Preference Change Triggered by Belief Change: A Principled Approach

Jérôme Lang[1] and Leendert van der Torre[2,⋆]

[1] Laboratoire d'Analyse et Modélisation des Systèmes pour l'Aide à la Décision (LAMSADE), Université Paris-Dauphine,
Place du Maréchal de Lattre de Tassigny, 75775 Paris, France
lang@irit.fr

[2] Computer Science and Communications, Université de Luxembourg, 6, rue Richard Coudenhove-Kalergi, 1359 Luxembourg, Luxembourg
leendert@vandertorre.com

Abstract. Various tasks need to consider preferences in a dynamic way. To evaluate and classify methods for preference change, we introduce eight properties for preferences evolving after some new fact has been learned. Four properties are concerned with persistence of preferences when something being preferred is (partly) satisfied or dissatisfied, and formalize that preference change indicates that the ideal state has not been reached or has become unreachable. Four other properties are concerned with persistence of preferences when, roughly, the agent learns something she already expected to hold, and formalizes that preference change is due to surprise. We define a family of preference change operators, parameterized by a revision function on epistemic states and a semantics for interpreting preferences over formulas, and we give conditions on the revision function and the semantics of preference for each of the eight conditions to hold.

1 Introduction

The behaviour of a rational agent — namely, the actions she decides to perform — is a function of her *beliefs* about the current state of the world and the predicted state of the world after performing such or such course of action, and of her *preferences* about those states of world she wants to bring about (or avoid to) and the actions she wants to perform (or avoid to). This classical distinction between beliefs and preferences comes from practical reasoning and decision theory, where the most common approach consists in modelling beliefs by probability distributions, and preferences by utility functions. Also in cognitive approaches one can find concepts corresponding to beliefs and preferences as they are used in decision theory, typically referring to concepts like knowledge or belief on the one hand, and preference, desire or goals on the other hand, together with other

⋆ We wish to thank the anonymous reviewers of a previous version of this paper for helpful comments. Jérôme Lang is supported by the ANR Project ANR-05-BLAN-0384. A much shorter and preliminary version of this paper appeared as [21].

G. Bonanno, B. Löwe, and W. van der Hoek (Eds.): LOFT 2008, LNAI 6006, pp. 86–111, 2010.

cognitive and social concepts like intentions, commitments, obligations, and so on. In this introduction we use this terminology "beliefs" and "preferences" with the meaning conveyed by practical reasoning and decision theory: beliefs refer to the uncertainty the agent has about the current and future states of the worlds, and preferences refer to her satisfaction when performing an action sequence and obtaining a sequence of states. For the formal framework and results in this paper, we discuss our assumptions in more detail in Section 1.3.

Thus, in contrast to some papers, we do not use "preferences" as a mere synonym of "ranking", independently of whether this ranking expresses relative plausibility of relative satisfaction. For example, Freund [13,14] investigates preference revision in the following meaning: how should an initial ranking (a "chain") over a set of worlds be revised by the addition, retraction of modification of the links of the chain? See also Chomicki [9] for a similar approach to preference revision, in a database querying context. In these papers, "preference" has to be understood as "ranking over a set of worlds" rather than its decision-theoretic sense, and the results apply indifferently whether the ranking is interpreted in terms of decision-theoretic preferences or in terms of comparative plausibility.

Our work is based on the fundamental distinction between belief and preference, since changes of preferences are often the repercussion of changes of beliefs. Beliefs are dynamic, because they change every time the agent learns something about the state of the world (notably via observations) or performs an action. The effects of learning some information or performing some action on the agent's *beliefs* has been extensively studied, not only in the classical Bayesian setting (where learning an information amounts to Bayesian conditioning) but also in logical and ordinal settings, where beliefs most often consist of rankings over worlds rather than probability distributions. Started by Alchourròn, Gärdenfors and Makinson [1], families of theory change operators have been axiomatically characterized, by means of representation theorems. There is now a huge literature on belief change (see, *e.g.*, Rott's recent survey [25]), and, to some extent, a general agreement about the meaning of the various classes of belief change operators such as revision or update.

Now, the question of whether preferences evolve, and how, is just as relevant, and yet the literature on preference change is much sparser than the one on belief change. A few recent articles have focused on this issue (see [4] for a recent overview) but much remains to be done. This particular paper provides an AGM-style approach to preference and preference change. A lot of the work in this paper is conceptual and programmatic, in the sense that we are searching for the right formulation of AGM-style axioms for preference change. Before we discuss the need for our approach, and how our approach can deal with problems that are not accessible to other approaches, we start with an overview of the work in preference change.

1.1 Kinds of Preference Change

While what "belief change" conveys is fairly agreed upon, the recent literature describes a variety of very different processes that can be called "preference

change", which roughly can be clustered in three groups, depending mainly on the nature of the mathematical object that changes and the nature of the input that leads this object to change.

"Direct" preference change, or revision of preferences by preferences

The first group consist in approaches viewing preference change as parallel to belief change: just as belief revision aims at incorporating newly acquired beliefs into an old belief state, this kind of preference change (that we may call *intrinsic preference revision*) consists in incorporating new preferences into an old preference state: *preferences are revised by preferences so as to lead to new preferences*. This kind of preference change can be modelled in a way that mirrors belief change, in the sense that preferences are revised by preferences, and lead to new preferences, without beliefs to intervene in the process. This kind of preference change has been given an in depth analysis by Hansson [17,18]. He addresses not only preference revision and contraction, but also preference addition (respectively subtraction), where preference evolves after an alternative is added to (respectively removed from) the set of alternatives. Preference change triggered by "commands" or "suggestions", as considered in van Benthem and Liu [5], can be seen as a variant of the former class of approaches, the difference being that the "input" is exogeneous: a command is an imperative from an authority ("see to it that φ!") whose effect is that the agent now prefers φ-worlds over $\neg\varphi$-worlds, and a suggestion is a milder kind of preference upgrade.

Example 1. [5] Let's take a trip!

Example 1 is a command or a suggestion, depending of whether the preference for a trip must hold in the resulting preference state or not. Van Benthem and Liu [5] build an extension of dynamic epistemic logic for reasoning both with beliefs and preferences, in which these two kinds of preference change are defined and studied. See also Girard [15] for a more extensive study of preference change in modal logics and [27] for preference-based deontic logic.

Preference change trigerred by belief change

The second group consist in approaches where *preferences change in response to changes in belief.*

Example 2. Initially, I desire to eat sushi from this plate (e). Then I learn that this sushi has been made with old fish ($\neg f$). Now I desire not to eat this sushi.

The event that trigerred the preference change does not primarily concern preference, but beliefs. Learning that the sushi was made from old fish made me believe that I could be sick, and as a consequence my initial preference for sushi has been replaced for the opposite preference.

A different kind of preference can be put in the same group: *preferences that change when the world changes*:

Example 3. It is a nice afternoon and I would like to have a walk. Then it starts to rain. I do not want to have a walk anymore.

Here the change in preferences is triggered by a change of the world, because it was not raining and now it does. Things are quite similar to change trigerred by belief revision as discussed just above, with the difference that the belief change process is not a revision, but an update [20].

Preference change as a result of belief change has been considered only recently. Bradley [8] argues that changes in preference can have two sorts of possible causes: "what might be called change in tastes" (cf. Example 5) and *change in beliefs*, where "preference change is induced by a redistribution of belief across some particular partition of the possibility space." Then he develops a Bayesian formalization of this principle. De Jongh and Liu [19] (see also Sections 3.4 and 3.5 of [24]) consider also preference change due to belief change. Preferences are induced from priorities (over formulas) and beliefs, using various possible strategies, as illustrated on their following example.

Example 4. [19,24] Alice is looking for a flat. She considers price more important than neighbourhood. She believes that flat d_1 has a low cost and is in a bad neighbourhood. She has no information about the price of flat d_2, and believes it is in a good neighbourhood.

For instance, the so-called "decision stategy" compares two alternatives by focusing on the most important criterion that one alternative is believed to satisfy and the other one is not, and here would lead to preferring d_1 over d_2. When beliefs change, preference may change as well: for instance, if Alice learns that d_2 has a low cost, she will now prefer d_2 over d_1. This preference change trigerred by belief change contrasts with preference change due to changes in her priorities (see also Example 6).

Preferences that change when the agent evolves

Example 5. [18] I grow tired of my favourite brand of mustard, A, and start to like brand B better.

Here, a change in preference reflects a modification of the agent's tastes, possibly due to an event the agent is subject to.

It could be discussed whether it is relevant to distinguish preference change due to the evolution of the rational agent to preference change due to the evolution of the world. This is primarily a choice to be made when we model the process, and it thus comes down to deciding whether the rational agent should be part of the world of not. Consider the following example from Liu [24], a variation of Example 4:

Example 6. [24] Alice is looking for a flat. She considers price more important than quality. After she wins a lottery prize of ten million dollars, she considered quality most important.

Depending on whether the agent is part of the world, this is an instance of a preference change triggered by a change in the world or by an evolution of the tastes of the agent.

1.2 Evaluating and Classifying Preference Change Methods

Our survey of the three kinds of preference change illustrates a wide variety in the kinds of preference change studied in the literature, even when we restrict ourselves to the notions of preference and belief studied in practical reasoning and decision theory. Our central research question is therefore:

How should we evaluate and classify preference change methods?

This breaks down in the following research questions:

1. Which language do we need to represent postulates of preference change?
2. Which postulates should we use to evaluate and classify methods for preference change?
3. How to use the postulates to evaluate or classify existing or new approaches to preference change, or to develop new approaches?

The success criterium of our postulates is that they are able to distinguish a variety of approaches. To illustrate our postulates and their use to evaluate or classify preference change methods, we propose a general family of operators for preferences evolving after a new fact has been learned, parameterized by a revision function on epistemic states and a semantics for interpreting preferences over formulas, and we give sufficient conditions on the revision function and the semantics of preference for each of these axioms to hold.

