Conditional Logics of Belief Change*

Nir Friedman
Stanford University
Dept. of Computer Science
Stanford, CA 94305-2140
nir@cs.stanford.edu

Joseph Y. Halpern
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120–6099
Stalnaker'shalpern@almaden.ibm.com

Abstract

The study of *belief change* has been an active area in philosophy and AI. In recent years two special cases of belief change, *belief revision* and *belief update*, have been studied in detail. Belief revision and update are clearly not the only possible notions of belief change. In this paper we investigate properties of a range of possible belief change operations. We start with an abstract notion of a *belief change system* and provide a logical language that describes belief change in such systems. We then consider several reasonable properties one can impose on such systems and characterize them axiomatically. We show that both belief revision and update fit into our classification. As a consequence, we get both a semantic and an axiomatic (proof-theoretic) characterization of belief revision and update (as well as some belief change operations that generalize them), in one natural framework.

Introduction

The study of *belief change* has been an active area in philosophy and in artificial intelligence (Gärdenfors 1988; Katsuno & Mendelzon 1991). The focus of this research is to understand how an agent should change his beliefs as a result of getting new information. In the literature, two types of belief change operation have been studied in detail: *belief revision* (Alchourrón, Gärdenfors, & Makinson 1985; Gärdenfors 1988) and *belief update* (Katsuno & Mendelzon 1991). Belief revision and update are two cases of belief change, but clearly not the only ones. In this paper we investigate properties of a range of possible belief change operations.

We start with the notion of a *belief change system* (BCS). A BCS contains three components: The set of possible *epistemic states* that the agent can be in, a *belief assignment* that maps each epistemic state to a set of beliefs, and a *transition function* that determines how the agent changes epistemic states as a result of learning new information. We assume some logical language $\mathcal{L}$ that describes the agent's world, and assume that the agent's beliefs are closed under deduction in $\mathcal{L}$. Thus, the belief assignment maps each state to a deductively closed set of formulas in $\mathcal{L}$. We make the assumption (which is standard in the literature) that the agent learns a formula in $\mathcal{L}$, i.e., that events that cause the agent to change epistemic state can be described by formulas. Thus, the transition function takes a formula in $\mathcal{L}$ and an epistemic state to another epistemic state.

The notion of a BCS is quite general. It is easy to show that any operator satisfying the axioms of belief revision or update can be represented as a BCS. However, by starting at this level of abstraction, we can more easily investigate the general properties of belief change. We do so by considering a language that reasons about the belief change in a BCS. The language contains two modal operators: a unary modal operator B for belief and a binary modal operator $>$ to represent change, where, as usual, $B\varphi$ should be read "the agent believes φ", while $\varphi > \psi$ should be read "after learning φ, the agent will be in an epistemic state satisfying ψ". We show that the language is expressive enough to capture the belief change process. More precisely, the set of (modal) formulas holding at a state uniquely determines the agent's beliefs after any sequence of events. Thus, it is possible to describe the agent's belief change behavior by specifying what formulas in the extended language of conditionals hold at the agent's initial state. We also characterize the class of all BCS's axiomatically in this language.

We then investigate an important class of BCS's that we call *preferential BCS's*. This class can be viewed as an abstraction of the semantic models considered in papers such as (Grove 1988; Katsuno & Mendelzon 1991; Boutilier 1992; Katsuno & Satoh 1991). Roughly speaking, a preferential BCS is a BCS where an epistemic state can be identified with a set of possible *worlds*, where a world is a complete truth assignment to $\mathcal{L}$, together with a *preference ordering* on worlds. An agent believes φ in epistemic state s exactly if φ is true in all the worlds considered possible at s, and the agent believes ψ after learning φ in epistemic state s exactly if ψ is true in all the minimal worlds that satisfy φ (according to the preference ordering at s).[1]

*Work supported in part by the Air Force Office of Scientific Research (AFSC), under Contract F49620-91-C-0080.

[1] We note that there is some confusion in the literature between the Ramsey conditional (i.e., $>$) and preference conditional that describes the agent preferences (see (Boutilier 1992) for example). There is a strong connection between the two in preferential BCS's, but even in that context they have different properties. We think it is important to distinguish them. (See also (Friedman & Halpern

The class of preferential BCS's includes, in a precise sense, the class of operators for belief revision and the class of operators for belief update, so it can be viewed as a generalization of these notions. We consider a number of reasonable properties that one can impose on preferential BCS's, and characterize them axiomatically. It turns out that both belief revision and update can be characterized in terms of these properties. As a consequence, we get both a semantic and an axiomatic (proof-theoretic) characterization of belief revision and update (as well as some belief change operations that generalize them), in one natural framework.

There are some similarities between our work and others that have appeared in the literature. In particular, our language and its semantics bear some similarities to others that have been considered in the literature (for example, in papers such as (Gärdenfors 1978; 1986; Grahne 1991; Lewis 1973; Stalnaker 1968; Wobcke 1992)), and our notion of a BCS is very similar to Gärdenfors' belief revision systems (Gärdenfors 1978; 1986; 1988). However, there are some significant differences as well, both philosophical and technical. We discuss these in more detail in the next section. These differences allow us to avoid Gärdenfors' triviality result 1986, which essentially says that there are no interesting BCS's that satisfy the AGM postulates (Alchourrón, Gärdenfors, & Makinson 1985).

