From belief change to preference change

Jérôme Lang
IRIT – CNRS & Université Paul Sabatier
Toulouse, France
lang@irit.fr

Leon van der Torre
Université de Luxembourg
Luxembourg
leon.vandertorre@uni.lu

Abstract

There is a huge literature on belief change. In contrast, preference change has been considered only in a few recent papers. There are reasons for that: while there is to some extent a general agreement about the very meaning of belief change, this is definitely not so for preference change. We discuss here the possible meanings of preference change, arguing that we should at least distinguish between four paradigms: preferences evolving after some new fact has been learned, preferences evolving as a result of an evolution of the world, preferences evolving after the rational agent itself evolves, and preferences evolving *per se*. We then develop in more detail the first of these four paradigms (which we think is the most natural). We give some natural properties that we think preference change should fulfill and define several families of preference change operators, parameterized by a revision function on epistemic states and a semantics for interpreting preferences over formulas.

1 Introduction

There is a huge literature on belief change, and after all those years it is reasonable to claim that there is — to some extent — a general agreement about the very meaning of belief revision and other belief change paradigms (especially belief update). In brief, belief revision consists of on agent changing her beliefs about the state of the world after learning some new information about this world, and belief update consists in an agent adjusting her beliefs after learning that a specific world-changing action or event takes place.

Now, the behaviour of an agent is function not only of her beliefs but also of her preferences about the possible states of the world. This raises the issue of when, why and how preferences evolve (if they ever do). At first glance, it seems that preference change occurs in many situations in the lifetime of an agent (just think of he natural language utterances "I have changed my mind", "I don't love you anymore", "I have had enough, I'm not hungry anymore", or "I used not to like beer and now I do so much", "Now that it's raining I don't want anymore to have a walk"). These refer to preference change, and yet they refer to very different processes, and it is not clear that theses processes can be modelled by preference change operators obtained by adapting belief change operators to preferences in a straightforward way. Consider first belief revision. Viewing preference revision as the exact replicate of belief revision would mean that the agent starts with some initial preferences, then "incorporates" a new preference and comes up with new preferences, while preference update, viewed in a similar way, would consist in projecting an agent's preferences after an preference-changing action or event. It is not clear at all what "incorporating a new preference" means, and similarly for a preference-changing action or event.

In the rest of the 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 agen't 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[1], suggesting that the variety of processes being covered by preference change might be larger than that covered by belief change.

The goal of this paper is to give a preliminary exploration of these different meanings conveyed by "preference change", to relate them to existing work (possibly totally outside the "belief change" area) and to discuss briefly the class of methods that could be used to model each of these families of processes. This is the subject of Section 3. Then, in Section 4 we pick the interpre-

[1] We have informally asked a few specialists of belief change around us about the meaning of "preference revision" and we have obtained *very* different answers.

Dagstuhl Seminar Proceedings 07351
Formal Models of Belief Change in Rational Agents
http://drops.dagstuhl.de/opus/volltexte/2007/1209

tation of preference change that we find most relevant and natural, namely, the evolution of an agent's preferences after a revision by a new fact (or belief), and we give more technical developments. Section **??** both discusses related issues, the importance of paying attention to belief change when desinging autonomous agents, and further important research directions.

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 satisfying a formula φ is denoted by $Mod(\varphi)$. We use the following notation for worlds: $a\bar{b}c$ denotes the world where a and c are assigned to *true* and b to *false*. The set of all worlds is denoted by W.

A weak order $\succeq$ is a reflexive, transitive and complete relation. The relations $\sim$ and $\succ$ are defined from $\succeq$ in the usual way: $s \sim s'$ if $s \succeq s'$ and $s \succeq s$ and $s' \succ s$ if $s \succeq s'$ and not $(s' \succeq s)$. If $X \subseteq W$, $Max_{\succeq}(X)$ 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 Preference change: a temptative taxonomy

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.

3.1 Preferences that change when beliefs are revised

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

This is clearly an example of preference change. Letting e for "eating (some of) the sushis", I had a preference for e, something happened, and as a result, I have now a preference for $\neg e$.The event that trigerred the preference change does not primarily concerns preference, but beliefs. Learning that the sushis were made from old fish made me belief that I could be sick, and as a consequence I change my mind about my future behaviour (as I will choose the action "doing nothing" rather than the action "eat").

