Erkenn (2009) 70:189–209
DOI 10.1007/s10670-008-9147-5

ORIGINAL ARTICLE

Iterated Belief Revision

Robert Stalnaker

Received: 30 August 2008 / Accepted: 17 September 2008 / Published online: 5 December 2008

Abstract This is a discussion of the problem of extending the basic AGM belief revision theory to iterated belief revision: the problem of formulating rules, not only for revising a basic belief state in response to potential new information, but also for revising one's revision rules in response to potential new information. The emphasis in the paper is on foundational questions about the nature of and motivation for various constraints, and about the methodology of the evaluation of putative counterexamples to proposed constraints. Some specific constraints that have been proposed are criticized. The paper emphasizes the importance of meta-information—information about one's sources of information—and argues that little of substance can be said about constraints on iterated belief revision at a level of abstraction that lacks the resources for explicit representation of meta-information.

1 Introduction

There is a relatively well-established formal theory of belief revision—the so-called AGM theory, named for Carlos Alchourrón, Peter Gärdenfors and David Makinson, who first proposed it in the 1980s. Agents' dispositions to revise their beliefs are represented, in this theory, by functions taking prior belief states into posterior belief states that are determined by an input proposition, the information that induces the revision. The theory states constraints on such a function, and the constraints that were proposed gave rise to a nice model-theoretic structure. The initial theory did not give any account of iterated belief revision, but it has been extended in various different ways, with constraints on the way that a posterior belief state might be further revised by a sequence of input propositions. The iterated revision theories are more varied, and more controversial. The constraints

R. Stalnaker (✉)
107 Upland Road, Cambridge, MA 02140, USA
e-mail: stal@mit.edu

proposed by different theorists are sometimes in conflict with one another, and they are (I will argue) less well motivated than at least the basic AGM postulates. So despite the existence of a number of very elegant proposals for solving the problem of iterated belief revision, I think the problem has not been solved. I will argue that we need to go back to the beginning and look more closely at the nature of the problem, and at the foundational assumptions about the phenomena that an iterated belief revision theory is trying to clarify. And I will argue that little of substance can be said about iterated belief revision if we remain at the level of abstraction at which such theorizing has been carried out. We need to distinguish different kinds of information, including meta-information about the agent's own beliefs and policies, and about the sources of his information and misinformation.

So I will start by going back to the beginning, setting up the problems of belief revision, and iterated belief revision. While my ultimate concern is with foundational issues, I will spend some time trying to get clear about the shape of the formal problem, and of the abstract structure of some of the solutions to it that have been proposed. I will then look critically at some of the postulates that have been proposed, at the general considerations that motivate them, and at some examples that raise problems for them. I will be concerned both with the assessment of the particular postulates, and with more general questions about the dialectic of critique and defense in such cases.

I will discuss the basic problems, and the various solutions, entirely in a model theoretic framework. The original AGM theory was given a syntactic formulation, and much of the subsequent discussion has followed this lead. Belief states were represented by sets of sentences, and belief changes were induced by sentences. But it was assumed from the beginning that the relevant sets were deductively closed, and that logically equivalent sentences played an equivalent role, so a model theoretic framework that represents belief states by sets of possible worlds, and information more generally by sets of possible worlds, can capture the same essential structure, and it was recognized early on that the AGM theory could be given a simple model-theoretic representation.[1] For several reasons, I think the model theoretic representation has conceptual as well as technical advantages, avoiding some potential for confusion that may arise from the syntactic formulation. First, the status of the language in terms of which information is represented in the syntactic formulation is unclear. It is not the language used by the subjects whose beliefs are being modeled (They do not generally speak in propositional calculus), though it is often treated as if it is. Second, the syntactic formulation brings a potential for use-mention confusion, and confusion of properties of the sentences with properties of what those sentences are being used to say.[2] Third, there are additional problems when the sentences in question are context-sensitive, expressing different propositions at different times or in different epistemic contexts. Conditional sentences, which are sometimes used to represent belief revision policies, and tensed sentences about what one believes at the present time are both

[1] See Grove (1988).

[2] I discuss some of the potential confusions, in the context of a discussion of theories of non-monotonic reasoning, in Stalnaker (1994).

obviously context-sensitive in these ways, and there is a potential for confusing the identity of a sentence, across contexts, with the identity of what it is being used to say.[3] Finally, a further reason to prefer the model-theoretic representation is that our concern is with knowledge, belief, and belief change, and not with speech, and we do not want to make implicit assumptions about the forms in which information is represented, or about the linguistic capacities of the subjects we are modeling. There may be rational agents who act, but do not speak, and more generally, there may be contents of knowledge and belief that the agent lacks the capacity to articulate. A general theory of belief revision should apply in these cases as well as to the case where what an agent knows corresponds to what he is in a position to say.

So in my formulation of the various theories, there will be no languages involved. Of course we use language to formulate our problems and theories, but so do geologists when they are talking about rocks. As with geologists, languages will form no part of the subject matter of our theory. The primitive elements of our basic structure are possible worlds, and the propositions that are the items of information that may be believed by an agent in a prior or posterior belief state will be sets of possible worlds. Propositions are thus identified with their truth conditions—with the distinction between the possibilities in which they are true and the possibilities in which they are false.[4]

One way to develop, in this model-theoretic context, a representation of knowledge and belief is to characterize propositions *about* the agent's knowledge and belief, as in Jaakko Hintikka's logics of knowledge and belief. This approach usually takes the form of a semantics for a language with knowledge and belief operators, but the language is dispensable: one can enrich one's model (the initial model being just a set of possible worlds) by giving it the structure necessary to define a function taking any proposition α (represented by a set of possible worlds) to the proposition that the agent knows α. The required structure is a Kripke structure, or frame: a set of possible worlds and a binary relation R on it, where 'xRy' says that possible world y is compatible with what the agent knows or believes in world x. But the AGM belief revision theory did not follow this path. In its model-theoretic version, a belief state is represented simply by a set of possible worlds—those compatible with what is believed. The theory is more abstract than the Hintikka-style theory in that there is no explicit representation of knowledge and belief about what the agent knows or believes. This is a significant difference, but it is important to recognize that this theoretical decision is not a decision to ignore information about the agent's knowledge and belief, or to restrict in any way the kind of information that might be relevant to inducing a belief change, or more generally in distinguishing between the

[3] I think some of the issues about updating vs. revision would be clearer if sentences were more clearly distinguished from what they are used to say.

[4] Any representation of cognitive states that uses this coarse-grained notion of proposition will be highly idealized. It will be assumed that agents know or believe all the logical consequences of their knowledge or beliefs. This is a familiar feature of all theories in this general ballpark, including probabilistic representations of degrees of belief. The syntactic formulations make an analogous idealization. There are various different ways of motivating the idealization, and of explaining the relation between theory and reality on this issue, but that is a problem for another occasion. Most will agree that the idealization has proved useful, despite its unrealistic character.

possible worlds. Rather, it is a decision to theorize at a level of abstraction at which nothing is said, one way or another, about the subject matter of the information. Meta-information about the agent's knowledge and belief is like information about politics. The theory lacks the resources to distinguish information about politics from information, for example, about astronomy, but of course information about politics or astronomy might be information that is compatible with an agent's beliefs, or that induces a change in belief. Similarly, the abstract AGM theory lacks the resources to distinguish meta-information from information about the physical environment, but nothing forecloses the possibility that information about informational states be relevant. We cannot generalize about the distinctive role of such information, if it has a distinctive role, but when we consider examples, where we describe a certain concrete scenario, and then construct a model to fit it, if the scenario is one in which meta-information is intuitively relevant, then we must include it in our model. (A corresponding point might be made about the syntactic formulation. The primitive sentence letters, in such a formulation, have no special status: they represent the general case. Any generalization that one makes should continue to apply if one enriches the language to include quantifiers, modal operators, epistemic operators, causal conditionals, or whatever.)

