Coherence and Conservatism in the Dynamics of Belief

Part I: Finding the right framework

Hans Rott

February 2, 1999

Abstract

In this paper I discuss the foundations of a formal theory of coherent and conservative belief change that is (a) suitable to be used as a method for constructing iterated changes of belief, (b) sensitive to the history of earlier belief changes, and (c) independent of any form of dispositional coherence. I review various ways to conceive the relationship between the beliefs actually held by an agent and her belief change strategies (that also deal with potential belief sets), show the problems they suffer from, and suggest that belief states should be represented by unary revision functions that take sequences of inputs. Three concepts of coherence implicit in current theories of belief change are distinguished: synchronic, diachronic and dispositional coherence. Diachronic coherence is essentially identified with what is known as conservatism in epistemology. The present paper elaborates on the philosophical motivation of the general framework; formal details and results are provided in a companion paper.

1. Introduction

Initiated by William Harper and Isaac Levi in the 1970s, the logical study of belief change took shape as a well-defined area of research in the hands of Alchourrón, Gärdenfors and Makinson during the first half of the 1980s.[1] The work of these authors has served as the starting point of a research program that has often been labelled the *AGM paradigm*. My considerations in this paper are of a very elementary character. They are critical of the state of reflection on some fundamental concepts and principles of the AGM paradigm, and they may indeed

[1]The classical reference is Alchourrón, Gärdenfors and Makinson (1985).

be taken as a reason to be suspicious of the whole research program. How could work in that area have gone on for about two decades if not even the most fundamental entities of the dynamics of belief and their mutual relationships are clear? Surely this cannot have been good philosophy.[2]

If the arguments to be presented in this paper are correct, it is true that there has been obscurity and confusion about the conceptualization of belief states and revision functions in the formal treatment of belief change. But this does not in itself disqualify the work that has been done in the field. For one thing, it is quite common in philosophy proper that even the most fundamental and widely used concepts are unclear and far from being understood in a uniform way. It does not follow that, say, the many papers dealing with analytic and synthetic judgements are worthless, just because there still is – more than 200 years after Kant's first *Critique* – no hope for consensus amongst philosophers about what analyticity *really* is and how it relates to necessity and apriority. And secondly, I do not even see the need for insisting that the logical work in belief revision in the last two decades has been good philosophy, since it is controversial if that work fits under the heading 'philosophy' at all. If someone prefers to class logic, even so-called philosophical logic, as a science of its own, distinct from philosophy, so be it. Then the following reflections are part of the philosophy of science, to wit, of the science of logic. We hope that we can improve our understanding of the concept of belief by analyzing the concepts and methods underlying the theory of belief revision, in precisely the same way as we have increased our understanding of nature by analyzing the concepts and methods used in the natural sciences. Philosophy need not give a foundation to the natural sciences, or to the logic of belief change, in order to dignify these fields as respectable sciences. In my view, things work rather the other way round. We first look at successful existing sciences (successful, say, in providing the means to build good cars, microwave ovens, or intelligent systems) and then try to understand what people working in these fields have really been doing all the time. So the fundamental, if perhaps controversial, assumption of this paper is that the current theories of belief revision (those developed in the AGM paradigm as well as more recent ones of broadly the same nature) are fairly interesting or successful theories. This paper does not attempt to give a philosophical *foundation* of belief revision, but a philosophical *analysis* and further *development* of the existing

[2] I think that similar feelings are partly responsible for the lack of communication between people working in belief revision and epistemology. Lehrer's (1990, pp. 141, 149, 194) theory of knowledge, for instance, uses precisely the concepts that are central in belief revision ("replacements" and "eliminations"), but he makes no recourse at all to the results achieved there (for a detailed discussion, see Rott 1996, Chapter 2). Cohen (1992, p. 28) voices a general suspicion against the literature on belief revision, but he does not take any effort to give reasons why the alleged philosophical naïvete might lead the logical theory of belief revision astray. The familiarity with the epistemological literature is not better developed among belief revisionists than the epistemologists' awareness of the research that has been done in belief revision.

practice of research in belief revision. Still I hope that conceptual analyses like those below will help us, in the long run, to come up with something like a philosophical foundation for the dynamics of belief.

The motivation for this project on coherent and conservative belief change is driven by four considerations.

(a) It seems to me that the basic *ontology*[3] of theories of belief change is still not sufficiently well-understood. Models of belief change feature two kinds of objects: *Belief states* (doxastic states) and *revisions of belief states*, mostly captured by belief revision functions.[4] At least the ontological placement of revision functions presents problems. Where do they come from, and how are we to understand them? The only clear interpretation of revision functions is the subjective one, according to which they are themselves part of the agent's state of mind. But then it seems difficult to understand how a belief revision function can at the same time be a *part of* a belief state and a function *operating on* belief states (including, presumably, the one which it is a part of). We propose a simple model that is free of these circularity problems. While in much of the literature, belief revision functions are taken to be two-place functions, I shall conceive of belief revision as best represented by single-argument functions that take sequences of inputs as their arguments.

(b) Another main concern of my project are *iterated changes of belief*. Repeated belief changes should surely be treated in every decent theory of belief change, yet Alchourrón, Gärdenfors and Makinson said almost nothing about that topic. Iterated changes also pose a special question for our ontology, because they require that the general "format" of a belief state after a revision has taken place is the same as that before the revision. This is the principle of categorial matching put forward in Gärdenfors and Rott (1995).

(c) We analyze basic philosophical *principles* guiding changes of belief as well as

[3]'Ontology' is here understood in its traditional philosophical sense as the theory of what there is, rather than as a collection of sundry kinds of methodological and modelling assumptions, as in Friedman and Halpern (1996).

[4]No "relational" (non-functional) approach is discussed in this paper. I take it to be the task of belief change theories to say what the posterior belief state should look like, given the prior belief state and a piece of input. It is not enough to say that there are a number of equally rational changes of belief. If there are multiple solutions, it is part of the task of belief revision theory to say what the agent should do with them. (This problem is analogous to the notorious multiple extension problem in nonmonotonic reasoning. Just saying that there are multiple extensions of a premise set does not tell us what to conclude from the premises.) Should he play dice? On the one hand, this does not seem to be a principled solution. On the other hand, ambiguities concerning the result of a revision are sometimes just what we expect. If the current belief state is incompletely specified (e.g., if only an unembellished set of beliefs is given), then different completions of the specification (e.g., by detailing an appropriate entrenchment ordering) will in general lead to different revision behaviour, and this ought to be reflected by a multiplicity of potentially rational posterior states.

the theorizing about them. It turns out that much can be said about the rationality of belief changes by taking into account three different concepts of coherence: a synchronic, a diachronic and a dispositional one. These concepts may serve as dimensions that allow us to locate various methods of belief revision that have been suggested in the literature. As compared with Alchourrón, Gärdenfors and Makinson, we shall impose much weaker (in fact: no) requirements of dispositional coherence, but much more demanding requirements of diachronic coherence. Diachronic coherence in the sense discussed here essentially comes down to a principle of minimal change or conservatism.[5]

(d) Finally, the present project is motivated by an interest in relations of doxastic *entrenchment* which have become one of the more important tools in the modelling of belief revision. In Rott (1998a), the companion paper to the present one, we want to find out about the preconditions for a sensible application and interpretation of entrenchment relations in belief change. We present a sound and complete axiomatization for the relations corresponding to basic belief revision functions (that need not be dispositionally coherent), and we set straight a usual (mis-)interpretation of such relations. Despite the formal connectedness of entrenchment relations, we allow for intuitive incomparabilities in the entrenchment of beliefs. In any case, entrenchments are a sufficient means for formulating, in non-quantitative terms, a strong principle of conservative belief change.

