Joint revision of belief and intention

Thomas Icard
Department of Philosophy
Stanford University

Eric Pacuit
Center for Logic and Philosophy of Science
Tilburg University

Yoav Shoham
Department of Computer Science
Stanford University

Abstract

We present a formal semantical model to capture action, belief and intention, based on the "database perspective" (Shoham 2009). We then provide postulates for belief and intention revision, and state a representation theorem relating our postulates to the formal model. Our belief postulates are in the spirit of the AGM theory; the intention postulates stand in rough correspondence with the belief postulates.

Introduction and Motivation

While there is an extensive literature developing logical models to reason about changing *informational* attitudes (eg., belief, knowledge, certainty), other mental states have received less attention[1]. However, this is changing with recent articles introducing dynamic logics of intention (van der Hoek, Jamroga, and Wooldridge 2007; Herzig and Lorini 2008)[2]. These papers take as a starting point logical frameworks derived from Cohen and Levesque's seminal paper (Cohen and Levesque 1990) aimed at formalizing Bratman's planning theory of intention (Bratman 1987). In this paper we take a different angle on intentions, focusing on intention revision as it relates to, and is intertwined with, belief revision.

We view the problem of intention revision as a database management problem (see (Shoham 2009) for more on the conceptual underpinnings of this standpoint). At any given moment, an agent must keep track of a number of facts about the current situation. This includes beliefs about the current state, beliefs about possible future states, which actions are available now and in the future, and also what the agent plans to do at future moments. It is important that all of this information be *jointly consistent* at any given moment and furthermore that it can be *modified* as needed while maintaining consistency.

In the following we introduce a simple logic that formally models such a "database". That is, *consistency* in this logic is meant to represent not only that the agent's beliefs are consistent and the agent's future plan is consistent, but also that the agent's beliefs and intentions together form a *coherent* picture of what may happen, and of how the agent's own actions will play a role in what happens. Many of the BDI-style logics emanating from (Cohen and Levesque 1990) can be viewed as addressing this issue (Rao and Georgeff 1992; Meyer, van der Hoek, and van Linder 1999, are two examples). Our primary contribution in this article (in line with the recent articles on dynamic BDI logics mentioned above) is to focus also on how the database is to be modified, and in the process to provide a clear picture of how intentions and beliefs relate.

What can cause an agent's database to change? In this paper, we focus on two main sources:

1. The agent makes some observation, e.g. from sensory input. If the new observation is inconsistent with the agent's beliefs, these beliefs will have to be revised to accommodate it. While we recognize the classical AGM theory (Alchourrón, Gärdenfors, and Makinson 1985) is not without problems, in particular when it comes to iterated revision,[3] our account of belief revision simply adopts this framework. The goal is thus to give general conditions on a single revision with new information *that the agent has already committed to incorporating*.
2. The agent forms a new intention. Here we focus on *future directed* intentions, understood as time-labelled actions that might make up a plan. Analogously to belief revision, it is assumed the agent has already committed to a new intention, so it must be accommodated by any means short of revising beliefs. The force of the theory is in restricting how this can be accomplished. To be more precise, we purport to model an intelligent database, which receives instructions from some planner (e.g. a STRIPS-like planner) that is itself engaged in some form of practical reasoning. The job of the database is to maintain consistency and coherence between intentions and beliefs.

This simple description, however, obscures some important subtleties in the interaction between beliefs and intentions, subtleties we would also like to capture.

The following will serve as a running example. Suppose an agent intends to drive to the city at 6:00 this evening.

[1] A notable exception is work on logics of preferences and preference change. See (van Benthem 2009) for a survey of recent work.

[2] See also a recent discussion of "goal dynamics" in (Castelfranchi and Paglieri 2007).

[3] Though, see (Darwiche and Pearl 1997; Boutilier 1996) for postulates concerning iterated revision.

Upon adopting this intention, the agent will come to have new beliefs based on the predicted success of this intention, e.g. that he will be in the city by 7:00. These further beliefs are important in the course of further planning, for instance, what he will do in the city. The intention is also supported by the absence of certain beliefs. It would be irrational to form this intention if the agent believed his car was not working and this was the only means of getting there. Likewise, even if originally the agent thought his car might be working, upon learning that it is not and lacking other ideas of how to get there, the intention to go to the city should be dropped. Yet, by dropping this intention that was based on the now-dropped belief, other beliefs, including the belief that he will be in the city by 7:00, should also be dropped, which may in turn force other intentions and beliefs to be dropped. And so on.

