A Unified View of Belief Revision and Update

Alvaro del Val and Yoav Shoham
Robotics Lab
Computer Science Department
Stanford University
Stanford, CA 94035
{delval,shoham}@cs.stanford.edu

Abstract

Belief *revision* and belief *update* have been considered as two distinct types of belief modification. In this paper, we show that both can be captured within an unified formalism, in which revision is encoded as *update of the mental state*. Specifically, we show that the framework for belief update presented in [6, 7] can be used to capture revision as well, by extending the underlying language to include explicit representations of the agent's beliefs. The resulting framework allows for the definition of a wide variety of revision operators. These include, in particular, all the AGM operators, for which we provide a representation theorem.

1 Introduction

Much recent work on belief revision has taken as starting point the work of Alchourrón, Gärdenfors, and Makinson [1, 11], who propose a number of normative conditions or "postulates" that every "rational" belief change operator should satisfy. Katsuno and Mendelzon [16] recently noted, however, that AGM revision is not adequate for reasoning about changes in the world, and suggested a distinction, that appears to have become accepted in the literature, between belief *revision* and *update*. Loosely speaking, the former says that the beliefs may have been wrong and in need of revision, whereas the latter says that the beliefs were correct, but the world has in the meanwhile evolved and the beliefs must be updated. Katsuno and Mendelzon [16] proposed a set of belief-update postulates, which are similar to, but distinct from, the belief-revision postulates; they also provide the model theory for these postulates, in the form of a representation theorem.

In [7] we argued that, since the fundamental problem in update is one of *reasoning about change*, any proposal for update should be defended in terms of its ability to solve traditional problems of non-monotonic temporal reasoning, such as the well-known frame and ramification problems.[1] In [6, 7], specifically, we encoded update in a theory of action allowing for explicit reasoning about default persistence (a generalization of the proposal of Lin and Shoham [19]), and showed that the framework thus

[1] A similar argument, restricted to the frame problem, has been advanced by Reiter in [22].

developed allows us to identify exact conditions for the applicability of KM semantics (in the form of a representation theorem), as well as circumstances under which we need to move beyond KM semantics.

The intuition underlying these results, however, appeared not to be straightforwardly applicable to belief revision, and thus left open the question of what is the relation between revision and update. The key idea linking both types of belief change was first formulated, in a different framework, by Grahne *et. al.* [13], and can be summarized by the slogan: *revision is update of the mental state*; in other words, the revision of the agent's beliefs can be seen as the update of a theory in which the agent's beliefs are explicitly represented. In this paper, we show that the framework for update of [7] can be used to capture this intuition, in a way that allows to axiomatically define a wide variety of revision operators. These include, in particular, all AGM operators, for which we provide a representation theorem in terms of theories of action.

The basic idea of our encoding of update in [7] is as follows. Although update is supposed to reflect changes that have taken place in the world over time, the update problem, like that of belief revision, has almost always been formulated in a language that incorporates no model of time or change. Our approach consists in translating the update problem into the richer language of situation calculus, so as to make explicit this temporal information. The initial database is taken to describe a particular situation, and the update formula is taken to describe the occurrence of a special action, denoted by A_μ^S, whose intuitive reading is "that action which, when taken in S, causes μ." The updated database is taken to describe the situation $result(A_\mu^S, S)$. A formal theory of action is then used to infer facts about the result of taking the particular action in the particular situation. Finally, anything inferred about the resulting situation can be translated back, if desired, to the timeless framework of belief update.

In order to apply these ideas to revision, we make the mental state of the agent part of our representation of the world, and add actions which affect this mental state. Specifically, we will first present a "mental situation calculus" (to use a term due to John McCarthy), which extends situation calculus with epistemic operators for belief and "believing only." The properties of both operators are then explicitly axiomatized. For example, we will require beliefs to be consistent and closed under logical consequence.[2] The initial database ψ, representing the agent's beliefs, is described in our language by stating that the agent *believes only* ψ. This notion, which is closely related to the concept of "all I know" in [15] and "believe only" in [18], is needed to capture the idea that the agent has no beliefs other than those entailed by the database. The revision of the initial set of beliefs is captured in terms of the effects of a particular action, namely the action of *learning* the new information, which causes the agent to believe it. Finally, this extended theory of action, coupled with the same circumscriptive policy used for update, is used to infer the agent's beliefs in the situation resulting from learning the new information, *i.e.*, the "revised database."

We remark that the mental situation calculus presented here is fairly restricted, in that it does not allow for nested beliefs, and thus does not provide the full expressive power of traditional modal logics of belief. However, in this our approach is not

[2] Other notions of belief could also be used, as long as they can be axiomatized in the style of section 4.

different from most approaches to belief revision in the literature, that have left the question of revision of beliefs about beliefs for the most part unexplored.

The structure of this paper is as follows. The next two sections review, respectively, the AGM approach to belief revision, and our encoding of update. In section 4 we present the mental situation calculus, which is then used to encode revision in a theory of action in section 5. Our main technical results can be found in section 6, where we provide a representation theorem for AGM operators in terms of this theory of action. These results, and the encoding in which they are based, use a translation between a propositional language and the situation calculus. In section 7 we show that this metalinguistic approach is unnecessary: all inferences about revision and update can be carried within the theory of action. Related work, in particular [13], is discussed in the concluding section.

2 Normative approaches to belief revision: Review

In this section we review the AGM approach to belief revision. For readability, we follow the presentation of [17]; as in [17], we assume in the rest of the paper a propositional language $\mathcal{L}$ with a finite set $\mathcal{P}_{\mathcal{L}}$ of variables. Given a database (sentence) ψ and a sentence μ, $\psi \circ \mu$ denotes the knowledge base obtained by *revising* ψ with the new information μ; $\psi \diamond \mu$ is used for update. $Mod(\phi)$ denotes the set of models of the formula ϕ. We also say that a binary relation $\leq$ is a total preorder iff it is reflexive, transitive, and connected. Finally, for any subdomain S of $\leq$, $Min(S, \leq)$ represents the set of elements of S minimal under $\leq$.

AGM postulates for revision:

(R1) $\psi \circ \mu$ implies μ.

