On the Semantics of Belief Revision Systems*†

Gösta Grahne‡, Alberto O. Mendelzon§, and Ray Reiter¶

Abstract

We consider belief revision operators that satisfy the Alchourrón-Gärdenfors-Makinson postulates, and present an epistemic logic in which, for any such revision operator, the result of a revision can be described by a sentence in the logic.

In our logic, the fact that the agent's set of beliefs is ϕ is represented by the sentence $\mathtt{O}\phi$, where $\mathtt{O}$ is Levesque's 'only know' operator. Intuitively, $\mathtt{O}\phi$ is read as 'ϕ is *all* that is believed.' The fact that the agent believes ψ is represented by the sentence $\mathtt{B}\psi$, read in the usual way as 'ψ is believed'. The connective $\diamond$ represents *update* as defined by Katsuno and Mendelzon. The revised beliefs are represented by the sentence $\mathtt{O}\phi \diamond \mathtt{B}\psi$. We show that for every revision operator that satisfies the *AGM* postulates, there is a model for our epistemic logic such that the beliefs implied by the sentence $\mathtt{O}\phi \diamond \mathtt{B}\psi$ in this model correspond exactly to the sentences implied by the theory that results from revising ϕ by ψ. This means that reasoning about changes in the agent's beliefs reduces to model checking of certain epistemic sentences.

The negative result in the paper is that this type of formal account of revision cannot be extended to the situation where the agent is able to reason about its beliefs. A fully introspective agent cannot use our construction to reason about the results of its own revisions, on pain of triviality.

1 Introduction

Consider an agent that holds a set of beliefs and is sometimes required to change these beliefs according to its perceptions or other inputs. The problem of *belief revision* is: what are rational (or desirable, or typical, or economical) ways of revising a set of beliefs

*A preliminary version of this paper appeared in: *Proceedings of the Fourth Conference on Theoretical Aspects of Reasoning about Knowledge*, Y. Moses (Ed.). Morgan Kaufman Publishers, San Mateo, Ca., 1992, pp. 132 –142

†This work was supported by the Natural Sciences and Engineering Research Council of Canada and by the Institute for Robotics and Intelligent Systems.

‡Department of Computer Science, P. O. Box 26 (Teollisuukatu 23), SF-00140 University of Helsinki, Finland. `grahne@cs.helsinki.fi`. Work performed while visiting University of Toronto, Department of Computer Science.

§Department of Computer Science, University of Toronto, Toronto, Canada, M5S 1A4. `mendel@db.toronto.edu`.

¶Department of Computer Science, University of Toronto, Toronto, Canada, M5S 1A4. `reiter@ai.toronto.edu`. Fellow of the Canadian Institute of Advanced Research.

given some input? These questions have been studied in philosophy [7], in statistics [15], in Artificial Intelligence [3], and in the theory of databases [5].

For concreteness, suppose, as in [16], that the agent's beliefs are represented by a propositional theory ϕ and that the new input is a propositional sentence ψ that must be incorporated into the theory. A growing body of work [2, 22, 24] takes as a departure point the *rationality postulates* proposed of Alchourrón, Gärdenfors and Makinson (the *AGM* postulates, [1]). These are rules that every adequate revision operator should be expected to satisfy. For example: the new fact ψ must be a consequence of the revised theory.

Katsuno and Mendelzon [17] reason that no such set of postulates will be adequate for every application. In particular, they make a distinction between two kinds of modifications to a theory. The first one, *update*, consists of bringing the theory up to date when the world described by it changes. For example, most database updates are of this variety, e.g. 'increase Joe's salary by 5%'. Another example is the incorporation into the theory of changes in the world caused by the actions of an agent such as a robot [9, 25, 26]. The second type of modification, *revision*, is used when the agent obtains new information about a static world. For example, the agent may be trying to diagnose a faulty circuit and wanting to incorporate into the knowledgebase the results of successive tests, where newer results may contradict old ones. Katsuno and Mendelzon conclude that the *AGM* postulates describe only revision and they give a modified set of postulates for update (the *KM* postulates).

As an aside, we note that distinctions of a similar nature have been suggested in other contexts. Gibbard [8] compares indicative and subjunctive conditionals of revision and update. In the context of Bayesian models of belief revision, Lewis [21] and Gärdenfors [7] propose *imaging* as an alternative to *conditionalization.* As suggested to us in private communications by H. Arló-Costa and P. Gärdenfors, imaging appears to be a probabilistic counterpart of update.

The purpose of this paper is to present an epistemic logic such that, for any revision operator that satisfies the *AGM* postulates, the result of a revision can be described by a sentence in the logic.

In our logic, the fact that the agent's set of beliefs is ϕ is represented by the sentence $\mathtt{O}\phi$, where $\mathtt{O}$ is Levesque's 'only know' operator [19]. Intuitively, $\mathtt{O}\phi$ is read as 'ϕ is *all* that is believed.' The fact that the agent believes ψ is represented by the sentence $\mathtt{B}\psi$, read in the usual way as 'ψ is believed'. The connective $\diamond$ represents *update* as defined in [17]. The revised beliefs are represented by the sentence $\mathtt{O}\phi \diamond \mathtt{B}\psi$. [1] We show that for every revision operator that satisfies the *AGM* postulates, there is a model for our epistemic logic such that the beliefs implied by the sentence $\mathtt{O}\phi \diamond \mathtt{B}\psi$ in this model correspond exactly to the sentences implied by the theory that results from revising ϕ by ψ. This means that reasoning about changes in the agent's beliefs reduces to model checking [13] of certain epistemic sentences.

Note that a revision in the agent's universe of discourse is translated into an update at the meta level. Intuitively, the reason is that update, and not revision, is the operator that serves to accommodate changes in the world. At the meta level, the world includes

[1] We shall leave out parenthesis symbols from sentences, with the convenience that unary operators bind stronger than binary, and that among the binary connectives, the update $\diamond$ has the strongest binding. Thus for instance $(((\mathtt{O}\phi) \diamond (\mathtt{B}\psi)) \rightarrow (\mathtt{B}\chi))$ is written as $\mathtt{O}\phi \diamond \mathtt{B}\psi \rightarrow \mathtt{B}\chi$.

the agent's set of beliefs, since we can assert that the agent does or does not believe certain sentences; hence update is appropriate to incorporate the change in the world that occurs when the agent changes its mind about something.