2 The AGM Belief Revision Theory

The basic belief revision problem is to model the result of the process of changing one's state of belief in response to the reception of a new piece of information. Assume that we have a set of possible worlds, B, representing a prior belief state, and a proposition α representing the new information. A belief revision function will be a function $B(\alpha)$ whose value will be the posterior belief state that would be induced by this new information. The problem is to specify the constraints on such a function that any rational belief revision must satisfy, and to motivate those constraints. Let me sketch the AGM theory's solution to this problem, as spelled out in the model-theoretic context, and then turn to the further problem of iterated belief revision.

To model a belief revision function, we start by specifying two sets, B and B*, the first a subset of the second, B represents the prior belief state, and B* is the set of possible worlds that are candidates to be compatible with some posterior belief state. An AGM belief revision function will be any function from propositions (subsets of B*) to posterior belief states (also subsets of B*) that meets the following four conditions for all propositions[5]: (I will use 'Λ' for the empty set of possible worlds.)

[5] The usual formulation of the AGM theory, in the syntactic context, has eight postulates. One of them (that logically equivalent input sentences have the same output) is unnecessary in the model-theoretic context, since logically equivalent propositions are identical. Analogues of my AGM1 and AGM4 are, in the usual formulation, each separated into two separate conditions. Finally, the first of the usual postulates is the requirement that the output of the revision function is a belief set, which is defined as a deductively closed set of sentences. This is also unnecessary in the model theoretic context. In the syntactic formulation, there is no analogue of our B*: it is assumed that every consistent and deductively closed set of sentences is an admissible belief set.

AGM1 $B(\alpha) \subseteq \alpha$
AGM2 If $B \cap \alpha \neq \Lambda$ then $B(\alpha) = B \cap \alpha$
AGM3 If $B^* \cap \alpha \neq \Lambda$, then $B(\alpha) \neq \Lambda$
AGM4 If $B(\alpha) \cap \beta \neq \Lambda$, then $B(\alpha \cap \beta) = B(\alpha) \cap \beta$

AGM4, as we will see, is less obvious and more problematic than the other postulates, so one might consider a more cautious theory, which I will call AGM− defined just by the first three postulates. One might also consider intermediate theories that are stronger than AGM−, but weaker than the full AGM.[6]

Any AGM function gives rise to a nice formal structure. It is equivalent to a formulation in terms of a binary ordering relation, or in terms of an ordinal ranking function on the possible worlds: worlds compatible with B get rank 0, the most plausible alternatives get 1, the next get 2, etc., with all worlds in B* getting some ordinal rank. The ranking of worlds determines a ranking of propositions: the rank of proposition α is the minimal rank of possible worlds within α. Any ranking function of this kind will determine an AGM revision function, and any AGM function will determine such a ranking function. In the ranking-function formulation, $B(\alpha)$ will be defined as$\{w \in \alpha: r(w) \leq r(\alpha)\}$.[7]

Why should these four postulates constrain any rational revision, no matter what the subject matter? To begin answering this, we need to say more about exactly what we are modeling. First, exactly what does it mean to believe, or accept a proposition? Second, what is the status of the argument of the function—the proposition that induces the change?

On the first question: I will assume that we are talking about a strong, unqualified doxastic state, although the postulates may be appropriate for various weaker notions of acceptance as well. Specifically, I will assume that to believe α is to take oneself to know it.[8] Belief in this sense differs from knowledge only in that believing α is compatible with α being in fact false, or a Gettier case of justified true belief. And of course believing α is (at least in some cases) compatible with later discovering that it is false, and so being forced to revise one's beliefs so that one believes the complement of α. Actual knowledge is in a sense unrevisable, since one cannot discover (come to know) that what one knows is false. But one may know something, while still having rational dispositions to respond to potential

[6] Hans Rott discusses such theories in Rott (2001).

[7] It should be emphasized that ranking functions of the kind I have defined are not the same as Wolfgang Spohn's ranking functions. Or more precisely, they are a special case of Spohn ranking functions. Spohn's ranking functions are richer structures than AGM ranking functions, since they allow the values of the function to be any ascending sequence of non-negative integers. In a Spohn function, it might happen that there are gaps in the ranking (a positive integer k such that some worlds have ranks greater than k, and some less, but none with rank k), and the gaps are given representational significance when the theory is extended to an iterated revision theory. But any Spohn ranking function determines a unique AGM structure, and any AGM structure determines a unique Spohn ranking function with no gaps. See Spohn (1988).

[8] By "taking oneself to know" I do not intend a reflective state of believing that one knows, but just a cognitive state that is like knowledge in its consequences for action. I also think that in the idealized context of belief revision theory, it is appropriate to make the kind of transparency assumptions relative to which taking oneself to know, in the relevant sense, entails believing that one knows, but that is a controversial question that I need not commit myself to here.

information that conflicted with one's knowledge. Suppose I know who won the gold medal in a certain Olympic event—Michael Phelps. Still, my system of beliefs would not collapse into contradiction if, contrary to fact, I discovered that I was mistaken. In such a counterfactual situation, I would revise my beliefs to accommodate the belief-contravening evidence, with exactly how I revise depending on the details of that evidence. Of course my rational dispositions to revise in response to contrary evidence won't in fact be exercised, since he did win the gold in the event in question, and I know that he did.[9]

On the second question: I will assume that the input proposition represents an item of information that the subject takes himself to have come to know (in a situation in which the disposition to revise is exercised), and that it is the total relevant information received. That is, the rational disposition being modeled is the disposition to shift to the posterior belief state upon coming to know (or taking oneself to have come to know) the input proposition, and no stronger new piece of information. The total evidence assumption is of course essential to the application of the theory, since any nontrivial belief revision function that permits revision by information that is incompatible with prior beliefs must be nonmonotonic. One may come to believe something on receiving information α that one would not come to believe on receiving information β that entails α.

Now let's look at the four AGM postulates in the light of these answers to our two questions. AGM1 is unassailable, given our answer to the second question. To question it is to reject the intended application of the theory. AGM2 I take to be implied by the answer to the first question, together with a proper account of what it is to know, or fully believe a proposition. The crucial assumption that I want to make here is that to fully accept something (to treat it as knowledge) is to accord it this privileged status: to continue accepting it unless evidence forces one to give up something. This is a relatively weak commitment: it says only that if one accepts α, then one will continue accepting α provided the evidence one receives is compatible with *everything* that one accepts. But it is a commitment that distinguishes full acceptance from any degree of belief, no matter how high, that is short of probability one.