Our search for the best representation of belief states and their dynamics will be presented in five steps. First we try to find out how to philosophically position the inhabitants of the ontology underlying belief revision theories, namely: belief states, inputs, and transitions between belief states that are occasioned by these inputs. Secondly, we have a fresh look at the collection of "rationality postulates" that have been suggested by Alchourrón, Gärdenfors and Makinson, and distinguish three distinct fundamental concepts of coherence encoded in them. We shall subsequently dispense with dispositional coherence and work in the context of what AGM call "basic belief change" for which we develop a theory of doxastic entrenchments. While this means a weakening of the AGM framework, we shall thirdly strengthen the latter along the dimension of diachronic coherence by providing the means of implementing a fully-fledged conservative attitude. This attitude uniquely determines a constructive strategy for iterated belief change. Fourthly, we provide a representation theorem that shows that the conservative strategy can be captured in the context of basic belief change by a single additional axiom. Finally, we compare the present approach to existing work in the literature. It turns out that this form of conservative belief change has been characterized semantically by Boutilier (1993, 1996) in the context of the full AGM theory that presupposes dispositional coherence. Unfortunately, a decisive weakness of Boutilier's approach carries over to our more general context. We

[5]In Rott (1998a), a fourth concept of coherence shows up: the temporal coherence of strategies to deal with the import of the evidence.

diagnose the inadequacy as due to a lack of temporal coherence in conservative belief change, compare it with other qualitative models of iterated belief change and indicate ways for further research.

Due to space limitations, the present project on the concepts of coherence and conservatism in belief revision is divided into two parts. Part I (the present paper) deals with foundational questions and principles, Part II (the companion paper, Rott 1998a) is of a mainly technical nature, develops the details of the requisite concepts and theorems, and includes all the proofs.

2. Binary belief revision functions, unary projections and coherence

If we are talking about the dynamics of belief, we need to be clear about the representation of beliefs. A picture of belief in flux must include a picture of belief at rest. A general framework in which belief revision has been studied for about 20 years is this. A prior (or simply "old") belief state is changed through the impact of some input that has to be accommodated; the result is a posterior (or simply "new") belief state. More formally, a *belief revision function* $*$ takes as an argument a pair consisting of an old state S and an input Υ and yields as a result the new state $*(S,\Upsilon)$, also written in the more common infix notation $S * \Upsilon$. It is important to note that this general framework does not in any way determine the format in which belief states and inputs are being represented. In order not to complicate the following considerations too much, however, we shall start with the assumption that the input is purely propositional in nature, or more precisely, that Υ just consists of a single sentence ϕ.[6] If we denote the set of sentences in a given (propositional) language by $\mathcal{L}$ and the set of all *belief states* by $\mathbb{S}$ then a belief revision function has the following signature:

$$* \ : \ \mathbb{S} \times \mathcal{L} \ \rightarrow \ \mathbb{S}$$

When considering a revision $*(S,\phi)$, one may intuitively think of the state S as the passive component that is being revised, and of the proposition ϕ as the active component that is accomplishing the revision. Formally, however, there is

[6] It is clear that for many real cases of belief change this is not an adequate representation. Inputs might be pictures, diagrams, sounds, signals, sets or sequences of sentences, sentences labelled with some degree of certainty or importance, relations over the set of sentences, structures of the same format as entire belief states, and much more. If we say that the input is propositional rather than sentential we mean that it is the content of ϕ rather than its syntactic structure that matters; logically equivalent sentences are supposed to express the same proposition. By saying that the input is propositional, we do of course *not* prescribe what exactly should be done with ϕ; sometimes we want ϕ to be accepted, sometimes we want it to be rejected, and sometimes we might want it to be accepted or rejected in a special way, say, with a certain degree of (im-)plausibility.

no such distinction, except for the condition that ϕ be accepted in the posterior belief state (see Section 6 below).

Now not just any belief revision function will qualify as 'rational' or 'coherent'. There are arguments that only functions that are constrained in a certain way ought to be admissible in the dynamics of belief. The most general idea to express conditions of rationality or coherence seems to be that $*$ should be a structure-preserving function (a morphism) in the sense that the values $*(S, \phi)$ and $*(T, \psi)$ stand in some special relation whenever the arguments $\langle S, \phi\rangle$ and $\langle T, \psi\rangle$ of $*$ stand in a special relation. It is not easy to come up with a plausible constraint in this very general guise in which the prior state and the input vary at the same time. If, however, we consider the projections of general, two-place revision functions $*$ that result from keeping one of their arguments fixed, then we find good examples of constraints of 'rationality' and 'coherence' in the literature.

Before looking at some important examples in a little more detail, we need to announce a further simplification with which we shall start our discussion. Perhaps the simplest and most common way of representing belief states is to model them as sets of beliefs, or more precisely, as sets of sentences that are believed (believed to be true, held as true, accepted). This fits together very well with our decision to represent inputs by means of sentences in $\mathcal{L}$. Let $\mathbb{K}$ stand for the set of all *belief sets*, i.e., the set of all sets of sentences that are closed under a logical consequence operation Cn. A revision function then has the following signature:

$$* \; : \; \mathbb{K} \times \mathcal{L} \; \rightarrow \; \mathbb{K}$$

Like many other authors in belief revision, we begin by identifying belief states with belief sets. However, it is debatable even for AGM whether they really wanted to identify belief states with belief sets, or whether they did not rather want to include into a belief state the selection function or preference relation suitable for guiding AGM-style constructions of belief change. Be that as it may, we shall give up the provisional identification of belief states with belief sets soon in the course of this paper.

There is yet another important restriction that we will impose on ourselves. We shall not consider any models that make essential use of quantitative or numerical information. So we do not touch upon many important approaches to belief representation, such as probabilistic models of a Bayesian kind or the 'conditional functions' or 'κ-rankings' introduced by Spohn (1988) and later used by numerous authors. On the one hand, numerical approaches are more versatile and powerful than purely qualitative models, since they can draw on the arithmetical operations of addition and multiplication. On the other hand, to unfold these superior powers, numerical approaches usually[7] need the input information

[7]But not always. The method favoured by Darwiche and Pearl (1994, 1997) is a counterexample. Also compare the remarks made by Spohn (1988, pp. 113–114).

to come in with some numerical value attached, and it is notoriously difficult to give a philosophically satisfactory answer to the question where such numerical values come from and how they can be justified. We shall therefore content ourselves with exploring the reach of qualitative models.

2.1. Varying belief states, fixing the input

Let us first consider the case where we keep the input sentence ϕ fixed. Given a two-place revision function $*$, we can define for any arbitrary but fixed sentence ϕ a unary revision function

$$*_\phi \ : \ \mathbb{S} \ \rightarrow \ \mathbb{S}$$

that takes varying prior belief states S to the posterior belief states $*_\phi(S) = *(S, \phi)$ that are obtained by revising the prior states by ϕ.

One may hold, with good arguments, that this is an attractive conception for a cognitive semantics for a language: The *meaning* of a sentence can be characterized by the change it brings about in the information states of the speakers of that language. This is the central idea of the *dynamic semantics* or *update semantics* championed by Stalnaker (1974), Gärdenfors (1984), Groenendijk and Stokhof (1991), and Veltman (1996). As far as I know, dynamic semantics up to now has taken into account only revisions by formulae that are consistent with the prior beliefs of the speaker. In the inconsistent case, revisions get more complicated since they are determined not only by the meaning of the input sentence ϕ, but also by internal factors like preferences associated with the speaker's mental state. These factors, however, may well be taken to be part and parcel of the belief states themselves, and are responsible for the fortunate fact that belief-contravening information is not indigestible information. Therefore I see no reason to exclude, from the (determination of the) meaning of ϕ, those changes that ϕ brings about in belief states with which it is logically incompatible.

