Synthese
DOI 10.1007/s11229-010-9863-7

Preference change and conservatism: comparing the Bayesian and the AGM models of preference revision

Till Grüne-Yanoff

Received: 10 March 2010 / Accepted: 9 December 2010

Abstract Richard Bradley's Bayesian model of preference kinematics is compared with Sven Ove Hansson's AGM-style model of preference revision. Both seek to model the revision of preference orders as a consequence of retaining consistency when some preferences change. Both models are often interpreted normatively, as giving advice on how an agent should revise her preferences. I raise four criticisms of the Bayesian model: it is unrealistic; it neglects an important change mechanism; it disregards endogenous information relevant to preference change, in particular about similarity and incompleteness; and its representational framework, when expanded with similarity comparisons, may give misleading advice. These criticisms are based on a principle of conservatism, and on two proposals of similarity metrics for the Bayesian model. The performance of the Bayesian model, with and without the similarity metrics, is then tested in three different cases of preference change, and compared to the performance of the AGM model.

Keywords Preference change · Attitude revision · Bayesianism · AGM · Consistency · Conservatism · Similarity · Rationality · Normative decision theory

This paper was presented at the third Decision, Games and Logic workshop (DGL'09) in Lausanne.

T. Grüne-Yanoff (✉)
Helsinki Collegium of Advanced Studies, University of Helsinki,
P.O. Box 4, 00014 Helsinki, Finland
e-mail: till.grune@helsinki.fi

Published online: 20 January 2011

0 Introduction

Bayesian and AGM-style models are both used for modelling preference revision.[1] In this paper, I compare and critically discuss both modelling approaches. In doing so, I focus on three cases of pure preference change, which will help to highlight the differences between the approaches. The aim of this comparison is to identify the respective advantages and disadvantages of the two accounts, and to suggest possible improvements.

AGM and Bayesian models share significant core features. They are both consistency-preservation models (Grüne-Yanoff and Hansson 2009, pp. 17–19), which proceed in three steps:

(1) The agent's mental state is represented by some formal structure. Certain rationality constraints (e.g., transitivity) are imposed on that representation.
(2) A local change is introduced into the representation. This change is commonly interpreted as a certain preference or probability judgement arising from some learning experience.
(3) The representation is adjusted to incorporate this local change, retain overall consistency with the constraints, and remain maximally similar to the prior representation.

Consistency-preserving models of this kind often have a normative function: they tell an agent how to revise her mental attitudes in order to successfully incorporate the local change while retaining consistency and remaining as close as possible to her prior judgements.

Yet the two approaches employ different modelling strategies. On the one hand, Bayesian models represent an agent's state of mind as a probability and a desirability function. The probability function represents the agent's beliefs, and the desirability function her preferences. Revision operations are defined on these two functions. AGM-models, on the other hand, represent an agent's state of mind as sets of sentences or set-theoretic models that make these sentences true. More specifically, knowledge sets or knowledge models represent the agent's certain beliefs, and preference sets or preference models represent her preferences. Revision operations are defined on these sets of sentences or on these models.

Whereas Bayesian and AGM models have been compared with respect to their modelling of epistemic change (e.g., Gärdenfors 1988, pp. 36–40), the respective models of preference revision have not, to the best of my knowledge, been compared in the literature thus far. This paper provides such a comparison, the aim being three-fold: first, to shed light on an important but rarely discussed problem in preference revision; second, to critically assess the performance of the respective approaches in modelling this problem; and third, to propose improvements to these models.

[1] AGM models are named after the three authors of an influential early model of belief change, *A*lchourrón, *G*ärdenfors and *M*akinson. The preference-revision models discussed here use the same approach and similar logical machinery. Because the AGM authors have not explicitly endorsed these preference-revision models, it is better to speak of AGM-*style* models. For reasons of simplicity, I simply use 'AGM model' in this article.

Fig. 1 The relations between AGM and Bayesian models

$$\begin{tikzcd}
u(\cdot) \arrow[rr, "(Y)"] & & u^*(\cdot) \\
\succeq \arrow[u] \arrow[rr, "(X)"] & & \succeq^* \arrow[u]
\end{tikzcd}$$

The comparison in this paper is restricted to pure preference changes, which concern changes of preference that are not driven by a change in beliefs. Further, I will restrict myself to those pure preference changes that are initiated by a change in a single preference between two alternatives. By way of illustration, take the following example. An agent prefers having dinner in an Afghan rather than a Bangladeshi restaurant, and prefers the Bangladeshi to a Chinese restaurant. Then, after having recently sampled both menus, she comes to prefer Chinese over Afghan cuisine. Yet her prior preferences require her, on pain of inconsistency, not to hold a preference for Chinese over Afghan. If she wants to hold the new preference she must re-adjust her other preferences as well. This re-adjustment is a case of pure preference change, as it is driven solely by the acquisition of a new preference and not by any change in beliefs.

One of the strengths of Bayesian models is that they treat preference and belief revision at the same time. It is nevertheless possible to conceptually separate the two within the Bayesian framework, and exclusively focus on the latter. This makes comparison between Bayesian and AGM models feasible. When the focus is on pure preference change, disregarding belief changes and preference changes driven by beliefs, the two models are based on two different representations of the same thing, namely preference orderings.[2] Figure 1 shows AGM models defining a revision operator (X) between preference relations. Bayesian models, on the other hand, represent these preference relations as desirability functions, and define the revision operator (Y) between them.

I compare two specific models in this paper—Bradley (2007, 2009a,b) Bayesian *preference kinematics* and Hansson (1995, 2001) AGM-style *preference revision*.[3] I will show that there are cases in which the Bayesian change operator (Y) gives more ambiguous advice than the AGM operator (X). Beyond this ambiguity, there also are cases in which the Bayesian operator gives incorrect advice and the AGM operator gives correct advice. I argue that this is partly attributable to differences in the representational frameworks of the two models.

