈ D \ B and ∃π ∈ Π(φ) s.t. A ∪ {π}
3.          satisfies items 1 and 2 of Definition 2
4.  do choose such a φ and π;
5.     A := A ∪ {π};
6.  end-while
```

Theorem 2 *Let $\mathcal{A}', \mathcal{B}, \mathcal{D}, pr$ be given, and suppose that $\mathcal{B}$ is consistent.*

1. *If $\mathcal{A}'$ is a set of plans satisfying conditions 1 and 2 of Definition 2, and activePlans$(\mathcal{A}', \mathcal{B}, \mathcal{D}, \mathcal{P})$ terminates, then the result is an active plan set.*
2. *If activePlans$(\emptyset, \mathcal{B}, \mathcal{D}, \mathcal{P})$ terminates, then the result is an active plan set.*
3. *Every finite active plan set $\mathcal{A}$ for $\mathcal{D}$, $\mathcal{B}$ and $\mathcal{P}$ can be generated by activePlans$(\emptyset, \mathcal{B}, \mathcal{D}, \mathcal{P})$.*
4. *Let $\mathcal{D} \setminus \mathcal{B}$ be finite, and $\mathcal{P}$ finite and acyclic. Then, activePlans$(\mathcal{A}', \mathcal{B}, \mathcal{D}, \mathcal{P}))$ terminates for any $\mathcal{A}'$, and for a finite $\mathcal{A}'$ it returns a finite set of plans.*

Proof

1. Let $\mathcal{A}'$ be a set of plans satisfying items 1 and 2 of Definition 2. The condition on lines `2.` and `3.` of the algorithm guarantees that $\mathcal{A}'$ will never be extended in the way that violates these items. Moreover, the stop condition of the `while` loop guarantees that the resulting set $\mathcal{A}$ satisfies item 3 of Definition 2, so that it is indeed an active plan set.
2. This follows immediately from (1): given that $\mathcal{B}$ is consistent, the empty set already satisfies the conditions 1 and 2 from Definition 2.
3. If $\mathcal{A}$ is finite, one can add the plans in it to the empty set one by one using the algorithm.
4. By Theorem 1.2, and the fact that $\mathcal{D} \setminus \mathcal{B}$ is finite, only a finite number of plans can be added to $\mathcal{A}'$. □

Function *activePlans* will become important in Sect. 4, where we discuss intention revision.

3.5 Describing mental states

A mental state

$$\Gamma = \langle \mathcal{B}, \mathcal{D}, \langle \mathcal{I}, \mathcal{A} \rangle, \mathcal{P} \rangle$$

collects the agent's beliefs, desires, intentions with active plans and reasoning rules. We will write $\mathcal{B}(\Gamma)$, $\mathcal{D}(\Gamma)$, $\mathcal{P}(\Gamma)$, $\mathcal{I}(\Gamma)$ and $\mathcal{A}(\Gamma)$ to denote the components of Γ when it is not clear from the context. In what follows, when we refer to a mental state, we will assume that the components of this state satisfy the constraints specified in the preceding sections; in particular we assume $\mathcal{I} = Int(\mathcal{A}, \mathcal{B})$. Here, we summarise the basic intuitions:

- beliefs are about the future, and must be mutually consistent;
- desires are also about the future, but need not be consistent;
- practical reasoning rules define how intentions can be decomposed into sub-intentions; and
- a pair $\langle$*intentions*, *active plans*$\rangle$ represents a complete, consistent picture of how an agent will act towards a chosen sub-set of his desires, that is maximal among those that are consistent with the agent's beliefs.

To describe mental states $\langle \mathcal{B}, \mathcal{D}, \langle \mathcal{I}, \mathcal{A} \rangle, \mathcal{P} \rangle$, we introduce a *mental state language*. This is a language used by us, as external observers, to characterise the state of agents. It will allow us to talk about an agent's beliefs, desires and intentions; in this sense the language is similar to the intention logic of Cohen–Levesque, the BDI logic of Rao–Georgeff, and other similar formalisms (Cohen & Levesque, 1990; Rao & Georgeff, 1998; Wooldridge, 2000). However, in this paper, we do not consider the issue of nesting mental state expressions. Thus, as we will see, an agent does not have beliefs *about* beliefs, or beliefs about desires, etc.[9]

Let $\Gamma = \langle \mathcal{B}, \mathcal{D}, \langle \mathcal{I}, \mathcal{A} \rangle, \mathcal{P} \rangle$ be a mental state, as described above, and, as before, let α be an objective formula, φ be a temporal formula, and let ω denote a formula which is either objective or temporal (i.e. ω is either of type α or φ). We now define the mental state language MS, with typical element γ:

$$\gamma := \mathrm{WB}\,\omega \mid \mathrm{SB}\,\omega \mid D\varphi \mid I\varphi \mid \neg\gamma \mid \gamma \vee \gamma.$$

Here, WB ω means that the agent *weakly believes* ω, while SB ω means that the agent *strongly believes* ω; we will discuss the distinction between weak and strong belief shortly. $D\varphi$ means that the agent desires φ. Finally, $I\varphi$ means that the agent has the intention φ, in the sense that he will act upon it, and not doing so will not guarantee φ.

To understand the distinction between weak and strong belief, consider that in the literature there has been a lot of discussion on the relation between beliefs and intentions. A property usually embraced, and considered seriously by us, is that, for "optional formulas" φ, one has $I\varphi \rightarrow \mathrm{WB}\varphi$. The rationale behind this property is as follows: if the agent intends $\Box p$, then all the actions he choses in the future will maintain p; if he intends $\Diamond p$ then the agent will sometime see to it that p; and if he intends $\bigcirc p$, then he has committed himself to acting in such a way now such that in the next state p. In other words, under this reading, φ is believed if, assuming the agent acts according to his commitments, φ holds. Thus, for example, $\Box p$ need be not believed because p is inevitable, but may as well be believed because in all futures *that the agent envisages*, p is true. It is easy to see that axiom $I\psi \rightarrow \mathrm{WB}\psi$ entails a very important feature of the agent: the future is under control of the agent. Or, more precisely, the agent can only intend things over which he can exert control. While this may seem a reasonable simplification at first glance, such an assumption is in fact ultimately unrealistic. For most (if not all) things one intends to achieve, there is no guarantee of success. For all human agents, for example, there is always a possibility that the agent will die in the next moment, which would bring the ultimate failure to all his intentions (unless he actually intended to die, of course). In consequence, we might say that intending to eventually die would be the only valid intention for such an agent; for other properties, the agent does not have sufficient control over them. The notions of weak and strong belief are introduced to find a way out of this conundrum. That is, apart from *weak* (or *optimistic*) belief operator WB that assumes that the agent's intentions are bound to succeed, we also define *strong* (or *pessimistic*) belief operator SB that sticks to the "hard facts" that the agent has in his belief base $\mathcal{B}$. We characterise the relationship between these notions of belief in Theorem 7, below.

We define the truth of a mental state formula γ in a mental state $\Gamma = \langle \mathcal{B}, \mathcal{D}, \langle \mathcal{I}, \mathcal{A} \rangle, \mathcal{P} \rangle$ in the following way (recall that $\vdash_{\mathrm{LTL}}$ denotes derivability in LTL).