AGM3 also seems unassailable, given the answers to the second question, and given our assumption about what the set B* represents. B*, by definition, contains all the possible situations that are compatible with any proposition that one might conceivably come to accept, and coming to accept such a proposition means coming to be in a coherent belief state.

The intention, in the attempts to justify the first three AGM postulates, has been to show that the postulates are constitutive of the concepts involved in the application of the theory. The claim is that these principles do not give substantive guidance to inductive reasoning, but are implicit in the relevant concept of

[9] So I am assuming a different conception of knowledge from that assumed by, for example, in Friedman and Halpern (1999). They assume that to take observations to be *knowledge* is to take them to be unrevisable. In terms of our notation, they are, in effect, identifying knowledge with what is true in all possible worlds in B*. Friedman and Halpern's way of understanding knowledge is common in the computer science literature, but I think it is a concept of knowledge that distorts epistemological issues. See Stalnaker (2006) for a discussion of some of the questions about logics of knowledge and belief.

acceptance, and the intended application of the theory. But I don't see how to give AGM4 the kind of justification that we can give for the other postulates. (One temptation is to justify this postulate in virtue of the nice formal structure that it determines, but this is a different kind of justification, and I think one that we should resist.) I will defer discussion of the motivation for AGM4 until we consider the iteration problem, since I think that if it can be justified at all, it will be in that context. So let me now turn to that problem.

3 The Iteration Problem

We started with simple belief states, but the belief state itself, as represented formally, was not a rich enough structure to determine how it should evolve in response to new evidence. So we added new structure, which yielded a posterior belief state for any given input proposition. One might put the point by saying that a full representation of an agent's cognitive situation should be represented, not by a simple belief state, but by a belief revision function. But the output of an AGM function is just a simple belief state, and so it does not yield a full representation of the agent's posterior cognitive situation. We need to enrich the theory further so that the output of our revision function is a new revision function. But of course if we enrich the representation of the cognitive situation further, then we need to be sure that we also get a structure of the enriched kind as our output. We need to satisfy a principle that Gärdenfors and Rott have called "the principle of categorical matching"[10]: roughly, that our representation of a cognitive situation must be a kind of structure such that given a proposition as input, yields a structure of the same kind as output.

Before looking at some particular proposals for solving this problem, let me describe a general form that any solution can take, and introduce some terminology for talking about the solutions.[11] Any solution to the iteration problem will yield a function taking finite sequences of propositions to belief states.[12] Call any such function a *belief system*. The idea is that the proposition that is the first term of the sequence induces a certain revision, and then the next term induces a revision of the belief state that results from the first revision, etc. The initial prior belief state of a belief system will be the value of the function for the empty sequence. A belief system, restricted to inputs that are one-term sequences will be a revision function (a function from propositions to posterior belief states). Say that a revision function R is *generated* by a belief system Ψ iff for some sequence of propositions, $\beta_1, \ldots \beta_n$, and for all propositions α, $R(\alpha) = \Psi(\beta_1, \ldots \beta_n, \alpha)$. Say that a belief system Ψ is an

[10] Gärdenfors and Rott (1995, p. 37).

[11] The general form I will describe was used by Lehmann (1995). A related formalism is used in Rott (1999).

[12] Recall that a belief state consists of a *pair* of sets of possible worlds, B and B*, the first being a subset of the second. In the general case of a belief system, it will be important to consider the case where B* as well as B may take different values for different arguments of the belief system. But to simplify the discussion, I will assume for now that for any given belief system there is a fixed B* for that system.

AGM belief system iff every revision function that it generates is an AGM function, and similarly for AGM− belief systems.[13]

The weakest and most unconstrained 'solution' to the iteration problem would be simply to note that an AGM belief system yields a belief state, and an AGM revision function for any input. (And any AGM− belief system yields an AGM− revision function for any input.), so we can just say that an agent's full cognitive situation is to be represented by such a belief system. But one would like to find some more interesting and revealing structure that determines such a system, and some constraints on the way that belief states may evolve.

The most constrained kind of solution would be to lay down postulates governing belief systems that are strong enough so that there will be a unique full belief system determined by any initial AGM belief revision function. In 1996, Craig Boutilier proposed a theory that meets this condition (Boutilier 1996). A more flexible approach would be to state and defend a set of postulates that constrains the evolution of a belief system, but that allow for alternative evolutions from the same initial revision function. I will describe Boutilier's system, and a classic proposal of the more flexible kind (by A. Darwiche and J. Pearl) before looking at the ways that such proposals are motivated and evaluated.

Boutilier's proposal states a rule for taking any prior AGM revision function and any (consistent) input proposition to a new AGM function. The idea can be put simply, in terms of the ranking-function representation of an AGM revision function. Let $r(w)$ be the rank of possible world w according to the prior AGM ranking function, and let $r_\alpha(w)$ be the rank of w in the posterior AGM ranking function induced by information α. The Boutilier rule (with one minor simplification) is this: $r_\alpha(w) = 0$ for all $w \in B(\alpha)$, and $r_\alpha(w) = r(w) + 1$ for all $w \notin B(\alpha)$.[14] Intuitively, the idea is to move the posterior state into the center of the nested spheres, but to leave the ordering of all the other worlds the same.

How might such a solution be justified? It is, in one clear sense, the uniquely minimal revision of an AGM revision function, and so it might be thought to be a natural extension of the minimality assumption that is implicit in postulate AGM2. But we suggested that the justification of AGM2 was not some general methodological principle of minimality, but rather an assumption about the kind of commitment that is constitutive of accepting, or fully believing a proposition. According to this way of understanding it, AGM2 does not in any way constrain a rational agent's response to an event that one might be tempted to describe as the

[13] As we defined AGM revision functions, the input proposition could be an impossible proposition (the empty set). In this vacuous limiting case, the postulates imply that the output is also the empty set. This was harmless in the simple theory. The empty set is not really a belief state, but it doesn't hurt to call it one for technical purposes. But we need to do some minor cleaning up when we turn to the extension to a full belief system. One simple stipulation would be to require that in the empty 'belief state', the B* is also empty. Alternatively, one might restrict the input sequences to sequences of nonempty propositions, or in the general case where B* may change with changes in the belief state, to sequences of nonempty subsets of the relevant B*.

[14] A pedantic qualification is needed to get this exactly right. In the case where $B(\alpha)$ is the set of *all* possible worlds of a certain rank greater than 0, the Boutilier rule, as stated, will result in a ranking function with a gap. If a simple AGM function is represented by a ranking function with no gaps, one needs to add that after applying the rule, gaps should be closed.

receipt of a certain piece of information. On the intended application of the belief revision theory, an event is *correctly* described as the receipt of the information that α only if the event is one in which the agent fully accepts α, and so undertakes the commitment. An event that for one agent, or in one context, constitutes receiving the information that α might not be correctly described in that way for another agent, or in another context. One might, for example, respond to an observation that it seems to be that α, or a report by a witness that α, with a commitment short of full belief, perhaps concluding that α has a very high probability, but less than one. AGM2 does not say that one should or should not respond in this way: it says only that *if* one responds by fully accepting the proposition, then this constitutes taking on a commitment to continue accepting it until one is forced to give something up.

This kind of justification for a postulate, as a constraint on any rational revision, does not extend to any kind of minimality assumption in the iterated case. It is not constitutive of having the epistemic priorities represented by a belief revision function that one is disposed to retain those priorities in response to surprising new information.