We can generalize this example to a class of situations that have in common the following: (a) the world is static; (b) the beliefs about the world are revised; (c) the agent's future behaviour is influenced by this belief change. We did not explicitly say that preferences changed. Whether they really change or not is actually a tricky question. To answer it, we are going to give now two distinct formalizations of our example.

In the first formalization, we have two propositional symbols: e (eating sushis) and f (fresh)[2] There are therefore four possibles states of the world, namely $S = \{ef, e\bar{f}, \bar{e}f, \bar{e}\bar{f}\}$. At the beginning of the process, it is reasonable to assume (even if this is not explicitly said) that I believe the sushis to be made out of fresh fish — or, at least, that I do not believe that the fish is not fresh (if I did, 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, then, even if I do not trust the source completely, it is reasonable to expect that my belief that the fish is fresh gets much lower. What about my preferences? If we are talking about preferences over *states* (as opposed to preferences over actions), then my initial preferences are likely to be

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

(I prefer eating fresh sushis over not eating sushis, and I prefer not eating sushis over eating sushis made out of old fish; if I do not eat the sushis I don't care whether the fish is old or not[3]. 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 know now that this world is impossible, ar, at least, highly implausible). to $\neg e$ and $\neg e$ to $e\bar{f}$. Thus, in this situation, *belief change, but preferences remain static.* Still, it is no less true that I used to intend to eat these sushis and I do not anymore. This is right, but we are now talking about *actions*, as opposed to *properties of the world.* Indeed, my future behaviour (that is, the action that I intend to do) has changed, but my preference between states of the world has not. This process is actually 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.

In the second formalization, we stil use two symbols e and f but we want to reason about the preference between e, seen as a propositional formula (corresponding to the set of states $\{ef, e\bar{f}\}$) and $\neg e$ (corresponding to the set of states $\{\bar{e}f, \bar{e}\bar{f}\}$). 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 state satisfying e to the most state satisfying $\neg e$. Of course, after learning the information about the fish, these typical states are now

[2] We could also introduce a third symbol s for "sick"; together with some belief that $\neg f \wedge e$ implies s, but this turns out to be unnecessary.

[3] One may argue that in a real situation $\bar{e}\bar{f}$ is preferred to $\bar{e}f$, because if $\bar{e}f$ is the case then I may experienced the regret of not having eaten the sushis, if I later learn that they were fresh. For the sake of simplicity we will not consider regret in our approach.

$e\bar{f}$ and $\bar{e}\bar{f}$, and after focusing, I prefer now the formula $\neg e$ to the formula e.

Therefore, whether preference change or not when our beliefs change depends of whether we talk about preferences over *states of the world*, *formulas* or *actions*. Preferences over states are static, but their lifting on formulas or actions change.

Finally, one may also argue that whether preferences over states change or not is also 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 change, and in this case, it is not possible to express that we learn $\neg f$, therefore the only wat of modeling the input is a "direct preference change" (see further): the world sends a "command" to the user, asking her to now prefer $\bar{e}$ to e.

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

3.2 Preferences that change when the world changes

Example 2 *Initially, I desire to eat sushis from this plate. Then I eat 50 sushis. After that, I desire not to eat sushis.*

Example 3 *It is a nice saturday afternoon and I'd like to have a walk. Then is starts to rain. After that I don't want to have a walk anymore.*

Example 3 clearly illustrates a preference change trigerred by a change of the world (it was not raining and now it does). So is Example 2 (I was hungry and now I am not), however there is a second way of interpreting this example (see Subsection 3.3).

Things are quite similar to the situation discussed in Subsection 3.1, 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. We have therefore *static preferences, dynamic world and dynamic beliefs*.

3.3 Preferences that change when the rational agent evolves

Example 4 *When I was a child I did not like cheese. Now I do.*

Here, a change in preference reflects a modification of the agent's tastes due to an event (or several events) the agent is subject to. In Example 2, that can be viewed as well as a change in the rational agent, we clearly see what the event is (eating 50 sushis). This is less clear with Example 4, as there is no clear, "namable" event that made the agent change his mind and start to like cheese. One may just say that this event is "growing up", or, going further in the granularity of events, and say that this change has resulted from a lot of micro-events (such as eating a little bit of cheese many times in several years).

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 (it is generally assumed not to — and this is not the place to enter this discussion).