$\Gamma \models \mathrm{WB}\varphi$ iff $\mathcal{B} \cup \mathcal{I} \vdash_{\mathrm{LTL}} \varphi$,
$\Gamma \models \mathrm{SB}\varphi$ iff $\mathcal{B} \vdash_{\mathrm{LTL}} \varphi$,
$\Gamma \models D\varphi$ iff $\varphi \in \mathcal{D}$,
$\Gamma \models I\varphi$ iff $\exists \varphi' \in \mathcal{I}$ s.t. $\vdash_{\mathrm{LTL}} \varphi \leftrightarrow \varphi'$.

[9] In Dennett's terminology, our agent is a *first-order intentional system* (Dennett, 1987). One can also see a close relationship to *0-level agents*, defined by Vidal and Durfee (1998).

We define validity in the conventional way: $\models \gamma$ means γ is true in all mental states Γ. Note that the optimistic and realistic beliefs with respect to the current moment coincide, i.e. $\models \mathrm{WB}\alpha \leftrightarrow \mathrm{SB}\alpha$ is a validity (where α is objective). This makes sense: if it is about the current state, it is too late to interfere; the agent's intentions will not be able to change the present.

We briefly give some validities concerning beliefs and intentions when considered separately.

Theorem 3

1. $\models \mathrm{SB}\varphi \rightarrow \mathrm{WB}\varphi$,
2. $\models \neg\mathrm{SB}\bot \wedge \neg\mathrm{WB}\bot$,
3. $\models \neg I\bot$,
4. $\not\models (I\varphi \wedge I\psi) \rightarrow I(\varphi \wedge \psi)$,
5. $\vdash_{\mathrm{LTL}} \varphi \rightarrow \psi$ *does not imply* $\models I\varphi \rightarrow I\psi$.

Proof

1. Follows from monotonicity of LTL: whatever is derivable from $\mathcal{B}$, is derivable from $\mathcal{B} \cup \mathcal{I}$.
2. By (1), it is sufficient to demonstrate that $\models \neg\ \mathrm{WB}\ \bot$. Let $\Gamma = \langle \mathcal{B}, \mathcal{D}, \langle \mathcal{I}, \mathcal{A}\rangle, \mathcal{P}\rangle$ be an arbitrary mental state. $\Gamma \models\ \mathrm{WB}\ \bot$ would mean that $\mathcal{I} \cup \mathcal{B} \vdash_{\mathrm{LTL}} \bot$, i.e. $\mathcal{I} \cup \mathcal{B}$ would be LTL-inconsistent. However, $\mathcal{B}$ is by definition consistent, and, by item 2 of Definition 2, also $\mathcal{B} \cup \mathcal{I}$ must be consistent (note that $\mathcal{I} = \mathit{Int}(\mathcal{A}) = \mathit{Nodes}^{+}(\mathcal{A}, \mathcal{B})$).
3. Since $\mathcal{B} \cup \mathcal{I}$ is consistent (see above), by monotonicity of LTL $\mathcal{I}$ is also consistent, which proves this property.
4. The point here is that $\mathcal{I}$ does not have to be closed under any kind of logical consequence. As an example, consider the tree π_1 in Fig. 2: if only validities are believed, and $\mathcal{A}$ would just contain π_1, then we have $I \bigcirc \mathit{milk}$ and $I \bigcirc \mathit{espresso}$, but there is no node representing $(\bigcirc \mathit{milk} \wedge \bigcirc \mathit{espresso})$, hence the latter conjunction is not an intention.
5. Again, this follows from the fact that the nodes in a plan do not have to form any logically closed set. A node could be labelled $\varphi = (\bigcirc \mathit{milk} \wedge \bigcirc \mathit{sugar})$, and if the agent does not have a rule to reduce this to the two parts $\psi = \bigcirc \mathit{milk}$ and $\bigcirc \mathit{sugar}$ that constitute this intention, we have $I\varphi$, and $\vdash_{\mathrm{LTL}} \varphi \rightarrow \psi$, but not necessarily $I\psi$. □

Item (1) of Theorem 3 illustrates why we refer to SB as strong belief and WB as weak belief. It says that if the agent believes that φ will happen no matter how he will act, then he also believes φ happens no matter what his intentions are. Item 2 guarantees that both weak and strong beliefs are consistent, and 3 says that this is also true for intentions. Item 4 says that combining two intentions does not necessarily yield a new one, and, finally, the last item of Theorem 3 says that intentions are not closed under LTL-consequence, which is in accordance to Bratman's and Cohen and Levesque's observation that agents need not intend all the expected side effects of their intentions (cf. p. 3).

Notice that we immediately get the following two properties, which again make the difference between the two types of beliefs clear. Suppose that the agent intends φ, i.e. we have $I\varphi$. This implies two opposite facts for his beliefs: first of all, since the agent intends φ, he then weakly believes φ. However, concerning his strong beliefs, i.e. his beliefs about what the world would look like if he did not act, we have $\neg\ \mathrm{SB}\varphi$. To wit, were the agent to already strongly believe that φ would come about, there would be no sense in adopting it as an intention! Thus, regarding connections between beliefs and intentions, we have the following.

Theorem 4

1. $\models I\varphi \rightarrow \mathrm{WB}\varphi$,
2. $\models I\varphi \rightarrow \neg\mathrm{SB}\varphi$,
3. $\models I\varphi \rightarrow \neg\mathrm{WB}\neg\varphi \wedge \neg\mathrm{SB}\neg\varphi$.

Proof

1. As in the proof of item 1 of Theorem 3, this follows from monotonicity of LTL.
2. Let $\Gamma = \langle \mathcal{B}, \mathcal{D}, \langle \mathcal{I}, \mathcal{A}\rangle, \mathcal{P}\rangle$ be a mental state and suppose that $\Gamma \models I\varphi$. This means that φ occurs in a tree $\pi \in \mathcal{A}$ for which $\lambda(\varphi, \pi, \mathcal{B}) = +$. But in order to receive this label, we have $\mathcal{B} \not\vdash_{\mathrm{LTL}} \varphi$, i.e. $\Gamma \models \neg\,\mathrm{SB}\varphi$.
3. By (1) and consistency of $\mathcal{B} \cup \mathcal{I}$, we get that $I\varphi \rightarrow \neg\,\mathrm{WB}\neg\varphi$. By Theorem 3.1 and contraposition, we also get that $I\varphi \rightarrow \neg\,\mathrm{SB}\neg\varphi$. □

Property (1) shows that weak beliefs define a very particular notion of realism (to which we refer in this paper as *optimism*): everything the agent intends to achieve is weakly believed to be achieved. Moreover, by (2), if the agent intends φ, he is aware that this implies a commitment from his side: he cannot assume that φ will be brought about no matter how he acts. Neither does he believe (item (3)) that φ is impossible: an agent who intends $\Diamond p$ does not strongly believe that $\Box\neg p$. Note that items 2 and 3 together yield $\models (\mathrm{SB}\,\varphi \vee \mathrm{SB}\,\neg\varphi) \rightarrow \neg\mathrm{I}\varphi$: if the agent has any strong beliefs about the truth value of φ, he cannot adopt it as an intention.

Another way of seeing weak beliefs is that they represent what the agent should believe under the *working assumption* that he will succeed. For instance (to use an example of Bratman (1987)), if you intend to play basketball this afternoon, then you can work on the assumption that you *will* play basketball—you do not assume that you will fail.