For example, in a distributed system an agent (a node in a network) may revise its beliefs about the state of a computation. An account of this revision is necessary to reason formally about properties of such systems. An interesting extension would be to include multiple agents and a notion of 'common knowledge', as in Fagin, Halpern, Moses and Vardi [4], since a formal analysis of their 'muddy children puzzle' requires the concept of belief revision to describe how the children's beliefs change as a consequence of the father's statements.

It might be illuminating at this point to consider Levesque's distinction between *subjective* and *objective* uses of logic: If the agent used logic to reason about its epistemic states, it would be using logic subjectively. The metalogic is used here objectively to *describe* the epistemic states of the agent.

2 Belief Revision Systems

In this section we define the language and the apparatus that the agent uses to represent, reason about, and change its beliefs. The language and reasoning is classical propositional logic, and the change mechanism is a revision operator.

Let $\mathcal{L}$ be a propositional language with a countably infinite set P of propositional letters. The set S of all (well formed) sentences (in $\mathcal{L}$) is defined in the usual way. Sentences will be denoted by $\phi, \psi, \chi, \ldots$. A *theory* (in $\mathcal{L}$) is a subset Γ of S, such that $Cn(\Gamma) = \Gamma$, where Cn denotes the propositional consequence operator. The set of all theories (in $\mathcal{L}$) is denoted T. A sentence ϕ is said to be a *theorem*, if $\phi \in Cn(\emptyset)$, and ϕ is said to be *consistent*, if $\neg\phi \notin Cn(\emptyset)$. A theory Γ is said to be consistent, if $\Gamma \neq S$. A theory Γ is said to be *finitely axiomatizable* (in $\mathcal{L}$), if there is a $\phi \in S$, such that $Cn(\{\phi\}) = \Gamma$.

The theory that an agent holds changes as new inputs are assimilated. To abstract the change mechanism we use a revision operator that maps a theory and a sentence to a new theory. Following Gärdenfors [7], a ($\mathcal{L}$-) *belief revision system* is a triple

$$\langle S, Cn, \oplus \rangle,$$

where $\oplus$ is a function from $T \times S$ to T, such that the following conditions hold:

(R1) $\phi \in \Gamma \oplus \phi$.

(R2) If $\Gamma \cup \{\phi\}$ is consistent, then $\Gamma \oplus \phi = Cn(\Gamma \cup \{\phi\})$,

(R3) If ϕ is consistent, then $\Gamma \oplus \phi$ is consistent.

(R4) If $\phi \leftrightarrow \psi$ is a theorem, then $\Gamma \oplus \phi = \Gamma \oplus \psi$.

(R5) $\Gamma \oplus (\phi \wedge \psi) \subseteq Cn(\Gamma \oplus \phi \cup \{\psi\})$.

(R6) If $\Gamma \oplus \phi \cup \{\psi\}$ is consistent, then $Cn(\Gamma \oplus \phi \cup \{\psi\}) \subseteq \Gamma \oplus (\phi \wedge \psi)$.

For an intuitive motivation of the conditions (the 'revision postulates', or 'rationality criteria'), see [1, 7].

In any practical application, the agent's beliefs Γ will be of the form $Cn(\phi)$, for some $\phi \in S$. This practical aspect will however be ruined, if the same is no longer true after a revision. Consider therefore the following condition.

(R0) $Cn(\{\phi\}) \oplus \psi$ is finitely axiomatizable.

If a belief revision $\langle S, Cn, \oplus\rangle$ also satisfies condition (R0), we say that that $\langle S, Cn, \oplus\rangle$ is a ($\mathcal{L}$-) *preservative* belief revision system.

In Section 4 we shall need the following lemma, the proof of which can be found in [7].

Lemma 2.1 *If $\psi \in \Gamma \oplus \phi$, and $\phi \in \Gamma \oplus \psi$, then $\Gamma \oplus \phi = \Gamma \oplus \psi$.*

The following lemma will also be needed later.

Lemma 2.2 *Let $Cn(\{\phi\}) \oplus \psi = Cn(\{\chi\})$, and let ξ be a sentence in S such that $\psi \in Cn(\{\xi\})$. Then $Cn(\{\phi\}) \oplus (\chi \vee \xi) = Cn(\{\chi\})$.*

Proof. If ψ is inconsistent the claim is obvious. Suppose then that ψ is consistent. From (R1) and (R4) it follows that $Cn(\{\phi\}) \oplus (\chi \vee \xi) = Cn(\{\phi\}) \oplus (\psi \wedge (\chi \vee \xi))$. By assumption $Cn(\{\phi\}) \oplus \psi \ \cup \ \{\chi \vee \xi\}$ is consistent. Therefore (R5) and (R6) entails that $Cn(\{\phi\}) \oplus (\psi \wedge (\chi \vee \xi)) = Cn(Cn(\{\phi\}) \oplus \psi \ \cup \ \{\chi \vee \xi\})$. By assumption the latter set is equal to $Cn(Cn(\{\chi\}) \ \cup \ \{\chi \vee \xi\})$ which is equal to $Cn(\{\chi\})$.

3 The Update Language and its Interpretation

We now present the epistemic language that we use to describe an agent's beliefs and how they change under revision. This language $\mathcal{L}_{\mathcal{L}}$ is constructed from the set P of propositional letters, the Boolean operators, and the following intensional connectives: a binary connective $\diamond$ (update), and the unary connectives $\mathtt{B}$ and $\mathtt{N}$ (believe at least and believe at most). Intuitively, $\mathtt{B}\phi$ will mean that the agent believes sentence ϕ (and perhaps other things too), while $\mathtt{N}\phi$ will mean that the agent does not believe anything that is not a consequence of ϕ. $\mathtt{N}\phi$ is only introduced for technical reasons; we will actually use $\mathtt{O}\phi$, which is defined in terms of $\mathtt{B}\phi$ and $\mathtt{N}\phi$ and means the agent believes precisely ϕ.