Belief change systems

A belief change system describes the possible states the agent might be in, the beliefs of the agent in each state, and how the agent changes state when receiving new information. We assume beliefs are described in some logical language $\mathcal{L}$ with a consequence relation $\models_{\mathcal{L}}$, which contains the usual truth-functional propositional connectives and satisfies the deduction theorem. We define a *belief change system* as a tuple $M = \langle S, \rho, \tau \rangle$, where S is a set of *states*, ρ is a *belief assignment* that maps a state $s \in S$ to a set of sentences $\rho(s)$ that is deductively closed (with respect to $\models_{\mathcal{L}}$), and τ is a function that maps a state $s \in S$ and sentence $\varphi \in \mathcal{L}$ to a new state $\tau(s, \varphi) \in S$. We differ from some work in the area of conditional logic (for example, (Grahne 1991; Lewis 1973; Stalnaker 1968)) in taking epistemic states rather than worlds as our primitive objects, while we differ from other work (for example, (Gärdenfors 1978; 1986)) by not identifying epistemic states with belief sets. In our view, while the $\mathcal{L}$-beliefs of an agent are certainly an important part of his epistemic state, they do not in general characterize it. Notice that because we do not identify belief sets with epistemic states, the function τ may behave differently at two epistemic states that agree on the beliefs in $\mathcal{L}$.[2]

A BCS describes how the agent's beliefs about the world change. We use a logical language we call $\mathcal{L}^{>}$ to reason about BCS's. As we said in the introduction, the language $\mathcal{L}^{>}$ augments $\mathcal{L}$ with a unary modal operator B and a binary modal operator $>$ to capture belief change. Formally, we take $\mathcal{L}^{>}$ be the least set of formulas such that if $\varphi \in \mathcal{L}$ and $\psi, \psi' \in \mathcal{L}^{>}$ then $B\varphi$, $B\psi$, $\neg\psi$, $\psi \wedge \psi'$, and $\varphi{>}\psi$ are in $\mathcal{L}^{>}$. A number of observations should be made with regard to the choice of language. First observe that $\mathcal{L}$ and $\mathcal{L}^{>}$ are disjoint languages. The language $\mathcal{L}$ consists intuitively of objective formulas (talking about the world), while $\mathcal{L}^{>}$ consists of subjective formulas (talking about the agent's epistemic state). Thus, the formula $\varphi \in \mathcal{L}$ is not in $\mathcal{L}^{>}$, although $B\varphi$ is. We view the states in a BCS as epistemic states, and thus use the language $\mathcal{L}^{>}$ for reasoning about BCS's. There is no notion of an "actual world" in a BCS (as there is, by way of contrast, in a Kripke structure), so we have no way in our semantic model to evaluate whether a formula $\varphi \in \mathcal{L}$ is true. Of course, we could augment BCS's in a way that would let us do this, but there is no need for the purposes of this paper. (In fact, this is done in (Friedman & Halpern 1994a; 1994b), where we examine a broader framework that models both the agent and world and allows us to evaluate objective and subjective formulas.) We could have also interpreted a formula $\varphi \in \mathcal{L}$ to mean "the agent believes φ" (as in (Gärdenfors 1978)), but it turns out to be technically more convenient to add the B operator, since it lets us distinguish between the agent believing $\neg\varphi$ and the agent not believing φ.

Another significant difference between our language and other languages considered in the literature for reasoning about belief change (for example, (Gärdenfors 1978; 1986; Grahne 1991; Wobcke 1992)) is that on the left-hand side of $>$, we only allow formulas in $\mathcal{L}$ rather than arbitrary formulas in $\mathcal{L}^{>}$. For example, $p{>}(q{>}Br)$ is in $\mathcal{L}^{>}$, but $(p{>}Bq){>}Br$ is not. Recall that the formula on the left-hand side of $>$ represents something that the agent could learn. It is not clear how an agent could come to learn a formula like $p{>}Bq$. Our intuition is that an agent learns about the external world, as described by $\mathcal{L}$, and not facts about the belief change process itself. Our language $\mathcal{L}^{>}$ is used to reason about the belief change process.[3]

We now assign truth values to formulas in $\mathcal{L}^{>}$. We write $(M, s) \models \varphi$ if φ holds in epistemic state s in the system M. We interpret $(M, s) \models \varphi$ to mean that the agent believes φ in epistemic state s. Since we take our agents to be introspective, we would expect that if $(M, s) \models \varphi$, then

1994b) for a discussion of this issue.)

[2] A similar distinction between epistemic states and belief sets can be found in (Rott 1990; Boutilier 1992). See also (Friedman & Halpern 1994b).

[3] Our position in this respect bears some similarity to that of (Levi 1988). However, Levi seems to be arguing against the agent learning *any* modal formula, while our quarrel is only with the agent learning modal formulas of the form $\varphi{>}\psi$. The formulas in $\mathcal{L}$ may be modal. It may seem to the reader familiar with the recent work of (Boutilier & Goldszmidt 1993) that they are dealing with precisely the problem of revising beliefs by formulas of the form $\varphi{>}\psi$. However, their interpretation of a formula such as $\varphi{>}\psi$ is "normally if φ is true then ψ is true". Although there is a relationship between the two interpretations of $>$ in the preferential BCS's we consider in the next section, they are distinct, and should be represented by two distinct modal operators. We would have no problem with normality formulas of the form considered by Boutilier and Goldszmidt appearing in $\mathcal{L}$, and thus on the left-hand side of $>$.

$(M, s) \models B\varphi$. Our semantics enforces this expectation. We have already given the intuition for $>$, namely, that $\varphi > \psi$ should hold precisely if ψ holds in the epistemic state that results after updating by ψ. Our semantics enforces this as well.

- $(M, s) \models B\varphi$ if $\varphi \in \rho(s)$ for $\varphi \in \mathcal{L}$
- $(M, s) \models B\psi$ if $(M, s) \models \psi$ for $\psi \in \mathcal{L}^>$
- $(M, s) \models \neg\varphi$ if $(M, s) \not\models \varphi$.
- $(M, s) \models \varphi \wedge \psi$ if $(M, s) \models \varphi$ and $(M, s) \models \psi$
- $(M, s) \models \varphi > \psi$ if $(M, \tau(s, \varphi)) \models \psi$.

