

# Minimal Belief Change and the Pareto Principle

Oliver Schulte  
Department of Philosophy  
University of Alberta  
Edmonton, AB T6G 2E5  
Canada

October 3, 2001

## Abstract

This paper analyzes the notion of a minimal belief change that incorporates new information. I apply the fundamental decision-theoretic principle of *Pareto-optimality* to derive a notion of minimal belief change, for two different representations of belief: First, for beliefs represented by a *theory*—a deductively closed set of sentences or propositions—and second for beliefs represented by an axiomatic *base* for a theory. Three postulates exactly characterize Pareto-minimal revisions of theories, yielding a weaker set of constraints than the standard AGM postulates. The Levi identity characterizes Pareto-minimal revisions of belief bases: a change of belief base is Pareto-minimal if and only if the change satisfies the Levi identity (for “maxichoice” contraction operators). Thus for belief bases, Pareto-minimality imposes constraints that the AGM postulates do not.

The Ramsey test is a well-known way of establishing connections between belief revision postulates and axioms for conditionals (“if  $p$ , then  $q$ ”). Pareto-minimal theory change corresponds exactly to three characteristic axioms of counterfactual systems: a theory revision operator that satisfies the Ramsey test validates these axioms if and only if the revision operator is Pareto-minimal.

## 1 Minimal Belief Change

New information changes our beliefs continually. How should we incorporate new assertions into a body of existing ones? This question arises in many situations of philosophical and practical interest. For example, if the new assertion describes evidence about the world, incorporating the evidence into current beliefs is scientific and inductive reasoning. If the new assertion is a datum presented to a database system, we face the question of how to update a database. If the assertion is a new law, the issue becomes how to revise legal codes.

In the last two decades or so, the following principle has attracted much interest among philosophers, logicians and computer scientists: Revise your

beliefs so as to *minimize the extent of change* from the original beliefs.<sup>1</sup> The aim of this paper is to analyze the notion of minimal belief change. I derive axioms for minimal belief change from basic principles of *decision theory*. The same decision-theoretic principles lead to different results for different ways of formally representing beliefs. Specifically, I consider two such representations: Belief modeled as a deductively closed set of sentences or propositions, and belief modeled by an axiomatic “belief base”.<sup>2</sup> The analysis goes like this.

There are two kinds of changes to a theory  $T$ , viewed as a set of sentences. First, we may add a sentence to  $T$ , and second, we may retract a sentence from  $T$ . I say that a theory  $T'$  *adds* a sentence to  $T$  if  $T'$  entails a sentence that  $T$  does not entail; similarly, a theory  $T'$  *retracts* a sentence from  $T$  if  $T$  entails a sentence that  $T'$  does not entail. Given a current theory  $T$  and two possible revisions  $T_1$  and  $T_2$ , I say that  $T_1$  *adds more* than  $T_2$  if  $T_1$  adds all the sentences to  $T$  that  $T_2$  adds, and  $T_1$  adds some sentences that  $T_2$  does not add. Similarly,  $T_1$  *retracts more* than  $T_2$  if  $T_1$  retracts all the sentences from  $T$  that  $T_2$  retracts, and  $T_1$  retracts some sentences that  $T_2$  does not retract. (See Figure 1 in Section 3.)

Next, I observe that in theory revision, retractions and additions *trade off* against each other. That is, typically it is possible to avoid additions to theories if we are willing to retract more from them, and vice versa. Decision theory provides some general principles for dealing with trade-offs between different kinds of “costs”. The *Pareto principle* says that if an option  $O$  is no worse than an alternative  $O'$  on all dimensions of interest, and better than  $O'$  on some, we ought to prefer  $O$ ; in that case we say that  $O$  *Pareto-dominates*  $O'$ . A *lexicographic* choice procedure ranks the dimensions of interest by importance, then eliminates all options that are not optimal by the most important criterion; we break ties among these by eliminating the ones that are not optimal by the second most important criterion, etc.

I consider the implications of both the Pareto principle and lexicographic choice for minimal theory change. First, I define a Pareto-minimal theory revision to be one that is not Pareto-dominated with respect to additions and retractions. Thus Pareto-minimal theory revisions are those that cannot be improved by adding less without retracting more, or by retracting less without adding more. As it turns out, there is a purely set-theoretic definition of Pareto-minimal theory revisions in terms of the symmetric set differences between the current theory and alternative revisions. The main theorem of this paper establishes that certain axioms for belief revision characterize Pareto-minimal theory changes, in the sense that a theory change is Pareto-minimal if and only if the change satisfies these axioms. The chief difference between Pareto-minimality and the standard AGM postulates [Gärdenfors 1988] arises in the case in which the current theory neither entails the new information nor its negation. In that case, the AGM revision is the result of adding the new information to the cur-

---

<sup>1</sup>Quine’s principle of “minimal mutilation” is an early philosophical precursor [Quine 1951]. [Harman 1986] endorses this idea as epistemic “conservatism”.

<sup>2</sup>[Alchourrón and Makinson 1982] is an early study of the differences between these two representations.

rent theory. Pareto-minimal revisions, however, may be logically *weaker* than the AGM revision.<sup>3</sup>

Second, suppose that we lexicographically assign more importance to avoiding retractions than to avoiding additions (which captures some aspects of the idea that we ought to avoid “loss of information”). I provide a set of axioms that are necessary and sufficient for theory changes to be minimal in this sense. These axioms agree with the AGM postulates when the current theory is consistent with the new information. They disagree when the current theory is inconsistent with the new information: Then the retraction-minimizing revisions must yield a contradiction or a complete (maximal) set of beliefs. Thus, the notion of minimal change based on the Pareto principle and the notion of minimal change based on ranking retractions over additions each agree with a different part of the relevant AGM axioms. (See Figure 3 in Section 6.)

Pareto-optimality leads to different results for minimal revisions of belief bases, sets of sentences or propositions that need not contain all of their logical consequences. If we distinguish between an agent’s “basic beliefs” and the beliefs that follow from the basic beliefs, it is natural to make a corresponding distinction between changes in the agent’s basic beliefs and changes in the logical consequences of his basic beliefs. Thus I say that a belief base  $B'$  *adds* a sentence to another belief base  $B$  if  $B'$  contains a sentence that  $B$  does not contain; similarly, a belief base  $B'$  *retracts* a sentence from another belief base  $B$  if  $B$  contains a sentence that  $B'$  does not contain. It turns out that for belief bases, there is no tension between avoiding additions and avoiding retractions, and Pareto-minimal revisions of belief bases are those that minimize both kinds of change. They don’t add anything to the agent’s basic beliefs (save the new information) and they minimize retractions. The well-known *Levi identity* characterizes Pareto-minimal changes of belief bases: They are *exactly* those that result from, first, retracting just enough basic beliefs to make the agent’s basic beliefs consistent with the new information (technically, a “maxichoice contraction” [Gärdenfors 1988, Ch.4.2]), and second, adding the new information to the basic beliefs contracted in this manner. Since AGM revisions may give up more beliefs than maxichoice contraction permits, this characterization shows that Pareto-minimality yields some constraints on the revision of belief bases that the AGM axioms do not require (cf. [Alchourrón and Makinson 1982]).

Belief revision theorists have proposed various principles for rational belief change that do not follow from Pareto-minimality, though they are consistent with it. I give a generalized definition of Pareto-minimal belief change that accommodates any constraints on belief revision that one may wish to impose. This definition suggests directions of application for the Pareto principle beyond those investigated in this paper, such as principles of suppositional reasoning and conditional axioms along the lines of [Levi 1996].

One of the interesting aspects of belief revision axioms is their connection with axioms for conditionals (“if  $p$ , then  $q$ ”). In the last part of the paper,

---

<sup>3</sup>In this respect, Pareto-minimal revisions agree with Katsuno and Mendelzon’s approach to “belief update” [Katsuno and Mendelzon 1991]; see Section 4.

I follow a well-known approach, based on the so-called “Ramsey test” (see [Gärdenfors 1988, Ch.7]), to establish the conditional axioms that correspond to Pareto-minimal theory change. It turns out that Pareto-minimal theory change corresponds to three axiom schemata of Lewis’ and Stalnaker’s systems of counterfactuals [Lewis 1973], [Stalnaker 1968].

## 2 Theories

I begin with the representation of an agent’s current beliefs as a deductively closed *theory*. There are two main ways to represent theories, syntactically or semantically. For a syntactic representation, I assume that some language  $L$  has been fixed, and take a theory to be a deductively closed set of sentences or formulas from  $L$ . On a semantic approach, we take theories to be (conjunctions of) propositions, where propositions are suitable abstract objects such as sets of possible worlds. In this paper, I represent theories syntactically to facilitate comparison with the large part of the literature on belief revision and conditionals that takes a syntactic approach. However, it should be noted that all of the developments to follow are valid in a purely semantic, propositional setting as well.

As is usual in belief revision theory, my assumptions about the structure of the language in which an agent formulates her beliefs are sparse; essentially, all I assume is that the language features the usual propositional connectives. I take as given a suitable consequence relation between sets of formulas in the language, obeying the standard Tarskian properties. The formal presuppositions are as follows.

A **language**  $L$  is a set of formulas satisfying the following conditions.

1.  $L$  contains a **negation operator**  $\neg$  such that if  $p$  is a formula in  $L$ , so is  $\neg p$ .
2.  $L$  contains a **conjunction connective**  $\wedge$  such that if  $p$  and  $q$  are formulas in  $L$ , so is  $p \wedge q$ .
3.  $L$  contains an **implication connective**  $\rightarrow$  such that if  $p$  and  $q$  are formulas in  $L$ , so is  $p \rightarrow q$ .

A **consequence operation**  $Cn : 2^L \rightarrow 2^L$  represents a notion of entailment between sets of formulas from a language  $L$ . A set of formulas  $\Gamma$  **entails** another set of formulas  $\Gamma'$ , written  $\Gamma \vdash \Gamma'$  iff  $Cn(\Gamma) \supseteq \Gamma'$ . A set of formulas  $\Gamma$  entails a formula  $p$ , written  $\Gamma \vdash p$ , iff  $p \in Cn(\Gamma)$ . I assume that  $Cn$  satisfies the following properties, for all sets of formulas  $\Gamma, \Gamma'$ .

**Inclusion**  $\Gamma \subseteq Cn(\Gamma)$ .

**Monotonicity**  $Cn(\Gamma) \subseteq Cn(\Gamma')$  whenever  $\Gamma \subseteq \Gamma'$ .

**Iteration**  $Cn(Cn(\Gamma)) = Cn(\Gamma)$ .

A **theory** is a deductively closed set of formulas. That is, a set of formulas  $T \subseteq L$  is a theory iff  $Cn(T) = T$ .

The entailment relation  $\vdash$  is related to the propositional connectives as follows.

