

Beliefs are in our heads, and so are our preferences: Some philosophical reflections on belief revision

Extended abstract for the ILLC workshop “Changing Minds”, 29 October 2004

Hans Rott, Regensburg

Theories of belief revision are typically based on two assumptions. The first one is that beliefs should conform to the demands of logic: They should be consistent and they should be closed under logical consequences (in sum, they should be “logically coherent”). In this sense, logic guides the set of beliefs held by the agent at a given point of time. On the other hand, theories of belief revision presume that logic is *not* sufficient to guide the *dynamics* of belief. The dynamics comes into play if the logically or otherwise well-balanced belief state of an agent is disturbed by an external input into the belief system. Usually, belief revision theory has assumed that the agent’s moves are based on, or determined by, some extra-logical structure, like preference relations or other means for effecting certain choices. These are necessary because non-trivial processes of belief revision involve choices: choices which beliefs to give up in the face of contradictory beliefs, or choices which scenarios to take seriously in the face of evidence contradicting the scenarios considered possible before.

My talk addresses the question how preferences are supposed to guide the agent’s moves in the space of potential belief states and to what extent this picture leaves an agent the freedom to choose her beliefs as she pleases. I hope that I will be pardoned if I address these issues by freely drawing from some conclusions I have reached in my own earlier research. (The simple reason is that this is the research I know best.)

Here are some important ideas that have driven traditional methods of belief revision, as encoded in the famous AGM postulates that are usually labelled (*1)–(*8) [Alchourrón, Gärdenfors and Makinson 1985, Gärdenfors 1988]:

- *Input-related criteria*
 - Success (priority of new information): (*2)
 - Syntax-independence (semantics of the input): (*6)
- *Coherence criteria*
 - Static coherence (logical coherence, semantics of the belief set): (*1) and (*5)
 - Dynamic coherence (minimum mutilation, conservatism, inertia, encoded very weakly in AGM): (*3) and (*4)
 - Dispositional coherence (rationality in the sense of rational choice theory, relationality): (*7) and (*8)

I will not elaborate on these well-known criteria in the present talk, since this has been done elsewhere [Hansson 1997, Rott 1999, 2000, 2001, 2003b].

I will also neglect the very interesting question in which format the input arrives. Does the agent receive plain propositions, or propositions together with some source signature, or propositions together with some quantitative signature specifying the certainty (reliability, “strength”) of the proposition to be accepted, or comparisons of propositions with regard to their relative certainty (reliability, “strength”), or maybe structures that are even more complicated (such as whole preference relations)? In the following, I shall refer to propositional input only; it remains to be thought through carefully whether the philosophical points below transfer to more complex or general cases.

Various representation theorems proved in the last two decades of belief revision research have shown that AGM methods are (reconstructible as) preference-based revisions, regardless of whether AGM’s recipes for partial meet contraction/revision, safe contraction/revision, possible-worlds models or contraction/revision or epistemic entrenchment contraction/revision are employed. Sven Ove Hansson [1995] has suggested that preferences and similar structures featuring in formal theories of belief revision are *hidden structures of belief*, and he illustrated his suggestion with a figure depicting a brain

in an opened skull. Borrowing from this picture, we can say that preferences are in our heads, just as beliefs are. Does that mean that we need to search for neural structures that correspond to the structures stipulated by formal theories of belief revision? Questions of this kind will be addressed in my paper.

The *ontology* of belief revision is simple. First, there are the *beliefs* of an agent – the field is called “belief revision” after all. Secondly, there are *functions* that map the agent’s prior beliefs to a set of posterior beliefs – posterior with regard to a revision incident occasioned by a certain piece of input. *Inputs* are the last items to be dealt with in belief revision theories.