In the set of well formed sentences of $\mathcal{L}_{\mathcal{L}}$ we shall not include for instance sentences of the form $\mathtt{BB}\phi$, which intuitively would mean 'the agent believes that it believes ϕ'. The reason for this will become clear in Section 5, where we consider the scenario where the agent is fully introspective.[2] The set of all well formed sentences of $\mathcal{L}_{\mathcal{L}}$ is defined as follows: Any propositional sentence (i.e. any sentence in S) is a $\mathcal{L}_{\mathcal{L}}$-sentence, and if If ϕ and ψ are propositional sentences (i.e. sentences in S), then $\mathtt{B}\phi$, $\mathtt{N}\phi$, $(\mathtt{B}\phi \wedge \mathtt{N}\phi)$ and $(\mathtt{B}\phi \wedge \mathtt{N}\phi) \diamond \mathtt{B}\psi$ are epistemic sentences. The set of $\mathcal{L}_{\mathcal{L}}$-sentences consists of all and only those sentences that are propositional, epistemic, or Boolean combinations of epistemic and propositional sentences.

[2]The further restrictions on the well formed sentences are placed in order to guarantee that we will be able to constructively satisfy the limit assumption. For a discussion of the limit assumption, see [20, 23].

Now let us define formally the interpretation of the $\mathcal{L}_{\mathcal{C}}$-sentences. An *update model* is a quadruple

$$\mathcal{M} = \langle W, R, \leq, [\![]\!] \rangle,$$

where W is a non-empty set of worlds $w, v, u, \ldots$, where R is a binary relation over W, where $\leq$ is a function that assigns a partial pre-order $\leq_w$ on W to each member w of W, and where $[\![]\!]$ is a function that assigns a subset $[\![\phi]\!]$ of W to each $\mathcal{L}_{\mathcal{C}}$-sentence ϕ. The relation R is an *accessibility relation*. We use it to model belief as in the standard semantics of modal logic, that is, ϕ will be believed at some world w iff ϕ is true at all possible worlds accessible from w. For any world $w \in W$ we shall use the notation R_w for the set of worlds accessible from w, that is, $\{v \in W : wRv\}$. The relation $\leq_w$ is a *comparative similarity ordering of worlds w.r.t. world w*. This relation will be used to define the semantics of update as in [10, 17]. The sentence $\phi \diamond \psi$ will be true at a world v if v satisfies ψ and there exists a world w that satisfies ϕ such that v is 'as close' to w as any other world that satisfies ψ, w.r.t. the comparative similarity ordering $\leq_w$.

In addition the following conditions have to be fulfilled.

(S1) For each propositional sentence ϕ, the set $min_{\leq_w}([\![\mathtt{B}\phi]\!])$ is non-empty, whenever $[\![\mathtt{B}\phi]\!]$ is non-empty. [3]

(S2) $min_{\leq_w}(W) = \{w\}$.

(VC) $[\![\neg\phi]\!] = W \setminus [\![\phi]\!]$.

(VA) $[\![\phi \wedge \psi]\!] = [\![\phi]\!] \cap [\![\psi]\!]$.

(VU) $[\![\phi \diamond \psi]\!] = \bigcup_{w \in [\![\phi]\!]} min_{\leq_w}([\![\psi]\!])$.

(VB) $[\![\mathtt{B}\phi]\!] = \{w \in W : R_w \subseteq [\![\phi]\!]\}$

(VN) $[\![\mathtt{N}\phi]\!] = \{w \in W : R_w \supseteq [\![\phi]\!]\}$

Using (VU) and a representation theorem in [17], the following property can be shown.

Theorem 3.1 *The $\diamond$-operator satisfies the KM postulates for update.*

In the sequel we shall use the following abbreviation:

$$\mathtt{O}\phi =_{df} \mathtt{B}\phi \wedge \mathtt{N}\phi.$$

It is obvious that

$$[\![\mathtt{O}\phi]\!] = \{w \in W : R_w = [\![\phi]\!]\}.$$

A sentence ϕ is said to be *valid in a model* $\mathcal{M} = \langle W, R, \leq, [\![]\!] \rangle$, if $[\![\phi]\!] = W$. This fact is denoted $\mathcal{M} \models \phi$.

[3] For $V \subseteq W$ and $w \in W$, $min_{\leq_w}(V)$ denotes the set $\{v \in V : \text{if } u \in V \text{ and } u \leq_w v, \text{ then } v \leq_w u\}$.

4 Interpreting Belief Revision Systems

In this section we show that the semantics of $\mathcal{L}$-preservative belief revision systems can be captured by update models to the extent that the question $\chi \in Cn(\{\phi\}) \oplus \psi$ can be reduced to deciding whether the $\mathcal{L}_{\mathcal{L}}$-sentence $\mathtt{O}\phi \diamond \mathtt{B}\psi \rightarrow \mathtt{B}\chi$ is valid in a particular $\mathcal{L}_{\mathcal{L}}$ update model $\mathcal{M}^{\oplus}$. Formally, we have the following.

Theorem 4.1 *For every $\mathcal{L}$-preservative belief revision system $\langle S, Cn, \oplus\rangle$, there is an $\mathcal{L}_{\mathcal{L}}$ update model $\mathcal{M}^{\oplus}$, such that for all sentences ϕ and ψ in S, we have*

(RU) $Cn(\{\phi\}) \oplus \psi = \{\chi \in S : \mathcal{M}^{\oplus} \models \mathtt{O}\phi \diamond \mathtt{B}\psi \rightarrow \mathtt{B}\chi\}$.

