From Belief Change to Preference Change

Jérôme Lang[1] and **Leendert van der Torre**[2]

Abstract. Various tasks need to consider preferences in a dynamic way. We start by discussing several possible meanings of preference change, and then focus on the one we think is the most natural: preferences evolving after some new fact has been learned. We define a family of such preference change operators, parameterized by a revision function on epistemic states and a semantics for interpreting preferences over formulas. We list some natural properties that this kind of preference change should fulfill and give conditions on the revision function and the semantics of preference for each of these properties to hold.

1 Introduction

Analyzing games requires a formal modelling of preferences, when the behaviour of a rational agent is a function of her beliefs and her preferences about the possible consequences of her actions. The mental state of an agent evolves in the course of the game, and these changes result also in a change of the agent's *preferences*. Quoting Liu [15], "preferences are not static, but they change through commands of moral authorities, suggestions from friends who give good advice, or just changes in our own evaluation of worlds and actions." The effects of learning some information or performing some action on the agent's *beliefs* has been extensively studied in the last two decades. Yet, while there is a huge literature on belief change and, to some extent, a general agreement about the meaning of the various classes of belief change operators such as revision or update, the literature on preference change is sparser.

In this paper we argue that the difficulty is that whereas belief change processes can reasonably be considered independent of an agent's preferences, it is generally not true that a preference change process is independent of the agent's beliefs. 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. A second difficulty is that "preference change" conveys more ambiguity than belief change, suggesting that the variety of processes being covered by preference change might be larger than that covered by belief change.

After discussing briefly the different meanings that "preference change" may convey, we focus on one of the most natural interpretations of preference change, namely, the evolution of an agent's preferences after revision by a new fact (or belief), and we give technical developments. We end the paper by discussing related work and further research directions.

[1] IRIT, Université Paul Sabatier, 31062 Toulouse, France; lang@irit.fr
[2] Université de Luxembourg; leendert@vandertorre.com

2 Notations

Throughout the paper we consider a propositional language formed from a fixed, finite set of propositional symbols and the usual connectives. This language will be enriched with modalities in Section 4. The set of all truth assignments (or valuations) satisfying a formula φ is denoted by $Mod(\varphi)$. 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. A *preference relation* is a weak order $\succeq$ (that is, a reflexive, transitive and complete relation, also called a total preorder) on the set of valuations. The relations $\sim$ and $\succ$ are defined from $\succeq$ in the usual way: for valuations s, s', s'', we have $s \sim s'$ if $s \succeq s'$ and $s' \succeq s$, and $s \succ s'$ if $s \succeq s'$ and not $(s' \succeq s)$. If $X \subseteq W$, then $Max_{\succeq}(X)$ is the set of maximal elements in X: $Max_{\succeq}(X) = \{w \in X \mid \text{there is no } w' \text{ such that } w' \succ w\}$.

3 Various kinds of preference change

We distinguish several kinds of preference change, depending mainly on the nature of the mathematical object that changes and the nature of the input that leads this object to change.

Example 1 *Initially, I desire to eat sushi from this plate. Then I learn that these sushi have been made with old fish. Now I desire not to eat any of these sushi.*

This is an instance of *preferences that change when beliefs are revised*. Learning that the sushi were made from old fish made me believe that I could be sick, and as a consequence I change my mind about my future behaviour. I will choose the action "doing nothing" rather than the action "eat".

Whether preferences have really changed is a complicated question. This primarily depends on what we mean by "preference". On the one hand, the *preference relation on complete states of the worlds* remains static – only the relative plausibility of these states of the world (and thus the agent's beliefs) change. Let $S = \{ef, e\bar{f}, \bar{e}f, \bar{e}\bar{f}\}$ be the set of possible states of the world.[3] At first, it is reasonable to assume (even if this is not said explicitly) that I believe the sushi to be made out of fresh fish — or, at least, that I do not believe that the fish is not fresh. 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

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

Now, *my preferences after learning that $\neg f$ is true or likely to be true are exactly the same*: I still prefer ef (even if I now consider

[3] 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".

this world hardly plausible) to $\bar{e}f$ and $\bar{e}\bar{f}$, and these two to $e\bar{f}$. Thus, *beliefs change, but preferences remain static*. Still, 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.g., 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.[4]

One may also argue 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.

The process that we have explained here on an example will be formalized in Section 4.