I develop my argument as follows. Sections 1 and 2 present the basics of the Bayesian and the AGM accounts, respectively. Section 3 introduces the principle of conservatism, and discusses why it is normatively relevant to preference revision. Section 4 proposes some possible similarity metrics that would allow the Bayesian

[2] This restriction is largely at the expense of the Bayesian model, which is designed to deal with informationally richer inputs. Nevertheless, the criticism derived from this partial comparison is still valid as long as the features neglected here do not promise to remedy the shortcomings detected in it.

[3] Bradley recently linked his model to AGM-style preference revision (Bradley 2009c, p. 245). I therefore refer to the particular kind of preference change modelled in both approaches as preference revision.

model to satisfy conservatism. Section 5 presents three cases, which help in testing the performance of the two models. The significance of these findings are discussed in the concluding section. In particular, I criticise the Bayesian model for neglecting pure preference change, for neglecting important endogenous information, and for offering unnecessarily ambiguous conclusions. Consequently, I propose expanding the Bayesian model with a model of pure preference change.

1 The AGM account

Hansson's model represents the motivational part of an agent's state of mind as a preference model $\mathbf{R}$. $\mathbf{R}$ is a set of preference relations. A preference relation R is a set of binary tuples $\langle X, Y \rangle$, in which each element X of such a tuple is a proposition of a Boolean algebra Ω. Each preference relation validates certain preference sentences. The set $[R]$ of these sentences is defined recursively according to the usual semantic interpretation of propositional logic:

$$A \succeq B \in [R] \Leftrightarrow \langle A, B \rangle \in R \tag{1}$$

$$\neg\alpha \in [R] \Leftrightarrow \alpha \notin [R] \tag{2}$$

$$\ldots$$

The preference model $\mathbf{R}$ then validates those sentences that are validated by each of its preference relations: $[\mathbf{R}] = \bigcap\{[R] | R \in \mathbf{R}\}$. $\mathbf{R}$ is called T-obeying if $[\mathbf{R}]$ is consistent and $[\mathbf{R}] = Cn_T([\mathbf{R}])$. Cn_T stands for the closure under propositional logic *and* preference rationality postulates.

When agents, as a result of some learning experience, adopt a new preference sentence that is not in $[\mathbf{R}]$, Hansson proposes Π-*prioritised revision* as the updating rule. This operation revises $\mathbf{R}$ by a preference sentence α, yielding the revised preference model $\mathbf{R}^*_{\mathbf{\Pi}}\alpha$. $\mathbf{R}^*_{\mathbf{\Pi}}\alpha$ must validate the sentence α and must be maximally similar to $\mathbf{R}$, giving priority in similarity comparison to preferences between propositions not in Π.[4]

Hansson constructs Π-similarity as follows. Π is the set of prioritised preferences that should be changed last. A preference relation R_1 is more Π-*similar* to R_2 than to R_3 if and only if the symmetric difference between R_1 and R_2 is smaller than between R_1 and R_3 for those R_is not in Π, or if these symmetric differences are equal, then for all R_is. More specifically,

(Π) Let μ be a numerical function such that if $\Phi \subset \Psi$, for any $\Phi, \Psi \in \Omega \times \Omega$, then $\mu(\Phi) < \mu(\Psi)$. Let $\Phi\Delta\Psi = (\Phi \setminus \Psi) \cup (\Psi \setminus \Phi)$. Then R_1 is more Π-*similar* to R_2 than to R_3 if and only if either:
$\mu(R_1 \Delta R_2) < \mu(R_1 \Delta R_3)$ for $R_1, R_2, R_3 \in \Pi$, or
$\mu(R_1 \Delta R_2) = \mu(R_1 \Delta R_3)$ for $R_1, R_2, R_3 \in \Pi$ and $\mu(R_1 \Delta R_2) < \mu(R_1 \Delta R_3)$.

[4] Multiple preference revisions are interpreted as α consisting of a conjunction of (finite) preference sentences (Hansson 2001, pp. 47–48).

$\mathbf{R}^*_{\Pi}\alpha$ has many properties that make it a plausible rule for preference revision (Hansson 2001, p. 50). Crucially, it explicitly adopts a version of conservatism in insisting on maximal similarity between $\mathbf{R}$ and $\mathbf{R}^*_{\Pi}\alpha$. The notion of conservatism is explained in Sect. 3, and an outline of its normative relevance is given.

2 The Bayesian account

Bradley's model represents an agent's state of mind as a pair of functions $\langle p, v\rangle$, defined on a Boolean algebra Ω of propositions. Thus, p is a probability measure of the agent's degrees of belief, and v is a real-valued (desirability) measure of the agent's preferences. This mode of representation is based on Jeffrey–Bolker decision theory (Jeffrey 1983).

When agents, as a result of some learning experience, change the probabilities or desirabilities of elements of a partition A_i of Ω, Bradley proposes *generalised conditioning* as the updating rule. In the following equations, $\langle p, v\rangle$ are the probability and desirability functions before the change, and $\langle p^*, v^*\rangle$ are the functions after the change. X is any proposition from Ω that is assigned a new probability and desirability according to these two rules.

$$v^*(X) = \sum_{i=1}^{n}[v(XA_i) + v^*(A_i) - v(A_i)] \times p^*(A_i|X) \tag{3}$$

$$p^*(X) = \sum_{i=1}^{n} p(X|A_i) \times p^*(A_i) \tag{4}$$

Equation 3 states that when the desirability and probability of elements of a partition A_i of Ω changes, then the posterior desirability of any proposition X in Ω is computed as the sum of the prior desirabilities of the conjunctions of X and A_i, plus the difference between the posterior and the prior desirabilities of A_i, weighted by the posterior probabilities of A_i, given X. Equation 4, in turn, states that when the probability of elements of a partition A_i of Ω changes, then the posterior probability of any proposition X in Ω is computed as the sum of the prior probabilities of X, given A_i, weighted by the posterior probability of A_i.

Generalised conditioning is the appropriate updating rule if and only if the preferences satisfy the Jeffrey–Bolker preference axioms, and the conditional preferences are rigid with respect to some partition A_i in Ω (Bradley 2007, p. 523). Rigidity of conditional preferences requires that $X \succeq_{Ai} Y \Leftrightarrow X \succeq^*_{Ai} Y$. In other words, the preference judgement $\succeq_{Ai}$, made on the supposition that A_i is true, must cohere with unconditional preferences after revision by A_i.

