BOUNDED REVISION: TWO-DIMENSIONAL BELIEF CHANGE BETWEEN CONSERVATISM AND MODERATION

Hans Rott

Department of Philosophy, University of Regensburg
hans.rott@psk.uni-regensburg.de

ABSTRACT: In this paper I present the model of 'bounded revision' that is based on two-dimensional revision functions taking as arguments pairs consisting of an input sentence and a reference sentence. The key idea is similar to the model of 'revision by comparison' investigated by Fermé and Rott (*Artificial Intelligence* 157, 2004). In contrast to the latter, however, bounded revision satisfies the AGM axioms as well as the Darwiche-Pearl axioms. Two one-dimensional special cases are obtained by setting one argument of the two-dimensional revision operation to certain extremal values. Bounded revision thus fills the space between conservative revision (also known as natural revision) and moderate revision (also known as lexicographic revision). I argue that two-dimensional revision operations add decisively to the expressive power of qualitative approaches that refrain from assuming numbers as measures of degrees of belief.

1. Introduction

In his joint work with Sten Lindström in the late 1980s and early 1990s, Wlodek Rabinowicz pioneered a new, relational approach to belief change. As they say in the introduction to their beautiful paper "Epistemic Entrenchment with Incomparabilities and Relational Belief Revision":

> Our proposal was to view belief revision as a relation rather than

Printed from: *Hommage à Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz.* Ed. T. Rønnow-Rasmussen, B. Petersson, J. Josefsson & D. Egonsson, 2007. www.fil.lu.se/hommageawlodek

> as a function on theories (or belief sets). The idea was to allow for there being several equally reasonable revisions of a theory with a given proposition. (Lindström and Rabinowicz 1991, p. 93)

This insightful idea turned out to be very fruitful and was further investigated, besides by Lindström and Rabinowicz themselves, by Krister Segerberg, John Cantwell, Alexander Bochman and others. Though the details differ significantly, the present paper may be seen as another way of putting to work the Lindström-Rabinowicz idea of allowing for several equally reasonable ways of accepting a piece of new information.[1]

Representations of belief states in terms of probability functions or ranking functions are very rich and powerful. However, it is often hard to come by the relevant numbers. Qualitative belief change in the style of AGM or its extensions to iterated belief change in the 1990s, on the other hand, does not need numbers, but is rather more restricted in its expressive powers. Fermé and Rott (2003) suggested a basically qualitative approach that is more flexible than AGM style models in that it allows a new piece of information to be accepted in various degrees or strengths. The key idea is that the input α does not come 'naked'. It does not come with a number either, but with a reference sentence δ expressing an antecedently held belief, and the agent is supposed to follow an instruction of the form:

'Accept α with a strength that at least equals that of δ.'

Revision by comparison so conceived presents a model that lies between the traditional qualitative and quantitative approaches.[2] In the context of revision by comparison, 'revising by α' is understood to mean revising with respect to some existing belief expressed by the reference sentence δ.

A drawback of the revision-by-comparison approach of Fermé and Rott is that it does not satisfy the Darwiche-Pearl postulates for iterated belief change. These postulates have a very appealing possible worlds semantics. Assuming that a belief state can be represented by a total pre-ordering of possible worlds

[1] I have admired Wlodek's width of interests, originality and sharpness since I first met him at the memorable Konstanz conference in 1989. Wlodek has become a dear friend since then, and it is a great pleasure for me indeed to dedicate this paper to him. Happy birthday, Wlodek!

[2] An earlier approach similar to revision by comparison is elaborated in Cantwell's (1997) 'raising' operation. Cantwell also has an interesting dual operation that he calls 'lowering'.

(graphically representable as a Lewis-Grove style system of spheres[3]) and the new piece of information is α, the Darwiche-Pearl conditions express that

- the ordering restricted to the α-worlds should be left untouched, and the same holds for the ordering restricted to the $\neg\alpha$-worlds

- no $\neg\alpha$-world should improve its position relative to some α-world

These very plausible semantic constraint conditions give rise to two conditions each, making up the set known as Darwiche-Pearl postulates.

The present paper introduces a model of belief change similar to revision-by-comparison that satisfies the Darwiche-Pearl postulates. I call this model *bounded revision*, because intuitively, δ serves as a bound for the acceptance of α. The reference sentence δ functions here as a measure of how firmly entrenched α should be in the agent's belief state after the change. Put differently, δ may be seen as a parameter picking out one of 'several equally reasonable revisions' (Lindström and Rabinowicz) of a belief state with a given proposition. The specification of a reference sentence thus makes explicit what is left external to the Lindström-Rabiniwicz theory.

Another welcome feature of bounded revision is that it offers a whole range of potential revisions that are, in a precise sense, *between* conservative revision and moderate revision.[4] The latter operations can be obtained by setting the parameter sentence to $\bot$ and $\top$, that is to FALSITY and TRUTH respectively. Both conservative and moderate revision, however, are defective in some way. Conservative revision accepts new evidence, but always accords the lowest possible entrenchment to it, so that it is immediately lost if a contradiction with another piece of evidence arises (see Rott 2003). Moderate revision in a way suffers from the opposite defect by accepting the new information too firmly. Every world in which the new information is true is considered more plausible

[3] Lewis (1973) and Grove (1988). More on this in Section 4.

[4] These are my names (Rott 2003). The operations I denote by these names are more well-known as *natural revision* in the sense of Boutilier (1993) and *lexicographic revision* as studied by Nayak (1994) and many others. It is odd that 'moderate' revision features as a limiting case below. From the perspective of the present paper, Booth and Meyer (2006, p. 149) are right in saying that "lexicographic [= moderate] revision is a formalisation of the 'most recent is best' approach to revision taken to its logical extreme." The still more extreme strategy called *radical revision* in Rott (2003) is not put on stage in the present paper because it fails to conform to the Darwiche-Pearl postulates.

than any world in which the new information is false. While conservative revision is too conservative, moderate revision is too radical. Bounded revision offers a wealth of options just in between conservative and moderate revision.

The idea of bounded revision can be expressed by the following command

> 'Accept α as long as δ holds along with α, and just a little more.'

In a way, the acceptance of α is *bounded* by δ. We can think of the reference sentence δ in two ways. First, we may suppose that it is a marker delineating the shape of a sphere of a Lewis-Grove system of spheres that characterizes the reasoner's belief state. Such sentences are independent of the new information. Or, second, we can just use δ as a sentence that is supposed to hold in a range of relatively plausible situations in which α holds. It would be nice if we could use the first option. Technically this is possible in a finitistic framework, but psychologically it is not very realistic that one can in general characterized a 'degree of belief' with respect to a given belief state by a single reasonably comprehensible sentence. So I will pursue the last option, in which δ is context-dependent (depending on the input sentence). Notice that the second option is more general in that it includes the first as a special case.[5]

So any sentence δ may serve as the parameter sentence for bounded revision, δ need not actually be believed to be true by the agent. However, the paradigm cases are those in which δ is cotenable with α to some extent, in the sense that a stretch of the more plausible ways of making α true are all ways that make δ true as well.[6] The greater the stretch where δ holds along with α, the firmer α gets accepted by a revision that is bounded by δ.

Often the intended cases of belief revision are those in which the input sentence α is not believed prior to the revision. However, bounded revision may well be used to increase the strength or entrenchment of a sentence believed true to begin with.[7]

[5] Compare the discussion of two ways of interpreting Goodman's 'cotenability' in Lewis (1973, pp. 69–70).

[6] Intuitively, not only the *most* plausible α-worlds, but also all those that are *sufficiently* plausible are moved center stage in a bounded revision by α, and it is precisely the task of the reference sentence δ to characterize what is meant 'sufficient.' In terms of doxastic entrenchment, one could characterize the intended case not only by $\neg\alpha < \alpha \rightarrow \delta$, but perhaps more graphically, by $\neg\alpha \ll \alpha \rightarrow \delta$ or by $\neg\alpha <_n \alpha \rightarrow \delta$, if the δ-worlds cover the n most plausible strata of α-worlds. For the idea of entrenchment, cf. section 4.

[7] This is the main idea of Cantwell's (1997) 'raising.'

2. Generalizing AGM to two-dimensional revisions of belief states

What is a belief state? For the purposes of this paper, belief states may be entities of any type whatsoever, neural states, holistic mental states, abstract machine states etc. We may even suppose that their full nature is inscrutable for us. At the same time, we will assume that the set of beliefs of an agent in a certain belief state is epistemically accessible. The beliefs are, so to speak, the visible tip of the iceberg that itself remains concealed from our eyes. Our assumption will be that belief states contain a rich structure that determines the development of the agent's belief sets in response to a sequence of inputs (it is not excluded that it determines more). Belief states may contain a lot more structure, but this is what we are interested in. Later, in Section 4, we shall introduce formal structures as representations of belief states that contain a lot more structure than a plain belief set, but are still abstractions from 'real' belief states.