Proof. Let $\mathcal{M}^{\oplus} = \langle W, R, \leq, [\![]\!]\rangle$, where $W = \{\langle i, \phi\rangle : i \subseteq P, \phi \in S\}$. Furthermore, for each propositional letter $p \in P$, define $[\![p]\!] = \{\langle i, \phi\rangle : p \in i, \phi \in S\}$. Then extend $[\![]\!]$ to all $\phi \in S$ through equations (VC) and (VA). After this, put $\langle i, \phi\rangle R \langle j, \psi\rangle$ if and only if $\langle j, \psi\rangle \in [\![\phi]\!]$. Thus $R_{\langle i,\phi\rangle} = [\![\phi]\!]$. Now extend $[\![]\!]$ through equations (VN), (VB), (VA) and (VC) until saturation.

For the sequel, it is worth keeping in mind that for any $\phi \in S$, we have $[\![\mathtt{B}\phi]\!] = \{\langle i, \psi\rangle : i \subseteq P,\ \phi \in Cn(\{\psi\})\}$, and $[\![\mathtt{O}\phi]\!] = \{\langle i, \psi\rangle : i \subseteq P,\ Cn(\{\psi\}) = Cn(\{\phi\})\}$.

Then, for each $\langle i, \phi\rangle \in W$, define $\langle j, \psi\rangle \leq_{\langle i,\phi\rangle} \langle k, \xi\rangle$ iff either

(a). $Cn(\{\psi\}) \neq Cn(\{\xi\})$ and $Cn(\{\phi\}) \oplus (\psi \vee \xi) = Cn(\{\psi\})$, or

(b). $Cn(\{\psi\}) = Cn(\{\xi\})$ and $(j \setminus i) \cup (i \setminus j) \subseteq (k \setminus i) \cup (i \setminus k)$.

Finally, extend $[\![]\!]$ to all $\mathcal{L}_{\mathcal{L}}$-sentences through equations (VC), (VA), (VU), (VB) and (VN).

We now have to verify that $\mathcal{M}^{\oplus}$ thus constructed indeed is an update model. The requirements for $[\![]\!]$ are obviously satisfied.

From the definition it immediately follows that $\leq_{\langle i,\phi\rangle}$ is reflexive. For transitivity, suppose that $\langle j, \psi\rangle \leq_{\langle i,\phi\rangle} \langle k, \xi\rangle$, and that $\langle k, \xi\rangle \leq_{\langle i,\phi\rangle} \langle m, \nu\rangle$. There are four cases to consider.

(i). $Cn(\{\psi\}) = Cn(\{\xi\})$, and $Cn(\{\xi\}) = Cn(\{\nu\})$.

Then it must be that $(j \setminus i) \cup (i \setminus j) \subseteq (k \setminus i) \cup (i \setminus k)$, and $(k \setminus i) \cup (i \setminus k) \subseteq (m \setminus i) \cup (i \setminus m)$. Thus $(j \setminus i) \cup (i \setminus j) \subseteq (m \setminus i) \cup (i \setminus m)$.

(ii). $Cn(\{\psi\}) = Cn(\{\xi\})$, $Cn(\{\xi\}) \neq Cn(\{\nu\})$, and $Cn(\{\phi\}) \oplus (\xi \vee \nu) = Cn(\{\xi\})$.

Then it follows from (R4) that $Cn(\{\phi\}) \oplus (\psi \vee \nu) = Cn(\{\psi\})$.

(iii). $Cn(\{\psi\}) \neq Cn(\{\xi\})$, $Cn(\{\phi\}) \oplus (\psi \vee \xi) = Cn(\{\psi\})$, and $Cn(\{\xi\}) = Cn(\{\nu\})$.

Again, it follows from (R4) that $Cn(\{\phi\}) \oplus (\psi \vee \nu) = Cn(\{\psi\})$.

(iv). $Cn(\{\psi\}) \neq Cn(\{\xi\})$, $Cn(\{\phi\}) \oplus (\psi \vee \xi) = Cn(\{\psi\})$, $Cn(\{\xi\}) \neq Cn(\{\nu\})$, and $Cn(\{\phi\}) \oplus (\xi \vee \nu) = Cn(\{\xi\})$.

First we claim that $(\psi \vee \xi) \in Cn(\{\phi\}) \oplus (\psi \vee \xi \vee \nu)$. If $\psi \in Cn(\{\phi\}) \oplus (\psi \vee \xi \vee \nu)$ the claim is true. Suppose therefore the contrary. Then $Cn(\{\phi\}) \oplus (\psi \vee \xi \vee \nu)\ \cup\ \{\neg\psi\}$ is consistent. From (R1) it then follows that $Cn(\{\phi\}) \oplus (\psi \vee \xi \vee \nu) \ \cup \ \{\xi \vee \nu\}$ is consistent. Then (R5), (R6), (R4), and our assumptions yield

$$Cn(Cn(\{\phi\}) \oplus (\psi \vee \xi \vee \nu) \ \cup \ \{\xi \vee \nu\}) =$$

$$Cn(\{\phi\}) \oplus ((\psi \vee \xi \vee \nu) \wedge (\xi \vee \nu)) =$$

$$Cn(\{\phi\}) \oplus (\xi \vee \nu) = Cn(\{\xi\}).$$

From the above it follows that $Cn(\{\phi\}) \oplus (\psi \vee \xi \vee \nu) \ \cup \ \{\psi \vee \xi\}$ is consistent. Using (R1) we now get

$$Cn(\{\phi\}) \oplus (\psi \vee \xi \vee \nu)$$

$$=$$

$$Cn(Cn(\{\phi\}) \oplus (\psi \vee \xi \vee \nu) \ \cup \ \{\psi \vee \xi\}) \ \cap \ Cn(Cn(\{\phi\}) \oplus (\psi \vee \xi \vee \nu) \ \cup \ \{\xi \vee \nu\}).$$

By (R5), (R6), and (R4) the last set is equal to

$$Cn(\{\phi\}) \oplus (\psi \vee \xi) \ \cap \ Cn(\{\phi\}) \oplus (\xi \vee \nu).$$

By (R1) $(\psi \vee \xi) \in Cn(\{\phi\}) \oplus (\psi \vee \xi)$, and since $Cn(\{\phi\}) \oplus (\xi \vee \nu) = Cn(\{\xi\})$, we have $(\psi \vee \xi) \in Cn(\{\phi\}) \oplus (\xi \vee \nu)$. Consequently the claim $(\psi \vee \xi) \in Cn(\{\phi\}) \oplus (\psi \vee \xi \vee \nu)$ is proved.