Furthermore, *all* preference revision can be modelled as generalised conditioning. For any two pairs of preferences $\langle p, v\rangle$ and $\langle p^*, v^*\rangle$, there exists some partition $\{A_i\}$ such that $\langle p^*, v^*\rangle$ is obtained from $\langle p, v\rangle$ through generalised conditioning on this partition (Bradley 2009a, p. 236).

In instances of *pure preference change*, learning experience only affects the agent's preferences, and leaves her beliefs intact. In such cases, $p^*(\cdot) = p(\cdot)$ and Eq. 3 of generalised conditioning can be simplified.

$$
\begin{aligned}
v^*(X) &= \sum_{i=1}^{n}[v(XA_i) + v^*(A_i) - v(A_i)] \times p(A_i|X) \\
&= \sum_{i=1}^{n}[v(XA_i) \times p(A_i|X) + (v^*(A_i) - v(A_i)) \times p(A_i|X)] \\
&= v(X) + \sum_{i=1}^{n}(v^*(A_i) - v(A_i)) \times p(A_i|X) \qquad (5)
\end{aligned}
$$

Thus, the revised desirability of a prospect X varies from the prior desirability by virtue of the change in taste for prospect A_i, and the probabilistic dependence of A_i on X (Bradley 2009b, p. 232). Equation 5 specifies how a pure preference change affects the evaluation of other propositions probabilistically dependent on the partition elements whose evaluation had changed.

What does this model say about the three-restaurant example given in the introduction? In that case, the question concerned how prior preferences are affected by a new preference for Chinese over Afghan cuisine. Equation 5 does not help here. The example starts with the change of a preference over two of the three elements of the partition {Afghan,Bangladeshi,Chinese}. It then investigates how the preference comparisons of the other elements are affected. Equation 5, in contrast, becomes trivial when X is identical to some A_i.[5] Bradley's approach thus merely suggests modelling the change as a re-assignment of desirabilities to (some of) the elements of the partition $\{A_i\}$. This can be done in at least three different ways: desirabilities may be assigned to the three elements such that either (i) Chinese is preferred to Afghan, Afghan is preferred to Bangladeshi, or (ii) Bangladeshi is preferred to Chinese, Chinese is preferred to Afghan, or (iii) Chinese is preferred to Bangladeshi, Bangladeshi is preferred to Afghan. Any of these options accommodates the new preference for Chinese over Afghan, but the Bayesian model does not indicate which way of redistributing the desirabilities is the correct one.

Yet it is exactly with respect to these bare-bone cases of pure preference change that the Bayesian model is comparable to the AGM model. The AGM model shows how **R** must be adjusted in order to validate some preference sentence α, while satisfying consistency and conservatism. It models the rational reaction to accommodating a perturbing judgement α into one's preference state. The Bayesian approach specifies—albeit in an ambiguous way—how the desirability function $v(\cdot)$ over Ω must be adjusted to be consistent with new desirability assignments on some elements of $\{A_i\}$.

The ambiguity of this model, I will argue, stems from its neglect of conservatism. It is this principle that I discuss next.

[5] If X is equal to e.g. $B \in \{A, B, C\}$, then $p(A|B) = p(C|B) = 0$ and Eq. 5 reduces to $v^*(B) = v(X) + \sum_{i=1}^{n}(v^*(B) - v(B)) \times p(B|B)$. But this is equivalent to $v^*(B) = v^*(B)$.

3 The principle of conservatism

The principle of conservatism has been discussed largely in the context of epistemic attitudes (see McCain 2008 and the references therein). Yet it also applies to evaluative attitudes such as preferences. In the latter context, the principle of preference conservatism could be defined as follows:

> **(PPC)** If S prefers A to B, and $A \succ B$ is consistent with S's preferences, then S is justified in retaining $A \succ B$, and remains so as long as $A \succ B$ is not defeated for S.

A preference $A \succ B$ is *defeated* for S if S has better reasons not to hold $A \succ B$, or equally good reasons not to hold $A \succ B$ as to hold it, and $\neg(A \succ B)$ is consistent with S' preferences. Thus, conservatism acknowledges that agents change their preferences as a consequence of changing reasons, and it acknowledges that in order to successfully adopt these changed preferences, other adjustments to the preference state may be necessary. But it mandates that no unnecessary adjustments be performed, and hence that the revision be minimal. Three main justifications are given in support of the principle of conservatism.

First, the *cognitive process* of revision is costly, and adhering to the principle keeps these costs at a minimum. Costs are incurred both through determining what one's rational commitments are, and through seeing to it that these commitments are honoured. Determining what needs to be adjusted in order to retain consistency may be computationally demanding, and hence costly. These costs rise rapidly in accordance with the number of adjustments necessitated. Honouring commitments is likely to be even more costly than computing them. Human agents often find it hard to 'give up' some of their preferences. Preferences driven by visceral factors, such as appetites, sexual desire and addiction are examples: agents may acknowledge that they should give them up (in this case for the sake of consistency), and may commit to doing so, yet find that they still influence their decisions and actions. Successfully removing such preferences from one's mental state can be extremely costly, involving large amounts of resources and occupying the agent for long periods of time. Adhering to the principle of conservatism helps in minimising both the computational and the commitment-honouring costs of preference revision.

Secondly, one's preferences constitute some kind of *accumulated capital* of one's past reasoning. For example, one may have forgotten the reasons why one adopted a preference, and the preference serves as a reminder of such past reasoning processes. Adhering to the principle of conservatism preserves this capital as much as possible.

Thirdly, preferences may fulfil certain *functions* that require their stability. For example, they constitute part of an individual's personal identity. To the extent that personal identity is a value in itself, one should avoid destabilising it by unnecessarily changing one's preferences. Furthermore, given that most people make long-range future plans, and regularly check their optimality by comparing planned results against current preferences, the unnecessary changing of preferences would make consistent long-term planning very difficult, or near-impossible.