Because Gärdenfors (Gärdenfors 1978; 1986) identifies each state, not with a set of beliefs in $\mathcal{L}$, but with a set of beliefs in $\mathcal{L}^>$, he cannot define $\models$ inductively as we do here. Rather, he puts constraints on the transition function τ so that $>$ satisfies the *Ramsey test*; i.e., he requires that $\varphi > \psi$ holds at epistemic state s if and only if ψ holds at $\tau(s, \varphi)$.

Notice that this condition amounts to the agent having positive introspection about his belief change protocol. One can imagine an agent who is unaware of his belief change protocol, so that although it is true that the agent will believe ψ after learning φ in epistemic state s, the agent is not aware of this, so that $\varphi > \psi$ does not hold at s. At the other extreme is an agent who is completely aware of his belief change protocol, so that if learning φ in state s results in the agent's believing ψ, then $\varphi > \psi$ holds at s, otherwise $\neg(\varphi > \psi)$ holds. We are implicitly assuming such complete introspective power on the part of the agent: Our semantics guarantees that one of $\varphi > \psi$ or $\neg(\varphi > \psi)$ must hold at every state s. Gärdenfors' semantics enforces positive introspection, but not complete introspection. As a result, his epistemic states may be incomplete with respect to conditional formulas; it is possible that neither $\varphi > \psi$ nor $\neg(\varphi > \psi)$ holds at a given epistemic state. It is not clear what the rationale is for this intermediate position.

Given a state s we define $\text{Bel}(s)$ to be the (extended) beliefs of the agent at s:

$$\text{Bel}(s) = \{\varphi \in \mathcal{L}^> | (M, s) \models \varphi\}$$

Intuitively, $\text{Bel}(s)$ describes the agent's beliefs when he is in state s, and how these belief change after each possible sequence of observations. This intuition is justified, since $(M, s) \models \varphi$ if and only if $(M, s) \models B\varphi$ for any $\varphi \in \mathcal{L}^>$.

It is easy to see that given $\text{Bel}(s)$ we can reconstruct $\rho(s)$, i.e., for $\varphi \in \mathcal{L}$, $\varphi \in \rho(s)$ if and only if $B\varphi \in \text{Bel}(s)$. Indeed, as the following results show, $\text{Bel}(s)$ completely characterizes the belief change process at s.

Proposition 1: *Let M be a BCS, s a state in M, and $\varphi \in \mathcal{L}$ a formula. Then $Bel(\tau(s, \varphi)) = \{\psi | \varphi > \psi \in Bel(s)\}$.*

Applying this result repeatedly we get

Corollary 2 : *Let M, M' be BCS structures, and let s, s' be states in M and M', respectively. $Bel(s) = Bel(s')$ if and only if for any sequence of observations $\varphi_1, \ldots, \varphi_n$ it is the case that $\rho(\tau(\ldots\tau(s, \varphi_1)\ldots, \varphi_n)) = \rho'(\tau'(\ldots\tau(s', \varphi_1)\ldots, \varphi_n))$.*

This implies that $\text{Bel}(s) = \text{Bel}(s')$ if and only if s and s' cannot be distinguished by the belief change process. Thus, the language $\mathcal{L}^>$ is appropriate for describing the belief change process; it captures all the details of the process, but no unnecessary details.

We next turn our attention to the problem of axiomatizing belief change. Given a BCS M, we say that $\varphi \in \mathcal{L}^>$ is *valid* in M, denoted $M \models \varphi$, if $(M, s) \models \varphi$ for every s. Let $\mathcal{M}$ be the class of all BCS structures, and let $\mathcal{N}$ be a subclass of $\mathcal{M}$. We say that $\varphi \in \mathcal{L}^>$ is *valid with respect to $\mathcal{N}$* if it is valid in all $M \in \mathcal{N}$. An axiom system is *sound* and *complete* for $\mathcal{L}^>$ with respect to $\mathcal{N}$ if φ is provable if and only if it is valid in $\mathcal{N}$. We are interested in characterizing various subclasses of $\mathcal{M}$ axiomatically. We start with $\mathcal{M}$ itself. Consider the following axiom system, which we call AX. In all the axioms and inference rules of AX, the formulas range over allowable formulas in $\mathcal{L}^>$ (so that when we write $\varphi > \psi$, we are implicitly assuming that $\varphi \in \mathcal{L}$ and that $\psi \in \mathcal{L}^>$):

B1. All substitution instances of propositional tautologies

B2. $B\varphi$, if $\varphi \in \mathcal{L}$ is $\mathcal{L}$-valid

B3. $B\varphi \wedge B(\varphi \Rightarrow \psi) \Rightarrow B\psi$

B4. $\varphi \Rightarrow B\varphi$

B5. $B\varphi \Rightarrow \neg B\neg\varphi$ for $\varphi \in \mathcal{L}^>$

B6. $\varphi > \mathit{true}_{\mathcal{L}^>}$

B7. $\varphi > \psi_1 \wedge \varphi > (\psi_1 \Rightarrow \psi_2) \Rightarrow \varphi > \psi_2$

B8. $\neg(\varphi > \psi) \equiv \varphi > \neg\psi$

RB1. From φ and $\varphi \Rightarrow \psi$ infer ψ

RB2. From $\psi_1 \Rightarrow \psi_2$ infer $\varphi > \psi_1 \Rightarrow \varphi > \psi_2$

