A Knowledge-Based Framework for Belief Change, Part II: Revision and Update*

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

Joseph Y. Halpern
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120–6099
halpern@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. In a companion paper [FH94b] we introduced a new framework to model belief change. This framework combines temporal and epistemic modalities with a notion of plausibility, allowing us to examine the changes of beliefs over time. In this paper we show how belief revision and belief update can be captured in our framework. This allows us to compare the assumptions made by each method and to better understand the principles underlying them. In particular, it allows us to understand the source of Gärdenfors' triviality result for belief revision [Gär86] and suggests a way of mitigating the problem. It also shows that Katsuno and Mendelzon's notion of belief update [KM91a] depends on several strong assumptions that may limit its applicability in AI.

1 INTRODUCTION

The study of *belief change* has been an active area in philosophy and AI. The focus of this research is to understand how an agent should change his beliefs as a result of getting new information. Two instances of this general phenomenon have been studied in detail. *Belief revision* [AGM85, Gär88] focuses on how an agent revises his beliefs when he adopts a new belief. *Belief update* [KM91a], on the other hand, focuses on how an agent should change his beliefs when he realizes that the world has changed. Both approaches attempt to capture the intuition that an agent should make minimal changes in his beliefs in order to accommodate the new belief. The difference is that belief revision attempts to decide what beliefs should be discarded to accommodate a new belief, while belief update attempts to decide what changes in the world led to the new observation.

In [FH94b] we introduce a general framework for modeling belief change. We start with the framework for analyzing knowledge in multi-agent systems, introduced in [HF89], and add to it a notion of plausibility ordering at each situation. We then define belief as truth in the most plausible situations. The resulting framework is very expressive; it captures both time and knowledge as well as beliefs. The representation of time allows us to reason in the framework about changes in the beliefs of the agent. It also allows us to relate the beliefs of the agent about the future with his actual beliefs in the future. Knowledge captures in a precise sense the non-defeasible information the agent has about the world he is in, while belief captures defeasible information. The framework allows us to represent a broad spectrum of notions of belief change. In this paper we show how belief revision and update can be represented. Doing this allows us to compare the assumptions implicit in each method and to understand the principles underlying them.

The explicit representation of time allows us to investigate some of the subtle differences between revision and update. For example, in the literature, belief revision has been described (in [KM91a], for example) as a process of changing beliefs about a *static world*, but this is slightly misleading. In fact, what is important for revision is not that the world is static, but that the propositions used to describe the world are static, i.e., their truth value does not change over time.[1] For example, "At time 0 the block is on the table" is a static proposition, while "The block is on the table" is not, since it implicitly references the *current* state of affairs. Belief update, on the other hand, deals with

*To appear in Principles of Knowledge Representation and Reasoning: Proc. Fourth International Conference (KR '94).

[1] This assumption is not unique to belief revision. Bayesian updating, for example, makes similar assumptions.

propositions whose truth depends on the current situation. It allows any proposition to change its truth value, and treats this as a change in the world rather than as a change in the agent's beliefs about the world.

This distinction allows us to better understand Gärdenfors' triviality result [Gär86]. This result states that the belief revision postulates cannot be applied to belief states that contain *Ramsey conditionals* of the form $\varphi > \psi$ with the interpretation "revising by φ will lead to a state where ψ is believed". Technically, this is because the AGM framework includes a postulate of *persistence*: if φ is consistent with the current beliefs, then no beliefs should be discarded to accommodate φ. Since the truth value of a Ramsey conditional depends on the current state of the agent, it is inappropriate to assume that it persists when that state changes. It should thus be no surprise that assuming persistence of such formulas leads to triviality. Indeed, this observation was essentially already made by Levi [Lev88]. Our solution to the triviality result is somewhat different from others that have been considered in the literature (e.g., [Rot89, Fuh89, LR92, Bou92]) in that it shifts the focus from postulates for the revision process to considerations of the appropriate logic of conditionals.

We then turn our attention to belief update. Our treatment enables us to identify implicit assumptions made in the update process. In particular, it brings out how update prefers to defer abnormalities to as late a time as possible. This allows us to clarify when update is appropriate. Essentially, it is appropriate if the agent always receives enough information to deduce the exact change in the state of the world, a condition unlikely to be met in most AI applications.

We are certainly not the first to provide semantic models for belief revision and update. For example, [AGM85, Gro88, GM88, Rot91, Bou92, Rij92] deal with revision and [KM91a, dVS92] deal with update. In fact, there are several works in the literature that capture both using the same machinery [KS91, GP92] and others that simulate belief revision using belief update [GMR92, dVS94]. Our approach is different from most in that we did not construct a specific framework to capture one or both belief change paradigms. Instead, we start from a natural framework to model how an agent's knowledge changes over time [HF89] and add to it machinery that captures a defeasible notion of belief. As we shall see, our framework allows us to clearly bring out the similarities and differences between update and revision. We believe that the insights gained into revision and update using our approach—particularly in terms of the assumptions that each makes about how an agent's plausibility ordering changes over time—provide further justification as to the usefulness of having such a framework.

2 THE FRAMEWORK

We now review the framework of [HF89] for modeling knowledge in multi-agent systems, and our extension of it [FH94b] for dealing with belief change.

The key assumption in this framework is that we can characterize the system by describing it in terms of a *state* that changes over time. Formally, we assume that at each point in time, the agent is in some *local state*. Intuitively, this local state encodes the information the agent has observed thus far. There is also an *environment*, whose state encodes relevant aspects of the system that are not part of the agent's local state.

A *global state* is a tuple (s_e, s_a) consisting of the environment state s_e and the local state s_a of the agent. A *run* of the system is a function from time (which, for ease of exposition, we assume ranges over the natural numbers) to global states. Thus, if r is a run, then $r(0), r(1), \ldots$ is a sequence of global states that, roughly speaking, is a complete description of what happens over time in one possible execution of the system. We take a *system* to consist of a set of runs. Intuitively, these runs describe all the possible behaviors of the system, that is, all the possible sequences of events that could occur in the system over time.

Given a system $\mathcal{R}$, we refer to a pair (r, m) consisting of a run $r \in \mathcal{R}$ and a time m as a *point*. If $r(m) = (s_e, s_a)$, we define $r_a(m) = s_a$ and $r_e(m) = s_e$. We say two points (r, m) and (r', m') are *indistinguishable* to the agent, and write $(r, m) \sim_a (r', m')$, if $r_a(m) = r'_a(m')$, i.e., if the agent has the same local state at both points. Finally, Halpern and Fagin define an *interpreted* system $\mathcal{I}$ to be a tuple $(\mathcal{R}, \pi)$ consisting of a system $\mathcal{R}$ together with a mapping π that associates with each point a truth assignment to the primitive propositions. In an interpreted system we can talk about an agent's knowledge: the agent knows φ at a point (r, m) if φ holds in all points (r', m') such that $(r, m) \sim_a (r', m')$. However, we can not talk about the agent's (possibly defeasible) beliefs at (r, m).