A *one-dimensional* belief revision operation is a function $*$ that takes a belief state $\mathcal{B}$ and an input sentence α and returns the new belief state $*(\mathcal{B}, \alpha)$ which is to denote the state $\mathcal{B}$ revised by α. A *two-dimensional* revision function is similar, except that the input is a pair of sentences $\langle \alpha, \delta \rangle$. The first sentence is the *input sentence*, the second sentence the *reference sentence*. Usually, I will use the variables α, β and γ etc. for input sentences, and the variables δ, ε, ζ etc. for reference sentences.[8]

We work with a finite propositional language. The set of possible worlds (interpretations, models) and the set of sets of logically equivalent sentences are then finite, too.[9] We use Cn to indicate a consequence operation governing the language. We suppose throughout this paper that the logic is Tarskian, that it includes classical propositional logic, and that it satisfies the deduction theorem.[10]

[8] The terminology of input and reference sentences is taken over from Fermé and Rott.

[9] We presuppose finiteness mainly as a matter of convenience, in order not to burden this paper with technical details distracting us from the main issues. An infinite language would not complicate things as long as we work with entrenchment relations, but when working with systems of spheres, infinity complicates the matter enormously. See, e.g., Rott and Pagnucco (1999, Section 8).

[10] By saying that the logic Cn is Tarskian, we mean that it is reflexive ($H \subseteq Cn(H)$), monotonic (if $H \subseteq H'$, then $Cn(H) \subseteq Cn(H')$), idempotent ($Cn(Cn(H)) \subseteq Cn(H)$) and compact (if $\alpha \in Cn(H)$, then $\alpha \in Cn(H')$ for some finite $H' \subseteq H$). The deduction theorem says that $\alpha \to \beta \in Cn(H)$ if and only if $\beta \in Cn(H \cup \{\alpha\})$. We may write $H \vdash \alpha$ for $\alpha \in Cn(H)$.

Notation: If Γ is a set of sentences and α and β are sentences, I shall write $\Gamma + \alpha$ for $\Gamma \cup \{\alpha\}$ and $\alpha + \beta$ for $\{\alpha, \beta\}$. For any belief state $\mathcal{B}$, $\ulcorner\mathcal{B}\urcorner$ is the set of beliefs held by a person in belief state $\mathcal{B}$ (more exactly: the beliefs that can be ascribed to the person, or the beliefs that the person is committed to). We assume that $\ulcorner\mathcal{B}\urcorner$ is logically closed. If $\mathcal{B}$ and $\mathcal{B}'$ are two belief states, then $\mathcal{B} \simeq \mathcal{B}'$ is short for $\ulcorner\mathcal{B}\urcorner = \ulcorner\mathcal{B}'\urcorner$.

For a one-dimensional revision function $*$, we write $\mathcal{B} * \alpha$ for $*(\mathcal{B}, \alpha)$. For a two-dimensional revision function $*$, we write $\mathcal{B} *_\delta \alpha$ for $*(\mathcal{B}, \langle\alpha, \delta\rangle)$. In contexts where it is clear that we deal with a two-dimensional function $*$ and where the reference sentence δ is irrelevant (i.e., may be arbitrarily chosen), we may simply write $\mathcal{B} * \alpha$ for $*(\mathcal{B}, \langle\alpha, \delta\rangle)$.

We base our considerations on the famous AGM postulates for one-step belief revision. We transfer them to the new notation that makes explicit that belief revision is about the revision of belief states.

(AGM1) $\ulcorner\mathcal{B} * \alpha\urcorner$ is logically closed.

(AGM2) $\ulcorner\mathcal{B} * \alpha\urcorner$ implies α

(AGM3) $\ulcorner\mathcal{B} * \alpha\urcorner$ is a subset of $\mathit{Cn}(\ulcorner\mathcal{B}\urcorner + \alpha)$

(AGM4) If α is consistent with $\ulcorner\mathcal{B}\urcorner$, then $\ulcorner\mathcal{B}\urcorner$ is a subset of $\ulcorner\mathcal{B} * \alpha\urcorner$

(AGM5) If α is consistent, then $\ulcorner\mathcal{B} * \alpha\urcorner$ is consistent

(AGM6) If α is equivalent with β, then $\ulcorner\mathcal{B} * \alpha\urcorner = \ulcorner\mathcal{B} * \beta\urcorner$

(AGM7) $\ulcorner\mathcal{B} * (\alpha \wedge \beta)\urcorner$ is a subset of $\mathit{Cn}(\ulcorner\mathcal{B} * \alpha\urcorner + \beta)$

(AGM8) If β is consistent with $\ulcorner\mathcal{B} * \alpha\urcorner$, then $\ulcorner\mathcal{B} * \alpha\urcorner$ is a subset of $\ulcorner\mathcal{B} * (\alpha \wedge \beta)\urcorner$

The AGM postulates as they are written down here apply to one-dimensional belief revision functions in the first place. For two-dimensional revision operations, replace '$*$' by '$*_\delta$'. We shall always assume that the revision functions we consider, whether one-dimensional or two-dimensional, satisfy the AGM postulates, *except* for (AGM5) which introduces an unnecessary loss of generality.[11]

[11] See Section 5. I recommend to use the following two conditions instead of (AGM5):
($\emptyset * 1$) If $\ulcorner\mathcal{B} * \alpha\urcorner$ is inconsistent, so is $\ulcorner\mathcal{B} * (\alpha \wedge \beta)\urcorner$, and
($\emptyset * 2$) If $\ulcorner\mathcal{B} * (\alpha \wedge \beta)\urcorner$ is inconsistent, then β is inconsistent with $\ulcorner\mathcal{B} * \alpha\urcorner$.
See the constraints for 'refusing to choose' and the corresponding conditions for belief revision in Rott (2001, pp. 149–153, 206 and 118).

There may be other sentences than logical falsehoods that are considered 'absolutely impossible' by the agent, and any revision by an impossible sentence leads to a belief state associated with the inconsistent belief set.

It follows from (AGM3) and (AGM4) that if $\ulcorner\mathcal{B}\urcorner$ is consistent, then $\ulcorner\mathcal{B}\urcorner = \ulcorner\mathcal{B} * \top\urcorner$. The beliefs in a consistent belief state $\mathcal{B}$ are exactly the same beliefs as after the revision of $\mathcal{B}$ by the tautology $\top$. And if $*$ is two-dimensional, this should not depend on the reference sentence δ.

3. Bounded revision as an operation for iterated belief change

The general condition for the two-dimensional operation of *bounded revision* is this:

(BoundRevIter)

$$\boxed{\mathcal{B} *_\delta \alpha *_\varepsilon \beta \;\simeq\; \begin{cases} \mathcal{B} *_\zeta (\alpha \wedge \beta) & \text{if } \ulcorner\mathcal{B} *_\delta (\alpha \wedge (\delta \rightarrow \beta))\urcorner + \beta \text{ is consistent} \\ \mathcal{B} *_\zeta \beta & \text{otherwise} \end{cases}}$$

Unfortunately, this condition is not very transparent. The rationale for it will become clear when we turn to the modellings in terms of systems of spheres and entrenchments.[12]

Look at what the condition of bounded revision gives for the important special case $\beta = \top$. We obtain, after a little simplification using (AGM6),

$$\mathcal{B} *_\delta \alpha *_\varepsilon \top \;\simeq\; \begin{cases} \mathcal{B} *_\zeta \alpha & \text{if } \ulcorner\mathcal{B} *_\delta \alpha\urcorner \text{ is consistent} \\ \mathcal{B} *_\zeta \top & \text{otherwise} \end{cases}$$

If $\ulcorner\mathcal{B} *_\delta \alpha\urcorner$ is consistent, the left hand side equals $\ulcorner\mathcal{B} *_\delta \alpha\urcorner$, by (AGM3) and (AGM4). Therefore, (BoundRevIter) implies that if $\ulcorner\mathcal{B} *_\delta \alpha\urcorner$ is consistent, it is identical with $\ulcorner\mathcal{B} *_\zeta \alpha\urcorner$ for any reference sentence ζ. By a symmetrical argument applied $\ulcorner\mathcal{B} *_\zeta \alpha\urcorner$, it also follows that if $\ulcorner\mathcal{B} *_\delta \alpha\urcorner$ is inconsistent so is $\ulcorner\mathcal{B} *_\zeta \alpha\urcorner$. We end up with the observation that in bounded revision, the *belief set* obtained

[12] There is an alternative definition that uses the condition $\ulcorner\mathcal{B} *_\delta (\alpha \wedge (\delta \rightarrow \beta))\urcorner \vdash \delta$ for the case distinction instead of the condition used in (BoundRevIter). This corresponds to the alternative options mentioned in Sections 5 and 6.

after a revision by input α does not at all depend on the reference sentence δ. Only the *belief states* obtained differ.[13] We keep for the record

$$\ulcorner\mathcal{B} *_\delta \alpha\urcorner = \ulcorner\mathcal{B} *_\varepsilon \alpha\urcorner \text{ for all } \delta \text{ and } \varepsilon$$

It is interesting to consider two limiting cases of reference sentences δ that are (i) never or (ii) always cotenable with α. Now that we know that the reference sentences are not important when comparing two belief sets obtained by one-step revision, we may suppress indication of the reference sentences for the last revision steps.