Consistency-preserving models of preference revision should therefore respect the principle of conservatism, the satisfaction of which, in PPC terms, ranks lexicographically after the maintenance of consistency, and after the successful incorporation of a local preference change.

4 Similarity measures for the Bayesian model

The application of the principle of conservatism depends on the availability of information about prior and posterior preference orderings, in particular similarity comparisons and defeaters. If the principle is normatively relevant, then the information needed for the similarity comparisons is also normatively relevant. Depending on how much of this relevant information each model makes use of, one can judge how well each model performs its normative function.

AGM models offer concrete proposals for the measurement of similarity between prior and posterior preference relations. In the following I use Hansson's Π-similarity, as presented in Sect. 1. Π-similarity satisfies PCC in the sense that it seeks to minimise the number of binary preference comparisons that need to be changed in a revision. Bayesian models, in contrast, do not offer similarity measures between prior and posterior desirabilities. For this reason, Bradley's Bayesian model does not give unambiguous advice on how to redistribute desirabilities in the three-restaurant example. This section proposes two possible similarity measures for the Bayesian model in order to fill this gap.

The first question to address when considering possible similarity measures concerns *what* is to be compared. Because the model is based on Jeffrey–Bolker decision theory, it is, in principle, possible to translate the desirability/probability functions into a preference ordering, and to compare these underlying prior and posterior orderings with respect to their similarity, in a similar way as the AGM model does. However, as Fig. 1 illustrates, the Bayesian model defines its revision operation on the desirability function, not the underlying preference ordering. It is therefore natural to seek a measure of similarity on this functional level, too. There are at least two different ways of doing this.

One kind of similarity measure compares desirability functions by the number of alternatives to which they assign the same desirability. A desirability function u is said to be more SA-similar to w than another function v if and only if u and w assign the same desirability to a larger number of alternatives than u and v do.[6] More precisely, it is defined as follows.

> **(SA)** Let u, v and w be desirability functions defined over the same domain of alternatives Ω. Let S^{uv} be the set of alternatives on which u and v assign different desirability: $S^{uv} = \{X \in \Omega | u(X) \neq v(X)\}$. Then u is more *SA-similar* to w than v if and only if $\#S^{uw} < \#S^{vw}$.

[6] This metric is related to the Levenshtein, Needleman-Wunsch and Smith-Waterman distances, known from information theory and computer science.

In the context of preference revision, let u and v be two posterior desirability functions, and w the prior desirability function. SA-revision chooses those posterior functions that are most SA-similar to the prior desirability function. It satisfies PPC in the sense that it seeks to minimise the number of desirability re-assignments in a preference revision.[7] This interpretation of PPC is most plausible if alternatives and desirabilities are seen as cognitively realistic: if alternatives and desires are real cognitive entities, the cost of revision consists in 'picking up' the relevant alternative, 'erasing' the assigned desirability and 'inscribing' a new one.

Another kind of similarity measure compares desirability functions by the differences between prior and posterior desirabilities. A desirability function u is said to be more SD-similar to w than another function v if and only if the sum of desirability differences between u and w is smaller than between v and w.[8] More precisely, it is defined here as follows.

> **(SD)** Let u, v and w be desirability functions defined over the same domain of alternatives Ω. Let $\delta^{uw} = \sum_i |\frac{u(x_i)}{U} - \frac{w(x_i)}{W}|$ be the average-weighted differences between the desirability functions u and w, where U and W are the average values of the sets $\{u(x_i)\}$ and $\{w(x_i)\}$, respectively. Then u is more *SD-similar* to w than v if and only if $\delta^{uw} < \delta^{vw}$.[9]

In the context of preference revision, let u and v be two posterior desirability functions of w. SD-revision chooses those posterior functions that are most SD-similar to the prior desirability function. It satisfies PPC in the sense that it seeks to minimise the overall difference in evaluation between the two functions. Similar approaches are to be found in curve-fitting techniques in statistics. This interpretation of PPC is most plausible if desirability degrees are seen as cognitively realistic: if they are the main cognitive entities, the cost of revision consists in 'distancing' the posterior from the prior degree.

There are numerous other ways of calculating the distance between two discrete functions through their respective values at each argument. For example, SD may be weighted by the probability of each alternative, hence giving priority to desirability differences between the alternatives that are more probable. Alternatively, instead of computing differences, one may compute weighted products, such as the cosine similarity $\cos(\theta^{uv}) = \frac{\sum u(x_i)v(x_i)}{\sqrt{\sum u(x_j)^2}\sqrt{\sum v(x_j)^2}}$. I do not discuss these alternatives in the following for two reasons. First, some of these metrics require inputs that are not readily available in the comparison made in this paper, such as the probability of alternatives. Secondly, the aim is not to resolve the technical question of which metric may

[7] In the most common interpretation, desirability functions are identical up to positive affine transformations. Thus, no special meaning is assigned to something having a desirability number 0, and there is no highest or lowest bound on a desirability function. Desirability reassignments, then, only serve to revise the *relative* positions of the prospects. SA-similarity measures the minimal number of necessary reassignments for each preference revision.

[8] Related metrics are the taxicab metric and Euclidean distance. More specifically, SD is an average-weighted taxicab distance. The weighting acts as a normalisation, which accommodates the identity of desirability functions up to positive affine transformations.

[9] I am indebted to Richard Bradley for suggesting this measure to me.

be most suitable. The focus is rather on the conceptual question of what kind of similarity comparison would be most appropriate, how these comparisons satisfy PPC, and how much they are influenced by the modes of representation of the respective AGM and Bayesian models. For this it matters whether similarity applies to binary preference comparisons, desirability reassignments, or desirability differences. Π-, SA- and SD-similarity are merely exemplars of these different kinds of similarity measures.

5 Comparing model performance

In the following I discuss three stylised cases of preference revision. Each case highlights how the particularities of the respective representational frameworks influence the modelling. In particular, these differences affect the kind of information about the perturbatory judgements and prior orderings that is available for similarity comparisons. I argue on the basis of these cases that the AGM model sometimes outperforms the Bayesian model in its normative function.