There are a number of stories that may be (re-)told to further illustrate and motivate these two notions of belief. In the story of *Little Nell* (McDermott, 1982), a hero, Dudley, is facing a disaster: little Nell is tied to a railway track, and a train is approaching. Dudley realises that if he does not act ($\neg a$), something horrible (h) will happen to Nell. So he devises a plan to get there quickly and rescue her. While he is travelling, he has time to reflect on his plan, and concludes that since he has decided to act, Nell will indeed be safe now after all. But this, in turn, undercuts the need for him to take action. And so he drops the intention to act. But then... In our framework, we do not run into the problem of iterated adoptions and rejections of the intention. To be more precise, we could specify that $\neg\mathrm{SB}\ \Box\neg h \wedge Ia \wedge \mathrm{WB}\ \Box\neg h$, which says that, given what is inevitable no matter what the agent does, he considers it possible that she will eventually be hurt ($\neg\mathrm{SB}\ \Box\neg h$ is equivalent to $\neg\mathrm{SB}\ \neg\Diamond h$), but, taking into account his own intentions, he believes Nell will indeed always be fine.

Another "paradox" that pops up when an agent bases his intentions on his beliefs and vice versa, is found in *knowledge-based programs* (Fagin, Halpern, Moses, & Vardi, 1995). Consider an agent that controls a variable x, which is currently set to 0. The agent then conditionally is programmed to write the value 1 in the next step in x, namely only if he knows the value of x will not be set to 1 in the next round. Again, this paradox is easily solved by distinguishing beliefs SB that refer to how the world would be no matter how the agent acts, and the beliefs WB in which the agent takes into account his intentions. And it seems reasonable to make the agent's actions conditional on both!

Having the two notions of belief may also shed some light on the difference between two influential approaches to the formalisation of intentions, beliefs and other mental attitudes, namely those of Cohen & Levesque (1990) and Rao & Georgeff (1991).

Cohen and Levesque (1990) takes the notion of *goal* as primitive, and distinguish goals from intentions. They embrace the property $BELIEVES\varphi \rightarrow GOAL\varphi$, and also define a

notion of realism: what the agent chooses must not be ruled out by his beliefs ($GOAL\varphi \rightarrow \neg BELIEVES\neg\varphi$). On the other hand (Rao & Georgeff, 1991) requires that $GOAL\varphi \rightarrow BELIEVES\varphi$. It seems that Rao and Georgeff have our weak notion of goal WB in mind, i.e. beliefs that take into account the agent's plans. At the same time, Cohen and Levesque's realism is related to our strong notion of belief: one should not intend to do anything he already knows will fail. Moreover, the notion of belief that guarantees $BELIEVES\varphi \rightarrow GOAL\varphi$ seems to correspond to our $\mathrm{WB}\,\varphi \wedge \neg \mathrm{SB}\,\varphi$: φ is believed to be true taking into account our intentions, but not believed to be true per se.

Regarding properties of desires, since in our approach desires can be an arbitrary set of sentences, they do not have to obey any logical property. They can be inconsistent, and need not be closed under any operation. Desires that are not chosen to become intentions (i.e. formulas φ for which $D\varphi \wedge \neg I\varphi$ holds) could have been left out from $\mathcal{I}$ for two reasons: either φ is already believed to be true, or else each plan π for φ contains some nodes that are believed to be incompatible with already chosen intentions. We summarise this in the following theorem.

Theorem 5 $\models D\varphi \rightarrow (I\varphi \vee \mathrm{SB}\,\varphi \vee \bigwedge_{\pi\in\Pi(\varphi)} \mathrm{WB}\,\neg \bigwedge_{\psi\in Nodes(\pi)} \psi)$.

Proof Suppose $\Gamma = \langle \mathcal{B}, \mathcal{D}, \langle \mathcal{I}, \mathcal{A}\rangle, \mathcal{P}\rangle \models D\varphi$. If $\varphi \notin \mathcal{I}$, then according to Definition 2, item 3, φ has no plan π such that $\mathcal{A} \cup \{\pi\}$ satisfies items 1 and 2. So, for any π, either item 2 does not hold (in which case we have $\lambda(\varphi, \pi, \mathcal{B}) = -$, i.e. $\Gamma \models \mathrm{SB}\varphi$), or else item 2 is false, and in this case $\mathcal{B} \cup \mathcal{I} \cup Nodes^{+}(\pi, \mathcal{B})$ must be inconsistent. By monotonicity of LTL, also $\mathcal{B} \cup \mathcal{I} \cup Nodes(\pi)$ is inconsistent, and thus $\mathcal{B} \cup \mathcal{I} \vdash \neg \bigwedge_{\psi\in Nodes(\pi)} \psi$. □

4 Dynamics of beliefs and intentions

Now, we come to the issue of *revising* intentions. We take the point of view that an agent should maintain his beliefs and intentions in a rational way. That is, after every revision of beliefs and intentions, the agent's mental state should be rational, in the sense that it should satisfy the properties we described above. However, an agent's desires are, so far as we are concerned, not under his control. In fact, we assume that the agent's desires remain fixed; we will also assume that his set of practical reasoning rules $\mathcal{P}$ is fixed.

Of course, these are all simplifying assumptions. A *learning agent* would typically also update his practical reasoning rules (cf. Skubch & Thielscher, 2005, where the strategies that are learned by the agent are in fact close to our reasoning rules), or at least want to discriminate them according to the likelyhood of the success. Moreover, regarding the agent's desires:

> Preferences are not static, but they can change through commands of moral authorities, suggestions from friends, or just changes in our own evaluation of the world and our possible actions (van Benthem & Liu, 2005).

Indeed, we think that, were a mechanism for dynamic desires at hand, we could relatively easy deal with that in our framework: as will become clear below, an agent's intentions depend on his desires, but not necessarily the other way around. As the quote above suggests, we think the change of desires of an agent will typically be triggered by something outside the agent, like a change in utility, or a new encouragement. Since a satisfying mechanism of this kind has not been proposed yet, we refrain from studying the dynamics of desires here.

Summarising, our agent should revise $\mathcal{B}$ and $\langle \mathcal{I}, \mathcal{A} \rangle$, given new information about the (future of the) world, and given his (constant) desires and reasoning rules. Suppose that the agent's mental state $\Gamma_0 = \langle \mathcal{B}_0, \mathcal{D}, \langle \mathcal{I}_0, \mathcal{A}_0 \rangle, \mathcal{P} \rangle$ is given. Now, the agent can do some observations, on the basis of which he will adjust his mental state to a new one, say $\Gamma_1 = \langle \mathcal{B}_1, \mathcal{D}, \langle \mathcal{I}_1, \mathcal{A}_1 \rangle, \mathcal{P} \rangle$. It is *this* adjustment that we want to describe here.

Example 3 Let $\Gamma_0 = \langle \mathcal{B}_0, \mathcal{D}, \langle \mathcal{I}_0, \mathcal{A}_0 \rangle, \mathcal{P} \rangle$ where $\mathcal{B}_0$ is as in Example 1 and $\mathcal{P}$ as in Example 2, and $\mathcal{A}_0$ contains the plans of Fig. 2. Suppose the agent discovers that it is impossible to obtain milk for tomorrow, and he adjusts his beliefs accordingly: $\mathcal{B}_1$ will contain $\neg \bigcirc \mathit{milk}$. Regarding the intentions, we expect that $\bigcirc \mathit{milk}$ will be given up, and so will $\bigcirc \mathit{nicebf}$. How about the intention to have cereals tomorrow? Since the (only) reason to have them was to have a nice breakfast, $\bigcirc \mathit{cereals}$ will be given up as well. But as to the intention to have nice coffee tomorrow, although it was needed for a nice breakfast, it is not given up, since it is also needed for the intention to be energetic tomorrow! Also note that, by the fact that agents intend to satisfy as many desires as possible, giving up the intention $\bigcirc \mathit{cereals}$ makes room for a new intention $\bigcirc \mathit{yogurt}$.