3.4 Direct preference change

[2] consider direct preference change (unrelated to anything else), trigerred by "commands" or "suggestions" (the difference both being a matter of strength).

Example 5 *[2] Let's take a trip!*

This kind of preference change mimics exactly belief change, in the sense that preferences are revised by preferences (so as to lead to new preferences), without any beliefs to intervene in the process. The situations in which this preference change *per se* occurs are those where another agent (or nature) can make an agent *believe* α by sending him a signal asking him, or leading him, to prefer α. A context where this happens is the context considered in Example 1 when f is not in the language: I can simply not make you revise your beliefs by $\neg f$, for the technical reason that $\neg f$ cannot be expressed, but I can instead ask you to revise your preferences in the same way that they would have evolved after incorporating the piece of evidence $\neg f$: "I order you to prefer $\neg e$."[4].

3.5 Other kinds of preference change?

There are at least two other kinds situation where we may want to say that preference change occur.

The first one is when revising (or updating) an agents' preferences by some new information about this agent's preferences. For instance: I am the system that sells you train tickets and when you ask me for a ticket from Paris to Toulouse I initially believe that you want to take the TGV and go through Bordeaux – until you tell me that you want to go through Limoges. This is however a pure belief revision process, in which the world on which we reason concerns your preferences, so this process is not about preference change, *belief change about another agent's preferences* – so this situation does not really have to be discussed in this paper, but it ought to be mentioned at some point.

The second one is when an agent is following a plan and has a desire for α to be satisfied because it is a means-end objective. When α is realized, after that I don't need α anymore and my preference for α disappears. See example 2: the primary goal is not to be

[4] A similar context where direct preference can be seenmore clearly is in dialogues such as the following one: "is there anything interesting to see in this town? – Oh no, you don't want to go here".

hungry any more, and eating sushis can be seen as a means (not the worst one, admittedly) to see to it that the goal is satisfied. (One can of course consider more complex plans with several actions in sequence.) This is clearly a variation on "preference change implies by a change in the world" (the world has changed because some subgoals have been satisfied), and also a variation on "preference change implies by a change in the rational agent" (the agent had an intention to see α satisfied, now that it has been satisfied he doesn't care anymore – think of those Casanovas who want to seduce all women).

A situation similar to the latter (but a little bit more complicated) is when I learn that α won't help me reach my goal. An example: I have the desire to prove a conjecture, which easily follows from the conjunction of two lemmas (Lemma 1 and Lemma 2). I initially have a preference for Lemma 1 to be proven and similarly for Lemma 2. I expect both lemmas to be true. However, then I find a counterexample for Lemma 1, and in this case Lemma 2 becomes useless, so my preference for Lemma 2 to be proven disappears. Anyway, again this can be seen as an instance of the classes of preference change developed in the previous subsection.

4 Preference change triggered by belief revision

4.1 Beliefs and preferences

We now consider in more details the scenario that we discussed informally in Subsection 3.1. 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 (or actions);
- the agent learns a new piece of information α about the world;
- the agent 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 [12], 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 don't 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 indifference relations $\sim_N$ and $\sim_P$ are defined as usual, as well as the strict relations $\succ_N$ and $\succ_P$, are defined as usual (see Section 2).

The model for Example 1 is visualized in Figure 1. 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. The most preferred worlds are the ones in which we are eating fresh sushi, which is preferred to not eating fresh sushi, and not eating not fresh sushi is preferred to eating not fresh sushi $ef \succ_P \bar{e}f \succ_P \bar{e}\bar{f} \succ_P e\bar{f}$.

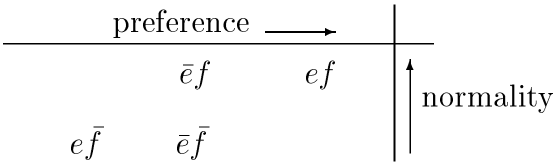

Figure 1: The sushi example. $fe \sim_N f\bar{e} \succ_N \bar{f}e \sim_N \bar{f}\bar{e}$ and $ef \succ_P \bar{e}f \succ_P \bar{e}\bar{f} \succ_P e\bar{f}$.

Again as in [4; 12], the language is built up from a finite set of propositional symbols, usual connectives, and two dyadic modalities: one for normality (N) and one for preference ($>$, and also P – the latter two being interdefinable, see further).