Example 2 *It is a nice afternoon and I'd like to have a walk. Then it starts to rain. I don't want to have a walk anymore.*

This is an instance of *preferences that changes when the world changes*: the preference change is triggered by a change of the world (it was not raining and now it does). Things are quite similar to the previous situation, with the difference that the belief change process is not a revision, but an update. Again, we argue that preference over states do not change (I prefer walking under the sun to not walking, and not walking to walking in the rain); only the state of world, and of course the agent's belief about the state of the world, do.

Example 3 *[11] I grow tired of my favourite brand of mustard, A, and start to like brand B better.*

This is an instance of *preferences that change when the agent evolves*. 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, as thus comes down to decide whether the rational agent should be part of the world of not. Consider the following example from [15]:

Example 4 *[15] 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. 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. Other examples can be found in [3], who consider preference change triggered by "commands" or "suggestions".A command is an input from an authority ("see to it that φ!") whose effect is that the agent now prefers φ-worlds over $\neg\varphi$-worlds. A suggestion is a milder kind of preference upgrade.

Example 5 *[3] Let's take a trip!*

See [3] for a dynamic epistemic logic formalization of preference upgrade via commands and suggestions.

4 Preference change triggered by belief revision

4.1 Beliefs and preferences

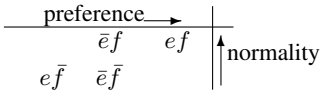
We now consider in more detail the scenario illustrated informally on Example 1. The general principle is the following: the agent has some initial beliefs and preferences over possible states of the world; these preferences over states can be lifted to preferences over formulas; then she learns a new piece of information α about the world; 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. Because one has to make choices, we stick to the ordinal way of modeling beliefs and preferences (which is common in the belief change literature). Thus, as in [4] and subsequently in [14], we use a normality ordering together with a preference ordering.

Definition 1 *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.*

$s \succeq_N s'$ 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 1 is visualized on 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 [4, 14], 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 2 (normality)

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

Things are less easy with preference, for two reasons.

First, there is no standard way of lifting preferences from the level of worlds to the level of sets of worlds (see, e.g., [12, 13]). We consider first the three following ways of lifting:[5]

[4] A related interpretation of this example, more in accordance with decision theory, consists in seeing e and $\neg e$ as actions (or as the postconditions of actions, as pointed out by an anonymous referee): my future behaviour (that is, the action that I intend to do) has changed, but my preference between states has not. This process is well-known in decision theory: after learning something, probabilities change, utilities of consequences remain unchanged but the expected utility of actions (that depend both on the probability of states and the utility of consequences) change.

[5] There is obviously a fourth one, corresponding to two existential quantifiers; however, this notion is much too weak, as it makes $P\varphi \wedge P\neg\varphi$ consistent. As suggested by a referee, we may also consider combinations of optimistic and pessimistic semantics. Alternative ways of lifting preference would also be worth considering, such as, for instance, *ceteris paribus* preferences [17] of other kinds of similarity-based preferences [11].

strong semantics $W_1 \gg_{str} W_2$ 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 semantics $W_1 \gg_{opt} W_2$ 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[6].

pessimistic semantics $W_1 \gg_{pess} W_2$ 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.

Second, as argued in [4, 14], 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. Let $\gg$ be one of $\gg_{str}$, $\gg_{opt}$, or $\gg_{pess}$.

1. "among the most normal ϕ-worlds, ψ is preferred to $\neg\psi$"[4]: $\mathcal{M} \models P_{\gg}(\psi|\varphi)$ iff $Max_{\succeq_N}(Mod(\varphi)) \cap Mod(\psi) \gg Max_{\succeq_N}(Mod(\varphi)) \cap Mod(\neg\psi))$.
2. "the most normal $\psi \wedge \phi$-worlds are preferred to the most normal $\neg\psi \wedge \phi$-worlds" [14]: $\mathcal{M} \models P_{\gg}(\psi|\varphi)$ iff $Max_{\succeq_N}(Mod(\varphi \wedge \psi)) \gg Max_{\succeq_N}(Mod(\varphi \wedge \neg\psi))$.

$P(\varphi|\top)$ is abbreviated in $P(\varphi)$.