Now let us leave dynamic semantics and look at two examples of coherence constraints for $*_\phi$ that are independent of the particular content or meaning of ϕ. For the start, we consider the simple case where belief states are logically closed belief sets, that is, where $\mathbb{S}$ is $\mathbb{K}$. First, it is a well-known fact that the validity of the following monotony condition characteristically distinguishes *updates* in the sense of Katsuno and Mendelzon (1992) from epistemic *revisions* in the sense of Alchourrón, Gärdenfors and Makinson:

(Distribute) $\quad *_\phi(K \cap K') = *_\phi(K) \cap *_\phi(K')$

Updates of belief states are occasioned by observed changes in the world, for instance, changes that result from some action that has been performed, whereas revisions are occasioned by new information about a static world. Formally, the distinction between updates and revisions is a qualitative analogue of the (probabilistic) distinction between imaging and conditionalization first made by Lewis (1976). The distinction is most striking when the input is consistent with the prior belief state, since updates violate the preservation condition ($*4$) discussed in Section 6 below.[8]

It has been pointed out repeatedly that the method of imaging alias updates can be used to save – and is in fact enforced by – certain analyses of conditionals in terms of changes of (qualitative or probabilistic) belief changes.[9] Notice finally that (Distribute) implies (but is not implied by) the following monotonicity condition which has played a prominent role in the literature on belief revision and its application to the analysis of epistemic conditionals:

(Monoton) If $K \subseteq K'$ then $*_\phi(K) \subseteq *_\phi(K')$

The second condition we want to consider has been isolated by Alchourrón and Makinson (1985, Observation 7.5) and Rott (1992, Section 7):

(Extern) If $\neg\phi \in K$, then $*_\phi(K) = *_\phi(\mathcal{L})$

This condition is satisfied by methods for belief change in which one and the same revision mechanism is applied in the context of every conceivable set of beliefs. One can dub the method "external belief change" because in the interesting case where the input is inconsistent with the currently held beliefs, there is no link at all between the revision method and these very beliefs, or any beliefs that might have been held earlier by the agent. Actually the condition just mentioned was first formulated for belief contractions rather than revisions. In this version (Extern) says that $K \dot{-} \phi = K \cap (\mathcal{L} \dot{-} \phi)$ whenever $\phi \in K$. Alchourrón and Makinson in effect proved that the condition is satisfied by so-called safe contractions based on reasonably well-behaved "hierarchies" over the set $\mathcal{L}$ of all sentences in the language, while Rott showed that it is satisfied by contractions based on a generalized relation of "epistemic entrenchment" over $\mathcal{L}$. Rott (1996, Section 7.9) notes that this version of (Extern) is also valid for contractions based on

---

[8] Gärdenfors characterizes probabilistic imaging by a quantitative analogue of (Distribute), viz., by

(P-Linearity) $*_\phi(P\alpha P') = (*_\phi(P))\ \alpha\ (*_\phi(P'))$

where P and P' are probability functions, $*_\phi(P)$ and $*_\phi(P')$ are the respective revised probability functions, and $P\alpha P'$ denotes the linear combination $\alpha P + (1-\alpha)P'$. Gärdenfors opts for a version of imaging that satisfies the preservation condition. If we heed the intuition of updates occasioned by changes in the world, however, it is implausible to preserve preservation.

[9] I refer to analyses based on the so-called Ramsey test. See for instance Lewis (1976), Gärdenfors (1982, 1988), Grahne (1991), Ryan and Schobbens (1997), and Crocco and Herzig (1997).

semantic or syntactic choice functions. An interesting corollary of the condition on contractions is that $(K \dot{-} \phi) \dot{-} \psi = (K \dot{-} \phi) \cap (K \dot{-} \psi)$ whenever $\psi \in K \dot{-} \phi$. For a systematic discussion of belief change constructions defined by methods that are essentially external to the set of current beliefs, see Freund and Lehmann (1994) and Areces and Becher (1998).

2.2. Fixing belief states, varying the input

Let us now consider the complementary projection of two-place revision functions where we keep a belief state S fixed. Given a two-place revision function $*$, we can define for any arbitrary but fixed belief state S a unary revision function

$$*_S \; : \; \mathcal{L} \; \rightarrow \; \mathbb{S}$$

that takes varying input sentences ϕ to posterior belief states $*_S(\phi) = *(S, \phi)$ that are obtained by revising S by these sentences. If we take again the simple case where $\mathbb{S}$ is $\mathbb{K}$, then this is – in my opinion – exactly what the classical AGM theory of belief revision, as well as its more recent variations, are about: potential changes of a single belief set by all kinds of propositional input. I claim that this is how we should interpret AGM theory, although Alchourrón, Gärdenfors and Makinson sometimes define belief change operations as two-dimensional functions. Almost nothing is said by these authors about the revision of varying belief sets.[10] And it is not the so-called "basic" postulates of AGM that make their theory interesting, but only their "supplementary" postulates and their variants and weakenings. This is particularly evident from the classic paper of Alchourrón, Gärdenfors and Makinson (1985), where it is beautifully demonstrated that there is a great variety of conditions relating the change by a conjunctive input $\phi \wedge \psi$ to the changes individually effected by the conjuncts ϕ and ψ. Here we can be content with listing the two official *supplementary postulates* of AGM. We adjust the notation to our present concerns, but we keep the original AGM labelling.

($*$7) $\quad *_K(\phi \wedge \psi) \subseteq Cn(*_K(\phi) \cup \{\psi\})$

($*$8) $\quad$ If $\neg\psi \notin *_K(\phi)$, then $*_K(\phi) \subseteq *_K(\phi \wedge \psi)$

We shall return to the interpretation of these postulates in Section 6.

[10] Alchourrón and Makinson (1985) are not interested in commenting on the intuitive adequacy of (Extern), and neither is Gärdenfors (1986, 1988) concerned with discussing (Monoton). The passages dealing with varying belief sets in the rather extensive work of AGM comprise only very few pages.

3. A two-component model of belief states

In the last section, we have provided a notion of a belief state – the set of sentences held true by the agent – and a notion of a belief revision function – a two-place function taking belief states and sentences as arguments. We have also got a first feeling of a double-edged notion of coherence that can be applied to two-place revision functions. Still the situation is not satisfactory. Of course, the set of beliefs of an agent is something that should count as a feature of his mental state. But what about the belief revision function itself? Where is its proper place in our ontology? Do we, *qua* belief revision theorists, have a right to say how agents ought to change their beliefs, since we know which ways of doing so are *objectively* right? Or is it essentially up to the agent himself how to change his beliefs, and the best we can do is to place certain constraints on the agent, to the effect that *if* the agent does *this* (in a given situation) he should *consequently* do *that* (in the same or another situation) as well? It seems quite obvious to me that the latter option is the right one. If we have, for instance, two sentences ϕ and ψ, there is no objective criterion telling us which of ϕ and ψ to give up in case of conflict. What we can say, though, are things like that: *If* the agent chooses to give up ϕ he should consequently give up his belief that $\phi \wedge \chi$ is true as well. I conclude that the belief revision function – or equivalently, some structure on which the belief revision function can be based[11] – is part of the agent's mental state, and since it is concerned with beliefs, it is part of the doxastic state (which in turn is part of the mental state).

It is natural, therefore, to think of an agent's doxastic state as a pair $\langle K, * \rangle$ where K is a set of beliefs and the two-place function $*$ represents the agent's *belief change strategy*. The first is the static, the second the dynamic component of a belief state, and it is important to see that each of the components can be thought of as independent of the other. That the set of beliefs K does not determine a belief revision strategy is evident; one can easily imagine two agents (or one agent at two different points of time) who entertain(s) the same beliefs but react(s) differently when confronted with belief-contravening information.[12] But, on the other hand, the belief change strategy $*$ does not determine the set of beliefs either.[13] It provides for ways of revising beliefs, no matter what the

[11]That there is an equivalence between unary revision functions and some such structures belongs to the core of the AGM lore. More specifically, relevant structures have been choice functions over maximal non-implying subsets (Alchourrón, Gärdenfors and Makinson 1985), over worlds (Grove 1988) and over sentences (Rott 1996). Somewhat less general structures are preference relations on which choices may be based, an instrument used for instance in Alchourrón and Makinson (1985), Gärdenfors and Makinson (1988) and Katsuno and Mendelzon (1991).

[12]This description presupposes, of course, that K is a set of "objective" beliefs, excluding higher-order beliefs about the agent's own beliefs and belief-revision behaviour.

[13]If one works with one-place revision functions $* : \mathcal{L} \to \mathbb{K}$, as AGM do according to my interpretation, then the belief change strategy $*$ may be viewed as determining the belief set

beliefs actually happen to be. This is as it should be. It is not implausible to hold that the agent's beliefs are to a large extent a matter of chance, dependent for instance on which articles he happens to have read in the morning newspaper. In contrast, his belief change strategy relies on his appreciation of certain kinds of beliefs and may be thought of as more stable and less susceptible to contingent inputs than the set of beliefs itself.[14] The function $*$ represents a comparatively permanent *disposition to change* beliefs, while many of the beliefs actually held are quite ephemeral. The coherence constraints of the kind mentioned above for $*$ (more specifically, for all the $*_\phi$'s and all the $*_K$'s) are conditions that make no reference to a particular set of actually held beliefs.