**Modus Ponens** If  $\Gamma \vdash p$ ,  $(p \rightarrow q)$ , then  $\Gamma \vdash q$ .

**Implication** If  $\Gamma \vdash q$ , then  $\Gamma \vdash (p \rightarrow q)$ .

**Deduction**  $\Gamma \cup \{p\} \vdash q$  iff  $\Gamma \vdash (p \rightarrow q)$ .

**Conjunction**  $\Gamma \vdash (p \wedge q)$  iff both  $\Gamma \vdash p$  and  $\Gamma \vdash q$ .

**Consistency** Suppose that  $\Gamma \not\vdash p$ . Then  $\Gamma \cup \{\neg p\} \not\vdash p$ .

**Inconsistency**  $\{p \wedge \neg p\} \vdash L$ .

**Double Negation**  $\Gamma \vdash p$  iff  $\Gamma \vdash \neg\neg p$ .

Classical propositional logic satisfies these assumptions. Belief revision theorists usually assume that the consequence relation  $Cn$  is compact; none of the results in this paper require compactness.<sup>4</sup>

For the remainder of this paper, assume that a language  $L$  and a consequence relation  $Cn$  (and hence an entailment relation  $\vdash$ ) have been fixed that satisfy the conditions laid down above.

### 3 Theory Change: Additions and Retractions

I now begin the analysis of what a minimal theory change is. An obvious approach to this question would be to define a metric  $\rho$  between theories, such that  $\rho(T_0, T_1)$  is a real number that measures the “distance” between two theories. If we had such a metric  $\rho$  at our disposal, we could define a minimal change from a current theory  $T_0$  to be another “closest” theory  $T_1$ , that is, a theory  $T_1$  such that there is no “closer” theory  $T_2$ . In symbols, a theory  $T_1$  is  $\rho$ -closest to  $T_0$  if there is no other theory  $T_2$  such that  $\rho(T_0, T_1) > \rho(T_0, T_2)$ . However, so far no satisfactory metric between theories has been designed.

A metric  $\rho$  between theories defines a *total order*  $\leq_T^\rho$  among possible new theories given a current theory  $T$ :  $T_1 \leq_T^\rho T_2$  iff  $\rho(T, T_1) \leq \rho(T, T_2)$ , where  $\leq$  denotes the standard ordering of the real numbers. My approach is to aim for less than a metric would provide, namely a *partial order*  $\prec_T$  where we read  $T_1 \prec_T T_2$  as “ $T_1$  is a smaller change from  $T$  than  $T_2$  is”. Since this ordering is partial, there may be possible changes that are incomparable. As far as a given partial order among theory changes goes, if two changes are incomparable, we should view neither as a smaller change than the other. However, a theory change  $T_2$  from an old theory  $T$  is *not* minimal if there is another, comparable,

<sup>4</sup>A consequence relation  $Cn$  is compact iff for all formulas  $p$  and sets of formulas  $\Gamma$ , we have that  $p \in Cn(\Gamma)$  only if  $p \in Cn(\Gamma')$  for some *finite* subset  $\Gamma'$  of  $\Gamma$ .

new theory  $T_1$  such that  $T_1 <_T T_2$ . Thus I shall take minimal changes from a current theory  $T$  to be the minimal elements in the given partial order  $<_T$ .

In these terms, the project of the first part of this paper is this: Define naturally motivated partial orders, and then characterize their minimal elements in terms of a belief revision operation  $*$ , such that  $*$  produces a minimal element if and only if  $*$  satisfies certain axioms.

I make use of decision-theoretic principles to define partial orders among theory changes. Let's begin by distinguishing two kinds of change: A *retraction* in which the old theory entails a formula that the new theory does not entail, and an *addition*, in which the new theory entails a formula that the old theory does not entail.

**Definition 1** *Let  $T, T'$  be two theories.*

1.  $T'$  **retracts** the formula  $p$  from  $T \iff T \vdash p$  and  $T' \not\vdash p$ .
2.  $T'$  **adds** the formula  $p$  to  $T \iff T \not\vdash p$  and  $T' \vdash p$ .

Next, I define two partial orders among theory changes by applying the principle of *dominance*. The first partial order defines a notion of a new theory  $T_1$  “retracting more” from a previous theory  $T$  than another new theory  $T_2$ , namely if  $T_1$  retracts all the formulas from  $T$  that  $T_2$  retracts from  $T$ , and  $T_1$  retracts at least one formula from  $T$  that  $T_2$  does not retract. The second partial order defines a notion of a new theory  $T_1$  “adding more” to a previous theory  $T$  than another new theory  $T_2$ , namely if  $T_1$  adds all the formulas from  $T$  that  $T_2$  adds to  $T$ , and  $T_1$  adds at least one formula to  $T$  that  $T_2$  does not add to  $T$ .

**Definition 2** *Let  $T, T_1, T_2$  be three theories.*

1.  $T_1$  **retracts more** formulas from  $T$  than  $T_2$  does  $\iff$ 
  - (a) for all formulas  $p$ , if  $T_2$  retracts  $p$  from  $T$ , then  $T_1$  retracts  $p$  from  $T$ , and
  - (b) for some formula  $p$ ,  $T_1$  retracts  $p$  from  $T$  and  $T_2$  does not retract  $p$  from  $T$ .
2.  $T_1$  **adds more** formulas to  $T$  than  $T_2$  does  $\iff$ 
  - (a) for all formulas  $p$ , if  $T_2$  adds  $p$  to  $T$ , then  $T_1$  adds  $p$  to  $T$ , and
  - (b) for some formula  $p$ ,  $T_1$  adds  $p$  to  $T$  and  $T_2$  does not add  $p$  to  $T$ .

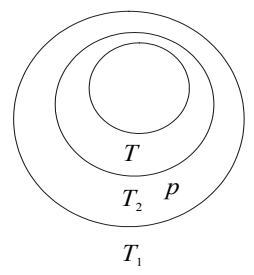
Thus  $T_1$  retracts more formulas from  $T$  than  $T_2$  iff  $T - T_2 \subset T - T_1$ , and  $T_1$  adds more formulas to  $T$  than  $T_2$  iff  $T_2 - T \subset T_1 - T$ , where  $\subset$  stands for proper set inclusion. Figure 1 illustrates these definitions.

We may think of the addition partial order and the retraction partial order as defining two distinct dimensions of “cost” in theory revision. If additions and retractions were linked such that minimizing one minimizes the other, this



$T_1$  retracts  $p$  from  $T$                        $T_2$  retracts nothing from  $T$

$T_1$  retracts more formulas from  $T$  than  $T_2$  does



$T_1$  adds  $p$  to  $T$                        $T_2$  does not add  $p$  to  $T$

$T_1$  adds more formulas to  $T$  than  $T_2$  does

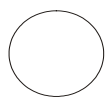
 : a theory = a deductively closed set of sentences  
 $T$

Figure 1: Dominance in Additions and Retractions

distinction would have no interesting consequences for the question of how to minimize theory change: we would just minimize both additions and retractions at once. What makes the distinction important is the fact that in general, additions and retractions *trade off* against each other. Typically, avoiding retractions entails adding more sentences than necessary, and avoiding additions entails retracting more sentences than necessary. An example will clarify this point.

**Example.** Imagine a cognitive scientist who believes that a certain AI system, say SOAR, is the only candidate for machine intelligence. This scientist believes that “if SOAR is not intelligent, there is no intelligent machine”. Letting  $s$  stand for “SOAR is intelligent” and  $m$  stand for “there is an intelligent machine”, the scientist believes the sentence  $p = \neg s \rightarrow \neg m$ . Suppose that the scientist believes only the consequences of  $p$ , that is, her current theory is  $T = Cn(\{p\})$ . In particular, the scientist neither believes that there is an intelligent machine ( $m$ ), nor does she believe that there is no intelligent machine ( $\neg m$ ). Now the scientist receives new information to the effect that SOAR is not intelligent. She has to revise her theory  $T$  on evidence  $\neg s$ . Let us consider two possible revisions,  $T_1$  and  $T_2$  (see Figure 2). Revision  $T_1$  adds the new information  $\neg s$  to  $T$  and accepts the deductive consequences of this addition; thus  $T_1 = Cn(\{p\} \cup \{\neg s\})$ . This revision  $T_1$  is logically stronger than  $T$  and hence retracts nothing from  $T$ . However, the revision adds the sentence  $\neg m$  (“there is no intelligent machine”), since  $p$  and  $\neg s$  entail  $\neg m$ .

Contrast this with a different revision  $T_2$  that retracts the scientist’s initial belief that SOAR is the only road to machine intelligence, and adds the new information that SOAR is not intelligent. That is,  $T_2 = Cn(\{\neg s\})$ . This revision  $T_2$  retracts more from  $T$  than  $T_1$  does. On the other hand,  $T_2$  adds less to  $T$  than  $T_1$  does, since  $T_2$  is strictly weaker than  $T_1$ . In particular,  $T_2$  continues to reserve judgment about whether machine intelligence is possible or not, whereas  $T_1$  concludes that it is impossible ( $\neg m$ ).

As the results below show, this example illustrates a general tension between avoiding additions and avoiding retractions; essentially, additions and retractions trade off against each other unless the current theory already entails the new information. When additions and retractions stand in conflict, how shall we make trade-offs between them? This is the topic of the next section.

## 4 Pareto-Minimal Theory Change

When a conflict arises between avoiding additions and avoiding retractions in belief revision, an agent may strike a subjective balance between them, as in any case of conflicting aims. She may assign one kind of change more subjective weight than the other, or favour some beliefs as more “entrenched” than others. I will come back to this idea in Section ???. But before we resort to subjective factors, we can look to decision theory for an objective constraint that applies to all agents seeking to minimize theory change. If avoiding changes is our aim, then we should avoid revisions that make more additions than necessary without



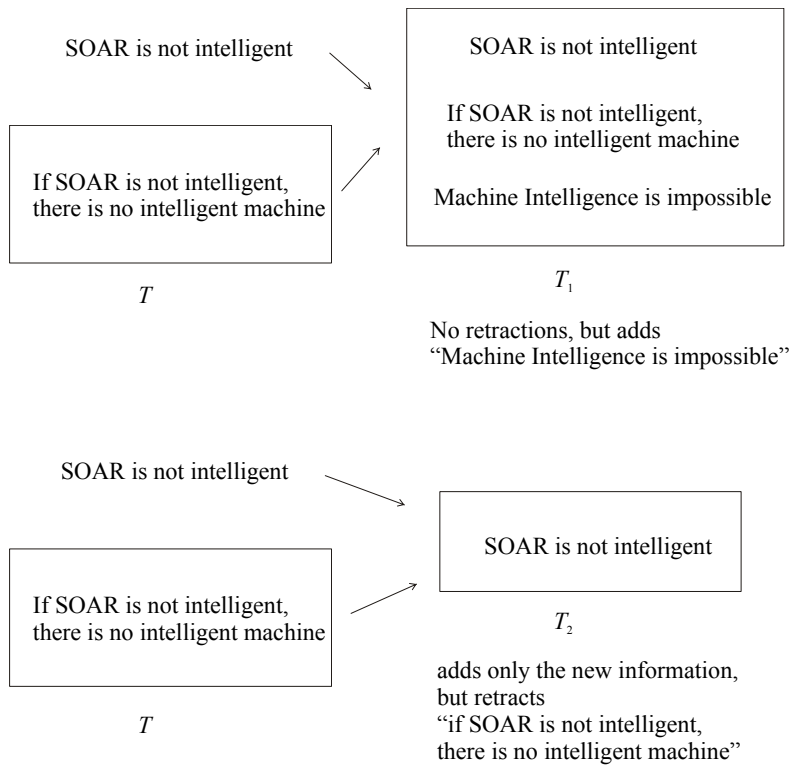


Figure 2: Additions and Retractions trade off against each other in theory revision: Revision  $T_1$  adds more to  $T$  than  $T_2$  does, and revision  $T_2$  retracts more from  $T$  than  $T_1$  does.

avoiding retractions, and we should avoid revisions that make more retractions than necessary without avoiding additions. This is an instance of the basic principle of *Pareto-optimality*. For minimal theory change, we can render it as follows.<sup>5</sup>

**Definition 3** *Let  $T, T_1, T_2$  be three theories.  $T_1$  is a **greater change** from  $T$  than  $T_2$  is  $\iff$*

1.  $T_1$  retracts more formulas from  $T$  than  $T_2$  does, and for all formulas  $p$ , if  $T_2$  adds  $p$  to  $T$ , then  $T_1$  adds  $p$  to  $T$ ; **or**
2.  $T_1$  adds more formulas to  $T$  than  $T_2$  does, and for all formulas  $p$ , if  $T_2$  retracts  $p$  from  $T$ , then  $T_1$  retracts  $p$  from  $T$ .