From (R1) it follows that $(\psi \vee \xi \vee \nu) \in Cn(\{\phi\}) \oplus (\psi \vee \xi)$. Then Lemma 2.1 entails $Cn(\{\phi\}) \oplus (\psi \vee \xi \vee \nu) = Cn(\{\phi\}) \oplus (\psi \vee \xi)$.

Since $Cn(\{\phi\}) \oplus (\psi \vee \xi \vee \nu) = Cn(\{\phi\}) \oplus (\psi \vee \xi)$, and since the assumption of the present case is that $Cn(\{\phi\}) \oplus (\psi \vee \xi) = Cn(\{\psi\})$, Lemma 2.2 gives the desired conclusion that $Cn(\{\phi\}) \oplus (\psi \vee \nu) = Cn(\{\psi\})$.

Thus $\leq_{\langle i,\phi\rangle}$ is transitive and indeed a partial pre-order. Regarding $\leq_{\langle i,\phi\rangle}$ it remains to show that (S1) and (S2) are satisfied.

For (S2), if $Cn(\{\phi\}) \neq Cn(\{\psi\})$, then it follows from (R2) and condition (a) of the definition of the similarity ordering that $\langle i,\phi\rangle \leq_{\langle i,\phi\rangle} \langle j,\psi\rangle$ and that $\langle j,\psi\rangle \not\leq_{\langle i,\phi\rangle} \langle i,\phi\rangle$. If $Cn(\{\phi\}) = Cn(\{\psi\})$ and $i \neq j$, then condition (b) entails that $\langle i,\phi\rangle \leq_{\langle i,\phi\rangle} \langle j,\phi\rangle$ and $\langle j,\phi\rangle \not\leq_{\langle i,\phi\rangle} \langle i,\phi\rangle$.

We then verify that (S1) holds for all sentences of the form $[\![\mathtt{B}\psi]\!]$, where $\psi \in S$. That is to say that the set $min_{\leq_{\langle i,\phi\rangle}}([\![\mathtt{B}\psi]\!])$ is nonempty. We shall show that if $Cn(\{\phi\}) \oplus \psi = Cn(\{\chi\})$, then $min_{\leq_{\langle i,\phi\rangle}}([\![\mathtt{B}\psi]\!]) = \{\langle i,\xi\rangle : Cn(\{\xi\}) = Cn(\{\chi\})\}$. Call the latter set A. From (R1) it follows that $A \subseteq [\![\mathtt{B}\psi]\!]$.

For minimality, first note that if $\langle i,\xi_1\rangle$ and $\langle i,\xi_2\rangle$ are both in A, then $\langle i,\xi_1\rangle \leq_{\langle i,\phi\rangle} \langle i,\xi_2\rangle$ and $\langle i,\xi_2\rangle \leq_{\langle i,\phi\rangle} \langle i,\xi_1\rangle$. Then choose an arbitrary world $\langle i,\xi\rangle \in A$ and an arbitrary world $\langle j,\nu\rangle \in [\![\mathtt{B}\psi]\!]$. Note that $\psi \in Cn(\{\nu\})$. If $Cn(\{\nu\}) = Cn(\{\xi\})$, then $\langle i,\xi\rangle \leq_{\langle i,\phi\rangle} \langle j,\nu\rangle$. Suppose then that $Cn(\{\nu\}) \neq Cn(\{\xi\})$. It now follows from Lemma 2.2 that $Cn(\{\phi\}) \oplus (\xi \vee \nu) = Cn(\{\xi\})$. Thus $\langle i,\xi\rangle \leq_{\langle i,\phi\rangle} \langle j,\nu\rangle$ and since $Cn(\{\nu\}) \neq Cn(\{\xi\})$, we have $\langle j,\nu\rangle \not\leq_{\langle i,\phi\rangle} \langle i,\xi\rangle$.

Thus $\mathcal{M}^{\oplus}$ is indeed an $\mathcal{L}_{\mathcal{L}}$ update model.[4] Now the equality (RU) can be verified.

To this end, we first show that if $Cn(\{\phi\}) \oplus \psi = Cn(\{\chi\})$, then $[\![\mathtt{O}\phi \diamond \mathtt{B}\psi]\!] = [\![\mathtt{O}\chi]\!]$.

[4]Note that if $\mathcal{L}$ is finitary, then $\mathcal{M}^{\oplus}$ is finite.

Let $\langle j, \xi\rangle \in [\![\mathtt{O}\phi\diamond\mathtt{B}\psi]\!]$. Then we have $\psi \in Cn(\{\xi\})$. Furthermore, there is a world $\langle i, \phi\rangle$ such that $\langle j, \xi\rangle \in min_{\leq_{\langle i,\phi\rangle}}([\![\mathtt{B}\psi]\!])$. Suppose now to the contrary that $\langle j, \xi\rangle \notin [\![\mathtt{O}\chi]\!]$. Then we have $Cn(\{\xi\}) \neq Cn(\{\chi\})$. Since $\psi \in Cn(\{\xi\})$, Lemma 2.2 gives us $Cn(\{\phi\})\oplus(\chi\vee\xi) = Cn(\{\chi\})$. Therefore, for any $k \subseteq P$ we have $\langle k, \chi\rangle \leq_{\langle i,\phi\rangle} \langle j, \xi\rangle$, and $\langle j, \xi\rangle \not\leq_{\langle i,\phi\rangle} \langle k, \chi\rangle$, which is a contradiction to the fact that $\langle j, \xi\rangle \in min_{\leq_{\langle i,\phi\rangle}}([\![\mathtt{B}\psi]\!])$.