Case 1 Let there be only three mutually exclusive alternatives, $\{A, B, C\}$. These propositions are ordered as follows:

As in the three-restaurant example given in the introduction, the agent comes to prefer C over A. How do the two modelling approaches recommend revising the prior ordering? Given that the preference $C \succ A$ must be represented in the posterior ordering, three solutions are possible, as presented in Table 1.

Hansson's AGM approach represents the prior ordering in Table 1 by the set

$$\mathbf{R} = \{\{\langle A, B\rangle, \langle B, C\rangle, \langle A, C\rangle\}\}$$

R can be manipulated in three ways, which correspond to the three solutions given in Table 2.

$$\begin{aligned}
\mathbf{R_i} &= \{\{\langle B, A\rangle, \langle B, C\rangle, \langle C, A\rangle\}\}\\
\mathbf{R_{ii}} &= \{\{\langle A, B\rangle, \langle C, B\rangle, \langle C, A\rangle\}\}\\
\mathbf{R_{iii}} &= \{\{\langle B, A\rangle, \langle C, B\rangle, \langle C, A\rangle\}\}
\end{aligned}$$

$\mathbf{R_i}$ and $\mathbf{R_{ii}}$ are more similar to **R** than $\mathbf{R_{iii}}$ is because they differ in only two elements from **R**, whereas $\mathbf{R_{iii}}$ differs in three elements. Given a Π-similarity interpretation of conservatism, with Π either empty or identical to Ω, the AGM approach can dismiss

Table 1 An ordering of three alternatives

	$v(X)$
A	3
B	2
C	1

Table 2 Three possible solutions

(i)	(ii)	(iii)	$v^*(x)$
B	C	C	3
C	A	B	2
A	B	A	1

solution **iii** as violating PPC. It does with reference to *endogenous* information—by counting the number of tuples that differ between the prior and the posterior orderings.

The Bayesian account, in contrast, does not explicitly offer any such measure. It represents the prior ordering with the desirability function $v(X)$ shown in the right-hand column in Table 2. Bradley merely suggests modelling the revision as the redistribution of desirabilities over the relevant partition of Ω, which in this case is $\{A, B, C\}$. Such a redistribution leaves unspecified the relative positions of C and A with respect to B on the desirability scale. Satisfying $v^*(C) > v^*(A)$, one may assign values to $v^*(C)$ and $v^*(A)$ that leave $v^*(B)$ (i) larger than $v^*(C)$, or (ii) smaller than $v^*(A)$, or (iii) smaller than $v^*(C)$ but larger than $v^*(A)$. The numbering of these possibilities corresponds to the solutions given in Table 3. From the Bayesian perspective, none of these solutions seem preferable over the others.

The Bayesian model can be expanded by one of the similarity measures discussed in Sect. 4. Table 3 shows the *distances* between the solutions given in Table 2 and the prior ordering. The respective similarities are the inverse of these distances.

I first explain how these numbers come about. The computation of Π- and SA-similarity is straightforward. According to Π, the symmetric difference between, say R and R_i is $R \Delta R_i = \{\langle A, B\rangle, \langle B, A\rangle, \langle A, C\rangle, \langle C, A\rangle\}$.[10] $R \Delta R_i$ has a cardinality of 4, as shown in the table. Now to SA: solution **i** is obtained by reassigning A a desirability lower than C. Hence the two functions differ only in one desirability assignment. The same holds for **ii**. Solution **iii**, however, requires reassigning two desirabilities.

The computation of SD is complicated by the way new desirabilities are assigned. Within the Jeffrey–Bolker framework, if desirabilities have been assigned to at least two alternatives, and if preferences and beliefs satisfy certain conditions, the desirabilities are assigned as follows. A gamble is constructed that includes the alternative requiring desirability reassignment. The probability of this gamble is set in such a way that the agent in question is indifferent between the gamble and one of those alternatives requiring no reassignment. Then, one solves for the desirability of the reassigned alternative as a function of the gamble's probability. Finally, the posterior desirability function is co-scaled with the prior function by normalising it to the same point of origin.

For example, solution **i** is obtained by reassigning a lower desirability to A than to C, such that C is now located between B and A. Therefore, the agent will be indifferent between C, and gamble $(A, p; B, 1-p)$ between A and B with some probability p.

[10] Here and in Table 5 I disregard exogenous information and assume Π to be either empty or containing all elements of Ω.

Table 3 Three possible solutions and their distances from the prior ordering

	i	**ii**	**iii**
Π-distance	4	4	6
SA-distance	1	1	2
SD-distance	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$

In other words, in the posterior ordering,

$$v^*(C) = v^*(A, p; B, 1-p) = p \times v^*(A) + (1-p) \times v^*(B) \tag{6}$$

Because the positions of C and B remain the same, $v^*(C) = v(C)$ and $v^*(B) = v(B)$. Substituting these into (6) yields

$$v^*(A) = 2 - \frac{1}{p} \tag{7}$$

In order to keep things simple I assume desirability equidistance between all alternatives. Cases 1 to 3 all start with equidistant prior orderings. In order to maintain this for posterior orderings I assume that the probability p of gamble $(A, p; B, 1-p)$ is $\frac{1}{2}$. From (7) it then follows that $v^*(A) = 0$.

In the final step, this posterior desirability function needs to be co-scaled with the prior function. In order to do that I calculate the desirability of the tautology $T \equiv A \vee B \vee C$. Assuming equiprobability of all options, $v(T) = \frac{\sum_i^n X_i}{n}$ for all X_i from the partition $\Omega = \{X_i\}$. Then for the prior function, $v(T) = 2$, and for solution **i**, $v_i^*(T) = 1$. I therefore normalise $v_i^*(\cdot)$ by adding a constant of $c = 1$ to the function. Hence the newly assigned posterior desirability function $v_i^*(\cdot)$ assigns $v_i^*(C) = 3$, $v_i^*(B) = 2$ and $v_i^*(A) = 1$. Comparing $v(\cdot)$ and $v_i^*(\cdot)$ through SD yields the numbers in the last row of Table 3.