Importantly, we will assume that this transition *does not take time*, i.e. that mental updates are instantaneous, and the time flow is only associated with the agent's external actions. Of course, the aim of the agent's intentions is to govern his external behaviour: he must *act* to get all his intentions fulfilled. Were we to model temporal change, enforced by execution of external actions, we would assume that executing an atomic action transforms mental state $\langle \mathcal{B}_1, \mathcal{D}, \langle \mathcal{I}_1, \mathcal{A}_1 \rangle, \mathcal{P} \rangle$ into a new state $\Gamma'_0 = \langle \mathcal{B}'_0, \mathcal{D}, \langle \mathcal{I}'_0, \mathcal{A}'_0 \rangle, \mathcal{P} \rangle$ in which the agent's beliefs, desires and active plans get "discounted" with a single time step: beliefs $\bigcirc \alpha$ from $\mathcal{B}_1$ re-appear as α in $\mathcal{B}'_0$, belief $\Diamond \alpha$ in $\mathcal{B}_1$ becomes $\alpha \vee \Diamond \alpha$ in $\mathcal{B}'_0$, whereas a belief $\Box \alpha$ does not change—and similarly for the other attitudes.

Also note that external actions usually (but not necessarily always!) bring new observations. In this case, the change triggered by executing such an action can be modelled as a composition of the temporal change *and* the mental state update with the observations.

4.1 Revising intentions

A key issue in the theory of intentions is that of the extent to which an agent remains *committed* to his intentions (Cohen & Levesque, 1990). That is, what rules should govern when an agent drops an intention? If the intentions are dropped too easily, we would be inclined to say that the agent did not really have an intention at all. But likewise, there *are* circumstances under which it makes sense to drop an intention: when one discovers that the intention has been fulfilled, or is impossible to fulfil, for example.

We want to capture some of these concerns in our language. Our design of the language is first of all motivated by simplicity. Note for instance that we still cannot express how the agents mental attitudes evolve over time; no mental state operator WB , D, or I is supposed to be in the scope of a temporal one. As a consequence, we cannot easily express the commitment strategies discussed in Rao & Georgeff (1991) in our object language. This does not mean that these strategies cannot be characterised at all: we show how a *blindly committed* agent can be specified at the end of Sect. 4.3. Nevertheless, we propose that, when it comes to the dynamics of beliefs and intentions, one should first give an account of the *reasons* to adopt or drop a belief or an intention, and the *way* this is done.

We now describe how an agent drops an intention. Suppose the initial mental state is $\Gamma_0 = \langle \mathcal{B}_0, \mathcal{D}, \langle \mathcal{I}_0, \mathcal{A}_0 \rangle, \mathcal{P} \rangle$, and the agent, given some observations, updates his beliefs from $\mathcal{B}_0$ to $\mathcal{B}_1$. The change in strong beliefs should trigger a change in intentions, which, in

Fig. 4 Cleaning an active plan

```
Algorithm: cleanup(·,·,·,·)
Input: B1, A0, D, and P
Output: a set of plans T

1. T := ∅;
2. forall π ∈ A0
3.    do re-label all nodes φ of π using B1;
4.    if λ(root(π), π, B1) = + AND π does not B1-fail
6.       do T := T ∪ {π}
6.       else skip
7.    end-if
8. end-forall
```

turn, should affect weak beliefs of the agent. We say that a plan π $\mathcal{B}$-*fails* if there is some $\varphi \in \mathit{Nodes}(\pi)$ such that $\mathcal{B} \vdash_{\mathrm{LTL}} \neg\varphi$, and there is no $\psi \in \mathit{Nodes}(\pi)$ such that $\psi \rightsquigarrow_\pi \varphi$ and $\mathcal{B} \vdash_{\mathrm{LTL}} \psi$. That is, π $\mathcal{B}$-fails if it has a node that is necessary (no ancestor of it is already believed to be true) but unachievable.

The aim of $\mathit{cleanup}(\mathcal{B}_1, \mathcal{A}_0, \mathcal{D}, \mathcal{P})$ is to produce a set of plans $\mathcal{T}$ that can be later expanded to a new active plan set $\mathcal{A}_1$. First of all, for every $\pi \in \mathcal{A}_0$, the algorithm checks what the status of every node φ in π is. Then, the resulting plan π is passed on to $\mathcal{T}$ if and only if (1) the root of π is not believed to be true yet, and (2) every node in π that is believed to be unachievable is only a means for some ψ that is already believed to be fulfilled.

Note that other update strategies may also make sense. For instance, the algorithm from Fig. 4 ignores every plan π that fails given the new beliefs $\mathcal{B}_1$. It might be also interesting to check whether, it is still possible to add alternative plans π' with $\mathit{root}(\pi') = \mathit{root}(\pi)$ to the final set of plans $\mathcal{T}$. This would ensure that the agent takes an effort to re-try desires for which a previous plan has been discovered to fail.

To compute the new pair $\langle \mathcal{I}_1, \mathcal{A}_1 \rangle$, we apply the procedure *activePlans* from Sect. 3.4 to $\mathcal{T}$:

$$\mathcal{A}_1 = \mathit{activePlans}(\mathit{cleanup}(\mathcal{B}_1, \mathcal{A}_0, \mathcal{D}, \mathcal{P}), \mathcal{B}_1, \mathcal{D}, \mathcal{P}), \text{ and}$$
$$\mathcal{I}_1 = \mathit{Int}(\mathcal{A}_1).$$

That is, given the new beliefs $\mathcal{B}_1$, the old active plans $\mathcal{A}_0$ and the constant sets $\mathcal{D}$ of desires and $\mathcal{P}$ of reasoning rules, to compute the new set $\mathcal{A}_1$ of active plans, we first remove all succeeded and failed plans from $\mathcal{A}_0$, and then fill the result until it is maximal in the sense that every possible desire is there (cf. Definition 2).