But one might be satisfied, in one's search for a solution to the iteration problem, with a more relaxed kind of justification. A demanding justification of the kind that I suggested could be given for the postulates of AGM− requires an argument that given what it means to fully believe something, and given the assumption that the revision function applies only when the input proposition is fully believed, it would be irrational to violate the postulates. Or perhaps it would be better to say that a successful justification of this kind would show that if you apparently violated the postulate in question, that would show that the conditions for the application of the theory were not met. (You didn't really fully accept either the input proposition or one of the propositions that characterize the prior belief state.) But even if one cannot give a justification that meets this demanding standard for the postulates of an iterated belief revision theory, one might argue that a weaker kind of justification would suffice. According to this line of thought, there may be very general and abstract substantive methodological principles that govern the appropriate formation and change of belief, and the project of developing belief revision theory is to find principles of this kind that seem to be descriptively adequate, and that seem to illuminate common sense reasoning and scientific practice. If one thinks of the project this way, than one should assess a proposed set of postulates by looking at its consequences for examples. If the consequences seem intuitively right, that is reason enough to accept them. This is a reasonable aim, though I think it is important to distinguish the more stringent kind of justification from the more relaxed kind.

Before considering this kind of evaluation of solutions to the iteration problem, I will sketch an influential proposal of the more flexible kind: a set of constraints that permits alternative evolutions of a given initial AGM revision function. The iteration postulates proposed in Darwiche and Pearl (1997) give rise to a nice formal structure; they allow Boutilier's theory as a special case, but avoid some of the problems that have been raised for that more constrained proposal.

Darwiche and Pearl propose four postulates to extend the AGM theory. I will state them in terms of the framework and terminology introduced above. A DP

belief system Ψ is an AGM belief system that also meets the following four conditions for any proposition $\beta_1, \ldots \beta_n$, α and ϕ:

(C1) If $\alpha \subseteq \phi$, then $\Psi(\beta_1 \ldots \beta_n, \phi, \alpha) = \Psi(\beta_1, \ldots \beta_n, \alpha)$

(C2) If $\alpha \subseteq \neg\phi$, then $\Psi(\beta_1 \ldots \beta_n, \phi, \alpha) = \Psi(\beta_1, \ldots \beta_n, \alpha)$

(C3) If $\Psi(\beta_1, \ldots \beta_n, \alpha) \subseteq \phi$, then $\Psi(\beta_1, \ldots \beta_n, \phi, \alpha) \subseteq \phi$.

(C4) If $\Psi(\beta_1, \ldots \beta_n, \alpha) \not\subset \neg\phi$, then $\Psi(\beta_1, \ldots \beta_n, \phi, \alpha) \not\subset \neg\phi$.

Think of an AGM belief system as a procedure for redefining the ranks of the possible worlds, at each stage of the iterated process, in response to the input proposition.[15] If $r(w)$ is the prior rank of world w, and $r_\alpha(w)$ is the posterior ranking induced by proposition α, then (C1) is equivalent to assuming that for all $w,w' \in \alpha$, $r(w) > r(w')$ iff $r_\alpha(w) > r_\alpha(w)'$. (C2) is equivalent to assuming that for all $w,w' \notin \alpha$, $r(w) > r(w')$ iff $r_\alpha(w) > r_\alpha(w')$. (C3) is equivalent to assuming that for all $w \in \alpha$ and $w' \notin \alpha$, if $r(w) < r(w')$, then $r_\alpha(w) < r_\alpha(w')$. (C4) is equivalent to assuming that for all $w \in \alpha$ and $w' \notin \alpha$, if $r(w) \leq r(w')$, then $r_\alpha(w) \leq r_\alpha(w')$. The idea is that in any belief revision induced by proposition α, all the α-worlds move down or stay the same in the ranking (with the minimal α-worlds taking posterior rank 0, and the relative order, within α, remaining the same); the non-α-worlds all move up in the ranking or stay the same (with no non-α-worlds remaining at rank 0, and the relative ranking within α remaining the same). Given these rules, the initial AGM ranking constrains the new ranking, but does not determine it, since it does not in all cases determine the relative ranking of α and non-α worlds.

Spohn's ranking function theory provides an iterated revision rule that conforms to the DP conditions, but that is more constrained. Spohn's iterated revision rule is a reductionist account in the sense that the posterior ranking function is fully determined by the prior ranking function, plus the input. But as mentioned in footnote 7, a Spohn ranking function is richer than an AGM ranking function, since gaps in the rankings are allowed, and the gaps are intended to provide information about the entrenchment of beliefs that goes beyond ordering information. In addition, in the Spohn theory, the inputs to the revision function contain more information than is provided just by specifying the input proposition. The input will be a pair consisting of a proposition and a positive integer, where the number is intended as a measure of the degree of robustness of the new information. Any new information will be treated as something fully believed, in the posterior state, but higher levels of robustness will mean that the new information becomes more fully entrenched in the posterior belief revision structure. The Spohn rule is this (using 'r' for the prior ranking function, and $r_{\langle\alpha,k\rangle}$ for the posterior ranking induced by new information α, with robustness measure k)[16]:

For all $w \in \alpha, r_{\langle\alpha,k\rangle}(w) = r(w) - r(\alpha)$.

For all $w \notin \alpha, r_{\langle\alpha,k\rangle}(w) = r(w) + k$.

[15] The ranking function representation is just a notational convenience. The theory, and all the DP constraints, could be stated without using numbers.

[16] This rule holds on the assumption that α is a proposition that is not believed in the prior state. The revision rule becomes a contraction rule when the value of k is 0.

Before moving on to look at concrete examples, let me mention one technical consequence of postulate (C1): it entails the fourth postulate of the simple AGM theory, the postulate that we said was more difficult to defend than the other three AGM postulates. More precisely, the result is that any AGM− belief system that satisfies postulate (C1) is an AGM− belief system.[17] So if the iterated revision postulate (C1) can be defended, that will suffice for a defense of AGM4.

These are beautiful theories, providing nice formal structures, and the commentary on them provides intuitive content to the functions and relations that are specified in the structures. But are the constraints that give rise to these nice structures defensible? I will consider some examples that seem to raise problems for several of the postulates, but the task of confronting an abstract theory with concrete examples is a subtle and difficult one, and it is often questionable whether a putative counterexample is really relevant to the abstract principle to which it is applied. So before looking at the details, I will digress to make some general comments about this methodological problem.

4 Evaluating Counterexamples

Let me start with a caricature of a fallacious use of an example to refute a principle: I see a man carrying a sign reading "THE WORLD WILL END TOMORROW" (the kind of event that takes place mostly in *New Yorker* cartoons). So I receive the information that the world will end tomorrow, but I don't revise my beliefs by coming to fully believe that the world will end tomorrow. I think the man is a nut, so my beliefs about the future of the world are unaffected by the message on his sign. Therefore, AGM1 (which says that the input proposition is fully believed in the posterior belief state) is not in general true.