Our overall methodology is inspired by the so-called AGM framework of theory change [1], a formal framework to evaluate and classify change methods, originally developed as a framework to describe and classify both normative systems and belief change (though only normative system change is mentioned explicitly in [1] as an example of theory change). The AGM framework studies how a set of propositions should change in view of possibly new conflicting information, by providing a set of postulates that the new theory should satisfy. Typically there are several operators that satisfy the conditions and no solution about which one to chose is provided.

Summarizing, we are searching for the right formulation of AGM-style postulates for preference change. AGM theory respects a number of postulates which may be useful in the setting of preference change, like minimality. However, it has been used for belief change only, not for preference or norm change. We might wonder whether the AGM postulates can still be of any use for preference change. Unfortunately, they do do not seem very helpful to define properties for preference change. For example, the most often discussed postulates, like success, do not make sense in a preference change setting when the trigger of the preference change is a belief change. We are therefore looking for other postulates.

Our ultimate goal is to characterize classes of preference revision operators by AGM-like axioms. Even if the revision of preferences by beliefs has been considered in several places, there exists so far no principled study. Such a study would allow to shed light on what these operators mean. Obtaining a full characterization is an ambitious goal, that we do not aim at reaching in this paper. Another long term goal is to use our theory to develop new preference change methods.

1.3 Three Assumptions of Our Language for Preference Change Postulates

The evaluation and classification of preference change methods are not accessible to other approaches discussed thus far, because most papers aim to define a precise notion of preference and preference change by fixing the meaning of the concepts. The AGM approach to theory change can be used to evaluate and classify belief change methods, because it is based on a minimal number of assumptions: a belief base is represented by a theory, the belief change is triggered by new incoming information, and the result is again a theory. The AGM theory is abstract, there is not even a reference to belief in the AGM approach to theory change. However, it is not the case that the most general framework is the best one to evaluate and classify methods, but there is a trade-off in generality: on the one hand we want to be general to cover a large class of approaches, on the other hand specialized enough to be able to represent useful properties. For example, the AGM postulate of success implies that unsuccessful revisions cannot be represented.

In the case of preference change, we have to make some additional assumptions on the language we use to represent the postulates, though we should not limit applicability by fixing the meaning of the concepts. The first assumption we make is that we accept a distinction between beliefs and preferences. This is a very weak assumption, which holds in most models that use the notion of preference. More precisely, we assume a language in which we can talk about beliefs and preferences. These two classes can be found in decision theory by probabilities and (expected) utilities, but also in cognitive science or in agent theories in computer science, which makes our postulates generally applicable. For example, whereas in decision theory there are additional assumptions on how utility and probability can be combined to calculate expected utilities, which are not made in cognitive science, we do not make any assumptions on the way belief and preference can be combined. Moreover, whereas in cognitive approaches it is assumed that belief and preference interact with other social-cognitive concepts like intentions or obligations, which is not assumed in decision theory, we do not make such assumptions either. Summarizing, in our approach it is not important what belief or preference precisely mean, since we give a general framework to compare existing methods, and to guide the development of future preference change methods.

The second assumption of our framework is that preference change is due to belief change. In other words, we focus on the second group of approaches,

namely, preference change trigerred by belief revision. The first reason is that we find it more natural and more widely applicable than other types of preference change. What triggers changes in the mental state of an agent (hence changing her present or future behaviour) generally consists of inputs that come from the world or from other agents (via observations, communication etc.) and *primarily affects the agent's beliefs*. We do not mean that these inputs do not affect in any way the agent's preferences, but that they often do so because they change her beliefs in the first place. The second reason that we consider only change of belief, is that we think that in most cases, and in particular in the class of situations that we consider here, preferences can be assumed to be static. This is analogous to approaches in decision theory, that assume that the utility function is fixed while probabilities change.

The third assumption we make is that belief change can be appropriately represented by the AGM approach to theory change, together with some more recent extensions to deal with iterated theory change (we do not need iteration *stricto sensu*, but we need to refer to revision operators acting on belief states, *i.e.*, plausibility or normality rankings on worlds). The reason of this assumption is that we need a framework of belief change, and the AGM framework is the most generally accepted one.

1.4 Postulates

In the AGM approach to theory change, an important guideline for finding postulates is to formalize the idea of minimal change, as, for example, formulated in the Ramsey test. This is also our first guideline, in the sense that all properties we consider in this paper are of the form: if there is a preference for α, and some other conditions hold, then after learning β, the preference for α still holds. They are therefore a kind of persistence properties, explaining a kind of minimal change properties for preference change.

Moreover, our second guideline to find the properties discussed in this paper is the interaction between belief and preference change. In belief change, an important distinction between so-called revision and expansion is that the former not only adds some new facts to the belief base, but when the new information is inconsistent with the previously held beliefs, it also has to drop some beliefs. Thus, we may say that revision also handles the surprise of learning something which the agent expected to be false. In the case of preference change, our properties represent whether preference change is due to surprise, where we distinguish between weak surprise (something new is learned as handled by both expansion and revision) and strong surprise (something is learned which was believed to be false, as handled by revision only).

The first four properties (P1 to P4) consider the case in which we learn that our preferences are (partly) satisfied or dissatisfied:

$$\text{(P1)} \qquad P\alpha \rightarrow [\star\alpha]P\alpha$$

(P1) intuitively means that learning something that is preferred leaves this preference unchanged. For instance, if I already prefer to have my paper accepted

then I still prefer this state of fact to the opposite after learning that it has actually been accepted. Alternatively, as another example, if I desire to be rich, then after becoming rich, this state does not become undesirable. This persistence of preference seems natural in most contexts (except perhaps for some pathological agents who are always unhappy with what they have).

This principle should not be confused with the dropping of so-called achievement goals, once they are fulfilled. Consider an agent who has the goal to run the marathon once in her life. After she achieves this goal, we may expect her to drop the goal from her goal base (this kind of reasoning is common in planning and BDI theories of agent action, see, e.g., [10]). This, however, does not conflict with our preference persistence postulate. In our setting, preferences bear on propositions, not on actions. "Run the marathon" is an action, whereas "having already run the marathon in one's life" is a proposition (denote it by m). If the agent initially has a preference for m, and m becomes true, then she keeps preferring m over $\neg m$ (which has now become unacessible). In other words, once she has run the marathon, she's happy with this state of fact and does not wish she had never done it.

By symmetry, things are similar when revising by a dispreferred formula:

$$\text{(P2)} \quad P\alpha \rightarrow [\star\neg\alpha]P\alpha$$

Suppose now that we learn that what we want to hold, in fact *partially* holds. In that case, it would be intuitive that the preference persists.

$$\text{(P3)} \quad P\alpha \wedge \neg N(\neg\beta|\neg\alpha) \rightarrow [\star(\alpha \vee \beta)]P\alpha$$

$$\text{(P3')} \quad P\alpha \wedge \neg N(\neg\beta|\neg\alpha) \wedge \neg N(\alpha|\alpha \vee \beta) \rightarrow [\star(\alpha \vee \beta)]P\alpha$$

The following example illustrates the normality condition in (P3).

Example 7. Consider the following version of property (P3) without normality condition: (P3*) $P\alpha \rightarrow [\star(\alpha \vee \beta)]P\alpha$. Assume there is a lottery in which you can win a car, and let α be "you win a BMW in the lottery" and let β be "you win a Mercedes in the lottery." Moreover, assume that it is unlikely that you win a car, but you desire to win one without a preference of either brand. Now if you receive the information that you win a car in the lottery, then you still do not care whether it is a BMW or a Mercedes. Before, you preferred to win a BMW since you compared the alternatives of winning a BMW and winning nothing, but with this new information that you won a car, you compare the alternatives of winning a BMW and winning a Mercedes. It satisfies the property (P3), since of the normal worlds in which you do not win the BMW, there is no world in which you win the Mercedes. In other words, normally you do not win the Mercedes but you win nothing, which explains why the preference does not persist. Therefore (P3*) is too strong.

(P4) and (P4') are similar to (P3) and (P3'):

$$\text{(P4)} \quad P\alpha \wedge \neg N(\neg\beta|\alpha) \rightarrow [*(\neg\alpha \vee \beta)]P\alpha$$

$$\text{(P4')} \quad P\alpha \wedge \neg N(\neg\beta|\alpha) \wedge \neg N(\neg\alpha|\neg\alpha \vee \beta) \rightarrow [*(\neg\alpha \vee \beta)]P\alpha$$

The next four properties are concerned with the case in which we learn something we expected to hold.

$$\text{(P5)} \quad P\alpha \wedge N\beta \rightarrow [\star\beta]P\alpha$$

Property P5 is logically equivalent to the principle that *preference change implies surprise*: $P\alpha \wedge \neg[\star\beta]P\alpha \rightarrow \neg N\beta$. Note that this is only a weak notion of surprise, since it does not imply that β is exceptional, only that β is not normal (a stronger notion of surprise is considered in (P8) below). Property (P5) expresses that if we learn something we already expected to hold, then none of our preferences should change. While this property seems intuitively satisfactory, it is sometimes too strong. Consider the following example.

Example 8. Take α to be "my paper is accepted" and β to be "my paper is bad", because, for instance, the proof of the main result is flawed (and I am not good enough to detect it myself). I initially prefer α, and since I hold myself in a bad self-esteem I believe β. Now, suppose there is a strong correlation between $\neg\alpha$ and β: if my paper is bad, then it is very likely that it will be rejected. I prefer my paper to be accepted (and in case I learn this I will revise my belief that the paper is bad and start believing the opposite), but I would surely not want my paper to be accepted if the main result is false; this is consistent with β being normal, provided that states where $\beta \wedge \alpha$ hold are at least as exceptional as states where $\neg\beta$ hold.