The model we have sketched now is fairly general and seems to provide all that is necessary for a formal analysis of the statics and dynamics of belief. But we have for the second time reached a state which seems satisfactory but isn't really.[15]

There are two problems. First, we must admit that there is some sort of unresolved *circularity* in our intuitive argument which is resolved in the model, but at a cost that may turn out to be too high. The circle is this. In the beginning of the paper, we said that beliefs states are *arguments* and *values* of (two-place) belief revision functions. Later we have argued that belief revision strategies should be viewed as *parts* of an agent's belief state, and that they should be represented by (two-place) belief revision functions. But then we do not know what to take as primitive, belief states or belief revision functions. Belief states are arguments and values of belief revision functions which in turn are parts of belief states. In the model we have reached now, the circle is broken by letting revision functions take sets of beliefs as arguments – but a set of beliefs is only one component of a fully-fledged belief state. We thus violate the fundamental idea with which we opened this paper.[16]

The second problem is related to the first and gets relevant if we consider *iterated belief change*. On the face of it, repeated changes of beliefs pose no problem for the present model. We said that a belief revision strategy $*$ embodies relatively stable dispositions, in a way that can be paraphrased as follows: "*If* the set of my current beliefs were K and if the input were ϕ, *then* I would proceed to the

K through the equation $K = *(\top)$. We shall return to this equation in footnote 24 below.

[14]Revisions are dependent on the values we attach to certain beliefs, or our preferences between them. Quite generally, our value judgements are not dependent on what we actually happen to possess. For better or worse, values tend to be more stable than possessions.

[15]At the end of his beautiful discussion of the "problem of deduction", Stalnaker (1984, p. 99) advocates a two-component model of acceptance states. However, the change functions he describes do not return acceptance states (in his own sense). Stalnaker does not seem to notice the shortcomings of the two-component model.

[16]It is not implied by what I have been saying that such a circularity would necessarily be vicious. For a proposal how to tame (a different kind of) self-reflexive circles in reasoning about information change, see Gerbrandy and Groeneveld (1997). Notice that I haven't yet dealt with the problem of the intuitive circularity; this will be done later.

posterior belief set $K * \phi$." Such hypothetical transitions are available for all K and all ϕ. So the problem of iterated belief change is solved if we assume – in the "normal" cases of belief change that are the intended applications of our model[17] – that only the belief set of an agent is subjected to change while his belief revision strategy may well remain the same. Given a sequence $\phi_1, \ldots, \phi_n$ of input sentences, an agent in the initial state $\langle K, * \rangle$ passes through the belief sets $K_1 = K * \phi_1$, $K_2 = (K * \phi_1) * \phi_2$, $K_3 = ((K * \phi_1) * \phi_2) * \phi_3$, and so on, till he finally arrives at $K_n = (\ldots((K * \phi_1) * \phi_2) * \ldots * \phi_{n-1}) * \phi_n$. The belief state at any step in the sequence then just is $\langle K_i, * \rangle$. Notice that in this sequence of revision steps the two-place revision function, which never alters itself, gets fed with different arguments in both places simultaneously.[18]

4. Unary belief revision functions as representations of belief states

Why can't we be satisfied with the picture afforded so far? As regards the first problem that I mentioned, the present model avoids the intuitive circularity by simplifying the concept of a belief revision function in such a way that in effect not full belief states, but just their propositional ("static") components get revised. Concerning the second problem we made a similar assumption, viz., that the input does not at all affect the belief revision strategy. Although belief revision strategies may be assumed to be more resistant to changes than beliefs, it seems a very strong assumption after all that they are totally unaffected throughout a long series of "normal" revisions. If that were so, a doxastic agent would have no sensitivity at all for the history of his belief changes. If he starts out from the belief set K and – after a long series of experiences giving rise to many changes of belief – by chance happens to find himself endorsing the same belief set K again, the (one-place) revision function $*_K$ applied to K is precisely the same as the one at the outset. Or, to make a different but related point, if an agent first learns that ϕ and much later learns that ψ, and if he then has a belief set K that contains ϕ and ψ, the order of learning ϕ and ψ does not at all matter. First learning ψ and much later learning ϕ has precisely the same effect, if only the resulting belief set is K. But intuitively we might imagine our agent pursuing quite different strategies in the temporal processing of information. For some reason, he may attach the greatest value to the things he learned a long time ago. Or alternatively, he may especially appreciate the most recent news. I do not want to recommend *a priori* any of these strategies, but I definitely think we should have a framework for the study of belief revision that is general enough to

[17] "Revolutionary" cases that involve shifts of conceptual schemes or scientific paradigms are outside the scope of any logical theory of belief change that has been developed so far.

[18] A very different, in a way complementary picture of sequential revisions is given in Areces and Rott (1998).

allow us to model at least some of these strategies. However, sensitivity to one's history of learning is impossible if we insist that an agent's belief state comprise a single pre-determined, immutable belief revision strategy, i.e., a two-place revision function, that does not change at all under the influence of experience. I propose to take the doxastic history of an agent seriously. We must not conceive of agents as having absolutely stable strategies for belief change (two-place revision functions). We need to make room for a dependence of the revision function not only on the current belief state,[19] but also on the history of belief changes (previous belief states as well as previous inputs).

My negative suggestion is therefore to *renounce the use of two-place revision functions altogether*. Whenever such a function is paired with a belief set K, it invariably yields a unique one-place revision function $*_K$ for K. And I have argued that this is just what we need to avoid: To fix a unique revision function for each potential belief set. In the modelling of belief changes, we need the flexibility of attaching different unary revision functions to any given set of beliefs.

My preliminary positive suggestion is to *identify belief states with unary belief revision functions*. The unary functions we have been considering so far have had the format

$$* \;:\; \mathcal{L} \;\rightarrow\; \mathbb{K}$$

A belief state in this conception thus is a function that responds to each conceivable propositional input ϕ by returning a belief set containing what would be believed if the information that ϕ is true were actually coming in. The set K of current beliefs can be obtained by applying $*$ to the trivial input $\top$, i.e., $K = *(\top)$. In this picture, a revision function does not *revise* a belief state – let alone potentially revise all possible belief states – but *a revision function is a belief state*. Actually, a revision function does not revise anything; in particular, there are no primitive entities in the study of belief revision that could be revised by such a function. Revision functions are themselves the primitive entities of the theory of belief revision.

The intuitive circularity mentioned above has not yet been solved. If unary revision functions are primitive and the appropriate formal representatives of doxastic states, how do *they* get revised by propositional inputs?[20] Before answering this question we shortly reflect on some desiderata for the revision of belief states in this sense. First of all, we want the changes to be as gentle and smooth as possible. A sequence of doxastic states represents the mental development of an agent in time, and as a matter of personal identity, his doxastic evolution should be comprehensible as coherent when looked at as a whole – as coherent as is compatible with the impact of the input. From this idea of *diachronic coherence*

[19]See the discussion of the AGM postulates $(*3)$ and $(*4)$ below.

[20]To the best of my knowledge, revisions of revision functions ("meta-revisions") have first been addressed as a subject of its own by Nayak et al. (1996).

it follows, or at least appears to follow, that the difference between successive belief states should be kept as small as possible. This conforms to the idea of *minimal change* or *conservatism*. Usually conservatism in the theory of belief revision is taken to mean that there should be minimal changes between successive *belief sets*. I shall argue that this idea of "minimum mutilation" (Quine's term) has played a much less important role in belief revision theory than most people think. Actually I think there is some myth about minimal change that ought to be deconstructed.