To remedy this deficiency, in [FH94b] we added plausibility orderings to interpreted systems. We can then say that the agent believes φ if φ is true at all the most plausible worlds. Formally, a *plausibility space* is a tuple $(\Omega, \preceq)$, where Ω is a set of points in the system, and $\preceq$ is a preorder (i.e., a reflexive and transitive relation) over Ω. As usual, we write $(r', m') \prec (r'', m'')$ if $(r', m') \preceq (r'', m'')$ and it is not the case that $(r'', m'') \preceq (r', m')$. Intuitively, $(r', m') \prec (r'', m'')$ if (r', m') is strictly more plausible than (r'', m'') according to the plausibility ordering. An *(interpreted) plausibility system* is a tuple $(\mathcal{R}, \pi, \mathcal{P})$ where, as before, $\mathcal{R}$ is a set of runs and π maps each point to a truth assignment, and where $\mathcal{P}$ is a *plausibility assignment function* mapping each point (r, m) to a plausibility space $\mathcal{P}(r, m) = (\Omega_{(r,m)}, \preceq_{(r,m)})$. Intuitively, the plausibility space $\mathcal{P}(r,m)$ describes the relative plausibility of points from the point of view of the agent at (r,m). In this paper we assume that $\Omega_{(r,m)}$ is a (possibly empty) subset of $\{(r',m')|(r,m) \sim_a (r',m')\}$. Thus, the agent considers plausible only situations that are possible according to his knowledge. We also assume that the plausibility space is a function of the agent's local state. Thus, if $(r,m) \sim_a (r',m')$ then $\mathcal{P}(r,m) = \mathcal{P}(r',m')$.[2]

We define the logical language $\mathcal{L}^{KPT}(\Phi)$ to be a propositional language over a set of primitive propositions Φ with the following modalities: $K\varphi$ (the agent knows φ is true), $\bigcirc\varphi$ (φ is true in the next time step), and $\varphi \rightarrow \psi$ (in all most-plausible situations where φ is true, ψ is also true).[3] We recursively assign truth values to formulas in $\mathcal{L}^{KPT}(\Phi)$ at a point (r,m) in a plausibility system $\mathcal{I}$. The truth of primitive propositions is determined by π, so that

$$(\mathcal{I},r,m) \models p \text{ if and only if } \pi(r,m)(p) = \textbf{true}.$$

Conjunction and negation are treated in the standard way, as is knowledge: The agent knows φ at (r,m) if φ holds at all points that he cannot distinguish from (r,m). Thus,

$$(\mathcal{I},r,m) \models K\varphi \text{ if } (\mathcal{I},r',m') \models \varphi \text{ for all } (r',m') \sim_a (r,m).$$

$\bigcirc\varphi$ is true at (r,m) if φ is true at $(r,m+1)$. Thus,

$$(\mathcal{I},r,m) \models \bigcirc\varphi \text{ if } (\mathcal{I},r,m+1) \models \varphi.$$

We would like $\varphi \rightarrow \psi$ to be true at (r,m) if the most plausible points in $\Omega_{(r,m)}$ that satisfy φ also satisfy ψ. The actual definition that we use, which is standard in the literature (see [Lew73, Bur81, Bou92]), captures this desideratum if there are most plausible points that satisfy φ (in particular, if $\Omega_{(r,m)}$ is finite), and also deals with the more general case where there may be a sequence of increasingly more plausible points, with none being most plausible, i.e., $\ldots s_3 \prec_{(r,m)} s_2 \prec_{(r,m)} s_1$. The actual definition says that $\varphi \rightarrow \psi$ is true at a point (r,m) if for every point (r_1,m_1) in $\Omega_{(r,m)}$ satisfying φ, there is another point (r_2,m_2) such that (a) (r_2,m_2) is at least as plausible as (r_1,m_1), (b) (r_2,m_2) satisfies $\varphi \wedge \psi$, and (c) each point satisfying φ that is at least as plausible as (r_2,m_2) also satisfies ψ.

$$(\mathcal{I},r,m) \models \varphi \rightarrow \psi \text{ if for every } (r_1,m_1) \in \Omega_{(r,m)} \text{ such that } (\mathcal{I},r_1,m_1) \models \varphi, \text{ there is a point } (r_2,m_2) \preceq_{(r,m)} (r_1,m_1) \text{ such that } (\mathcal{I},r_2,m_2) \models \varphi \wedge \psi, \text{ and there is no } (r_3,m_3) \preceq_{(r,m)} (r_2,m_2) \text{ such that } (\mathcal{I},r_3,m_3) \models \varphi \wedge \neg\psi.$$

We now define a notion of *belief.* Intuitively, the agent believes φ if φ is true in all the worlds he considers most plausible. Formally, we define $B\varphi \Leftrightarrow$ *true*$\rightarrow\varphi$. In [FH94b] we prove that, in this framework, knowledge is an S5 operator, belief is a KD45 operator, and the interactions between knowledge and belief are captured by the axioms $K\varphi \Rightarrow B\varphi$ and $B\varphi \Rightarrow KB\varphi$.

In a plausibility system, the agent's beliefs change from point to point because his plausibility space changes. The general framework does not put any constraints on how the plausibility space changes. In this paper, we identify the constraints that correspond to belief revision and update.

3 BELIEF CHANGE SYSTEMS

In the rest of this paper, we focus on a certain class of systems that we call *belief change systems*, in which we can capture both belief revision and belief update. These systems describe agents that change their local state at each round according to new information they receive (or learn). Both revision and update assume that this information is described by a formula.[4] Thus, they describe how the agent's beliefs change when the new information is captured by a formula φ. Implicitly they assume that φ is the only factor that affects the change. We now make this assumption precise.

We start with some language $\mathcal{L}(\Phi)$ that describes the worlds. We assume that $\mathcal{L}(\Phi)$ contains the propositional calculus and has a consequence relation $\vdash_{\mathcal{L}}$ that satisfies the deduction theorem. The set Φ denotes the primitive propositions in the $\mathcal{L}(\Phi)$. We can think of $\vdash_{\mathcal{L}}$ as a description of *state constraints* that govern the language. We assume that the agent is described by a *protocol*. The protocol describes how the agent changes state when receiving new information. Formally, a protocol is a tuple $P = (S, s_0, \tau)$, where S is the set of local states the agent can attain, s_0 is the *initial state* of the agent, and τ is a *transition function* that maps a state and a formula in $\mathcal{L}(\Phi)$ to another state. We take $\tau(s,\varphi)$ to be the local state of the agent after learning φ in local state s. We sometimes write $s \cdot \varphi$ instead of $\tau(s,\varphi)$.

To clarify the concept of protocol, we examine a rather simple protocol that we use below in our representation of update. The protocol P^* is defined as follows:

[2] The framework presented in [FH94b] is more general than this, dealing with multiple agents and allowing the agent to consider several plausibility spaces in each local state. The simplified version we present here suffices to capture belief revision and update.

[3] It is easy to add other temporal modalities such as until, eventually, since, etc. These do not play a role in this paper.

[4] This is a rather strong assumption, since it implies that the language in question can capture, in a precise manner, the information content of the change. Our framework can also be used to describe situations where this assumption does not hold.

The agent's local state is simply the sequence of observation made. Thus, S is the set of sequences of formulas in $\mathcal{L}(\Phi)$. Initially the agent has not made any observations, so $s_0 = \langle\rangle$. The transition function simply appends the new observation to the agent's state: $\tau(\langle\varphi_0,\ldots,\varphi_n\rangle,\psi) = \langle\varphi_0,\ldots,\varphi_n,\psi\rangle$. This simple definition describes an agent that remembers all his observations.[5]

Given a protocol P, we define $\mathcal{R}(P)$ to be the system consisting of all runs in which the agent runs P as follows. Recall that in our framework we need to describe the local states of the agent and the environment at each point. We use the environment state to represent which of the propositions in Φ is true and what observation the agent makes. We represent a truth assignment over Φ by the set Ψ of propositions that are true. We say that a truth assignment Ψ is *consistent* according to $\vdash_{\mathcal{L}}$ if for every $\varphi_1,\ldots\varphi_n \in \Psi$ and $\psi_1\ldots\psi_m \notin \Psi$, it is not the case that $\vdash_{\mathcal{L}} \neg(\bigwedge_{i=1} n\varphi_i \wedge \bigwedge_{i=m} k\neg\psi_i)$. Formally, we take the environment state to be a pair $r_e(m) = \langle\Psi,\varphi\rangle$, such that $\Psi \subseteq \Phi$ is a consistent truth assignment to Ψ and φ is the observation that the agent makes in the transition from (r,m) to $(r,m+1)$. If $r_e(m) = \langle\Psi,\varphi\rangle$, then we define $\mathit{world}(r,m) = \Psi$ and $\mathit{obs}(r,m) = \varphi$.

We take $\mathcal{R}(P)$ to be the set of all runs satisfying the following conditions for all $m \geq 0$:

- $r_a(m) \in S$
- $r_e(m) = \langle\Psi,\varphi\rangle$, where $\Psi \subseteq \Phi$ is consistent according to $\vdash_{\mathcal{L}}$ and $\varphi \in \mathcal{L}(\Phi)$
- $r_a(0) = s_0$
- $r_a(m+1) = \tau_a(r_a(m), \mathit{obs}(r,m))$.