Mathias Hild, (1998) discussed examples something like this, and he seemed to be using them as counterexamples to a principle analogous to AGM1. (Hild's first example was of testimony from a person that the subject believes is trying to be misleading, and so she not only fails to believe the statement made, but takes the fact that it was made as decisive evidence that it is false.) I am not sure what Hild had in mind with these examples, but he seemed to be suggesting that we should characterize the evidence, observation or information that induces a belief change independently of its effect on the cognitive situation of the subject whose beliefs are changed. But I am not sure how to do this, in general, and I don't think it is a good idea in any case, to try. It is not the job of belief revision theory to say what a person

[17] The proof is simple: By AGM2 (and the assumption that Ψ is an AGM− system),
1. If $\Psi(\beta_1, \ldots \beta_n,\phi)\cap\alpha\cap\phi \neq \Lambda$, $\Psi(\beta_1, \ldots \beta_n,\phi,\alpha\cap\phi) = \Psi(\beta_1, \ldots \beta_n,\phi)\cap\alpha\cap\phi$
But by AGM1, $\Psi(\beta_1, \ldots \beta_n,\phi)\cap\alpha\cap\phi = \Psi(\beta_1, \ldots \beta_n,\phi)\cap\alpha$, so it follows from 1 that
2. If $\Psi(\beta_1, \ldots \beta_n,\phi)\cap\alpha \neq \Lambda$, $\Psi(\beta_1, \ldots \beta_n,\phi,\alpha\cap\phi) = \Psi(\beta_1, \ldots \beta_n,\phi)\cap\alpha$
By (C1),
3. $\Psi(\beta_1, \ldots \beta_n,\phi,\alpha\cap\phi) = \Psi(\beta_1, \ldots \beta_n,\alpha\cap\phi)$
Therefore, by 2 and 3,
4. If $\Psi(\beta_1, \ldots \beta_n,\phi)\cap\alpha \neq \Lambda$, $\Psi(\beta_1, \ldots \beta_n,\alpha\cap\phi) = \Psi(\beta_1, \ldots \beta_n,\phi)\cap\alpha$
But 4 is just the claim that AGM4 holds in all revision functions generated by Ψ.

should come to believe on having certain kinds of experiences. The task is to say how new beliefs interact with a prior belief state to determine a posterior state.

Hild says "receiving total evidence *B*, as I understand it, means to be presented (precisely) with the information that *B* has occurred, or to receive (precisely) the input that *B*."[18] But this is subject to different interpretations. Perhaps in a sense, when someone tells me (in speech, or on a sign) that P, I am presented with the information (or misinformation) that P. The statement is surely *evidence*, but a statement that P is not (necessarily) evidence that P—it is evidence that someone said that P.

What is meant by '*evidence*'? Perhaps in one sense, evidence may consist of things such as hair samples, pieces of a broken vase, a gun with fingerprints on it, that one might label and put in plastic bags until the trial. Such things *carry* information, and that is why they might be entered into evidence, but they are not themselves items of information, and that is what an input to a belief revision function must be. Timothy Williamson argues persuasively (2001) that evidence should be identified with knowledge. An input proposition need not become known, in the posterior belief state that it induces (and so need not be evidence, according to Williamson's thesis), since rational agents may make mistakes. But we are assuming that any input proposition is (in the induced posterior belief state) taken by the subject to be knowledge, or at least that it becomes fully believed. This is not a claim made by the theory, but a constraint on its application.

But there are more subtle cases. Suppose I am disposed to take the word of either of two potential informants, Alice and Bert (at least on certain subjects). If either Alice or Bert were to tell me that the Red Sox won, I would take myself to have learned that they did. Both Alice and Bert are a lot better informed than I am about current events, sports, etc., so even in some cases where I have a prior opinion, (say I believe that the US won more medals than China in the 2008 Olympics), I am disposed to change that belief if Alice or Bert were to tell me that I was wrong. Against this background, suppose that Alice tells me that the Red Sox won the game, so I revise by coming to believe that they did. But a little later, Bert comes and tells me that they lost, so I revise again, coming to believe that they lost. As Darwiche and Pearl say, "when two contradictory pieces of evidence arrive, the last one prevails."[19] But why should this be? It was only an accident that Alice arrived first. If she had been slightly delayed, so that I heard first from Bert, then my settled opinion, in the end, would have been the opposite. Surely it is unreasonable to require that any rational agent have revision policies that are constrained in this way. (Though the example is a case of iterated belief revision, the assumption being challenged is, again, AGM1.)

But this argument is also fallacious. The AGM principles do not say that we must give the last witness priority over earlier ones; they permit the reasonable response, which, in most case fitting this description, would be treat the witnesses symmetrically. The assumption was only that *in the prior state*, I was disposed to take Alice's word, or Bert's, about the game. The principles imply only that if, after receiving Bert's report, I am *still* disposed to take Alice's word on the outcome of the game, then I should give Alice the last word. To have this belief revision policy

[18] Hild (1998, p. 334).

[19] Darwiche and Pearl (1997, p. 11).

is to give Alice priority over Bert, which might be perfectly reasonable, in some cases. But in such a case, I would probably reach the same conclusion, in the end, even if Alice had come first. For in that case, in the intermediate belief state (after receiving Alice's report, but before receiving Bert's), I will no longer be disposed to take Bert's word for it in cases where he contradicts what I learned from Alice.

If the epistemic status of the reporters is symmetrical, as suggested in the example, then I will most likely take neither to be reliable when contradicted by the other, so my end state will be to suspend judgment. Nevertheless, it is *consistent* for me to have the policy of letting the last reporter decide my belief, whichever one comes last. Such a policy might, in some contrived circumstances, be a reasonable one. But the postulates do not require such a policy, and that is enough to defeat the counterexample. The general moral of the story is that a concrete evidential situation (being told by Alice that P) may count as receiving information that P in one context, even if the very same situation would not be correctly described in that way in another context.

For a piece of information to be the input that induces a belief change in a proper application of the theory, it must not only be a proposition that the subject come fully to believe, it must also be the *total* evidence—the strongest proposition that is the new information received. This condition on the application of the theory to a concrete case raises a serious general problem, since in any concrete situation where one receives evidence, one receives a lot of extraneous information along with the information that is in focus. When Alice told me that the Red Sox won, she said it in English, slowly, with a slight southern accent. She was wearing blue jeans and a grey tee shirt with a picture on it of Albert Einstein sticking out his tongue. Strictly speaking, total evidence should include all this information, everything that I take in while receiving Alice's report. We can't be expected to include all of this information in our models of concrete situations, but it will suffice to include all of the *relevant* information that is received. The problem is that it may be controversial whether some evidence is relevant, and one might reject a putative counterexample on the ground that it left out some relevant evidence. In response, the critic might vary the example, and show that the result claimed for the example holds across a variety of cases, where the information claimed to be irrelevant varies. The example will be unconvincing if the intuitions about the concrete case seem to be relying on information that does not get included in the model.

A further problem in the application of the abstract theory to concrete cases, concerns timing. Evidence comes in continuously, and takes time to process. In some cases, one might represent a situation either as a single belief revision, made after some complex new information is taken in and appreciated, or as a sequence of belief revisions as one takes in part of the information, revises, and then takes in more of it. It may matter, for the evaluation of iteration principles, which way one represents the case, but the choice may seem arbitrary. It would be a point against a proposed iteration principle if it had different consequences for different formal representations that seem equally faithful to the concrete case in question.[20]

[20] Thanks to Hans Rott for raising this issue in his comments on a version of this paper. The issue deserves more discussion.