A weaker form of (P5) is that preference for α should remain unchanged if we learn something that is normal *both given* α *and given* $\neg\alpha$.

$$\text{(P6)} \quad P\alpha \wedge N(\beta|\alpha) \wedge N(\beta|\neg\alpha) \rightarrow [\star\beta]P\alpha$$

While Example 8 conflicts with (P5), it does not conflict with (P6), because $N(\beta|\alpha)$ does not hold (if my paper is accepted then it is likely that it is good). Having β normal both given α and given $\neg\alpha$ ensures that when comparing α and $\neg\alpha$, the most normal worlds, that is, the ones I focus one, remain the same, which is a strong reason for α to remain preferred.

Another weaker form of (P5), which is when one learns something which is believed (normal), and the preference bears on a proposition α such that neither α nor $\neg\alpha$ is exceptional.

$$\text{(P7)} \quad P\alpha \wedge N\beta \wedge \neg N\alpha \wedge \neg N\neg\alpha \rightarrow [*\beta]P\alpha$$

Again, the reason why we need α and $\neg\alpha$ to be non-exceptional is that together with β being believed, it guarantees some stability of the most normal α-worlds and most normal $\neg\alpha$-worlds before and after revision by β. Consider Example 8 again; it does not contradict (P7), because $N\alpha$ does not hold ($N\beta$ and $N(\neg\alpha|\beta)$ imply $N\neg\alpha$).

This condition that both β and $\neg\beta$ are non-exceptional is intuitively desirable in many contexts, especially when β (and $\neg\beta$) refers to something that is controllable by the agent. For instance, on Example 2: $\mathcal{M} \models Pe \wedge \neg N\neg e \wedge \neg N\neg e \wedge Nf$: the agent initially believes that the fish is fresh and, of course, does not considers eating, nor not eating, as exceptional. As a result, after learning that the fish is fresh, she still prefers eating the sushi.

Now, when revising by something that *is not exceptional* (not disbelieved), we would expect some form of preservation of preference as well.

$$\text{(P8)} \quad P\alpha \wedge \neg N(\neg\beta|\alpha) \wedge \neg N(\neg\beta|\neg\alpha) \rightarrow [\star\beta]P\alpha$$

(P8) means that if α is initially preferred and is no longer preferred after learning β, then not only β should not be normal, but it should be exceptional, either given α or given $\neg\alpha$.

Again we need these two conditions $\neg N(\neg\beta|\alpha)$ and $\neg N(\neg\beta|\neg\alpha)$. Suppose for instance that $\neg N(\neg\beta|\neg\alpha)$ holds (as in Example 8). Then revising by β may change radically the meaning of $\neg\alpha$.

There are dependencies between the postulates. However, stating and proving them formally needs us to be more precise anout the semantics of normality, hence we leave this to Section 3. We just state here informally that under the usual semantics for normality, the following relationships hold: (P5) implies (P6), (P5) implies (P7) and (P8) implies (P6). (For any other pair of postulates (Pi), (Pj) than these three, (Pi) and (Pj) are independent.)

1.5 Using Our Postulates

To illustrate our postulates and their use to evaluate or classify preference change methods, we propose a general family of operators for preferences evolving after some new fact has been learned, parameterized by a revision function on epistemic states and a semantics for interpreting preferences over formulas, and we give conditions on the revision function and the semantics of preference for each of our postulates to hold. We give here an informal presentation of the operators (formal details will be given in Section 2). In order to express preference revision trigerred by belief revision we need to consider both relative plausibility between worlds (or normality) and preference between worlds in our model. While a classical decision-theoretic approach would model plausibility and preference by probability distributions and utility functions respectively, we stick here to a purely ordinal modelling, following a long-standing tradition in the belief change community. Our models consist of two orderings on a set of worlds, one for normality and one for preference, as illustrated on Figure 1. On the left hand side, the mental state of the agent is represented by two complete weak orders expressing respectively normality and preference, and new incoming information $\overline{f}$ results in the shift of $\overline{f}$ worlds towards normality, leaving the preference order unchanged. On the right hand side, the two complete weak orders are visualized more compactly by a two-dimensional structure. The striking out parts of the right hand side show the normality shift of the $\bar{f}$-worlds.

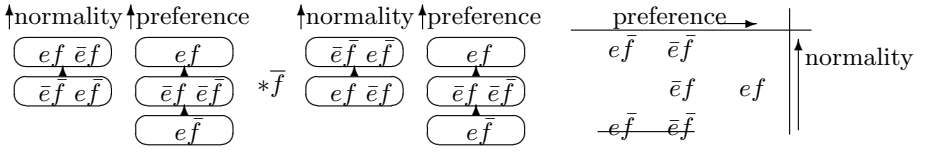


Fig. 1. Learning $\overline{f}$: $\overline{f}$ becomes most normal, preference remains unchanged

The layout of the rest of the paper is as follows. In Section 2 we define the general model for the evolution of an agent's preferences after revision by a new fact. In Section 3 we investigate the range of validity of our postulates within this family of preference change operators. We conclude and evoke further developments in Section 4.

2 Preference Change Triggered by Belief Change: A General Model

2.1 Notations

Throughout the paper we consider a propositional language $\mathcal{L}$ formed from a fixed, finite set of propositional symbols and the usual connectives (this language is enriched with modalities in the following subsection). Propositional symbols are denoted by p, q, etc. Propositional formulas of $\mathcal{L}$ are denoted by α, β, φ etc. The set of all truth assignments (or valuations) satisfying a formula φ is denoted by $\mathrm{Mod}(\varphi)$. Valuations are denoted by w, w' etc. We use the following notation for valuations: $a\bar{b}c$ denotes the valuation where a and c are assigned to true and b to false. The set of all valuations is denoted by W.

A complete weak order is a reflexive, transitive and complete relation $\succeq$ on the set of valuations. $L(W)$ denotes the set of all complete weak orders on W. The relations $\sim$ and $\succ$ are defined from $\succeq$ in the usual way: for any valuations s, s', s'', we have $s \sim s'$ iff $s \succeq s'$ and $s' \succeq s$ and $s \succ s'$ iff $s \succeq s'$ and not $(s' \succeq s)$. If $X \subseteq W$, $\mathrm{Max}_{\succeq}(X)$ is the set of maximal elements in X: $\mathrm{Max}_{\succeq}(X) = \{w \in X \mid \text{there is no } w' \text{ such that } w' \succ w\}$.

Below we shall make use of *two* complete weak orders on the set of worlds: a *normality* relation $\succeq_N$ and a *preference* relation $\succeq_P$ (where 'preference' is here employed in its decision-theoretic meaning, cf. the first paragraph of Section 1).

2.2 Beliefs and Preferences

We now consider in more detail the scenario illustrated informally on Example 2.

Whether preferences have really changed is a complicated question. This primarily depends on what we mean by "preference". The *preference relation on complete states of the worlds* remains static – only the relative plausibility of these states of the world change, and thus the agent's beliefs. Let $S = \{ef, e\bar{f}, \bar{e}f, \bar{e}\bar{f}\}$ be the set of possible states of the world. Some may argue that

e is an action rather than a static proposition. To resolve this ambiguity, just consider that e precisely refers to "being in the process of eating".

At first, it is reasonable to assume that I believe the sushi to be made out of fresh fish, or, at least, to assume that I do not believe that the fish is not fresh, even if this is not said explicitly. The reason is that if I already believed that the fish is not fresh, then the new information would have had no impact on my beliefs, and likewise, no impact on my future behaviour. After I am told that the fish is not fresh, it is reasonable to expect that my belief that the fish is fresh gets much lower.

As for my preferences, they may initially be

$$ef \succ_P \bar{e}f \sim_P \bar{e}\bar{f} \succ_P e\bar{f}$$

i.e., I prefer eating fresh sushi over not eating sushi, and I prefer not eating sushi over eating sushi made out of old fish; if I do not eat the sushi, then I do not care whether the fish is old or not. Now, *my preferences after learning that* $\neg f$ *is true or likely to be true are exactly the same.* even if I now consider ef hardly plausible, I still prefer this world to $\bar{e}f$ and $\bar{e}\bar{f}$, and these two to $e\bar{f}$. Thus, *beliefs change, but preferences remain static.*

Still, it is no less true that I used to prefer e over $\neg e$ and I no longer do. However, e and $\neg e$ are not single states, but formulas or, equivalently, sets of states (e corresponds to the set of states $\{ef, e\bar{f}\}$) and $\neg e$ to $\{\bar{e}f, \bar{e}\bar{f}\}$). When expressing an initial preference for e I mean that when I focus on those states where e is true, I see ef as the most plausible state, and similarly when I focus on those states where $\neg e$ is true, I see $\bar{e}f$ as the most plausible state. Because I prefer ef to $\bar{e}f$, I naturally prefer e to $\neg e$: in other terms, I prefer e to $\neg e$ because I prefer the most plausible state satisfying e to the most plausible state satisfying $\neg e$. Of course, after learning the information about the fish, these typical states are now $e\bar{f}$ and $\bar{e}\bar{f}$, and after focusing, I now prefer $\neg e$ to e.

One may argue also that whether preferences over states change or not is a question of language granularity. If both e and f are in the language, then preference over states do not change, but if the language contains only the propositional symbol e, then they do change, and in this case, it is not possible to express that we learn $\neg f$, therefore the only way of modeling the input is a "direct preference change": the world sends a "command" to the user, asking her to now prefer $\bar{e}$ to e.

This informal discussion on Example 2 allows us state the general principle of preference change triggered by belief change:

- the agent has some initial beliefs and preferences over possible states of the world; these preferences over states can be lifted to preferences over formulas;
- then she learns a new piece of information α about the world;
- therefore she revises her prior beliefs by α and keeps the same preference on states; however, preferences over formulas may change in reaction to the change of beliefs.

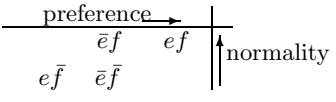
We see that a formalization needs at least two semantical structures: one for beliefs and one for preferences. For the sake of simplicity, and because we have

to start with some specific assumptions, we stick to the ordinal way of modeling beliefs and preferences, which is common in the belief change literature. Thus, as in Boutilier [7] and subsequently in Lang, van der Torre and Weydert [23], we use a normality ordering together with a preference ordering.