For inclusion in the other direction, let $\langle j, \xi\rangle \in [\![\mathtt{O}\chi]\!]$. Note that $Cn(\{\xi\})$ must be equal to $Cn(\{\chi\})$. From (R1) we get $\langle j, \xi\rangle \in [\![\mathtt{B}\psi]\!]$. Towards a contradiction, suppose then that $\langle j, \xi\rangle \notin min_{\leq_{\langle j,\phi\rangle}}([\![\mathtt{B}\psi]\!])$. Then there is a sentence $\nu \in S$ and a $k \subseteq P$, such that $\psi \in Cn(\{\nu\})$, $\langle k, \nu\rangle \leq_{\langle i,\phi\rangle} \langle j, \xi\rangle$, and $\langle j, \xi\rangle \not\leq_{\langle i,\phi\rangle} \langle k, \nu\rangle$. From the construction of the similarity ordering it then follows that $Cn(\{\phi\}) \oplus (\nu \vee \xi) = Cn(\{\nu\}) \neq Cn(\{\xi\})$. Since $Cn(\{\nu\}) = Cn(\{\chi\})$ we have a contradiction.

It remains to verify (RU). Let $Cn(\{\phi\}) \oplus \psi = Cn(\{\chi\})$, and let $\xi \in Cn(\{\chi\})$. By the preceding elaboration it suffices to verify that $\mathcal{M}^\oplus \models \mathtt{O}\chi \to \mathtt{B}\xi$. But this follows directly from the fact that $\xi \in Cn(\{\chi\})$ and the construction of $\mathcal{M}^\oplus$.

Finally, let ξ be a sentence in S such that $\mathcal{M}^\oplus \models \mathtt{O}\chi \to \mathtt{B}\xi$. Then we have $\xi \in Cn(\{\chi\})$ directly from the construction of $\mathcal{M}^\oplus$.

Corollary 4.1 *$Cn(\emptyset) = \{\phi \in S : \mathcal{M}^\oplus \models \mathtt{B}\phi\}$.*

5 On the Limitations of Introspection

There is considerable evidence to the fact that revision is a meta level concept, not expressible in the object language whose theories are being revised. The most striking piece of evidence is the *triviality theorem* of Gärdenfors [7]. To arrive at his theorem Gärdenfors invokes the so called *Ramsey rule* that relates conditional propositions and theory change. The Ramsey rule is summarized by Gärdenfors as follows:

> Accept a proposition of the form 'If A then B' in a state of belief K if and only if the minimal change of K needed to accept A also requires accepting B.

In other words, if $\psi > \chi$ represents the conditional 'If ψ were true, then χ would also be true', then Gärdenfors' interpretation of the Ramsey rule can be expressed as

(RR') $\psi > \chi \in Cn(\{\phi\})$ iff $\chi \in Cn(\{\phi\}) \oplus \psi$.

Gärdenfors has shown that this rule is incompatible with revision postulates (R1)–(R3).[5] The incompatibility means that if a belief revision system satisfies (RR'), and (R1)–(R3), then the system is trivial.[6]

The fundamental conflict seems to be between (RR') and (R2). Much discussion has centered around the question whether to therefore abandon (RR') or (R2) (for an overview, see [7]). However, Grahne [10] has constructed a non-trivial logic that has both

[5] As a matter of fact, Gärdenfors obtains the result already for certain weakened versions of (R1)–(R3).

[6] A belief revision system $\langle S, Cn, \oplus\rangle$ is said to be *non-trivial*, if there are sentences ϕ, ψ, and χ in S, and a theory Γ in T, such that the sentences ϕ, ψ, and χ are consistent, the sentences $\phi\wedge\psi$, $\psi\wedge\chi$, and $\phi\wedge\chi$ are inconsistent, and the theories Γ, $\Gamma\oplus\phi$, $\Gamma\oplus\psi$, and $\Gamma\oplus\chi$ are consistent. Otherwise, the system is *trivial* [7].

a conditional connective $>$ and an update operator $\diamond$ in the object language, such that $\diamond$ satisfies the *KM*-postulates for update, and $>$ has the interpretation

$$[\![\psi > \chi]\!] = \{w \in W : min_{\leq_w}(\psi) \subseteq [\![\chi]\!]\},$$

while the two operators are connected by the bidirectional derivation rule

(RR) $\psi > \chi \in Cn(\{\phi\})$ iff $\chi \in Cn(\{\phi \diamond \psi\})$,

which is nothing less than the Ramsey rule. In particular, our current update language $\mathcal{L}_{\mathcal{L}}$ could be extended with a $>$-operator and the following semantic version of the Ramsey rule: $\mathcal{M} \models \phi \to \psi > \chi$ if and only if $\mathcal{M} \models \phi \diamond \psi \to \chi$, for all update models $\mathcal{M}$. Since update and conditional thus belong to the same language level, and revision becomes an update on the meta level, it should not be too surprising that confusing the two levels by adopting (RR') and the revision postulate (R2) leads to triviality.

From a semantic viewpoint, the hierarchical approach of Theorem 4.1 is well justified. In particular, the equation (RU) in Theorem 4.1 could be justifiable as a postulate to require for any belief revision system. From a proof theoretic viewpoint, Levi [18] arrives at a similar meta level approach. In his approach, Levi distinguishes between the agent's beliefs (called the *corpus*), and a set of statements about the agent's beliefs (called the *metacorpus*). That is, if the agent's beliefs are represented by a theory Γ, then the metacorpus, called $Poss(\Gamma)$, contains $\mathtt{B}\phi$ iff $\phi \in \Gamma$, and contains $\neg\mathtt{B}\phi$ iff $\phi \notin \Gamma$. Furthermore, Γ is to be a subset of $Poss(\Gamma)$. In particular, Levi points out that the relationship between $Poss(\Gamma)$ and $Poss(\Gamma \oplus \psi)$ is *not* that of a revision (but he does not say exactly what that relationship is), and that $Poss(\Gamma)$ is formulated in a language that properly extends the language in which Γ is formulated.