For the first limiting case, let δ be $\perp$ or $\neg\alpha$. Then the definition reduces to

$$\mathcal{B} *_\perp \alpha * \beta \;\simeq\; \begin{cases} \mathcal{B} * (\alpha \wedge \beta) & \text{if } \ulcorner\mathcal{B} * \alpha\urcorner + \beta \text{ is consistent} \\ \mathcal{B} * \beta & \text{otherwise} \end{cases}$$

which characterizes *conservative* (or Boutilier's 1993 *natural*) *revision.* The upper line follows already from the AGM postulates (AGM7) and (AGM8) for one-step revisions.

For the second limiting case, let δ be $\top$ or α. Then the clause for the upper line of the definition reduces to $\neg\beta \notin \ulcorner\mathcal{B} * (\alpha \wedge \beta)\urcorner$, which given the success, closure and consistency conditions of AGM, is in turn equivalent to the consistency of $\mathcal{B} * (\alpha \wedge \beta)$. So we get

$$\mathcal{B} *_\top \alpha * \beta \;\simeq\; \begin{cases} \mathcal{B} * (\alpha \wedge \beta) & \text{if } \ulcorner\mathcal{B} * (\alpha \wedge \beta)\urcorner \text{ is consistent} \\ \mathcal{B} * \beta & \text{otherwise} \end{cases}$$

which characterizes *moderate* (or Nayak's 1994 *lexicographic*) *revision.* Had we stipulated that the consistency postulate (AGM5) be satisfied, the upper line could be made conditional on the simple requirement that $\alpha \wedge \beta$ be consistent.

Now we turn to the *Darwiche-Pearl postulates* for iterated belief change.

[13] This is quite different from the situation with Fermé and Rott's revision by comparison. There the relative strengths of input and reference sentence do matter. Revision by comparison is not a pure operation of revision but has also features of contraction. – The identity of belief sets also shows that bounded revision is not a direct instantiation of the Lindström-Rabinowicz idea of relational revision described in the introduction.

(DP1) If β implies α, then $\ulcorner(\mathcal{B} * \alpha) * \beta\urcorner = \ulcorner\mathcal{B} * \beta\urcorner$

(DP2) If β is inconsistent with α, then $\ulcorner(\mathcal{B} * \alpha) * \beta\urcorner = \ulcorner\mathcal{B} * \beta\urcorner$.

(DP3) If α is in $\ulcorner\mathcal{B} * \beta\urcorner$, then α is in $\ulcorner(\mathcal{B} * \alpha) * \beta\urcorner$.

(DP4) If $\neg\alpha$ is not in $\ulcorner\mathcal{B} * \beta\urcorner$, then $\neg\alpha$ is not in $\ulcorner(\mathcal{B} * \alpha) * \beta\urcorner$.

Notice that these postulates make statements only about *belief sets*, but since they concern iterations, they implicitly talk about one-step changes of *belief states* as well. (More about this later.) If the postulates are meant to apply to two-dimensional revision functions, any occurrence of '$*$' should be replaced by a subscripted star, '$*_\delta$', with the reference sentences δ being allowed to vary arbitrarily even within a single postulate.

As Darwiche and Pearl have shown, these postulates correspond one by one to very appealing constraints on total preorderings of possible worlds (interpretations, models).[14] As already mentioned, the first pair of postulates essentially says that a revision by α should not mess up the preordering *within the α-worlds*, nor should it mess up the preordering *within the $\neg\alpha$-worlds*. The second pair of postulates says that the relative position of an α-world with respect to a $\neg\alpha$-world must not be worse after a revision of the belief state by α. I take it this semantics recommends that the Darwiche-Pearl postulates be obeyed by reasonable iterated belief revision operators.[15]

Lemma. Let Φ be a condition entailing that $\ulcorner\mathcal{B} * (\alpha \wedge \beta)\urcorner$ is consistent. Then any iterated revision recipe of the form

$$(+) \qquad \mathcal{B} * \alpha * \beta \ \simeq \ \begin{cases} \mathcal{B} * (\alpha \wedge \beta) & \text{if } \ldots\Phi\ldots \\ \mathcal{B} * \beta & \text{otherwise} \end{cases}$$

[14] In symbols:

(DPO1) For any two α-worlds w and w', $w \leq w'$ iff $w \leq^*_\alpha w'$.

(DPO2) For any two $\neg\alpha$-worlds w and w', $w \leq w'$ iff $w \leq^*_\alpha w'$.

(DPO3) For any α-world w and $\neg\alpha$-world w', if $w < w'$, then $w <^*_\alpha w'$.

(DPO4) For any α-world w and $\neg\alpha$-world w', if $w \leq w'$, then $w \leq^*_\alpha w'$.

Note that worlds that are smaller according to $\leq$ are *more* plausible, or *closer* to the agent's beliefs than worlds that are greater according to $\leq$.

[15] Note, however, that the correspondence between the DP postulates with their semantic 'counterparts' depends on the satisfaction of other conditions. Papini (2001, pp. 292–293), for instance, shows that her reverse-lexicographic belief change operator $\circ_\triangleleft$ satisfies all semantic properties, but fails to satisfy (DP1) and (DP2).

satisfies the Darwiche-Pearl postulates.

A proof of this lemma can be found in the Appendix.

Notice that, given (AGM2), such a condition Φ also implies that $\alpha \wedge \beta$ is consistent, but, by (AGM7), Φ may be weaker than the condition used for conservative revision, viz., that $\ulcorner \mathcal{B} * \alpha \urcorner + \beta$ is consistent.

Now we show that bounded revision satisfies the Darwiche-Pearl postulates. In order to do that, we just have to verify that the definition (BoundRevIter) is of the form (+). Suppose that $\ulcorner \mathcal{B} * (\alpha \wedge (\delta \rightarrow \beta)) \urcorner$ is consistent with β. We need to check that $\ulcorner \mathcal{B} * (\alpha \wedge \beta) \urcorner$ is consistent. But by (AGM6), $\ulcorner \mathcal{B} * (\alpha \wedge \beta) \urcorner$ is identical with $\ulcorner \mathcal{B} * ((\alpha \wedge (\delta \rightarrow \beta)) \wedge \beta) \urcorner$, and the latter set is a subset of $Cn(\ulcorner \mathcal{B} * (\alpha \wedge (\delta \rightarrow \beta)) \urcorner + \beta)$, by (AGM7). Since this latter set was supposed to be consistent, we are done.

To give an impression of the scope of the lemma, we give another example of its application. Given (AGM7), the general format of (+) also covers restrained revision (Booth and Meyer (2006, p. 142) which is characterized by the following condition.

$$\mathcal{B} * \alpha * \beta \; \simeq \; \begin{cases} \mathcal{B} * (\alpha \wedge \beta) & \text{if } \ulcorner \mathcal{B} * \alpha \urcorner + \beta \text{ is consistent} \\ & \quad \text{or } \ulcorner \mathcal{B} * \beta \urcorner + \alpha \text{ is consistent} \\ \mathcal{B} * \beta & \text{otherwise .} \end{cases}$$

4. Representing belief states as order relations

In what sense do equations like the above characterize an iterated revision function for belief states? How can we get from $\mathcal{B}$ and an input of the form α or $\langle \alpha, \beta \rangle$ to the revised belief state? The main point is that the *belief sets* obtained after potential second revision steps give sufficient evidence about the *belief state* the agent is in after the first revision step.

We assume that the one-step part of belief change satisfies the AGM postulates except (AGM5). We will work with two different forms of representation of belief states that are sufficient to determine the set of beliefs held after any sequence of inputs.[16] The first is a total pre-ordering of possible worlds. Such preorderings can equivalently be presented graphically in the form of Lewis-Grove style systems of spheres (s.o.s.) of possible worlds. This is by far the

[16] There is a third representation in terms of prioritized belief bases which is particularly attractive for the operation of revision by comparison; cf. Rott (2006b).

most easily comprehensible representation.[17] For this paper, we assume that an s.o.s. \$ is a non-empty, finite set of sets of possible worlds such that for any two sets S and S' in \$, either $S \subseteq S'$ or $S' \subseteq S$ (that is, the elements of \$ are 'nested', or form a chain with respect to $\subseteq$). Intuitively, the most plausible worlds are contained in the smallest (graphically, 'innermost') sphere of \$, the second most plausible worlds are contained in the second smallest sphere, and so on. Worlds not contained in any sphere are called inaccessible according to \$. The set of sentences true at all the worlds contained in the innermost sphere express the beliefs held true by an agent in belief state \$; this is the agent's belief set and denoted by $\ulcorner \$ \urcorner$.[18]

The second way of representing a belief state is by a total ordering $\leq$ of sentences, usually called 'entrenchment relation' (Gärdenfors and Makinson 1988, Rott 2001, 2003). Such an ordering can roughly be thought of as reflecting the degree or strength of belief in the respective sentences, where all non-beliefs get assigned minimal entrenchment. More precisely, an entrenchment relation is a relation of comparative retractability. It is required to respect logical structure in two ways. First, if α implies β, then the entrenchment of α cannot be higher as that of β (*dominance*). In this respect entrenchments behave like probabilities. Second, the conjunction $\alpha \wedge \beta$ is not less entrenched than the weaker of α and β (*conjunctiveness*). In this respect entrenchments are different from probabilities. The set of sentences that are more than minimally entrenched are the beliefs held in belief state $\leq$; this is the agent's belief set and denoted by $\ulcorner \leq \urcorner$.[19]

[17] Lewis (1973), Grove (1988). "All necessary reasoning without exception is diagrammatic" said Charles Sanders Peirce (1903, p. 212, thanks to Ralf Busse for bringing this quote to my attention).