Axioms B3–B5 capture the standard properties of introspective belief. Notice that B4 relies on the fact that all formulas are taken to be subjective, that is, statements about the agent's beliefs. Although it may appear that B2 should follow from B1 and B4, it does not, since $\varphi \Rightarrow B\varphi$ is not an instance of B4 if $\varphi \in \mathcal{L}$ (since it is not a formula in $\mathcal{L}^>$). B5 states that the agent's beliefs about subjective formulas are always consistent. This follows naturally from our semantics. For any $\varphi \in \mathcal{L}^>$, either φ or $\neg\varphi$ is true at a state s, and thus only one of them will be believed. It is important to note that this axiom does not force the agent's beliefs about the world to be consistent. More precisely, let $\mathit{false}_{\mathcal{L}}$ be $p \wedge \neg p$ for some $p \in \mathcal{L}$, and let $\mathit{false}_{\mathcal{L}^>}$ be $Bp \wedge \neg Bp$. Clearly, $\mathit{false}_{\mathcal{L}} \in \mathcal{L}$ and $\mathit{false}_{\mathcal{L}^>} \in \mathcal{L}^>$. Axiom B5 states that $\neg B\mathit{false}_{\mathcal{L}^>}$ is valid, but it does *not* imply that $\neg B\mathit{false}_{\mathcal{L}}$ is valid. In fact, $B\mathit{false}_{\mathcal{L}}$ is satisfiable in our semantics. (Of course, the formula $\mathit{true}_{\mathcal{L}^>}$ used in B6 is the valid $\mathcal{L}^>$ formula $\neg\mathit{false}_{\mathcal{L}^>}$; we take $\mathit{true}_{\mathcal{L}}$ to be $\neg\mathit{false}_{\mathcal{L}}$.) B8 follows from the fact that we have assumed the transition function τ is deterministic. Axiom B8 is known as *law of conditional excluded middle* (Stalnaker 1968). This axiom has been controversial in the literature (Lewis 1973; Harper, Stalnaker, & Pearce 1981). It does not seem as problematic here, since we are applying it to only subjective formulas, rather than objective formulas.

The following result shows that AX does indeed characterize belief change.

Theorem 3: *AX is a sound and complete axiomatization of $\mathcal{L}^>$ with respect to $\mathcal{M}$.*

It is interesting to compare our axiomatization with the system CM discussed in (Gärdenfors 1978). All of his axioms are sound in our framework. We have some extra axioms due to the fact that our language includes a B operator, but this could be easily added to Gärdenfors' framework as well. A more interesting difference is our axiom B8, which does not hold in CM. B8 essentially says that Bel(s) is complete for each epistemic state s. As we already observed, Gärdenfors does not require completeness for formulas of the form $\varphi{>}\psi$, so B8 is not valid for him.

Preferential BCS's

Up to now we examined a very abstract notion of belief change. The definition of BCS puts few restrictions on the belief change process and does not provide much insight into the structure of such processes. We now describe a more specific class of systems that has a semantic representation similar to that of (Grove 1988; Katsuno & Mendelzon 1991; Boutilier 1992; Katsuno & Satoh 1991). The basic intuition is the following. We introduce *possible worlds*. Each possible world describes a way the world can be. We then associate with each epistemic set a set of possible worlds and a *preference* (or *plausibility*) ordering on worlds. The set of possible worlds associated with a state s defines the agent's beliefs at s in the usual manner, and the agent's epistemic state after learning φ corresponds to the minimal (i.e., most plausible) worlds satisfying φ.

We proceed as follows. A *preferential interpretation* of a BCS $\langle S, \rho, \tau\rangle$ is a tuple $\langle W, \pi, K, R\rangle$, where W is a set of possible *worlds*, π is a function mapping each world $w \in W$ to a maximally consistent subset of $\mathcal{L}$ (i.e., $\pi(w)$ must be consistent, and have the additional property that for each formula $\varphi \in \mathcal{L}$, either $\varphi \in \pi(w)$ or $\neg\varphi \in \pi(w)$), K is a mapping from S to subsets of W, and R is a function that maps each state $s \in S$ to a relation $\preceq_s$ over W.

The set $K(s)$ associated with each $s \in S$ describes the worlds considered possible when the agent is in state s. The ordering associated with each $s \in S$ describes a plausibility measure, or preference, among worlds. We define $\prec_s$ in the usual manner: $w \prec_s w'$ if $w \preceq_s w'$ and $w' \not\preceq_s w$. We require that $\preceq_s$ be smooth, i.e., for every $\varphi \in \mathcal{L}$ there are no infinite sequences of worlds $\ldots \prec_s w_1 \prec_s w_0$ such that $\varphi \in \pi(w_i)$ for all i. Following (Lewis 1973), we define $W_s = \{w \in W | \exists w' \in W, w \preceq_s w'\}$ as the set of worlds considered plausible when the agent is in state s. We require that $\preceq_s$ be a pre-order (i.e., reflexive and transitive relation) over W_s. Given φ, the set min(s, φ) is the set of minimal worlds in W_s that satisfy φ, i.e., $w \in$ min(s, φ) if $\varphi \in \pi(w)$, $w \in W_s$ and there is no $w' \prec_s w$ such that $\varphi \in \pi(w')$.

We want preferential interpretations to satisfy several consistency requirements that ensure that they satisfy the intuition we outlined above. Formally, we require that for all $s \in S$ the following hold:

- $\varphi \in \rho(s)$ if and only if $\varphi \in \pi(w)$ for all $w \in K(s)$.
- If $s' = \tau(s, \varphi)$ then $K(s') =$ min(s, φ).

Thus, each belief set is characterized by the set of worlds considered possible and belief change is described through the preference ordering associated with each belief set. A BCS is *preferential* if it has a preferential interpretation. Let $\mathcal{M}^P$ be the class of preferential belief structures.