It is easy to extract the above ontology from formal theories of belief change. What we cannot extract from such theories is their interpretation. Most of them are intended as normative theories specifying how *ideally rational* agents should or would change their beliefs. But suppose that Maria is such an ideally rational agent. So again, where do her beliefs and her belief revision functions reside, and where do they come from? Are they really to be found “in her head”, or are they to be found elsewhere?

The situation we encounter in formal theories of belief change seems to give rise to a *paradox*: On the one hand, belief revision functions in the full sense should be functions revising *belief states* (whatever such states may be – in any case it seems to little to require that such functions revise only sets of plain beliefs). On the other hand, belief revision functions (or the structures on which they are based) themselves appear to be parts of belief states. But they surely cannot take themselves as arguments.

Let us see how the apparent paradox can be tackled. At first it seems that we should base our considerations on two-component models, with pairs consisting of a belief set and a two-place belief revision function taking (prior) belief sets and input propositions as arguments and returning (posterior) belief sets, or alternatively, with pairs consisting of a belief set and some structure powerful enough to determine a belief revision function. I have argued in [Rott 1999] that although it makes good sense to think of the two components as independent of each other, such two-component models do not work because they can’t resolve the paradox and they would leave (in the most natural way of employing them) revision function unchanged in the light of any evidence, so that genuine learning becomes impossible.

My conclusion in [1999] was that the right response to the paradox is to renounce the use of two-place revision functions altogether and work with unary functions instead. In order to secure the means for iterated changes of belief, however, the revision function has to be ready to take sequences of propositions as inputs (giving rise to iterated changes of beliefs) rather than only one-shot revision functions (as introduced by AGM). In this model, a belief state is represented by a revision function that assigns to any sequence of inputs the set of sentences that would be believed if the agent actually received this sequence of inputs. The set of beliefs held at the start can be obtained from such a revision function by feeding in the empty sequence as an argument.

In the light of this account, the term “revision function” is a misnomer: Revision functions do not *revise* doxastic states, but rather *are* doxastic states. One might even say that revision functions in this sense do not revise anything. But in a derivative way the set of current beliefs does indeed change when input comes in. Compared with the revision function itself, the set of beliefs held at a given time is only a comparatively minor aspect of a doxastic state, a surface phenomenon so to speak. The revision function itself changes, too, but it does so only in a trivial way. If $*$ is the initial revision function and the input is $\langle \phi_1, \dots, \phi_n \rangle$, then the posterior revision function $*'$ is simply defined by $*'(\langle \psi_1, \dots, \psi_m \rangle) = *(\langle \phi_1, \dots, \phi_n, \psi_1, \dots, \psi_m \rangle)$. At a deeper level, one might wish to say that the revision function itself never really changes. At a given point of time its potential values are just restricted by the agent’s past experiences which fix the initial segment of the sequence of inputs.

In [Rott 2003a], I suggested a *conservative* method of revising preference relations (in this case, so-called “entrenchment relations” that may be viewed as the preferences revealed by certain syntactic choices). I contrasted this method with other simple methods of preference revision that I called

external revision, *radical* revision and *moderate* revision. Each of these methods specifies a recipe for the revision of preference relations in response to some propositional input (again, the format of the input does not seem to be crucial for the point I wish to make here). This suggests a particular strategy of filling in the program sketched in [1999]. The sequence of inputs $\langle \phi_1, \dots, \phi_n \rangle$ gets broken up and processed in a stepwise manner. Fixing an initial preference relation \leq and a method of (external, radical, conservative or moderate) one-step preference revision, the belief set obtained in response to the input $\langle \phi_1, \dots, \phi_n \rangle$ can be retrieved from the n -fold revised relation $((\dots(\leq^*_{\phi_1})^*_{\phi_2}\dots)^*_{\phi_{n-1}})^*_{\phi_n}$. (The kind of preference relations I am talking about in this paper always allows one to retrieve the set of current beliefs). If a method for changing preferences is fixed, then possessing a preference relation is *equivalent with* possessing a unary iterated belief revision function.