Notice that because $\mathcal{R}(P)$ contains all runs that satisfy these conditions, for each sequence of world states $\Psi_0,\ldots,\Psi_m$ and observations $\varphi_0,\ldots\varphi_m$ there is a run in $\mathcal{R}(P)$ such that $r_e(i) = \langle\Psi_i,\varphi_i\rangle$ for all $0 \leq i \leq m$.

We introduce propositions that allow us to describe the observations the agent make at each step. More formally, let Φ^* be the set of primitive propositions obtained by augmenting Φ with all primitive propositions of the form $\mathit{learn}(\varphi)$, where $\varphi \in \mathcal{L}(\Phi)$. Intuitively, $\mathit{learn}(\varphi)$ holds if the agent has just learned (or observed) φ. We now define a truth assignment π on the points in $\mathcal{R}(P)$ in the obvious way: For $p \in \Phi$, we define $\pi(r,m)(p) = \mathbf{true}$ if and only if $p \in \mathit{world}(r,m)$. Since the formula $\mathit{learn}(\varphi)$ is intended to denote that the agent has just learned φ, we define $\pi(r,m)(\mathit{learn}(\varphi)) = \mathbf{true}$ if and only if $\mathit{obs}(r,m) = \varphi$.

These definitions set the background for our presentation of belief revision and belief update. Our description is still missing a plausibility assignment function that describes the plausibility ordering of the agent at each point. This function requires a different treatment for revision and update. Indeed, the plausibility function is the main source of difference between the two notions.

4 REVISION

Belief revision attempts to describe how a rational agent incorporates new beliefs. As we said earlier, the main intuition is that as few changes as possible should be made. Thus, when something is learned that is consistent with earlier beliefs, it is just added to the set of beliefs. The more interesting situation is when the agent learns something inconsistent with his current beliefs. He must then discard some of his old beliefs in order to incorporate the new belief and remain consistent. The question is which ones?

The most widely accepted notion of belief revision is defined by the AGM theory [AGM85, Gär88]. The agent's epistemic state is represented as a *belief set*, that is, a set of formulas in $\mathcal{L}(\Phi)$ closed under deduction. There is also assumed to be a revision operator $\circ$ that takes a belief set A and a formula φ and returns a new belief set $A \circ \varphi$, intuitively, the result of revising A by φ. The following AGM postulates are an attempt to characterize the intuition of "minimal change":

(R1) $A \circ \varphi$ is a belief set
(R2) $\varphi \in A \circ \varphi$
(R3) $A \circ \varphi \subseteq Cl(A \cup \{\varphi\})$[6]
(R4) If $\neg\varphi \notin A$ then $Cl(A \cup \{\varphi\}) \subseteq A \circ \varphi$
(R5) $A \circ \varphi = Cl(\mathit{false})$ if and only if $\vdash_L \neg\varphi$
(R6) If $\vdash_L \varphi \Leftrightarrow \psi$ then $A \circ \varphi = A \circ \psi$
(R7) $A \circ (\varphi \wedge \psi) \subseteq Cl(A \circ \varphi \cup \{\psi\})$
(R8) If $\neg\psi \notin A \circ \varphi$ then $Cl(A \circ \varphi \cup \{\psi\}) \subseteq A \circ (\varphi \wedge \psi)$.

The essence of these postulates is the following. After a revision by φ the belief set should include φ (postulates R1 and R2). If the new belief is consistent with the belief set, then the revision should not remove any of the old beliefs and should not add any new beliefs except these implied by the combination of the old beliefs with the new belief (postulates R3 and R4). This condition is called *persistence*. The next two conditions discuss the coherence of beliefs. Postulate R5 states that the agent is capable of incorporating any consistent belief and postulate R6 states that the syntactic form of the new belief does not affect the revision process. The last two postulates enforce a certain coherency on the outcome of revisions by related beliefs. Basically they state that if ψ is consistent with $A \circ \varphi$ then $A \circ (\varphi \wedge \psi)$ is just $A \circ \varphi$ combined with ψ. This ensures that revision is coherent regarding the outcome of revision by similar formulas (e.g., φ and $\varphi \wedge \psi$).

[5] We remark that P^* is similar to protocols used to model knowledge bases in [FHMV94].

[6] $Cl(A) = \{\varphi | A \vdash_{\mathcal{L}} \varphi\}$ is the deductive closure of a set of formulas A.

While there are several representation theorems for belief revision, the clearest is perhaps the following [Gro88, KM91b]: We associate with each belief set A a set W_A of possible worlds. Intuitively, the worlds in W_A are all those that are consistent with the agent's beliefs, in that W_A consists of all those worlds in which all formulas in A are true. Thus, an agent whose belief set is A believes that one of the worlds in W_A is the real world. An agent that performs belief revision behaves as though in each belief state A he has a *ranking*, i.e., a total preorder, over all possible worlds such that the minimal (i.e., most plausible) worlds in the ranking are exactly those in W_A. The ranking prescribes how the agent revises his beliefs. When revising by φ, the agent chooses the minimal worlds satisfying φ in the ranking and constructs a belief set from them. It is easy to see that this procedure for belief revision satisfies the AGM postulates. Moreover, in [Gro88, KM91b] it is shown that any belief revision operator can be described in terms of such a ranking.

This representation suggests how we can capture belief revision in our framework. We want to define a family of belief systems that captures all the revision operators consistent with the AGM postulates. Since revision assumes that the primitive propositions are static, we assume that world state is constant throughout the run. To capture this, we use a variant of our definition from Section 3: Given P, let $\mathcal{R}^R(P)$ be the runs in $\mathcal{R}(P)$ such that $\mathit{world}(r,m) = \mathit{world}(r,0)$ for all $m \geq 0$.

All that remains to define a plausibility system is to define the plausibility assignment function $\mathcal{P}$. We take $\mathcal{S}^R$ to be the set of systems of the form $(\mathcal{R}^R(P), \pi, \mathcal{P})$ for some protocol P, in which $\mathcal{P}(r,m) = (\Omega_{(r,m)}, \preceq_{(r,m)})$ satisfies the following conditions:

- $\Omega_{(r,m)} = \emptyset$ if $\mathit{obs}(r,m)$ is inconsistent; otherwise $\Omega_{(r,m)} = \{(r',m') : (r,m) \sim_a (r',m')\}$, the set of all points the agent considers possible.
- $\preceq_{(r,m)}$ is a ranking, i.e., for any two point $(r',m'), (r'',m'') \in \Omega_{(r,m)}$, either $(r',m') \preceq_{(r,m)} (r'',m'')$ or $(r'',m'') \preceq_{(r,m)} (r',m')$.
- $\preceq_{(r,m)}$ compares points examining only the state of the world, so that if (r',m') and (r'',m'') are in $\Omega_{(r,m)}$ and $\mathit{world}(r',m') = \mathit{world}(r'',m'')$, then (r',m') and (r'',m'') are equivalent according to $\preceq_{(r,m)}$.
- if $\mathit{obs}(r,m)) = \varphi$ is consistent, then $(r', m'+1)$ is a $\preceq_{(r,m+1)}$-minimal point if and only if (r',m') is a $\preceq_{(r,m)}$-minimal point satisfying φ.

Note that our assumptions correspond closely to those of [Gro88, KM91b]. The difference is that we have time explicitly in the picture and that our states have more structure. That is, in [Gro88, KM91b], a state is just a truth assignment. For us, the truth assignment is still there, as part of the environment's state, but we have also added the agent's local state. Of course, we can associate a belief set with each local state, since an agent's local state determines his beliefs over $\mathcal{L}(\Phi)$. That is, if $r_a(m) = r'_a(m)$, then it is easy to check that $(\mathcal{I}, r, m) \models B\varphi$ if and only if $(\mathcal{I}, r', m') \models B\varphi$ for any $\varphi \in \mathcal{L}(\Phi)$. Thus, we can write $(\mathcal{I}, s_a) \models B\varphi$, where s_a is the agent's local state $r_a(m)$. Define the belief set $\mathrm{Bel}\,(\mathcal{I}, s_a)$ to be $\{\varphi \in \mathcal{L}(\Phi) : (\mathcal{I}, s_a) \models B\varphi\}$. It is easy to show that every AGM revision operator can be represented in our framework. Recall that we sometimes write $s_a \cdot \varphi$ for $\tau(s_a, \varphi)$.