(R2) If $\psi \wedge \mu$ is satisfiable then $\psi \circ \mu$ is equivalent to $\psi \wedge \mu$.

(R3) If μ is satisfiable then $\psi \circ \mu$ is also satisfiable.

(R4) If $\models \psi_1 \equiv \psi_2$ and $\models \mu_1 \equiv \mu_2$ then $\psi_1 \circ \mu_1$ is equivalent to $\psi_2 \circ \mu_2$.

(R5) $(\psi \circ \mu) \wedge \phi$ implies $\psi \circ (\mu \wedge \phi)$.

(R6) If $(\psi \circ \mu) \wedge \phi$ is satisfiable then $\psi \circ (\mu \wedge \phi)$ implies $(\psi \circ \mu) \wedge \phi$.

Some support for the postulates derives from Katsuno and Mendelzon's characterization of AGM revision operators [17], provided by the following representation theorem. A *revision assignment* is a function which assigns to each propositional formula ψ a binary relation $\leq_\psi$ over the set of all interpretations $\mathcal{W}$ of $\mathcal{L}$. The revision assignment is said to be *faithful* iff (i) $Min(\mathcal{W}, \leq_\psi) = Mod(\psi)$ for any satisfiable ψ; and (ii) $\leq_\psi = \leq_\phi$ whenever $\models \psi \equiv \phi$.

Theorem 1 ([17]) A revision operator $\circ$ satisfies conditions (R1)-(R6) iff there exists a faithful revision assignment that maps each formula ψ to a total preorder $\leq_\psi$ such that:

$$Mod(\psi \circ \mu) = Min(Mod(\mu), \leq_\psi)$$

3 Update in theories of action: Review

The motivation behind our encoding of update in [7] has been briefly discussed in the introduction; in this section we describe the theory of action we use. We cover it in some detail, since it is by extending this theory that we later encode revision. Our initial language, to be extended later, is that of situation calculus, a sorted predicate calculus with action, situation, and fluent sorts, a subsort of which is propositional fluents. (Belief fluents, the other subsort, is introduced later.) There are four binary predicates, *holds*, *holds'*, *persistent* and *persistent'*, all of them with arguments of sort fluent and situation, and a binary function *result*, with arguments of sort action and situation, and value of sort situation. Closed action terms are all terms of the form A^S_μ, for μ a satisfiable formula of $\mathcal{L}$ and S a situation term, with the intuitive meaning described above. Closed situation terms are S_0, intuitively denoting the initial situation, and those formed with *result*. The set $\mathcal{P}$ of closed propositional fluent terms, finally, is simply defined to be the propositional language $\mathcal{L}$, where connectives are treated as propositional fluent valued functions, and appropriately axiomatized.[3] Lower case Greek letter variables range over fluents; we will keep this convention when belief fluents are introduced, unless it is explicitly indicated or the context requires that they range over propositional fluents.

Formulas of the form $holds(p, S)$ are used to capture the set of facts that hold at a given situation. Updates are captured in terms of the effects of actions, which are governed by the axiom schema:

$$holds(\mu, result(A^S_\mu, S)).$$

This schema captures the direct effects of updates, but it does not say anything about whether other facts should or should not change in the resulting situation. For this purpose, we use a circumscription policy that minimizes changes from situation to situation with respect to a set of fluents declared to be "persistent by default" in the given situation, using the predicate *persistent*. In order to do this, we introduce an "abnormality predicate" $ab(\sigma, a, s)$ as an abbreviation for:

$$persistent(\sigma, s) \wedge (holds(\sigma, s) \equiv \neg holds(\sigma, result(a, s))).$$

The idea is then to minimize the extension of $ab(\sigma, a/A, s/S)$ for each situation S and action A. (The notation "a/A," etc., is used for substitution.) Clearly, the result of this minimization will vary according to which fluents are "persistent." We assume therefore that there is an (update) *persistence axiom* (P_U) stating which facts should be treated as persistent in each situation. This axiom might, for example, state that the set of persistent fluent terms is, in every situation, the set $\mathcal{P}_\mathcal{L}$ of "primitive" fluent terms (i.e. the propositional vocabulary of $\mathcal{L}$). The choice of this persistence axiom is left as a parameter in the theory of action, subject to only some quite weak "well-behavedness" conditions;[4] *different choices determine different "update operators."*

[3] *E.g.*, $\forall s.\ holds(\neg p, s) \equiv \neg holds(p, s)$, etc. This slightly abusive notational convention will not result in any ambiguity and makes for much easier reading.

[4] The formal statement of these conditions, however, is complicated, and will be omitted. Their goal is to ensure that only named fluents are in the extension of *persistent*, that (P_U) "is only about" persistence (via a conservativeness requirement), and that the set of persistent fluents at any situation depends only on the current or past states of the world.

The previous axioms constitute our "theory of action" T. In order to apply the circumscription, we need to add unique names axioms $(N1)$ for fluents and $(N2)$ for situations, and the formula (with free variable s):

$$\forall\sigma.\ holds(\sigma, s) \equiv holds'(\sigma, s) \wedge \forall\sigma.\ persistent(\sigma, s) \equiv persistent'(\sigma, s) \qquad (1)$$

Let $W(s)$ be obtained from T by adding $(N1)$, $(N2)$, and (1). The completion of the theory of action, denoted $Comp(T)$, stands for the union over all closed terms A and S of the appropriate sorts of:

$$Circum(W(S); ab(\sigma, a/A, s/S); holds, persistent),$$

i.e., the circumscription of $ab(\sigma, a/A, s/S)$ in $W(S)$ with $holds$ and $persistent$ allowed to vary. Intuitively, what this circumscription policy does is minimize changes one situation at a time. For any situation S, the minimization will allow $holds$ to vary at any other point except at S, since $holds'$ is kept fixed, and similarly for $persistent$.

Finally, we can define an update operator $\diamond$, for any database ψ and satisfiable formula μ, as follows:

Definition 1 $\psi \diamond \mu \models \sigma$ iff $Comp(T) \models holds(\psi, S_0) \supset holds(\sigma, result(A_\mu^{S_0}, S_0))$.