4.2 Updating mental states

A rational agent's mental state will not change spontaneously, except, perhaps, to note the passage of time and its consequences. Typically, an agent's state changes because he has made *observations* of his environment; these observations cause changes in beliefs as described by the Alchourrón–Gärdenfors–Makinson theory of belief revision, or the Katsuno–Mendelzon approach to updates. We will describe the effect of incorporating the observation ω in the belief set $\mathcal{B}$ by $\mathcal{B} \star \omega$. The difference between *revising* a belief set and *updating* it (see Katsuno & Mendelzon, 1991), is that the information used for revision is seen as a new information about a *static* world, thereby adjusting the accuracy of the old beliefs. Information used for update, however, is interpreted as new information about a *dynamic* world: if the

old belief set does is not consistent with the new information, it means the old belief set was out of date.[10]

We are interested in how a mental state $\langle \mathcal{B}_0, \mathcal{D}, \langle \mathcal{I}_0, \mathcal{A}_0 \rangle, \mathcal{P} \rangle$ transforms into another mental state $\langle \mathcal{B}_1, \mathcal{D}, \langle \mathcal{I}_1, \mathcal{A}_1 \rangle, \mathcal{P} \rangle$ as a result of some observations. To investigate this, we have to be precise about what an *observation* is. Here, we assume that observations either regard the current state, or the future: formally, an *observation* ω is either a propositional formula α or a temporal formula φ. We also allow for updating the agent's beliefs with a *finite set* of observations Ω. To indicate that mental state $\Gamma_1 = \langle \mathcal{B}_1, \mathcal{D}, \langle \mathcal{I}_1, \mathcal{A}_1 \rangle, \mathcal{P} \rangle$ can result from a rational update of mental state $\Gamma_0 = \langle \mathcal{B}_0, \mathcal{D}, \langle \mathcal{I}_0, \mathcal{A}_0 \rangle, \mathcal{P} \rangle$ with observations Ω, we write

$$\Gamma_0 \stackrel{\Omega}{\rightarrow} \Gamma_1.$$

As for the beliefs, we assume to have a belief change function $\star$ with some basic properties discussed in Sect. 4.3, and we define $\mathcal{B}_1 = \mathcal{B}_0 \star \Omega = \mathcal{B}_0 \star \bigwedge_{\omega \in \Omega} \omega$. To obtain the new pair of intention and active plan sets $\langle \mathcal{I}_1, \mathcal{A}_1 \rangle$, we use the ideas outlined in Sect. 4.1. Formally, we define relation $\langle \mathcal{B}_0, \mathcal{D}, \langle \mathcal{I}_0, \mathcal{A}_0 \rangle, \mathcal{P} \rangle \stackrel{\Omega}{\rightarrow} \langle \mathcal{B}_1, \mathcal{D}, \langle \mathcal{I}_1, \mathcal{A}_1 \rangle, \mathcal{P} \rangle$, between two mental states, as follows:

1. $\mathcal{B}_1 = \mathcal{B}_0 \star \bigwedge_{\omega \in \Omega} \omega$, and $\mathcal{B}_1 \neq \mathcal{B}_\perp$.
2. $\mathcal{A}_1 = \mathit{activePlans}(\mathit{cleanup}(\mathcal{B}_1, \mathcal{A}_0, \mathcal{D}, \mathcal{P}), \mathcal{B}_1, \mathcal{D}, \mathcal{P})$.
3. $\mathcal{I}_1 = \mathit{Int}(\mathcal{A}_1)$.

Note that $\Gamma_0 \stackrel{\Omega}{\rightarrow} \Gamma_1$ only defines a *partial* function: if $\mathcal{B}_0 \star \bigwedge_{\omega \in \Omega} \omega$ is inconsistent, there is no result. Otherwise, the resulting state Γ_1 is enforced by the new beliefs.

Now, we want a language that will allow us to explicitly capture this dynamics. To this end, we use some notation from dynamic logic (Harel, Kozen, & Tiuryn, 2000), and introduce a *dynamic update operator* $[\Omega]\varphi$, meaning "after the agent has updated on the basis of observations Ω, it must be the case that φ". We denote a typical member of this dynamic language by δ:

$$\delta := \gamma \mid \neg\delta \mid \delta \vee \delta \mid [\Omega]\gamma,$$

where γ is a mental state formula. We write $[\omega]\gamma$ as a shorthand for $[\{\omega\}]\gamma$.

We define the truth conditions for the language with respect to a mental state Γ: the rules for interpreting formulae γ of the mental state language remain unchanged, and since the rules for negation and disjunction are entirely standard, we only need give the rule for the dynamic observation-update operator:

$$\Gamma \models [\Omega]\gamma \;\text{ iff }\; \text{for every } \Gamma' \text{ such that } \Gamma \stackrel{\Omega}{\rightarrow} \Gamma' \text{ we have } \Gamma' \models \gamma.$$

Thus, the dynamic observation-update operator is a "normal modal box", which behaves very much like the box operator of propositional dynamic logic PDL (Harel et al., 2000). Following the modal tradition, we will also write $\langle \Omega \rangle \gamma$ for $\neg[\Omega]\neg\gamma$, which says that there is some mental state Γ_1 reachable by observing Ω in Γ_0, and such that γ holds in Γ_1.

[10] One referee brought to our attention that it would be nice to see how the difference between revising a belief and updating it would affect our intention revision. If beliefs are about the future, the difference between updating and revising them raises interesting philosophical questions again. For instance, one might take the point of view that the future is determined, so that we can only revise our strong beliefs about it. However, we do not want to make any claims about belief change in the context of temporal logic in the paper, and in fact, the properties that we assume in Sect. 4.3 for our belief change operator $\star$, are properties both commonly adopted for belief revision and belief update (cf. Katsuno & Mendelzon, 1991).

4.3 Properties of the dynamic update

The fact that $\star$ is an AGM-style belief update operator suggests the following (semantic) postulates on the properties of belief change, together with their characterisations in our language of dynamic update:

1. Given LTL-consistent observations Ω, $\mathcal{B}$ can be updated with Ω in the way that $\mathcal{B} \star \Omega \neq \mathcal{B}_\perp$:

$$\models \neg[\Omega]\perp \quad \text{for all consistent } \Omega.$$

2. (*Success*) For every Ω we have $\Omega \subseteq \mathcal{B} \star \Omega$:

$$\models [\Omega] \bigwedge_{\omega \in \Omega} \mathrm{SB}\, \omega.$$

3. (*Expansion*) What is consistent with the beliefs, can be added to them, i.e. if Ω is consistent with $\mathcal{B}$, then $\mathcal{B} \star \Omega = \mathcal{B} \cup \{\Omega\}$:

$$\models (\neg \mathrm{SB}\, \neg\Omega \wedge \mathrm{SB}\, \varphi) \rightarrow [\Omega](\mathrm{SB}\, \varphi \wedge \mathrm{SB}\, \Omega).$$

The postulates yield several immediate corollaries:

Theorem 6

1. *Inconsistency of updates results exactly from the inconsistency of observations. That is, $\mathcal{B} \star \Omega = \mathcal{B}_\perp$ iff $\vdash_{\mathrm{LTL}} \neg\Omega$:*

$$\models \neg \bigwedge_{\omega \in \Omega} \omega \leftrightarrow [\Omega]\perp.$$

2. (No surprise) *Updating with what was already believed does not change the beliefs, i.e. if $\Omega \subseteq \mathcal{B}$ then $\mathcal{B} \star \Omega = \mathcal{B}$:*

$$\models \mathrm{SB}\, \Omega \rightarrow (\mathrm{SB}\, \psi \leftrightarrow [\Omega]\mathrm{SB}\, \psi).$$

3. (No un-necessary removal) *Updating with Ω, that is consistent with the current beliefs does not remove any belief, i.e. $\mathcal{B} \star \Omega \supseteq \mathcal{B}$:*
4. *Every successful update yields a belief set:*

$$\models [\Omega]\neg\mathrm{SB}\perp.$$

Proof Property 1: the equivalence follows one way from postulate 1, and the other way from the *Success* postulate. Properties 2 and 3 follow directly from the *Expansion* postulate. Property 4 follows from the semantics of $[\Omega]\gamma$. □

We can now show how strong and weak beliefs are linked via observations: the next theorem expresses that believing on the assumption that all your intentions will be successful (i.e. weakly believing) is equivalent to believing what is inevitable, *given that you have accepted that all what you intend will be brought about*.