An equivalent purely set-theoretic definition is that  $T_1$  is a greater change from  $T$  than  $T_2$  is iff  $T_2 \triangle T \subset T_1 \triangle T$ , where  $\subset$  denotes proper inclusion and  $\triangle$  is symmetric difference ( $A \triangle B = A - B \cup B - A$ ). (This definition is due to Norman Foo.)

Thus the principle of Pareto-Optimality defines a partial relation  $\prec_T$  between theories:  $T_2 \prec_T T_1$  iff  $T_1$  is a greater change from  $T$  than  $T_2$  is. It seems that we can now take a minimal change from  $T$  to be a minimal theory in the  $\prec_T$ -ordering. But on that definition, the only minimal change from  $T$  is  $T$  itself! Of course, it is generally true that the smallest change is no change, on any acceptable notion of “small change” (cf. [Levi 1988, p.52, Condition (1)], [Lewis 1976, p.313]). What we want is a minimal change that satisfies *additional constraints*. In the case of belief update, the additional constraint is that the minimal theory change should incorporate the new information. Accordingly, I define a Pareto-minimal theory change from  $T$ , given new information  $p$ , as a theory that is minimal in the  $\preceq_T$ -ordering among the theories that entail  $p$ .

**Definition 4** *Let  $T, T_1$  be two theories, and let  $p$  be a formula. Then  $T_1$  is a **Pareto-minimal change** from  $T$  that incorporates  $p$   $\iff$*

1.  $T_1 \vdash p$ , and
2. there is no other theory  $T_2$  such that  $T_2 \vdash p$  and  $T_1$  is a greater change from  $T$  than  $T_2$  is.

Now we are ready for the main result of this paper: Necessary and sufficient conditions for a theory revision to be a Pareto-minimal change. It is not difficult to see that the following three conditions are necessary. Let us write  $T * p$  for the revision of theory  $T$  given new information  $p$ . First, it is our basic constraint that the revision  $T * p$  must entail  $p$ . Second, since the least change of a theory  $T$  is  $T$  itself, we don’t change the current theory at all if it already entails the new information  $p$ ; in symbols,  $T * p = T$ . Third, the revision  $T * p$  must follow from

---

<sup>5</sup>To obtain the appropriate definition for the propositional setting, replace the word “formulas” by “propositions” in the following definition.

the result of simply adding the new information to the old theory; formally, it must be the case that  $T \cup \{p\} \vdash T * p$ . For suppose that a revision  $T * p$  does not satisfy this condition. Then  $T * p$  entails a sentence  $q$  that is not entailed by  $T \cup \{p\}$ , and hence not by  $T$ . Consider the theory  $T'$  that entails a sentence  $r$  just in case both  $T * p$  and  $T \cup \{p\} \cup \{\neg q\}$  entail  $r$ . Clearly  $T'$  adds less to  $T$  than  $T * p$  does because  $T'$  is weaker than  $T * p$ ; in particular,  $T'$  does not add  $q$  to  $T$  whereas  $T * p$  does. Furthermore,  $T'$  retracts from  $T$  exactly those sentences that  $T * p$  retracts from  $T$ . For let  $r$  be a sentence entailed by  $T$  but not by  $T'$ . Then  $T \cup \{p\} \cup \{\neg q\}$  entails  $r$  and so by the definition of  $T'$ , it must be the case that  $T * p$  does not entail  $r$ . This argument shows that  $T * p$  is a greater change from  $T$  than  $T'$  is. Hence  $T * p$  is not a Pareto-minimal change unless  $T \cup \{p\}$  entails  $T * p$ . The proof of Theorem 5 in Section 12 formalizes these considerations, and shows that the three conditions listed are sufficient as well, that is, any theory revision that satisfies them is Pareto-minimal. Thus we have the following characterization of Pareto-minimal theory change that incorporates a given piece of new information (see Figure 3 in Section 6).

**Theorem 5** *Let  $T$  be a theory and let  $p$  be a formula. A theory revision  $T * p$  is a Pareto-minimal change from  $T$  that incorporates  $p \iff$*

1.  $T * p \vdash p$ , and
2.  $T \cup \{p\} \vdash T * p$ , and
3. if  $T \vdash p$ , then  $T * p = T$ .

The theorem shows that the tension between additions and retractions arises whenever the agent's current theory does not already entail the new information. When this is the case, the revisions that make Pareto-acceptable trade-offs run in strength from adding the evidence to the current theory ( $T \cup \{p\}$ ) to entailing nothing but the evidence and its consequences ( $\{p\}$ ).

## 5 Retraction-Minimal Theory Change and the AGM Axioms

Mention retraction-minimality. Also mention the AGM axioms.

This account of minimal change distinguishes sharply between the case in which the current theory already entails the new information and the case in which it does not. The standard AGM axioms also make a sharp distinction, but along a different line: They distinguish between the case in which the evidence is consistent with the current theory (but not necessarily already part of it) and the case in which the evidence is inconsistent with the current theory.

In my notation, the AGM axioms for theory revision are the following, for a given theory  $T$  and sentences  $p, q$  [Gärdenfors 1988, Ch.3.3]. For comparison with Pareto-minimal theory change, the relevant axioms are K\*1–K\*4; I will discuss the other axioms later.

**K\*1**  $T * p$  is a theory.

**K\*2**  $T \vdash p$ .

**K\*3**  $T \cup p \vdash T * p$ .

**K\*4** If  $T \cup p$  is consistent, then  $T * p \vdash T \cup p$ .

**K\*5**  $T * p$  is inconsistent just in case  $p$  is inconsistent.

**K\*6** If  $p$  and  $q$  are logically equivalent, then  $T * p = T * q$ .

**K\*7**  $T * p \cup \{q\} \vdash T * p \wedge q$ .

**K\*8** If  $T * p \cup \{q\}$  is consistent, then  $T * p \wedge q \vdash T * p \cup \{q\}$ .

## 6 Retraction-Minimal Theory Change

In some circumstances, we may think of our current theory as representing “valuable information”. This may be the case when the theory records reliable evidence reports, or when it contains entries in a data base, for example. In a Bayesian setting, a theory may be a record of the evidence on which a Bayesian conditioned her degrees of belief. From the Bayesian’s point of view, this evidence is “certain fact” rather than a “mere conjecture” in the sense that the Bayesian assigns probability 0 to the possibility of the evidence being false. There are other circumstances in which our theory represents our “best conjecture” or “current hypothesis” rather than “certain, valuable information”. This applies to many scientific theories, as well as the sort of tentative conclusions and generalizations on which we rely in everyday life.

When it is appropriate to regard our current theory as containing “valuable information”, we may want to change our theories in such a way that we give up as little of this information as possible. Gärdenfors calls this the “principle of information economy”.<sup>6</sup> One way of formulating this idea in decision-theoretic terms is to apply the principle of dominance to retractions: I say that a revision is *retraction-minimal* if no other revision retracts less.

**Definition 6** Let  $T, T_1$  be two theories, and let  $p$  be a formula. Then  $T_1$  is a *retraction-minimal change* from  $T$  that incorporates  $p \iff$

1.  $T_1 \vdash p$ , and
2. there is no theory  $T_2$  such that  $T_2 \vdash p$  and  $T_1$  retracts more from  $T$  than  $T_2$  does.

---

<sup>6</sup> “The next postulate for expansions can be justified by the ‘economic’ side of rationality. The key idea is that, when we change our beliefs, we want to retain as much as possible of our old beliefs—information is in general not gratuitous, and unnecessary losses of information are therefore to be avoided. This heuristic criterion is called the criterion of *information economy*.” [Gärdenfors 1988, p.49]; emphasis is Gärdenfors’.

Clearly any revision  $T * p$  that is at least as strong as the original theory  $T$  is retraction-minimal because it retracts nothing from  $T$ . Conversely, if  $q$  is any sentence that  $T * p$  retracts from  $T$ , then  $T * p$  retracts more from  $T$  than  $T * p \cup \{q\}$  does. Hence retraction-minimal revisions are those don't give up any beliefs.

**Proposition 7** *Let  $T$  be a theory and let  $p$  be a formula. A theory revision  $T * p$  is a retraction-minimal change from  $T$  that incorporates  $p \iff T * p \vdash T \cup \{p\}$ .*

If our aim is minimal change, it is natural to strengthen the principle of avoiding retractions by selecting among the retraction-minimal revisions those theories that minimize additions. In decision-theoretic terms, this amounts to lexicographically assigning highest priority to avoiding retractions, and second highest to avoiding additions. To be precise, say that a revision  $T * p$  is **addition-minimal among retraction-minimal revisions** iff  $T * p$  is retraction-minimal,  $T * p$  entails  $p$ , and there is no other retraction-minimal revision  $T'$  entailing  $p$  such that  $T'$  adds less than  $T * p$  to  $T$ . Thus if  $T * p$  is addition-minimal among retraction-minimal revisions, then  $T * p$  is Pareto-minimal. Let  $T * p$  be such a revision. By Proposition 7, we have that  $T * p \vdash T \cup \{p\}$ . And it follows from Theorem 5 that conversely  $T * p$  must entail  $T \cup \{p\}$ . Hence  $T * p$  is just the result of adding the new information  $p$  to the previous theory  $T$ .

**Proposition 8** *Let  $T$  be a theory and let  $p$  be a formula. A theory revision  $T * p$  is addition-minimal among retraction-minimal revisions  $\iff T * p = Cn(T \cup \{p\})$ .*

Figure 3 summarizes the characterizations of Pareto-minimal and retraction-minimal theory change in comparison with the AGM postulates.

When the new information contradicts the current theory, Proposition 7 implies that retraction-minimality will lead an agent to adopt an inconsistent theory. The obvious answer to this problem is that we should reinterpret the “principle of information economy”: what we want is a retraction-minimal *consistent* revision. Formally, we can express this idea by rephrasing Definition 6 such that a theory change  $T * p$  is retraction-minimal just in case  $T * p$  entails  $p$  and there is no *consistent* theory  $T'$  such that  $T'$  entails  $p$  and  $T * p$  retracts more from  $T$  than  $T'$  does. However, rather surprisingly it follows from a result of [Alchourrón and Makinson 1982, Observation 3.2] that under this definition, avoiding retractions requires the revision  $T * p$  to be a *complete* theory when  $T$  is inconsistent with  $p$ .<sup>7</sup> For this reason, many belief revision theorists advise against applying retraction-minimality when the new information contradicts the agent’s current beliefs (see [Gärdenfors 1988, pp.58–59] and [Levi 1996,

<sup>7</sup>Briefly, the reason is this. Since the previous theory  $T$  entails  $\neg p$ , it also entails  $p \rightarrow q, p \rightarrow \neg q$  for all atomic formulas  $q$ . Consistency requires that an agent remove one of the pairs  $p \rightarrow q, p \rightarrow \neg q$ , but the aim of minimizing retractions prevents an agent from removing both.