[18] The equivalence between an s.o.s. \$ and a total pre-ordering $\preceq$ of worlds is established by the bridge principle that w is more plausible than w', in symbols, $w \prec w'$, if and only if there is a sphere S in \$ such that $w \in S$ but $w' \notin S$. – The equivalence is not perfect, though. An s.o.s. \$ actually contains a little more information than an 'equivalent' total pre-ordering of possible worlds. If $\emptyset$ is in \$, then the belief set associated with \$ is inconsistent, if the set W of all possible worlds is in \$, then no world is considered impossible (or inaccessible). From on ordering point of view, it does not matter whether the empty set $\emptyset$ and whether the full set W of worlds is included as a member of \$; an ordering does not tell whether the most plausible worlds satisfy the agent's beliefs, or whether the least plausible worlds are excluded as impossible.

[19] The belief set $\ulcorner \leq \urcorner$ associated with a non-trivial entrenchment ordering $\leq$ is *always* consistent (an entrenchment relation is non-trivial if it does not relate all sentences of the language, or equivalently, if $\bot < \top$). Sometimes, if the agent has inconsistent beliefs, one must think of her entrenchment ordering $\leq$ as supporting the set $\{\alpha : \bot \leq \alpha\}$ which is the set of all sentences, by the dominance condition for $\leq$.

What does it mean to say that a system of spheres or an entrenchment ordering represents a belief state? This is not a trivial question. As we said before, a formal structure like an s.o.s. $\$$ or an entrenchment relation $\leq$ is still an abstraction from a 'real' belief state. But we can say that it *represents* a belief state if it reproduces just that aspect of belief states we have access to, i.e., the development of the agent's beliefs. More precisely, for systems of spheres this means that

$$\ulcorner(((\mathcal{B} * \alpha) * \beta) * \gamma) * \ldots\urcorner = \ulcorner(((\$^*_{\alpha})^*_{\beta})^*_{\gamma})^* \ldots\urcorner$$

or

$$\ulcorner(((\mathcal{B} *_{\delta} \alpha) *_{\varepsilon} \beta) *_{\zeta} \gamma) * \ldots\urcorner = \ulcorner(((\$^*_{\alpha,\delta})^*_{\beta,\varepsilon})^*_{\gamma,\zeta})^* \ldots\urcorner$$

for all finite sequences of inputs $\langle \alpha, \beta, \gamma, \ldots \rangle$ or $\langle \langle \alpha, \delta \rangle, \langle \beta, \varepsilon \rangle, \langle \gamma, \zeta \rangle, \ldots \rangle$, respectively. The definition for entrenchment relations is similar.

Ordering representations of belief states determine revision functions specifying, for each potential input sentence α, the belief set that would result from revising by α. Conversely, given such a revision function satisfying certain 'rationality postulates', one can reconstruct the ordering that can be thought of as underlying the revision function. All this is well-known from the belief revision literature.

For the connection between revised belief sets[20] and systems of spheres, we can make use of the following transitions (compare Grove 1988):

(From $\$$ to $\ulcorner * \urcorner$) β is in $\ulcorner \mathcal{B} * \alpha \urcorner$ if and only if there is a sphere containing some α-worlds which are all β-worlds, or there is no sphere containing any α-worlds.

(From $\ulcorner * \urcorner$ to $\$$) A set X of possible worlds is in $\$$ if and only if there is a sentence α such that $X = \{w \in W :$ for some β, w satisfies all sentences in $\ulcorner \mathcal{B} * (\alpha \vee \beta) \urcorner\}$.

For the connection between revised belief sets and entrenchments, we can use the following transitions (compare Gärdenfors and Makinson 1988, Rott 2001, Ch. 8):

(From $\leq$ to $\ulcorner * \urcorner$) β is in $\ulcorner \mathcal{B} * \alpha \urcorner$ if and only if $\neg\alpha < \alpha \rightarrow \beta$ or $\top \leq \neg\alpha$.

[20] Note that the following connections, as well as the following ones concerning revised belief sets and entrenchments, do not appeal to the belief *states*.

(From $\ulcorner * \urcorner$ to $\leq$) $\alpha \leq \beta$ if and only if α is not in $\ulcorner \mathcal{B} * \neg(\alpha \wedge \beta) \urcorner$ or $\ulcorner \mathcal{B} * \neg(\alpha \wedge \beta) \urcorner$ is inconsistent.

Although it is not necessary for the purposes of this paper, it may help to point out that the s.o.s. modelling and the entrenchment modelling are equivalent in quite a strong sense. One can easily go full circle and define an entrenchment relation from an s.o.s. (or: an s.o.s. from an entrenchment relation) that generates exactly the same revision function. The relevant transitions are (compare Rott and Pagnucco 1999):

(From $\$$ to $\leq$) $\alpha \leq \beta$ if and only if for all S in $\$$, if α is true throughout S, then β is true throughout S as well.

(From $\leq$ to $\$$) A non-empty set X of possible worlds is in $\$$ if and only if there is a sentence α such that $X = \{w \in W : w$ satisfies all sentences β with $\alpha \leq \beta\}$.[21]

These three bridges 'fit together' very well.

Since there is no danger of confusion, we shall occasionally allow ourselves to say that an s.o.s. or an entrenchment relation *is* a belief state rather than saying that it is an abstraction from, or a representation of, a belief state.

5. Bounded revision as an operation on systems of spheres

In the last section we have seen that by knowing the revised belief sets $\ulcorner \mathcal{B} * \alpha \urcorner$ for all inputs α, one can reconstruct a representation of the belief state $\mathcal{B}$. This throws a new light on the equations for iterated belief change from Section 3. By knowing the revised belief sets $\ulcorner (\mathcal{B} * \alpha) * \beta \urcorner$ for all inputs β, one can similarly reconstruct a representation of the belief state $\mathcal{B} * \alpha$. That is to say that these equations in effect specify transitions from representations of $\mathcal{B}$ to representations of $\mathcal{B} * \alpha$.

Bounded revision functions can thus be viewed as functions applying to formal representations of belief states (not on belief sets which contain too little information, not on belief states which are inscrutable). In this and the next section, we give a direct account of the relevant transitions, as applying on

[21] I neglect the problem of adding the empty set to systems of spheres. Cf. footnote 18.

systems of spheres and entrenchment relations respectively. We begin with the representation of belief states in terms of systems of spheres.

Let $[\alpha]$ denote the sets of possible worlds in which α is true. Let α intersect \$, i.e., let there be one sphere in \$ has a non-empty intersection with $[\alpha]$. $S_{\alpha;\delta}$ be the largest sphere S in \$ such that $S \cap [\alpha] \subseteq [\delta]$; if there is no such sphere, put $S_{\alpha;\delta} = \emptyset$. Let $S_{\alpha,\delta}$ be the smallest sphere S in \$ such that $S \cap [\alpha] \not\subseteq [\delta]$; if there is no such sphere, take $S_{\alpha;\delta}$ to be the largest sphere in \$. Except for the limiting cases, $S_{\alpha;\delta}$ and $S_{\alpha,\delta}$ are neighbouring spheres, the latter is just a little larger than the former.

Let $\$^*_{\alpha;\delta}$ and $\$^*_{\alpha,\delta}$ denote systems of spheres that result from revising the prior s.o.s. \$ by an input sentence α, bounded by reference sentence δ. There are two potential definitions of bounded revision as an operation on systems of spheres. Both apply for the case in which α intersects \$. If α does not intersect \$, then we simply put $\$^*_{\alpha,\delta} = \$^*_{\alpha;\delta} = \$ \cup \{\emptyset\}$. Here are the two variants.

$$\boxed{\$^*_{\alpha;\delta} = \{S \cap [\alpha] : S \in \$, S \cap [\alpha] \neq \emptyset \text{ and } S \subseteq S_{\alpha;\delta}\} \cup \{S \cup ([\alpha] \cap S_{\alpha;\delta}) : S \in \$\}}$$

and

(BoundRevSS)

$$\boxed{\$^*_{\alpha,\delta} = \{S \cap [\alpha] : S \in \$, S \cap [\alpha] \neq \emptyset \text{ and } S \subseteq S_{\alpha,\delta}\} \cup \{S \cup ([\alpha] \cap S_{\alpha,\delta}) : S \in \$\}}$$

The two ideas are obviously similar. But we shall restrict our attention to the second method in this paper.[22] This is for three reasons. First, the first method violates the success condition (AGM2) if there is no S in \$ such that $S \cap [\alpha] \subseteq \delta$. Second, while we get that α covers more spheres in the posterior s.o.s. than δ if we use the second method using $S_{\alpha,\delta}$, we get no such relation for the first method using $S_{\alpha;\delta}$. Third, the second method is the one that allows us to reconstruct both conservative and moderate revision as limiting cases. The label (BoundRevSS) is attached to the second method only.