Let AX^P be AX combined with the following axioms:

P1. $\varphi{>}B\varphi$

P2. $(\varphi_1{>}B\psi) \wedge (\varphi_1{>}B\varphi_2) \Rightarrow (\varphi_1 \wedge \varphi_2){>}B\psi$ if ψ, φ_1 and φ_2 are in $\mathcal{L}$

P3. $(\varphi_1{>}B\psi) \wedge (\varphi_2{>}B\psi) \Rightarrow (\varphi_1 \vee \varphi_2){>}B\psi$ if ψ, φ_1 and φ_2 are in $\mathcal{L}$

P4. $\varphi{>}B\psi \equiv \varphi'{>}B\psi$ if $\varphi \equiv \varphi'$ is $\mathcal{L}$-valid and $\psi \in \mathcal{L}$.

Theorem 4: *AX^P is a sound and complete axiomatization of $\mathcal{L}^{>}$ with respect to $\mathcal{M}^P$.*

We shall also be interested in subclasses of $\mathcal{M}^P$ that satisfy additional properties; these will help us capture belief revision and update.

The first property of interest is that the most preferred worlds according to the ordering $\preceq_s$ are precisely the worlds in $K(s)$. Formally, we say that the ordering $\preceq_s$ in a preferential interpretation is *faithful* if $K(s) =$ min$(s, true_{\mathcal{L}})$. If $\preceq_s$ is faithful, then $K(\tau(s, \varphi)) = K(s)$ if $\varphi \in \rho(s)$, so that an agent does not modify his beliefs if he learns something that he already believes. A preferential interpretation is *faithful* if $\preceq_s$ is faithful for every $s \in S$. This definition implies that once the agent is in an inconsistent state (i.e., one such that $K(s) = \emptyset$) he cannot leave it, i.e., min$(\emptyset, \varphi) = \emptyset$, for any φ.[4] This leads us to define a slightly weaker notion: A preferential interpretation is *weakly faithful* if $\preceq_s$ is faithful for all $s \in S$ such that $K(s) \neq \emptyset$. A preferential BCS is (weakly) faithful if it has a (weakly) faithful preferential interpretation. (Similarly, for other properties of interest, we say below that a preferential BCS has the property if it has a preferential interpretation that has it.)

We can characterize faithful and weakly faithful BCS's (in a sense made precise by Theorem 5 below) by the axioms PF and PW, respectively:

PF. $B\varphi \equiv (true_{\mathcal{L}}{>}B\varphi)$ for $\varphi \in \mathcal{L}$.

PW. $\neg B(false_{\mathcal{L}}) \Rightarrow (B\varphi \equiv (true_{\mathcal{L}}{>}B\varphi))$ for $\varphi \in \mathcal{L}$.

Notice that these axioms say only that in a (weakly) faithful BCS, the agent believes φ if and only if learning a valid formula results in him believing φ.

The property of faithfulness guarantees that if the agent learns something that he currently believes, then he still maintains all of his former $\mathcal{L}$-beliefs. What happens if he learns something *consistent* with his current beliefs, although not necessarily in the belief set? The next condition guarantees that the agent does not remove any of his previous beliefs in this case. A preferential structure is *ranked* if $\preceq_s$ is a total pre-order over W_s for every epistemic state s, i.e., for every $w, w' \in W_s$, either $w \preceq_s w'$ or $w' \preceq_s w$. Combining ranking with faithfulness guarantees that if the agent learns something that is consistent with

[4] This is one of the differences between revision and update (Katsuno & Mendelzon 1991); in revision the agent can "escape" the inconsistent state by revision with a consistent formula, and in update he cannot.

what he believes—i.e., if $\varphi \in \pi(w)$ for some $w \in K(s)$—then it must be the case that $K(\tau(s,\varphi)) \subseteq K(s)$, since the most preferred worlds (with respect to $\preceq_s$) where φ holds are precisely those worlds in $K(s)$ where φ is true. To see this, note that in a ranked and faithful ordering it must be the case that if $w \in K(s)$ and $w' \notin K(s)$, then $w \prec_s w'$. It follows that, in this case, $\rho(\tau(s,\varphi)) \supseteq \rho(s)$. Thus, if an agent learns something consistent with his current beliefs, he maintains all of his current $\mathcal{L}$-beliefs. Ranked BCS's can be characterized by the following axiom:

PR. $((\varphi_1 \vee \varphi_2) > B\neg\varphi_2) \Rightarrow ((\varphi_2 \vee \psi) > B\neg\varphi_2) \vee ((\varphi_1 \vee \psi) > B\neg\psi)$ if $\varphi_1, \varphi_2, \psi \in \mathcal{L}$.[5]

Axiom PR is an analogue of a standard axiom of conditional logic that captures the ranking condition (Burgess 1981). We must restrict the axiom here to $\mathcal{L}$-beliefs, whereas the corresponding axiom in conditional logic need not be restricted. This difference is rooted in the fact that we take epistemic states as the primitive objects, while standard conditional logic takes worlds to be the primitive objects.[6]

What happens when the agent learns something inconsistent with his current beliefs? The next condition puts another (rather weak) restriction on the set $\min(s,\varphi)$ in this case: a preferential structure is *saturated* if for every s and for every consistent $\varphi \in \mathcal{L}$, $\min(s,\varphi)$ is not empty. Thus, in a saturated preferential BCS, as long as what the agent learns is consistent, then his belief set will be consistent. Saturated BCS's can be characterized by the following axiom:

PS. $\neg(\varphi > B(\mathit{false}_{\mathcal{L}}))$ if $\varphi \in \mathcal{L}$ is consistent.

Typically we are interested in axiom schemes that are recursive (or at least r.e.). This scheme, however, may not be. It depends on how hard it is to check consistency in $\mathcal{L}$. For example, if $\mathcal{L}$ is first-order logic, this scheme is co-r.e.

Belief revision and belief update assume that the belief change process depends only on the agent's $\mathcal{L}$-beliefs. This is clearly a strong assumption. We feel that a more reasonable approach is to have the revision process depend on the full epistemic state, not just on the agent's $\mathcal{L}$-beliefs. Nevertheless, we can capture the assumption that all that matters are the agent's $\mathcal{L}$-beliefs quite simply. A BCS M is *propositional* if for all epistemic states $s, s' \in M$, we have that $\rho(s) = \rho(s')$ implies $\rho(\tau(s,\varphi)) = \rho(\tau(s',\varphi))$ for all $\varphi \in \mathcal{L}$.