Theorem 7 *For every $\Gamma = \langle \mathcal{B}, \mathcal{D}, \langle \mathcal{I}, \mathcal{A} \rangle, \mathcal{P} \rangle$, we have the following characterisation of optimistic beliefs:*

$$\Gamma \models \mathrm{WB}\, \omega \leftrightarrow [\mathcal{I}]\mathrm{SB}\, \omega.$$

Proof Suppose $\Gamma_0 = \langle \mathcal{B}_0, \mathcal{D}, \langle \mathcal{I}_0, \mathcal{A}_0 \rangle, \mathcal{P} \rangle \models \text{WB}\ \omega$, which means that $\mathcal{B}_0 \cup \mathcal{I}_0 \vdash_{\text{LTL}} \omega$. We know that $\mathcal{B}_0 \cup \mathcal{I}_0$ is consistent (item 1, Definition 2). Hence $\mathcal{I}_0$ is consistent with $\mathcal{B}_0$. By *Expansion*, we have $\mathcal{B}_1 = \mathcal{B}_0 \star \mathcal{I}_0 = \mathcal{B}_0 \cup \mathcal{I}_0$, and hence, $\mathcal{B}_1 \vdash_{\text{LTL}} \omega$, i.e. $\Gamma_0 \models [\mathcal{I}_0]\text{SB}\omega$. For the other direction, suppose $\Gamma_0 \models [\mathcal{I}_0]\text{SB}\omega$. Since $\mathcal{I}_0$ is consistent with $\mathcal{B}_0$, this means (again, use *Expansion*) that $\mathcal{B}_0 \cup \mathcal{I}_0 \vdash_{\text{LTL}} \omega$. But this is exactly the definition of being a weak belief, i.e. $\Gamma_0 \models \text{WB}\ \omega$. □

More relevant for the purposes of this paper are the following properties, which relate the update of intentions to observations.

Theorem 8

1. *An agent drops intentions that he believes to have been accomplished:*

$$\models [\Omega]\text{SB}\varphi \rightarrow [\Omega]\neg I\varphi.$$

2. *Dropping accomplished intentions is successful:*

$$\models I \bigwedge_{\omega \in \Omega} \omega \rightarrow \neg[\Omega]\bot.$$

3. *An agent drops intentions that he observes impossible to achieve:*

$$\models [\Box\neg\alpha]\neg(I\Box\alpha \vee I\Diamond\alpha \vee I \bigcirc \alpha) \wedge [\bigcirc\neg\alpha]\neg(I\Box\alpha \vee I \bigcirc \alpha) \wedge [\Diamond\neg\alpha]\neg I\Box\alpha.$$

4. *More generally: an agent drops intentions that he believes impossible to achieve:*

$$\models [\Omega]\text{SB}\ \neg\varphi \rightarrow [\Omega]\neg I\varphi.$$

Proof

1. Suppose $\Gamma_0 \models [\Omega]\text{SB}\varphi$. This means that in every Γ_1 that results as an update with Ω in Γ_0, we have $\mathcal{B}_1 \vdash_{\text{LTL}} \varphi$. But then, by Theorem 4.2, we obtain $\Gamma_1 \models \neg I\varphi$.
2. By Definitions 2 and 3, $\mathcal{I}$ must be LTL-consistent, so Ω is also consistent. Thus, by Theorem 6.6, we get $\neg[\Omega]\bot$.
3. We do the reasoning for the conjunct $[\Diamond\neg\alpha]\neg I\Box\alpha$; the other cases are analogous. If the observation $\Diamond\neg\alpha$ is inconsistent, the result is trivial, since then we have $\models [\Diamond\neg\alpha]\psi$ for any ψ. Assume that $\Diamond\neg\alpha$ is consistent. By *Success*, we have $\Gamma \models [\Diamond\neg\alpha]\text{SB}\ \Diamond\neg\alpha$. But we also have $\vdash_{\text{LTL}} \Diamond\neg\alpha \rightarrow \neg\Box\alpha$. Since beliefs are closed wrt logical consequence, $\Gamma \models [\Diamond\neg\alpha]\text{SB}\ \neg\Box\alpha$ too. Now, if $\Gamma \models [\Diamond\neg\alpha]\neg I\Box\alpha$, then the intentions would be inconsistent with the beliefs, which cannot be the case.
4. From Theorem 4, item 2, we have $\models \text{SB}\ \neg\varphi \rightarrow \neg I\varphi$. Since this is a validity, it will hold in any Γ_1, so we have $\models [\Omega](\text{SB}\ \neg\varphi \rightarrow \neg I\varphi)$. Since $[\Omega]$ is a normal modal operator, we get $\models [\Omega]\text{SB}\ \neg\varphi \rightarrow [\Omega]\neg I\varphi$. □

The literature on belief revision has always emphasised the principle of *minimal change*: the new information should disturb the mental state as "little as possible". In this context, we have the following:

Theorem 9 *An agent does not drop an intended desire if he does not believe it has been established or become impossible:*

$$\models (D\varphi \wedge I\varphi \wedge \neg[\Omega]\text{SB}\varphi \wedge \neg[\Omega]\text{SB}\ \neg\varphi) \rightarrow [\Omega]I\varphi.$$

Proof Assume $\Gamma_0 \models D\varphi \wedge I\varphi \wedge \neg[\Omega]\mathrm{SB}\,\varphi \wedge \neg[\Omega]\mathrm{SB}\,\neg\varphi$. Since φ is intended, there is are no strong beliefs wrt φ: we have $\lambda(\varphi, \pi, \mathcal{B}_0) = +$ for some plan π. Let us take the sub-plan π' of π with $root(\pi') = \varphi$. As φ is a desire, and π' satisfies items 1 and 2 of Definition 2, then by item 3 of the same definition, π' must be also in $\mathcal{A}_0$. Now, $\Gamma_0 \models \neg[\Omega]\mathrm{SB}\varphi \wedge \neg[\Omega]\mathrm{SB}\,\neg\varphi$ guarantees that also in Γ_1, φ is still not believed, nor is its negation. That means that also we have $\lambda(\varphi, \pi', \mathcal{B}_1) = +$. So, we have $\Gamma_1 \models I\varphi$, and hence $\Gamma_0 \models [\delta]I\varphi$. □

Theorem 9 does not hold for intentions in general, i.e. if we are not guaranteed that φ is a desire: an agent that uses plan π_2 in Fig. 2 to become energetic tomorrow, might come to believe $\bigcirc$*nicecoffee*, and hence give up the intention $\bigcirc$*espresso*—not because $\bigcirc$*espresso* is now believed to be guaranteed or impossible, but because it is not needed for the intention $\bigcirc$*nicecoffee* anymore.

Now we can finally come back to the commitment strategies from Rao & Georgeff (1991). For instance, a *blindly committed agent* is defined there as one who only gives up an intention φ if he believes that φ is achieved. In our framework, we can model blindly committed agents by demanding the validity of the scheme:

$$(I\varphi \wedge [\omega]\neg I\varphi) \rightarrow [\omega]\mathrm{SB}\,\varphi.$$

This property says that if an observation ω makes the agent abandon an intention, it must be because he believes after the observation that the intention is already guaranteed. Note that this is not saying that our framework only models blindly committed agents. On the contrary, one can easily model agents who do not satisfy the above requirement.