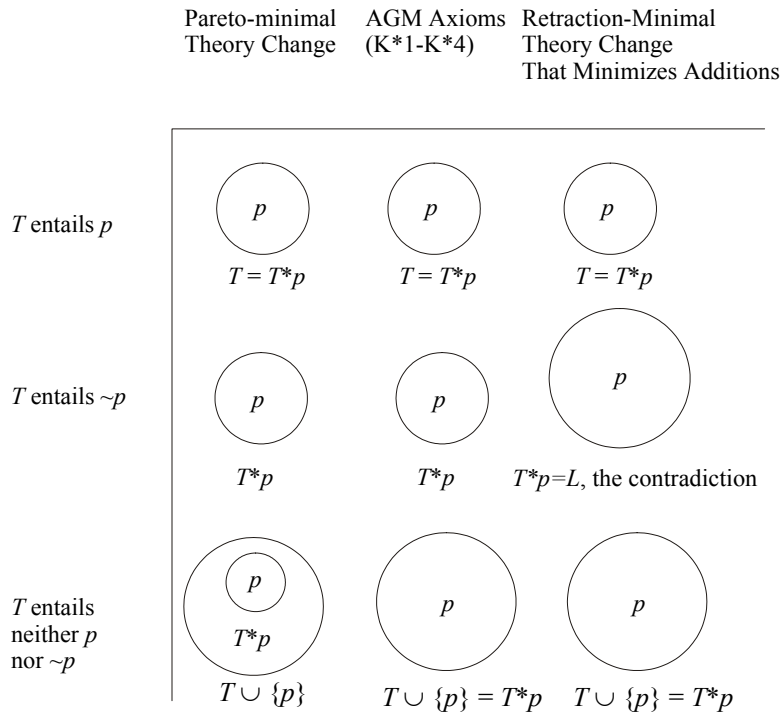


Figure 3: The behaviour of Pareto-minimal, retraction-minimal and AGM belief revision operators, depending on whether the current theory entails the new information, entails its negation, or is consistent with both.

p.22])).<sup>8</sup> In an important study, Alchourrón and Makinson suggested a different approach: If we refine the representation of an agent’s beliefs, minimizing retractions no longer leads to complete theories. The next section explores this idea.

## 7 Pareto-Minimal Revision of Belief Bases

So far I have treated all of an agent’s beliefs as equally important. A more refined representation of the agent’s epistemic state may distinguish between a “basic” set of beliefs  $B$ , and the consequences of  $B$  that the agent might be said to hold because he believes  $B$ .<sup>9</sup> [Hansson 1998] endorses the distinction between a basic set of beliefs and their consequences as a “small step toward capturing the justificatory structure” of an agent’s beliefs.<sup>10</sup> I shall take a **base** for a theory  $T$  to be a set of formulas  $B$ , which may or may not be deductively closed, such that  $B \vdash T$ .

Unlike [Alchourrón and Makinson 1982], I do not require that belief bases be “irredundant” in the sense that none of the basic beliefs follow from the others. The goal is not a compact representation of an agent’s beliefs but rather differentiating among his beliefs to capture some of their justificatory relationships. For example, a scientist may wish to distinguish between holding a belief  $p$  because her current scientific theory  $H$ —included among her basic beliefs—predicts  $p$ , and holding a belief  $p$  after testing her theory and observing  $p$ . In the second case, she may well include  $p$  among her basic beliefs along with other observation data; but then if  $H$  also remains among her basic beliefs, her basic belief  $p$  will follow from other basic beliefs (namely  $H$ ).<sup>11</sup>

We could formulate the notion of a base in a purely propositional setting by taking bases to be sets of propositions (sets of sets of possible worlds), rather than just a single proposition (the intersection of all propositions believed by the agent). Representing belief as the intersection of all propositions believed by the agent amounts to choosing a single basic belief—namely the conjunction of his beliefs—to represent his epistemic state; this may not reflect the finer justificatory structure of his beliefs. Mathematically, all the results to follow are

<sup>8</sup>[Levi 1996, Ch.2.1] presents a theory of how an agent may minimize the loss of “damped informational value”. In my terms, this is advice for how to retract some beliefs to avoid adding too many.

<sup>9</sup>A paradigm example is a database, where we may distinguish between the records that are explicitly stored in the database and what follows from the explicitly stored information.

<sup>10</sup>Some authors cite psychological plausibility and computational feasibility as other reasons for introducing this distinction. [Harman 1986] proposes to distinguish between the set of beliefs that an agent may hold “explicitly” at a given time—of which in some sense she may be “conscious” or “aware”—and the consequences of these beliefs which she holds “implicitly” at that time. [Alchourrón and Makinson 1982, p.21] argue that we ought to think of an agent’s belief set as being “generated” from finitely many basic beliefs. Since my motivation for using the notion of a belief base is not psychological plausibility or computational feasibility, I make no assumptions about whether a given belief base is finite or even recursively enumerable.

<sup>11</sup>This is not a “foundations theory” in the sense of [Harman 1986] because I do not require agents to keep a record of their epistemic history. Nor do I assume that an agent’s basic beliefs are themselves justified by other beliefs or by anything else. See [Gärdenfors 1988, Ch.2].

valid for belief bases represented by sets of propositions (substitute propositions for sentences in the definitions and arguments below).

Setting apart basic beliefs from beliefs that follow from them is relevant to the theory of minimal belief change because it naturally suggests a corresponding distinction between changes in the basic beliefs and changes in the consequences of those beliefs. Changes in the belief base may be very small even when many changes in the agent’s overall beliefs occur. Some examples will illustrate this phenomenon. Suppose that an agent has beliefs that are widely indeterminate in the sense that they entail few other beliefs, but at the same time, settling one more question will lead him to very strong beliefs. For a formal example, we may take the agent’s (basic) beliefs to be the set  $\{p \rightarrow q, \neg p \rightarrow \neg q : q \text{ is an atomic formula}\}$  for some fixed atomic formula  $p$ . Although currently the agent does not accept any atomic formula, if she next comes to learn either  $p$  or not  $\neg p$ , and includes this information in her basic beliefs, her basic beliefs will entail a complete theory. For an informal example, consider a physicist who assumes that a physical system under study satisfies a certain differential equation. Before any observations, the differential equation is consistent with a wide range of system trajectories; but given an initial condition (or more generally, a small number of observations), the differential equation may determine a unique trajectory, leading to a definite prediction about the system’s behaviour for each future time.

In these situations, how great is the extent of the agent’s change in his beliefs after adding the relevant evidence ( $p$  resp.  $\neg p$  or the initial condition)? One answer is that since the agent adopted many new beliefs, the “distance” of the new theory from his previous one is great. Another view is that the belief change is small, because only one piece of new information was added to the agent’s *basic* beliefs. Although after adding the new information, the agent’s basic beliefs entail many new assertions, we may choose not to count these because they “just follow” from the agent’s basic beliefs. Both views of minimal belief change seem to be defensible. I considered the first beforehand and now turn to the investigation of the latter.

I begin again with two ways of making a change to a belief base. If  $B, B'$  are two bases, I say that  $B'$  **retracts** the formula  $p$  from  $B$  iff  $p \in B$  and  $p \notin B'$ , and that  $B'$  **adds** the formula  $p$  to  $B$  iff  $p \notin B$  and  $p \in B'$ . The definition of “adding more” and “retracting more” from a base is just like that for theories (cf. Definition 2). Thus  $B_1$  **retracts more** formulas from  $B$  than  $B_2$  iff  $B - B_2 \subset B - B_1$ , and  $B_1$  **adds more** formulas to  $B$  than  $B_2$  iff  $B_2 - B \subset B_1 - B$ , where  $\subset$  stands for proper set inclusion.

As with Definition 4, we can apply the principle of Pareto-optimality to define a partial comparison of base revisions with respect to the extent of change that they induce.

**Definition 9** *Let  $B, B_1, B_2$  be three bases. Then  $B_1$  is a **greater change** from  $B$  than  $B_2$  is  $\iff$*

1.  $B_1$  *retracts more formulas from  $B$  than  $B_2$  does, and for all formulas  $p$ , if  $B_2$  adds  $p$  to  $B$ , then  $B_1$  adds  $p$  to  $B$ ; or*



2.  $B_1$  adds more formulas to  $B$  than  $B_2$  does, and for all formulas  $p$ , if  $B_2$  retracts  $p$  from  $B$ , then  $B_1$  retracts  $p$  from  $B$ .

As with Definition 3, an equivalent purely set-theoretic definition is that  $B_1$  is a greater change from  $B$  than  $B_2$  is iff  $B_2 \triangle B \subset B_1 \triangle B$ .

One fundamental difference between the Pareto-minimal revision of basic beliefs and Pareto-minimal theory change is this: Pareto-minimal base revisions never add beliefs to the basic ones other than the new information. For suppose that a revision  $B * p$  adds a belief  $q$  to a base  $B$ ; then  $B * p - \{q\}$  adds less to  $B$  and retracts no more. Hence  $B * p$  is not a Pareto-minimal change of  $B$ . Another way to put the point is that, in contrast with theories, for bases the conflict between additions and retractions does not arise: it is possible to minimize both additions and retractions at the same time. In the case in which the new information contradicts the current basic beliefs, this will lead an agent to hold inconsistent beliefs. Since many writers accept as a general norm of epistemic rationality that an agent ought to avoid inconsistent beliefs, I shall restrict Pareto-minimal revisions to consistent bases.

**Definition 10** *Let  $B, B_1$  be two bases, and let  $p$  be a formula. Then  $B_1$  is a **Pareto-minimal consistent change** from  $B$  that incorporates  $p \iff$*

1.  $p \in B_1$ , and
2.  $B_1$  is consistent, and
3. there is no other consistent base  $B_2$  such that  $p \in B_2$  and  $B_1$  is a greater change from  $B$  than  $B_2$  is.

Since for basic beliefs, there is no tension between additions and retractions, Pareto-minimality yields the same result as lexicographically assigning the highest weight to avoiding retractions: Pareto-minimal revisions of belief bases are exactly the retraction-minimal revisions that minimize additions. I define a retraction-minimal base revision that produces consistent bases as follows (cf. Definition 6). Let  $B, B_1$  be two theories, and let  $p$  be a formula. Then  $B_1$  is a **retraction-minimal consistent change** from  $B$  that incorporates  $p$  iff (1)  $p \in B_1$ , and (2)  $B_1$  is consistent, and (3) there is no other consistent base  $B_2$  such that  $p \in B_2$  and  $B_1$  retracts more formulas from  $B$  than  $B_2$  does. The next proposition asserts that Pareto-minimality applied to changes in basic beliefs collapses into first minimizing retractions and then minimizing additions.