To deal with these subtleties, we treat intention-contingent beliefs, or simply *contingent beliefs*, and concrete "physical" beliefs, or *non-contingent beliefs*, separately. Non-contingent beliefs concern the world as it is, independent of the agent's future plans, but including what (sequences of) actions will be physically possible. Thus, in addition to non-contingent atomic facts, the agent will have beliefs about what the preconditions and postconditions of actions are, and about which sequences of actions might be possible. Our treatment of contingent beliefs is similar to the notion of *weak belief* in (van der Hoek, Jamroga, and Wooldridge 2007), but differs in an important respect. We assume that the postconditions of intended actions are believed in this stronger sense, but that the preconditions need not be believed. The intuition behind this decision is that, from the perspective of a planner, the postconditions of intended actions are justifiably believed *merely by the fact that the agent has committed to bringing them about*. In this way, these beliefs are *contingent* on the success of the agent's plans. The preconditions, on the other hand, may still present a practical problem yet to be solved by the planner. To say that they are believed underrates the fact that they are not *directly justified* by any intended future action. Hence, contingent beliefs are simply derived from the agent's non-contingent beliefs by adding the postconditions (and all consequences) of any intended actions. These kinds of beliefs might also be called "optimistic" beliefs, since the agent assumes the success of the action without ensuring the preconditions hold.

In this way, our account avoids the potentially infinite regress alluded to above by allowing belief revision to trigger intention revision, but restricting intention revision to trigger belief revision only in this stronger, derivative sense of contingent belief. Our postulates will reflect this fact.

In the next section, we describe the belief and intention revision postulates on an informal level before going into formal details and definitions. We then define the logic underlying the database, as a simple temporal logic with transitions labeled by actions. The models of this logic are then used to give a semantic characterization of our revision operations, which are shown in the next section to represent our main postulates. Finally in the last section before the conclusion, we define a notion of contingent beliefs, as described above, and provide postulates for revision of these beliefs, as derived from the separate postulates for beliefs and intentions.

A Preview of the Postulates

The main aim of our framework is to implement the "database perspective" on intentions in the form of a dual logical theory of belief and intention revision. In this section, before going into the formalities of our framework, we offer a preview of the revision postulates that we will be working toward for the rest of the paper. Relevant definitions of key terms like *coherence* will come later.

If B is a set of non-contingent beliefs (i.e. a set of formulas, which by definition does not depend on intentions) and I is a set of intentions (which shall be action/time pairs (a, t), including an empty pair ϵ), we shall define a class of intention revision operators $\circ$ that adhere to the following restrictions when $\langle B, I\rangle \circ (a, t) = \langle B', I'\rangle$ for some proposed new intention (a, t).

1. $\langle B', I'\rangle$ is coherent;
2. If $\langle B, \{(a, t)\}\rangle$ is coherent, then $(a, t) \in I'$;
3. If $\langle B, I \cup \{(a, t)\}\rangle$ is coherent, then $I \cup \{(a, t)\} \subseteq I'$;
4. $I' \subseteq I \cup \{(a, t)\}$;
5. $B' = B$.

Revision of non-contingent beliefs in AGM is in many ways analogous to intention revision. However, in a sense, intention revision is subordinate to belief revision. By 5 above, intention revision does not change the (non-contingent) belief set. But it is dependent on the belief set. Conversely, belief revision should not be dependent on the intention set, but it should in general change the intention set. To deal with this, we assume that implicit in any belief revision operator $*$ is an underlying intention revision operator $\circ^*$. We will define a class of belief revision operators that satisfy the following postulates, where again $\langle B, I\rangle * \varphi = \langle B', I'\rangle$.

1. $\langle B', I'\rangle = \langle B', I\rangle \circ^* \epsilon$, where $\circ^*$ satisfies the aforementioned intention revision postulates (ensuring coherence);
2. φ is consistent, iff $\varphi \in B'$;
3. If $\neg\varphi \notin B$, then $Cl(B \cup \{\varphi\}) = B'$;
4. If φ and ψ are equivalent and $\langle B, I\rangle * \psi = \langle B'', I''\rangle$, then $B' = B''$ and $I' = I''$;
5. $B' = Cl(B')$;
6. If $\neg\psi \notin B'$ and $\langle B, I\rangle * \psi = \langle B'', I''\rangle$, then we have $Cl(B' \cup \{\psi\}) \subseteq B''$;
7. If $\langle B, I''\rangle * \varphi = \langle B'', I'''\rangle$, then $B' = B''$.

Essentially, these postulates can be seen as (a slight variation of) AGM plus the intention revision postulates above.

For the rest of the paper we shall make precise how the postulates are to be represented, and in the last section investigate how these postulates look for contingent beliefs.