Note that 1. and 2. 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.[7] They 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 6 modalities using the superscript B (for item 1. above) or LTW (for item 2. above), and one of the three subscripts str, opt or $pess$. For instance, P_{opt}^{LTW} refers to the semantics in [14] and the optimistic way of lifting preferences. However we will try to avoid using these subscripts and superscripts whenever possible.[8]

4.2 The impact of belief revision on preferences

4.2.1 Revising a pre-order

Given a model $\mathcal{M} = \langle W, \succeq_N, \succeq_P \rangle$, its revision by belief α is a new model $\mathcal{M}' = \mathcal{M} \star \alpha$ consists 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., [6, 16]).

[6] Recall that the set of truth asignments is finite; therefore, there cannot be any infinite ascending chains of worlds, anf our definition always makes sense. An equivalent definition, which does not need the finiteness assumption, is: $\forall w' \in W_2 \exists w \in W_1$ such that $w \prec_P w'$.

[7] The two approaches are be based on distinct intuitions. In 2., 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 both approaches (already discussed in [14]) is a matter of choosing the worlds to focus on.

[8] 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 [11]).

Definition 3 *A* revision function $\star$ *maps each complete weak order over W and each 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. In the rest of the paper we need the following ones. A revision function $\star$ satisfies

- *acceptance* if for every $\succeq_N$ and every satisfiable α, $Max(\succeq_N^{\star\alpha}, W) \subseteq [\alpha]$ (most normal worlds after revising by α satisfy α).
- *positive uniformity* (called (CR1) in [6]) if for any two worlds w, w' such that $w \models \alpha$ and $w' \models \alpha$, $w \succ_N^{\star\alpha} w'$ iff $w \succ_N w'$;
- *negative uniformity* (called (CR2) in [6]) if for any two worlds w, w' such that $w \models \neg\alpha$ and $w' \models \neg\alpha$, $w \succ_N^{\star\alpha} w'$ iff $w \succ_N w'$.
- *weak (resp. strong) responsiveness* if for any two worlds w, w' such that $w \models \alpha$ and $w' \models \neg\alpha$ then $w \succeq_N w'$ implies $w \succeq_N^{\star\alpha} w'$ (resp. $w \succ_N^{\star\alpha} w'$).

Definition 4 *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*

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

4.2.2 AGM style postulates

Perhaps the easiest way to describe the behavior of the preference change, is to aim for an AGM style representation with postulates. To do so, we use dynamic modalities to refer to revisions, as in [7, 2].

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

We are now going to look into the logical properties of preference change under newly learned beliefs (that is, the relationships between $\mathcal{M}$ and $\mathcal{M} \star \alpha$), depending on the belief revision operator $\star$ used and the choice of the semantics for interpreting preference.

For readability we only give the properties for unconditional preferences like $P(\alpha)$, but they can be extended to conditional ones like $P(\alpha|\beta)$ in a straightforward way.

4.2.3 Preference satisfaction (or dissatisfaction)

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

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

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

Proposition 1 *(P1) is satisfied:*

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

Let us give a quick proof. Because $\star$ is positively (resp. negatively) uniform, the most normal α-worlds (resp. $\neg\alpha$-worlds) are the same before and after revision by α. Now, for any lifting operator, whether $P\alpha$ holds in the LTW semantics depends only on the preference between most normal α-worlds and most normal $\neg\alpha$-worlds, from which the result 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. Suppose 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) does not hold either for Boutilier's semantics, because $[\alpha]$ or $[\neg\alpha]$ may become empty after revision by α.

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

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

Proposition 2 *(P2) is satisfied:*

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

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:

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

Proposition 3 *(P3) is satisfied:*

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

The proof goes as follows. By positive uniformity, $\alpha \vee \beta$-worlds are shifted uniformly. This applies in particular to α-worlds, therefore the most normal α-worlds remain the same. Because $\mathcal{M} \models \neg N(\neg\beta|\neg\alpha)$, at least one most normal $\neg\alpha$-world w satisfies β. After revision by $\alpha \vee \beta$, this world is still a most normal $\neg\alpha$-world. To see this, assume $w' \succ_N^{\star(\alpha\vee\beta)} w$. 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 weak responsiveness, $w' \succ_N w$, which again contradicts w being a most normal $\neg\alpha$-world. Therefore, 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 $\succeq_N$. From this we get that $\mathcal{M} \models (\alpha \gg_{str} \neg\alpha) \rightarrow [\star(\alpha \vee \beta)](\alpha \gg_{str} \neg\alpha)$, and similarly for $\gg_{opt}$.