Conservatism as applied to *belief states* has not been dealt with very extensively in the literature so far. After having made precise what this sort of conservatism is supposed to mean, we shall explore some of its good and bad properties. In our modelling, the watchword is: Keep as much of the structure of the prior revision function as possible! Unfortunately, it is not at all obvious how the similarity between successive revision functions is to be measured. In the companion paper to the present one (Rott 1998a), we make this idea concrete by using a detour via a particular kind of structure that may be used for – and is in fact "equivalent with" – unary revision functions.[21]

5. The right representation of belief states: Unary iterated belief revision functions

Now all these considerations are relevant and valid, I submit, so long as we consider only singular changes occasioned by one input sentence. We have decided, however, to take the concerns of iterated belief revision seriously. Interestingly enough, this will give us the key to solving the intuitive circularity we described above. I think the only way of dealing with all the problems we have come across so far is by generalizing the notion of a (unary) revision function to one that can take not only single input sentences, but *sequences of input sentences*. I suggest that the right format for a revision function is

$$* \ : \ \mathcal{L}^\omega \ \longrightarrow \ \mathbb{K}$$

Here $\mathcal{L}^\omega = \bigcup_{n=0,1,2,\ldots} \mathcal{L}^n$ is the set of all finite (possibly empty) sequences of $\mathcal{L}$-sentences.

In order to keep formulations as simple as possible, we introduce the following bits of notation (which will be most useful for the companion paper, Rott 1998a). The empty sequence $\langle \ \rangle$ is denoted by $\mathbf{0}$, the sequences $\langle \phi_1, \ldots, \phi_n \rangle$ and $\langle \psi_1, \ldots, \psi_m \rangle$ are denoted by Φ and Ψ. Sequence concatenation gets represented by $\cdot$, so the

[21]Revision functions of this signature are "equivalent" with any of the kinds of structures that are necessary and sufficient to construct unary belief revision functions according to some given construction recipe. Compare footnote 11.

sequence $\langle\phi_1,\ldots,\phi_n,\psi_1,\ldots,\psi_m\rangle$ for instance is written as $\Phi\cdot\Psi$. A sequence $\langle\phi\rangle$ of length 1 is identified with the sentence ϕ. We define the current belief state K to be the state "arrived at" through the empty input sequence $\mathbf{0}$, i.e., $K = *(\mathbf{0})$. While a two-place revision function *takes various belief sets as arguments*, a unary revision function may in contrast be thought of as being *associated with the belief set* $K = *(\mathbf{0})$. Instead of $*(\Phi)$ we shall usually write $K * \Phi$. This set denotes what might more explicitly be written as $(\ldots((K*\phi_1)*\phi_2)*\ldots*\phi_{n-1})*\phi_n$. But we need to bear in mind that the various unary one-stepped revision functions denoted by '$*$' in the latter expression are in general different from one another! The belief sets vary in response to incoming input, and the one-stepped revision functions applied at the respective points of time vary with them. The idea of conservatism as applied to belief states, however, dictates that the variation of the one-stepped functions should not be greater than necessary, and this requirement will get encoded as a constraint on revision functions in the above format.

This conception of revision functions indeed helps to solve the circularity problem. We need not worry any more about how the revision of belief states is to be effected – because the revision is obvious. If the prior belief state is $*$ and the input is a sequence Φ, then the posterior belief state naturally is the function $*'$ that is defined by

$$*'(\Psi) \;=\; *(\Phi\cdot\Psi)$$

for all sequences Ψ. The new, revised revision function $*'$ with respect to the belief set $*'(\mathbf{0}) = *(\Phi)$ could be written as $*^*_\Phi$, but we will not adopt this confusing notation. Any sequence of inputs to the posterior state leads to the same result as the same sequence appended to Φ in the prior state. This is just what it means that the posterior state $*'$ is reached from the prior state $*$ through revising the latter by Φ. Summing up, if doxastic states are represented by iterated belief revision functions, then the problem of revising doxastic states takes care of itself. Since the dynamics of belief are implicit in the representations of the statics, we can fully concentrate on the *structure* of belief states. The idea of *diachronic coherence* or *conservatism* can now be encoded as a constraint on iterated belief revision functions – i.e., on belief *states* rather than on the revision of belief states.

Before moving on, we should pause a little and compare the result of our search for the right framework with what appears to be the most similar approach in the literature. Lehmann (1995) introduces a "revised framework" for belief revision that is similarly based on unary revision functions that take sequences of sentences as input. However, this remarkable coincidence should not conceal the fact that the intentions of Lehmann's paper and the present one are entirely different. First, Lehmann speaks of the central "concept of a belief state *resulting* from a finite sequence of revisions" (Lehmann 1995, p. 1535, emphasis added), whereas we *identify* belief states with unary, iterated revision functions. Lehmann explicitly renounces any ambition to contribute to the "epistemology of science" or the “ontology of belief revision” (p. 1538). His paper is complementary to the present one in that he *starts* with revision functions of a format that we have been labouring hard to *justify*. Lehmann proposes a semantics for belief change that is based on a fixed, static ranking of models which is to be conceived as external to the agent's state of mind, and he introduces a corresponding set of postulates that we shall find reason criticize as counterintuitive (in the companion paper to the present one). He consciously deviates from the conservative Principle of Minimal Change in favour of two other postulates that he also subsumes under the heading of ‘informational economy’ (p. 1539). There are many details in Lehmann's paper with which I disagree, but his choice of framework for the representation of belief change is indeed striking and meets the requirements for the modelling of iterated belief change.

6. Basic belief change: Interpreting rationality postulates in terms of coherence

We shall now review some elements of the classical belief revision theory – the well-known AGM postulates – and isolate three different concepts of coherence that they can be seen as embodying.

A *belief set* is a set of sentences of a given language $\mathcal{L}$, usually consistent, that is closed under logical consequences. We use $\vdash$ and Cn to indicate the consequence relation and operation governing $\mathcal{L}$, respectively. We reserve the letter ‘K’ for belief sets.

Alchourrón, Gärdenfors and Makinson developed their theories for *unary one-stepped belief revision functions*.[22] Such a function $*$ is associated with a belief set K and assigns, for each input sentence ϕ, the revision $K * \phi$ of K that assimilates ϕ. So formally the revision function $*$ for K is a function with domain $\mathcal{L}$ and range $\mathbb{K}$ (the set of all belief sets).

A function $*$ is supposed to satisfy the following conditions. In the belief revision literature these conditions (and those that will follow) are usually called “rationality postulates.” We use the AGM labels to refer to them.

[22]In their classic paper, Alchourrón, Gärdenfors and Makinson (1985) sometimes *formally* use revision functions $*$ as binary functions taking various belief sets as their first argument. But it is obvious that they are interested only in revision functions for some given belief set K, and they do not offer any constraints regarding the revisions of varying belief sets in that paper. So at least “in spirit” the first argument of the two-place functions (i.e., the belief set) may be taken as fixed, and AGM investigate essentially only the unary projections $*_K$. In contrast to other writers (e.g., Lehmann 1995, Arló-Costa 1998), I do not think it is appropriate to charge them with the view that an agent with the same belief set at different times is (or: two agents with the same belief set are) committed to revise this very belief set in the same ways. I have explained in Section 4 why I think that the conception of binary revision functions is not appropriate as a general framework in which to study iterated belief change.

(∗1) $K * \phi = Cn(K * \phi)$ *(Closure)*

(∗2) $\phi \in K * \phi$ *(Success)*

(∗5) If $\phi \nvdash \bot$, then $K * \phi \nvdash \bot$ *(Consistency)*

(∗6) If $\phi \dashv\vdash \psi$, then $K * \phi = K * \psi$ *(Extensionality)*

In this paper, I shall treat these postulates as absolutely fundamental. Roughly speaking, they say that revisions should be made in a way that is *successful* (i.e., the input is actually accepted in the posterior belief state – (∗2)), *inferentially coherent* (i.e., the posterior belief set is logically closed and consistent – (∗1) and (∗5)) and *content-driven* (i.e., the result does not depend on variations in surface grammar of the input sentence – (∗6)). We call the set consisting of (∗1), (∗2), (∗5) and (∗6) the set of *basic postulates for belief revision*, and revision functions satisfying them *basic revision functions*.

Postulates (∗1) and (∗5) taken together embody a notion of *synchronic coherence*. Synchronic notions of coherence are important for belief change, but if they are supposed to be the only notions of coherence that are relevant, they tend to deprive the theory of belief revision (in the usual sense) of its very task. *Theory change gets reduced to theory choice*: Just the best, most coherent theory will then be chosen, regardless of the predecessor theories. Belief change on this account ceases to be a relational matter (i.e., to be grounded on inter-theory relations between prior and posterior belief sets), but is rather driven solely by the structure and properties of the posterior theory. The theory chooser jumps to the theory with the best overall characteristics that fits the data, without any commitment to earlier theories.