Logical Preliminaries

Our aim in this section is to develop a simple logical system that will represent the database describing the agent's beliefs about the current moment and future moment and actions that may be performed. We start with a number of simplifying assumptions about time, actions and states. First of all, we assume time is discrete and infinite in both directions, let $\mathbb{Z}$ denote the set of time-points, or moments. Nothing we say crucially depends on this assumption. Second, at each moment, some subset of the set of atomic sentences $\mathsf{Prop} = \{p, q, r, \ldots\}$ are true (intuitively, the generated propositional language describes different *ground facts* about the current state of affairs). Third, there is a finite set of primitive action symbols $\mathsf{Act} = \{a, b, c, \ldots\}$

Entries in the database will be represented by the formal language $\mathcal{L}$ given by the following grammar:

$$\varphi \ := \ p_t \mid pre(a)_t \mid post(a)_t \mid Do(a)_t \mid \Box\varphi \mid \varphi \wedge \varphi \mid \neg\varphi$$

with $p \in \mathsf{Prop}, a \in \mathsf{Act}$, and $t \in \mathbb{Z}$. Intuitively, p_t means that the atomic formula p is true at time t and $Do(a)_t$ means the agent will do (or did) action a at time t. To every action and every time we associate formulas $pre(a)_t$ and $post(a)_{t+1}$, which we treat as distinguished propositional variables, and are understood as the preconditions and postconditions of a at time t. The modal operator is interpreted as historic necessity. The other boolean connectives and the dual modal operator $\Diamond$ are defined as usual.

Definition 1 (Paths)**.** Let P be the set

$$\mathcal{P}(\mathsf{Prop} \cup \{pre(a), post(a) : a \in \mathsf{Act}\}).$$

A *path* $\pi : \mathbb{Z} \to (P \ \times \ \mathsf{Act})$ assigns to each time t the set of proposition-like formulas true at that time, and the next action a on the path. Let $\pi(t)_1$ denote the left projection and $\pi(t)_2$ denotes the right projection. A path is called *appropriate* if the following obtains:

$$\text{If } \pi(t)_2 = a, \text{ then } post(a) \in \pi(t+1)_1.$$

There is a natural equivalence relation on a set Π of paths: we write $\pi \sim_t \pi'$ if for all $t' \leq t$, $\pi(t') = \pi'(t')$. Intuitively, $\pi \sim_t \pi'$ if π and π' represent the same situation up to time t. We extend the definition of *appropriate* to sets of paths by declaring Π to be appropriate if all paths $\pi \in \Pi$ are appropriate and moreover satisfy the following condition:

$$\text{If } pre(a) \in \pi(t)_1, \text{ then there is some } \pi' \sim_t \pi \text{ such that } \pi'(t)_2 = a.$$

Definition 2 (Truth Definition)**.** The truth relation $\models_\Pi$ is defined relative to some underlying appropriate set of paths Π. For convenience we leave off the relativizing subscript.

$\pi, t \models \alpha_{t'}$, iff $\alpha \in \pi(t')_1$, with $\alpha \equiv p, pre(a)$, or $post(a)$.
$\pi, t \models Do(a)_{t'}$, iff $\pi(t')_2 = a$.
$\pi, t \models \Box\varphi$, iff for all $\pi' \in \Pi$, if $\pi \sim_t \pi'$ then $\pi', t \models \varphi$.
$\pi, t \models \varphi \wedge \psi$, iff $\pi, t \models \varphi$ and $\pi, t \models \psi$.
$\pi, t \models \neg\varphi$, iff $\pi, t \not\models \varphi$.

The usual logical notions of satisfiability and validity are defined as usual. We next present a simple sound and complete logic where consistent sets are meant to represent the agent's database describing the "view" of the current situation. The proof of this theorem is by standard techniques.

Theorem 1 (The logic $\mathbf{L}_{Path}$ of paths)**.** *The following logic is sound and strongly complete with respect to the class of all appropriate sets of paths. We call this logic* $\mathbf{L}_{Path}$*.*

1. *Propositional Tautologies;*
2. **S5** *axioms and rules for* $\Box$ *(*$\Box\varphi \to \varphi$*,* $\Box\varphi \to \Box\Box\varphi$*,* $\Diamond\varphi \to \Box\Diamond\varphi$ *and Necessitation: from* φ *infer* $\Box\varphi$*);*
3. $\bigvee_{a \in \mathsf{Act}} Do(a)_t$*;*
4. $Do(a)_t \to \bigwedge_{b \neq a} \neg Do(b)_t$*;*
5. $Do(a)_t \to post(a)_{t+1}$*;*
6. $pre(a)_t \to \Diamond Do(a)_t$*;*
7. *Modus Ponens.*

Modeling Revision

Beliefs in our framework are represented by sets of $\mathbf{L}_{Path}$-consistent formulas of $\mathcal{L}$, or equivalently, as (appropriate) sets of paths. Given a set of formulas B, we can consider the set of paths on which all formulas of B hold at time 0,[4] denoted $\rho(B)$. Conversely, given a set of paths Π, we let $\beta(\Pi)$ be defined as the set of formulas valid at 0 in all paths in Π.[5] We will use this correspondence in the representation theorem. For now we restrict our attention to sets of paths, and in particular we will represent beliefs by the minimal set under a total preorder on paths. Intentions in our models will simply be action/time pairs.