In [7] we prove a representation theorem for KM update semantics in terms of a number of restrictions on the persistence axiom analogous to those provided in section 5. We refer the reader to this paper for further details and motivation.

Example 1 Let $\forall s, \sigma.\ persistent(\sigma, s) \equiv (\bigvee_{p \in \mathcal{P}_\mathcal{L}} \sigma = p)$ be the persistence axiom, *i.e* the set of persistent fluents in every situation is identical to the set $\mathcal{P}_\mathcal{L}$ of symbols of the initial propositional language. The update operator determined by definition 1 with this axiom is identical to Winslett's PMA update operator [24]. As noted in [3, 7], however, more complex persistence policies are often needed, for example to properly deal with non-deterministic actions. ∎

4 A mental situation calculus

In this section, we extend the situation calculus to include explicit representations of beliefs; the next section uses the extended language to encode revision. The agent's beliefs will be expressed by formulas such as $holds(bel(\psi), S)$, intuitively meaning that the agent believes ψ in situation S. We consider only a language without nested beliefs, and thus, as said, weaker than traditional modal logics of belief. Specifically, the set of belief fluent terms is the set of terms of the form $bel(\psi)$, where bel is an unary function with argument of sort *propositional* fluent and value of sort *belief* fluent. Letting Cn stand for standard propositional consequence, the properties of belief are given by the following axioms:

$$\forall s, \sigma.\ \neg holds(bel(\sigma \wedge \neg\sigma), s) \qquad (2)$$

$$\forall s.\ holds(bel(\sigma), s), \text{ for any } \sigma \in Cn(\emptyset) \qquad (3)$$

$$\forall s, \sigma_1, \sigma_2.\ holds(bel(\sigma_1), s) \wedge holds(bel(\sigma_2), s) \supset holds(bel(\sigma_1 \wedge \sigma_2), s) \qquad (4)$$

$$\forall s, \sigma_1, \sigma_2.\ (holds(bel(\sigma_1), s) \wedge holds(bel(\sigma_1 \supset \sigma_2), s) \supset holds(bel(\sigma_2), s) \qquad (5)$$

Thus, the agent's beliefs are assumed to be consistent, to include all propositional tautologies, and to be closed under conjunction and modus ponens.[5]

The function *bel* is not sufficient to uniquely fix the agent's set of beliefs, in the sense that $holds(bel(\psi), S)$ leaves open whether the agent has other beliefs besides those that logically follow from ψ. For this, as already anticipated, we need the notion of *believing only*. Let *belonly* be a unary function with same sorts as *bel*. The intent is that the agent believes only ψ at situation S, written $holds(belonly(\psi), S)$, iff its set of beliefs is identical to the set of logical consequences of ψ. First, we need the axiom:

$$\begin{aligned}&\forall s, \psi.\ holds(belonly(\psi), s) \equiv \\ &\qquad (\forall \sigma.\ holds(bel(\sigma), s) \equiv \forall s'.\ holds(bel(\psi), s') \supset holds(bel(\sigma), s'))\end{aligned} \tag{6}$$

The right hand side of the innermost equivalence is intended to exactly capture the circumstances in which $\sigma \in Cn(\psi)$ (see part 1 of proposition 1 below); for it to achieve its intended effect, however, it must be supplemented with an "existence of situations" axiom for every *satisfiable* ψ:

$$\exists s.\ holds(bel(\psi), s). \tag{7}$$

This completes the axiomatization of the epistemic operators. Before stating some of its properties, we need the notion of belief state, intuitively a complete specification of the agent's beliefs at a given situation. Formally, we say that R is a *belief state of situation* S iff there exists a set of propositional fluent terms $T(R)$ such that:

$$R = \{holds(bel(\sigma), S) \mid \sigma \in T(R)\} \cup \{\neg holds(bel(\sigma), S) \mid \sigma \in \mathcal{P} \setminus T(R)\},$$

and R is consistent with axioms 2–5. It follows that the set $T(R)$ (treated as a set of propositional formulas) is closed under logical consequence; we will refer to it as "the theory believed in R," and often treat it as a propositional theory. Notice also that every model M of these axioms uniquely determines a belief state $M_B(S)$ for each situation S.

Proposition 1 For every ψ and σ:

1. $[2\text{–}7 \models \forall s.\ holds(bel(\psi), s) \supset holds(bel(\sigma), s)]$ if and only if $\sigma \in Cn(\psi)$.
2. $[2\text{–}7 \models \forall s.\ holds(belonly(\psi), s) \supset holds(bel(\sigma), s)]$ if and only if $\sigma \in Cn(\psi)$.
3. $6 \models (\forall s.\ holds(belonly(\psi), s) \supset holds(bel(\sigma), s)) \lor$
   $(\forall s.\ holds(belonly(\psi), s) \supset \neg holds(bel(\sigma), s))$.
4. For every belief state R of a situation S, there exists some formula ψ_R such that 2–7,$R \models holds(belonly(\psi_R), S)$, with $T(R) = Cn(\psi_R)$.

[5]There will often be some beliefs that act as "constraints," *i.e.* that we want to preserve no matter what new information arrives. These "protected beliefs" can easily be handled by strengthening the consequence operator Cn. Given a set of protected beliefs Γ, we define the consequence operation Cn' so that for any set of formulas Σ, $\sigma \in Cn'(\Sigma)$ iff $\sigma \in Cn(\Sigma \cup \Gamma)$, and substitute Cn' for Cn in the axioms.