Two observations arise from Table 3. First, SA-similarity and Π-similarity make the same distinctions: they both identify **i** and **ii** as more conservative than **iii**. This is because with only three alternatives in the partition, Π's focus on pair-wise comparisons and SA's focus on reassignments are equivalent. This is not the case with more than three alternatives, as I show in the next case.

Secondly, SA and Π-similarity on the one hand, and SD-similarity on the other, diverge. SD-similarity assigns the same distance to all three solutions, in contrast to the intuition that **iii** is a stronger revision of the prior ordering than either **i** or **ii**.

Let me go back to the illustrative example. If one preferred Afghan to Bangladeshi and Bangladeshi to Chinese cuisine, and then came to prefer Chinese to Afghan, one could accommodate this new preference in the three ways described. Why would one choose **iii** and give up *all* of one's prior preferences? In the absence of defeaters, such a choice violates PPC. It would create more computation and realisation costs, destroy more preference capital, and obstruct the functions of preference more than necessary. The AGM model satisfies this intuition. The Bayesian model does not, and when expanded by similarity metrics, only the SA model gets it right.

Of course, there may be defeaters present. Suppose, for instance, that prospects A and B are very similar (both restaurants have a Pakistani chef, say), whereas C is different from both. Then coming to prefer C over A could be grounds for changing ones preference between B and C as well. In this case, the solution Hansson rejected is in fact the one that best respects conceptual coherence.

The answer is that in this case the similarity of A and B are taken to be a *reason* for changing ones preference between B and C. Such a reason constitutes a defeater of $B \succ C$. Hence PPC does not apply, and the models *must not* exclude **iii** on the basis of similarity considerations.

Admittedly, the AGM model does not easily accommodate such forms of defeaters. It can introduce supplementary information through adjusting Π, but this does not work for conditional restrictions such as 'if $A \succ B$ is changed, then $B \succ C$ must also be changed'. Thus there is room for improvement in the AGM model as well. However, the main lesson of this case remains: in the absence of defeaters, AGM and the SA-expanded Bayesian model satisfy PPC, whereas the standard and the SD-expanded Bayesian model do not.

Case 2 Let there be only four mutually exclusive alternatives, $\{A, B, C, D\}$. The order of these propositions is shown in Table 4.

The Bayesian model represents this ordering with the desirability function $v(X)$ as shown in the right-hand column of Table 4. This ordering is represented in Hansson's model by the preference model

$$\mathbf{R} = \{\{\langle A, B\rangle, \langle A, C\rangle, \langle A, D\rangle, \langle B, C\rangle, \langle B, D\rangle, \langle C, D\rangle\}\}$$

Now the agent comes to prefer C over A. How do the two modelling approaches recommend revising the prior ordering? Given that the preference $C \succ A$ must be validated by the posterior ordering, twelve solutions are possible (see Table 5).

I will point out the most notable aspects of this table by comparing the results of the different distance measures in pairs. First, the Π- and SA-distances—although agreeing on the most distant solution (**xii**) and on some of the most similar solutions, (**i** and **iii**)—diverge on **xi**: SA identifies it as a most similar solution, whereas Π does not. Why is that?

SA-distance counts the number of alternative positions that have to be altered against the whole ordering. This is a technically simple way to describe the revision, but it is intuitively difficult to see which changes actually occur in such a positional shift. Π-distance, in contrast, counts the number of binary comparisons that are changed.

Table 4 An ordering of four alternatives

	$v(X)$
A	4
B	3
C	2
D	1

Table 5 Twelve possible solutions and their distances to the prior ordering

	i	**ii**	**iii**	**iv**	**v**	**vi**	**vii**	**viii**	**ix**	**x**	**xi**	**xii**
	C	C	B	D	B	D	C	C	C	C	B	D
	A	A	C	C	D	B	B	B	D	D	C	C
	B	D	A	A	C	C	A	D	A	B	D	B
	D	B	D	B	A	A	D	A	B	A	A	A
Π-distance	4	6	4	10	8	10	6	8	8	10	6	12
SA-distance	1	2	1	2	2	2	2	2	2	2	1	3
SD-distance	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{2}{5}$	$\frac{4}{5}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	$\frac{4}{5}$	$\frac{3}{5}$	$\frac{4}{5}$

Only A's position is changed in solution **xi**, which is why SA identifies it as maximally similar. In fact, though, three pair-wise preference comparisons are changed through this positional shift, namely $A \succ B$, $A \succ C$, and $A \succ D$. This is why Π-similarity does not identify **xi** as a maximally similar solution.

Secondly, the SA- and SD-distances, although in accord with regard to some of the most similar solutions (**i** and **iii**), and some of the most distant (**xii**), are not with regards to solutions **vii** and **xi**: SA identifies **xi** as the most similar but SD does not, whereas SD identifies **vii** as the most similar but SA does not.[11] I have already explained the SA-result for **xi**. SD picks up on the large distance between the prior and posterior desirability of A, and hence seems the better indicator here. Nevertheless, **vii** is an interesting case that exposes the irreconcilability of similarity intuition focusing on pair-wise comparisons, desirability reassignments and desirability distances, respectively. In **vii**, A and C swap places. This creates a sufficiently small desirability difference for SD to group it with **i** and **iii** as the most similar. Yet for SA, this swap involves reassigning desirabilities to two alternatives, and for Π, it involves changing three pair-wise comparisons.

Third, Π-distance offers five levels of differentiation, whereas both SA and SD only offer three. This follows directly from the fact that there are only four different desirability levels, while there are six binary preference relations in each solution. This mostly affects the middle-sections of SA and SD, which lump together what Π differentiates, but it also evident in additional identifications of the most similar and most distant solutions, with regard to both.

Finally, Π is in accord with both SA and SD on the most similar and most distant solutions in the very same cases where SA and SD are in accord. It is therefore tempting to consider those solutions as correct that are identified by *both* SA *and* SD, or that are identified by Π.