The fact that postconditions of actions always hold on a path, but that preconditions may not, is a direct implementation of our proposal that preconditions, unlike postconditions, need not be believed when an action is intended. Even if all of the paths in some (minimal) set include action a being taken at time t, it need not be that the preconditions also hold along all paths at t. We might therefore think of our belief model as, in some sense, one of "optimistic" or "imaginary" beliefs. On the other hand, we do put a slightly weaker requirement on sets of paths, that the preconditions hold on *some* path in the set. Where again I is a set of pairs (a, t), we require that the joint preconditions of all intended actions not be *disbelieved* by the agent. This is our notion of coherence.

Definition 3 (Coherence)**.** The pair (Π, I) is said to be *coherent* (at time 0) if there is some path $\pi \in \Pi$,

$$\pi, 0 \models \Diamond \bigwedge_{(a,t) \in I} pre(a)_t.$$

Intuitively, intentions cohere with beliefs if the agent considers it possible to carry out all of the intended actions. This is a kind of minimal requirement on *rational balance* between the two mental states.

Remark 1. A word is in order concerning this choice of coherence conditions. Consider our example of the agent that intends to go to the city at 6:00. As we pointed out, it is not actually necessary that the agent believe his car is working; only that he does not believe his car is not working.

[4] As our framework is absent of operations that move time forward, we may assume it is "always" time 0.

[5] In general, $\beta(\rho(B)) = B$, but $\rho(\beta(\Pi)) \neq \Pi$.

Anticipating our treatment of contingent beliefs, we can also ask, what can be our agent's working assumptions about the future, upon adopting this intention? In so far as the agent is *committing* himself to this action, we may assume that he *will* go to the city at 6:00. If we then consider the subset of paths in our belief set on which this action is taken at 6:00, the postconditions will hold along all of them. However, to allow that the preconditions may not yet be believed, we admit paths on which the preconditions do not strictly hold. We only require that they hold on *some* path in the set, so that the agent cannot stray too far from reality.

Indeed, this is arguably closer to how we reason about future actions. We often commit to actions without explicitly considering the path that will lead us there. Eventually this decision will have to be made, but there is nothing incoherent about glossing over these details at the current moment. Our example agent should assume he will be in the city by 7:00 and can continue making plans about what he will do in the city once he is there. But he should not assume the preconditions will hold until he has made further, specific plans for bringing them about. This topic will be revisited in the penultimate section.

From here on we assume a coherent pair (Π, I), and define revision operations on these sets that preserve coherence. These operations will be used to represent our revision postulates. Selection functions, defined here, are simply the intention revision postulates given in the first section, under a different guise.

Definition 4 (Selection Function)**.** A *selection function* γ is a function that assigns an intention set to a tuple consisting of a set of paths, an intention set and a pair (a, t) satisfying the following conditions. If $\gamma(\Pi, I, (a, t)) = I'$ then,

1. (Π, I') is coherent;
2. If $(\Pi, \{(a, t)\})$ is coherent, $(a, t) \in I'$;
3. If $(\Pi, I \cup \{(a, t)\}$ is coherent, then $I' = I \cup \{(a, t)\}$.
4. $I' \subseteq I \cup \{(a, t)\}$.

In the simple case of the empty intention pair ϵ, this reduces merely to requiring coherence.

Definition 5 (Belief Sets)**.** Suppose Π is an appropriate set of paths. If we define a total preorder $\leq$ on Π, then the *belief set* of $(\Pi, \leq)$ is the set $\{\pi \in \Pi : \pi \leq \pi' \text{ for all } \pi' \in \Pi\}$. We denote this by $\min_{\leq}(\Pi)$, or just $\min(\Pi)$ when the ordering is understood from context.

Definition 6 (Belief Intention Model)**.** A belief-intention model is a triple $(\Pi, \leq, I, \gamma)$ where Π is a set of paths, $\leq$ is a total preorder on Π, I is a finite set of pairs (a, t) with $a \in \mathsf{Act}$ and $t \in \mathbb{Z}^+$, $(\min(\Pi), I)$ is coherent and γ is a selection function.