An important design decision[23] concerns the question how to treat inaccessible

[22] We essentially decided to take the second option already in Section 3. Compare footnote 12.

[23] Most relevant for 'moderate' belief change, see below.

worlds. I suggest that once a world is inaccessible, it cannot be made accessible by bounded revision. This preservation-of-inaccessibility condition is violated in purely lexicographic revisions that give absolute priority to the most recent information. But this does not seem desirable. Suppose a world in which humans have seven heads is considered doxastically inaccessible. Then the information that, say, Jarosław Kaczyński is the prime minister of Poland, should not make a world with humans having seven heads and Kaczyński being the prime minister of Poland more plausible than a world in which the opposite is true. This is why we stick to the preservation of inaccessibility.

Figures 1 and 2 show what happens when an s.o.s. gets revised according to the two variants of bounded revision. The numbers used in these figures are just there to indicate the relative plausibilities of (regions of) possible worlds, where '1' designates the most plausible worlds, '2' the second most plausible worlds, and so on. '∞' designates the doxastically impossible or inaccessible worlds.

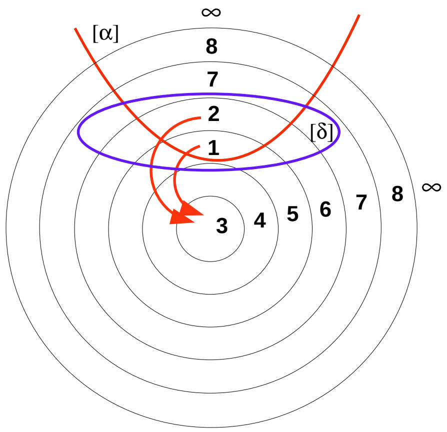

Fig. 1: Bounded revision using $S_{\alpha;\delta}$

The pictures bring out the fact that bounded revision tends to increase the number of spheres in an s.o.s. thus making plausibility distinctions finer. This contrasts with the revision-by-comparison operation of Fermé and Rott that tends to decrease the number of spheres and thus removes plausibility distinctions.

The s.o.s. presentation is generally to be preferred to an equivalent total preordering of possible worlds, because it is much easier to visualize. However, the *Darwiche-Pearl postulates* written as conditions for the change of s.o.s.s are rather less intuitive than those for orderings (compare footnote 14). As

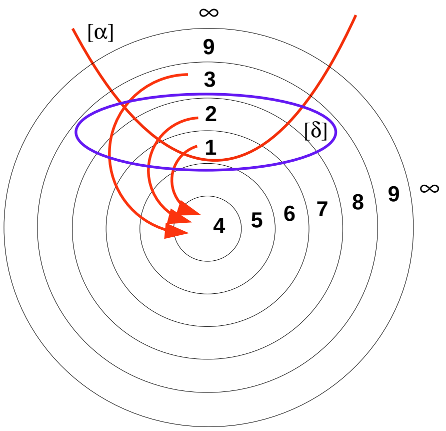

Fig. 2: Bounded revision using $S_{\alpha,\delta}$

I am not aware that the Darwiche-Pearl postulates have been represented as conditions for the change of systems of spheres elsewhere, let us give them in this form here.

We need a few preparatory definitions. If \$ is a system of spheres and α a sentence then $\$ \cap [\alpha]$ is short for $\{S \cap [\alpha] : S \in \$ \text{ and } S \cap [\alpha] \neq \emptyset\}$. If X is a set of worlds, let $C_\$(X)$ denote the cover of X in \$, i.e., the minimal sphere S in \$ such that $X \subseteq S$. Finally, let $\$'$ be short for $\$^*_\alpha$. Here now are the encodings of the Darwiche-Pearl postulates in s.o.s. language.

(DPS1) $\$' \cap [\alpha] = \$ \cap [\alpha]$

(DPS2) $\$' \cap [\neg\alpha] = \$ \cap [\neg\alpha]$

(DPS3) For every S in \$, $C_{\$'}(S \cap [\alpha]) \subseteq S \cup [\alpha]$.

(DPS4) For every S' in $\$'$, $C_\$(S' \cap [\neg\alpha]) \subseteq S' \cup [\neg\alpha]$.

A short proof of the equivalence of these DP sphere postulates with the original DP ordering postulates is given in the Appendix. We leave it to the reader to verify directly that (BoundRevSS) satisfies this version of the Darwiche-Pearl postulates.

Now let us have a look at the limiting cases for (BoundRevSS).[24] If δ is $\perp$ (or $\neg\alpha$), then $S_{\alpha,\delta}$ is the smallest sphere S in \$ such that $S \cap [\alpha] \neq \emptyset$; let us denote

[24] I neglect the case of an impossible α now.

this sphere by S_α. Then we get conservative revision:

$$\$^*_{\alpha,\delta} \;=\; \{S_\alpha \cap [\alpha]\} \;\cup\; \{S \cup (S_\alpha \cap [\alpha]) : S \in \$\}$$

If δ is $\top$ (or α), then $S_{\alpha,\delta}$ is the largest sphere S in \$; let us denote this sphere by S_{max}. Then we get moderate revision:

$$\$^*_{\alpha,\delta} \;=\; \{S \cap [\alpha] : S \in \$ \text{ and } S \cap [\alpha] \neq \emptyset\} \;\cup\; \{S \cup (S_{max} \cap [\alpha]) : S \in \$\}$$

Fig. **??** gives a picture of these (basically one-dimensional) limiting cases of bounded revision.

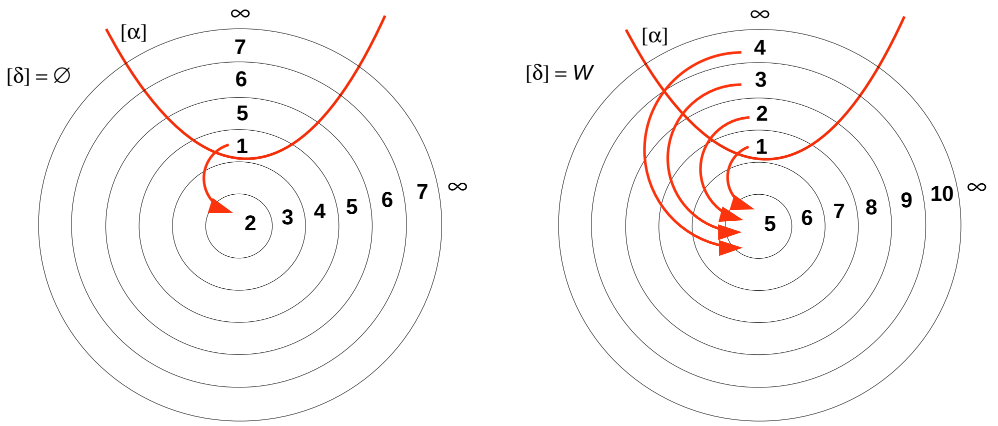

Fig. 3: Bounded revision using $S_{\alpha,\delta}$ with $\delta = \bot$ and $\delta = \top$: conservative and moderate revision

Assuming that belief states $\mathcal{B}$ can be represented by systems of spheres \$, we get the following characterization theorem:

Observation. (i) The two-dimensional revision function $*$ determined by (BoundRevSS) satisfies (AGM1)-(AGM8) and (BoundRevIter).

(ii) If the two-dimensional revision function $*$ satisfies (AGM1) - (AGM8) and (BoundRevIter), then there is, for each belief state $\mathcal{B}$, a system of spheres \$ such that at each state of the iterated revision of $\mathcal{B}$, the set of beliefs accepted in the belief state is identical with the set of beliefs determined by the system of spheres as transformed according to (BoundRevSS):

$\ulcorner\mathcal{B}\urcorner = \ulcorner\$\urcorner$

$\ulcorner\mathcal{B} *_\delta \alpha\urcorner = \ulcorner\$^*_{\alpha,\delta}\urcorner$

$\ulcorner\mathcal{B} *_\delta \alpha *_\varepsilon \beta\urcorner = \ulcorner(\$^*_{\alpha,\delta})^*_{\beta,\varepsilon}\urcorner$

… and so on.

The proof of this theorem has to be supplied elsewhere.