I subscribe to this position. Whereas considering the relative position of an alternative vis-à-vis a whole set of options is often a convenient way of summarising one's preferences or making a preference-informed choice, pair-wise preference representation is the more natural and intuitive way of making similarity comparisons

[11] As in Case 1, I assume desirability equidistance and equiprobability of alternatives in order to simplify the computation.

between preference orderings. In particular, it is the number of changed binary preference judgements that is salient for considering the costs of preference revision, not the relative position of an option in relation to everything else.

AGM's Π-similarity captures this intuition in a way that the Bayesian model achieves only when the two proposed similarity metrics correct each other. As the above two cases show, both of these similarity measures are suspect in themselves. Under the assumption that conservatism is normatively relevant, and that the above intuitions are correct, I therefore conclude that the AGM model contains more normatively relevant information than the Bayesian model in these kinds of cases, and hence performs its normative function better.

Case 3 So far, I have discussed only cases with complete orderings. However, the explicit representation of preference incompleteness can provide relevant information on how preferences should be revised. The AGM account is capable of handling incomplete preference orderings. AGM belief states, for example, are explicitly not required to contain every sentence or its negation (cf. Gärdenfors 1988). In a similar fashion, it does not require that every alternative is part of an ordering.

Similarly, the Bayesian model can be expanded to represent incomplete preferences, by representing an agent's state of mind as a *set* of functions $\langle p_i, v_i \rangle$ (Bradley 2009c). Each of these function pairs can be thought of as a permissible sharpening of an agent's actual beliefs and desires. Bradley refers to an equivalence class of such a sharpening as an *avatar* of an agent. However, this framework cannot represent the special case of acyclical, non-transitive preferences discussed here. According to Bradley (2009c, p. 242), for all avatars i of an agent:

$$X \succeq Y \Leftrightarrow v_i(X) \geq v_i(Y) \tag{8}$$

However, $Y \succeq Z$ implies $v_i(Y) \geq v_i(Z)$ for all i, and from that and (7) it follows that $v_i(X) \geq v_i(Z)$, from which it follows that $X \succeq Z$. Consequently, Bradley's expanded model cannot represent preference incompleteness where this implies intransitivity (but not cyclicity). I will now discuss a case in which the representation of incomplete preference makes a difference for the modelling of minimal revision.

Let there be only three mutually exclusive alternatives $\{A, B, C\}$. Now there are two agents. Agent 1 has explicit preferences between A and B and between B and C, but not between A and C. Agent 2 has the same explicit preferences as agent 1, and also holds an explicit preference between A and C.[12]

Thus, their preference orderings are represented differently in Hansson's AGM model. The Bayesian model, for the reasons discussed above, cannot distinguish between these two cases. The two orderings give rise to the same desirability function (Table 6).

[12] Representing $\mathbf{R_1}$ and $\mathbf{R_2}$ differently requires giving up the claim that $\mathbf{R}$ is transitive. As discussed in Sect. 1, Hansson's definition (1) implies that if $\mathbf{R_1} = \{\{\langle A, B\rangle, \langle B, C\rangle\}\}$, then $A \succ C \notin [\mathbf{R}]$, and according to his definition (2) it follows that $\neg(A \succ C) \in [\mathbf{R}]$. Yet if transitivity is part of T, $A \succ C \in Cn_T\{\langle A, B\rangle, \langle B, C\rangle\}$, hence $Cn_T = \perp$. If T includes acyclicity but not transitivity, however, then $\neg(C \succeq A) \in Cn_T\{\langle A, B\rangle, \langle B, C\rangle\}$, but not $A \succ C \in Cn_T\{\langle A, B\rangle, \langle B, C\rangle\}$, and hence the above inconsistency is avoided.

Table 6 Two preference bases

$$\mathbf{R_1} \quad = \{\langle A, B\rangle, \langle B, C\rangle\}$$
$$\mathbf{R_2} \quad = \{\langle A, B\rangle, \langle B, C\rangle\, \langle A, C\rangle\}$$

	$v(X)$
A	3
B	2
C	1

Now both agents revise their preference ordering by $C \succ B$. This is trivial for both the AGM and the Bayesian model. From the AGM perspective there is a unique most similar preference model that contains $\langle C, B\rangle$ for both agents, whereas from the Bayesian perspective, either the desirability of C is raised or the desirability of B is reduced, neither of which affects the relation of B and C to A (Table 7).

However, now both agents revise their preference orderings by $B \succ A$. From the AGM perspective this is trivial only for agent 1: for her, there is a unique preference model that includes $\langle B, A\rangle$ and is most similar to her prior preference model, namely $\mathbf{R_1^{**}} = \{\langle B, A\rangle, \langle C, B\rangle\}$.

For agent 2, however, the same problems with transitivity as in Case 1 arise. From the AGM perspective, there are two consistent preference models that include $\langle B, A\rangle$ and are most similar to her prior preference model:

$$\mathbf{R_2^{**}} = \{\langle B, A\rangle, \langle B, C\rangle, \langle A, C\rangle\}$$
$$\mathbf{R_2^{***}} = \{\langle B, A\rangle, \langle C, B\rangle\, \langle C, A\rangle\}$$

From the Bayesian perspective, three possible and equally legitimate solutions exist (Table 8).

Without the use of a similarity metric, even under incomplete preferences, the Bayesian framework does not give any further advice on how to choose a solution. With regard to SA the Bayesian model arrives at the same conclusion as the AGM model under complete preferences, whereas with regard to SD it consider all solutions equally similar. Nevertheless, I believe that given the information about preference

Table 7 Two preference bases revised by $C \succ B$

$$\mathbf{R_1^*} \quad = \{\langle A, B\rangle, \langle C, B\rangle\}$$
$$\mathbf{R_2^*} \quad = \{\langle A, B\rangle, \langle C, B\rangle\, \langle A, C\rangle\}$$

	$v(X)$
A	3
C	2
B	1

Table 8 Three solutions for $v^*(x)$

(i)	(ii)	(iii)	$v^*(x)$
B	C	B	3
A	B	C	2
C	A	A	1

incompleteness, and in the absence of any defeaters, $\mathbf{R}_1^{**}$ is the solution that satisfies PPC best. There is no reason to change $C \succ B$ rather than $A \succ C$, because the agent does not hold $A \succ C$. Hence the accommodation of $B \succ A$ does not give rise to problems related to realising the elimination of $A \succ C$, its loss as reasoning capital, or for its functioning in plans and identity.