Definition 9. *A model $\mathcal{M}$ is a triple $\langle W, \succeq_N, \succeq_P \rangle$, where W is a set of valuations of a set of propositions, and $\succeq_N$ and $\succeq_P$ are total pre-orders on W. We do not distinguish worlds from valuations, so each valuation occurs precisely once, and W is finite.*

$s \succeq_N s'$ means that s is at least as plausible (or normal) as s', whereas $s \succeq_P s'$ means that s is at least as preferred as s'.

The model for Example 2 is visualized in the figure below. The normality ordering is visualized vertically, where higher worlds are more normal. The most normal worlds are worlds in which the fish is fresh, and exceptional worlds are worlds in which the fish is not fresh: $fe \sim_N f\bar{e} \succ_N \bar{f}e \sim_N \bar{f}\bar{e}$. Preferences are visualized horizontally, where the more to the right are the more preferred worlds. Eating fresh sushi is preferred to not eating sushi, which itself is preferred to eating not fresh sushi: $ef \succ_P \bar{e}f \sim_P \bar{e}\bar{f} \succ_P e\bar{f}$.



As in Boutilier [7] and Lang, van der Torre and Weydert [23], we extend the propositional language with two dyadic modalities: N for normality and P for preference.

As usual, $N(\psi|\varphi)$ is true if the most normal φ-worlds are ψ-worlds. $N(\varphi|\top)$ is abbreviated as $N(\varphi)$.

Definition 10 (normality)

$$\mathcal{M} \models N(\psi|\varphi) \textit{ iff } \mathrm{Max}_{\succeq_N}(\mathrm{Mod}(\varphi)) \subseteq \mathrm{Mod}(\psi)$$

Things are less easy with preference, for two reasons.

First, there are several ways of lifting preferences from the level of worlds to the level of sets of worlds. A canonical family of lifting operators is obtained by comparing two sets of worlds A and B by comparing the best (or the worst) elements of A with the best (or the worst) elements of B – this gives four ways of alternating quantifiers, henceforth, four lifting operators. This principle is classical, as it has been used in many places, including [16,2,22,5]. There are other families of lifting operators, notably *ceteris paribus* preferences [28,6] and other kinds of similarity-based preferences [18]. While these would also be worth considering, in this paper we restrict our study to the "canonical" lifting operators defined below.

strong lifting
$W_1 \gg_{\text{str}} W_2$ if $W_1 \neq \emptyset, W_2 \neq \emptyset$, and $\forall w \in W_1\ \forall w' \in W_2 : w \succ_P w'$: the worst worlds in W_1 are preferred to the best worlds in W_2, or equivalently, every world in W_1 is preferred to every world in W_2.

optimistic lifting
$W_1 \gg_{\text{opt}} W_2$ if $W_1 \neq \emptyset, W_2 \neq \emptyset$, and $\exists w \in W_1$ such that $\forall w' \in W_2, w \succ_P w'$: the best worlds in W_1 are preferred to the best worlds in W_2 (or equivalently, the best $\varphi \vee \psi$ worlds are $\neg\psi$ worlds).

pessimistic lifting
$W_1 \gg_{\text{pess}} W_2$ if $W_1 \neq \emptyset, W_2 \neq \emptyset$, and $\forall w \in W_1 \exists w' \in W_2$ such that $w \succ_P w'$: the worst worlds in W_1 are preferred to the worst worlds in W_2.

We deliberately omit to define the fourth case, corresponding to two existential quantifiers, because the resulting lifting operator is much too weak, as it makes $P\varphi \wedge P\neg\varphi$ consistent. This weak lifting operator is left out of our study.

Recall that the set of truth assignments is finite; therefore, there cannot be any infinite ascending or descending chains of worlds, and the last two definitions always make sense. An equivalent definition of $\gg_{\text{opt}}$, which does not need the finiteness assumption, is: $\forall w' \in W_2 \exists w \in W_1$ such that $w \prec_P w'$.

Second, as argued in [7,23], in the presence of uncertainty or normality expressed by $\succeq_N$, preferences cannot be interpreted from $\succeq_P$ alone, but from $\succeq_P$ and $\succeq_N$. There are at least two ways of interpreting a preference for φ over $\neg\varphi$ in this context, that we name B and LTW after their authors.[1] Let $\gg$ be one of $\gg_{\text{str}}$, $\gg_{\text{opt}}$, or $\gg_{\text{pess}}$.

B

"among the most normal ϕ-worlds, ψ is preferred to $\neg\psi$"[7]:

$$\mathcal{M} \models P_{\gg}(\psi|\varphi) \text{ iff}$$
$$\text{Max}_{\succeq_N}(\text{Mod}(\varphi)) \cap \text{Mod}(\psi) \gg \text{Max}_{\succeq_N}(\text{Mod}(\varphi)) \cap \text{Mod}(\neg\psi))$$

LTW

"the most normal $\psi \wedge \phi$-worlds are preferred to the most normal $\neg\psi \wedge \phi$-worlds" [23]:

$$\mathcal{M} \models P_{\gg}(\psi|\varphi) \text{ iff}$$
$$\text{Max}_{\succeq_N}(\text{Mod}(\varphi \wedge \psi)) \gg \text{Max}_{\succeq_N}(\text{Mod}(\varphi \wedge \neg\psi))$$

$P_{\gg}(\varphi|\top)$ is abbreviated in $P_{\gg}(\varphi)$.

Note that B and LTW are not equivalent, because either the most normal $\psi\wedge\phi$ worlds or the most normal $\neg\psi\wedge\phi$ worlds may be exceptional among the ϕ worlds. The two approaches are based on distinct intuitions. In LTW, the intuition is that an agent is comparing two alternatives, and for each alternative he is considering the most normal situations. Then he compares the two alternatives and expresses a preference of the former over the latter. The difference between

[1] Another way, for example, is to compare all worlds in the preference ranking up to minimal rank of $\text{Max}_{\succeq_N}(\psi \wedge \phi)$ and $\text{Max}_{\succeq_N}(\neg\psi \wedge \phi)$.

both approaches, already discussed in [23], is a matter of choosing the worlds to focus on. The two approaches coincide if there exist both most normal $\psi \wedge \phi$-worlds and most normal $\neg\psi \wedge \phi$-worlds, that is, if $\neg N(\psi|\phi) \wedge \neg N(\neg\psi|\phi)$ holds.

We have thus defined *six* semantics for interpreting $P(.|.)$, since we have three ways of lifting preference from worlds to formulas, and two ways of focusing on normal worlds. We denote the corresponding six modalities using the superscript B or LTW, and one of the three subscripts str, opt or pess. For instance, $P^{\mathrm{LTW}}_{\mathrm{opt}}$ refers to the semantics in [23] and the optimistic way of lifting preferences. However we shall try to avoid using these subscripts and superscripts when it is clear from the context. From the P modality we may also define a dyadic $>$ modality, where $\varphi > \psi$ means "I prefer φ to ψ", defined by

$$(\varphi > \psi) \equiv P(\varphi|(\varphi \wedge \neg\psi) \vee (\psi \wedge \neg\varphi))$$

$P(.|.)$ and $. > .$ are interdefinable (see [18]).

2.3 Revision Functions

Given a model $\mathcal{M} = \langle W, \succeq_N, \succeq_P \rangle$, its revision by belief α is a new model

$$\mathcal{M}' = \mathcal{M} \star \alpha$$

consisting of the same W, the same $\succeq_P$ (since preferences over worlds do not change), and the revision of the initial plausibility ordering $\succeq_N$ by α. This requires the prior definition of a revision function $\star$ acting on plausibility orderings. Such functions have been extensively considered in the literature of iterated belief revision (e.g., [11,26]).

Definition 11. *A* revision function $\star$ *is a mapping from* $L(W) \times \mathcal{L}$ *to* $L(W)$, i.e.*, it maps a complete weak order over* W *and a formula* α *into a complete weak order over* W*.*

For the sake of notation we note $\succeq_N^{\star\alpha}$ instead of $\succeq_N \star\alpha$.

Revision functions on plausibility orderings are usually required to obey some properties. For example, $\star$ satisfies *success* iff for every $\succeq_N$ and every satisfiable α, $\mathrm{Max}(\succeq_N^{\star\alpha}, W) \subseteq [\alpha]$. Hence, the most normal worlds after revising by α satisfy α. In the rest of the paper we need the following properties. A revision function $\star$ satisfies

- *positive uniformity* iff for any two worlds w, w' such that $w \models \alpha$ and $w' \models \alpha$, $w \succeq_N^{\star\alpha} w'$ iff $w \succeq_N w'$;
- *negative uniformity* iff for any two worlds w, w' such that $w \models \neg\alpha$ and $w' \models \neg\alpha$, $w \succeq_N^{\star\alpha} w'$ iff $w \succeq_N w'$.
- *responsiveness* iff for any two worlds w, w' such that $w \models \alpha$ and $w' \models \neg\alpha$, $w \succeq_N w'$ implies $w \succ_N^{\star\alpha} w'$.
- *stability* iff the following holds: if all most normal worlds in $\succeq_N$ satisfy α then $\succeq_N^{\star\alpha} = \succeq_N$;

– *top-stability* iff the following holds: if all most normal worlds in $\succeq_N$ satisfy α then $\mathrm{Max}(\succeq_N^{\alpha}, W) = \mathrm{Max}(\succeq_N, W)$;

Positive and negative uniformity are named respectively (CR1) and (CR2) by Darwiche and Pearl [11]. Note that success implies that $\mathcal{M} \star \alpha \models N\alpha$. Top-stability is weaker than stability, and top-stability is implied by positive uniformity together with responsiveness.