There is another pair of postulates that AGM also call basic, but that are somewhat more problematic than those in the first group. They relate the revision function to a belief set K and express principles of *minimal change* for the case where the input ϕ is consistent with K.

(∗3) $K * \phi \subseteq Cn(K \cup \{\phi\})$ *(Expansion)*

(∗4) If $\neg\phi \notin K$, then $K \subseteq K * \phi$ *(Preservation)*

Here we are presented with a substantial recommendation of how to perform revisions by inputs that are consistent with the prior beliefs. (∗3) states that the agent should not acquire more beliefs than are necessary on the strength of (∗1) and (∗2); (∗4) tells him not to give up more beliefs than are necessary on the strength of $(*5)$.[23] Postulates $(*3)$ and $(*4)$ are vacuously satisfied if the input ϕ is inconsistent with the belief set K (i.e., if $\neg\phi \in K$). They may be regarded as restricted principles of *diachronic coherence* – restricted, that is, to the consistent case. This relational notion of coherence must clearly be distinguished from the synchronic one codified in $(*1)$ and $(*5)$ which pertains to the properties of a single (posterior) belief state. The idea of diachronic coherence is that prior and posterior belief state (or more generally, the members in a sequence of belief states) somehow "hang together." In this sense, conservativity may be interpreted as a strategy aiming at a certain kind of coherence. We call revision functions satisfying $(*3)$ and $(*4)$ *c-conservative* (with respect to K, "c" for "consistent").

Although $(*3)$ and $(*4)$ look very straightforward, it is not obvious that they ought to be satisfied. In the important form of belief "updates" that are occasioned by changes in the world, $(*4)$ gets violated (Katsuno and Mendelzon 1992). The same is true in the approach to foundational belief change advocated in Rott (1996, Chapter 5), and there are reasons against identifying consistent revisions ("additions") with expansions if the object language contains autoepistemic operators or conditionals (Rott 1989, 1991). Further interesting arguments against Preservation are put forward by Rabinowicz (1995), Levi (1996, Chapters 2 and 3) and Arló-Costa (1998). I am not going to argue for or against $(*3)$ and $(*4)$ here, but I do want to draw the reader's attention to their being more open to controversy than the four postulates I call basic.

Finally, there are the "supplementary" AGM postulates $(*7)$ and $(*8)$ which we presented already in Section 2, using a somewhat different notation. Here are the originals:

(∗7) $\qquad K * (\phi \wedge \psi) \subseteq Cn((K * \phi) \cup \{\psi\})$

(∗8) $\qquad$ If $\neg\psi \notin K * \phi$, then $K * \phi \subseteq K * (\phi \wedge \psi)$

It has frequently been pointed out that $(*7)$ implies $(*3)$ and that $(*8)$ implies $(*4)$ – provided that we assume that $K = K * \top$.[24] But saying this tends

[23] The original AGM conditions actually have as the fourth condition some kind of converse of $(*3)$, viz.,

$(*4')$ $\qquad$ If $\neg\phi \notin K$, then $Cn(K \cup \{\phi\}) \subseteq K * \phi$

The additional strength of $(*4')$ over $(*4)$, however is derivable from the conditions $(*1)$ and $(*2)$. In order to avoid redundancies in the axioms, we use the more elementary Preservation condition $(*4)$. – The pair $(*3)$ and $(*4)$ may be taken to express the requirement that (unary) belief revision functions as applied to the current belief state should be *faithful* to that belief state. Notice that both $(*3)$ and $(*4)$ are vacuously satisfied if K is the inconsistent belief set $\mathcal{L}$.

[24] If K is consistent, the identity $K = K * \top$ can itself be derived from $(*3)$ and $(*4)$. An alternative idea, taking unary revision functions as the only primitives in belief revision, is to

to obscure the fact that $(*3)$ and $(*4)$ really deal with something completely different from what $(*7)$ and $(*8)$ are about. The former pair compares the prior and the posterior belief set in the case of a revision by an input that is consistent with the prior state. The subject matter of the latter pair is orthogonal, as it were, in that it compares revisions by two different, but logically related input sentences, to wit, ϕ and $\phi \wedge \psi$ (compare Figure 1).[25] Important results in belief change theory have shown that $(*7)$ and $(*8)$ are equivalent to the existence of a well-behaved, "rationalizing" structure that is ascribed to the agent's mental state and thought to govern his belief changes. Thus $(*7)$ and $(*8)$ are about the agent's *dispositions* to change his beliefs in response to *potential* inputs. I call $(*7)$ and $(*8)$ *dispositional postulates*, and revision functions satisfying $(*7)$ and $(*8)$ *dispositional revision functions*. There are many variations on $(*7)$ and $(*8)$, and amongst these, $(*7)$ and $(*8)$ have turned out to be particularly strong dispositional postulates.[26] In many respects one can say that only they – or perhaps some weakenings of them – make the AGM theory of belief revision powerful and interesting. However, they say nothing about the relation between prior and posterior belief states. It is one of the purposes of the companion paper to show that neither they nor even some weakenings of them are necessary to establish strong results about diachronic coherence or conservatism in belief change.

Revision functions satisfying $(*1)$ through $(*8)$ are called *AGM revision functions.* Notice that Alchourrón, Gärdenfors and Makinson impose *no conditions whatsoever* that encode a requirement of minimal change for $K*\phi$ in relation to K for the (much more interesting) case where ϕ is inconsistent with K.[27] *It is a pure myth that minimal change principles are the foundation of the existing theories of*

interpret the equation $K = K * \top$ as the *definition* of the current belief set. In the orthodox AGM theory, however, the equation is satisfied only when K is consistent, and it is certainly less natural than our equation $K = K * \mathbf{0}$.

[25] Robert van Rooy has drawn my attention to the fact that there is a tradition in dynamic semantics that interprets *any* change by a conjunction as a sequential change by the two conjuncts (see Groenendijk and Stokhof 1991, pp. 47, 54). In such a context of "dynamic conjunction", $(*7)$ and $(*8)$ would of course say something about diachronic coherence. The concept of conjunction that is being used in his paper, however, is the classical, symmetrical one.

[26] See Rott (1996, Chapter 4). In this respect, I very much disagree with Boutilier (1996, p. 272) who finds these postulates "quite mild." While it is true that AGM say next to nothing about the problem of iterated belief change, their conditions $(*7)$ and $(*8)$, *as conditions constraining revisions by different inputs*, are very powerful indeed. They basically imply that all beliefs in a belief set are comparable with one another in terms of entrenchment, a requirement that Boutilier accepts but that will be deliberately avoided in the companion paper.

[27] Boutilier (1996, p. 264) and Darwiche and Pearl (1997, p. 2) call the "principle of informational economy" or the "principle of minimal belief change" the hallmark of the AGM theory. They are echoing familiar prejudices here, introduced by AGM themselves and repeated time and again in the literature.

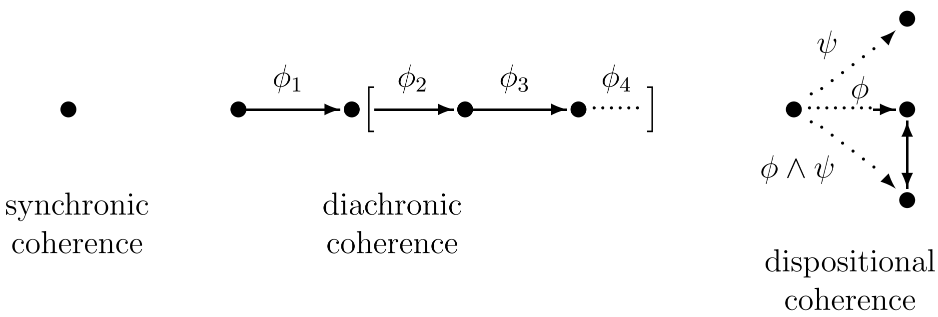

Fig. 1. *The relata of the three types of coherence*

belief revision, at least as far as the AGM tradition is concerned. This is already evident from the fact that the revision function which sets $K * \phi = Cn(\{\phi\})$ in the inconsistent case (and $K * \phi = Cn(K \cup \{\phi\})$ in the consistent case) perfectly satisfies all the AGM postulates.[28]

7. Conclusion

In this paper I have tried to offer some conceptual clarifications for the modelling of belief change, and in particular iterated belief change. First I suggested an abstract approach that allows us to represent coherence constraints for the dynamics of doxastic states by means of certain morphisms. The problem with this approach, though, was that is was unclear where belief revision functions should be ontologically positioned.