5 Encoding belief revision

The basic idea in our encoding of belief revision is to add a "causal axiom" indicating that the new information is believed as a result of "learning" it, and to apply the circumscription policy used in section 3 to capture update. Formally, let *learn* be a unary action-valued function with argument of sort propositional fluent. The causal axiom for *learn* is:

$$\forall s, \mu.\ (\exists s'.\ holds(bel(\mu), s')) \supset holds(bel(\mu), result(learn(\mu), s))^{6} \qquad (8)$$

As in the case of update, the causal axiom only captures direct effects of learning; it does not say anything about whether other beliefs should be preserved, added, or dropped in the situation resulting from learning. These "indirect effects" will be captured, as before, by minimizing abnormalities from situation to situation, relative to a set of facts declared to be persistent by default. In this case, however, the facts of interest are facts about the agent's mental state. To specify which beliefs are treated as persistent in each situation, we use an additional (revision) persistence axiom (P_R), subject to similar conditions as in the case of update.[7] Such axiom might, for example, be a collection of sentences of the form

$$\forall s.\ holds(belonly(\psi), s) \supset \forall \sigma.\ persistent(bel(\sigma), s) \equiv Q(\sigma),$$

associating to each database ψ a set (defined through some formula Q) of belief fluents to be treated as persistent in situations in which the agent believes only ψ.

As it is readily seen from the definition of the *ab* predicate, minimizing *ab* in this context amounts to minimizing changes from belief to non-belief, and *vice versa.* However, minimizing both types of change simultaneously can be shown to lead to very counterintuitive results (for reasons similar to those in [9]). Hence we will require (P_R) to be consistent with the axiom:

$$\forall s, \sigma.\ persistent(bel(\sigma), s) \supset holds(bel(\sigma), s). \qquad (9)$$

In view of this axiom, the role of *persistent* in the context of revision is to select a set of worth-preserving or "basic" beliefs among the agent's beliefs, since minimizing *ab* will result in dropping as few as possible "persistent" beliefs.

Let then T_R be the theory obtained from T by adding all the axioms from this and the previous section, let $W_R(s)$ be obtained from T_R as $W(s)$ was from T (after suitable extension of $(N1)$ to deal with belief fluents), and let $Comp(T_R)$ be defined by replacing $W(S)$ by $W_R(S)$ in the definition of $Comp(T)$.

We now define a revision operator $\circ$, for any *satisfiable* propositional formulas μ and ψ as:

Definition 2 $\psi \circ \mu \models \sigma$ iff

$$Comp(T_R) \models holds(belonly(\psi), S_0) \supset holds(bel(\sigma), result(learn(\mu), S_0)).$$

[6] The antecedent of this axiom prevents learning unsatisfiable formulas.

[7] The introduction of belief fluents requires an additional condition: only propositional fluent terms (belief fluent terms, respectively) should occur as arguments of *holds* and *persistent* in (P_U) (respectively, (P_R)).

This definition exactly mirrors the definition of update in section 3. As in the case of update, the result of revision under this definition will depend on the choice of the persistence axiom. The ability to make this choice provides much flexibility in the definition of revision operators.

Example 2 If the persistence axiom P_R is such that $P_R \models holds(belonly(a \wedge b), S_0) \supset \forall\sigma.\ persistent(bel(\sigma), S_0) \equiv (\sigma = a \vee \sigma = b)$, then, under definition 2, we obtain $Cn((a \wedge b) \circ \neg a) = Cn(\neg a \wedge b)$. In contrast, if instead $P_R \models holds(belonly(a \wedge b), S_0) \supset \forall\sigma.\ persistent(bel(\sigma), S_0) \equiv (\sigma = a \vee \sigma = (a \supset b))$, then $Cn((a \wedge b) \circ \neg a) = Cn((a \supset b) \wedge \neg a) = Cn(\neg a)$. Intuitively, the first persistence axiom captures the case in which b is believed "independently" of a, while the second axiom corresponds to the situation in which the only *reason* to believe b is that a is also believed. ∎

6 The properties of belief revision

Definition 2, therefore, characterizes a family of revision operators, depending on the persistence axiom (P_R). Interesting subfamilies can be defined by placing additional restrictions on (P_R). We will consider four such restrictions.[8] The first one (TI_B, for "time independence") requires persistent beliefs to be independent of the situation:

Definition 3 ($\mathbf{TI}_B$) A theory of action satisfies the TI_B condition iff (P_R) is of the form $\forall s.\ P(s)$, where $P(s)$ contains no situation term other than s.

The next restriction, SDP_B, requires every belief *s*tate to uniquely *d*etermine the set of *p*ersistent beliefs in that state.

Definition 4 ($\mathbf{SDP}_B$) A theory of action satisfies the SDP_B condition iff for any situation S, any belief state R of S consistent with T_R, and any propositional fluent term σ, either $T_R, R \models persistent(bel(\sigma), S)$ or $T_R, R \models \neg persistent(bel(\sigma), S)$.

Nothing said so far prevents us from *e.g.* treating *none* of the agent's beliefs as worth preserving. The next condition, PAB, prevents this by requiring that the set of *p*ersistent beliefs associated with a given belief state provide an *a*xiomatization for the *b*eliefs held at that state.

Definition 5 (PAB) A theory of action satisfies the PAB condition iff for any situation S, any belief state R of S, and any set Ψ of propositional fluent terms: if the set $\{persistent(bel(\sigma), S) \mid \sigma \in \Psi\} \cup \{\neg persistent(bel(\sigma), S) \mid \sigma \in \mathcal{P} \setminus \Psi\}$ is consistent with $T_R \cup R$ then $Cn(\Psi) = T(R)$.

A theory of action satisfying PAB is such that the "persistent beliefs" provide an axiomatization of the agent's set of beliefs; thus this set can be seen as partitioned into a set of basic or persistent beliefs, on the one hand, and all other beliefs, on the other, where the latter are derivable from, and in that sense justified by the basic beliefs.

A final, somewhat unintuitive condition to consider is the following, which we call the TO (for "total order") condition.

[8] Some of the restrictions are phrased in terms of the theory of action as a whole, not just the persistence axiom. As the latter is however the only element of the theory of action that is left as a parameter, the conditions are properly seen as restricting the persistence axiom.

Definition 6 (TO) A theory of action satisfies the TO condition iff for any situation S, any belief state R of situation S, and any set of propositional fluent terms Ψ: if the set $\{persistent(bel(\sigma), S) \mid \sigma \in \Psi\} \cup \{\neg persistent(bel(\sigma), S) \mid \sigma \in \mathcal{P} \setminus \Psi\}$ is consistent with $T_R \cup R$, then $\Psi = \{\sigma_i \mid 1 \leq i \leq n\}$ for some n, where the sequence $\sigma_1, \ldots, \sigma_n$ is such that $\sigma_{i+1} \in Cn(\sigma_i)$ for $1 \leq i \leq n$.