Definition 7 (Adding an Intention)**.** Let $(\Pi, \leq, I, \gamma)$ be a belief-intention model. Adding the intention (a, t) results in the model $(\Pi, \leq, I', \gamma')$ where $I' = \gamma(\min(\Pi), I, (a, t))$ and $\gamma' = \gamma$. We denote this model by $(\Pi, \leq, I, \gamma) \bullet (a, t)$.[6]

[6]Notice that this setup allows the possibility that $\gamma' \neq \gamma$, so that after revision the selection function itself can change. Of course this would only become interesting in the iterated case

Definition 8 (Adding a Belief)**.** Let $(\Pi, \leq, I, \gamma)$ be a belief-intention model. Adding a (consistent) belief φ results in the model $(\Pi, \leq', I', \gamma')$, where $\gamma' = \gamma$, $I' = \gamma(\min_{\leq'}(\Pi), I, \epsilon)$, and $\leq'$ is defined so that $\pi \leq' \pi'$, if and only if one of the following holds:

1. $\pi, 0 \vDash \varphi$ and $\pi', 0 \nvDash \varphi$;
2. $\pi, 0 \vDash \varphi$ and $\pi', 0 \vDash \varphi$, and $\pi \leq \pi'$;
3. $\pi, 0 \nvDash \varphi$ and $\pi', 0 \nvDash \varphi$, and $\pi \leq \pi'$.

This is the so-called *lexicographic* reordering operation, familiar from the belief revision and dynamic epistemic logic literatures. We denote the new belief-intention model by $(\Pi, \leq, I, \gamma) \star \varphi$.

Remark 2. Lexicographic reordering is only one of many possible choices one could make here, and we adopt it only for concreteness. When we go on in future work to consider the problem of iterated revision, this decision will become more important. For now, it is sufficient to choose any revision policy that obeys the AGM postulates, as belief revision *per se* is not our central concern.

Representation of Revision Postulates

We are now ready to represent the postulates in full detail. In the following let $Cl(X)$ denote the closure of a set X of $\mathcal{L}$ formulas under consequence in $\mathbf{L}_{Path}$. And if I is a finite set of pairs (a, t), with $a \in \mathsf{Act}$ and $t \in \mathbb{Z}^+$, define,

$$Cohere_I := \Diamond \bigwedge_{(a,t)\in I} pre(a)_t.$$

Definition 9 (Belief Intention Base)**.** A belief intention base is a pair $\langle B, I\rangle$, where:

- B is a consistent set of formulas such that $Cl(B) = B$.
- I is a finite set of pairs (a, t).

Definition 10 (Coherence)**.** A belief-intention base $\langle B, I\rangle$ is *coherent* if $\neg Cohere_I \notin B$.

We then have the following obvious correspondence.

Lemma 1. *$\langle B, I\rangle$ is coherent, iff $(\rho(B), I)$ is coherent.*

Now having provided all of the necessary formal details, we repeat our postulates for intention and belief revision.

Definition 11 (Intention Revision)**.** Suppose $\langle B, I\rangle \circ (a, t) = \langle B', I'\rangle$. The operator $\circ$ is called *proper* if the following conditions obtain.

1. $\langle B', I'\rangle$ is coherent;
2. If $\langle B, \{(a, t)\}\rangle$ is coherent, then $(a, t) \in I'$;
3. If $\langle B, I \cup \{(a, t)\}\rangle$ is coherent, then $I \cup \{(a, t)\} \subseteq I'$;
4. $I' \subseteq I \cup \{(a, t)\}$;
5. $B' = B$.

The first postulate simply says that intention revision should restore coherence. The second postulate says that the new intention (a, t) takes precedence over all other currently held intentions; it should be added if it is possible to maintain coherence, even if this means discarding current intentions. The third postulate, taken together with the fourth postulate, says that if it is possible to maintain coherence by

simply adding the new intention, then this is the only change that is made. The fourth in addition guarantees that, unlike in the case of belief revision below, no extraneous intentions are ever added.[7] Finally, the fifth postulate says that non-contingent beliefs do not change with intention revision.

Recall that we assume every belief revision operator $*$ is given with its own intention revision operator $\circ^*$, so that a belief revision may trigger an intention revision.

Definition 12 (Belief Revision)**.** Suppose $\langle B, I\rangle * \varphi = \langle B', I'\rangle$. The operator $*$ is called *proper* if the following conditions obtain.

1. $\langle B', I'\rangle = \langle B', I\rangle \circ^* \epsilon$, where $\circ^*$ is proper;
2. φ is consistent, iff $\varphi \in B'$;
3. If $\neg\varphi \notin B$, then $Cl(B \cup \{\varphi\}) = B'$;
4. If $\mathbf{L}_{Path} \vdash \varphi \leftrightarrow \psi$ and $\langle B, I\rangle * \psi = \langle B'', I''\rangle$, then $B' = B''$ and $I' = I''$;
5. $B' = Cl(B')$;
6. If $\neg\psi \notin B'$ and $\langle B, I\rangle * \psi = \langle B'', I''\rangle$, then we have $Cl(B' \cup \{\psi\}) \subseteq B''$;
7. If $\langle B, I''\rangle * \varphi = \langle B'', I'''\rangle$, then $B' = B''$.