Let us briefly reflect on the properties ascribed to intentions by Bratman, Cohen and Levesque (see p. 3). Indeed, our agents need to determine ways of achieving intentions, using practical reasoning rules but also taking into account what they already believe and what the other current intentions are (see our Definition 2). Intentions provide a "filter" for adopting other intentions, see item 1 of Definition 2. Our agents track the success of their intentions: see the algorithm *cleanup* of Fig. 4. Our agents believe their intentions are possible: according to Theorem 4, an intention of φ implies both that $\mathrm{WB}\,\varphi$ and that $\neg\mathrm{SB}\,\neg\varphi$, which, in terms of modal doxastic logic says that the agent considers φ possible but not guaranteed. Agents do not believe they will not bring about their intentions both in the weak and the strong sense: $I\varphi \rightarrow \neg\mathrm{WB}\,\neg\varphi$ and $I\varphi \rightarrow \neg\mathrm{SB}\neg\varphi$; both properties follow from Theorem 4. Moreover, agents believe they will bring about their intentions in the sense of Theorem 4.1. Finally, we have already seen that our agents do not intend all the expected side effects of their intentions: see Theorem 3.5.

5 Conclusions

In this paper, we have started to investigate the difficult issue of intention revision. Intention revision is hard to handle theoretically for several reasons: first, because there is no clear agreement on what exactly an intention is, and while there are several competing models of intentions, none is without drawbacks; and second, because intentions cannot be studied in isolation—they must be considered in the wider context of an agent's mental state. Our strategy in this paper has been to build up a model of intentions and intention revision operators, then to construct a logic to talk about these structures, and finally, to characterise properties of intentions and intention update as formulae of this logic. We hope that we have convincingly argued that an agent needs to have beliefs both about what would happen if he abstracts from his intentions, and beliefs that take into account his intentions.

The literature to which this work can be related or compared is vast. Obviously, our models of agents fit well in the stream of work triggered by Rao and Georgeff's BDI logic (Rao & Georgeff, 1991). Also the KARO framework deals with informational (in the case of KARO: knowledge-related) and motivational (desires- and goals-related) attitudes, and—rather than a temporal dimension—it facilitates explicit reasoning about actions (van der Hoek, van Linder & Meyer, 1999). There is some work on the dynamics of knowledge in KARO (van Linder, van der Hoek, & Meyer, 1995) and it might be feasible to extend it to desires, goals and/or intentions.[11] Extending practical reasoning rules to handle actions would probably yield a clean and expressive framework to reason about both intentions *to be* (as in this paper), and intentions *to do*. Moreover, our notion of mental states and formulas is somewhat similar to that of 3APL (Hindriks, de Boer, van der Hoek, & Meyer, 1999). In 3APL, agents can adopt and drop goals (which are similar to our notion of intentions), and the overall requirement complies with our second item of Theorem 8: an agent can never have a goal, that is already believed to be satisfied.

Since the focus in this paper is on intention revision, we assumed a mechanism for belief revision that satisfies the well-known postulates of AGM put forward in Alchourrón (1985). Currently, there is a lot of interest in generalising AGM to Dynamic Epistemic Logic DEL (Baltag & Moss, 2004; van Ditmarsch, van der Hoek, & Kooi, 2003), with a multi-modal language enabling to analyse the effect of specific *epistemic actions*—like public announcements or secret communication—that directly change the knowledge of the agent. The intention-oriented counterpart for such actions can be seen in *motivational actions* like getting inspiration, receiving a command, or encouragement. But, as we indicated in the introduction, the notions of beliefs and intentions are tightly connected, and even a simple observation about the future, or the repeated failure of an undertaken plan, might lead the agent to change his intentions—either directly, or as a side effect of the belief change.

We do not, of course, claim that this work represents the final statement on intentions and intention revision. In particular, a future model would consider the status and semantics of practical reasoning rules in more detail, perhaps drawing on work on the semantics of agent programming languages (Rao, 1996) and (abductive) logic programming. Also, our treatment of intentions and desires is rather syntactical. This is mainly motivated by the urge to avoide the side-effect problem mentioned earlier, but we agree that it would be interesting to interpret desires with some preference relation over states to give an account that sometimes people have preferences for which they cannot come up with a syntactical description (cf. Kaplan & Kaplan, 1989). Moreover, the analysis in this paper is restricted to the single-agent case: we do not attempt to model the multi-agent setting. In such a setting, one would expect the agents to perform communicative actions about their intentions and beliefs, and maybe even practical reasoning rules (perhaps the framework DEL might give some hints how the effect of such actions can be described). Another extremely interesting question that arises in the multi-agent setting is the issue of *group intentions* (cf. Cohen & Levesque, 1991; Jennings, 1993; Wooldridge & Jennings, 1999), and the way their dynamics should relate to the intention change performed by individual agents. Nevertheless, we believe that the model presented here is a useful first step towards understanding these difficult and important processes.

Acknowledgements We thank Fabio Paglieri for his support and the three anonymous referees for their helpful comments.

[11] We note also that *commitments* provide one way for an agent to persist in his intentions; the dynamics of commitments in KARO has been studied in Meyer, van der Hoek, and van Linder (1999).

References

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

Allen, J. F., Hendler, J., & Tate, A. (Eds.). (1990). *Readings in planning*. San Mateo, CA: Morgan Kaufmann Publishers.

Aristotle (1998). *The nicomachean ethics*. Oxford, UK: Oxford University Press. Translated by D. Ross, J. R. Ackrill and O. Urmson.

Atkinson, K. (2005). What should we do? Ph.D Thesis, Computer Science Department, the University of Liverpool. www.csc.liv.ac.uk/∼katie/ThesisFinal.pdf, retrieved 19 July 2006.

Baltag, A., & Moss, L. S. (2004). Logics for epistemic programs. *Synthese, 139*, 165–224. Knowledge, Rationality & Action 1–60.

Bratman, M. E. (1987). *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Bratman, M. E. (1990). What is intention? In P. R. Cohen, J. L. Morgan, & M. E. Pollack (Eds.), *Intentions in communication* (pp. 15–32). Cambridge, MA: The MIT Press.

Bratman, M. E., Israel, D. J., & Pollack, M. E. (1988). Plans and resource-bounded practical reasoning. *Computational Intelligence, 4*, 349–355.

Cohen, P. R., & Levesque, H. J. (1990). Intention is choice with commitment. *Artificial Intelligence, 42*, 213–261.

Cohen, P. R., Levesque, H. J. (1990). Rational interaction as the basis for communication. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in communication* (pp. 221–256). Cambridge, MA: The MIT Press.

Cohen, P. R., & Levesque, H. J. (1991). Teamwork. *Nous, 25*(4), 487–512.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: The MIT Press.

Dignum, F., Dunin-Keplicz, B., & Verbrugge, R. (2001). Agent theory for team formation by dialogue. In C. Castelfranchi, & Y. Lesperance (Eds.), *Proceedings agent theories, architectures and languages ATAL*, number 1986 in LNAI (pp. 150–166). Berlin: Springer.

Dignum. F., & Conte, R. (1998). Intentional agents and goal formation. In M. P. Singh, A. Rao, & M. J. Wooldridge (Eds.), *Intelligent agents IV (LNAI Vol. 1365)* (pp. 231–244). Berlin, Heidelberg, New York, Germany: Springer-Verlag.

Emerson, E. A. (1990). Temporal and modal logic. In J. van Leeuwen (Ed.), *Handbook of theoretical computer science volume B: Formal models and semantics* (pp. 996–1072). Amsterdam, The Netherlands: Elsevier Science Publishers B.V.

Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. Y. (1995). *Reasoning about knowledge*. Cambridge, MA: The MIT Press.