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

Definition 2 (normality)
$\mathcal{M} \models N(\psi|\varphi)$ 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 world level to the formula level (see [**?**; 11]). We consider here the three following ways of lifting [11][5]

$\mathcal{M} \models \varphi \gg^P_{mM} \psi$
if $\forall w \in Mod(\varphi)\ \exists w' \in Mod(\varphi)$ such that $w \succ_P w'$
that is, if the worst φ-worlds are preferred to the best ψ-worlds (or, equivalently, every φ-world is preferred to every ψ-world).

$\mathcal{M} \models \varphi \gg^P_{MM} \psi$
if $\exists w \in Mod(\varphi)$ such that $\forall w' \in Mod(\varphi)$, $w \succ_P w'$
that is, if the best φ-worlds are preferred to the best ψ-worlds (or equivalently, the best $\varphi \vee \psi$ worlds are $\neg\psi$ worlds).

$\mathcal{M} \models \varphi \gg^P_{mm} \psi$
if $\forall w \in Mod(\varphi) \exists w' \in Mod(\varphi)$ such that $w \succ_P w'$
that is, if the worst φ-worlds are preferred to the worst ψ-worlds.

[5] There is obviously a fourth one (Mm), corresponding to two existential quantifiers; however, this notion is much too weak, as it makes $P\varphi \wedge P\neg\varphi$ consistent.

Alternative ways of lifting preference would also be worth considering, such as, for instance, *ceteris paribus* preferences [14] of other kinds of similarity-based preferences [10]. However, for the sake of brevity, in this paper we stick to these three ways of lifting preferences.

Second, as argued in [4; 12], 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 p over $\neg p$ in the presence of uncertainty or normality. Let $\gg$ be one of $\gg^P_{mM}$, $\gg^P_{Mm}$, or $\gg^P_{MM}$.

1. "among the most normal q-worlds, p is preferred to $\neg p$," [4]:
$\mathcal{M} \models P(\psi|\varphi)$ iff
$Max_{\succeq_N}(Mod(\varphi)) \cap Mod(\psi) \gg Max_{\succeq_N}(Mod(\varphi)) \cap Mod(\neg\psi))$

2. "the most normal $p \wedge q$-worlds are preferred to the most normal $\neg p \wedge q$-worlds" [12]:
$\mathcal{M} \models P(\psi|\varphi)$ 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 $p \wedge q$ worlds or the most normal $\neg p \wedge q$ worlds may be exceptional among the q worlds[6].

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 LVW (for item 2. above), and one the three subscripts MM, mm, or mM. For instance, P^{LVW}_{MM} refers to the semantics in [12] and the optimistic way of lifting preferences (which is the semantics studied in detailed in [12]). However we will try to avoid using these heavy subscripts and superscripts whenever possible.

Now, from the P modality (where $P(\varphi|\psi)$ means "given ψ, I have a preference / a desire for φ" we define a $>$ 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 also [10])[7]:

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

4.2 Belief revision, and its impact on preferences

Revising a pre-order

Given a model $M = \langle W, \succeq_N, \succeq_P \rangle$, the revision by belief α is a new model $M' = M \star \alpha$ consists in 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 belief revision (and especially iterated revision, see e.g. [6]).

Definition 3 *Given a set of worlds W, a revision function $\star$ is a function that maps each complete weak order over W into a complete weak order over W, and that satisfies the acceptance property: for every $\succeq_N$ and every consistent α, $Max_{\succeq_N \star \alpha}(W) \subseteq Mod(\alpha)$ – in other words, most normal worlds after revising by α should satisfy α*

Given a model $M = \langle W, \succeq_N, \succeq_P \rangle$, a revision function $\star$, and a formula α, the revision of M by α, is the model $M \star \alpha$ defined by

$$M \star \alpha = \langle W, \succeq_N \star\alpha, \succeq_P \rangle$$

Note that acceptance implies that $\mathcal{M} \star \alpha \models N\alpha$. Apart of acceptance, revision functions on plausibility orderings are usually required to obey some other properties. A common one is the uniform shifting of p worlds[8]:

Definition 4 *A revision operator $\star$ satisfies:*

- positive uniformity *if for any two worlds w, w' such that $w \models \alpha$ and $w' \models \alpha$ then $w \succ_N^{\star\alpha} w'$ iff $w \succ_N w'$;*
- negative uniformity *if for any two worlds w, w' such that $w \models \neg\alpha$ and $w' \models \neg\alpha$ then $w \succ_N^{\star\alpha} w'$ iff $w \succ_N w'$.*