Our results in the previous section verify that observation: If we assume that the agent's corpus is represented by a sentence ϕ, then Levi's metacorpus would correspond to the sentence $\mathtt{O}\phi$ interpreted in the model $\mathcal{M}^{\oplus}$. As we have seen, the relationship between the 'metacorpus' for $Cn(\{\phi\})$ and for $Cn(\{\phi\}) \oplus \psi$ is that of an *update*, and not a revision. We could thus, in analogy with Levi, relate revision and conditional through the following reasoning:

$$\chi \in Cn(\{\phi\}) \oplus \psi \text{ iff } \mathcal{M}^{\oplus} \models \mathtt{O}\phi \diamond \mathtt{B}\psi \to \mathtt{B}\chi \text{ iff } \mathcal{M}^{\oplus} \models \mathtt{O}\phi \to \mathtt{B}\psi > \mathtt{B}\chi.$$

We could thus also account for *iterated* conditionals, that is propositions of the form 'If A then (if B then C)'. Let φ be a sentence in S such that $Cn(\{\phi\}) \oplus \psi = Cn(\{\varphi\})$. Then we would simply have $\mathcal{M}^{\oplus} \models \mathtt{O}\phi \to \mathtt{B}\psi > (\mathtt{B}\chi > \mathtt{B}\xi)$ iff $\mathcal{M}^{\oplus} \models \mathtt{O}\phi \diamond \mathtt{B}\psi \to \mathtt{B}\chi > \mathtt{B}\xi$ iff $\mathcal{M}^{\oplus} \models (\mathtt{O}\phi \diamond \mathtt{B}\psi) \diamond \mathtt{B}\chi \to \mathtt{B}\xi$ iff $\mathcal{M}^{\oplus} \models (\mathtt{O}\varphi) \diamond \mathtt{B}\chi \to \mathtt{B}\xi$ iff $\xi \in Cn(\{\varphi\}) \oplus \chi$ iff $\xi \in (Cn(\{\phi\}) \oplus \psi) \oplus \chi$.

Returning to the triviality issue, we saw that adopting (RR') and (R2) gives the agent some abilities to reason about its revision strategies. We shall now consider the scenario where the agent has such abilities as a consequence of *introspection*: Suppose that instead of propositional logic, the agent is reasoning in some *epistemic* logic, for example in the logic *K45* of belief. This modal logic is propositional logic augmented with a modal operator $\Box$. The axioms and derivation rules of *K45* are those of modal logic K plus the axiom of *positive introspection*

$$\Box\phi \to \Box\Box\phi,$$

and the axiom of *negative introspection*

$$\neg\Box\phi \rightarrow \Box\neg\Box\phi.$$

If the operator $\Box$ is taken to mean belief, then the positive introspection axiom 'says' that if the agent believes ϕ then it believes that it believes ϕ. The axiom for negative introspection says that if the agent does not believe ϕ then it believes that it does not believe ϕ. In case the agent's logic is both positively and negatively introspective, the agent is said to be *fully introspective*. For a more detailed account of *K45* see e.g. [12, 14].

By invoking a representation theorem of Grove [11], it is possible to show that Theorem 4.1 also holds for *K45*. Note however that the operators $\mathtt{B}$ and $\Box$ will be given different meanings by this construction. If the agent believes the sentence $\Box\phi$, this will be translated into the meta level sentence $\mathtt{B}\Box\phi$, meaning intuitively that 'the agent believes $\Box\phi$', which is not the same as 'the agent believes $\mathtt{B}\phi$,' or 'the agent believes "the agent believes ϕ" '. Suppose we try to eliminate this difference by identifying the object level $\Box$ with the meta level $\mathtt{B}$, allowing the agent to hold beliefs about its own beliefs. We show below that this attempt is doomed to end in triviality, as also happens when (RR') and (R2) are adopted. A result with a similar flavour but somewhat different assumptions has been obtained by Fuhrmann [6]. To state our result formally, we shall need the following definition: A $\mathcal{L}$-belief revision system is said to be *strongly trivial* if all sentences are complete.[7]

We can now show that if the difference between $\mathtt{B}$ and $\Box$ is erased, the construction of Theorem 4.1 is only applicable to strongly trivial belief revision systems.

Theorem 5.1 *Let $\langle S, Cn, \oplus\rangle$ be a $\mathcal{L}$-belief revision system satisfying at least (R0)–(R3), where $\mathcal{L}$ and Cn extend propositional logic, the consequence operator is finitistic*[8]*, and where $\mathtt{B}\phi \rightarrow \mathtt{BB}\phi \in Cn(\emptyset)$, and $\neg\mathtt{B}\phi \rightarrow \mathtt{B}\neg\mathtt{B}\phi \in Cn(\emptyset)$. Suppose furthermore that the equation (RU) of Theorem 4.1 holds. Then the belief revision system is strongly trivial.*

Proof. Suppose to the contrary that the belief revision system is not strongly trivial, and let ϕ and ψ be sentences in $\mathcal{L}$ such that $\psi \notin Cn(\{\phi\})$, and $\neg\psi \notin Cn(\{\phi\})$. Thus, by (R2), $\psi \notin Cn(\{\phi\}) \oplus \phi$. Then we have by (RU)

$$\mathcal{M}^{\oplus} \not\models \mathtt{O}\phi \diamond \mathtt{B}\phi \rightarrow \mathtt{B}\psi,$$

and from the definition of $\diamond$ we get

$$\mathcal{M}^{\oplus} \not\models \mathtt{O}\phi \rightarrow \mathtt{B}\psi.$$

It now follows from the definition of the $\mathtt{O}$-operator that

$$\mathcal{M}^{\oplus} \models \mathtt{O}\phi \rightarrow \neg\mathtt{B}\psi,$$

and by negative introspection we get

$$\mathcal{M}^{\oplus} \models \mathtt{O}\phi \rightarrow \mathtt{B}\neg\mathtt{B}\psi,$$

[7] A sentence ϕ is complete if for any other sentence ψ, either $\psi \in Cn(\{\phi\})$, or $\neg\psi \in Cn(\{\phi\})$.