**Proposition 11** *Let  $B$  be a base and let  $p$  be a formula. Then  $B * p$  is a **Pareto-minimal consistent change** from  $B$  that incorporates  $p \iff B * p$  is a **retraction-minimal consistent change** from  $B$  that incorporates  $p$  and minimizes additions.*

An important insight of [Alchourrón and Makinson 1982] is that when new information contradicts the agent's current basic beliefs, the consistent retraction-minimal revisions are not necessarily complete theories. For the simplest example, let  $B = \{p\}$  be the agent's current beliefs, and suppose that the agent learns

$\neg p$ . Then the consistent Pareto-minimal revision of  $B$  is  $B * \neg p = \{\neg p\}$ , clearly not a complete theory. The difference to Proposition 7 is this:  $B$  entails all material implications of the form  $\neg p \rightarrow q$ ,  $\neg p \rightarrow r$ , etc. for all atomic formulas  $q, r, \dots$ . The revision  $B * \neg p$  no longer entails these implications. However, since these implications are “only” consequences of  $B$ , not themselves basic beliefs, retracting them is not retracting a basic belief. Thus  $B * \neg p = \{\neg p\}$  minimizes retractions of the agent’s basic beliefs even though it retracts many of the logical consequences of the agent’s basic beliefs.

What are the characteristic properties of Pareto-minimal base revisions? It turns out that a version of a proposal originally due to Levi amounts to necessary and sufficient conditions for a base revision to be Pareto-minimal and consistent. The proposal is to think of a Pareto-minimal revision of a belief base  $B$  on new information  $p$  as proceeding in two steps: First, remove just enough beliefs from  $B$  to obtain a belief base  $B'$  that is consistent with  $p$ ; then add  $p$  to  $B'$ . Formally, we require that  $B'$  be a belief base that is consistent with  $p$ —thus  $B' \not\vdash \neg p$ —and removes as few beliefs from  $B$  as possible. Hence I define a retraction-minimal *contraction* of a belief base as follows.

**Definition 12** *Let  $B, B_1$  be two bases, and let  $p$  be a formula. Then  $B_1$  is a **retraction-minimal contraction** from  $B$  on  $p \iff$*

1.  $B_1 \subseteq B$ , and
2.  $B_1 \not\vdash p$ , and
3. there is no other base  $B_2$  such that  $B_2 \not\vdash p$  and  $B_1$  retracts more from  $B$  than  $B_2$  does.

Retraction-minimal contractions of a base  $B$  on new information  $p$  have a simple characterization: They are exactly those subsets of  $B$  that cannot be expanded without entailing  $p$ .

**Lemma 13** *Let  $B, B_1$  be two bases such that  $B_1 \subseteq B$ , and let  $p$  be a formula. Then  $B_1$  is a **retraction-minimal contraction** from  $B$  on  $p \iff$  for all formulas  $q$ , if  $B_1$  retracts  $q$  from  $B$ , it is the case that  $B_1 \cup \{q\} \vdash p$ .*

Thus retraction-minimal contractions are those that belief revision theorists refer to as “maxichoice contractions” [Gärdenfors 1988, Ch.4.2]. The **Levi identity** says that minimal revisions of a belief set  $K$  given new information  $p$  are the result of adding  $p$  after contracting  $K$  on  $\neg p$  (see [Gärdenfors 1988, Ch.3.6]). The next theorem shows that the Levi identity for retraction-minimal (maxichoice) contractions characterizes Pareto-minimal revisions of belief bases that lead to consistent belief bases.

**Theorem 14** *Let  $B$  be a base and let  $p$  be a formula. Suppose that a revision  $B * p$  contains  $p$ . Then  $B * p$  is a Pareto-minimal consistent change from  $B$  that incorporates  $p \iff$  there is a retraction-minimal contraction  $B'$  from  $B$  on  $\neg p$  such that  $B * p = B' \cup \{p\}$ .*

Alchourrón and Makinson conjectured that “when applied to bases that are irredundant, choice contraction and revision functions serve as good formal representations of the corresponding intuitive processes” [Alchourrón and Makinson 1982, p.21]. Theorem 14 establishes a formal version of this conjecture, in which bases need not be irredundant and Pareto-minimality takes the place of “intuition”.

In view of Theorem 14, it is not difficult to see that Pareto-minimal consistent revisions of belief bases satisfy the AGM axioms K\*1–K\*5 (interpreted for base revisions with  $\supseteq$  in place of  $\vdash$ ; see also [Alchourrón and Makinson 1982, Part II]).<sup>12</sup> The converse is not true, however: Pareto-minimality places more constraints on the revision of belief bases than K\*1–K\*5, since AGM revisions need not be the result of maxichoice contractions and hence may give up more beliefs than Pareto-minimal revisions. On the other hand, Pareto-minimality does not require compliance with the other AGM postulates (K\*6–K\*8). The next section shows how to combine Pareto-minimality with any desired extra constraints on belief revision.

## 8 A Generalized Definition of Pareto-Minimal Belief Change

Belief revision theorists appeal to the principle of minimal belief change as well as to general considerations of epistemic rationality to justify norms for belief revision.<sup>13</sup> For example, the requirement that an agent should not adopt inconsistent beliefs (K\*5) and the principle that logically equivalent information should lead to the same revisions (K\*6) appear to express general principles of rational belief revision rather than means of minimizing the extent of belief change. Since the smallest change is no change, if an agent happens to have inconsistent beliefs, which trivially entails the new information, the minimal change of his beliefs will be none; that is, the minimal revision of the inconsistent theory will always be the inconsistent theory. K\*5 however requires that an agent remedy inconsistencies (the so-called “success postulate”; see [Arló Costa 1990] for a critique of this postulate). So it seems that sometimes at least belief revision theorists are willing to put general epistemic rationality before the aim of minimizing the extent of belief change.<sup>14</sup> Moreover, we saw in Section 7 that some of the most interesting applications of the Pareto principle come from combining it with rationality principles such as logical consistency. The next definition

<sup>12</sup>For K\*2 I require that  $p \in B * p$ . For K\*5 we must assume that the underlying consequence relation  $\vdash$  is consistent in the sense that  $\emptyset \not\vdash L$ ; otherwise there is no consistent base. When the new information  $p$  is inconsistent, there is no consistent revision on  $p$ ; in that case I require that  $B * p$  is an inconsistent base in accordance with K\*5.

<sup>13</sup>For example, Gärdenfors writes with reference to the AGM postulates that “these [belief] changes are characterized by a number of *rationality postulates*” [Gärdenfors 1988, p.3], emphasis Gärdenfors’. Or in another passage: “my main goal in this section is to delimit the meaning of ‘minimal change’ by formulating some rationality postulates that apply to revisions of belief sets” [Gärdenfors 1988, p.53].

<sup>14</sup>[Rott 1998a] discusses an interpretation of the AGM framework in terms of the general theory of rational choice.

allows us to apply the Pareto Principle together with any rationality principles of interest. Recall that Definition 9 derives from the Pareto principle a ternary partial relation between belief bases “ $B_1$  is a greater change from  $B$  than  $B_2$ ”.

**Definition 15** *Let  $B, B_1$  be two bases, and let  $p$  be a formula. Let  $C$  be any constraint relating  $B, B_1$  and  $p$ .<sup>15</sup> Then  $B_1$  is a **Pareto-minimal change** from  $B$  satisfying  $C \iff$*

1.  $B_1$  satisfies  $C$ , and
2. there is no base  $B_2$  satisfying  $C$  such that  $B_1$  is a greater change from  $B$  than  $B_2$  is.

Of course we may define Pareto-minimal change given an arbitrary constraint for theories in just the same way by using Definition 3. Some examples to illustrate Definition 15: In Definition 4 I took the constraint  $C$  to be “ $B_1$  must entail  $p$ ”. Definition 10 is a special case of Definition 15 with the constraint “ $B_1$  must contain  $p$  and be consistent”. Considering theories, we may also take the constraint  $C$  as “ $T_1$  satisfies the AGM postulates for revisions of  $T$  on  $p$ , except for possibly K\*3”. By arguments similar to those in Section 4, we would then obtain K\*3 as expressing the characteristic condition of Pareto-minimal theory change. According to this interpretation, K\*4—the preservation principle—would be justified on grounds of general epistemic rationality, but—in contrast with K\*3—not as a universal principle for minimizing the extent of theory change.<sup>16</sup> (Section 7 suggested a different interpretation of the preservation principle as part of minimal revisions of belief *bases*.)

Definition 15 points to new applications of the Pareto principle. Here’s one: Motivated by ideas of Frank Ramsey, Isaac Levi has suggested an analysis of suppositional reasoning (“suppose that  $p$ ”) in which an agent first, as it were, clears his mind with respect to  $p$ , and then adds  $p$  to his “stock of beliefs” [Levi 1996]. Formally, we may take this as the constraint that a change  $B_1$  from original beliefs  $B$  on  $p$  must neither entail  $p$  nor  $\neg p$ . An interesting open question is what axioms of suppositional reasoning characterize Pareto-minimal belief change satisfying this constraint.

We may ask the same question assuming that the agent directly adds the supposition  $p$  to his beliefs. This leads to a familiar path that connects belief revision postulates with axioms for conditionals via the so-called “Ramsey test”. In the final part of this paper, I combine the Ramsey test with Pareto-minimal theory change to derive principles for reasoning about conditionals.

<sup>15</sup>Formally,  $C$  is a ternary relation  $C(B, B', p)$ , where  $B, B'$  are arbitrary bases and  $p$  is an arbitrary formula; a base  $B'$  satisfies  $C$  given a base  $B$  and formula  $p$  iff  $C(B, B', p)$  holds.

<sup>16</sup>[Gärdenfors 1988, Ch. 7.4] supports the preservation principle by citing the “criterion of informational economy” and the fact that the principle is “strongly endorsed within the Bayesian tradition”. [Levi 1988] too defends it as a principle of general epistemic rationality. [Putnam 1963] refers to the principle as “tenacity” and [Kelly *et al.* 1995] label it “stubbornness”; these papers ask whether the principle helps or hinders an agent in reliably arriving at correct beliefs. In a nutshell, the answer is that the preservation principle does not help but need not prevent a sufficiently careful agent from finding the truth. See also [Martin and Osherson 1998].

## 9 Axioms for Conditionals

It is well-known among philosophers that material implication does not capture the meaning of many conditionals that we use in nonmathematical discourse, for example conditionals that express causal connections or counterfactuals. One reason for this is that ordinary conditionals typically are *defeasible*, whereas material implication is not. That is, if  $p \rightarrow q$ , then  $p \wedge r \rightarrow q$ , for any claim  $r$ ; but it seems to be quite consistent to assert that “if the match is struck, it will light” at same time as “if the match is struck and it is wet, it will not light” (to use a classic example from Nelson Goodman). Thus there is good reason to add a conditional operator  $>$  to our formal language  $L$ ; we read  $p > q$  as “if  $p$ , then  $q$ ”. From now on, assume that we have fixed a consequence relation  $Cn_{>}$  extended to the language  $L_{>}$  with conditionals that satisfies the properties of consequence relations specified in Section 2. For the remainder of the paper,  $\vdash$  denotes the entailment relation induced by  $Cn_{>}$ .