Theorem 4.1: *Let $\circ$ be an AGM revision operator. There is a system $\mathcal{I}_\circ \in \mathcal{S}^R$ such that for all $\psi \in \mathcal{L}(\Phi)$, we have*

$$\mathit{Bel}\,(\mathcal{I}_\circ, s_a) \circ \psi = \mathit{Bel}\,(\mathcal{I}_\circ, s_a \cdot \psi). \qquad (1)$$

What about the converse? That is, given a system $\mathcal{I} \in \mathcal{S}^R$, can we define a belief revision operator $\circ_{\mathcal{I}}$ on belief sets such that (1) holds? The answer is no. In general, $\circ_{\mathcal{I}}$ would not be well defined: It is not hard to find a system $\mathcal{I} \in \mathcal{S}^R$ and two local states s_a and s'_a such that $\mathrm{Bel}\,(\mathcal{I}, s_a) = \mathrm{Bel}\,(\mathcal{I}, s'_a)$, but $\mathrm{Bel}\,(\mathcal{I}, s_a \cdot \psi) \neq \mathrm{Bel}\,(\mathcal{I}, s'_a \cdot \psi)$. That is, the agent can believe exactly the same propositional formulas at two points in $\mathcal{I}$ and yet revise his beliefs differently at those points. Our framework makes a clear distinction between the agents' belief state and his local state, which we can identify with his epistemic state. In any $S \in \mathcal{S}^R$, the agent's belief set does not determine how the agent's beliefs will be revised; his local state does.

We could put further restrictions on $\mathcal{S}^R$ to obtain only systems in which the agent's belief state determines how his beliefs are revised. That is, we could consider only systems $\mathcal{I}$, where $\mathrm{Bel}\,(\mathcal{I}, s_a \cdot \varphi) = \mathrm{Bel}\,(\mathcal{I}, s'_a \cdot \varphi)$ whenever $\mathrm{Bel}\,(\mathcal{I}, s_a) = \mathrm{Bel}\,(\mathcal{I}, s'_a)$. If we restrict to such systems, we can obtain a converse to Theorem 4.1, but this seems to us the wrong way to go. We believe it is inappropriate to equate belief sets with epistemic states in general. For example, the agent's local state determines his plausibility ordering, but his belief set does not. Yet surely how an agent revises his beliefs is an important part of his epistemic state.

We believe that there are two more appropriate ways to deal with this problem. The first is to modify the AGM postulates to deal with epistemic states, not belief sets. The second is to enrich the language to allow richer belief sets. We deal with these one at a time.

We can easily modify the AGM postulates to deal with epistemic states. We now assume that we start with a space of abstract epistemic states, $\circ$ maps an epistemic state and a formula to a new epistemic state, and Bel maps epistemic states to belief sets. We then have analogues to each of the AGM postulates, obtained by replacing each belief set by the beliefs of the corresponding epistemic state. For example, we have:

(R1′) $E \circ \varphi$ is an epistemic state

(R2′) $\varphi \in \mathrm{Bel}\,(E \circ \varphi)$
(R3′) $\mathrm{Bel}\,(E \circ \varphi) \subseteq Cl(\mathrm{Bel}\,(E) \cup \{\varphi\})$

and so on, with the obvious transformation.[7]

There is a clear correspondence between systems in our framework and belief revision functions that use abstract epistemic states. Using this correspondence we show that $\mathcal{S}^R$ captures belief revision according to R1′–R8′:

Theorem 4.2: *$\mathcal{I} \in \mathcal{S}^R$ if and only if the corresponding belief revision function satisfies R1′–R8′.*

As we said earlier, there is a second approach to dealing with this problem: extending the language. We defer discussion of this approach to Section 6.

Our representation brings out several issues. The revision literature usually does not address the relations between the agent's beliefs and the "real" world. (This point is explicitly discussed in [Gär88, pp 18–20].) In fact, revision does not assume any correlation between what the agent learns and the state of the world. For example, revision allows the agent to learn (revise by) φ and then learn $\neg\varphi$. Moreover, after learning φ the agent may consider worlds where $\neg\varphi$ is true as quite plausible (although he will not consider them to be the most plausible worlds). In this case, most observations that are not consistent with his beliefs will lead him to believe $\neg\varphi$. These examples are two aspects of a bigger problem: The AGM postulates put very weak restrictions on the ordering that the agent has after a revision step (see [Bou93, DP94]). Essentialy, the only requirement is that after learning φ, the most plausible worlds must be ones where φ is true. While this is an important and reasonable constraint on how beliefs should change, it does not capture all our intuitions regarding how beliefs change in many applications. We believe that by introducing more structure it should be possible to derive reasonable constraints that will make revision a more useful tool.

5 UPDATE

The notion of update originated in the database community [KW85, Win88]. The problem is how a knowledge base should change when something is learned about world, such as "A table was moved from office 1 to office 2". Katsuno and Mendelzon [KM91a] suggest a set of postulates that any update operator should satisfy.

The update postulates are expressed in terms of formulas, not belief sets. This is not unreasonable, since we can identify a formula φ with the belief set $Cl(\varphi)$. Indeed, if Φ is finite (which is what Katsuno and Mendelzon assume) every belief set A can be associated with some formula φ_A such that $Cl(\varphi) = A$; we denote this formula $desc(A)$.

(U1) $\vdash_{\mathcal{L}} \varphi \diamond \mu \Rightarrow \mu$
(U2) If $\vdash_{\mathcal{L}} \varphi \Rightarrow \mu$, then $\vdash_{\mathcal{L}} \varphi \diamond \mu \Leftrightarrow \varphi$
(U3) $\vdash_{\mathcal{L}} \neg\varphi \diamond \mu$ if and only if $\vdash_{\mathcal{L}} \neg\varphi$ or $\vdash_{\mathcal{L}} \neg\mu$
(U4) If $\vdash_{\mathcal{L}} \varphi_1 \Leftrightarrow \varphi_2$ and $\vdash_{\mathcal{L}} \mu_1 \Leftrightarrow \mu_2$ then $\vdash_{\mathcal{L}} \varphi_1 \diamond \mu_1 \Leftrightarrow \varphi_2 \diamond \mu_2$
(U5) $\vdash_{\mathcal{L}} (\varphi \diamond \mu) \wedge \theta \Rightarrow \varphi \diamond (\mu \wedge \theta)$
(U6) If $\vdash_{\mathcal{L}} \varphi \diamond \mu_1 \Rightarrow \mu_2$ and $\vdash_{\mathcal{L}} \varphi \diamond \mu_2 \Rightarrow \mu_1$, then $\vdash_{\mathcal{L}} \varphi \diamond \mu_1 \Leftrightarrow \varphi \diamond \mu_2$
(U7) If φ is complete then $\vdash_{\mathcal{L}} (\varphi \diamond \mu_1) \wedge (\varphi \diamond \mu_2) \Rightarrow \varphi \diamond (\mu_1 \vee \mu_2)$[8]
(U8) $\vdash_{\mathcal{L}} (\varphi_1 \vee \varphi_2) \diamond \mu \Leftrightarrow (\varphi_1 \diamond \mu) \vee (\varphi_2 \diamond \mu)$.

Update tries to capture the intuition that there is a preference for runs where all the observations made are true, and where changes from one point to the next along the run are minimized. To capture the notion of "minimal change from world to world", we use a *distance function* d on worlds.[9] Given two worlds w and w', $d(w, w')$ measures the distance between them. Distances might be incomparable, so we require that d maps pairs of worlds into a *partially ordered* domain with a unique minimal element 0 and that $d(w, w') = 0$ if and only if $w = w'$.