The attempt to solve this problem led us to two-component models of belief change, composed of a belief set and a belief change strategy that are independent of one another. This model was found wanting and was thus rejected. We came to the conclusion that the best representation of doxastic state is afforded by a *unary, iterated belief revision function*. In contrast with the idea that revision functions should take pairs consisting of a belief state and an input sentence, the modelling advocated here does justice to the intuition that the revision of belief sets is in general dependent on the doxastic history of agent. Our new format for revision functions solves the problem of circularity in earlier characterizations of the static and dynamic components of belief states, and it allows us to capture truly dynamic constraints on belief change (constraints concerning strategies for repeated changes of belief) by features of the structure of the initial belief state.

[28]In the AGM theory of belief *contraction*, the – very controversial – postulate of Recovery may be regarded as a partial explication of minimal change. Its effects, however, vanish completely if contractions are used only as an intermediate for the construction of revisions. For a detailed discussion of the myth of minimal change in belief revision theory, see Rott (1998b).

In the last part of the paper I have identified three different concepts of coherence that help formulate important principles for solving the problem of rational belief change: *synchronic coherence* (inferential coherence, reflective equilibrium), *diachronic coherence* (minimal change) and *dispositional coherence* (rationalizability by preference orderings).[29] Because of the relative independence of these concepts, I think it is justified to conceive of them as three different *dimensions* of coherence (see Figure 2).[30] We found that Alchourrón, Gärdenfors and Makinson

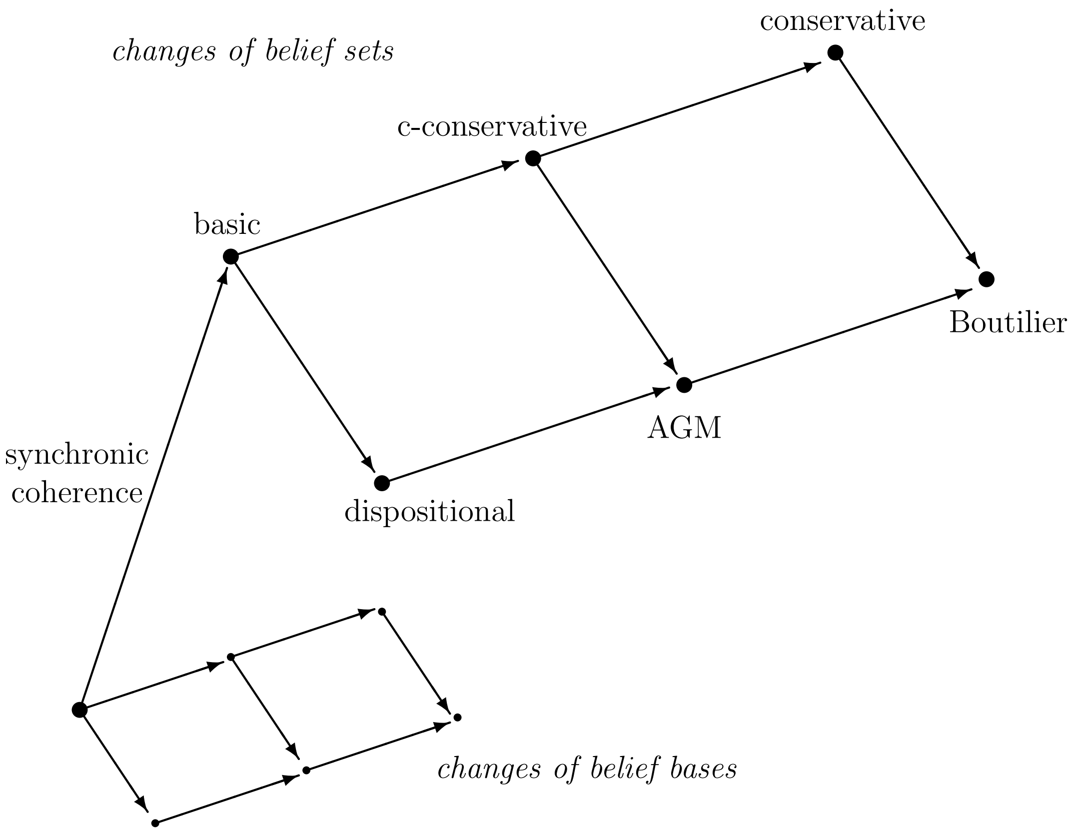

Fig. 2. *The three-dimensional space of coherence*

went quite far as regards inferential and dispositional coherence, but that they

[29]I do not, of course, mean to suggest that an analysis of the (basic) AGM postulates exhausts all there is to synchronic and diachronic coherence. Horacio Arló-Costa (personal communication) has pointed out to me that *suppositional coherence* is an important kind of synchronic coherence (*synchronic*, because "merely" suppositional change is essentially different from "genuine" change due to new information).

[30]In Rott (1997), I discuss the question to which extent the three concepts of coherence are actually conflicting with one another.

said surprisingly little about diachronic coherence which we identify with the idea of minimal change or conservatism.

The task before us in the second part of the present project (Rott 1998a) is, first of all, to develop a notion of conservative belief change for the situation where the input is inconsistent with the current beliefs – this being the more interesting situation for which the theory of belief revision was developed in the first place. We shall present an account of conservative belief change that is at the same time (a) suitable to be used as a method for constructing iterated changes of belief, (b) sensitive to the history of earlier belief changes, and (c) independent of any form of dispositional coherence. We shall thus be addressing belief change located at the upper right node of Figure 2. It turns out that Boutilier (1993, 1996) has studied a semantical modelling of the special case of conservative belief change in which the strong supplementary AGM postulates for dispositional coherence are satisfied as well.

The structures we shall use for our analyses, viz., relations of doxastic entrenchment, can perfectly well be developed in the context of basic one-stepped revision functions. In the companion paper (Rott 1998a), we give axiomatic characterizations of entrenchment for the basic and the c-conservative case, where belief change need not comply with the dispositional postulates characteristic for the AGM theory (see Section 6). Then we develop a simple method of conservatively revising entrenchment relations, formulate an extra postulate for fully conservative iterated belief change, and prove a representation theorem for the suggested construction method of conservative belief change. Finally we discuss related approaches, as well as a serious shortcoming of conservative belief change that is due to a violation of a fourth type of coherence – *temporal coherence* – that will be the subject of future research (Areces and Rott 1998).

Acknowledgement

I would like to thank Johan van Benthem, Gabriella Crocco, Michael Freund, Andreas Herzig, Isaac Levi, Pierre Livet, David Makinson, Abhaya Nayak, Erik Olsson, Robert van Rooy, Wolfgang Spohn, Bernard Walliser, and other people from audiences in Amsterdam, Konstanz and Paris for discussions of various presentations of this paper. Comments by Carlos Areces and Horacio Arló-Costa on written versions were particularly helpful. Of course I take the blame for the remaining blunders.

References

Alchourrón, Carlos, Peter Gärdenfors and David Makinson: 1985, 'On the Logic of Theory Change: Partial Meet Contraction Functions and Their Associated Revision Functions', *Journal of Symbolic Logic* **50**, 510–530.

Alchourrón, Carlos, and David Makinson: 1985, 'On the Logic of Theory Change: Safe Contraction', *Studia Logica* **44**, 405–422.

Areces, Carlos, and Verónica Becher: 1998, 'Iterable AGM Functions', to appear in Mary-Anne Williams and Hans Rott (eds.), *Frontiers of Belief Revision*, Kluwer 1999.