6. Bounded revision as an operation on entrenchment relations

We now turn to the direct account of transitions from representations of $\mathcal{B}$ to representations of $\mathcal{B} * \alpha$ in terms of entrenchment relations. Bounded revision thus becomes an iterable revision operation operating on entrenchment relations.

Let $\leq^*_{\alpha;\delta}$ and $\leq^*_{\alpha,\delta}$ denote entrenchment relations that result from revising the prior entrenchment relation $\leq$ by an input sentence α, bounded by reference sentence δ. There are two potential definitions of bounded revision as an operation on entrenchment relations, both applying (at least) for cases in which $\neg\alpha$ is less entrenched than the tautology $\top$.[25] Here are the two variants, defined by comparing any two sentences β and γ:

$$\boxed{\beta \leq^*_{\alpha;\delta} \gamma \;\text{ iff }\; \begin{cases} \alpha\to\beta \leq \alpha\to\gamma & \text{, if } \alpha\to(\beta\wedge\gamma) < \alpha\to\delta \\ \beta \leq \gamma & \text{, otherwise} \end{cases}}$$

and

(BoundRevEnt)

$$\boxed{\beta \leq^*_{\alpha,\delta} \gamma \;\text{ iff }\; \begin{cases} \alpha\to\beta \leq \alpha\to\gamma & \text{, if } \alpha\to(\beta\wedge\gamma) \leq \alpha\to\delta \\ & \quad\text{and } \alpha\to(\beta\wedge\gamma) < \top \\ \beta \leq \gamma & \text{, otherwise} \end{cases}}$$

[25] If, on the other hand, $\top \leq \neg\alpha$, then (the lower lines of) the following conditions rule that the entrenchment relation should not change at all. In order to satisfy the 'success' condition (AGM2) when α is impossible, we would have to stipulate that in this case $\ulcorner\leq\urcorner$ be the inconsistent set $\{\phi : \bot \leq \phi\}$.

The two ideas are obviously similar. In fact the two options correspond to the options that we identified with systems of spheres, and prefer to pursue the second for essentially the same reasons.[26] The label (BoundRevEnt) applies to this second option only.

With (BoundRevEnt), we get the posterior ordering $\delta <^*_{\alpha,\delta} \alpha$, that is $\delta \leq^*_{\alpha,\delta} \alpha$ and not $\alpha \leq^*_{\alpha,\delta} \delta$ (no such relation is obtained by the first option $\leq^*_{\alpha;\delta}$). But α surpasses δ after the revision only to the slightest possible degree: There is no sentence ϕ for which $\delta <^*_{\alpha,\delta} \phi <^*_{\alpha,\delta} \alpha$. Thus one could say that (BoundRevEnt) defines some kind of 'revision by comparison', in that it implements a reasonable way of minimally accepting the condition $\delta < \alpha$.

We need to make sure that the definitions lead from entrenchment relations to entrenchment relations.

Lemma. (BoundRevEnt) and its variant (the first option) define entrenchment relations, i.e. transitive relations that satisfy dominance and conjunctiveness.

The proof of this lemma has to be supplied elsewhere.

As mentioned before, we can allow arbitrary sentences to take the role of δ, but then we have to keep in mind that δ cannot be interpreted as specifying a degree of belief relative to the current belief state (characterized by $\leq$). In fact δ need not even be a belief at all in this interpretation.

We now turn to the limiting cases in which the reference sentence is a logical falsehood or a logical truth.

First limiting case. Let δ be $\bot$ (or $\neg\alpha$). Then (BoundRevEnt) reduces to:

$$\beta \leq^*_{\alpha,\bot} \gamma \ \text{ iff } \begin{cases} \alpha\to\beta \leq \alpha\to\gamma & \text{if } \alpha\to(\beta\wedge\gamma) \leq \neg\alpha \text{ and } \alpha\to(\beta\wedge\gamma) < \top \\ \beta \leq \gamma & \text{otherwise} \end{cases}$$

As already noted, this is 'natural' revision (Boutilier 1993) or 'conservative'

[26] There are some residual differences that cannot be fixed here. (BoundRevEnt) violates the success condition (AGM2) if $\top \leq \neg\alpha$ and we use the usual definition of $\ulcorner \leq^*_{\alpha,\delta} \urcorner$. What should happen after revision by an 'impossible' input is that the belief set comes out inconsistent, but, as mentioned before, the belief set associated with every non-trivial entrenchment relation is consistent. While we can change s.o.s.s by adding $\emptyset$ to \$ without changing the corresponding ordering of worlds, we do not have an analogous trick for entrenchment relations. So in such cases, one would have to follow the 'sometimes' recommendation of footnote 19. – Related to the partial violation of 'success' of (BoundRevEnt) is the partial violation of the Triangle property of Rott (2003, p. 120).

revision (Rott 2003).[27]

Second limiting case. Let δ be $\top$ (or α). Then (BoundRevEnt) reduces to:

$$\beta \leq^*_{\alpha,\top} \gamma \text{ iff } \begin{cases} \alpha \to \beta \leq \alpha \to \gamma & \text{if } \alpha \to (\beta \wedge \gamma) < \top \\ \beta \leq \gamma & \text{otherwise} \end{cases}$$

As already noted, this is 'lexicographic' revision (Nayak et al., 1994 and later) or 'moderate revision' (Rott 2003).[28]

Assuming that belief states $\mathcal{B}$ can be represented by entrenchment relations $\leq$, we get the following characterization theorem:

Observation. (i) The two-dimensional revision function $*$ determined by (BoundRevEnt) satisfies (AGM1)-(AGM8) and (BoundRevIter).

(ii) If the two-dimensional revision function $*$ satisfies (AGM1) - (AGM8) and (BoundRevIter), then there is, for each belief state $\mathcal{B}$, an entrenchment relation $\leq$ such that at each state of the iterated revision of $\mathcal{B}$, the set of beliefs accepted in the belief state is identical with the set of beliefs determined by the entrenchment relation as transformed according to (BoundRevEnt):

$$\ulcorner \mathcal{B} \urcorner = \ulcorner \leq \urcorner$$

$$\ulcorner \mathcal{B} *_\delta \alpha \urcorner = \ulcorner \leq^*_{\alpha,\delta} \urcorner$$

$$\ulcorner \mathcal{B} *_\delta \alpha *_\varepsilon \beta \urcorner = \ulcorner (\leq^*_{\alpha,\delta})^*_{\beta,\varepsilon} \urcorner$$

... and so on.

A sketch of the proof of this theorem can be found in the appendix.

7. Conclusion

We have studied a two-dimensional operation of belief revision which lies 'between' quantitative and qualitative approaches in that it does not use numbers

[27] Notice that $\alpha \to (\beta \wedge \gamma) \leq \neg\alpha$ means, as usual in the AGM paradigm, the same as $\beta \wedge \gamma \notin K * \alpha$. If this condition is satisfied $\alpha \to \beta \leq \alpha \to \gamma$ can be simplified to $\alpha \to \beta \leq \neg\alpha$.

[28] Had we presumed that only logical truths get top entrenchment (an assumption corresponding to (AGM5), then $\alpha \to (\beta \wedge \gamma) < \top$ had meant the same as $\beta \wedge \gamma \notin Cn(\alpha)$.

and is yet able to specify the extent or degree to which a new piece of information is to be accepted. It does so by specifying a reference sentence with the idea that the input has to be accepted as long as, and just a little further than, the reference sentence holds along with the input sentence. As a consequence, the input sentence is accepted just a little more strongly than the reference sentence after the revision has been performed. In both senses, the acceptance of the input sentence may be said to be bounded by the reference sentence.

In these respects bounded revision is similar to the operations of 'raising' and 'lowering' of Cantwell (1997) and of 'revision by comparison of' Fermé and Rott (2004). But there are substantial differences. Since only revision by comparison was suggested as an operation of belief *revision*, we just summarize the differences between this approach and the present one.

First, bounded revision is 'successful' in the sense that the input sentence always gets accepted, independently of which reference sentence is used. Revision by comparison, in contrast, embodies not only an operation of belief revision, but also an operation of belief contraction.[29] If the reference sentence is not more entrenched than the negation of the input sentence, then the former gets lost rather than the latter gets accepted.

Second, though both models have interesting one-dimensional belief change functions as limiting cases, these limiting cases are quite different. Taking a logical truth as the reference sentence gives 'irrevocable revision' (see Segerberg 1998 and Rott 2006a) for revision by comparison, while it gives 'moderate revision' for bounded revision. Taking a logical falsity as the reference sentence does not produce any change for revision by comparison, but generates 'conservative revision' for bounded revision. Fixing the input sentence to a logical falsity gives a 'severe withdrawal' of the reference sentence in revision by comparison. In bounded revision, it does not produce any changes in the ordering representing the belief state, but it generates an inconsistent belief set.[30]

Third, revision by comparison violates the Darwiche-Pearl postulates, since it wipes out some distinctions between worlds in which the input sentence is false.

[29] Terminologically speaking, of 'severe withdrawal.'