Postulate 1 simply says that if intention revision is necessary to retain coherence, this revision is itself proper. Postulate 2 is a slight variation of the AGM success postulate, which we adopt on a par with intention revision postulate 2. In this setting it only makes sense to adopt a new belief if it is non-contradictory. Postulates 3-6 fill out the rest of the AGM theory, and postulate 7 says that the underlying intention set is irrelevant to belief revision.

We can now represent these postulates in terms of the belief intention models of Definition 6.

Theorem 1 (Representation Theorem)**.** *For every belief intention base $\langle B, I\rangle$, with proper revision functions $*$ and $\circ^*$, there is a belief intention model $(\Pi, \leq, I, \gamma)$, such that:*

1. *$\rho(B) = min_{\leq}(\Pi)$;*
2. *I is the same set in the base and in the model;*
3. *For all $\varphi \in \mathcal{L}$: If $(\Pi, \leq, I, \gamma) \star \varphi = (\Pi, \leq', I', \gamma')$ and $\langle B, I\rangle * \varphi = \langle B', I''\rangle$, then,*

$$\rho(B') = min_{\leq'}(\Pi), \text{ and } I' = I''.$$

The proof of this theorem simply rides on the proof of the representation theorem for AGM in terms of the "system of spheres" interpretation (Grove 1988), with the intention revisions simply going along for the ride.

Contingent Beliefs

Definition 13. A contingent belief set B^I is derived from a belief-intention base $\langle B, I\rangle$ in the following way:

$$B^I = Cl(B \cup \{Do(a)_t : (a, t) \in I\}).$$

That is, one believes everything that was already believed non-contingently, and moreover that any actions the agent has committed to will in fact be carried out, in addition to everything that follows from this assumption, including that the postconditions of all intended actions will hold. In fact, B^I itself gives rise to a well defined belief base. This proposition follows directly from Definition 10 and the logic $\mathbf{L}_{Path}$.

Proposition 1. *If $\langle B, I\rangle$ is a coherent belief-intention base, then B^I is a consistent belief set.*

Notably the reverse direction of Proposition 1 does not hold. This is because of the nonparallel we have drawn between believing in preconditions and believing in postconditions (see Remark 1).

Now that we may treat B^I as a kind of belief base in its own right, we can consider what the postulates on belief and intention revision look like on the single set. The following proposition shows how the revision operators in Definitions 11 and 12 manifest themselves in the set of contingent beliefs. We give the postulates solely in terms of the set B^I itself (with no mention of the set B from which it is derived). Some information is lost with this restriction, including the distinction between non-contingently believed formulas and formulas that were added because of intentions. But arguably, this represents the kind of information the planner would solicit from the database. We shall write $B^I \circ (a, t)$ for the set $B'^{I'}$ where $\langle B, I\rangle \circ (a, t) = \langle B', I'\rangle$, and likewise for $B^I * \varphi$. We make no claim to completeness here, but verification of soundness is straightforward.

Proposition 2. *The following postulates hold for any a, t, and φ, assuming $\circ$ and $*$ are proper.*

Intention Revision

1. $B^I \circ (a, t)$ is consistent;
2. If $\neg Cohere_{I\cup\{(a,t)\}} \notin B^I$, then $B^I \circ (a, t) = Cl(B^I \cup \{post(a)_{t+1}\})$;
3. If $\varphi \notin B$ and $post(a)_{t+1} \rightarrow \varphi \notin B^I$, then $\varphi \notin B^I \circ (a, t)$;
4. If $\varphi \in B^I$ and $\varphi \wedge \bigwedge_{(b,u)\in I} \neg post(b)_{u+1}$ is consistent, then $\varphi \in B^I \circ (a, t)$.

Belief Revision:

1. $B^I * \varphi$ is consistent.
2. If $\neg\varphi \notin B^I$ and $\varphi \rightarrow \neg Cohere_I \notin B^I$, then $B^I * \varphi = Cl(B^I \cup \{\varphi\})$;
3. If φ is consistent, $\varphi \in B^I * \varphi$.
4. $B^I * \varphi = Cl(B^I * \varphi)$;
5. If φ and ψ are $\mathbf{L}_{Path}$-equivalent, then $B^I * \varphi = B^I * \psi$.