Consequently, the AGM model with relaxed transitivity makes most use of normatively relevant information about conservative preference revision, the AGM model with transitivity ranks second (possibly joint by an SA-expanded Bayesian model), and the standard Bayesian model makes the least use of such information of all three approaches.

6 Conclusion

The menu of my favourite lunch spot contains 23 items. I have fairly extensive preferences over these. One day I realise that I have changed preferences over two of the options. The question arises how this affects my other preferences related to the menu, given that I would like to remain consistent.

Bayesian models of preference revision do not answer this question. They rather require the complete and consistent redistribution of desirabilities over all menu items. As I have argued in this paper, this approach makes the Bayesian revision model unrealistic, negligent of important change mechanisms, negligent of important endogenous information—and hence unnecessarily ambiguous—and possibly leads to incorrect conclusions when one tries to expand the model.

The model is unrealistic because it commences modelling preference revision with a desirability redistribution over a complete partition. In contrast, I hope to have shown in the examples that preference change often starts with a change in a single binary preference comparison. Whereas the AGM approach can model such an initial impulse to preference revision, the Bayesian model cannot.

The Bayesian approach neglects important change mechanisms, exactly because it starts with a desirability redistribution over a complete partition. This begs the question of how a single preference reversal affects other preferences over the partition. Because the specification of a desirability function over the partition implies the consistency of the underlying preference ordering, the Bayesian model 'exogenises' an important part of the question: the effect of the single preference reversal on the ordering of the partition must be answered *before* the model is used. In contrast, the AGM model explicitly models the selection of the posterior ordering that accommodates the preference reversal.

The standard Bayesian model ignores important endogenous information that would help in modelling the effect of a single preference reversal on the ordering of the partition. As I have argued, conservatism (in the form of PPC) is a normatively relevant principle for preference revision, and is implemented through similarity comparisons between prior and posterior preference orderings. The standard Bayesian model does not offer such a similarity comparison, and hence ignores information that is normatively relevant for conservatism. This does not exclude the possibility that such similarity measures can be developed, but defenders of Bayesianism have not proposed any such measure so far. Furthermore, I have argued (Case 3) that information about incomplete preferences—in particular acyclical but intransitive preferences—is normatively relevant for preference revision. Although expanded versions of the Bayesian model are capable of representing incomplete preferences, they are not able to represent acyclical, intransitive preferences. Thus, the Bayesian representational framework makes only limited use of information that is relevant for preference revision, and therefore gives unnecessarily ambiguous advice.

Finally, attempts to expand the Bayesian model with similarity metrics may possibly lead to incorrect conclusions. The Bayesian approach models preference revision through changing desirability functions. Possible similarity metrics are therefore likely to be defined on these functions. This puts the focus of similarity comparisons on the number of desirability reassignments, or on desirability differences, as the SA and SD examples show. The AGM model, in contrast, focuses on pair-wise comparisons. As I have argued, similarity comparisons based on these two perspectives sometimes yield different results. I also point out that the arguments for conservatism favour the pair-wise perspective: as far as the costs of preference revision are concerned, the number of changed binary preference judgements is salient. Thus, correct application of the conservatism principle requires information that is more readily available from the representational framework of the AGM model than from the representational framework of the Bayesian model.

My comparison of the AGM and the Bayesian models has highlighted these concerns about the latter, but it should not hide the shortcomings of the former. Obvious concerns include the difficulty in defining a proper representation of PPC defeaters, as mentioned at the end of Case 1. Another criticism is the difficulty of efficiently dealing with instrumental influences on preference changes in AGM. Finally, the Bayesian model contains a lot of normatively relevant information about degrees of desirability and belief, and about the connection between the two, which the AGM model does not.

The goal in this paper thus was not so much to promote the AGM over the Bayesian approach, but rather to open up perspectives on possible improvements to Bayesian models. One such possibility is to model the effect of a single preference reversal on the ordering of the partition, and another is to develop a similarity metric for desirability functions that satisfies PPC. In both these perspectives, Bayesians could do worse than study the AGM approach in more detail.

Acknowledgments I would like to thank the participants of DGL'09, and in particular Richard Bradley, Franz Dietrich, Sven Ove Hansson and two anonymous referees, for extensive and very helpful comments. The usual disclaimers apply.

References

Bradley, R. (2007). The kinematics of belief and desire. *Synthese, 156*, 513–535.

Bradley, R. (2009a). Becker's thesis and three models of preference change. *Politics, Philosophy and Economics, 8*(2), 223–242.

Bradley, R. (2009b). Preference kinematics. In T. Grüne-Yanoff & S.-O. Hansson (Eds.), *Modelling preference change. Approaches from philosophy economics and psychology* (pp. 221–242). New York, NY: Springer, Theory and Decision Library.

Bradley, R. (2009c). Revising incomplete attitudes. *Synthese, 171*, 235–256.

Gärdenfors, P. (1988). *Knowledge in Flux. Modeling the dynamics of epistemic states*. Cambridge, MA: MIT Press.

Grüne-Yanoff, T., & Hansson, S.-O. (2009). Preference change. An introduction. In T. Grüne-Yanoff & S.-O. Hansson (Eds.), *Modelling preference change. Approaches from Philosophy economics and psychology* (pp. 1–27). New York, NY: Springer, Theory and Decision Library.

Hansson, S.-O. (1995). Changes in preferences. *Theory and Decision, 38*, 1–28.

Hansson, S.-O. (2001). *The structure of values and norms*. Cambridge, MA: Cambridge University Press.

Jeffrey, R. C. (1983). *The logic of decision*. Chicago, IL: University of Chicago Press.

McCain, K. (2008). The virtues of epistemic conservatism. *Synthese, 164*, 185–200.