[30] In terms of systems of spheres, revising by an inconsistency according to (BoundRevSS) only adds the empty sphere as the new innermost sphere (which means no change to the corresponding ordering of possible worlds). In terms of entrenchments, revising by an inconsistency according to (BoundRevEnt) introduces no change in the ordering of sentences, but the belief set obtained is not determined as the sentences *more* entrenched than $\perp$, but those *at least as entrenched* as $\perp$ – and those are all sentences. Cf. footnotes 18 and 19.

In contrast, bounded revision satisfies these postulates.

Fourth, while revision by comparison tends to make distinctions between possible worlds or beliefs coarser and coarser in a series of revisions (the number of spheres in the agent's system of spheres and the number of layers in her entrenchment relation tend to decrease), bounded revision has just the opposite effect and tends to make distinctions finer and finer.

These four factors are not independent of each other. But I think that together they make a good case for regarding bounded revision as a very useful supplement to revision by comparison in particular, and to the inventory of two-dimensional revision methods in general. Going two-dimensional gives a lot of power to approaches that refrain from assuming meaningful numbers as measuring degrees of belief. As the research in belief revision progresses, an increasing number of potentially rational methods for revising one's belief states emerges. What is needed in practice though, for both numerical and non-numerical approaches alike, is a general methodology telling us when to apply which operations.

References

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

Booth, Richard, and Thomas Meyer: 2006, 'Admissible and Restrained Revision', *Journal of Artificial Intelligence Research* **26**, 127–151.

Boutilier, Craig: 1993, 'Revision sequences and nested conditionals', in *IJCAI-93 – Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, ed. R. Bajcsy, pp. 519–525.

Cantwell, John: 1997, 'On the Logic of Small Changes in Hypertheories', *Theoria* **63**, 54–89.

Darwiche, Adnan, and Judea Pearl: 1997, 'On the Logic of Iterated Belief Revision', *Artificial Intelligence* **89**, 1–29.

Fermé, Eduardo, and Hans Rott: 2004, 'Revision by Comparison', *Artificial Intelligence* **157**, 5–47.

Gärdenfors, Peter, and David Makinson: 1988, 'Revisions of Knowledge Systems Using Epistemic Entrenchment', in Moshe Vardi (ed.), *TARK'88 – Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge*, Morgan Kaufmann, Los Altos, pp. 83–95.

Grove, Adam: 1988, 'Two Modellings for Theory Change', *Journal of Philosophical Logic* **17**, 157–170.

Konieczny, Sébastien, and Ramón Pino Pérez: 2000, 'A Framework for Iterated Revision', *Journal of Applied Non-Classical Logic* **10**, 339–367.

Lewis, David: 1973, *Counterfactuals*, Blackwell, Oxford.

Lindström, Sten, and Wlodzimierz Rabinowicz: 1991, 'Epistemic Entrenchment with Incomparabilities and Relational Belief Revision', in André Fuhrmann and Michael Morreau (eds.), *The Logic of Theory Change*, Springer LNAI **465**, Berlin etc., pp. 93–126.

Nayak, Abhaya C.: 1994, 'Iterated Belief Change Based on Epistemic Entrenchment', *Erkenntnis* **41**, 353–390.

Nayak, Abhaya C., Maurice Pagnucco and Pavlos Peppas (2003): 'Dynamic Belief Revision Operators', *Artificial Intelligence* **146**, 193–228.

Papini, Odile: 2001, 'Iterated Revision Operations Stemming from the History of an Agent's Observations', in Mary-Anne Williams and Hans Rott (eds.), *Frontiers of Belief Revision*, Dordrecht: Kluwer, 279–301.

Peirce, Charles S.: 1903, 'The Nature of Meaning', Harvard Lecture delivered on 7 May 1903, published in *The Essential Peirce*, Vol. 2 (1803-1913), ed. by the Peirce Edition Project, Indiana University Press, Bloomington 1998, pp. 208–225.

Rott, Hans: 2001, *Change, Choice and Inference*, Oxford: Oxford University Press.

Rott, Hans: 2003, 'Coherence and Conservatism in the Dynamics of Belief. Part II: Iterated Belief Change Without Dispositional Coherence', *Journal of Logic and Computation* **13**, 111–145.

Rott, Hans: 2006a, 'Revision by Comparison as a Unifying Framework: Severe Withdrawal, Irrevocable Revision and Irrefutable Revision', *Theoretical Computer Scince* **355**, 228–242.

Rott, Hans: 2006b, 'Shifting Priorities: Simple Representations for 27 Iterated Theory Change Operators', in Henrik Lagerlund, Sten Lindström and Rysiek Sliwinski (eds.), *Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg*, Uppsala Philosophical Studies **53**, pp. 359–383.

Rott, Hans, and Maurice Pagnucco: 1999, 'Severe Withdrawal (and Recovery)', *Journal of Philosophical Logic* **28**, 501–547. (Printing corrupted. Correct, complete reprint in the February 2000 issue of the *Journal of Philosophical Logic*.)

Appendix

Proof of the lemma of Section 3. Revision functions according to (+) satisfy

(DP1) Suppose that β implies α. Then $\alpha \wedge \beta$ is equivalent with β, so by (AGM6), both lines of (+) entail that $\ulcorner(\mathcal{B} * \alpha) * \beta\urcorner = \ulcorner\mathcal{B} * \beta\urcorner$, as desired.

(DP2) Suppose that β is inconsistent with α. Then, by condition Φ and (AGM2), the lower line of (+) applies, so $\ulcorner(\mathcal{B} * \alpha) * \beta\urcorner = \ulcorner\mathcal{B} * \beta\urcorner$, as desired.

(DP3) Suppose for reductio that α is in $\ulcorner \mathcal{B} * \beta \urcorner$, but not in $\ulcorner (\mathcal{B} * \alpha) * \beta \urcorner$. So $\ulcorner (\mathcal{B} * \alpha) * \beta \urcorner \neq \ulcorner \mathcal{B} * \beta \urcorner$, so the upper line of (+) must apply. But by (AGM2) and (AGM1), α is in $\ulcorner \mathcal{B} * (\alpha \wedge \beta) \urcorner$, and we get a contradiction.

(DP4) Suppose for reductio that $\neg\alpha$ is not in $\ulcorner \mathcal{B} * \beta \urcorner$, but in $\ulcorner (\mathcal{B} * \alpha) * \beta \urcorner$. So $\ulcorner (\mathcal{B} * \alpha) * \beta \urcorner \neq \ulcorner \mathcal{B} * \beta \urcorner$, so the upper line of (+) must apply. But by (AGM2) and (AGM1), α is in $\ulcorner \mathcal{B} * (\alpha \wedge \beta) \urcorner$ as well, so $\ulcorner \mathcal{B} * (\alpha \wedge \beta) \urcorner$ is inconsistent, and according to condition Φ the upper line must not apply. Again we have got a contradiction. QED

Proof of the equivalence of the (DPS) postulates with the (DPO) formulations (Section 5).

For the proof, we use the bridge principle mentioned in footnote 18. Let $\preceq'$ be short for $\preceq^*_\alpha$

(DPS1) and (DPS2) are trivial.

For DPS3, we need to show that it is equivalent to

(DPO3) For any α-world w and $\neg\alpha$-world w', if $w <_{\mathcal{B}} w'$, then $w <_{\mathcal{B}*\alpha} w'$.

(DPS3) implies (DPO3). Let $w \in [\alpha]$, $w' \in [\neg\alpha]$ and $w \prec w'$. The latter means, by the bridge principle, that there is a sphere S in \$ such that w but not w' is in S. Take this sphere S. We know that $w \in S \cap [\alpha] \subseteq C_{\$'}(S \cap [\alpha])$. Now $w' \notin S \cup [\alpha]$, since $w' \notin S$. So by (DPS3), $w' \notin C_{\$'}(S \cap [\alpha])$. So $C_{\$'}(S \cap [\alpha])$ separates w and w' in $\$'$, i.e., $w \prec' w'$, as desired.

(DPO3) implies (DPS3). Let $w \in C_{\$'}(S \cap [\alpha])$. In our finite setting this is equivalent to saying that $w \in \bigcup\{C_{\$'}(\{w'\}) : w' \in S \cap [\alpha]\}$. This in turn means, by the bridge principle, that $w \preceq' w'$ for some $w' \in S \cap [\alpha]$. Suppose that $w \notin [\alpha]$, that is $w \in [\neg\alpha]$. For the claim of (DPS3), we need to show that $w \in S$. Suppose for reductio that $w \notin S$. Then $w' \prec w$, separated by S. Since $w' \in [\alpha]$ and $w \in [\neg\alpha]$, it follows by (DPO3) that $w' \prec' w$, and we have a contradiction.