Areces, Carlos, and Hans Rott: 1998, 'Revising by Sequences', Manuscript, ILLC, University of Amsterdam, December 1998.

Arló-Costa, Horacio: 1998, 'Belief Revision Conditionals: Basic Iterated Systems', *Annals of Pure and Applied Logic*.

Boutilier, Craig: 1993, 'Revision Sequences and Nested Conditionals', in R. Bajcsy (ed.), *IJCAI-93 – Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, 519–525.

Boutilier, Craig: 1996, 'Iterated Revision and Minimal Change of Conditional Beliefs', *Journal of Philosophical Logic* **25**, 263–305.

Cohen, L. Jonathan: 1992, *An Essay on Belief and Acceptance*, Clarendon Press, Oxford.

Crocco, Gabriella, and Andreas Herzig: 1997, 'The Ramsey Test as an Inference Rule', in Paul Weingartner, Gerhard Schurz, and Georg Dorn (eds.), *The Role of Pragmatics in Contemporary Philosophy: Contributions of the Austrian Ludwig Wittgenstein Society*, Vol. 5, Austrian Ludwig Wittgenstein Society, Kirchberg.

Darwiche, Adnan, and Judea Pearl: 1994, 'On the Logic of Iterated Belief Revision', in Ronald Fagin, ed., *TARK'94 – Proceedings of the Fifth Conference on Theoretical Aspects of Reasoning About Knowledge*, Morgan Kaufmann, Pacific Grove, Cal., pp. 5–23.

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

Freund, Michael, and Daniel Lehmann: 1994, 'Belief Revision and Rational Inference', Technical Report TR-94-16, Institute of Computer Science, Hebrew University, Jerusalem.

Friedman, Nir, and Joseph Y. Halpern: 1996, 'Belief Revision: A Critique', in Luigia Carlucci Aiello, Jon Doyle and Stuart C. Shapiro (eds.), *Principles of Knowledge Representation and Reasoning. Proceedings of the Fifth International Conference (KR'96)*, Morgan Kaufmann, San Mateo, Cal., 421–431. Extended and revised version to appear in the *Journal of Logic, Language and Information*.

Gärdenfors, Peter: 1982, 'Imaging and Conditionalization', *Journal of Philosophy* **79**, 747–760.

Gärdenfors, Peter: 1984, 'The Dynamics of Belief as a Basis for Logic', *British Journal for the Philosophy of Science* **35**, 1–10.

Gärdenfors, Peter: 1986, 'Belief Revisions and the Ramsey Test for Conditionals', *Philosophical Review* **95**, 81–93.

Gärdenfors, Peter: 1988, *Knowledge in Flux. Modeling the Dynamics of Epistemic States*, Bradford Books, MIT Press, Cambridge, Mass.

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.

Gärdenfors, Peter, and Hans Rott: 1995, 'Belief revision', in D. M. Gabbay, C. J. Hogger, and J. A. Robinson (eds.), *Handbook of Logic in Artificial Intelligence and Logic Programming Volume IV: Epistemic and Temporal Reasoning*, Oxford University Press, pp. 35–132.

Gerbrandy, Jelle, and Willem Groeneveld: 1997, 'Reasoning about Information Change', *Journal of Logic, Language and Information* **6**, pp. 147–169.

Grahne, Gösta: 1991, 'Updates and Counterfactuals', in J. Allen, R. Fikes and E. Sandewall (eds.), *Principles of Knowledge Representation and Reasoning. Proceedings of the 2nd International Conference*, Morgan Kaufmann, San Mateo, Cal., 269–276.

Groenendijk, Jeroen, and Martin Stokhof: 1991, 'Dynamic Predicate Logic', *Linguistics and Philosophy* **14**, 39–101.

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

Katsuno, Hirofumi, and Alberto O. Mendelzon: 1991, 'Propositional Knowledge Base Revision and Minimal Change', *Artificial Intelligence* **52**, 263–294.

Katsuno, Hirofumi, and Alberto O. Mendelzon: 1992, 'On the Difference between Updating a Knowledge Base and Revising it', in Peter Gärdenfors (ed.), *Belief Revision*, Cambridge University Press, Cambridge, pp. 183–203.

Lehmann, Daniel: 1995, 'Belief Revision, Revised', in *IJCAI'95 – Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Mateo, pp. 1534–1540.

Lehrer, Keith: 1990, *Theory of Knowledge*, Routledge, London.

Levi, Isaac: 1996, *For the Sake of Argument*, Cambridge University Press, Cambridge.

Lewis, David: 1976, 'Probabilities of Conditionals and Conditional Probabilities', *Philosophical Review* **85**, 297–315.

Lindström, Sten, and Wlodzimierz Rabinowicz: 1992, 'Belief Revision, Epistemic Conditionals and the Ramsey Test', *Synthese* **91**, 195–237.

Makinson, David: 1985, 'How to Give it Up: A survey of Some Formal Aspects of the Logic of Theory Change', *Synthese* **62**, 347–363.

Nayak, Abhaya C., Norman Y. Foo, Maurice Pagnucco and Abdul Sattar: 1996, 'Changing Conditional Beliefs Unconditionally', in Yoav Shoham (ed.), *TARK'96 – Proceedings of the Sixth Conference on Theoretical Aspects of Rationality and Knowledge*, pp. 119–135.

Rabinowicz, Wlodzimierz: 1995, "Stable Revision, or Is Preservation Worth Preserving?", in André Fuhrmann and Hans Rott (eds.), *Logic, Action and Information: Essays on Logic in Philosophy and Artificial Intelligence*, de Gruyter, Berlin, pp. 101–128.

Rott, Hans: 1989, 'Conditionals and theory change: Revisions, expansions, and additions', *Synthese* **81**, 91–113.

Rott, Hans: 1991, 'A Non-monotonic Conditional Logic for Belief Revision I', in André Fuhrmann and Michael Morreau (eds.), *The Logic of Theory Change*, Lecture Notes in Computer Science **465**, Springer, Berlin etc., pp. 135–181.

Rott, Hans: 1992, 'Preferential Belief Change Using Generalized Epistemic Entrenchment', *Journal of Logic, Language and Information* **1**, 45–78.

Rott, Hans: 1996, *Making Up One's Mind: Foundations, Coherence, Nonmonotonicity*, Habilitationsschrift, Department of Philosophy, University of Konstanz, October 1996. To appear under the title "Change, Choice and Inference" with Oxford University Press.

Rott, Hans: 1997, 'Drei Kohärenzbegriffe in der Dynamik kognitiver Systeme', to appear in Julian Nida-Rümelin (ed.), *Rationalität, Realismus, Revision*, GAP3 – Proceedings des dritten internationalen Kongresses der Gesellschaft für Analytische Philosophie, de Gruyter, Berlin und New York.

Rott, Hans: 1998a, "Coherence and Conservatism in the Dynamics of Belief. Part II: Iterated belief change without dispositional coherence', Manuscript, Department of Philosophy and ILLC, University of Amsterdam, May 1998.

Rott, Hans: 1998b, 'Two Dogmas of Belief Revision', mimeographed in the *Proceedings of the Third Conference on Logic and the Foundations of the Theory of Games and Decisions*, International Centre for Economic Research, Torino, December 1998.

Ryan, Mark, and Pierre-Yves Schobbens: 1997, 'Counterfactuals and Updates as Inverse Modalities', *Journal of Logic, Language and Information* **6**, 123–146.

Spohn, Wolfgang: 1988, 'Ordinal Conditional Functions', in William L. Harper and Brian Skyrms (eds.), *Causation in Decision, Belief Change, and Statistics*, Vol. II, Reidel, Dordrecht, pp. 105–134.

Stalnaker, Robert: 1984, 'Pragmatic Presuppositions', in M. Munitz and P. Unger (eds.), *Semantics and Philosophy*, New York University Press, 197–213.

Stalnaker, Robert: 1984, *Inquiry*, Bradford Books, MIT Press, Cambridge, Mass.

Veltman, Frank: 1996, 'Defaults in Update Semantics', *Journal of Philosophical Logic* **25**, 221–261.

Department of Philosophy / ILLC
University of Amsterdam
Nieuwe Doelenstraat 15
1012 CP Amsterdam
The Netherlands