A stronger version of P4 holds in propositional preferential structures. We no longer have to restrict to $\mathcal{L}$-beliefs. Thus we get:

PP1. $\varphi > \psi \equiv \varphi' > \psi$ if $\varphi \equiv \varphi'$ is $\mathcal{L}$-valid

In propositional preferential structures that are (weakly) faithful, we need to strengthen axioms PF and PW in an analogous way. Call these strengthened axioms PF$'$ and PW$'$, respectively.

[5]Alternatively, we can use the *rational monotonicity* axiom (Kraus, Lehmann, & Magidor 1990) $(\varphi > B\psi_1) \wedge \neg(\varphi > B\neg\psi_2) \Rightarrow (\varphi \wedge \psi_2 > B\psi_1)$, which is similar to what has been used by (Grahne 1991; Katsuno & Satoh 1991) to capture ranked structures.

[6]For similar reasons, the axioms P2 and P3 are restricted while their counterparts in conditional logic (see (Lewis 1973)) are not.

These changes do not suffice to characterize propositional preferential structures. To do that, we need some additional machinery. We are interested in formulas that describe epistemic states. Given a belief set $E \subseteq \mathcal{L}$, we say that φ_E describes E if for all preferential BCS's, $(M,s) \models \varphi_E$ if and only if $\rho(s) = E$. We say that a formula is a *state description* if it describes some belief set. Note that the inconsistent belief state is always describable by $B(\mathit{false}_{\mathcal{L}})$, the describability of other states depends on the logic $\mathcal{L}$. It is easy to see that if $\mathcal{L}$ is a propositional logic over a finite number of primitive propositions, then all belief states are describable, while if $\mathcal{L}$ is propositional logic with infinitely many primitive propositions, then the inconsistent set is the only describable belief set. We remark if $\mathcal{L}$ included an *only knowing operator* of (Levesque 1990) (as in (Rott 1989; Boutilier 1992)), then more belief sets would be describable.

The following axiom, together with PP1 (and PF$'$ and PW$'$, if we are considering (weakly) faithful structures), characterizes propositional preferential BCS's:

PP2. $(\varphi \wedge (\psi_1 > \cdots > \psi_k > \varphi)) \Rightarrow ((\varphi_1 > \varphi_2) \equiv \psi_1 > \cdots > \psi_k > \varphi_1 > \varphi_2)$ if φ is a state description.

Axiom PP2 says that if $\varphi_1 > \varphi_2$ holds in the current state and φ characterizes the agent's current beliefs, then if after learning a number of facts the agent reaches a state with exactly the same beliefs, then $\varphi_1 > \varphi_2$ also holds in that state.

The next condition we consider says that the ordering $\preceq_s$ is determined by orderings $\preceq_w$ associated with worlds $w \in K(s)$. This corresponds to the intuition of (Katsuno & Mendelzon 1991) that in belief update, we do the update pointwise (so that if we consider a set of worlds possible, we update each of them individually). Formally, we say that a preferential interpretation is *decomposable* if there is a mapping that associates each $w \in W$ with an ordering $\preceq_w$ such that $\preceq_w$ is a pre-order on $W_w = \{w' | \exists w'', w' \preceq_w w''\}$ and the following condition is satisfied: for all $s \in S$, such that $K(s) \neq \emptyset$, we have $w \prec_s w'$ if and only if $w \prec_v w'$ for all $v \in K(s)$. It easy to show that this definition implies that $\min(s,\varphi) = \bigcup_{v \in K(s)} \min(v,\varphi)$, (where $\min(v,\varphi)$ is defined similarly to $\min(s,\varphi)$) matching the condition of (Katsuno & Mendelzon 1991) for update.

Characterizing decomposable BCS's is nontrivial. However, in two cases we have (different) characterizations of decomposable BCS's. When we examine decomposable BCS's that are also (weakly) faithful and ranked we need the following two axioms:

PD1. $((\neg B\neg\varphi \wedge (\psi_1 > B\psi_2)) \Rightarrow \varphi > \psi_1 > B\psi_2$ if $\varphi, \psi_2 \in \mathcal{L}$

PD2. $(B(\varphi_1 \vee \ldots \vee \varphi_k) \wedge (\wedge_{j=1}^{k}(\varphi_j > \psi_1 > B\psi_2))) \Rightarrow \psi_1 > B\psi_2$ if $\varphi_1, \ldots, \varphi_k, \psi_1, \psi_2 \in \mathcal{L}$.

Both axioms rely on the property of ranked and (weakly) faithful structures that if φ is consistent with $\rho(s)$ then $K(\tau(s,\varphi)) \subseteq K(s)$. Another situation where we can characterize decomposable structures is where we also assume that the structures are propositional. In this case we can use state descriptions and the fact that all subsets that are equivalent in terms of belief sets also revise in the same manner.

We get two axioms PD1$'$ and PD2$'$ that are analogues of PD1 and PD2. We omit them here for lack of space; they are described in the technical report.

Finally, we say that a BCS M is *complete* if for each belief set E, there is some state s in M such that $\rho(s) = E$. We have no axiom to characterize completeness, and we do not need one. As we shall see, in structures of interest to us, completeness does not add extra properties.

Let A be a subset of $\{f, w, r, s, p, d, c\}$. We denote by $\mathcal{M}_A^P$ the class of preferential BCS's that satisfy the respective subset of {faithful, weakly faithful, ranked, saturated, propositional, decomposable, complete}. For example, $\mathcal{M}_{r,s}^P$ is the class of ranked and saturated preferential BCS's.