(DPO4) is equivalent with: For any α-world w and $\neg\alpha$-world w', if $w' <' w$, then $w' < w$. So it is clear that this case is analogous to the case of (DP3), with \$ and $\$'$ and α and $\neg\alpha$ changing their roles. QED

Sketch of proof for the completeness part of the Observation of Section 6

We derive (BoundRevIter) from (BoundRevEnt) and the bridge principles connecting one-step revisions and entrenchment relations. This shows that if the initial entrenchment relation, obtained from one-step belief revision though ($\leq$ from $\ulcorner * \urcorner$), develops in accordance with (BoundRevEnt), then it generates, through ($\ulcorner * \urcorner$ from $\leq$), exactly the development of $\ulcorner \mathcal{B} \urcorner$ according to (BoundRevIter). This is exactly what the completeness part of the theorem claims. Notice that the bridge principles connecting two-dimensional one-step revisions and entrenchment relations do not depend on the reference sentences, due to the condition that $\ulcorner \mathcal{B} *_\delta \alpha \urcorner = \ulcorner \mathcal{B} *_\varepsilon \alpha \urcorner$ for all δ and ε.

$\phi \in \ulcorner \mathcal{B} *_\delta \alpha *_\varepsilon \beta \urcorner \quad$ iff (1: From $\leq$ to $\ulcorner * \urcorner$)

(i) $\neg\beta <^*_{\alpha,\delta} \beta \rightarrow \phi$ or (ii) $\top \leq^*_{\alpha,\delta} \neg\beta$

We now consider (i) first. Applying in the second step (2: BoundRevEnt), we get for (i)

$$\left\{ \begin{array}{ll} \alpha \rightarrow \neg\beta < \alpha \rightarrow (\beta \rightarrow \phi) & \text{if } \alpha \rightarrow (\neg\beta \wedge (\beta \rightarrow \phi)) \leq \alpha \rightarrow \delta \\ & \text{and } \alpha \rightarrow (\neg\beta \wedge (\beta \rightarrow \phi)) < \top \\ \neg\beta < \beta \rightarrow \phi & \text{otherwise} \end{array} \right\} \text{ iff (logic)}$$

$$\left\{ \begin{array}{ll} \alpha \rightarrow \neg\beta < \alpha \rightarrow (\beta \rightarrow \phi) & \text{if } \alpha \rightarrow \neg\beta \leq \alpha \rightarrow \delta \\ & \text{and } \alpha \rightarrow \neg\beta < \top \\ \neg\beta < \beta \rightarrow \phi & \text{otherwise} \end{array} \right\} \text{ iff (3: From } \ulcorner * \urcorner \text{ to } \leq)$$

$$\left\{ \begin{array}{ll} \multicolumn{2}{l}{\alpha \rightarrow (\beta \rightarrow \phi) \in \ulcorner \mathcal{B} *_\zeta \neg((\alpha \rightarrow \neg\beta) \wedge (\alpha \rightarrow (\beta \rightarrow \phi))) \urcorner \neq \mathcal{L}} \\ & \text{if } \alpha \rightarrow \neg\beta \notin \ulcorner \mathcal{B} *_\delta \neg((\alpha \rightarrow \neg\beta) \wedge (\alpha \rightarrow \delta)) \urcorner \\ & \text{or } \ulcorner \mathcal{B} *_\delta \neg((\alpha \rightarrow \neg\beta) \wedge (\alpha \rightarrow \delta)) \urcorner = \mathcal{L}, \\ & \text{and } \ulcorner \mathcal{B} *_\delta \neg((\alpha \rightarrow \neg\beta)) \urcorner \neq \mathcal{L} \\ \beta \rightarrow \phi \in \ulcorner \mathcal{B} *_\zeta \neg(\neg\beta \wedge (\beta \rightarrow \phi)) \urcorner \neq \mathcal{L} & \text{otherwise} \end{array} \right\} \text{ iff (logic)}$$

$$\left\{ \begin{array}{ll} \multicolumn{2}{l}{\alpha \rightarrow (\beta \rightarrow \phi) \in \ulcorner \mathcal{B} *_\zeta (\alpha \wedge \beta) \urcorner} \\ & \text{if } \alpha \rightarrow \neg\beta \notin \ulcorner \mathcal{B} *_\delta (\alpha \wedge (\delta \rightarrow \beta)) \urcorner \\ & \text{or } \ulcorner \mathcal{B} *_\delta (\alpha \wedge (\delta \rightarrow \beta)) \urcorner = \mathcal{L}, \\ & \text{and } \ulcorner \mathcal{B} *_\delta (\alpha \wedge \beta) \urcorner \neq \mathcal{L} \\ \beta \rightarrow \phi \in \ulcorner \mathcal{B} *_\zeta \beta \urcorner \neq \mathcal{L} & \text{otherwise} \end{array} \right\} \text{ iff (AGM1,AGM2)}$$

$$\left\{ \begin{array}{ll} \phi \in \ulcorner \mathcal{B} *_\zeta (\alpha \wedge \beta) \urcorner & \text{if } \neg\beta \notin \ulcorner \mathcal{B} *_\delta (\alpha \wedge (\delta \rightarrow \beta)) \urcorner \text{ or } \ulcorner \mathcal{B} *_\delta (\alpha \wedge (\delta \rightarrow \beta)) \urcorner = \mathcal{L}, \\ & \text{and } \ulcorner \mathcal{B} *_\delta (\alpha \wedge \beta) \urcorner \neq \mathcal{L} \\ \phi \in \ulcorner \mathcal{B} *_\zeta \beta \urcorner \neq \mathcal{L} & \text{otherwise} \end{array} \right.$$

Except for some limiting cases, we thus get a confirmation of (BoundRevIter). To deal with the limiting cases satisfactorily, we need to follow the recommendation of footnote 11 and stipulate that ($\emptyset * 1$) be valid. Then the condition that $\ulcorner \mathcal{B} *_\delta (\alpha \wedge \beta)\urcorner$ is consistent implies that $\ulcorner \mathcal{B} *_\delta (\alpha \wedge (\delta \rightarrow \beta))\urcorner$ is consistent, too. On the other hand, using (AGM6)–(AGM8), we can see that $\neg\beta \notin \ulcorner \mathcal{B} *_\delta (\alpha \wedge (\delta \rightarrow \beta))\urcorner$ implies that $\ulcorner \mathcal{B} *_\delta (\alpha \wedge \beta)\urcorner$ is consistent. So the final condition (i) as a whole reduces to:

$$\begin{cases} \phi \in \ulcorner \mathcal{B} *_\zeta (\alpha \wedge \beta)\urcorner & \text{if } \neg\beta \notin \ulcorner \mathcal{B} *_\delta (\alpha \wedge (\delta \rightarrow \beta))\urcorner \\ \phi \in \ulcorner \mathcal{B} *_\zeta \beta\urcorner \neq \mathcal{L} & \text{otherwise} \end{cases}$$

Now we consider condition (ii), $\top \leq^*_{\alpha,\delta} \neg\beta$. Applying in the second step (2: BoundRevEnt), we get for (ii)

$$\left\{ \begin{array}{ll} \alpha \rightarrow \top \leq \alpha \rightarrow \neg\beta & \text{if } \alpha \rightarrow (\top \wedge \neg\beta) \leq \alpha \rightarrow \delta \\ & \text{and } \alpha \rightarrow (\top \wedge \neg\beta) < \top \\ \top \leq \neg\beta & \text{otherwise} \end{array} \right\} \text{ iff (logic)}$$

$$\left\{ \begin{array}{ll} \top \leq \neg(\alpha \wedge \beta) & \text{if } \neg(\alpha \wedge \beta) \leq \alpha \rightarrow \delta \\ & \text{and } \neg(\alpha \wedge \beta) < \top \\ \top \leq \neg\beta & \text{if } \alpha \rightarrow \delta < \neg(\alpha \wedge \beta) \text{ or } \top \leq \neg(\alpha \wedge \beta) \end{array} \right\}$$

But the upper line is inconsistent with its condition of application, so we remain with the lower line. Since $\top \leq \neg\beta$ implies $\top \leq \neg(\alpha \wedge \beta)$, this the lower line reduces to

$\top \leq \neg\beta$ iff (From $\ulcorner * \urcorner$ to $\leq$)

$\top \notin \ulcorner \mathcal{B} *_\zeta \neg(\top \wedge \neg\beta)\urcorner$ or $\ulcorner \mathcal{B} *_\zeta \neg(\top \wedge \neg\beta)\urcorner = \mathcal{L}$ iff

$\ulcorner \mathcal{B} *_\zeta \beta\urcorner = \mathcal{L}$

Putting together the two conditions for (i) and (ii), we finally get that $\phi \in \ulcorner \mathcal{B} *_\delta \alpha *_\varepsilon \beta\urcorner$ if and only if

$$\begin{cases} \phi \in \ulcorner \mathcal{B} *_\zeta (\alpha \wedge \beta)\urcorner & \text{if } \neg\beta \notin \ulcorner \mathcal{B} *_\delta (\alpha \wedge (\delta \rightarrow \beta))\urcorner \\ \phi \in \ulcorner \mathcal{B} *_\zeta \beta\urcorner & \text{otherwise} \end{cases}$$

which is exactly (BoundRevIter).

QED