We can now describe how update is captured in our framework. The construction is very similar to the one we used for revision. The major difference is in how the preorders are constructed. For our discussion it is enough to consider agents that follow the simple protocol P^* of Section 3. We take $\mathcal{S}^U$, the set of plausibility systems for update, to consist of all systems of the form $(\mathcal{R}(P^*), \pi, \mathcal{P}_d)$, where $\mathcal{P}_d$ is determined by a distance function d in a manner we now describe. Recall that update has a preference for runs where the observations are all true. We say that a point (r, m), is *consistent* if $obs(r, j)$ is true in the world $world(r, j{+}1)$, for $0 \le j < m$. We take $\mathcal{P}(r, m) = (\Omega_{(r,m)}, \preceq_{(r,m)})$, where $\Omega_{(r,m)}$ consists of all points that the agent considers possible that are consistent, and $\preceq_{(r,m)}$ is a preorder defined as follows: suppose $(r', m), (r'', m) \in \Omega_{(r,m)}$.[10] Roughly speaking, we prefer (r', m) to (r'', m) if, at the first point where they differ, r' makes the smaller change. Formally, if $n > 0$ is the first point where r'

[7] The only problematic postulate is R6. The question is whether R6′ should be "If $\vdash_{\mathcal{L}} \varphi \Leftrightarrow \psi$ then $\mathrm{Bel}\,(E \circ \varphi) = \mathrm{Bel}\,(E \circ \psi)$" or "If $\vdash_{\mathcal{L}} \varphi \Leftrightarrow \psi$ then $E \circ \varphi = E \circ \psi$". Dealing with either version is straightforward. For definiteness, we use the first definition here.

[8] A belief set A is *complete* when for every $\varphi \in \mathcal{L}(\Phi)$ either $\varphi \in A$ or $\neg\varphi \in A$. A formula φ is *complete* if $Cl(\varphi)$ is complete.

[9] Katsuno and Mendelzon identify a "world" with a truth assignment to the primitive propositions. For us, this is just a component of the environment state.

[10] Note that the definition of P^* implies that if $(r, m) \sim_a (r', m')$ then $m = m'$ since the agent's local state encodes the time m by the length of the sequence.

and r'' have different world states (i.e., the first point where $world(r', n) \neq world(r'', n)$) then $(r', m) \prec_{(r,m)} (r'', m)$ if and only if $d(world(r', n-1), world(r', n)) < d(world(r'', n-1), world(r'', n))$. (Note that this definition is independent of (r, m).) Thus, update focuses on the *first* point of difference. The run that makes the smaller change at that point is preferred, even if later it makes quite abnormal changes. This point is emphasized in the example below. However, we first show that $\mathcal{S}^U$ captures all possible update operators.

Theorem 5.1: *$\diamond$ is a KM update operator if and only if there is a system $\mathcal{I}_\diamond \in \mathcal{S}^U$ such that for all $\psi \in \mathcal{L}(\Phi)$, we have $desc(Bel(\mathcal{I}_\diamond, s_a)) \diamond \psi \Leftrightarrow desc(Bel(\mathcal{I}_\diamond, s_a \cdot \psi))$.*

Notice that for update, unlike revision, the systems we consider are such that the belief state does determine the result of the update, i.e., if $B(\mathcal{I}, s_a) = B(\mathcal{I}, s'_a)$, then for any φ we get that $B(\mathcal{I}, s_a \cdot \varphi) = B(\mathcal{I}, s'_a \cdot \varphi)$.

How reasonable is the notion of update? As the definition of $\preceq_{(r,m)}$ given above suggests, it has a preference for deferring abnormal events. This makes it quite similar to Shoham's *chronological ignorance* [Sho88] (a point already noted by del Val and Shoham [dVS92, dVS93]), and it suffers from some of the same problems. Consider the following story, that we call the *borrowed-car example.*[11] (1) The agent leaves his car in a valet parking lot, (2) sits for an hour in a cafe, (3) returns to the car and starts driving home. Since the agent does not observe the car while he is in the cafe, there is no reason for him to revise his beliefs regarding the car's location. Since he finds it in the parking lot at step (3), he still has no reason to change his beliefs. Now, what should he believe when (4) he notices, during his drive, that the car has been driven 50 miles since he left home? The common sense explanation is that the valet took the car out for a joy ride. But update prefers to defer abnormalities, so it will conclude that the mileage must have jumped, for inexplicable reasons, since he left the parking lot. To see this, note that runs where the valet took the car have an abnormality at time (2), while runs where the car did not move at time (2) but the mileage suddenly changed, have their first abnormality at time (4) and thus are preferred! (See Figure 1.)

We emphasize that the counterintuitive conclusion drawn in this example is not an artifact of our representation, but inherent in the definition of update. We can formalize the example using propositions such as *car-in-lot, high-mileage,* etc. The observation of *high-mileage* at step (4) must be explained by some means, and an update operator will explain it in terms of a change that occurred in states consistent with the beliefs at step (3) (i.e., *car-in-lot, $\neg$high-mileage*). The exact change assumed will depend on the distance function embodied by the update operator. The key point is that update will not go back and revise the earlier beliefs about what happened between steps (1) and (2).

In an effort to understand the difficulty here, we look at the belief change process more generally. In a world w, the agent has some beliefs that are described by, say, the formula φ. These beliefs may or may not be *correct* (where we say a belief φ is correct in a world w if φ is true of w). Suppose something happens and the world changes to w'. As a result of the agent's observations, he has some new beliefs, described by φ'. Again, there is no reason to believe that φ' is correct. Indeed, it may be quite unreasonable to expect φ' to be correct, even if φ is correct. Consider the borrowed-car example. Suppose that while the agent was sitting in the cafe, the valet did in fact take the car out for a joy ride. Nevertheless, the most reasonable belief for the agent to hold when he observes that the car is still in the parking lot after he leaves the cafe is that it was there all along.

The problem here is that the information the agent obtains at steps (2) and (3) is insufficient to determine what happened. We cannot expect all the agent's beliefs to be correct at this point. On the other hand, if he does obtain sufficient information about the change and his beliefs were initially correct, then it seems reasonable to expect that his new beliefs will be correct. But what counts as *sufficient* information? In the context of update, we can provide a precise formulation.

We say that φ provides *sufficient information* about the change from w to w' if there is no world w'' satisfying φ such that $d(w, w'') < d(w, w')$. In other words, φ is sufficient information if, after observing φ in world w, the agent will consider the real world (w') one of the most likely worlds. Note that this definition is monotonic, in that if φ is sufficient information about the change then so is any formula ψ that implies φ (as long as it holds at w'). Moreover, this definition depends on the agent's distance function d. What constitutes sufficient information for one agent might not for another. We would hope that the function d is realistic in the sense that the worlds judged closest according to d really are the most likely to occur.

We can now show that update has the property that if the agent had correct beliefs and receives sufficient information about a change, then he will continue to have correct beliefs.

Theorem 5.2: *Let $\mathcal{I} \in \mathcal{S}^U$. If the agent's beliefs at (r, m) are correct and $obs(r, m)$ provides sufficient information about the change from $world(r, m)$ to $world(r, m+1)$, then the agent's beliefs at $(r, m+1)$ are correct.*

As we observed earlier, we cannot expect the agent to always have correct beliefs. Nevertheless, it seems

[11] This example is based on Kautz's stolen car story [Kau86], and is due to Boutilier, who independently observed this problem [private communication, 1993].

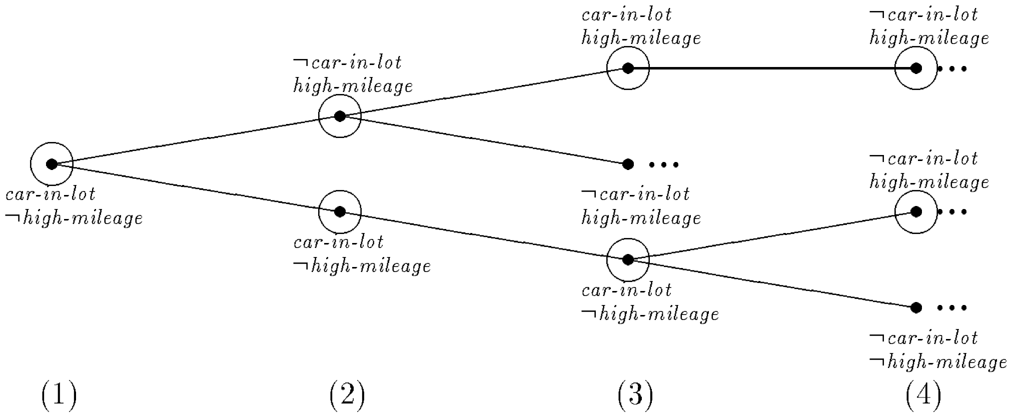

Figure 1: Runs in the borrowed car example. Lower branches are considered more likely than higher ones. The circles mark points that are consistent with the agent's observations at each time step.

reasonable to require that if the agent does (eventually) receive sufficiently detailed information, then he should realize that his beliefs were incorrect. This is precisely what does *not* happen in the borrowed-car example. Intuitively, once the agent observes that the car was driven 50 miles, this should be sufficient information to eliminate the possibility that the car remained in the parking lot. Roughly speaking, because update focuses only on the current state of the world, it cannot go back and revise beliefs about the past.