We can now state precisely the sense in which the axioms characterize the conditions we have described. Roughly, the axiom system contains AX^P and for each one of $\{f, w, r, s\}$ in $\mathcal{A}$, the matching axiom described above. When $\mathcal{A}$ contains d the axiom system may also contain PD1 and PD2 (depending on the contents of $\mathcal{A}$). When $\mathcal{A}$ contains p, the axiom system also contains PP1 and PP2 and the strengthened versions of the axioms corresponding to f and w. Moreover, PD1$'$ and PD2$'$ are required to deal with d. This is captured by the following theorem.

Theorem 5: *Let $\mathcal{A}$ be a subset of $\{f, w, r, s\}$, let $\mathcal{B}$ be a subset of $\{d\}$, and let $\mathcal{C}$ be a subset of $\{c\}$. Let A be the subset of {PF, PW, PR, PS} corresponding to $\mathcal{A}$, let A' be the subset of {PF$'$, PW$'$, PR, PS} corresponding to $\mathcal{A}$, let*

$$B = \begin{cases} \{\textit{PD1,PD2}\} & \textit{if } d \in \mathcal{B},\ r \in \mathcal{A},\ \{f, w\} \cap \mathcal{A} \neq \emptyset \\ \emptyset & \textit{otherwise,} \end{cases}$$

and let

$$B' = \begin{cases} \{\textit{PD1}',\textit{PD2}'\} & \textit{if } d \in \mathcal{B} \\ \emptyset & \textit{otherwise,} \end{cases}$$

Then AX^P $\cup A \cup B$ is a sound and complete axiomatization of $\mathcal{L}^{>}$ with respect to $\mathcal{M}^P_{\mathcal{A}\cup\mathcal{B}\cup\mathcal{C}}$, and AX^P $\cup A' \cup B' \cup \{PP1, PP2\}$ is a sound and complete axiomatization of $\mathcal{L}^{>}$ with respect to $\mathcal{M}^P_{\mathcal{A}\cup\mathcal{B}\cup\mathcal{C}\cup\{p\}}$.

Belief revision and belief update

The standard approach to defining belief revision and belief update is in terms of functions mapping deductively closed subsets of $\mathcal{L}$ and formulas in $\mathcal{L}$ to deductively closed subsets of $\mathcal{L}$, satisfying certain properties. We do not describe these properties here due to lack of space, but they can be found in (Gärdenfors 1988; Katsuno & Mendelzon 1991).

Given an update or revision operator f, we can associate with it a BCS $M_f = (S, \rho, \tau)$ in a straightforward way: the elements of S are all the deductively closed subsets of $\mathcal{L}$, for $s \in S$, we define $\rho(s) = s$, and we define $\tau(s, \varphi) = f(\rho(s), \varphi)$. It is not hard to show that f is a revision (resp. update) operator if and only if $M_f \in \mathcal{M}^P_{w,r,s,p,c}$ (resp. $M_f \in \mathcal{M}^P_{f,s,p,d,c}$). We might also hope to show that every system in $\mathcal{M}^P_{w,r,s,p,c}$ is of the form M_f for some revision operator f, so that $\mathcal{M}^P_{w,r,s,p,c}$ characterizes revision operators (and similarly for $\mathcal{M}^P_{f,s,p,d,c}$ and update operators). However, we have a slight technical problem, since even a propositional a BCS might contain more than one state with the same belief set, while M_f contains each belief set exactly once. This turns out to be not such a serious problem. We say that two BCS's M and M' are *equivalent* if for every $s \in M$ there is an $s' \in M'$ such that $\text{Bel}(s) = \text{Bel}(s')$ and vice versa. It follows from Proposition 1 that if $\text{Bel}(s) = \text{Bel}(s')$, then $\text{Bel}(\tau(s, \varphi)) = \text{Bel}(\tau(s', \varphi))$ for all $\varphi \in \mathcal{L}$. Hence, we can identify two equivalent BCS's (and, in particular, the same formulas are valid in equivalent BCS's).

Theorem 6:

(a) f is a belief revision operator if and only if $M_f \in \mathcal{M}^P_{w,r,s,p,c}$. Moreover, $M \in \mathcal{M}^P_{w,r,s,p,c}$ if and only if M is equivalent to M_f for some belief revision operator f.

(b) f is a belief update operator if and only if $M_f \in \mathcal{M}^P_{f,s,p,d,c}$. Moreover, $M \in \mathcal{M}^P_{f,s,p,d,c}$ if and only if M is equivalent to M_f for some belief update operator f.

This theorem, which can be viewed as a complete characterization of belief revision and belief update in terms of BCS's, is perhaps not so surprising, since it is in much the same spirit as other characterizations of belief revision and update (Grove 1988; Katsuno & Mendelzon 1991). On the other hand, when combined with Theorem 5, it means we have a complete axiomatization of belief change under belief revision and belief update.

It is interesting to compare this result to the work of (Gärdenfors 1978; 1986). In Theorem 6, the belief revision functions learned only formulas in $\mathcal{L}$, not $\mathcal{L}^{>}$. It follows from the theorem that in structures in $\mathcal{M}^P_{w,r,s,p,c}$, the AGM postulates hold, if we consider revision with respect to formulas in $\mathcal{L}$ and take belief sets to be subsets of $\mathcal{L}$, not $\mathcal{L}^{>}$. Because we restrict to belief sets in $\mathcal{L}$ and revise only by formulas in $\mathcal{L}$, we avoid the triviality problem that occurs when applying the AGM postulates to conditional beliefs (Gärdenfors 1986) or to nested beliefs (Levi 1988; Fuhrmann 1989). We remark that this approach to dealing with the triviality problem is in the spirit of suggestions made earlier (Levi 1988; Rott 1989; Boutilier 1992).