[6] 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 [12]) is a matter of choosing the worlds to focus on: when we are asked to compare two (incomplete) alternatives, we focus on typical situations that satisfy each of these alternatives and then we compare these situations. The approach in [4] first focuses on most normal worlds *independently* of the choice between the two alternatives. This has the consequence that the comparison becomes void when either $p \wedge q$ or $\neg p \wedge q$ is exceptional, because, wlog in the case where $p \wedge q$, there is no most normal $p \wedge q$-world to compare with most normal $\neg p \wedge q$-worlds. Consider q = taking the airplane, p = the airplane crashes. Because most normal q-worlds satisfy $\neg p$, there can be no preference for $\neg p$ given q. Both definitions ([4] and [12]) coincide iff there exist both most normal $p \wedge q$-worlds and most normal $\neg p \wedge q$-worlds, that is, if $\neg N(p|q) \wedge \neg N(\neg p|q)$ holds.

[7] This interdefinability needs a special treatment of limit cases where either $\varphi \wedge \psi$ or $\varphi \wedge \neg\psi$ is unsatisfiable – see [10]. In this paper we omit the treatment of these limit cases, which are of little interest anyway.

[8] These properties are named respeticely (CR1) and (CR2) in [6]

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 a modal logic to refer to updates [7].

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

We will now investigate a few key properties concerning preference change, depending on the belief revision operator $\star$ used and the choice of the semantics for interpretaing preference.

Properties 1: preference satisfaction

We are now going to look into the logical properties of preference change under newly learned beliefs. The properties we may expect can be derived from the properties of belief revision and preference logic. For example, whereas in belief revision newly learned beliefs that are no surprises do not change the old beliefs, we may consider whether newly learned beliefs which are not exceptional do not change the preferences. We do so in the following section, but we start with a simpler pattern.

Suppose we learn that what we want to hold, in fact holds. In that case, it would be intuitive that the preference still holds, i.e. persists in time. This property holds provided that $\star$ satisfies uniformity.

Proposition 1 (learning the preferred) *Suppose that $*$ satisfies positive and negative uniformity. Then for any formula p the following are true in any model $\mathcal{M}$:*

1. $p >_{MM} \neg p \to [*p](p >_{MM} \neg p)$;
2. $p >_{mm} \neg p \to [*p](p >_{mm} \neg p)$;
3. $p >_{mM} \neg p \to [*p](p >_{mM} \neg p)$;

Let us give a quick proof for 1 (the proof is similar for 2 and 3). Suppose $\mathcal{M} \models p >_{MM} \neg p$, which by definition is equivalent to: for any $w \in Max_{\succeq_P}(Max_{\succeq_N}(Mod(p))$ and $w' \in Max_{\succeq_P}(Max_{\succeq_N}(Mod(\neg p))$, we have $w \succeq_P w'$. Now, positive uniformity implies that $Max_{\succeq_{N*p}}(Mod(p)) = Max_{\succeq_N}(Mod(p))$, and negative uniformity, that $Max_{\succeq_{N*p}}(Mod(\neg p)) = Max_{\succeq_N}(Mod(\neg p))$: the most normal p-worlds are the same before and after revision by p, and similarly for the most normal $\neg p$-worlds. Therefore,
$Max_{\succeq_P}(Max_{\succeq_{N*p}}(Mod(p))) = Max_{\succeq_P}(Max_{\succeq_N}(Mod(p)))$
and
$Max_{\succeq_P}(Max_{\succeq_{N*p}}(Mod(\neg p))) = Max_{\succeq_P}(Max_{\succeq_N}(Mod(\neg p)))$
hence the result. (Note that it would also with a ceteris paribus semantics of preferences, or more generally any semantics of preference)

The positive and negative uniformity conditions are necessary. Consider for instance drastic revision operator that preserves the relative ranking of α-worlds and then push all $\neg\alpha$-worlds towards the bottom, irrespectively of their relative initial ranking: $w \succeq_N^{*\alpha} w'$ 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:

$$pq \succ_N \bar{p}\bar{q} \succ p\bar{q} \succ \bar{p}q$$

$$\bar{p}q \succ_P pq \succ_P \bar{p}\bar{q} \succ p\bar{q}$$

after revision by p:

$$pq \succ_N^{*p} p\bar{q} \succ \bar{p}q \sim \bar{p}\bar{q}$$

We have $\mathcal{M} \models p >_{MM} \neg p$ and $\mathcal{M} \models [*p]\neg p >_{MM} p$.