These postulates closely mirror those for intention revision and belief revision separately. Take first the postulates for intention revision. 1 says that the new set should be consistent. 2 says that the new intention should simply be added if it is possible to do so and still maintain consistency (that is, coherence of the underlying belief-intention base). 3 ensures that no extraneous beliefs result from adding a new

[7] This postulate, in particular, could be lifted depending on the application. Since we are modeling a database, we do not want the database to engage in any kind of planning. Adding new intentions when old intentions become inconsistent amounts to planning.

intention. And 4 guarantees that beliefs unrelated to the intention set, in particular those in B that have nothing to do with I, should remain, i.e. that an intention revision should not change the non-contingent beliefs.

It is interesting to note that when considering B^I, a proper belief revision operator $*$ will not, strictly speaking, satisfy all of the AGM postulates. For example, we see in Belief Revision Postulate 2 the need for an extra condition. Even if $\neg\varphi \notin B^I$, we may not simply take the closure of B^I and φ, since adding φ may trigger removal of intentions, which in turn may trigger removal of beliefs from B^I. So postulate 3 of Definition 12 requires the extra hypothesis that $\neg Cohere_I \notin B^I$. Otherwise, the postulates follow the same spirit as the AGM postulates we had for belief intention bases. Postulate 1 perfectly mirror the corresponding postulates for intention revision, ensuring consistency, and simple addition in the case that the new belief can be consistently added. Postulates 3-5 are directly inherited from the AGM postulates we had in Definition 12.

It would be possible to obtain even more detailed postulates, were we to label formulas in B^I by their "justifications". For example, $post(a)_{t+1}$ could be in B^I either because it is believed non-contingently or because (a, t) is in I. Labeling formulas in this way would amount to separating B and I as we do in belief-intention bases, so we leave this possibility aside. B^I allows for a slightly simpler, if also conflated, picture of how beliefs and intentions conspire to give rise to contingent beliefs.

Related Work

Starting with Cohen and Levesque's classic paper (Cohen and Levesque 1990), many logical systems have been developed for reasoning about informational and motivational attitudes, including intentions, in a dynamic environment (see, Meyer and Veltman 2007 and van der Hoek and Wooldridge 2003, for surveys). The central issues in this literature are (i) how to characterize the process of intention *generation*, i.e. certain kinds of practical reasoning, and (ii) how to model the *persistence* of the agents' intentions over time (see, Herzig and Lorini 2008, for a survey of the philosophical and logical literature surrounding these two issues). The problem addressed in this paper, namely how an agent should revise beliefs and intentions together given new information or a change of plans, has received relatively little attention (cf. Georgeff and Rao 1995; van der Hoek, Jamroga, and Wooldridge 2007; Lorini et al. 2009; Roy 2009; Shoham 2009).[8]

Broadly speaking, the logical framework we use in this paper falls into the category of the so-called "BDI logics" mentioned above, in the sense that we model an agent using the mental states of *belief* and *intention* (we leave out *desires*). We do not have the space to go into a detailed comparison with the many different BDI approaches. Instead we highlight some key details about our logical system that will help place it in this literature. Our semantics (Definition 1) is closest to the branching-time models of Rao and Georgeff (1992). However, one important difference is that we focus on the intention to perform an action *at a specific moment in time*. The benefits of this are discussed at length in (Shoham 2009). Our treatment also shares some features with (van der Hoek, Jamroga, and Wooldridge 2007), which also proposes a formal model of intention and belief revision. Some of the basic intuitions are similar (eg., contingent beliefs are quite similar to their *weak beliefs* – however, see above), but there are also fundamental differences. Van der Hoek, Jamroga and Wooldridge extend a BDI logic with a dynamic modal operator describing what is true after the agent makes an observation. Thus, intention revision and belief revision are characterized in the formal language as validities in their logic.[9] More importantly, we differ on a number of basic conceptual issues. For example, in this paper, plans are not explicitly part of the framework, but, as a feature of the database perspective, are conceived of in the background as a recipe describing precisely what actions the agent will perform at specific moments in time. In their framework a plan describes what needs to be true in order to fulfill some desire, and consequently they focus on the problem of revising intentions and beliefs in the presence of new information and less on the effect adopting new intentions has on beliefs.

Conclusions and Future Work

We have presented a framework for reasoning about joint revision of beliefs and intentions. Already in the case of a single revision a number of subtle issues arise. We have chosen to address these issues by adopting a particular stance on what intentions are and how they relate to beliefs, which we have called the database perspective (Shoham 2009). By viewing the problem of joint belief and intention revision as a database management problem, we have been able to bypass some of the more vexing problems about intention familiar from the philosophical literature, while at the same time confronting some basic logical problems of practical significance.