This condition might appear obscure, so it is worthwhile pointing out that it could be equivalently formulated using a set of abnormality predicates and applying prioritized circumscription, so that the prioritization induces a total order on the persistent beliefs.[9] We choose not to introduce prioritization in order to keep our basic set up as simple as possible.

Example 3 Consider the instantiation of these conditions in which R is the belief state such that $T_R, R \models holds(belonly(a \wedge b), S_0)$. Both persistence axioms from example 2 satisfy PAB and SDP_B with respect to R, but violate TO. Consider, on the other hand, persistence axioms P_R^1, P_R^2, P_R^3 such that:

$$\begin{aligned}
P_R^1 &\models holds(belonly(a \wedge b), S_0) \supset ((\forall \sigma.\ persistent(bel(\sigma), S_0) \equiv (\sigma = a \vee \sigma = b)) \vee \\
&\qquad\qquad (\forall \sigma.\ persistent(bel(\sigma), S_0) \equiv (\sigma = a \vee \sigma = a \supset b))) \\
P_R^2 &\models holds(belonly(a \wedge b), S_0) \supset (\forall \sigma.\ persistent(bel(\sigma), S_0) \equiv \sigma = a) \\
P_R^3 &\models holds(belonly(a \wedge b), S_0) \supset (\forall \sigma.\ persistent(bel(\sigma), S_0) \equiv (\sigma = a \vee \sigma = a \wedge b))
\end{aligned}$$

If T_R has P_R^1 as persistence axiom (and assuming that T_R does not entail anything stronger than the sentence given for P_R^1) then T_R violates SDP_B, but satisfies PAB with respect to R; if T_R contains instead P_R^2, then it violates PAB, but satisfies SDP_B wrt. R; finally, if it contains P_R^3 then it satisfies PAB, SDP_B and TO wrt R. ∎

To recap, we have so far provided the properties of belief for a simple logically omniscient agent, axiomatized in an almost trivial way the effects of learning new information, and applied the same circumscription policy used for update. Coupled with the just listed conditions on the persistence axiom, the result is a very robust and expressive account of revision, as the following two theorems attest. Taken together, they provide a representation theorem for AGM revision operators, with the only proviso that our operators are only defined for satisfiable formulas.[10]

Theorem 2 If the operator $\circ$ of definition 2 is based on a theory of action satisfying the PAB, TI_B, SDP_B, and TO conditions, then $\circ$ satisfies (R1)–(R6).

Proof (Sketch.) In the presence of SDP_B, for every satisfiable ψ there is a unique set Ψ_ψ such that $T_R \models holds(belonly(\psi), S_0) \supset \forall \sigma.\ persistent(bel(\sigma), S_0) \equiv P_\psi(\sigma)$, where $P_\psi(\sigma)$ holds iff $\sigma \in \Psi_\psi$. It is possible to show then that $Cn(\psi \circ \mu) = \bigcap_{\Gamma \in \Psi_\psi \downarrow \neg\mu} Cn(\Gamma \cup \{\mu\})$, where $\Psi_\psi \downarrow \neg\mu$ is the set of maximal subsets of Ψ_ψ which do not entail $\neg\mu$. (See [21, 1] for similar

[9] This is analogous, therefore, to Nebel's syntax-based "unambiguous prioritized revision" [21]. Alternatively, $\{\sigma_1, \ldots, \sigma_n\}$ in the statement of the TO condition is such that $\{Mod(\sigma_1), \ldots, Mod(\sigma_n)\}$ is a "nested system of spheres" in the sense of Grove's [14] representation theorem for AGM revision.

[10] That is, we could add the requirement that ψ is also satisfiable to the antecedent of (R3). Complete proofs of the results in this paper, as well as a more general analysis of the properties of our definition of revision, can be found in [3].

constructions.) This in turn induces a partial preorder $\leq_\psi$ associated to each database ψ over the set $\mathcal{W}$ of interpretations of $\mathcal{L}$, such that the models of the revised database can be defined as in theorem 1 in section 2. This preorder becomes total in the presence of TO, and, because of PAB and TI_B, satisfies the faithfulness conditions of theorem 1. The conclusion follows then from theorem 1. ∎

Theorem 3 For every revision operator $\circ'$ satisfying (R1)-(R6), there exists an operator $\circ$ based on definition 2 and satisfying the TI_B, SDP_B, PAB, and TO conditions such that for every satisfiable ψ and μ:

$$Mod(\psi \circ \mu) = Mod(\psi \circ' \mu).$$

Proof (Sketch.) The key step of the proof is to provide an appropriate persistence axiom, which is done using a construction very similar to that used in the analogous theorem for KM update operators (theorem 2 in [6]). For every non empty set $W \subseteq \mathcal{W}$, choose a formula θ_W such that $Mod(\theta_W) = W$, and let Θ be the (finite) set of all θ_W. Because $\circ'$ satisfies (R1)–(R6), it can be defined as in theorem 1 in terms of a total preorder $\leq_\psi$ associated to each database ψ. For each $\theta_W \in \Theta$, let $\Psi_{\theta_W} = \{\theta_{W'} \mid W' = \{I \mid I \leq_{\theta_W} J\} \textit{ for some } J\}$, and let $P_W(s)$ be a finite axiomatization of "$persistent(bel(\sigma), s)$ iff $\sigma \in \Psi_{\theta_W}$." Define the operator $\circ$ based on the theory of action T_R with persistence axiom:

$$\forall s. \bigwedge_{\theta_W \in \Theta} (holds(belonly(\theta_W), s) \supset P_W(s)).$$

The TI_B and SDP_B conditions are satisfied by construction, PAB because of faithfulness, and TO because the preorders are total.