By symmetry, things are the same if we revise by a dispreferred formula:

Proposition 2 (learning the dispreferred) *Suppose that $*$ satisfies positive and negative uniformity. Then the following are true in any model $\mathcal{M}$:*

1. $p >_{MM} \neg p \to [*\neg p](p >_{MM} \neg p)$;
2. $p >_{mm} \neg p \to [*\neg p](p >_{mm} \neg p)$;
3. $p >_{mM} \neg p \to [*\neg p](p >_{mM} \neg p)$;

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 still holds, i.e. persists in time. However, suppose that we prefer p and we learn that $p \vee q$. In that case we are shifting the p worlds uniformly, but not necessarily the $\neg p$ worlds. All we know is that when some of the most normal $\neg p$ worlds are $\neg p \wedge q$ worlds, then these $\neg p \wedge q$ worlds will become the most normal p worlds. This property therefore holds provided that $\star$ satisfies uniformity, $\neg N(\neg q|\neg p)$ holds, and only for two of the three preferences.

Proposition 3 (learning the partly preferred) *Suppose that $*$ satisfies positive and negative uniformity. Then the following are true in any model $\mathcal{M}$:*

1. $p >_{MM} \neg p \wedge \neg N(\neg q|\neg p) \to [*p \vee q](p >_{MM} \neg p)$;
2. $p >_{mM} \neg p \wedge \neg N(\neg q|\neg p) \to [*p \vee q](p >_{mM} \neg p)$

Let us give a quick proof for 1 (the proof is similar for 2). Suppose $\mathcal{M} \models p >_{MM} \neg p$, which by definition is equivalent to: for any $w \in Max_{\succeq_P}(Max_{\succeq_N}(Mod(p))$ and $w' \in Max_{\succeq_P}(Max_{\succeq_N}(Mod(\neg p))$, we have $w \succeq_P w'$. Now, assume in addition that $\neg N(\neg q|\neg p)$, which by definition is $Max_{\succeq_N}(Mod(\neg p \wedge q)) \subseteq Max_{\succeq_N}(Mod(\neg p))$, then positive uniformity implies $Max_{\succeq_{N*p}}(Mod(p)) = Max_{\succeq_N}(Mod(p))$, and negative uniformity implies analogously that $Max_{\succeq_{N*p}}(Mod(\neg p)) \subseteq Max_{\succeq_N}(Mod(\neg p))$: the most normal p-worlds are the same before and after revision by p, and the most normal $\neg p$-worlds will be a subset. Therefore,
$Max_{\succeq_P}(Max_{\succeq_{N*p}}(Mod(p))) = Max_{\succeq_P}(Max_{\succeq_N}(Mod(p)))$
and for $w \in Max_{\succeq_P}(Max_{\succeq_{N*p}}(Mod(\neg p)))$ and $w' \in Max_{\succeq_P}(Max_{\succeq_N}(Mod(\neg p)))$ we have $w \succeq_P w'$.
hence the result: if the best world among these worlds used to be a p world, then it remains a p world. (note that it does not hold for mm, since if the worst world

used to be a $\neg p$ world, after the revision the worst world may be a p world.)

By symmetry, things are the same if we revise by a dispreferred formula:

Proposition 4 (learning the partly dispreferred) *Suppose that $*$ satisfies positive and negative uniformity. Then the following are true in any model $\mathcal{M}$:*

1. $p >_{mM} \neg p \wedge \neg N(q|p) \rightarrow [*\neg p \vee q](p >_{mM} \neg p)$
2. $p >_{mm} \neg p \wedge \neg N(q|p) \rightarrow [*\neg p \vee q](p >_{mm} \neg p)$

Properties 2: surprises

We may expect that preferences don't change when we revise by something normal (i.e., expected). However, for P this holds only under the assumption that the normality ordering remains the same when we revise by a normal formula:

Proposition 5 (learning the normal, 1)