[8] A consequence operator Cn is said to be finitistic, if for all sets of sentences X we have $Cn(X) = \bigcup\{Cn(Y) : Y \subseteq X \text{ and } Y \text{ is finite}\}$

Applying (RU) again we get

$$\neg \mathtt{B}\psi \in Cn(\{\phi\}).$$

From (R2) it now follows that

$$\neg \mathtt{B}\psi \in Cn(\{\phi\}) \oplus \psi.$$

On the other hand, from the definition of $\diamond$ it follows that

$$\mathcal{M}^{\oplus} \models \mathtt{O}\phi \diamond \mathtt{B}\psi \rightarrow \mathtt{B}\psi,$$

and applying positive introspection we get

$$\mathcal{M}^{\oplus} \models \mathtt{O}\phi \diamond \mathtt{B}\psi \rightarrow \mathtt{BB}\psi.$$

Then (RU) gives

$$\mathtt{B}\psi \in Cn(\{\phi\}) \oplus \psi.$$

This means that $Cn(\{\phi\}) \oplus \psi$ is inconsistent. From (R3) it then follows that ψ must be inconsistent. Therefore $\neg\psi \in Cn(\emptyset)$, and consequently $\neg\psi \in Cn(\phi)$, a contradiction to the assumption that the belief revision system is not strongly trivial.

We see that the agent cannot be fully introspective with respect to the operator B. Whatever introspection means, it cannot thus mean introspection with respect to the way the agent revises its beliefs.

6 Conclusion

We have given a metalanguage in which it is possible to reason formally about how an agent changes its beliefs. The agent is performing revisions, that is, it is adjusting its beliefs to accommodate new, possibly conflicting information about a static world. At the meta level, the world—which includes the agent's state of beliefs—is changing as the agent changes its mind. These changes are modeled by updates. Levesque's 'only know' modal operator is used at the meta level to describe the totality of beliefs held by an agent.

The negative result in the paper is that this type of formal account of revision cannot be extended to the situation where the agent is able to reason about its beliefs. A fully introspective agent cannot use our construction to reason about the results of its own revisions, on pain of triviality.

Acknowledgements

We would like to thank David Israel for pointing out several relevant references, and Hirofumi Katsuno and Wolfgang Nejdl for pointing out an error in a previous version of this paper. Hirofumi Katsuno also supplied comments that helped simplify proofs. Early discussions with Craig Boutilier were very helpful in crystallizing the ideas presented here.

References

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

[2] M. Dalal. Investigations into a theory of knowledge base revision: Preliminary Report. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pages 475–479, 1988.

[3] J. Doyle. Rational belief revision (preliminary report). In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, pages 163–174, 1991.

[4] R. Fagin, J. Y. Halpern, Y. Moses, & M. Y. Vardi. *Reasoning about Knowledge.* To appear in 1992.

[5] R. Fagin, J. D. Ullman & M. Y. Vardi. On the semantics of updates in databases. In *Proceedings of the Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 352–365, 1983.

[6] A. Fuhrmann. Reflective modalities and theory change. *Synthèse*, (81) 115–1344, 1989.

[7] P. Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States.* Bradford Books, MIT Press, Cambridge, MA, 1988.

[8] A. Gibbard. Two recent theories of conditionals. In: *Ifs: Conditionals, Belief, Decision, Chance and Time*, W. L. Harper, R. Stalnaker & G Pearce (eds.). D. Reidel Publ. Co., Dordrecht, 1981, pp. 211–248.

[9] M. L. Ginsberg. Counterfactuals. *Artificial Intelligence*, (30) 35–79, 1986.

[10] G. Grahne. Updates and Counterfactuals. In: *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, pages 269–276, 1991.

[11] A. Grove. Two modelings for theory change. *Journal of Philosophical Logic*, (17) 157–70, 1988.

[12] J. Y. Halpern & Y. Moses. A guide to the modal logics of knowledge and belief. In: *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 480–490, 1985.

[13] J. Y. Halpern & M. Y. Vardi. Model checking vs. theorem proving: a manifesto. In: *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, pages 325–334, 1991.

[14] G. E. Hughes & M. J. Cresswell. *A Companion to Modal Logic.* Methuen, London, 1984.

[15] W.L.Harper & C.H. Hooker (eds). *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science.* D. Reidel Publ. Co., Dordrecht, 1976.

[16] H. Katsuno & A. O. Mendelzon. Propositional knowledgebase revision and minimal change. *Artificial Intelligence*, (52) 263–294, 1991.

[17] H. Katsuno & A. O. Mendelzon. On the difference between updating a knowledge base and revising it. In: *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, pages 387–394, 1991.

[18] I. Levi. Iteration of Conditionals and the Ramsey test. *Synthèse*, (76) 49–81, 1988.

[19] H. J. Levesque. All I Know: A study in autoepistemic logic. *Artificial Intelligence*, (42) 263–309, 1990.

[20] D. K. Lewis. *Counterfactuals.* Blackwell, Oxford, 1973.

[21] D. K. Lewis. Probabilities of conditionals and conditional probabilities. *Philosophical Review* (85) 297–315, 1976.

[22] B. Nebel. A knowledge level analysis of belief revision. In: *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 301–311, 1989.

[23] Donald Nute. Conditional logic. In Dov M. Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic*, D. Reidel, Dordrecht, 1984.

[24] A. S. Rao & N. Y. Foo. Minimal change and maximal coherence: A basis for belief revision and reasoning about action. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 966–971, 1989.

[25] M. Winslett. Reasoning about action using a possible models approach. In: *Proceedings of the Seventh National Conference on Artificial Intelligence*, pages 89–93, 1988.

[26] M. Winslett. *Updating Logical Databases.* Cambridge University Press, 1990.