We show that $\circ$ is identical to $\circ'$. Write $I \sim_{\theta_W} J$ iff $I \leq_{\theta_W} J$ and $J \leq_{\theta_W} I$, and write $|I| = \{J \mid I \sim_{\theta_W} J\}$. $\sim_{\theta_W}$ is an equivalence relation, and since $\leq_{\theta_W}$ is a total preorder, the equivalence classes can be linearly ordered by: $|I| <^\star_{\theta_W} |J|$ iff $I \leq_{\theta_W} J$ and $J \not\leq_{\theta_W} I$. Let $|I_1|, \ldots, |I_n|$ be the set of all equivalence classes, and suppose wlog that $|I_i| <^\star_{\theta_W} |I_j|$ iff $i < j$. Then Ψ_{θ_W} can be written as $\{\sigma_1, \ldots, \sigma_n\}$, where $Mod(\sigma_i) = \bigcup_{1 \leq j \leq i} |I_j|$. Note that for every μ, $\Psi_{\theta_W} \downarrow \neg\mu$ is always the singleton whose unique element is the set $\Gamma_k = \{\sigma_k, \sigma_{k+1}, \ldots, \sigma_n\}$, where k is the minimum integer for which $\sigma_k \wedge \mu$ is satisfiable. Also, $Cn(\Gamma_k) = Cn(\sigma_k)$.

From the proof of theorem 2, the revision operator determined by this persistence axiom satisfies $Cn(\theta_W \circ \mu) = \bigcap_{\Gamma \in \Psi_{\theta_W} \downarrow \neg\mu} Cn(\Gamma \cup \{\mu\}) = Cn(\Gamma_k \cup \{\mu\}) = Cn(\sigma_k \wedge \mu)$. Since $(\bigcup_{1 \leq j < k} |I_j|) \cap Mod(\mu) = \emptyset$, and $(\bigcup_{1 \leq j \leq k} |I_j|) \cap Mod(\mu) \neq \emptyset$, it follows that $Mod(\theta_W \circ \mu) = |I_k| \cap Mod(\mu)$, where $|I_k|$ is the unique minimum element under $<^\star_{\theta_W}$ such that some model in $|I_k|$ satisfies μ. The conclusion follows from this and theorem 1. ∎

Similar representation results can be obtained for operators defined as in theorem 1 in terms of partial rather than total preorders, by removing the TO condition; the proofs of these more general results use identical constructions. Partial preorder revision operators were characterized by Katsuno and Mendelzon in [16]. Some implications for non-monotonic reasoning of this ability to capture partial preorder operators are studied in a simplified setting in [4].

7 A unified view of revision and update

The results of the previous section show that our approach does indeed exhibit properties traditionally associated with revision, under suitable restrictions on the persistence axiom, and that it is expressive enough to capture a wide family of revision operators. These results are not yet, however, completely satisfactory, in that revision is characterized by translating back and forth between the propositional language and situation calculus. Our goal is to be able to perform all reasoning about revision and update within the theory of action.

There are two aspects to consider. First, our definition of revision leaves open whether the agent *believes only* in the revised database. But surely, if the revised database is intended to represent the set of beliefs of the agent after revision, it should not have a different status than the initial database within the theory of action. Second, as long as we use definitions 1 and 2, it does not matter whether *learn* has any effect on propositional facts, or the actions A_μ^S on epistemic facts, but this is unacceptable given the aforementioned goal.

As for the first point, our theory of action, as it stands, does *not* yield such desirable consequence: There are, in general, circumscription-minimal models of the theory of action which determine different belief states in the situation $result(learn(\mu), S_0)$. In order to fix this, we need to minimize not only the persistent beliefs dropped by revision, but also the beliefs added. For this purpose, we introduce a new abnormality predicate $ab_I(\sigma, a, s)$ as an abbreviation for:

$$\neg holds(bel(\sigma), s) \wedge holds(bel(\sigma), result(a, s)),$$

and minimize its extension with lower priority than ab. Formally, let $Comp_I(T_R)$ stand for the union over all closed terms A and S of the appropriate sorts of:

$$Circum(W_R(S); ab(\sigma, a/A, s/S) > ab_I(\sigma, a/A, s/S); holds, persistent).$$

It can be shown that substituting $Comp_I(T_R)$ for $Comp(T_R)$ in definition 2 determines an identical revision operator, and thus this theory preserves the results of the previous section. Furthermore, we can show:

Theorem 4 A revision operator $\circ$ satisfies (R1)–(R6) *if and only if* $\circ$ can be defined in terms of a theory of action satisfying the SDP$_B$, PAB, and TO conditions, such that for every satisfiable ψ and μ:

$$\psi \circ \mu \equiv \psi' \text{ if and only if}$$
$$Comp_I(T_R) \models holds(belonly(\psi), S_0) \supset holds(belonly(\psi'), result(learn(\mu), S_0)).$$

Proof (Sketch.) Suppose $\circ'$ satisfies (R1)–(R6). Let T_R be the theory of action defined in the proof of theorem 3 from $\circ'$, let $\psi = \theta_W$, and let Ψ_ψ, Γ_k, and σ_k be as before. Consider any model M of $Comp_I(T_R)$ such that $M \models holds(belonly(\psi), S_0)$. It suffices to show that (*) $T(M_B(result(learn(\mu), S_0))) = Cn(\sigma_k \wedge \mu)$ for any such M, since this means that $Comp_I(T_R) \models holds(belonly(\psi), S_0) \supset holds(belonly(\sigma_k \wedge \mu), result(learn(\mu), S_0))$, and we know from the proof of theorem 3 that $Cn(\sigma_k \wedge \mu) = Cn(\psi \circ' \mu)$.

By construction of T_R and the hypothesis on M, M satisfies $persistent(\sigma)$ iff $\sigma \in \Psi_\psi$. Thus, the right to left inclusion of (*) follows from the logical closure of belief states, together with the fact that $T(M_B(result(learn(\mu), S_0)))$ must contain both Γ_k, in order to have a minimal extension of $ab(\sigma, a/learn(\mu), s/S_0)$, and μ, to satisfy the causal axiom for $learn(\mu)$. The converse inclusion follows from the minimization of ab_I.