Discussion

We have analyzed belief change systems, starting with a very abstract notion of belief change and adding structure to it. The main contribution of this work lies in giving a logical (proof-theoretic) characterization of belief change operators and, in particular, belief revision and belief update. Our analysis shows what choices, in terms of semantic properties, lead to these two notions, and gives us a natural class of belief change operators that generalizes both.

Our work is also relevant to the problem of iterated belief revision. It is clear that the axiomatization we provide for belief revision captures all the properties of iterated AGM belief revision. This axiomatization highlights the fact the

AGM postulates put few restrictions on iterated belief revision. (Boutilier 1993) and (Darwiche & Pearl 1994) suggest strengthening belief revision by adding postulates on iterated belief change. In the full paper we show that these constraints can be easily axiomatized in our language, thus providing a proof system for iterated belief revision.

An important aspect of our work is the distinction between objective statements about the world and subjective statements about the agents beliefs. To analyze belief change we need to examine only the latter, and this is reflected in our choice of language. However, we believe that it is important to study belief change in frameworks that describe both the world and the agent's beliefs, and how both change over time. This type of investigation, which we are currently undertaking (see (Friedman & Halpern 1994a; 1994b)), should provide guidance in selecting the most reasonable and useful properties of belief change.

Acknowledgements

The authors are grateful to Craig Boutilier, Ronen Brafman, Adnan Darwiche, Daphne Koller, Alberto Mendelzon, and the anonymous referees for comments on drafts of this paper and useful discussions relating to this work.

References

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

Boutilier, C., and Goldszmidt, M. 1993. Revising by conditional beliefs. In *Proc. National Conference on Artificial Intelligence (AAAI '93)*, 648–654.

Boutilier, C. 1992. Normative, subjective and autoepistemic defaults: Adopting the Ramsey test. In *Principles of Knowledge Representation and Reasoning: Proc. Third International Conference (KR '92)*. San Francisco, CA: Morgan Kaufmann.

Boutilier, C. 1993. Revision sequences and nested conditionals. In *Proc. Thirteenth International Joint Conference on Artificial Intelligence (IJCAI '93)*, 519–525.

Burgess, J. 1981. Quick completeness proofs for some logics of conditionals. *Notre Dame Journal of Formal Logic* 22:76–84.

Darwiche, A., and Pearl, J. 1994. On the logic of iterated belief revision. In Fagin, R., ed., *Theoretical Aspects of Reasoning about Knowledge: Proc. Fifth Conference*. San Francisco, CA: Morgan Kaufmann. 5–23.

Friedman, N., and Halpern, J. Y. 1994a. A knowledge-based framework for belief change. Part I: Foundations. In Fagin, R., ed., *Theoretical Aspects of Reasoning about Knowledge: Proc. Fifth Conference*. San Francisco, CA: Morgan Kaufmann. 44–64.

Friedman, N., and Halpern, J. Y. 1994b. A knowledge-based framework for belief change. Part II: revision and update. In Doyle, J.; Sandewall, E.; and Torasso, P., eds., *Principles of Knowledge Representation and Reasoning: Proc. Fourth International Conference (KR '94)*. San Francisco, CA: Morgan Kaufmann.

Fuhrmann, A. 1989. Reflective modalities and theory change. *Synthese* 81:115–134.

Gärdenfors, P. 1978. Conditionals and changes of belief. *Acta Philosophica Fennica* 20.

Gärdenfors, P. 1986. Belief revision and the Ramsey test for conditionals. *Philosophical Review* 91:81–93.

Gärdenfors, P. 1988. *Knowledge in Flux*. Cambridge, UK: Cambridge University Press.

Grahne, G. 1991. Updates and counterfactuals. In *Principles of Knowledge Representation and Reasoning: Proc. Second International Conference (KR '91)*. San Francisco, CA: Morgan Kaufmann. 269–276.

Grove, A. 1988. Two modelings for theory change. *Journal of Philosophical Logic* 17:157–170.

Harper, W.; Stalnaker, R. C.; and Pearce, G., eds. 1981. *Ifs*. Dordrecht, Netherlands: Reidel.

Katsuno, H., and Mendelzon, A. 1991. On the difference between updating a knowledge base and revising it. In *Principles of Knowledge Representation and Reasoning: Proc. Second International Conference (KR '91)*. San Francisco, CA: Morgan Kaufmann. 387–394.

Katsuno, H., and Satoh, K. 1991. A unified view of consequence relation, belief revision and conditional logic. In *Proc. Twelfth International Joint Conference on Artificial Intelligence (IJCAI '91)*, 406–412.

Kraus, S.; Lehmann, D. J.; and Magidor, M. 1990. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence* 44:167–207.

Levesque, H. J. 1990. All I know: A study in autoepistemic logic. *Artificial Intelligence* 42(3):263–309.

Levi, I. 1988. Iteration of conditionals and the Ramsey test. *Synthese* 76:49–81.

Lewis, D. K. 1973. *Counterfactuals*. Cambridge, MA.: Harvard University Press.

Rott, H. 1989. Conditionals and theory change: revision, expansions, and additions. *Synthese* 81:91–113.

Rott, H. 1990. A nonmonotonic conditional logic for belief revision. In A, F., and Morreau, M., eds., *The Logic of Theory Change*. Springer-Verlag. 135–181.

Stalnaker, R. C. 1968. A theory of conditionals. In Rescher, N., ed., *Studies in logical theory*, number 2 in American Philosophical Quarterly monograph series. Blackwell, Oxford. Also appears in *Ifs*, (ed., by W. Harper, R. C. Stalnaker and G. Pearce), Reidel, Dordrecht, 1981.

Wobcke, W. 1992. On the use of epistemic entrenchment in nonmonotonic reasoning. In *10th European Conference on Artificial Intelligence (ECAI'92)*, 324–328.