Definition 12. *Given a model $\mathcal{M} = \langle W, \succeq_N, \succeq_P \rangle$, a revision function $\star$, and a formula α, the revision of $\mathcal{M}$ by α, is the model $\mathcal{M} \star \alpha$ defined by*

$$\mathcal{M} \star \alpha = \langle W, \succeq_N^{\star\alpha}, \succeq_P \rangle$$

3 Back to the Postulates

As explained in Section 1.4, perhaps the easiest way to describe the behavior of preference change, is to aim for an AGM style representation with postulates. To do so, we use dynamic modalities to refer to revisions, as by van Ditmarsch, van der Hoek and Kooi [12] and van Benthem [3].

$$M, w \models [\star\alpha]\varphi \text{ iff } M \star \alpha, w \models \varphi$$

In Section 1 we introduced the following eight postulates that preference change may fulfill. All properties are concerned with conditions in which a preference for α persists when new information is learned. The first four properties P1-P4 consider the case in which we learn that our preferences are (partly) satisfied or dissatisfied, and the following four properties P5-P8 are concerned with the case in which we learn something which we expected.

(P1) $P\alpha \rightarrow [\star\alpha]P\alpha$
(P2) $P\alpha \rightarrow [\star\neg\alpha]P\alpha$
(P3) $P\alpha \wedge \neg N(\neg\beta|\neg\alpha) \rightarrow [\star(\alpha \vee \beta)]P\alpha$
(P3') $P\alpha \wedge \neg N(\neg\beta|\neg\alpha) \wedge \neg N(\alpha|\alpha \vee \beta) \rightarrow [\star(\alpha \vee \beta)]P\alpha$
(P4) $P\alpha \wedge \neg N(\neg\beta|\alpha) \rightarrow [*(\neg\alpha \vee \beta)]P\alpha$
(P4') $P\alpha \wedge \neg N(\neg\beta|\alpha) \wedge \neg N(\neg\alpha|\neg\alpha \vee \beta) \rightarrow [*(\neg\alpha \vee \beta)]P\alpha$
(P5) $P\alpha \wedge N\beta \rightarrow [\star\beta]P\alpha$
(P6) $P\alpha \wedge N(\beta|\alpha) \wedge N(\beta|\neg\alpha) \rightarrow [\star\beta]P\alpha$
(P7) $P\alpha \wedge N\beta \wedge \neg N\alpha \wedge \neg N\neg\alpha \rightarrow [*\beta]P\alpha$
(P8) $P\alpha \wedge \neg N(\neg\beta|\alpha) \wedge \neg N(\neg\beta|\neg\alpha) \rightarrow [\star\beta]P\alpha$

The relationships between the postulates are the following. When saying that (Pi) implies (Pj) we mean that (Pi) implies (Pj) *whatever the chosen semantics for preference, provided that the semantics for normality is fixed to the classical semantics for normality* (as defined above). We state these relationships without proof (they are straightforward).

– (P5) implies (P6) and (P7);
– (P8) implies (P6).

Any two properties (Pi) and (Pj) other than the ones above are independent.

We are now going to look for sufficient conditions, on the belief revision operator $\star$ used and the choice of the semantics for interpreting preference, for each of these postulates to be satisfied.

3.1 Preference Satisfaction (or Dissatisfaction)

We first consider (P1).

$$\text{(P1)} \quad P\alpha \rightarrow [\star\alpha]P\alpha$$

or, equivalently: if $\mathcal{M} \models P\alpha$ then $\mathcal{M} \star \alpha \models P\alpha$.

Proposition 13. *(P1) is satisfied:*

- *if $\star$ satisfies positive and negative uniformity, and*
- *for any lifting operator $\gg \in \{\gg_{\text{str}}, \gg_{\text{opt}}, \gg_{\text{pess}}\}$, with the LTW semantics.*

Proof. Positive uniformity implies that $\text{Max}(\succeq_N^{\star\alpha}, [\alpha]) = \text{Max}(\succeq_N, [\alpha])$, and negative uniformity that $\text{Max}(\succeq_N^{\star\alpha}, [\neg\alpha]) = \text{Max}(\succeq_N, [\neg\alpha])$: the most normal α-worlds are the same before and after revision by α, and similarly for the most normal $\neg\alpha$-worlds. Now, for any lifting operator, whether $P\alpha$ holds in the LTW semantics depends only on the preference between the most normal α-worlds and the most normal $\neg\alpha$-worlds, from which the result follows. Let us give the details for $\gg_{\text{opt}}$ (things are similar for the proof for $\gg_{\text{str}}$ and $\gg_{\text{pess}}$, and for any lifting operator). We have (1) $\text{Max}(\succeq_P, \text{Max}(\succeq_N^{\star\alpha} [\alpha])) = \text{Max}(\succeq_P, \text{Max}(\succeq_N, [\alpha]))$ and (2) $\text{Max}(\succeq_P, \text{Max}(\succeq_N^{\star\alpha} [\neg\alpha])) = \text{Max}(\succeq_P, \text{Max}(\succeq_N, [\neg\alpha]))$. Suppose $\mathcal{M} \models P\alpha$, *i.e.*, (3) $\text{Max}(\succeq_P, \text{Max}(\succeq_N, \alpha)) \gg_{\text{opt}} \text{Max}(\succeq_P, \text{Max}(\succeq_N, \neg\alpha))$. From (1), (2) and (3) we get $\text{Max}(\succeq_P, \text{Max}(\succeq_N^{\star\alpha}, \alpha)) \gg_{\text{opt}} \text{Max}(\succeq_P, \text{Max}(\succeq_N^{\star\alpha}, \neg\alpha))$, that is, $\mathcal{M} \star \alpha \models P\alpha$ follows. □

Positive and negative uniformity are necessary. Consider for instance the drastic revision operator that preserves the relative ranking of α-worlds and then pushes all $\neg\alpha$-worlds towards the bottom, irrespectively of their relative initial ranking: $w \succeq_N^{\star\alpha} w'$ iff (a) $w \models \alpha$, $w' \models \alpha$ and $w \succeq_N w'$; or (b) $w \models \alpha$ and $w' \models \neg\alpha$. $*$ satisfies positive uniformity, but not negative uniformity. In Figure 2 we initially have $pq \succ_N \bar{p}\bar{q} \succ p\bar{q} \succ \bar{p}q$ and $\bar{p}q \succ_P pq \succ_P \bar{p}\bar{q} \succ p\bar{q}$. After revision by p we have $pq \succ_N^{\star p} p\bar{q} \succ \bar{p}q \sim \bar{p}\bar{q}$, therefore, with the optimistic lifting we have $\mathcal{M} \models Pp$ and $\mathcal{M} \models [\star p]P\neg p$.

(P1) is meaningless or arbitrary for Boutilier's semantics, because we have the property $\neg[\star\alpha]P\alpha$ for satisfiable α: If the most normal worlds become α worlds, then the intersection of most normal worlds and $\neg\alpha$ worlds is empty. Moreover, this property also suggests that there is no other property in the spirit of (P1) we can define for B semantics.

By symmetry, things are similar when revising by a dispreferred formula:

$$\text{(P2)} \quad P\alpha \rightarrow [\star\neg\alpha]P\alpha$$

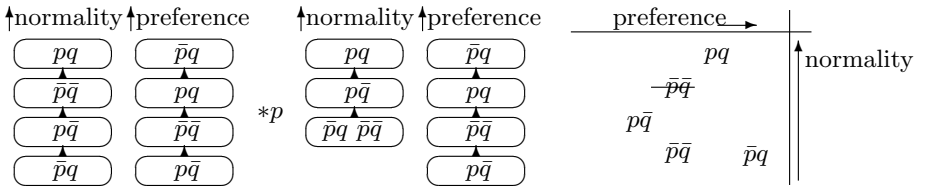


Fig. 2. Learning p: $\overline{p}$ becomes least normal, preference remains unchanged

Proposition 14. *(P2) is satisfied:*

- *if $\star$ satisfies positive and negative uniformity, and*
- *for any lifting operator with the LTW semantics.*

Proof. Positive uniformity implies that the most normal $\neg\alpha$-worlds are the same before and after revision by $\neg\alpha$, and negative uniformity that the most normal α-worlds are the same before and after revision by $\neg\alpha$. The rest of the proof is exactly as in the proof of Proposition 13. □

(P2) does not hold for B semantics, for similar reasons as (P1) does not hold: we have the property $\neg[\star\neg\alpha]\alpha$ for satisfiable $\neg\alpha$.

We now consider (P3), (P3'), (P4) and (P4').

$$\text{(P3)} \quad P\alpha \wedge \neg N(\neg\beta|\neg\alpha) \rightarrow [\star(\alpha \vee \beta)]P\alpha$$

$$\text{(P3')} \quad P\alpha \wedge \neg N(\neg\beta|\neg\alpha) \wedge \neg N(\alpha|\alpha \vee \beta) \rightarrow [\star(\alpha \vee \beta)]P\alpha$$

(P3) is equivalent to $P\alpha \wedge \neg[\star(\alpha \vee \beta)]P\alpha \rightarrow N(\neg\beta|\neg\alpha)$, which expresses that preference change in case of partial preference satisfaction is due to abnormality of β in case of $\neg\alpha$.

Proposition 15. *(P3) is satisfied:*

- *if $\star$ satisfies positive and negative uniformity, and responsiveness, and*
- *for strong or optimistic lifting with the LTW semantics.*

(P3') is satisfied:

- *if $\star$ satisfies positive and negative uniformity, and responsiveness, and*
- *for strong or optimistic lifting with the B semantics.*

Proof. Consider first the proof of (P3). By positive uniformity, $\alpha \vee \beta$-worlds are shifted uniformly when revising by $\alpha \vee \beta$. This applies in particular to α-worlds, therefore (1) the most normal α-worlds remain the same.

Assume $\mathcal{M} \models \neg N(\neg\beta|\neg\alpha)$: then at least one most normal $\neg\alpha$-world satisfies β. Let w be such a world. After revision by $\alpha \vee \beta$, w is still a most normal $\neg\alpha$-world. To see this, assume there exists a world w' such that $w' \models \neg\alpha$ and $w' \succ_N^{\star(\alpha\vee\beta)} w$.