How does the conditional  $>$  behave? Logicians have studied this question in depth; for example, the well-known Lewis-Stalnaker approach specifies truth-conditions for conditionals in terms of distances between possible worlds [Stalnaker 1968, Lewis 1973]. This section considers the question: What is the relationship between the conditionals that an agent accepts and the ways in which she revises her beliefs? Formally, what connections are there between the conditional  $>$  and the belief revision operator  $*$ ? An interesting proposal is that if an agent believes  $p > q$ , then upon learning  $p$ , the agent should accept  $q$  as true. That is, if  $T \vdash p > q$ , then  $T * p \vdash q$ . If we require the converse as well, we obtain what has come to be called the *Ramsey test*: An agent should believe a conditional  $p > q$  just in case after revising his beliefs on the assertion that  $p$ , the agent believes  $q$ . The interpretation of the Ramsey test depends on how we view conditionals. If we accept that conditionals have a meaning (truth-conditions) independent of the Ramsey test—a la Lewis and Stalnaker—then we may think of the Ramsey test as defining a kind of diachronic consistency, a coherence requirement on how theory changes relate to beliefs about conditionals. More ambitiously, we may view the Ramsey test or a similar connection between belief revision and accepting conditionals as part of the meaning of the conditional  $>$ . Or we might say that the Ramsey test applies to suppositional reasoning, rather than actual updates. [Arló Costa 1998] and [Levi 1996] are recent treatments of these issues. I don’t intend to take a stance on these questions, but leave the mathematical results that follow open to philosophical interpretation.

Here is the formal definition of the Ramsey test.

**Definition 16 (Gärdenfors)** *A belief revision system  $\mathcal{T} = \langle \mathbf{T}, * \rangle$  is a set of theories in a language  $L_{>}$  with conditionals and a belief revision operator  $* : \mathbf{T} \times L_{>} \rightarrow \mathbf{T}$  such that  $\mathbf{T}$  is closed under additions. That is, if  $T \in \mathbf{T}$  and  $p \in L_{>}$ , then  $T \cup \{p\} \in \mathbf{T}$ .*

**Definition 17** *A belief revision system  $\mathcal{T} = \langle \mathbf{T}, * \rangle$  satisfies the Ramsey test iff  $T \vdash p > q \Leftrightarrow T * p \vdash q$ , for all theories  $T \in \mathbf{T}$ , and formulas  $p, q \in L_{>}$ .*

Consider a belief revision system  $\langle \mathbf{T}, * \rangle$  that satisfies the Ramsey test. Suppose we place constraints on the belief revision operator  $*$ ; in particular, suppose we require that  $*$  is a Pareto-minimal belief revision operator. Then are there axioms governing the conditional  $>$  that are valid in  $\langle \mathbf{T}, * \rangle$  in the sense that all theories in the system entail them? The next theorem establishes that there are such axioms and what they are. Let's first give a definition of what it is for a formula to be valid in a belief revision system.

**Definition 18** *A formula  $p$  is valid in a belief revision system  $\mathcal{T} = \langle \mathbf{T}, * \rangle$  iff all theories  $T \in \mathbf{T}$  entail  $p$ .*

Now we are ready to establish which conditional axioms exactly characterize Pareto-minimal theory change.

**Theorem 19** *Let  $\mathcal{T} = \langle \mathbf{T}, * \rangle$  be a belief revision system satisfying the Ramsey test. Then  $*$  is a Pareto-minimal belief revision operator  $\iff \mathcal{T}$  validates*

1.  $p > p$ , and
2.  $(p > q) \rightarrow (p \rightarrow q)$ , and
3.  $(p \wedge q) \rightarrow (p > q)$

for all formulas  $p, q, r$ .

The formal proof is in Section 12. The arguments proceed along lines familiar from similar results (cf. [Arló Costa 1995], [Gärdenfors 1988, Ch.7]).

## 10 Comparison with Other Axiom Systems for Conditionals

Logicians have proposed many axioms systems for the conditional  $>$ . I will compare the three axioms from Theorem 19 with Lewis' axiomatization of counterfactuals, his *VC* system. The three conditional axioms for minimal theory revision are part of Lewis' system. In addition, Lewis has the principle  $(\neg p > p) \rightarrow (q > p)$ . This principle is valid for Pareto-minimal theory changes that lead into contradiction only from contradiction, or when an inconsistent new assertion has to be incorporated (cf. [Arló Costa 1990]). The aim of avoiding contradictions provides a rationale for adding this consistency requirement to Pareto-minimal theory change, along the lines of Section 8.

*VC* features one more axiom schema:  $(p > \neg q) \vee (((p \wedge q) > r) \equiv (p > (q \rightarrow r)))$ . This axiom may fail to be valid in a belief revision system that satisfies the Ramsey test with a Pareto-minimal belief revision operator.

I shall briefly compare the approach to conditionals that comes from connecting Pareto-minimal belief revision and the Ramsey test to similar results in the literature. Gärdenfors shows that  $K^*4$ , the preservation principle, is incompatible with the Ramsey test; for the details see [Gärdenfors 1988, Ch.7,

Sec. 4] (cf. also [Arló Costa 1998]). For that reason, when he investigates what conditional axioms correspond to the AGM postulates for belief revision, he replaces  $K^*4$  by the following weaker principle  $K^*4w$ : If  $T \vdash p$  and  $T$  is consistent, then  $T * p \vdash T$ . A Pareto-minimal belief revision operator  $*$  satisfies  $K^*4w$ , since  $T * p = T$  whenever  $T \vdash p$  by Theorem 5, Clause 3. If we want to use the Ramsey test to connect belief revision postulates with acceptance criteria for conditionals, Pareto-minimality has a clear advantage over the AGM axioms: The characteristic conditions for Pareto-minimal theory change translate directly into interesting axioms for conditionals, whereas the AGM approach has to drop the preservation principle  $K^*4$ —the very postulate that is the main difference between AGM and Pareto-minimal theory change.<sup>17</sup>

Gärdenfors defines validity in a belief revision system differently from my Definition 18. His definition is as follows. First, a formula  $p$  is **satisfiable in a belief revision system**  $\mathcal{T} = \langle \mathbf{T}, * \rangle$  iff there is some consistent theory  $T \in \mathbf{T}$  such that  $T$  entails  $p$ . Second, a formula  $p$  is **negatively valid in a belief revision system**  $\mathcal{T} = \langle \mathbf{T}, * \rangle$  iff  $\neg p$  is not satisfiable in  $\mathcal{T}$ . (I take the term “negative validity” from [Arló Costa 1995].)

Validity in my sense implies negative validity. For let  $\mathcal{T} = \langle \mathbf{T}, * \rangle$  be a belief revision system, and suppose that  $p$  is valid in  $\mathcal{T}$  in the sense of Definition 18. Then any theory  $T \in \mathbf{T}$  that entails  $\neg p$  also entails  $p$  and hence is not consistent. So  $\neg p$  is unsatisfiable in  $\mathcal{T}$ . Conversely, suppose that  $p$  is negatively valid in  $\mathcal{T} = \langle \mathbf{T}, * \rangle$ . Let  $T$  be any theory in  $\mathbf{T}$ , and suppose for reductio that  $T$  does not entail  $p$ . Then by Consistency,  $T \cup \{\neg p\}$  is consistent. Since  $\mathbf{T}$  is closed under expansions,  $T \cup \{\neg p\} \in \mathbf{T}$ , and clearly  $T \cup \{\neg p\}$  entails  $\neg p$ . So  $\neg p$  is satisfiable in  $\mathcal{T}$  and hence  $p$  is not negatively valid, contrary to hypothesis. This shows that a sentence  $p$  is negatively valid in a belief revision system  $\mathcal{T}$  iff  $p$  is valid in the sense of Definition 18.

## 11 Conclusion

What is minimal belief change? This paper applied fundamental concepts from decision theory to give a principled answer to this question. The main idea is to treat both adding and retracting beliefs as a kind of “cost” to be avoided in minimal belief revision. The Pareto principle says that we ought to eliminate options that do worse in some respects without doing better in others. Applying this norm to belief change, I defined a belief revision to be *Pareto-minimal* if there is no alternative revision that adds fewer beliefs without retracting more, or that retracts fewer beliefs without adding more. This definition raises a precise mathematical question: What are necessary and sufficient conditions for a theory revision to be Pareto-minimal? The answer consists of three theory revision postulates that characterize Pareto-minimal theory changes incorporating new information. First, the revision must entail the new information. Second,

<sup>17</sup>However, there are proposals to reconcile preservation and the Ramsey test by abandoning the idea that conditional sentences are parts of an agent’s beliefs in the same way that nonconditional sentences are [Arló Costa 1995], [Arló Costa 1998], [Levi 1988].

the revision must follow from the conjunction of the new information and the previous theory. Third, if the previous theory already entails the new information, no change occurs, since the smallest possible change from the previous theory is the previous theory itself.

Thus the main difference between Pareto-minimal theory change and the customary AGM postulates arises when the new information neither contradicts nor is entailed by the current theory. In this case, the AGM postulate K\*4—the *preservation principle*—stipulates that the revised theory should entail the conjunction of the new information and the agent’s current theory. Thus the AGM axioms select the logically strongest Pareto-minimal theory revision, the only one that retracts no beliefs. Pareto-minimality by contrast encompasses the continuum of logical strength between the AGM update as the strongest Pareto-minimal revision and the bare consequences of the new information as the weakest Pareto-minimal revision. From this point of view, the preservation principle appears not as an aspect of minimal theory change, but (arguably) as a general principle of epistemic rationality (perhaps related to “informational economy”).

We found another interpretation of the preservation principle by considering the consequences of the Pareto principle for revising *belief bases*—sets of sentences that represent an agent’s “basic beliefs” and that need not be closed under logical consequence. If we apply the Pareto principle to changes in the agent’s basic beliefs (discounting changes in the consequences of his basic beliefs), then it can be shown that the preservation principle is part of Pareto-minimal changes in basic beliefs. The well-known *Levi identity* provides a full characterization of Pareto-minimal base revisions: Pareto-minimal base revisions are exactly those that result from first retracting just enough basic beliefs to make the current basic beliefs consistent with the new information, and then adding the new information. Since the AGM axioms countenance belief revisions that retract more beliefs than are necessary to make the new information consistent with the current beliefs, it follows that AGM revisions may give up more basic beliefs than Pareto-minimal revisions do.

The full set of AGM axioms contains some that are not necessarily part of Pareto-minimal belief revisions (K\*6–K\*8). I showed how to generalize the definition of Pareto-minimal belief revisions to accommodate any extra constraints on belief change that we may wish to impose. This generalized definition suggests further applications of the Pareto principle by considering constraints other than those explored in this paper (for example, constraints motivated by principles for suppositional reasoning such as those from [Levi 1996]).

The so-called Ramsey test connects principles for belief revision with axioms for reasoning about conditionals (“if  $p$ , then  $q$ ” statements). We saw that the principles of Pareto-minimal theory change correspond exactly to three well-known axioms for counterfactuals (which are part of Lewis’ and Stalnaker’s systems): A belief revision system that satisfies the Ramsey test validates these axioms if and only if the system has a Pareto-minimal belief revision operator. Thus there is a tight connection between Pareto-minimal theory change and prominent axioms for conditionals. As part of a theory of acceptance condi-



tions for conditionals, Pareto-minimal theory change is immune to the triviality results that pose a problem for applying the AGM account of minimal theory change to the logic of conditionals.