In light of all these considerations, let me examine in some detail[21] an example that Hans Rott gave as a counterexample to a range of AGM principles.[22] Rott's example is a case where the information in focus concerns an appointment to be made in a philosophy department. The position is a metaphysics position, and there are many candidates, but three in particular who are of interest: Andrews, Becker and Cortez. Andrews is clearly the best metaphysician of the three, but is weak in logic. Becker is a very good metaphysician, also good in logic. Cortez is a brilliant logician, but relatively weak in metaphysics. Now we consider two possible scenarios (these are alternatives, not successive scenarios). In scenario one, our subject Paul is told by the dean, that the chosen candidate is either Andrews or Becker. Since Andrews is clearly the better metaphysician of the two, Paul concludes that the winning candidate will be Andrews. But in alternative scenario two, the Dean tells Paul that the chosen candidate is either Andrews, Becker or Cortez. "This piece of information sets off a rather subtle line of reasoning. Knowing that Cortez is a splendid logician, but that he can hardly be called a metaphysician, Paul comes to realize that his background assumption that expertise in the field advertised is the decisive criterion for the appointment cannot be upheld. Apparently, competence in logic is regarded as a considerable asset by the selection committee."[23] But while Paul concludes that logic is being given more weight than he initially thought it would be given, he still thinks that Cortez will not win out in the end, given his weakness in the advertised field, so he concludes that the winning candidate will be Becker.

Now if we assume that the total relevant information received in scenarios one is (A or B), and the total relevant information received in scenario two is (A or B or C), then we have a violation of AGM4 (as well as of some weaker principles that are implied by AGM4, but not by AGM1, 2 and 3 together). A natural response to the example is to say that some relevant evidence has been left out in the model of what is learned in scenario two. Strictly speaking, what Paul learned was that the Dean told him that the winning candidate will be either Andrews, Becker or Cortez. Given Paul's background belief that the Dean speaks the truth, he thereby learns that either Andrews, Becker or Cortez will be the winning candidate, but since he also learns that this is what the dean said, he learns something that he does not learn (because it is not true) in scenario one. So properly modeled, the example is not a counterexample to any of the principles.

Rott considers this objection to his counterexample. Here is how he puts the objection: "When the dean says that either Andrews or Becker or Cortez will be offered the job, isn't she in some sense saying *more* than when she says that either Andrews or Becker will be the winner of the competition? Namely that it is *possible* that Cortez will be offered the position, while the latter message, at least implicitly,

[21] Though not in as much detail as Rott provides about the example. One should consult his discussion in Rott (2004). He includes one additional scenario, since he aims to raise problems for six different principles, as well as more detail to the example. But I think my sketch of the example includes everything relevant to the issue I am raising about it.

[22] Most of the principles Rott considers are weaker than AGM4, but entailed by the full set of AGM postulates.

[23] Rott (2004, p. 230).

excludes this possibility." I don't think this is quite the right way to put the point. It is not that the dean *said* anything more than the winning candidate was one of those three. What is important is not what was said, but what total information Paul receives in the encounter. The point is that Paul learned that the dean said this, and he did not learn that in scenario one. The issue is whether this extra information is relevant. It seems to me that it is clear from the way Rott tells the story that it is taken by Paul to be relevant, since it is part of the story that Paul drew a conclusion from the fact that the dean included the Cortez possibility in his disjunctive statement. In scenario two, Paul concluded that Cortez was being taken seriously by the selection committee, and it is clear that the basis for this conclusion was a fact about what the Dean did and did not say.

Rott goes on to say that "nothing in the story commits us to the view that either the dean or Paul actually believes that Cortez stands a chance of being offered the position."[24] This is right, but the story, as Rott tells us, does commit us to the view that Paul believes (in scenario two) that Cortez is being taken seriously by the selection committee, and this is a conclusion he reaches only from the information received in scenario two, but not in scenario one. It is also clear from the story Rott tells that this conclusion (that Cortez is being taken seriously by the selection committee) plays an essential role in Paul's further inference that Becker will be the selected candidate. So this extra evidence is clearly relevant evidence.

Now this extra relevant evidence is meta-information—information that Paul acquires about what he learns, and how he learns it. As it is sometimes described, it is auto-epistemic information. Rott says, "As is common in the literature on belief formation, we presuppose in the paper that our language does not include the means to express autoepistemic possibility, something like $\Diamond c$ (read as 'for all I believe, c is possible'."[25] But the fact that the language does not include the resources to express an autoepistemic possibility operator does not mean that information about autoepistemic possibilities is not information that must be taken into account, if it is relevant. The simple language of belief revision theory also does not have resources to express counterfactuals, as a function of antecedent and consequent, or propositions about what other people know. But the primitive sentence letters (in a syntactic representation of a belief revision theory) are unrestricted. The information they are interpreted to express could be anything, including information about how I learn some of the things I learn, about the sources of my information, or about what I believe about what I believe and don't believe. If the story we tell in an example makes certain information about any of these things relevant, then it needs to be included in a proper model of the story, if it is to play the right role in the evaluation of the abstract principles of the model.

In the end, I am not really disagreeing with Rott on the correct diagnosis of the example; if we disagree, it is about exactly what the example shows about the confrontation of abstract theory with realistic phenomena. As I read Rott's conclusion, he is using the example, as I am, to bring out the importance of meta-information. In drawing the moral of his story, he says "What Paul really learns, in

[24] Ibid., p. 233.

[25] Ibid., p. 233.

scenario 2 is not just that either Andrews, Becker or Cortez will get the job, but that the dean says so.… The fact that a piece of information comes from a certain origin or source or is transmitted by a certain medium conveys information of its own."[26] But Rott seems to take the point about meta-information to explain why the example conflicts with the theoretical principles, whereas I want to conclude that it shows why the example does *not* conflict with the theoretical principles, since I take the relevance of the meta-information to show that the conditions for applying the principles in question are not met by the example. Rott says, "Once we make logic more realistic in the sense that it captures patterns of everyday reasoning, there is no easy way of saving any of the beautiful properties that have endeared classical logic to students of the subject from Frege on." Perhaps there is no *easy* way, but I think proper attention to the relation between concrete examples and the abstract models will allow us to reconcile some of the beautiful properties with the complexity of concrete reasoning. Rott's particular example shows that we need to take account of a richer body of information than is done in a simple model, but it does not threaten the principles of AGM theory, which should continue to apply, even when one enriches the representational resources of the language for characterizing the evidential situation. But I do agree with Rott's main point in this paper, that there are tradeoffs between the simplicity and precision that can be achieved at a higher level of abstraction, and the realism that requires a more complex immersion in the messy details of a case.

5 Evaluating Iteration Postulates

So examples need to be handled with care, and the defender of an abstract principle against a putative counterexample has various moves to make to explain it away. That is why neither the defense nor the refutation of abstract principles is easy. Still, we do need to confront our theories with their consequences for the phenomena. I will consider some examples that I don't think can be explained away as easily as the examples we have discussed so far.

First, let me take a quick look, in the light of our methodological worries about examples, at a case used by Darwiche and Pearl to criticize Boutilier's iterated revision rule. In this case, I think this example stands up to scrutiny.