Note that it does not hold for the pessimistic semantics, since if the worst world used to be a $\neg\alpha$-world, after the revision the worst world may be a α-world. It does not hold either for the B-semantics, because after revision by $\alpha \vee \beta$ the $\neg\alpha$-worlds may disappear from the top cluster.

Symmetrically, we may consider the following.

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

Proposition 4 *(P4) is satisfied if:*

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

4.2.4 Preference change implies surprise

Whereas in belief revision, learning a fact that is not disbelieved does not affect the old beliefs, we may wonder whether newly learned beliefs which are not exceptional do not change the preferences. However, this holds only under the assumption that the normality ordering remains the same when we revise by a normal formula:

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

Proposition 5 *(P5) is satisfied:*

- *if $\star$ satisfies* stability*: if all most normal worlds in $\succeq_N$ satisfy α then $\succeq_N^{\star\alpha}$=$\succeq_N$;*
- *for any lifting operator with the LTW semantics.*

or

- *if $\star$ satisfies* top-stability*: if all most normal worlds in $\succeq_N$ satisfy α then $Max(\succeq_N^{\alpha}, W) = Max(\succeq_N, W)$;*
- *for any lifting operator with the B semantics.*

Note that top-stability is implied by positive uniformity and weak responsiveness.

In the first case, the validity of $N\alpha \wedge P\varphi \rightarrow [*\alpha]P\varphi$ comes simply from the fact that $\succeq_N$ does not change after revision by α. In the second case , the fact that $N\alpha$ is true implies that all most normal worlds satisfy α, therefore revising by α leaves these most normal worlds (that is, $Max_{\succeq_N}(W)$) unchanged; since the truth of $P(.|.)$ depends only on $Max_{\succeq_N}(W)$, preferences remain unchanged.

However, 1. no longer holds if $\star$ does not satisfy stability, because revising by α may have an impact on the most normal β-worlds or on the most normal $\neg\beta$-worlds (but never on both). For example: $\succeq_N$: $pq \succ p\bar{q} \succ \bar{p}\bar{q} \succ \bar{p}q$; $\succeq_P$: $\bar{p}q \succ pq \succ \bar{p}\bar{q} \succ p\bar{q}$; and $\star$ such that that in $\succeq_N^{\star\alpha}$, all α-worlds are ranked above all $\neg\alpha$-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.

A weaker form of the previous property is that preference for φ should remain unchanged if we learn something that is normal *both given φ and given $\neg\varphi$*:

$$(P6) \quad N(\alpha|\varphi) \wedge N(\alpha|\neg\varphi) \wedge P\varphi \rightarrow [*\alpha]P\varphi$$

Proposition 6 *(P6) is satisfied:*

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

or

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

The proof is easy: when $N(\alpha|\varphi) \wedge N(\alpha|\neg\varphi)$ holds, the most normal φ-worlds are $\alpha \wedge \varphi$-worlds and the most normal $\neg\varphi$-worlds are $\alpha \wedge \neg\varphi$-worlds, therefore, the most normal φ-worlds remain the same after learning α, and similarly for the most normal $\neg\varphi$-worlds.

Still a stronger form of (1) which is incomparable with (2) is when one learns something which is believed (normal) and the preference bears on something which is not exceptional.

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

Proposition 7 *(P7) is satisfied:*

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

Indeed, the most normal φ-worlds are also α-worlds and hence remain the same after learning α, and similarly for the most normal $\neg\varphi$-worlds. This condition that both φ and $\neg\varphi$ are non-exceptional is intuitively desirable in many contexts, especially when φ (and $\neg\varphi$) refers to something that is controllable by the agent. For instance, on Example 1: $\mathcal{M} \models Pe \wedge \neg N\neg e \wedge \neg N\neg e \wedge Nf$: 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.

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

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

Proposition 8 *(P8) is satisfied:*

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

This holds because at least one most normal $\alpha \wedge \varphi$-world remains in the set of most normal $\alpha \wedge \varphi$-worlds after learning α.

However this no longer holds with the other kinds of lifting, as it 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 $\gg = \gg_{opt}$ or $\gg = \gg_{pess}$. After learning q, for any "reasonable" revision operator $\star$, including drastic revision, $pq \succ_N^{\star q} p\bar{q}$ 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 $\mathcal{M} \models [\star q](P\neg p \wedge \neg Pp)$.