If $w' \models \neg\alpha \wedge \beta$ then by positive uniformity, $w' \succ_N w$, which contradicts w being a most normal $\neg\alpha$-world. If $w' \models \neg\alpha \wedge \neg\beta$ then by responsiveness, $w' \succ_N w$, which again contradicts w being a most normal $\neg\alpha$-world. Analogously, for any other most normal $\neg\alpha$-world w', i.e. $w' \succeq_N^{\star(\alpha\vee\beta)} w$, if $w \models \neg\alpha \wedge \beta$ then $w' \succeq_N w$ by positive uniformity, and if $w \models \neg\alpha \wedge \neg\beta$ then $w' \succeq_N w$ by responsiveness. Therefore, (2) the set of most normal $\neg\alpha$-worlds in $\succeq_N^{\star(\alpha\vee\beta)}$ is contained in the set of most normal $\neg\alpha$-worlds in $\succeq_N$.

Therefore, if w_1 is a most normal α-world in $\succeq_N^{\star(\alpha\vee\beta)}$ and w_2 is a most normal $\neg\alpha$-world in $\succeq_N^{\star(\alpha\vee\beta)}$, then (1) implies that w_1 is a most normal α-world in $\succeq_N$, and (2) implies that w_2 be a most normal $\neg\alpha$-world in $\succeq_N$.

$\gg=\gg_{\text{str}}$

Assume $\mathcal{M} \models P\alpha$. Let $w_1 \in \text{Max}(\succeq_N^{\star(\alpha\vee\beta)}, [\alpha])$ and $w_2 \in \text{Max}(\succeq_N^{\star(\alpha\vee\beta)}, [\neg\alpha])$, which implies $w_1 \in \text{Max}(\succeq_N, [\alpha])$ and $w_2 \in \text{Max}(\succeq_N, [\neg\alpha])$. From $\mathcal{M} \models P\alpha$, we now have $w_1 \succ_P w_2$. Therefore, every most normal α-world in $\succeq_N^{\star(\alpha\vee\beta)}$ is preferred to every most normal $\neg\alpha$-world in $\succeq_N^{\star(\alpha\vee\beta)}$, that is, $\mathcal{M} \models [\star(\alpha \vee \beta)]P\alpha$.

$\gg=\gg_{\text{opt}}$

Assume (3) $\mathcal{M} \models P\alpha$. Let (4) $w_1 \in \text{Max}(\succeq_P, \text{Max}(\succeq_N^{\star(\alpha\vee\beta)}, [\alpha]))$ and (5) $w_2 \in \text{Max}(\succeq_N^{\star(\alpha\vee\beta)}, [\neg\alpha]))$. Again, from (4) and (5) we get (6) $w_1 \in \text{Max}(\succeq_N, [\alpha]))$ and (7) $w_2 \in \text{Max}(\succeq_N, [\neg\alpha])$. Suppose now that w_1 is not a most preferred world in $w_1 \in \text{Max}(\succeq_N, [\alpha]))$, that is, that there exists $w_3 \in \text{Max}(\succeq_N, [\alpha]))$ such that (8) $w_3 \succ_P w_1$. Because w_1 and w_3 are both most normal in $\succeq_N$, we have $w_1 \sim_N w_3$, which by positive uniformity (and because w_1 and w_3 both satisfy $\alpha \vee \beta$) implies $w_1 \sim_N^{\star(\alpha\vee\beta)} w_3$, which, together with (8), contradicts (4). Therefore we have (9) $w_1 \in \text{Max}(\succeq_P, \text{Max}(\succeq_N, [\alpha]))$. Now, from (3), the most preferred worlds in $\text{Max}(\succeq_N, [\alpha])$ are preferred to the most preferred worlds in $\text{Max}(\succeq_N, [\neg\alpha])$, therefore they are preferred to *all* worlds in $\text{Max}(\succeq_N, [\neg\alpha])$, which implies that $w_1 \succ_P w_2$, from which the result follows.

Consider now the proof of (P3'). Assume in addition that $\mathcal{M} \models \neg N(\alpha|\alpha \vee \beta)$, i.e. there is a $\neg\alpha \wedge \beta$ world among the most normal $\alpha \vee \beta$ worlds - let us call it w''. The proof is analogous, with the extra condition that from $\mathcal{M} \models P\alpha$ it follows that the most normal α worlds of $\mathcal{M}$ and the most normal $\neg\alpha$ worlds of $\mathcal{M}$ are among the most normal worlds of $\mathcal{M}$, and we have to prove that a similar condition holds for $M \star (\alpha \vee \beta)$. Due to positive uniformity and strong responsiveness, it follows that the most normal α worlds of $\mathcal{M} \star (\alpha \vee \beta)$ as well as w'' are among the most normal worlds of $\mathcal{M} \star (\alpha \vee \beta)$. From the inclusion of w'' it follows that the most normal $\neg\alpha$ worlds of $\mathcal{M} \star (\alpha \vee \beta)$ are among the most normal worlds of $\mathcal{M} \star (\alpha \vee \beta)$. □

Note that (P3) does not hold for the pessimistic semantics, since if the worst world used to be an $\neg\alpha$-world, then after the revision the worst world may be an α-world. Nor does it hold for the B-semantics, because after revision by $\alpha \vee \beta$ the $\neg\alpha$-worlds may disappear from the top cluster.

The case for (P4) and (P4') is similar.

$$\text{(P4)} \quad P\alpha \wedge \neg N(\neg\beta|\alpha) \rightarrow [*(\neg\alpha \vee \beta)]P\alpha$$

$$\text{(P4')} \quad P\alpha \wedge \neg N(\neg\beta|\alpha) \wedge \neg N(\neg\alpha|\neg\alpha \vee \beta) \rightarrow [*(\neg\alpha \vee \beta)]P\alpha$$

Proposition 16. *(P4) is satisfied if:*

- $\star$ *satisfies positive and negative uniformity, and responsiveness, and*
- $\star$ *= strong or pessimistic lifting with the LTW semantics.*

(P4') is satisfied if:

- $\star$ *satisfies positive and negative uniformity, and responsiveness, and*
- $\star$ *= strong or pessimistic lifting with the B semantics.*

The proof is similar to the proof of Proposition 15.

3.2 Preference Change Implies Surprise

We start by (P5).

$$\text{(P5)} \quad P\alpha \wedge N\beta \rightarrow [\star\beta]P\alpha$$

Proposition 17. *(P5) is satisfied:*

- *if* $\star$ *satisfies stability, and*
- *for any lifting operator with the LTW semantics.*

or

- *if* $\star$ *satisfies top-stability, and*
- *for any lifting operator with the B semantics.*

Proof

1. take any lifting operator with the LTW semantics. and assume that $\star$ satisfies stability. Assume $\mathcal{M} \models N\beta$. Then stability implies that $\succeq_N$ does not change after revision by β, that is, $\succeq_N^{\star\beta}=\succeq_N$. Therefore, most normal α-worlds are the same before and after revision by β, and similarly for $\neg\alpha$-worlds, from which we get that $\mathcal{M} \models P\alpha$ implies $\mathcal{M} \models [\star\beta]P\alpha$.
2. take any lifting operator with the B semantics. and assume that $\star$ satisfies top-stability. If $\mathcal{M} \models N\beta$ then all most normal worlds in $\succeq_N$ satisfy β, therefore revising by β leaves these most normal worlds (that is, $\text{Max}(\succeq_N, W)$) unchanged; since the truth of $P(.|.)$ depends only on $\text{Max}_{\succeq_N}(W)$, preferences remain unchanged after revision by β, therefore $\mathcal{M} \models P\alpha$ implies $\mathcal{M} \models [\star\beta]P\alpha$. □

Figure 3 illustrates that item 1. of the proof of Proposition 17 no longer holds if $\star$ does not satisfy stability, because revising by β may change the most normal α-worlds or the most normal $\neg\alpha$-worlds. We have $\succeq_N$: $pq \succ p\bar{q} \succ \bar{p}\bar{q} \succ \bar{p}q$; $\succeq_P$: $\bar{p}q \succ pq \succ \bar{p}\bar{q} \succ p\bar{q}$; and $\star$ such that that in $\succeq_N^{\star\beta}$, all β-worlds are ranked above all $\neg\beta$-worlds. That is: $\succeq_N^{\star q}$: $pq \succ \bar{p}q \succ p\bar{q} \succ \bar{p}\bar{q}$. Before learning q, the most normal p-world is pq and the most normal $\neg p$-world is $\bar{p}\bar{q}$, therefore $\mathcal{M} \models Pp$ for any kind of lifting. After learning q, the most normal p-world is still pq and the most normal $\neg p$-world is $\bar{p}q$, therefore $\mathcal{M} \models P\neg p$, again for any kind of lifting.

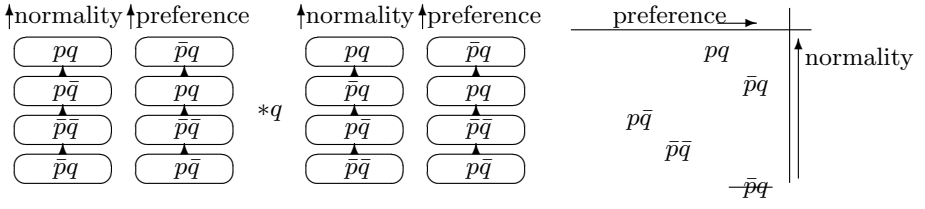


Fig. 3. Learning q: $\overline{q}$ becomes least normal, preference remains unchanged

We now consider (P6).

$$(P6) \quad P\alpha \wedge N(\beta|\alpha) \wedge N(\beta|\neg\alpha) \rightarrow [\star\beta]P\alpha$$

Proposition 18. *(P6) is satisfied:*

- *if $\star$ satisfies positive uniformity and responsiveness, and*
- *for any lifting operator with the LTW semantics.*

or

- *if $\star$ satisfies top-stability, and*
- *for any lifting operator with the B semantics.*

Proof

1. take any lifting operator with the LTW semantics, and assume that $\star$ satisfies positive uniformity and responsiveness. Moreover, assume $\mathcal{M} \models N(\beta|\alpha) \wedge N(\beta|\neg\alpha)$. Because $\mathcal{M} \models N(\beta \mid \alpha)$, the most normal α-worlds in $\succeq_N$ are $\beta \wedge \alpha$-worlds. This, together with positive uniformity, implies that (1) the most normal $\beta \wedge \alpha$-worlds are the same before and after learning β. Indeed, let $w \in \mathrm{Max}(\succeq_N, [\beta \wedge \alpha])$ and $w' \models \beta \wedge \alpha$. We have $w \succeq_N w'$, therefore, by positive uniformity, $w \succeq_N^{\star\beta} w'$. Similarly, again using positive uniformity, (2) the most normal $\beta \wedge \neg\alpha$-worlds are the same before and after learning β.