First we judge that a strange new animal that we observe is a bird. Second, when we come closer we see that it is red. But then an expert informs us that the animal is in fact not a bird. Should we take back our judgment that it is red (suspending judgment on that question? Boutilier's rule implies that we must, but this is unreasonable. The source of our information that the animal was red, in this case, was independent of the earlier judgment that the animal was a bird, and so should not have been affected by the overturning of this earlier judgment. If we had observed the color first, and then later reached the conclusion that the animal was a bird, then the earlier judgment would be maintained (by the Boutilier rule). But the order in which the independent pieces of evidence are received seems intuitively to be irrelevant.

[26] Ibid., p. 237.

Now we added some detail to the story, beyond the fact that certain pieces of information are received in a certain sequence, namely the meta-information that the judgments were based on independent evidence. But in this case, I don't think the added information compromises the counterexample. Even if it is added (at stage two, when it is observed that the animal is red) that I also come to believe that my observation of the color is in no way dependent on the assumption that the animal is a bird, the theory will still require that, when the belief that the animal is a bird is overturned, I must give up my belief about the animal's color. So this is a robust example that succeeds in bringing out a flaw in the Boutilier rule.

I think there are similar problems with the more flexible DP rules. The clearest problem is with (C2), but let me start with an example that has been used against (C1) (and also against AGM4).[27]

As with Rott's example, discussed above, there will be two alternative scenarios. Here is the first: there are three switches wired in such a way that the light they control is on if and only if either all three switches are up, or all three switches are down. I don't know whether the light is on or off, but I get reports from three independent observers who I take to be completely reliable. Alice tells me that switch A is up, Bert tells me that switch B is down, and Carla tells me that switch C is up, so this is what I believe. But now I learn that the light is on. I could be wrong about the wiring arrangement, but let us suppose that my belief about that is more robust than my belief about the reliability of my informants. So I conclude that either Bert is wrong, or else both Alice and Carla are wrong. A bold believer might reason as follows: the hypothesis of two independent errors is less plausible than the hypothesis of one error, so I should conclude that all three switches are up. But it seems too strong to force this conclusion on our subject. A more cautious believer might grant that this possibility is more likely, but refrain from reaching a definite conclusion. This more cautious policy does not seem unreasonable.

But now let us add to the story that there is also a secondary light that is controlled by just two of the switches, A and B. This light is on if and only if either both of these switches are up, or both are down. So if the main light is on, the secondary light will be on as well, but the secondary light might be on even if the main one is not. Suppose that I knew all along about the secondary light and the way it is wired, but this does not affect the story, as told so far, since I have no prior information about whether the secondary light is on or off. It also should not affect the story, so far, if we add that the information I receive, at the last stage, is that *both* lights are on.

Now scenario two: this time, after receiving the three independent reports, but before learning anything about state of the primary light, I learn that the secondary light is on. This new information is not relevant to Carla's report about switch C, so

[27] An example with the structure of the example I will describe was originally given, in the context of a discussion of counterfactuals, in Ginsberg (1986). Ginsberg's story was a variation on an old example used by Quine for a different purpose. Ginsberg's example is discussed in Stalnaker (1994) and in Rott (2001, p. 188ff).

The example raises a problem for the basic AGM theory, and not just for the iteration rule, but as I suggested above, I think the postulate AGM4 is really an iteration principle, and should stand or fall with the DP rule, (C1).

my belief based on Carla's report (that switch C is up) should stand. (To give it up in response to this new information would be just like giving up my belief that the animal was red, in DP's example against Boutilier.) So I conclude that either Alice is wrong, or Bert is: either A and B are down, and C up, or else all three are up, but I don't know which. After this revision, I receive the information that the primary light is also on. Since this information is compatible with my prior belief state, all I need is AGM2 to tell me that I should conclude that all three switches are up, and in any case this seems reasonable.

In scenario two, I receive the same information as in scenario one, but in two stages: first I receive a part of the information, and then all of it. (C1) requires that I reach the same final belief state in these two scenarios, but the example suggests that in the two-stage scenario the subject should reach a more decisive conclusion. But does this example stand up to scrutiny? I find it genuinely puzzling, since I don't see that any intuitively relevant information is being excluded or misdescribed in the model. The example seems robust, in that it is easy to vary the details of the concrete story in a way that preserves the basic structure, and the variations seem equally persuasive. So the example is not easily dismissed, but on the other hand, the principle seems well motivated. As in the case discussed above in which the order in which Alice and Bert arrived with their reports seemed irrelevant, it seems that it should not make a difference if I receive a piece of information in two stages, rather than one. Of course it might make a *psychological* difference, but the issue is whether rational methodology puts different normative constraints on belief revision in the two cases. And of course it is always possible that the meta-information that I receive the information in two stages is relevant information. (My background knowledge might be such that the fact that some informant bothered to give me a partial report tells me something that is relevant.) But that does not seem to be what is going on in this example. If there is a rebuttal to this kind of example, I don't know what it is. Nor do I have any explanation for why difference between the one and two stage scenarios in this kind of case might be epistemically relevant.

Just to add to the puzzlement: there is a stronger version of (C1), implicitly rejected by DP, but proposed as one of a different set of revision postulates by Daniel Lehmann.[28] In the belief system notation, this is the principle:

$$\begin{aligned}(\text{C1}+)\quad &\text{If } \alpha \subseteq \phi, \text{ then } \Psi(\beta_1, \ldots \beta_n, \phi, \alpha, \beta_{n+1}, \ldots \beta_m) \\ &= \Psi(\beta_1, \ldots \beta_n, \alpha, \beta_{n+1}, \ldots \beta_m)\end{aligned}$$

The difference is that the stronger principle says that the whole belief system that results from the change, and not just the resulting belief state, should be the same in the two-stage procedure as it is in the corresponding one stage procedure. The intuitive motivation for (C1) seems to apply to the stronger principle as well: why should it matter for your posterior belief revision policies whether you get partial information before getting the same whole information? But the stronger principle cannot consistently be added to the DP postulates.

The apparent counterexamples to (C1) are puzzling, but I think that there are clearer and more decisive problems for (C2). Consider this example: Fair coins are

[28] Lehmann (1995).

flipped in each of two rooms. Alice and Bert (who I initially take to be reliable) report to me, independently, about the results: Alice tells me that the coin in room A came up heads, while Bert tells me that the coin in room B came up heads, and so this is what I believe at stage one. Because my sources were independent, my belief revision policies, at stage one, will give priority to the H_AT_B and T_AH_B possibilities over the T_AT_B possibility (Were I to learn that Bert was wrong, I would continue to believe that Alice was right, and vice versa).[29] But now Carla and Dora, also two independent witnesses whose reliability, in my view, trumps that of Alice and Bert, give me information that conflicts with what I heard from Alice and Bert. Carla tells me that the coin in room A came up tails, and Dora tells me the same about the coin in room B. These two reports are also given independently, though we may assume simultaneously.[30] This is stage two. Finally (stage three), Elmer, whose reliability trumps everyone else, tells me that that the coin in room A in fact landed heads (So Alice was right after all). What should I now believe about the coin in room B? DP's postulate (C2) requires that I return to the original belief that the coin in room B came up heads.[31] Even though Dora's information had overturned Bert's information about the second coin, and even though Elmer provided no information at all about the result of the coin flip in room B (we may assume he knew nothing about it), Elmer's information still forces us to change our belief about this result (if we follow the DP constraint, C2). This seems just as unreasonable as the result in the counterexample to Boutilier. In both cases, information is lost which, intuitively, should have been preserved.