5 Related research

Preference change was be given an in depth analysis by Hansson [10, 11], who defines preference change in a way that is parallel to belief change: preferences are revised by preferences so as to lead to new preferences. He addresses not only preference revision and contraction, but also preference addition (resp. subtraction), where preference evolve after an alternative is added to (resp. removed from) the set of alternatives.

Preference change as a result of belief change has only been considered only recently. Bradley [5] argues that changes in preference can have two sorts of possible causes: "what might be called change in tastes" (cf. Example 3) 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. Starting from similar intuitions, our work goes in another direction than ours and connects the interaction between belief change and preference change to the existing body of research in belief revision.

Liu [15] also considers preference change due to belief change, that she contrasts with preference change due to changes in her priorities (see Example 4). Then she goes in another direction than ours, by building an extension of dynamic epistemic logic for reasoning both with beliefs and preferences. Van Benthem and Liu [3] discuss and study two kinds of preference change in a DEL setting, namely preference upgrade via commands and suggestions. A command is an input from an authority ("see to it that φ!") whose effect is that the agent now prefers φ-worlds over $\neg\varphi$-worlds. A suggestion is a milder kind of preference upgrade. See Example 5.

Freund [8, 9] investigates preference revision in the following meaning: how should an initial ranking over a set of worlds be revised by the addition, retraction of modification of the links of the chain? In these papers, "preference" has to be understood as "ranking over a set of worlds" and the results apply indifferently whether the ranking is interpreted in terms of decision-theoretic preferences or in terms of comparative plausilibity. In contrast, our work makes a fundamental distinction between preference and plausibility, since changes of preferences are the repercussion of changes of beliefs.

6 Conclusion

In this paper we have given a first investigation of the properties of preference change in reponse to belief change, depending on the choice of 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 result is a long-term goal that does not seem easy at all, due to the high number of parameters that can vary.

Acknowledgements

We wish to thank the anonymous revewers for helpful comments. Jérôme Lang is supported by the ANR Project ANR-05-BLAN-0384.

REFERENCES

[1] C. Alchourròn, P. Gärdenfors and D. Makinson, On the logic of theory change: Partial meet functions for contraction and revision. *J. of Symbolic Logic*, 50, 510-530, 1985.

[2] J. van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics* 17, 129-156, 2007.

[3] J. van Benthem and F. Liu, Dynamic Logic of Preference Upgrade. In *Journal of Applied Non-Classical Logic*, Vol.17, No.2, 2007.

[4] C. Boutilier, Toward a Logic for Qualitative Decision Theory. *Proceedings of KR94*, 75-86, 1994.

[5] R. Bradley, The kinematics of belief and desire, *Synthese* 156 (3), 513-535, 2007.

[6] A. Darwiche and J. Pearl, On the logic of iterated belief revision. *Artificial Intelligence* 89, 1-29, 1997.

[7] H. van Ditmarsch, W. van der Hoek and B. Kooi, *Dynamic Epistemic Logic*, Springer's Synthese Library, 2007.

[8] M. Freund, On the revision of preferences and rational inference processes, *Artificial Intelligence*, 2004.

[9] M. Freund, Revising preferences and choices, *Journal of Mathematical Economics* 41, 229-251, 2005.

[10] S. O. Hansson, Changes in preferences. *Theory and Decision* 38, 1-28, 1995.

[11] S. O. Hansson, *The structure of values and norms*, Cambridge University Press, 2001.

[12] J. Halpern. Defining relative likelihood inpartially ordered preferential structures. *Journal of Artificial Intelligence Research* 7, 1-24, 1997.

[13] J. Lang, L. van der Torre and E. Weydert, Utilitarian Desires. *International Journal on Autonomous Agents and Multi-Agent Systems*, 5, 329-363, 2002.

[14] J. Lang, L. van der Torre and E. Weydert, Hidden Uncertainty in the Logical Representation of Desires, *Proceedings of IJCAI2003*, pages 685-690. 2003.

[15] F. Liu. Changing for the Better. Preference Dynamics and Agent Diversity. PhD Thesis, University of Amsterdam, 2008.

[16] H. Rott, Shifting Priorities: Simple Representations for Twenty-seven Iterated Theory Change Operators. In: *Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg*. Uppsala Philosophical Studies, pp. 359-384, 2006.

[17] H.H. von Wright, *The logic of preference*, Edinburgh University Press, 1963