In a sense, one can see the AGM framework for belief revision as identifying what the problem of belief revision is in the first place. The standard postulates can be taken as *constitutive* of a particular kind of doxastic action, according to which the agent has committed to believing some new piece of information and must integrate this new belief with old beliefs. The interesting questions, on this view, arise when we ask how this simple picture can be embellished, to deal with iterated belief revision, interaction with other mental states and actions, and so on.[10] In the same way, one can view our treatment of joint intention and belief revision in

[8] This list contains only papers that focus on logical systems that explicitly represent how an agent's intentions (and other mental attitudes) can change in the presence of new information. Indeed, philosophers and computer scientists have discussed a number of issues relevant to the problem we study in this paper. A complete survey of such issues is outside the scope of this paper (Shoham 2009, has pointers to some relevant papers).

[9] See (van Benthem 2004) for a comparison between these two modeling styles for belief change, *vis-à-vis* AGM-style postulates versus modal languages with model change operators.

[10] A view like this is taken, for example, in (Stalnaker 2009).

this paper as a proposal to define what the problem is about, and to propose a framework in which further questions can be fruitfully asked and explored. Indeed, there are many directions from here that should be explored. A few of the main directions would include:

- We have mentioned several times the problem of iterated revision. This is an important and difficult topic that already comes up with belief revision by itself, and is of great interest both practically and theoretically. A large literature already exists on this problem (see, e.g. (Darwiche and Pearl 1997; Boutilier 1996)), but there is still further work to be done (c.f. (Stalnaker 2009)).
- In this paper only atomic actions are considered. However, agents typically reason with more elaborate representations of plans, and these more elaborate representations would undoubtedly interact with beliefs in subtle ways. For example, it may not be immediately clear how our definition of coherence should be adapted to a setting in which one has conditional intentions (e.g. 'Action a, if φ, b otherwise'). But such intentions are crucial for agents planning in uncertain environments.
- Other mental attitudes, like goals, desires and preferences, we have left out completely. This is not because we assume they are unimportant, but rather because we want to focus on these particular issues that arise in the interaction between belief and intention. To be sure, other interesting issues surface when belief and intention are treated together with other attitudes.

We think these are all exciting and important questions, and there are many more (see (Shoham 2009) for a longer list). They are all left for future work.

References

Alchourrón, C. E.; Gärdenfors, P.; and Makinson, D. 1985. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic* 50(2):510 – 530.

Boutilier, C. 1996. Iterated revision and minimal change of conditional beliefs. *Journal of Philosophical Logic* 25(3).

Bratman, M. 1987. *Intention, Plans and Practical Reason*. Harvard University Press.

Castelfranchi, C., and Paglieri, F. 2007. The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions. *Synthese* 155:237 – 263.

Cohen, P. R., and Levesque, H. 1990. Intention is choice with committment. *Artificial Intelligence* 42(3):213 — 261.

Darwiche, A., and Pearl, J. 1997. On the logic of iterated belief revision. *Artificial Intelligence* 89:1–29.

Georgeff, M. P., and Rao, A. S. 1995. The semantics of intention maintenance for rational agents. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, 704–710.

Grove, A. 1988. Two modellings for theory change. *Journal of Philosophical Logic* 17.

Herzig, A., and Lorini, E. 2008. A logic of intention and attempt. *Synthese* 163(1):45 – 77.

Lorini, E.; Dastani, M.; van Ditmarsch, H.; Herzig, A.; and Meyer, J.-J. 2009. Intention and assignments. In He, X.; Horty, J.; and Pacuit, E., eds., *Logic, Rationality and Interaction*, volume 5834 of *Lecture Notes in Computer Science*. Springer.

Meyer, J.-J., and Veltman, F. 2007. *Handbook of Modal Logic*. Elsevier. chapter Intelligent Agents and Common Sense Reasoining.

Meyer, J.-J.; van der Hoek, W.; and van Linder, B. 1999. A logical approach to the dynamics of commitments. *Artificial Intelligence* 113:1 – 40.

Rao, A. S., and Georgeff, M. 1992. Modeling rational agents within a BDI-architecture. In Fikes, R., and Sandewall, E., eds., *Proceedings of Knowledge Representation and Reasoning (KR & R)*.

Roy, O. 2009. A dynamic-epistemic hybrid logic for intentions and information changes in strategic games. *Synthese* 171:291 – 320.

Shoham, Y. 2009. Logical theories of intention and the database perspective. *Journal of Philosophical Logic* 38(6).

Stalnaker, R. 2009. Iterated belief revision. *Erkenntnis* 70.

van Benthem, J. 2004. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics* 14.

van Benthem, J. 2009. For better or for worse: Dynamic logics of preference change. In *Preference Change*, volume 42 of *Theory and Decision Library A*. Springer. 57 – 84.

van der Hoek, W., and Wooldridge, M. 2003. Towards a logic of rational agency. *Logic Journal of the IGPL* 11(2):135 – 160.

van der Hoek, W.; Jamroga, W.; and Wooldridge, M. 2007. Towards a theory of intention revision. *Synthese* 155:265 – 290.