How can we capture the intuition of a sequence of observations providing sufficient information about the changes that have occurred? Here it is most convenient to take full advantage of our framework with runs. Intuitively, we say that a sequence of observations provides sufficient information about the changes that occurred if, after observing the sequence, the agent will consider as most plausible runs where the real changes occurred. More precisely, we say that a sequence of observations $\varphi_1, \ldots, \varphi_n$ *provides sufficient information about the sequence of changes* $w_0, \ldots, w_n$ if for any run r such that $\mathit{world}(r,i) = w_i$ and $\mathit{obs}(r, i-1) = \varphi_i$ for $i = 1, \ldots, n$, there does not exist a run r' such that $(r,n) \sim_a (r',n)$ and $(r',n) \preceq_{(r,n)} (r,n)$. This definition is a natural generalization of the definition of sufficient information about a single change. We would like to state a theorem similar to Theorem 5.2, i.e., that if the agent has correct beliefs at (r,m) and receives sufficient information about the changes from (r,m) to $(r,m+n)$, then the agent's beliefs at $(r,m+n)$ are correct. However, the problem with update is that once the agent has incorrect beliefs, *no* sequence of observations can ever provide him with sufficient information about the changes that have occurred. More precisely, once the agent has incorrect beliefs, no sequence can satisfy the technical requirements we describe. This is in contrast to our intuition that some sequences (as in the example above) should provide sufficient information about the changes. Note, that this result does not imply that once the agent has incorrect beliefs then he continues to have incorrect beliefs. It is possible that he regains correctness after several observation (for example if the agent is told the exact state of the world). However, it is always possible to construct examples (like the one above) where the agent receives sufficient information about the changes in the rest of the run, and yet has incorrect beliefs in the rest of the run.

Our discussion of update shows that update is guaranteed to be safe only in situations where there is always enough information to characterize the change that has occurred. While this may be a plausible assumption in database applications, it seems somewhat less reasonable in AI examples, particularly cases involving reasoning about action.[12] In Section 7, we discuss how update might be modified to take this observation into account.

6 SYNTHESIS

In previous sections we analyzed belief revision and belief update separately. We provided representation theorems for both notions and discussed issues that are specific to each notion. In this section we try to identify some common themes and points of difference.

Some of the work has already been done for us by Katsuno and Mendelzon [KM91a], who identified three significant differences between revision and update:

1. Revision deals with static propositions, while update allows propositions that are not static. As we

[12] Similar observations were independently made by Boutilier [Bou94], although his representation is quite different than ours.

noted in the introduction this difference is in the types of propositions that these notions deal with, rather than a difference in the type of situations that they deal with.

2. Revision and update treat inconsistent belief states differently. Revision allows the agent to "recover" from an inconsistent state after observing a consistent formula. Update dictates that once the agent has inconsistent beliefs, he will continue to have inconsistent beliefs.
3. Revision considers only total preorders, while update allows partial preorders.

How significant are these differences? While the restriction to static propositions may seem to be a serious limitation of belief revision, notice that we can always convert a dynamic proposition to a static one by adding "time-stamps". That is, we can replace a proposition p by a family of propositions p^m that stand for "p is true at time m". Thus, it is possible to use revision to reason about a changing world. (Of course, it would then be necessary to capture connections between propositions of the form p^m, but specific revision operators could certainly do this.)

As far as the other differences go, we can get a better understanding of them, and of the relationship between revision and update, if we return to the general belief change systems described in Section 3 and use a language that allows us to explicitly reason about the belief change process. Although this involves a somewhat extended exposition, we hope the reader will agree that the payoff is worthwhile.

The language we consider for reasoning about the belief change process is called $\mathcal{L}^{>}(\Phi)$. It uses two modal operators, a binary modal operator $>$ to capture belief change and a unary modal operator B that captures belief, just as before. The formula $\varphi{>}\psi$ is a *Ramsey conditional.* It is intended to capture the *Ramsey test*: we want $\varphi{>}\psi$ to hold if ψ holds in the agent's epistemic state after he learns φ.

Formally, we take $\mathcal{L}^{>}(\Phi)$ be the least set of formulas such that if $\varphi \in \mathcal{L}(\Phi)$ and $\psi, \psi' \in \mathcal{L}^{>}(\Phi)$ then $B\varphi$, $B\psi$, $\neg\psi$, $\psi \wedge \psi'$, and $\varphi{>}\psi$ are in $\mathcal{L}^{>}(\Phi)$. All formulas in $\mathcal{L}^{>}(\Phi)$ are *subjective*, that is, their truth is determined by the agent's epistemic state. In particular, this means that $\mathcal{L}(\Phi)$ is not a sublanguage of $\mathcal{L}^{>}(\Phi)$. For example, the primitive proposition p is not in $\mathcal{L}^{>}(\Phi)$ (although Bp is). Since formulas in $\mathcal{L}^{>}(\Phi)$ are subjective, this means that all formulas on the right-hand side of $>$ are subjective. This seems reasonable, since we intend these formulas to represent the beliefs of the agent after learning. On the other hand, notice that the only formulas that can appear on the left-hand side of $>$ are formulas in $\mathcal{L}(\Phi)$. This is because only formulas in $\mathcal{L}(\Phi)$ can be learned.[13]

We give formulas in $\mathcal{L}^{>}(\Phi)$ semantics in interpreted plausibility systems. The semantics for $B\varphi$ is just as it was before, so all we need to do is define the semantics for $\varphi{>}\psi$. Notice that since each formula in $\mathcal{L}^{>}(\Phi)$ is subjective, its truth depends just on the agent's epistemic state. We can give the following natural definition for conditionals with respect to epistemic states, based on our desire to have them satisfy the Ramsey test.

$$(\mathcal{I}, s_a) \models \varphi{>}\psi \text{ if } (\mathcal{I}, s_a \cdot \varphi) \models \psi.$$

As expected, we then define $(\mathcal{I}, r, m) \models \varphi{>}\psi$ if and only if $(\mathcal{I}, r_a(m)) \models \varphi{>}\psi$.

The language $\mathcal{L}^{>}(\Phi)$ is actually a fragment of $\mathcal{L}^{KPT}(\Phi)$. As the following lemma shows, we can express Ramsey conditionals by using the modal operators for time and knowledge.

Lemma 6.1: *Let $\mathcal{I}$ be a belief change system, let $\varphi \in \mathcal{L}(\Phi)$ and let $\psi \in \mathcal{L}^{>}(\Phi)$. Then*

$$(\mathcal{I}, r, m) \models \varphi{>}\psi \Leftrightarrow K(\mathit{learn}(\varphi) \Rightarrow \bigcirc\psi).$$

Despite Lemma 6.1, there is a good reason to consider $\mathcal{L}^{>}(\Phi)$ rather than $\mathcal{L}^{KPT}(\Phi)$. As we now show, it is the "right" language for capturing the belief change process. Suppose we consider belief sets over the language $\mathcal{L}^{>}(\Phi)$ rather than $\mathcal{L}(\Phi)$. In analogy to our definition of Bel , define $\mathrm{Bel}^{>}(\mathcal{I}, s_a) = \{\varphi \in \mathcal{L}^{>}(\Phi) \mid (\mathcal{I}, s_a) \models \varphi\}$.[14] However, this lead to technical We define an *extended belief set* to be any set of the form $\mathrm{Bel}^{>}(\mathcal{I}, s_a)$. The following lemma shows the extended belief set captures exactly the epistemic state of the agent with regard to belief change.