A constraint for all suggested methods for transforming preferences is the so-called Triangle property. It states that the revision of a doxastic state (represented by a preference relation) needs to be made in such a way that the belief set retrievable from the posterior state (represented by another preference relation) is identical with the belief set obtained by a construction of ordinary AGM-style one-step revision. (AGM's recipe for revisions was *not* usable for iterated belief change.)

This design of iterated belief change is not the only option to make the architecture of [1999] more concrete. An essentially different strategy, not followed in any of the papers I have mentioned, is to pre-process the sequence of inputs, and then take the result $f(\langle \phi_1, \dots, \phi_n \rangle)$ – which may be, e.g., a single proposition – and perform some AGM-style one-step revision with respect to this result.

However we flesh out the architecture of [1999] – and, for that matter, all other preference-based models of belief revision I am aware of [e.g., Katsuno and Mendelzon 1991, Nayak 1994, Bochman 2001, Lehmann, Magidor and Schlechta 2001, Andreka, Ryan and Schobbens 2002, Nayak, Pagnucco and Peppas 2003, Freund 2004] –, there is one very basic common feature of these models. Belief revision theory is *deterministic* in the following sense:

Given a certain method for belief revision based on preference revision, having doxastic preferences is equivalent to being able to perform iterated revisions of belief; the mature agent's potential doxastic development in time is completely determined by her momentary preferences; her actual doxastic development in time is completely determined by her momentary preferences and the experiences she makes, i.e., the sequence of inputs arriving at her.

Behaviour governed by preferences is rational behaviour. Often preferences represent (combinations of) values. In our case we need to consider the *cognitive values* that can be attached to beliefs, such as the value of truth and the value of information [Levi 1967, Levi 2004, Rott to appear]. Sometimes, belief change behaviour appears to be such that it cannot be governed by preferences in the usual ways, and yet we would be very reluctant to call that behaviour “irrational” [Rott 2004].

The talk of rationality and the talk of choices and preferences in the semantics of belief revision are very common. This wide-spread vocabulary suggests that the agent has a *free choice* how to change her beliefs. But what does “free” mean here? Does it mean that the agent's beliefs get changed according to *her own* preferences (rather than according to some externally prescribed recipe)? But perhaps these preferences are themselves determined by external causes. So should the agent be free in some more thorough-going sense?

“Free” cannot mean that the agent's choices are not at all determined, neither by external nor by internal factors. Many philosophers have maintained that free choices are choices that are susceptible to reasons. In belief revision theory, reasons take the form of inputs to belief states (though sometimes the inputs are only imagined ones). Bieri [2001, pp. 74–83, 187–191 and *passim*] presents an extended example of a would-be emigrant who is torn between fighting his home country's terror regime and fleeing with his family into a safe place abroad. Bieri takes this situation to be a paradigm example of a free decision, even though (or: just because) the man's decision at the critical point of time depends on what he happens to see (or to imagine) last. When he looks at his old friend of the Resistance, he is

about to decide to stay, but when he sees photos of deported women and children (or when he only imagines such a horrible situation), he decides to set out for the escape. This is an instance of “free” and “rational” choice, one might say, exactly because adequate inputs lead to adequate responses. This can be so even if the hesitating refugee’s choice is fully determined by the sum of external and internal factors acting upon him.

Considerations like these raise critical questions for the field of belief revision theory. Can we rest content with a fully deterministic picture of belief change? Should there be some room for truly creative moves in belief change, moves that are not completely determined by external or internal forces (inputs and preferences)? For instance, should we think of the agent as free to choose her favoured method of revising preferences? If so, what can we say about these choices on the meta-level? If there are *norms* for choosing a suitable method of preference revision, shouldn’t these norms be studied as part and parcel of belief revision theory? If there is no freedom and thus no place for norms in belief revision, should we expect to find natural laws that *describe* what the deterministic processes of revision and meta-revision looks like? Formal models ought to cast light on material problems. Some twenty years after AGM, it seems to me that we have not even begun to link our studies of logical methods for belief change with the hard philosophical questions.