The connections between belief revision, conditionals and nonmonotonic reasoning form one of the most active and intriguing areas of research in philosophical logic. The results in this paper show that Pareto-minimality provides a fruitful and principled decision-theoretic foundation for postulates guiding minimal belief revision.

## Acknowledgments

I owe many valuable insights into belief revision and conditionals to Horacio Arló Costa. I thank two anonymous referees and the participants of the LOFT3 meeting for useful comments and suggestions.

## 12 Proofs

**Theorem 5** *Let  $T$  be a theory and let  $p$  be a formula. A theory revision  $T * p$  is a Pareto-minimal change from  $T$  that incorporates  $p \iff$*

1.  $T * p \vdash p$ , and
2.  $T \cup \{p\} \vdash T * p$ , and
3. if  $T \vdash p$ , then  $T * p = T$ .

**Proof.** ( $\Rightarrow$ ) Part 1: Immediate from Definition 4.

Part 2: I show the contrapositive. Suppose that  $T \cup \{p\} \not\vdash T * p$ . Then there is a formula  $q$  in  $T * p$  such that  $T \cup \{p\} \not\vdash q$ . So (a)  $T \not\vdash q$  by Monotonicity. Now consider  $T' = (T * p) \cap Cn(T \cup \{p\} \cup \{\neg q\})$ . First I note that  $T'$  is closed under deductive consequence. For let  $r \in Cn((T * p) \cap Cn(T \cup \{p\} \cup \{\neg q\}))$ . Then by Monotonicity,  $r \in Cn(T * p)$  and  $r \in Cn(Cn(T \cup \{p\} \cup \{\neg q\}))$ . We assumed that  $T * p$  is closed under consequence, and Iteration implies that  $Cn(Cn(T \cup \{p\} \cup \{\neg q\})) = Cn(T \cup \{p\} \cup \{\neg q\})$ ; thus  $r \in T * p \cap Cn(T \cup \{p\} \cup \{\neg q\})$ . This shows that  $Cn(T') = T'$ .

Next, note that (b)  $T' \not\vdash q$  because  $Cn(T \cup \{p\} \cup \{\neg q\}) \not\vdash q$  by Consistency (applied to  $T \cup \{p\}$ ) and Iteration; thus from Monotonicity and the fact that  $T' \subseteq Cn(T \cup \{p\} \cup \{\neg q\})$ , it follows that  $T' \not\vdash q$ . Moreover, we have from Monotonicity and the fact that  $T' \subseteq T * p$  as well that (c) if  $T'$  adds a formula to  $T$ , so does  $T * p$ . From (a), (b) and (c) it follows that (d)  $T * p$  adds more formulas to  $T$  than  $T'$ .

Now I show that (e)  $T'$  retracts from  $T$  exactly the formulas that  $T * p$  retracts from  $T$ . Monotonicity implies immediately that if  $T * p$  retracts a formula from  $T$ , so does  $T'$ . For the converse, suppose that  $T'$  retracts a formula  $r$  from  $T$ . Since  $Cn(T \cup \{p\} \cup \{\neg q\}) \vdash T$ , this implies that  $r \notin (T * p)$ . And that means that  $T * p$  retracts  $r$  from  $T$  as well.

Finally, we have that (f)  $T' \vdash p$ , since  $T * p \vdash p$  by Part 1 and clearly  $Cn(T \cup \{p\} \cup \{\neg q\}) \vdash p$ .

Together, (a)–(f) establish that  $T'$  incorporates  $p$  and  $T * p$  is a greater change from  $T$  than  $T'$  is. Hence  $T * p$  is not a Pareto-minimal change.

Part 3: Immediate, since every theory other than  $T$  retracts or adds more formulas to  $T$  than  $T$  itself does.

( $\Leftarrow$ ) Suppose that  $T * p$  satisfies conditions 1, 2 and 3. Then the claim is immediate if  $T \vdash p$  and  $T * p = T$ ; suppose that  $T \not\vdash p$ . I show that  $T * p$  is not a greater change from  $T$  than any other change  $T'$  that incorporates  $p$ .

First, suppose that  $T * p$  retracts a formula  $q$  from  $T$  but  $T'$  does not, such that  $T' \vdash q$ . Then  $T' \vdash (p \wedge q)$  by Conjunction, whereas  $T * p \not\vdash (p \wedge q)$  by Conjunction as well. Since we supposed that  $T \not\vdash p$ , it follows that  $T \not\vdash (p \wedge q)$  by Conjunction once more. So  $T'$  adds a formula to  $T$ —namely  $p \wedge q$ —that  $T * p$  does not add to  $T$ , and hence  $T * p$  is not a greater change from  $T$  than  $T'$  is. (This is the situation of the cognitive scientist discussed in Section 3.)

Second, suppose that  $T * p$  adds a formula  $q$  to  $T$ , but  $T' \not\vdash q$ . Condition 2 asserts that  $T \cup \{p\} \vdash T * p$  and hence  $Cn(T \cup \{p\}) \vdash q$ . By Deduction, we have that (a)  $T \vdash p \rightarrow q$ . Moreover, Implication implies that (b)  $T * p \vdash p \rightarrow q$ . Also, (c)  $T' \not\vdash p \rightarrow q$ . For suppose that on the contrary,  $T' \vdash p \rightarrow q$ . Then since  $T' \vdash p$ , it follows from Modus Ponens that  $T' \vdash q$ , contrary to assumption. From (a), (b) and (c) we have that  $T'$  retracts a formula from  $T$ —namely  $p \rightarrow q$ —that  $T * p$  does not retract from  $T$ . Thus  $T * p$  is not a greater change from  $T$  than  $T'$  is.

These arguments establish that if  $T * p$  satisfies conditions 2 and 3, then there is no theory  $T'$  incorporating  $p$  such that  $T * p$  is a greater change from  $T$  than  $T'$  is. From Condition 1 it follows that  $T * p$  is a Pareto-minimal change from  $T$  that incorporates  $p$ . ■

**Proposition 11** *Let  $B$  be a base and let  $p$  be a formula. Then  $B * p$  is a Pareto-minimal consistent change from  $B$  that incorporates  $p \iff B * p$  is a retraction-minimal consistent change from  $B$  that incorporates  $p$  and minimizes additions.*

**Proof.** ( $\Leftarrow$ ) Immediate since all retraction-minimal consistent changes that minimize additions are Pareto-minimal consistent changes. ( $\Rightarrow$ ) Suppose that  $B * p$  is a Pareto-minimal consistent change from  $B$ . Suppose for reductio that  $B * p$  adds a sentence  $q \neq p$  to  $B$ ; then  $B * p - \{q\}$  adds less to  $B$  and does not retract anything from  $B$  that  $B * p$  does not retract. So  $B * p$  is not Pareto-minimal, contrary to supposition. This contradiction shows that  $B * p$  does not add anything to  $B$  except possibly  $p$ . Hence  $B * p$  minimizes additions. For retraction-minimality, suppose for reductio that there is a consistent base  $B_2$  containing  $p$  such that  $B * p$  retracts more from  $B$  than  $B_2$  does. Let  $q \in (B_2 \cap B) - B * p$  be a sentence that  $B * p$  retracts from  $B$  but  $B_2$  does not. Consider  $B' = (B_2 \cap B) \cup \{p\}$ . Since  $B_2$  is consistent and contains  $p$ ,  $B'$  is consistent (by Monotonicity) and contains  $p$ . Clearly  $B'$  does not add anything to  $B$  other than  $p$ . Moreover, for all sentences  $r$ , if  $B'$  retracts  $r$  from  $B$ , so

does  $B_2$ . So if  $B'$  retracts  $r$  from  $B$ , so does  $B * p$ . Finally,  $B'$  does not retract  $q$  from  $B$  but  $B * p$  does. So  $B * p$  is a greater change from  $B$  than  $B'$  contrary to the supposition that  $B * p$  is a Pareto-minimal consistent change from  $B$  that incorporates  $p$ . This contradiction shows that  $B * p$  is a retraction-minimal consistent change from  $B$  that incorporates  $p$  and minimizes additions. ■

**Lemma 13** *Let  $B, B_1$  be two bases such that  $B_1 \subseteq B$ , and let  $p$  be a formula such that  $B_1 \not\vdash p$ . Then  $B_1$  is a **retraction-minimal contraction** from  $B$  on  $p \iff$  for all formulas  $q$ , if  $B_1$  retracts  $q$  from  $B$ , it is the case that  $B_1 \cup \{q\} \vdash p$ .*

**Proof.** ( $\Rightarrow$ ) Otherwise  $B_1 \cup \{q\}$  doesn't entail  $p$  and retracts less from  $B$  than  $B_1$ . ( $\Leftarrow$ ) Assume the right-hand side, and suppose for reductio that there is a contraction  $B_2 \subseteq B$  that retracts less from  $B$  than  $B_1$  does, and that  $B_2 \not\vdash p$ . Let  $q \in B_2 - B_1$ . Then by hypothesis,  $B_1 \cup \{q\} \vdash p$ . Hence by Monotonicity,  $B_2 \not\supseteq B_1$  since  $q \in B_2$ . Let  $r$  be any sentence in  $B_1 - B_2$ . Since  $B_1$  is a subset of  $B$ ,  $r$  is in  $B$  and hence  $B_1$  does not retract more than  $B_2$  does. ■

**Theorem 14** *Let  $B$  be a base and let  $p$  be a formula. Suppose that a revision  $B * p$  contains  $p$ . Then  $B * p$  is a Pareto-minimal consistent change from  $B$  that incorporates  $p \iff$  there is a retraction-minimal contraction  $B'$  from  $B$  on  $\neg p$  such that  $B * p = B' \cup \{p\}$ .*

**Proof.** ( $\Rightarrow$ ) Since by Proposition 11, Pareto-minimal changes minimize additions, we have that  $B * p \subseteq B \cup \{p\}$ .

Case 1:  $p \in B$ . Then we have that  $B * p \subseteq B \cup \{p\} = B$ . Now suppose for reductio that  $B * p$  is not a retraction-minimal contraction on  $\neg p$  from  $B$ . Then by Lemma 13 there is a sentence  $q \in B - B * p$  such that  $B * p \cup \{q\} \not\vdash \neg p$ . So by Consistency,  $B * p \cup \{q\}$  is consistent, contains  $p$ , retracts less from  $B$  than  $B * p$ , and does not add anything to  $B$ . This contradicts the assumption that  $B * p$  is a Pareto-minimal consistent change from  $B$  on  $p$ , and so  $B * p$  is a retraction-minimal contraction on  $\neg p$  from  $B$ . Since  $B * p$  contains  $p$ ,  $B * p \cup \{p\} = B * p$  and so  $B * p = B'$  witnesses the claim of the theorem.