Lemma 6.2: *If $\mathcal{I}, \mathcal{I}'$ are belief change systems, then $Bel^{>}(\mathcal{I}, s_a) = Bel^{>}(\mathcal{I}', s'_a)$ if and only if for every sequence $\varphi_1, \ldots, \varphi_n$ it is the case that*

$$Bel(\mathcal{I}, s_a \cdot \varphi_1 \cdot \ldots \cdot \varphi_n) = Bel(\mathcal{I}', s'_a \cdot \varphi_1 \cdot \ldots \cdot \varphi_n).$$

This implies that if two states have the same extended belief set, then they cannot be distinguished by the belief change process.

We can now define the obvious belief change operation on extended belief sets in terms of the Ramsey test:

$$E \cdot \varphi =_{\mathrm{def}} \{\psi \mid \varphi{>}\psi \in E\}, \qquad (2)$$

for an extended belief sets E. Thus, $\cdot$ maps an extended belief set and a formula to an extended belief

[13] This type of a right-nested language is also considered, for similar reasons, in [Bou93, EG93].

[14] We might have defined $\mathrm{Bel}^{>}(\mathcal{I}, s_a)$ as $\{\varphi \in \mathcal{L}^{>}(\Phi) \mid (\mathcal{I}, s_a) \models B\varphi\}$, which would have been even more in the spirit of our definition of $\mathrm{Bel}\,(\mathcal{I}, s_a)$. This definition agrees with our definition except when $(\mathcal{I}, s_a) \models B(\mathit{false})$. In this case, our definition does not put all formulas of the form $\varphi{>}\psi$ into the belief set, which seems to us the more appropriate behavior.

set. We have deliberately used the same notation here as for the mapping $\cdot$ ¿From local states and formulas to local states. The following lemma shows that these two mappings are related in the expected way:

Lemma 6.3: *If $\mathcal{I}$ be a belief change system and $\varphi \in \mathcal{L}(\Phi)$, then $Bel^{>}(\mathcal{I}, s_a \cdot \varphi) = Bel^{>}(\mathcal{I}, s_a) \cdot \varphi$.*

Although the proof of this result is easy, it has important implications. It shows that we have a well-defined notion of belief change on extended belief sets. Thus, it can be viewed as another way of solving the problem raised in Section 4. If we consider systems in $\mathcal{S}^R$, then extended belief sets, unlike belief sets, uniquely determine the outcome of revision.

Our notion of extended belief sets is similar to a notion introduced by Gärdenfors [Gär78, Gär86, Gär88]. He also considers revision of belief sets that contain conditionals of the form $\varphi > \psi$. However, he attempts to apply the AGM revision postulates to these sets (and then obtains his well known triviality result), while we define revision of extended belief sets directly in terms of the Ramsey test. Of course, our notion of belief revision does not satisfy the AGM postulates (although it does when restricted to $\mathcal{L}(\Phi)$). Indeed, we cannot expect it to, given Gärdenfors' triviality result. As we argued in the introduction, this should not be viewed as a defect. We do not want persistence (i.e., R4) to hold for formulas such as Ramsey conditionals whose truth depends on the current state of the agent. Indeed, as we argue in the full paper, other postulates, such as R8, should also not hold for conditional beliefs.

Once we adopt the Ramsey test we can in fact discard postulates R1–R8 altogether, and simply define revision using Eq. (2). That is, we shift the focus from finding a set of postulates for belief revision, as done by previous researchers [AGM85, Gär88, Bou92, Fuh89, Rot89], to that of finding a logical characterization of revision in terms of the properties of $>$.

In [FH94a], it is shown that this general approach of characterizing belief change in terms of characterizing the behavior of the $>$ operator in a class of plausibility structures is relevant for all reasonable notions of belief change, not just belief revision. In particular, this is the case for belief update. The results of [FH94a] enable us to completely characterize the differences between revision and update axiomatically. There is an axiom that holds for revision (and not for update) that captures the fact that revision focuses on total preorders, there is an axiom that holds for update (and not for revision) that captures the intuition that update works "pointwise", and there is an axiom for update that must be weakened slightly for revision because revision can recover from inconsistencies. We refer the reader to [FH94a] for further details.

Thinking in terms of $>$ helps us see connections between revision and update beyond those captured in our axioms. For one thing, it helps us make precise the intuition that both revision and update are characterized by the plausibility ordering at each state. In an arbitrary belief change system, there need be no connections between an agent's beliefs before and after observing a formula φ. We say that the belief change at a point (r, m) in a belief change system is *compatible with the plausibility ordering* if for any $\varphi, \psi \in \mathcal{L}(\Phi)$, we have $(\mathcal{I}, r, m) \models \varphi > B\psi$ if and only if $(\mathcal{I}, r, m) \models \bigcirc\varphi \rightarrow \bigcirc\psi$. That is, the agent believes ψ after learning φ exactly when ψ is true at the next time step in all the most plausible situations in which φ is true at the next time step. The next theorem shows that belief change is compatible with the plausibility ordering in both systems for revision and update (except that in revision, belief change is not compatible with the agent's plausibility ordering at states where the agent's beliefs are inconsistent; this is due to the fact that an agent with inconsistent beliefs may have consistent beliefs again after revision).

Theorem 6.4: *If $\mathcal{I} \in \mathcal{S}^R$ then belief change is compatible with the plausibility ordering at every point (r, m) such that $(\mathcal{I}, r, m) \not\models B(false)$. If $\mathcal{I} \in \mathcal{S}^U$ then belief change is compatible with the plausibility ordering at every point.*

We note that since propositions do not change their values in $\mathcal{S}^R$ we get the following corollary.

Corollary 6.5: *Let $\mathcal{I} \in \mathcal{S}^R$, let $\varphi, \psi \in \mathcal{L}(\Phi)$, and let (r, m) be a point in $\mathcal{I}$ such that $(\mathcal{I}, r, m) \not\models B(false)$. Then $(\mathcal{I}, r, m) \models (\varphi > B\psi) \Leftrightarrow (\varphi \rightarrow \psi)$.*

Consistency provides some connection between $\rightarrow$ and $>$. For example, as this corollary shows, in revision systems they are essentially identical for depth-one formulas. In general, however, they are different. For example, it is not hard to show that $p > (true > Bq)$ is not equivalent to $p \rightarrow (true \rightarrow q)$ in revision systems. Indeed, as noted above while revision guarantees minimal change in propositional (or base) beliefs, it does not put any such restrictions on changes in the ordering at each epistemic state. Thus, there is no necessary connection between $\rightarrow$ and iterated instances of $>$.[15] In our representation of update, on the other hand, we can make such a connection. Indeed, in $\mathcal{S}^U$, $\rightarrow$ completely captures the behavior of $>$.

Lemma 6.6: *Let $\mathcal{I} \in \mathcal{S}^U$. For all sequences of formulas $\varphi_1, \ldots, \varphi_n \in \mathcal{L}(\Phi)$ and all $\psi \in \mathcal{L}(\Phi)$, $(\mathcal{I}, r, m) \models (\varphi_1 > \cdots > \varphi_n > B\psi) \Leftrightarrow ((\bigcirc\varphi_1 \land \cdots \land \bigcirc^n \varphi_n) \rightarrow \bigcirc^n \psi)$.*

This result can be explained by the fact that in update systems, the plausibility ordering at $s_a \cdot \varphi$ is determined by the plausibility ordering at s_a. More precisely, after learning φ, $\Omega_{(r,m+1)} = \{(r, m +$

[15] Thus, we use $>$ rather than $\rightarrow$ for belief revision, contrary to the suggestion implicit in [Bou92].

$1)|(r,m) \in \Omega_{(r,m)}, (r,m) \models \mathit{learn}(\varphi) \wedge \bigcirc\varphi\}$, and $(r',m+1) \preceq_{(r,m+1)} (r'',m+1)$ if and only if $(r',m) \preceq_{(r,m)} (r'',m)$. (In the terminology of [FH94b], this means that the plausibility space $(r, m+1)$ can be understood as the result of conditioning on the plausibility space at (r,m).)