Galliers, J. R. (1988). A strategic framework for multi-agent cooperative dialogue. In *Proceedings of the eighth european conference on artificial intelligence (ECAI-88)*, Munich, Federal Republic of Germany, pp. 415–420.

Galliers, J. R. (1988). *A theoretical framework for computer models of cooperative dialogue, acknowledging multi-agent conflict*. Ph.D thesis, Open University, UK.

Georgeff, M. P., & Lansky, A. L. (Eds.), (1986). *Reasoning about actions & plans—proceedings of the 1986 workshop*. San Mateo, CA: Morgan Kaufmann Publishers.

Georgeff, M. P., & Lansky, A. L. (1987). Reactive reasoning and planning. In *Proceedings of the sixth national conference on artificial intelligence (AAAI-87)*, Seattle, WA, pp. 677–682.

Georgeff, M. P., & Rao, S. (1995). The semantics of intention maintenance for rational agents. In *Proceedings IJCAI* (pp. 704–710). San Mateo, CA: Morgan Kaufmann.

Harel, D., Kozen, D., & Tiuryn, J. (2000). *Dynamic logic*. Cambridge, MA: The MIT Press.

Hindriks, K. V., de Boer, F. S., van der Hoek, W., & Meyer, J.-J. Ch. (1999). Agent programming in 3APL. *Autonomous Agents and Multi-Agent Systems, 2*(4), 357–402.

Jennings, N. R. (1992). On being responsible. In E. Werner & Y. Demazeau (Ed.), *Decentralized AI 3—proceedings of the third european workshop on modelling autonomous agents in a multi-agent world (MAAMAW-91)* (pp. 93–102). Amsterdam, The Netherlands: Elsevier Science Publishers B.V.

Jennings, N. R. (1992). Towards a cooperation knowledge level for collaborative problem solving. In *Proceedings of the tenth european conference on artificial intelligence (ECAI-92)*, Vienna, Austria, pp. 224–228.

Jennings, N. R. (1993). Commitments and conventions: The foundation of coordination in multi-agent systems. *The Knowledge Engineering Review, 8*(3), 223–250.

Jennings, N. R. (1993). Specification and implementation of a belief desire joint-intention architecture for collaborative problem solving. *Journal of Intelligent and Cooperative Information Systems, 2*(3), 289–318.

Kaelbling, L. P., & Rosenschein, S. J. (1990). Action and planning in embedded agents. In P. Maes (Eds.), *Designing autonomous agents* (pp. 35–48). Cambridge, MA: The MIT Press.

Kaplan, R., & Kaplan, S. (1989). *The experience of nature: A psychological perspective*. Cambridge: Cambridge University Press.

Katsuno, H., & Mendelzon, A. (1991). On the difference between updating a knowledge base and revising it. In J. A. Allen, R. Fikes, & E. Sandewall (Eds.), *Principles of knowledge representation and reasoning* (pp. 387–394). San Mateo, CA: Morgan Kaufmann.

Kinny, D., & Georgeff, M. (1991). Commitment and effectiveness of situated agents. In *Proceedings of the twelfth international joint conference on artificial intelligence (IJCAI-91)*, Sydney, Australia, pp. 82–88.

Levesque, H. J., Cohen, P. R., & Nunes, J. H. T. (1990). On acting together. In *Proceedings of the eighth national conference on artificial intelligence (AAAI-90)*, Boston, MA, pp. 94–99.

McDermott, D. (1982). A temporal logic for reasoning about processes and plans. *Cognitive Science, 6*, 101–155.

Meyer, J.-J. Ch., van der Hoek, W., & van Linder, B. (1999). A logical approach to the dynamics of commitments. *Artificial Intelligence, 113*, 1–40.

Meyer, J.-J. Ch., & van der Hoek, W. (1995). *Epistemic logic for ai and computer science*. Cambridge, England: Cambridge University Press.

Pollack, M. E. (1992). The uses of plans. *Artificial Intelligence, 57*(1), 43–68.

Rao, A. S. (1996). AgentSpeak(L): BDI agents speak out in a logical computable language. In W. Van de Velde & J. W. Perram (Eds.), *Agents breaking away: Proceedings of the seventh european workshop on modelling autonomous agents in a multi-agent world, (LNAI Volume 1038)* (pp. 42–55). Berlin, Heidelberg, New York, Germany: Springer-Verlag.

Rao, A. S. & Georgeff, M. (1998). Decision procedures for BDI logics. *Journal of Logic and Computation, 8*(3), 293–344.

Rao, A. S., & Georgeff, M. P. (1991). Modeling rational agents within a BDI-architecture. In R. Fikes, & E. Sandewall (Eds.), *Proceedings of knowledge representation and reasoning (KR&R-91)* (pp. 473–484). San Mateo, CA: Morgan Kaufmann Publishers.

Rao, A. S., & Georgeff, M. P. (1992). An abstract architecture for rational agents. In C. Rich, W. Swartout, & B. Nebel (Eds.), *Proceedings of knowledge representation and reasoning (KR&R-92)* (pp. 439–449). San Mateo, CA: Morgan Kaufmann.

Searle, J. R. (2001). *Rationality in action*. Cambridge, MA, USA: MIT Press.

Singh, M. P. (1992). A critical examination of the Cohen–Levesque theory of intention. In *Proceedings of the tenth european conference on artificial intelligence (ECAI-92)*, Vienna, Austria, pp. 364–368.

Skubch, H., & Thielscher, M. (2005). Strategy learning for reasoning agents. In J. Gama et al. (Eds.), *ECML 2005*, number 3720 in LNAI (pp. 733–740). Berlin, Heidelberg, New York: Springer.

van Benthem, J., & Liu, F. (2005). Dynamic logic of preference upgrade. Technical Report PP-2005–2029, University of Amsterdam, 2005. To appear in *Journal of Applied Non-Classical Logic, 17*(2), 2007

van der Hoek, W., van Linder, B., & Meyer, J.-J. Ch. (1999). An integrated modal approach to rational agents. In M. Wooldridge, & A. Rao (Eds.), *Foundations of rational agency* (pp. 133–168). Dordrecht, The Netherlands: Kluwer Academic Publishers.

van Ditmarsch, H. P., van der Hoek, W., & Kooi, B. (2003). Concurrent dynamic epistemic logic. In V. F. Hendricks, K. F. Jørgensen, & S. A. Pedersen (Eds.), *Knowledge contributors*, *Synthese Library* (Vol. 322, pp. 45–82). Dordrecht, The Netherlands: Kluwer.

van Linder, B., van der Hoek, W., & Meyer, J.-J. Ch. (1995). Actions that make you change your mind. In A. Laux, & H. Wansing (Eds.), *Knowledge and belief in philosophy and artificial intelligence* (pp. 103–146). Berlin: Akademie Verlag.

Vidal, J., & Durfee, E. (1998). Learning nested agent models in an information economy. *Journal of Experimental and Theoretical Artificial Intelligence, 10*, 291–308.

Wooldridge, M. (2000). *Reasoning about rational agents*. Cambridge, MA: The MIT Press.

Wooldridge, M., & Jennings, N. R. (1999). The cooperative problem solving process. *Journal of Logic and Computation, 9*(4), 563–592.

Wooldridge, M., & Parsons, S. D. (1999). Intention reconsideration reconsidered. In J. P. Müller, M. P. Singh & A. S. Rao (Eds.), *Intelligent agents V (LNAI Volume 1555)* (pp. 63–80). Berlin, Heidelberg, New York, Germany: Springer.