Case 2:  $B$  does not contain  $p$ . Let  $B' = B * p - \{p\}$ . Since  $B * p \subseteq B \cup \{p\}$ , we have that  $B * p - \{p\} \subseteq B$ . Suppose for reductio that  $B'$  is not a retraction-minimal contraction on  $\neg p$  from  $B$ . Then by Lemma 13 there is a sentence  $q \in B$  but not in  $B'$  such that  $B' \cup \{q\} \not\vdash \neg p$ . So by Consistency,  $B' \cup \{q, p\}$  is consistent, contains  $p$ , retracts less from  $B$  than  $B * p$ , and does not add any more. This contradicts the assumption that  $B * p$  is a Pareto-minimal consistent change from  $B$  on  $p$ , and so  $B'$  is a retraction-minimal consistent contraction on  $\neg p$  from  $B$ . Since  $B' \cup \{p\} = (B * p - \{p\}) \cup \{p\} = B * p$ , it follows that  $B' = B * p - \{p\}$  witnesses the claim of the theorem.

So in either case there is a retraction-minimal contraction  $B'$  on  $\neg p$  from  $B$  such that  $B * p = B' \cup \{p\}$ .

( $\Leftarrow$ ) Let  $B'$  be a retraction-minimal contraction from  $B$  on  $\neg p$  such that  $B * p = B' \cup \{p\}$ . Let  $B_1$  be any other consistent base containing  $p$ . I show that  $B * p$  is not a greater change from  $B$  than  $B_1$  is. Since  $B'$  is a contraction of  $B$ , it follows that (a) the only sentence that  $B * p$  might add to  $B$  is  $p$  (i.e.,

$B * p \subseteq B \cup \{p\}$ ), and  $B_1$  contains  $p$  as well. Hence  $B_1$  does not add less to  $B$  than  $B * p$ . So suppose for reductio that (b)  $B_1$  retracts less from  $B$  than  $B * p$  and that (c) if  $B_1$  adds a sentence  $q$  to  $B$ , then so does  $B * p$ .

Given (a), from (c) follows (d): that  $B_1 \subseteq B \cup \{p\}$ . From (b) we have that (e)  $B_1$  retracts less from  $B$  than  $B'$  since  $B' \subseteq B * p$ . Since  $B_1$  is consistent and contains  $p$ , it follows by Inconsistency that (f)  $B_1 \not\vdash \neg p$ .

Case 1:  $p \in B$ . Then  $B \cup \{p\} = B$ , and so by (d) and (f)  $B_1$  is a contraction of  $B$  on  $\neg p$ . By (e),  $B_1$  retracts less from  $B$  than  $B'$  does, contrary to the assumption that  $B'$  is a retraction-minimal contraction from  $B$  on  $\neg p$ .

Case 2:  $p \notin B$ . Then  $B_1 - \{p\}$  retracts from  $B$  exactly those sentences that  $B_1$  retracts. Hence by (e)  $B_1 - \{p\}$  retracts less from  $B$  than  $B'$  does. Furthermore, it follows from (f) (given Monotonicity) and (d) that  $B_1 - \{p\}$  is a contraction of  $B$  on  $\neg p$ . This contradicts the assumption that  $B'$  is a retraction-minimal consistent contraction from  $B$  on  $\neg p$ .

Thus in either case assuming that  $B * p$  is a greater change from  $B$  than  $B_1$  leads to a contradiction. Since  $B_1$  was chosen as an arbitrary consistent base containing  $p$ , this establishes that  $B * p$  is a Pareto-minimal consistent base revision from  $B$ . ■

**Theorem 19** *Let  $\mathcal{T} = \langle \mathbf{T}, * \rangle$  be a belief revision system satisfying the Ramsey test. Then  $*$  is a Pareto-minimal belief revision operator  $\iff \mathcal{T}$  validates*

1.  $p > p$ , and
2.  $(p > q) \rightarrow (p \rightarrow q)$ , and
3.  $(p \wedge q) \rightarrow (p > q)$

for all formulas  $p, q, r$ .

**Proof.** ( $\implies$ ) Let  $T \in \mathbf{T}$  be an arbitrary theory in the belief revision system, and let  $p, q, r \in L_{>}$  be arbitrary formulas.

Axiom 1: Since  $*$  is a Pareto-minimal revision operator, we have that  $T * p \vdash p$ . Since  $\mathcal{T}$  satisfies the Ramsey test, this implies that  $T \vdash p > p$ . Hence  $\mathcal{T}$  validates  $p > p$  for all formulas  $p \in L$ .

Axiom 2: Let  $T' = Cn(T \cup \{p > q\})$ . By the Ramsey Test,  $T' * p \vdash q$ . Since  $*$  is a Pareto-minimal change revision operator, we have that  $T' \cup \{p\} \vdash T' * p$ . Hence  $T' \cup \{p\} \vdash q$ . Thus  $T' \vdash p \rightarrow q$  by Deduction. Again by Deduction, it follows that  $T \vdash (p > q) \rightarrow (p \rightarrow q)$ . Hence  $\mathcal{T}$  validates  $(p > q) \rightarrow (p \rightarrow q)$  for all formulas  $p, q \in L$ .

Axiom 3: Let  $T' = Cn(T \cup \{p \wedge q\})$ . Then  $T' * p = T'$  since by Conjunction  $T'$  entails  $p$  and  $*$  is a Pareto-minimal change revision operator. Since again by Conjunction  $T'$  entails  $q$ , it follows from the Ramsey Test that  $T' \vdash p > q$ . By Deduction, we have that  $T \vdash (p \wedge q) \rightarrow (p > q)$ . Thus  $\mathcal{T}$  validates  $(p \wedge q) \rightarrow (p > q)$  for all formulas  $p, q, r$ .

( $\impliedby$ ) Suppose that  $\mathcal{T}$  validates Axioms 1–3 for all formulas  $p, q, r$ . I show that  $*$  is a Pareto-minimal belief revision operator. Let  $T \in \mathbf{T}$  be an arbitrary theory.

Step 1: Since  $T \vdash p > p$ , the Ramsey Test implies that  $T * p \vdash p$ , for any formula  $p$ .

Step 2: Let  $q \in T * p$  be any formula in the revision of  $T$  by  $p$ ; so by Ramsey Test,  $T \vdash p > q$ . I show that  $T \cup \{p\} \vdash q$ . By Axiom 2,  $T \vdash (p > q) \rightarrow (p \rightarrow q)$ . By Modus Ponens, this implies that  $T \vdash p \rightarrow q$ . By Deduction, it follows that  $T \cup \{p\} \vdash q$ , as required. Since this holds for any formula  $q$ , we have for all revisions  $T * p$  that  $T \cup \{p\} \vdash T * p$ , for all formulas  $p$ .

Step 3: Suppose that  $T \vdash p$  for some formula  $p$ .

I show that  $T * p \vdash T$ . Let  $q \in T$ . Then  $T \vdash (p \wedge q)$  by Conjunction, and  $T \vdash (p \wedge q) \rightarrow (p > q)$  by Axiom 3. By Modus Ponens,  $T \vdash p > q$ . By Ramsey test,  $T * p \vdash q$ . So  $T * p \vdash T$ . From Step 2 we have that  $T = T \cup \{p\} \vdash T * p$ . All together, this implies that  $T * p = T$  whenever  $T \vdash p$ .

Steps 1,2 and 3 establish that  $*$  is a Pareto-minimal belief revision operator.

■

## References

- [Alchourrón and Makinson 1982] Alchourrón, C.E. and Makinson, D.: 1982, ‘The logic of theory change: Contraction functions and their associated revision functions’, *Theoria* **48**,14–37.
- [Arló Costa 1995] Arló Costa, H.: 1995, ‘Epistemic Conditionals, Snakes and Stars’, in *Conditionals, from Philosophy to Computer Science*, vol. 5 of Studies in Logic and Computation, eds. Fariñas del Cerro, G. Crocco, A. Herzig. Oxford University Press, Oxford.
- [Arló Costa 1998] Arló Costa, H.: 1998, ‘Belief revision conditionals: *basic* iterated systems’, forthcoming in *Annals of Pure and Applied Logic*.
- [Arló Costa 1990] Arló Costa, H.: 1990, ‘Conditionals and monotonic belief revisions: the success postulate’, *Studia Logica* **49**,557-566.
- [Gärdenfors 1988] Gärdenfors, P.: 1988, *Knowledge In Flux: modeling the dynamics of epistemic states*. MIT Press, Cambridge,Mass.
- [Hansson 1998] Hansson, S.O.: 1998, ‘Editorial: Belief Revision Theory Today’, *Journal of Logic, Language and Information* Vol.7, No.2,123–126.
- [Harman 1986] Harman, G.: 1986, *Change in View: Principles of Reasoning*. MIT Press, Cambridge, Mass.

- [Katsuno and Mendelzon 1990] Katsuno, H. and Mendelzon, A.O. 1990: *On the difference between updating a knowledge base and revising it*, Technical Report on Knowledge Representation and Reasoning, KRR-TR-90-6, University of Toronto, Department of Computer Science.
- [Katsuno and Mendelzon 1991] Katsuno, H. and Mendelzon, A.O. 1991: ‘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*, pp.387–394, Morgan Kaufmann Publishers, Los Altos, CA.
- [Kelly *et al.* 1995] Kelly, K., Schulte O. and Hendricks, V.: 1995, ‘Reliable Belief Revision’, *Proceedings of the IX International Joint Congress for Logic, Methodology and the Philosophy of Science*, Kluwer, Dordrecht.
- [Levi 1988] Levi, I.: 1988, ‘Iteration of conditionals and the Ramsey test’, *Synthese* **76**, 49–81.
- [Levi 1996] Levi, I.: 1996, *For the sake of the argument: Ramsey test conditionals, Inductive Inference, and Nonmonotonic Reasoning*, Cambridge University Press, Cambridge.
- [Lewis 1973] Lewis, D.: 1973, *Counterfactuals*, Blackwell, Oxford.
- [Lewis 1976] Lewis, D.: 1976, ‘Probabilities of Conditionals and Conditional Probabilities’, *The Philosophical Review* LXXXV, **3**, 297–315.
- [Martin and Osherson 1998] Martin, E. and Osherson, D.: 1998, *Elements of Scientific Discovery*, MIT Press, Cambridge, Mass.
- [Nayak 1994] Nayak, A.: 1994, ‘Iterated Belief Change Based on Epistemic Entrenchment’, *Erkenntnis* **41**, 353-390.
- [Osherson *et al.* 1986] Osherson, D., Stob, M. and Weinstein, S. 1986: *Systems That Learn*, MIT Press, Cambridge, Mass.

- [Putnam 1963] Putnam, H.: 1963, ‘Degree of Confirmation’ and Inductive Logic’, in *The Philosophy of Rudolf Carnap*, ed. A. Schilpp, Open Court, La Salle, Ill.
- [Rott 1998a] Rott, H.: 1998, ‘Logic and Choice’, in Gilboa, I. ed., *Proceedings of the Seventh Conference on Theoretical Aspects of Rationality and Knowledge*, pp.235–248. Morgan Kaufmann, San Francisco.
- [Rott 1998b] Rott, H. 1998: ‘Two Dogmas of Belief Revision’, *Proceedings of the Third Conference on Logic and the Foundations of Game and Decision Theory*. ICER, Turin, Italy.
- [Quine 1951] Quine, W.: 1951, ‘Two Dogmas of Empiricism’, *Philosophical Review* **60**, 20–43.
- [Stalnaker 1968] Stalnaker, R.: 1968, ‘A Theory of Conditionals’, in *Studies in Logical Theory*: 98–112 American Philosophical Quarterly Monograph Series, no. 2, N. Resher, ed., Blackwell, Oxford.