   Now, the most normal α-worlds are also most normal $\beta \wedge \alpha$-worlds (because $\mathcal{M} \models N(\beta \mid \alpha)$, therefore, these worlds remain among the most normal α-worlds after revising by β. We now have to prove that no other world can

become a most normal $\beta \wedge \alpha$-world after learning β. Let $w \in \mathrm{Max}(\succeq_N, [\alpha])$ and assume there is a $w' \in \mathrm{Max}(\succeq_N^{\star\beta}, [\alpha])$ such that $w' \notin \mathrm{Max}(\succeq_N, [\alpha])$. Either $w' \models \beta \wedge \alpha$ or $w' \models \neg\beta \wedge \alpha$. If $w' \models \beta$ then by (1), $w' \in \mathrm{Max}(\succeq_N, [\alpha])$, a contradiction. If $w' \models \neg\beta \wedge \alpha$ then $w \succ_N w'$, because $\mathcal{M} \models N(\beta \mid \alpha)$. Then, we have $w \models \beta$, $w' \models \neg\beta$ and $w \succ_N w'$, therefore by responsiveness we get $w \succ_N^{\star\beta} w'$, which contradicts $w' \in \mathrm{Max}(\succeq_N^{\star\beta}, [\alpha])$. Therefore, the most normal α-worlds before and after revision by β are the same. Similarly, we show that the most normal $\neg\alpha$-worlds before and after revision by β are the same. The result then follows.
2. consequence of the second part of Proposition 17, using the fact that (P5) implies (P6). □

We now consider (P7).

$$\text{(P7)} \quad P\alpha \wedge N\beta \wedge \neg N\alpha \wedge \neg N\neg\alpha \rightarrow [*\beta]P\alpha$$

Proposition 19. *(P7) is satisfied:*

- *if $\star$ satisfies top-stability, and*
- *for any lifting operator with the LTW semantics.*

or

- *if $\star$ satisfies top-stability, and*
- *for any lifting operator with the B semantics.*

Proof. The second part (with the B semantics) is a direct consequence of Proposition 17 together with the fact that (P5) implies (P7). As for the first part, take any lifting operator with the LTW semantics and let $\star$ satisfying top-stability. Assume $\mathcal{M} \models N\beta \wedge \neg N\alpha \wedge \neg N\neg\alpha \wedge P\alpha$. Top-stability and $\mathcal{M} \models N\beta$ imply that (1) the most normal worlds are the same in $\succeq_N$ and in $\succeq_N^{\star\beta}$. Now, all most normal worlds satisfy β; moreover, because $\mathcal{M} \models \neg N\alpha \wedge \neg N\neg\alpha$, at least one of these satisfy α and one of these satisfies $\neg\alpha$. Therefore, $\mathrm{Max}(\succeq_N, [\alpha]) = \mathrm{Max}(\succeq_N, [\beta \cap \alpha]) \subseteq \mathrm{Max}(\succeq_N, [\beta])$ and similarly $\mathrm{Max}(\succeq_N, [\neg\alpha]) \subseteq \mathrm{Max}(\succeq_N, [\beta])$. This, together with (1), implies that the most normal α-worlds are the same before and after revision by β, and similarly for $\neg\alpha$-worlds, from which the result follows. □

This condition that both β and $\neg\beta$ are non-exceptional is intuitively desirable in many contexts, especially when β (and $\neg\beta$) refers to something that is controllable by the agent. For instance, on Example 2: $\mathcal{M} \models Pe \wedge \neg N\neg e \wedge \neg N\neg e \wedge Nf$: the agent initially believes that the fish is fresh and, of course, does not considers eating, nor not eating, as exceptional. As a result, after learning that the fish is fresh, he still prefers eating the sushi.

Lastly, we consider (P8).

$$\text{(P8)} \quad P\alpha \wedge \neg N(\neg\beta|\alpha) \wedge \neg N(\neg\beta|\neg\alpha) \rightarrow [\star\beta]P\alpha$$

Proposition 20. *(P8) is satisfied:*

- *if $\star$ satisfies positive uniformity and responsiveness, and*
- *for the strong lifting operator with either the LTW or the B semantics.*

Proof

1. Take first the strong lifting operator with the LTW semantics, and assume (1) $\mathcal{M} \models \neg N(\neg\beta \mid \alpha)$, (2) $\mathcal{M} \models \neg N(\neg\beta \mid \neg\alpha)$ and (3) $\mathcal{M} \models P\alpha$. (1) implies that there exists a world w_1 in $\mathrm{Max}(\geq_N, [\alpha]) \cap [\beta]$. (2) implies that there exists a world w_2 in $\mathrm{Max}(\geq_N, [\neg\alpha]) \cap [\beta]$. Let $w_3 \in \mathrm{Max}(\geq_N^{\star\beta}, [\alpha])$, which implies $w_3 \succeq_N^{\star\beta} w_1$. Two cases:
   - $w_3 \models \beta$. In this case, $w_3 \succeq_N^{\star\beta} w_1$, together with $w_1 \models \beta$ and positive uniformity, implies $w_3 \succeq_N w_1$.
   - $w_3 \models \neg\beta$. In this case, $w_3 \succeq_N^{\star\beta} w_1$, together with $w_1 \models \beta$ and responsiveness, implies $w_3 \succeq_N w_1$.

   Therefore, $w_3 \succeq_N w_1$. Together with $w_1 \in \mathrm{Max}(\geq_N, [\alpha])$, this implies (4) $w_3 \in \mathrm{Max}(\geq_N, [\alpha])$.

   Similarly, let $w_4 \in \mathrm{Max}(\geq_N^{\star\beta}, [\neg\alpha])$, then we show in the very same way (using (2) instead of (1)) that (5) $w_4 \in \mathrm{Max}(\geq_N, [\neg\alpha])$.

   Lastly, from (3), (4) and (5) we get $w_3 \succ_P w_4$. This being true for any $w_3 \in \mathrm{Max}(\geq_N^{\star\beta}, [\alpha])$ and any $w_4 \in \mathrm{Max}(\geq_N^{\star\beta}, [\neg\alpha])$, we conclude that $\mathcal{M} \models [\star\beta]P\alpha$.
2. Take now the strong lifting operator with the B semantics, and assume (1), (2) and (3) hold. Again, (1) and (2) imply $\mathrm{Max}(\geq_N, [\alpha \wedge \beta]) \neq \emptyset$ and $\mathrm{Max}(\geq_N, [\neg\alpha \wedge \beta]) \neq \emptyset$. Moreover, let $w \in \mathrm{Max}(\succeq_N^{\star\beta}) \cap [\alpha]$ and $w' \in \mathrm{Max}(\succeq_N^{\star\beta}) \cap [\neg\alpha]$. From positive uniformity, responsiveness, and the nonemptiness of $\mathrm{Max}(\geq_N, [\alpha \wedge \beta])$ and of $\mathrm{Max}(\geq_N, [\neg\alpha \wedge \beta])$ (which follows from (1) and (2) respectively), we have that $w \in \mathrm{Max}(\geq_N, [\alpha \wedge \beta])$ and $w' \in \mathrm{Max}(\geq_N, [\neg\alpha \wedge \beta])$; from (3) we have $w \succ_P w'$, and the result follows. □

However this no longer holds with the other kinds of lifting, as can be seen on the following example: $\succeq_N$: $pq \sim p\bar{q} \succ \bar{p}q \sim \bar{p}\bar{q}$ and $\succeq_P$: $p\bar{q} \succ \bar{p}q \succ pq \succ \bar{p}\bar{q}$. We have $\mathcal{M} \models Pp$ for any of $\gg\!=\!\gg_{\mathrm{opt}}$ or $\gg\!=\!\gg_{\mathrm{pess}}$. After learning q, for any "reasonable" revision operator $\star$, including drastic revision, we have $pq \succ_N^{\star q} p\bar{q}$ and $\bar{p}q \succ \bar{p}\bar{q}$. Therefore, the most normal p-world is pq and the most normal $\neg p$-world is $\bar{p}q$, which implies that we have $\mathcal{M} \models [\star q](P\neg p \wedge \neg Pp)$.

4 Conclusion

There is a wide variety in the kinds of preference change studied in the literature, even when we restrict ourselves to the notions of preference and belief studied in practical reasoning and decision theory. Since the AGM approach to theory change can be used to evaluate and classify belief change methods, because it is based on a minimal number of assumptions, we propose an analogous approach to evaluate and classify preference change methods. We assume a distinction

between beliefs and preferences, without assuming that they can be combined (like probabilities and utilities can be combined in expected utility) or extended with other concepts (like beliefs and desires can be extended with intentions in cognitive theories). Moreover, we assume that preference change is due to belief change, because we find it more natural and more widely applicable than other approaches we discussed in the introduction of this paper, and because we think that in most cases, preferences can be assumed to be static (like the utility function is fixed while probabilities change). Finally, we assume that belief change can be appropriately represented by the AGM approach to theory change, together with some more recent extensions to deal with iterated theory change, because the AGM framework is the most generally accepted one for belief change.