These results, and those of [FH94a], support the thesis that this language, which lets us reason about plausibility (and thus belief), belief change, time, and knowledge, is the right one with which to study belief change.

7 CONCLUSION

We believe that our framework, with its natural representation of both time and belief, gives us a great deal of insight into belief revision and belief update. Of course, revision and update are but two points in a wide spectrum of possible types of belief change. Our ultimate goal is to use this framework to understand the whole spectrum better and to help us design belief change operations that overcome some of the difficulties we have observed with revision and update. In particular, we want belief change operations that can handle dynamic propositions, while still being able to revise information about the past.

One approach to doing this, much in the spirit of update, would be to use a distance function that relates not just worlds, but sequences of worlds (of the same length). We could then easily modify the definition of update so as to handle the borrowed-car problem correctly. Such a modification, however, comes at a cost. It is much simpler to represent (and do computations with) a distance function that applies to worlds than to sequences of worlds. A natural question to ask is whether we can get away with a simpler distance function (that, perhaps, considers only certain features of the sequence of worlds, rather than the sequence itself). Of course, the answer to this will depend in part on how we make "get away with" precise.

Whether or not this particular approach turns out to be a useful one, it is clear that these are the types of questions we should be asking. We hope that our framework will provide a useful basis for answering them.

Finally, we note that our approach is quite different from the traditional approach to belief change [AGM85, Gär88, KM91a]. Traditionally, belief change was considered as an abstract process. Our framework, on the other hand, models the agent and the environment he is situated in, and how both change in time. This allows us to model concrete agents in concrete settings (for example, diagnostic systems are analyzed in [FH94b]), and to reason about the beliefs and knowledge of such agents. We can then investigate what plausibility ordering induces beliefs that match our intuitions. By gaining a better understanding of such concrete situations, we can better investigate more abstract notions of belief change.

Acknowledgements

The authors are grateful to Craig Boutilier, Ronen Brafman, Adnan Darwiche, Moises Goldszmidt, Adam Grove, Phil Hubbard Alberto Mendelzon, Alvaro del Val, and particularly Daphne Koller and Moshe Vardi, for comments on drafts of this paper and useful discussions relating to this work.

References

[AGM85] C. E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: partial meet functions for contraction and revision. *Journal of Symbolic Logic*, 50:510–530, 1985.

[Bou92] C. Boutilier. Normative, subjective and autoepistemic defaults: Adopting the Ramsey test. In *Principles of Knowledge Representation and Reasoning: Proc. Third International Conference (KR '92)*. 1992.

[Bou93] C. Boutilier. Revision sequences and nested conditionals. In *Proc. Thirteenth International Joint Conference on Artificial Intelligence (IJCAI '93)*, pages 519–525, 1993.

[Bou94] C. Boutilier. An event-based abductive model of update. In *Proc. Tenth Biennial Canadian Conference on Artificial Intelligence (AI '94)*, 1994.

[Bur81] J. Burgess. Quick completeness proofs for some logics of conditionals. *Notre Dame Journal of Formal Logic*, 22:76–84, 1981.

[DP94] A. Darwiche and J. Pearl. On the logic of iterated belief revision. In R. Fagin, editor, *Theoretical Aspects of Reasoning about Knowledge: Proc. Fifth Conference*, pages 5–23. 1994.

[dVS92] A. del Val and Y. Shoham. Deriving properties of belief update from theories of action. In *Proc. National Conference on Artificial Intelligence (AAAI '92)*, pages 584–589, 1992.

[dVS93] A. del Val and Y. Shoham. Deriving properties of belief update from theories of action (II). In *Proc. Thirteenth International Joint Conference on Artificial Intelligence (IJCAI '93)*, pages 732–737, 1993.

[dVS94] A. del Val and Y. Shoham. A unified view of belief revision and update. *Journal of Logic and Computation*, Special issue on Actions and Processes, to appear, 1994.

[EG93] T. Eiter and Gottlob G. The complexity of nested counterfactuals and iterated knowledge base revisions. In *Proc. Thirteenth International Joint Conference on Artificial Intelligence (IJCAI '93)*, pages 526–531, 1993.

[FH94a] N. Friedman and J. Y. Halpern. Conditional logics of belief change. Technical report, 1994. Submited to AAAI-94.

[FH94b] N. Friedman and J. Y. Halpern. A knowledge-based framework for belief change. Part I: Foundations. In R. Fagin, editor, *Theoretical Aspects of Reasoning about Knowledge: Proc. Fifth Conference*, pages 44–64. 1994.

[FHMV94] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge.* MIT Press, to appear, 1994.

[Fuh89] A. Fuhrmann. Reflective modalities and theory change. *Synthese*, 81:115–134, 1989.

[Gär78] P. Gärdenfors. Conditionals and changes of belief. *Acta Philosophica Fennica*, 20, 1978.

[Gär86] P. Gärdenfors. Belief revision and the Ramsey test for conditionals. *Philosophical Review*, 91:81–93, 1986.

[Gär88] P. Gärdenfors. *Knowledge in Flux.* Cambridge University Press, 1988.

[GM88] P. Gärdenfors and D. Makinson. Revisions of knowledge systems using epistemic entrenchment. In M. Y. Vardi, editor, *Proc. Second Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 83–95. 1988.

[GMR92] G. Grahne, A. Mendelzon, and R. Rieter. On the semantics of belief revision systems. In Y. Moses, editor, *Theoretical Aspects of Reasoning about Knowledge: Proc. Fourth Conference*, pages 132–142. 1992.

[GP92] M. Goldszmidt and J. Pearl. Rank-based systems: A simple approach to belief revision, belief update and reasoning about evidence and actions. In R. Parikh, editor, *Principles of Knowledge Representation and Reasoning: Proc. Third International Conference (KR '92)*, pages 661–672. 1992.

[Gro88] A. Grove. Two modelings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.

[HF89] J. Y. Halpern and R. Fagin. Modelling knowledge and action in distributed systems. *Distributed Computing*, 3(4):159–179, 1989.

[Kau86] H. A. Kautz. Logic of persistence. In *Proc. National Conference on Artificial Intelligence (AAAI '86)*, pages 401–405, 1986.

[KM91a] H. Katsuno and A. Mendelzon. On the difference between updating a knowledge base and revising it. In *Principles of Knowledge Representation and Reasoning: Proc. Second International Conference (KR '91)*, pages 387–394. 1991.

[KM91b] H. Katsuno and A. Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52(3):263–294, 1991.

[KS91] H. Katsuno and K. Satoh. A unified view of consequence relation, belief revision and conditional logic. In *Proc. Twelfth International Joint Conference on Artificial Intelligence (IJCAI '91)*, pages 406–412, 1991.

[KW85] A. M. Keller and M. Winslett. On the use of an extended relational model to handle changing incomplete information. *IEEE Transactions on Software Engineering*, SE-11(7):620–633, 1985.

[Lev88] I. Levi. Iteration of conditionals and the Ramsey test. *Synthese*, 76:49–81, 1988.

[Lew73] D. K. Lewis. *Counterfactuals.* Harvard University Press, 1973.

[LR92] S. Lindström and W. Rabinowicz. Belief revision, epistemic conditionals and the Ramsey test. *Synthese*, 91:195–237, 1992.

[Rij92] M. de Rijke. Meeting some neighbors. Research Report LP-92-10, University of Amsterdam, 1992.

[Rot89] H. Rott. Conditionals and theory change: revision, expansions, and additions. *Synthese*, 81:91–113, 1989.

[Rot91] H. Rott. Two methods of constructing contractions and revisions of knowledge systems. *Journal of Philosophical Logic*, 20:149–173, 1991.

[Sho88] Y. Shoham. Chronological ignorance: experiments in nonmonotonic temporal reasoning. *Artificial Intelligence*, 36:271–331, 1988.

[Win88] M. Winslett. Reasoning about action using a possible models approach. In *Proc. National Conference on Artificial Intelligence (AAAI '88)*, pages 89–93, 1988.