For the other direction of the theorem, note that the same reasoning can be applied to any theory of action satisfying SDP$_B$, PBA, and TO to show that the revision operator it defines satisfies the conditions of the theorem. ∎

Thus, for AGM operators there is no need to define "the revised database" outside the language; rather, the revised database is simply whatever the agent believes only in, in the situation resulting from learning the new information. One consequence of this is that the theory of action given in the proof of theorem 3 captures the *temporal evolution* of the agent's beliefs under any given AGM operator, by virtue of TI_B. The persistence axiom provided associates a unique set of persistent beliefs to each belief state, independent of the situation, and it follows from the previous theorem that the revised database corresponds to a unique belief state consistent with the circumscription. Further revising this revised database will yield identical results as if it were the initial database, since identical beliefs will be persistent in both situations.

We remark that this result does not seem to hold for non-AGM operators, in particular for partial-preorder revision operators, or in the absence of TO (modulo variations in the persistence axiom that have no effect on revision). In this case, the theory of action will in some cases only entail that there are several alternative "revised databases" in which the agent could believe only, but the theory cannot tell which. This does not mean that we need to move outside the theory of action to reason about revision; rather, it simply introduces a degree of non-determinism in revision, analogous to the proposal of [8]. In fact, the connection between deterministic revision and AGM is very tight in our framework: theorem 4 tells us that any AGM operator can be captured in a theory of action that uniquely determines the revised database, and it can be shown that any theory of action that yields such deterministic behavior satisfies the AGM postulates.

As for the second issue raised at the beginning of this section, the interaction between both types of action can easily be reduced to none by adding the axioms:

$$\begin{gathered}\forall s, a, \sigma.\ (\forall \mu.\ a \neq learn(\mu)) \supset (holds(bel(\sigma), s) \equiv holds(bel(\sigma), result(a, s)))\\ \forall s, \sigma_p, \mu.\ holds(\sigma_p, s) \equiv holds(\sigma_p, result(learn(\mu), s)),\end{gathered}$$

where the subscript in σ_p is used to indicate that the variable ranges over *propositional* fluents. In this way, we ensure that revision (*i.e.* update of mental facts) and update (of propositional facts) are independent within the theory of action.

Let us finish this section by considering briefly our claim that our results provide a *reduction* of revision to update, as revision is defined in terms of update of the mental state. To sharpen this intuition, suppose we extend the initial propositional vocabulary $\mathcal{P}_{\mathcal{L}}$ by adding new symbols $belonly(\varphi)$ and $bel(\varphi)$ for every satisfiable $\varphi \in \mathcal{L}$. Let $\mathcal{P}_{\mathcal{L}}^{\mathcal{B}}$ be the extended vocabulary, and $\mathcal{L}_{\mathcal{B}}$ the propositional language obtained from $\mathcal{P}_{\mathcal{L}}^{\mathcal{B}}$. When the theory of action for update of section 3 is formulated by taking $\mathcal{L}_{\mathcal{B}}$ instead of $\mathcal{L}$ as initial propositional language, the definition 1 of *update* can be specialized to:

$$\begin{gathered}belonly(\psi) \diamond bel(\mu) \models bel(\sigma) \quad \text{ if and only if}\\ Comp(T) \models holds(belonly(\psi), S_0) \supset holds(bel(\sigma), result(A^{S_0}_{bel(\mu)}, S_0)).\end{gathered} \tag{10}$$

Since we only consider satisfiable μ, the action $A^{S_0}_{bel(\mu)}$ is essentially the action $learn(\mu)$ by another name, and thus the right hand side of this equation is at least superficially identical to the definition 2 of *revision*. To ensure that this is a true identity, we simply need to provide axioms for the new symbols in $\mathcal{P}_{\mathcal{L}}^{\mathcal{B}} \setminus \mathcal{P}_{\mathcal{B}}$ analogous to those of the mental situation calculus.[11] With this addition, revision defined in terms of the right hand side of expression 10 has the same effects as definition 2, and thus we have:

$$\psi \circ \mu \models \sigma \text{ iff } \mathit{onlybel}(\psi) \diamond \mathit{bel}(\mu) \models \mathit{bel}(\sigma).$$

8 Discussion

In [6, 7] we reduced belief update to theories of action. In this paper, we have reduced belief revision to theories of action through a definition of revision as update applied to the mental state. The resulting framework allows for a unified treatment of revision and update within a single temporal formalism, and provides a natural and intuitive way to axiomatically define update and revision operators in terms of default persistence of facts (mental or otherwise).

As we have seen, this is a very expressive framework, in which we can capture a wide variety of revision operators even in the rather restrictive setting imposed by the various conditions on the persistence axiom. Given the well known connection between belief revision and non monotonic "preferential logics" [23], furthermore, it is easy to see that we can capture any non-monotonic preferential relation by a simple application of the (omitted) results for partial preorder revision operators. Specifically, it can be shown that for any preferential consequence relation $\models_{\leq}$ defined by a partial preorder $\leq$ over $\mathcal{W}$, there exists a theory of action such that: $\mu \models_{\leq} \sigma$ if and only if $Comp(T_R) \models holds(belonly(\psi_T), S_0) \supset holds(bel(\sigma), result(learn(\mu), S_0))$, where ψ_T is such that $Mod(\psi_T) = Min(\mathcal{W}, \leq)$. This topic is addressed in more detail in [4] in a simplified (atemporal) setting. A simple mapping between the framework of the present paper and that of [4] derives from the fact that the persistence axiom here and the basic beliefs function there play identical roles.

The mental situation calculus we have presented is restricted in that it allows no nested beliefs. Lifting this restriction would require modifying the axioms governing *belonly*. In the presence of positive and negative introspection, it is easily shown that belief states are stable sets in the sense of [20], and thus it appears possible to axiomatically capture a notion of "believing only" which exactly matches the concept of "all I know" from [15]. The details however remain to be investigated.

As mentioned in the introduction, Grahne *et. al.* [13] were the first to reduce belief revision to update of the mental state. Specifically, they present a logic with operators for update $\diamond$, belief B, and "believe only" O, and show that for any AGM revision operator $\circ$ there exists a model M° of this logic such that

$$\psi \circ \mu \models \sigma \text{ iff } M^\circ \models O\psi \diamond B\mu \supset B\sigma.$$

This result is the analogous to our theorem 3, but they provide no analogous to theorem 2, and thus fall short of providing a representation result. They do not provide either a way to axiomatically define revision operators.