One might object that I should respond as I did to Hans Rott's example, discussed above, by saying that relevant information from the intuitive story has been left out of the model. At each stage, I received the information, not only that the coins landed a certain way, but also that I received this information, in some cases from independent sources, and that meta-information is clearly relevant to intuitive judgments about the way that beliefs should be revised. But in this case, adding this additional information would not help to avoid the consequences of the iteration principle. The full information that I receive from Elmer's report still contradicts the total information that I had previously received from the reports of Carla and Dora, so the argument still holds.

Unlike the situation with (C1), with (C2) it is not hard to see what goes wrong at a general level. (C2) directs us to take back the totality of any information that is overturned. Specifically, if we first receive information α, and then receive information that conflicts with α, we should return to the belief state we were

[29] The full AGM system requires a total ordering of the possibilities, so the assumption that neither HT nor TH has priority over the other implies that they are tied. It might be more reasonable to assume that they are incomparable (and that the ordering of possibilities is only a partial order). But all we need for the current argument is that both HT and TH have priority over TT.

[30] I am assuming here that the conjunction of the two simultaneously received reports count as a single input, since the subject comes to believe, at one time, that both reports are true. One might question this assumption, but nothing the theory as it stands provides any constraint on what counts as a single input, or any resources for representing the independence of sources.

[31] Here is the argument: Since H_A contradicts T_AT_B, $\Psi(H_AH_B, T_AT_B, H_A) = \Psi(H_AH_B, H_A)$ by (C2). But since H_AH_B entails H_A, $\Psi(H_AH_B, H_A) = \Psi(H_AH_B)$.

previously in before learning α. But this directive is too strong. Even if the new information conflicts with the information just received, it need not necessarily cast doubt on *all* of that information. The previous information might have been quite complex, consisting of independent parts, and it might be that the new evidence threatens only one of those independent parts. To take an extreme case, just to dramatize the problem, suppose I start out knowing almost nothing, and then receive, at once, a fat encyclopedia of information. After I have digested it, I am told that there is a typo on p. 576: the Battle of Hastings was in 1066, rather than 1606. Should I then throw out the encyclopedia and go back to my prior state of almost total ignorance? That does not seem reasonable. The problem parallels the general problems of counterfactuals and of belief-contravening evidence. In all these cases the problem is to decide what to give up and what to keep when one must give up something to maintain consistency. This kind of problem arises, not only for my whole prior belief state, when the evidence conflicts with it, but also for the specific information just received, when the still newer information conflicts with it. As we have all known since Nelson Goodman gave up on the problem of counterfactuals more than 60 years ago, there is no general formula for answering this kind of question. This lesson gives us reason to reject (C2), despite the fact that we thereby lose the nice formal structure of the DP theory and of the Spohn iteration procedure, with its structural parallel with generalized conditionalization. If we do give it up, we will be left, at best, with a relatively weak set of constraints on iterated belief revision.

So where do we go from here? The examples we have discussed bring out the importance of various kinds of meta-information. In many examples, assumptions about the sources of our information, in particular about the independence of sources, played a central role in motivating intuitive judgments about how to repair our belief state in the face of evidence that something we took ourselves to know is in fact false. To decide how to revise in a particular case in which we learn that we were mistaken, we need to think about what went wrong, and how to explain it. In discovering that an item of presumed knowledge was in fact not knowledge, we discover that there was some defect or corruption at some point in the channels through which information was transmitted to us, or in the reasoning we used to reach a conclusion from the information we received. The more we know about the source of the problem, the better repair job we can do. We, as believers revising our beliefs, need to think about explanations, not only for the things that we take ourselves to know, but also for the fact that we know them, or in the case of putative knowledge that was overturned, the fact that we did not know them. And as theorists, we might be able to say something more substantive about this kind of process if we had the resources to represent meta-information, and its relation to ground-level information about the subject matter that is the primary focus. There are different kinds of independence—conceptual, causal and epistemic—that interact, and one might be able to say more about constraints on rational belief revision if one had a model theory in which causal-counterfactual and epistemic information could both be represented. There are familiar problems, both technical and philosophical, that arise when one tries to make meta-information explicit, since it is self-locating (and auto-epistemic) information, and information about changing

states of the world. But the problems are not intractable; a good theory of iterated belief revision will need to tackle them, and I would expect a theory that did to yield interesting results.

Acknowledgements Thanks to the participants in the Konstanz workshop on conditionals and ranking functions, in particular to Thony Gillies, Franz Huber, Hans Rott and Wolfgang Spohn, for discussion of the issues about iterated belief revision and ranking functions, and for advice about the relevant literature. Thanks also to the editors and an anonymous referee for helpful comments and corrections. I am particularly grateful to Hans Rott, who gave me detailed comments and suggestions on a previous draft, leading to some revision of my beliefs about belief revision, and I hope to significant improvements in the paper.

References

Boutilier, C. (1996). Iterated revision and minimal change of conditional beliefs. *Journal of Philosophical Logic, 25*, 262–305.

Darwiche, A., & Pearl, J. (1997). On the logic of iterated belief revision. *Artificial Intelligence, 89*, 1–29.

Friedman, N., & Halpern, J. (1999). Belief revision: A critique. *Journal of Logic, Language, and Information, 8*, 401–420.

Gärdenfors, P., & Rott, H. (1995). Belief revision. In D. Gabbay, et al. (Eds.), *Handbook of logic in artificial intelligence and logic programming IV: Epistemic and temporal reasoning* (pp. 35–132). Oxford: Oxford University Press.

Ginsberg, M. (1986). Counterfactuals. *Artificial Intelligence, 30*, 35–80.

Grove, A. (1988). Two modellings for theory change. *Journal of Philosophical Logic, 17*, 157–170.

Hild, M. (1998). Auto-epistemology and updating. *Philosophical Studies, 92*, 321–361.

Lehmann, D. (1995). Belief revision, revised. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (pp. 1534–1540).

Rott, H. (1999). Coherence and conservativism in the dynamics of belief. *Erkenntnis, 50*, 387–412.

Rott, H. (2001). *Change, choice and inference*. Oxford Logic Guides (Vol. 42). Oxford: Clarendon Press.

Rott, H. (2004). A counterexample to six fundamental principles of belief formation. *Synthese, 139b*, 225–240.

Spohn, W. (1988). Ordinal conditional functions: A dynamic theory of epistemic states. In W. Harper & B. Skyrms (Eds.), *Causation in decision, belief change and statistics* (Vol. 2, pp. 105–134). Dordrecht: Reidel.

Stalnaker, R. (1994). What is a nonmonotonic consequence relation? *Fundamenta Informaticae, 21*(1/2), 7–12.

Stalnaker, R. (2006). On logics of knowledge and belief. *Philosophical Studies, 120*, 169–199.

Williamson, T. (2001). *Knowledge and its limits*. Oxford: Oxford University Press.