References:

- Alchourrón, Carlos, Peter Gärdenfors and David Makinson, “On the Logic of Theory Change: Partial Meet Contraction Functions and Their Associated Revision Functions,” *Journal of Symbolic Logic* 50, 1985, 510–530.
- Andreka, Hajnal, Mark Ryan and Pierre-Yves Schobbens: “Operators and Laws for Combining Preference Relations”, *Journal of Logic and Computation* 12, 2002, 13–53.
- Bieri, Peter, *Das Handwerk der Freiheit. Über die Entdeckung des eigenen Willens*, Munich: Hanser 2001, reference to the paperback edition Frankfurt a.M.: Fischer 2003.
- Bochman, Alexander: *A Logical Theory of Nonmonotonic Inference and Belief Change*, Berlin: Springer, 2001.
- Freund, Michael, “On the Revision of Preferences and Rational Inference Processes”, *Artificial Intelligence* 152, 2004, 105–137.
- Gärdenfors, Peter: *Knowledge in Flux. Modeling the Dynamics of Epistemic States*, Cambridge, Mass.: Bradford Books, MIT Press, 1988.
- Hansson, Sven Ove, “Hidden Structures of Belief,” in André Fuhrmann and Hans Rott, eds., *Logic, Action and Information*, Berlin: de Gruyter, 1995, pp. 79–100.
- , ed., “Non-Prioritized Belief Revision”, special issue of *Theoria* 63, 1997, 1–134.
- Katsuno, Hirofumi, and Alberto Mendelzon, “Propositional knowledge base revision and minimal change “, *Artificial Intelligence* 52, 1991, 263 – 294.
- Lehmann, Daniel, Menachem Magidor and Karl Schlechta, “Distance Semantics for Belief Revision”, *Journal of Symbolic Logic* 66, 2001, 295–317.
- Levi, Isaac. *Gambling With Truth: An Essay on Induction and the Aims of Science*. New York: Alfred Knopf, 1967.
- Levi, Isaac. *Mild Contraction: Evaluating Loss of Information due to Loss of Belief*. Oxford: Oxford University Press, 2004.
- Nayak, Abhaya: 1994, “Iterated Belief Change Based on Epistemic Entrenchment”, *Erkenntnis* 41, 353–390.
- Nayak, Abhaya, Maurice Pagnucco and Pavlos Peppas: 2003, “Dynamic Belief Revision Operators”, *Artificial Intelligence* 146, 193–228.
- Rott, Hans, “Coherence and Conservatism in the Dynamics of Belief. Part I: Finding the Right Framework”, *Erkenntnis* 50, 1999, 387–412.
- , “Two Dogmas of Belief Revision”, *Journal of Philosophy* 97, 2000, 503–522.
- , *Change, Choice and Inference: A Study of Belief Revision and Nonmonotonic Reasoning*, Oxford: Oxford University Press 2001.
- , “Coherence and Conservatism in the Dynamics of Belief. Part II: Iterated Belief Change without Dispositional Coherence”, *Journal of Logic and Computation* 13, 2003a, 111–145.
- , “Economics and Economy in the Theory of Belief Revision”, in: *Knowledge Contributors*, eds. Vincent F. Hendricks, Klaus F. Jørgensen and Stig A. Pedersen, Dordrecht: Kluwer 2003b, 57–86.
- , “A Counterexample to Six Fundamental Principles of Belief Formation”, *Synthese* 139, 2004, 225–240. (= *Knowledge, Rationality and Action* 1, 61–76.)
- , “The Value of Truth and the Value of Information: On Isaac Levi’s Epistemology”, *Knowledge and Inquiry: Essays on the Pragmatism of Isaac Levi*, ed. Erik J. Olsson, Cambridge: Cambridge University Press (to appear).