We introduce a standard language to represent postulates for preference change triggered by belief change, based on a dyadic modal operator for normality or belief operator, represented by $N(\alpha \mid \beta)$ for "α is normal or believed given β," and $P(\alpha \mid \beta)$ for "α is preferred given β." Moreover, to represent the updates, we extend this modal language with an update operator, represented by $[*\alpha]\beta$ for "after learning the new information α, β holds."

We introduce the following eight postulates to evaluate and classify preference change methods. All properties are concerned with conditions in which a preference for α persists when new information is learned. The first four properties P1-P4 consider the case in which we learn that our preferences are (partly) satisfied or dissatisfied, and the following four properties P5-P8 are concerned with the case in which we learn something which we expected or which did not surprise us.

(P1) $P\alpha \rightarrow [\star\alpha]P\alpha$
(P2) $P\alpha \rightarrow [\star\neg\alpha]P\alpha$
(P3) $P\alpha \wedge \neg N(\neg\beta|\neg\alpha) \rightarrow [\star(\alpha \vee \beta)]P\alpha$
(P3') $P\alpha \wedge \neg N(\neg\beta|\neg\alpha) \wedge \neg N(\alpha|\alpha \vee \beta) \rightarrow [\star(\alpha \vee \beta)]P\alpha$
(P4) $P\alpha \wedge \neg N(\neg\beta|\alpha) \rightarrow [*(\neg\alpha \vee \beta)]P\alpha$
(P4') $P\alpha \wedge \neg N(\neg\beta|\alpha) \wedge \neg N(\neg\alpha|\neg\alpha \vee \beta) \rightarrow [*(\neg\alpha \vee \beta)]P\alpha$
(P5) $P\alpha \wedge N\beta \rightarrow [\star\beta]P\alpha$
(P6) $P\alpha \wedge N(\beta|\alpha) \wedge N(\beta|\neg\alpha) \rightarrow [\star\beta]P\alpha$
(P7) $P\alpha \wedge N\beta \wedge \neg N\alpha \wedge \neg N\neg\alpha \rightarrow [*\beta]P\alpha$
(P8) $P\alpha \wedge \neg N(\neg\beta|\alpha) \wedge \neg N(\neg\beta|\neg\alpha) \rightarrow [\star\beta]P\alpha$

Moreover, we show how to use our postulates to evaluate and classify preference change methods. We define a family of operators for preferences evolving after some new fact has been learned, parameterized by a revision function on epistemic states and a semantics for interpreting preferences over formulas. Moreover, we give conditions on the revision function and the semantics of preference for each of these conditions to hold, as listed in Table 1. Roughly, all of them hold for LTW semantics under some conditions, whereas (P1) and (P2) are not meaningful for B semantics, whereas some of the others need stronger or other conditions.

Summarizing, in this paper we have given an investigation of the properties of preference change in reponse to belief change, depending on the choice of

Table 1. Results for some operators: PU = positive uniformity, NU = negative uniformity, R = responsiveness, S = stability, TS = top-stability

	LTW		B	
(P1)	PU, NU	$\gg_{\text{str}}, \gg_{\text{opt}}, \gg_{\text{pess}}$	not applicable	
(P2)	PU, NU	$\gg_{\text{str}}, \gg_{\text{opt}}, \gg_{\text{pess}}$	not applicable	
(P3)	PU, NU, R	$\gg_{\text{str}}, \gg_{\text{opt}}$		
(P3')			PU, NU, R	$\gg_{\text{str}}, \gg_{\text{opt}}$
(P4)	PU, NU, R	$\gg_{\text{str}}, \gg_{\text{pess}}$		
(P4')			PU, NU, R	$\gg_{\text{str}}, \gg_{\text{pess}}$
(P5)	S	$\gg_{\text{str}}, \gg_{\text{opt}}, \gg_{\text{pess}}$	TS	$\gg_{\text{str}}, \gg_{\text{opt}}, \gg_{\text{pess}}$
(P6)	PU, R	$\gg_{\text{str}}, \gg_{\text{opt}}, \gg_{\text{pess}}$	TS	$\gg_{\text{str}}, \gg_{\text{opt}}, \gg_{\text{pess}}$
(P7)	TS	$\gg_{\text{str}}, \gg_{\text{opt}}, \gg_{\text{pess}}$	TS	$\gg_{\text{str}}, \gg_{\text{opt}}, \gg_{\text{pess}}$
(P8)	PU, R	$\gg_{\text{str}}$	PU, R	$\gg_{\text{str}}$

a revision operator and the choice of a semantics of semantics for preference. Even if we have obtained sufficient conditions for several significant properties of preference change, what is still missing is a series of representation theorems of the form: this list of properties is satisfied *if and only if* $\star$ satisfies this set of properties and $\gg$ this other set of properties. Obtaining such a result is a long-term goal due to the high number of parameters that can vary.

References

1. Alchourròn, C., Gärdenfors, P., Makinson, D.: On the Logic of Theory Change: Partial Meet Functions for Contraction and Revision. J. Symb. Log. 50 (1985)
2. Barberà, S., Bossert, W., Pattanaik, P.: Ranking Sets of Objects. In: Barberà, S., Hammond, P., Seidl, C. (eds.) Handbook of Utility Theory, pp. 895–978. Kluwer Academic Publishers, Dordrecht (2004)
3. van Benthem, J.: Dynamic Logic for Belief Revision. J. Appl. Non-Class. Log. 17(2), 129–156 (2007)
4. van Benthem, J.: For Better or for Worse: Dynamic Logics of Preference. In: Grüne-Yanoff, T., Hansson, S.O. (eds.) Preference Change. Approaches from Philosophy, Economics and Psychology. Theory and Decision Library A, vol. 42, pp. 57–84. Springer, Netherlands (2009)
5. van Benthem, J., Liu, F.: Dynamic Logic of Preference Upgrade. J. Appl. Non-Class. Log. 17(2), 157–182 (2007)
6. van Benthem, J., Roy, O., Girard, P.: Everything Else Being Equal: A Modal Logic Approach to *ceteris paribus* Preferences. J. Philos. Log. 38(1), 83–125 (2009)
7. Boutilier, C.: Toward a Logic for Qualitative Decision Theory. In: Doyle, J., Sandewall, E., Torasso, P. (eds.) KR 1994, pp. 75–86. Morgan Kaufmann, San Francisco (1994)
8. Bradley, R.: The Kinematics of Belief and Desire. Synthese 156(3), 513–535 (2007)
9. Chomicki, J.: Database Querying under Changing Preferences. Ann. Math. Artif. Intell. 50(1-2), 79–107 (2007)
10. Cohen, P., Levesque, H.: Intention is Choice + Commitment. Artif. Intell. 42(2-3), 213–261 (1990)

11. Darwiche, A., Pearl, J.: On the Logic of Iterated Belief Revision. Artif. Intell. 89, 1–29 (1997)
12. van Ditmarsch, H., van der Hoek, W., Kooi, B.: Dynamic Epistemic Logic. Synthese Library, vol. 337. Springer, Heidelberg (2007)
13. Freund, M.: On the Revision of Preferences and Rational Inference Processes. Artif. Intell. 152(1), 105–137 (2004)
14. Freund, M.: Revising Preferences and Choices. J. Math. Econ. 41, 229–251 (2005)
15. Girard, P.: Modal Logic for Belief and Preference Change. Ph.D. thesis, Stanford University, ILLC Publications DS-2008-04 (2008)
16. Halpern, J.: Defining Relative Likelihood Inpartially Ordered Preferential Structures. J. Artif. Intell. Res. 7, 1–24 (1997)
17. Hansson, S.O.: Changes in Preferences. Theory and Decision 38, 1–28 (1995)
18. Hansson, S.O.: The structure of Values and Norms. Cambridge University Press, Cambridge (2001)
19. de Jongh, D., Liu, F.: Optimality, Belief and Preference. In: Artemov, S., Parikh, R. (eds.) Proceedings of the Workshop on Rationality and Knowledge. ESSLLI, Malaga (2006)
20. Katsuno, H., Mendelzon, A.: On the Difference Between Updating a Knowledge Base and Revising it. In: Allen, J.F., Fikes, R., Sandewall, E. (eds.) KR 1991, pp. 387–394. Morgan Kaufmann, San Francisco (1991)
21. Lang, J., van der Torre, L.: From Belief Change to Preference Change. In: Ghallab, M., Spyropoulos, C.D., Fakotakis, N., Avouris, N.M. (eds.) ECAI 2008, pp. 351–355. IOS Press, Amsterdam (2008)
22. Lang, J., van der Torre, L., Weydert, E.: Utilitarian Desires. J. Auton. Agents Multi-Agent Syst. 5, 329–363 (2002)
23. Lang, J., van der Torre, L., Weydert, E.: Hidden Uncertainty in the Logical Representation of Desires. In: Gottlob, G., Walsh, T. (eds.) IJCAI 2003, pp. 685–690. Morgan Kaufmann, San Francisco (2003)
24. Liu, F.: Changing for the Better. Preference Dynamics and Agent Diversity. Ph.D. thesis, Universiteit van Amsterdam, ILLC Publications DS-2008-02 (2008)
25. Rott, H.: Change, Choice and Inference: A Study of Belief Revision and Nonmonotonic Reasoning. Oxford University Press, Oxford (2001)
26. Rott, H.: Shifting Priorities: Simple Representations for Twenty-seven Iterated Theory Change Operators. In: Makinson, D., Malinowski, J., Wansing, H. (eds.) Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg. Uppsala Philosophical Studies, vol. 53, pp. 359–384. Uppsala Universitet (2006)
27. van der Torre, L.: Reasoning About Obligations: Defeasibility in Preference-based Deontic Logic. Ph.D. thesis, Erasmus University Rotterdam (1997)
28. von Wright, G.: The Logic of Preference. Edinburgh University Press (1963)