Most importantly, whereas they take AGM as given, our construction can be used to shed light on the nature of revision and to identify conditions under which various kinds of revision operators are acceptable, in a way similar to our analysis of KM

[11] In the framework of the theory of action for update of section 3, these axioms would have the role of *constraints* or *protected formulas*, see [5, 7].

update semantics in [7]. For example, it is easy to see that many natural revision operators definable in our construction will fail to satisfy the TO condition. Similarly, the TI condition might be too restrictive in that it fails to recognize that the same set of beliefs might have to be revised differently in different situations, as advocates of the foundational theory of belief revision insist (*e.g.* [21]). Indeed, as the example of section 5 suggests, our approach is very much foundational in style, though this does not prevent us from capturing the whole family of AGM operators. Unlike other foundational approaches, however, we keep the specification of the set of persistent or basic beliefs separate from the axiomatization of the database, and thus our approach avoids the undesirable dependence on syntax of other proposals.

It is more difficult to imagine circumstances in which it would be reasonable for the PAB condition to fail. A theory of action with a persistence axiom such as P_R^2, from example 3, entails, for instance, that $(a \wedge b) \circ \mathit{true} \equiv a$, thus "arbitrarily" dropping the belief in b. If our preliminary analysis of introspection is correct, however, failure of PAB can provide a simple way out of the paradoxes associated to revision of introspective theories [10, 13], for example by treating nested beliefs as non persistent. This persistence policy is justified in so far as belief states become stable sets in the presence of introspection, and are thus uniquely determined by their propositional subset.

Finally, failure of SDP_B corresponds to a situation in which there is uncertainty about which beliefs should be persistent; as it were, several revision operators coexist within the theory of action (one for each set of persistent beliefs that can be associated with a given belief state in a way consistent with the persistence axiom). The situation here is analogous to that of conditional logics for revision such as that of [2]: a theory in these logics will in general be incomplete, *i.e.* have multiple "revision models," each defining a different revision operator. The conclusions sanctioned by incomplete conditional theories can obviously be much weaker than those sanctioned by theories with a single model. In order to avoid this problem in conditional logic without recurring to unwieldy complete theories, one must either semantically provide a unique revision model, or appeal to further semantic constraints such as those that allow [12] to single out a unique revision model from an incomplete set of conditionals. In our framework, we achieve essentially the same effect *axiomatically*, by ensuring that the SDP_B condition is satisfied. And because all AGM operators can be captured in our construction, any formal device used to single out a unique revision model from a set of conditionals can also be used as a preliminary stage in the axiomatization of the theory of action.

Acknowledgements

Thanks to John McCarthy and Ray Reiter for useful discussions on the subject of this paper.

References

[1] Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet functions for contraction and revision. *Journal of Symbolic Logic*, 50:510–530, 1985.

[2] Craig Boutillier. Normative, subjunctive and autoepistemic defaults: Adopting the Ramsey test. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, 1992.

[3] Alvaro del Val. *Belief Revision and Update*. PhD thesis, Stanford University, 1993.

[4] Alvaro del Val. On the relation between the coherence and foundations theories of belief revision. In *Proceedings of the Twelfth Conference of the AAAI*, 1994.

[5] Alvaro del Val and Yoav Shoham. Deriving properties of belief update from theories of action. In *Proceedings of the Tenth Conference of the AAAI*, 1992.

[6] Alvaro del Val and Yoav Shoham. Deriving properties of belief update from theories of action (II). In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, 1993.

[7] Alvaro del Val and Yoav Shoham. Deriving properties of belief update from theories of action. *Journal of Logic, Language, and Information*, 1994. To appear. Also as Technical Report STAN-CS-TR-94-1501, Computer Science Department, Stanford University.

[8] Ronald Fagin, Gabriel M. Kuper, Jeffrey D. Ullman, and Moshe Y. Vardi. Updating logical databases. *Advances in Computing Research*, 3, 1986.

[9] Ronald Fagin, Jeffrey D. Ullman, and Moshe Y. Vardi. On the semantics of updates in databases. In *Proceedings of the Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 1983.

[10] André Fuhrmann. Reflective modalities and theory change. *Synthese*, 81:115–134, 1989.

[11] Peter Gärdenfors. *Knowledge in Flux*. The MIT Press, 1988.

[12] Moisés Goldszmidt and Judea Pearl. Rank-based systems: A simple approach to belief revision, belief update and reasoning about evidence and actions. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, 1992.

[13] Gösta Grahne, Alberto Mendelzon, and Raymond Reiter. On the semantics of belief revision systems. In *Proceedings of the Third Conference on Theoretical Aspects of Reasoning about Knowledge*, 1992.

[14] Adam Grove. Two modelings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.

[15] Joseph Halpern and Yoram Moses. Towards a theory of knowledge and ignorance: Preliminary report. In K. Apt, editor, *Logics and Models of Concurrent Systems*. Springer Verlag, 1985.

[16] Hirofumi Katsuno and Alberto O. Mendelzon. On the difference between updating a knowledge database and revising it. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, 1991.

[17] Hirofumi Katsuno and Alberto O. Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52:263–294, 1991.

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

[19] Fanghzen Lin and Yoav Shoham. Provably correct theories of action (preliminary report). In *Proceedings of the Ninth Conference of the AAAI*, 1991.

[20] R. C. Moore. Semantical considerations on non-monotonic logic. *Artificial Intelligence*, 25:75–94, 1985.

[21] Bernhard Nebel. Belief revision and default reasoning: Syntax-based approaches. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, 1991.

[22] Raymond Reiter. On specifying database updates. Technical Report KRR-TR-92-5, Computer Science Department, University of Toronto, 1992.

[23] Yoav Shoham. A semantical approach to non-monotonic logics. In Matthew L. Ginsberg, editor, *Readings in Non-Monotonic Reasoning*. Morgan Kaufmann, 1987.

[24] Marianne Winslett. Reasoning about action using a possible models approach. In *Proceedings of the Seventh Conference of the AAAI*, 1988.