1. *for Boutilier's semantics, under any of the four definitions of lifting, the following formula is valid:*

$$N\alpha \wedge P\varphi \rightarrow [\star\alpha]P\varphi$$

2. *for LTW's semantics, under the four definitions of lifting, the latter formula is valid* provided *that $\star$ satisfies the following inertia property: if $Max_{\succeq_N}(W) \subseteq Mod(\alpha)$ then $\succeq \star\alpha = \succeq$.*

In case 2, the validity of $N\alpha \wedge G\varphi \rightarrow [*\alpha]G\varphi$ comes simply from the fact that $\succeq_N$ does not change. In case 1, the fact that $N\alpha$ is true implies that all most normal worlds satisfy α, therefore revising by α lead these most normal worlds (that is, $Max_{\succeq_N}(W)$) unchanged; since the truth of $G(.|.)$ depends only on $Max_{\succeq_N}(W)$, preferences remain unchanged.

However, 1. no longer holds if $\star$ does not satisfy inertia, 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 $\neg\varphi$*:

Proposition 6 (learning the normal,2) *For LVT as well as Boutilier's semantics, and for any kind of lifting, the following formula is valid:*

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

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.

Proposition 7 (learning the normal,3) *For LVT as well as Boutilier's semantics, and for any kind of lifting, the following formula is valid:*

$$N\alpha \wedge \neg N\varphi \wedge \neg N\neg\varphi \wedge P\varphi \rightarrow [\star\alpha]P\varphi$$

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

Now, when revising by something that *is not disbelieved*, we would expect some form of preservation of preference as well. We consider this forst form of revision by the non-exceptional (non-disbelieved):

Proposition 8 (learning the non-exceptional,1) *For LVW (as well as Boutilier – CHECK) semantics, and for the $\forall\forall$ lifting (mM), the following formula is valid:*

$$\neg N(\neg\alpha|\varphi) \wedge \neg N(\neg\alpha|\neg\varphi) \wedge P\varphi \rightarrow [\star\alpha]P\varphi$$

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 MM, mm and Mm, as it can be seen on the following example:

$\succeq_N$: $pq \sim p\bar{q} \succ \bar{p}q \sim \bar{p}\bar{q}$

$\succeq_P$: $p\bar{q} \succ \bar{p}q \succ pq \succ \bar{p}\bar{q}$

We have $\mathcal{M} \models Pp$ for any of $\{MM, mm, Mm\}$. 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, or related issues such as goal change, has been considered under various forms in a few works that so far are unrelated to each other.

Bradley [5] argues that changes in preference can have two sorts of possible causes: change in beliefs (corresponding to the situation we described in Subsection 3.1) and "what might be called change in tastes" (which corresponds to the situation we described in Section 3.3). (It is not clear in which of both whether the situation of Subsection 3.2 should be classified.) He further refines the first case into two kinds of situations where learning B makes our desirability of A change: (a) A is preferentially dependent on B; (b) B is preferentially dependent on A, and there is a probabilistic dependency between A and B. Then he develops a Bayesian formalization of these ideas. Our work goes further in this direction and connects the interaction between belief change and preference change to the existing body of research in belief revision.

Van Benthem and Liu [2; 13] give a dynamic epistemic logic formalization of 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. Both kinds of preference change considered in this stream of works refer to the situation described in our Subsection 3.4 (direct preference change).

Freund [8; 9] investigates preference revision in the following meaning: how should an initial ranking (called a "chain") over a set of worlds be revised by the addition, retraction of modification of the links of the chain? In these two papers, "preference" has to be understood in its technical sense (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 plausilibity. In contrast, our work makes a fundamental distinction between preference and plausibility, and changes of preferences are viewed as the repercussion of changes of beliefs.

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 and F. Liu: Dynamic Logic of Preference Upgrade. In *Journal of Applied Non-Classical Logic*, Vol.17, No.2, 2007.

[3] J. van Benthem, O. Roy, and P. Girard, Everything else being equal: A modal logic approach to *ceteris paribus* preferences.

[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*, 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, *The structure of values and norms*, Cambridge University Press, 2001.

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

[12] J. Lang, L. van der Torre and E. Weydert, Hidden Uncertainty in the Logical Representation of Desires, *Proceedings of Eighteenth International Joint Conference on Artificial Intelligence (IJCAI2003)*, pages 685-690. 2003.

[13] F. Liu: Preference Change and Information Processing. In *Proceedings 7th Conference on Logic and the Foundations of Game and Decision Theory (LOFT 06)*, Liverpool, 2006.

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