Chapter 8
From Belief Revision to Preference Change

Till Grüne-Yanoff and Sven Ove Hansson

Abstract We propose to model the consistency-preserving aspect of preference change after the fashion of belief revision. First, we discuss the formal properties of the preference notion. Second, we discuss the various consistency requirements imposed on preference sets. Third, we discuss representations of consistency-driven preference change and compare them to models of belief change. Last, we discuss the specific needs of introducing a priority index in models of preference change. We conclude that while the general input-assimilating framework from belief change can be transferred to preference change, several modifications are necessary. In particular, the input model has to be complicated with the introduction of a distinction between primary (non-linguistic) and secondary (linguistic) inputs. Sentential representation has to be used with somewhat more caution for preferences than for beliefs. The priority-setting mechanism has to be adjusted, and priority-related information must be included in the inputs.

8.1 Introduction

How should a formal theory of preference change be constructed? In order to get a systematic grip on that issue, we have chosen to attack it from two sides. First and most obviously, we draw on previous studies of preferences, both preference logic and more informal discussions on preferences in the social sciences. Secondly, we compare preference change to belief change. There are important similarities, but – as we will soon see – also major differences between these two areas of formalized philosophy. Contrary to preference change, belief change is an established field. In formal epistemology and theoretical computer science, a large variety of formal models of belief change have been developed, for both descriptive and normative

T. Grüne-Yanoff
Helsinki Collegium of Advanced Studies and Royal Institute of Technology, Stockholm
e-mail: till.grune@helsinki.fi

S.O. Hansson (✉)
Royal Institute of Technology, Stockholm
e-mail: soh@kth.se

T. Grüne-Yanoff and S.O. Hansson (eds.), *Preference Change: Approaches from Philosophy, Economics and Psychology*, Theory and Decision Library A 42,

purposes (Gärdenfors 1988; Hansson 1999). In this contribution we intend to identify the central aspects of standard belief change models, discuss their applicability to preference change, and in this way put focus on several important issues in preference change. As we will show, some of the central features of belief change models can be used in preference change, but a number of extra features are also required that distinguish preference change models from the main models of belief revision.

Mental changes, or changes in mind, can take many forms, of which changes in beliefs and preferences are only two. Our norms, our emotions, our patterns of argumentation, our ideologies, etc. are also subject to change. Preference changes should not be seen as independent of these other types of change. Some of the most important questions that we need to clarify concern the interconnections between changes in these various compartments of the mind. How are for instance preferences affected by changes in beliefs, and the other way around? In order to investigate such issues we need models that represent larger parts of a state of mind than its preferences.

It is probably a good strategy to develop workable models of preference change before we try to develop such larger models. Discussing the makeup of such a simple model will be the main aim of this paper. But it is also advisable to construct a model of preference change such that it is embeddable into larger models of changes in mental states. This is a factor that we will pay particular attention to in what follows.

In Section 8.2 we investigate how preferences relate to four other conceptual categories that are important in the dynamics of mind and action, namely values, norms, choices, and beliefs. In Section 8.3 we discuss how preferences and their relata should be represented, and in Section 8.4 we investigate what general rationality constraints should apply to all preference states. Section 8.5 is devoted to fundamental issues in the representation of change, such as the typology of change operations and the role of inputs in these operations. In Section 8.6 we discuss what mechanisms should be used to select among alternative outcomes that all satisfy the rationality constraints. Section 8.7 concludes.

8.2 Preferences, Values, Norms, Choices, and Beliefs

First of all, the category of preferences has to be positioned among the various other categories that may be subject to formal analysis. We have chosen to focus on four such categories that seem to be particularly relevant for the comprehensive formal characterization of preferences, namely values, norms, choices, and beliefs.

8.2.1 Values

Preferences are expressions of values. From a structural point of view, the value concepts that we use in ordinary language as well as in more specialized discourse can be divided into two major categories. The *monadic* (classificatory) value concepts,

such as 'good', 'very bad', and 'worst', evaluate a single object. The *dyadic* (comparative) value concepts, such as 'better', 'worse', 'at least as good as', and 'equal in value to' make a value-based comparison between two objects. Preferences, as they are usually conceived, have their place in this category. To say that someone prefers A to B is synonymous with saying that according to some of that person's values, A is better than B.[1]

The common dyadic value concepts are usually taken to be interdefinable, hence it is assumed that A is better than B if and only if it is both the case that A is at least as good as B and that A is not equal in value to B. The subject-matter of preference logic is usually taken to cover all the dyadic value concepts. This will be our approach here. In other words, we will consider the topic of preference change to cover not only changes in what the subject prefers to something else but also in what the subject considers to be of equal value as something else.

There are close structural relationships between monadic and dyadic values. It would seem paradoxical to claim both that A is better than B and that B is best. It would be almost equally strange to claim both that A is better than B and that A is bad whereas B is good. It is generally accepted in formal studies of preferences that the monadic value predicate 'best' can be defined in terms of dyadic value predicates. An object A is best among a group of objects if it is better than all other objects in that group (or alternatively if no object in that group is better than A). At the other end of the value-scale, 'worst' is defined analogously (Hansson 2001a, pp. 115–116). Proposals are also available for the definition of 'good' and 'bad' in dyadic terms. According to one such proposal, a value-object is good if and only if it is better than its negation, and it is bad if and only if it is worse than its negation (Brogan 1919). According to the other major proposal, a value-object is good if and only if it is better than some proposition that is neither better nor worse than its negation. Similarly, it is bad if and only if it is worse than some proposition that is neither better nor worse than its negation (Chisholm and Sossa 1966). Both these definitions have the disadvantage of only being applicable to negatable value objects.

If monadic values can be defined in terms of dyadic values, then an account of preference change, i.e. change in dyadic values, will generate a derivative account of changes in monadic values, so that the logic of preference change becomes a general logic of value change.

8.2.2 *Norms*

Even among philosophers otherwise committed to fine linguistic distinctions, the distinction between norms and values is often overlooked. There is in fact an

[1] For some technical purposes, value predicates with more than two referents may be useful, such as the four-termed "x is preferred to y more than z is preferred to w" (Packard 1987). Our focus here will be on the dyadic concepts.

essential difference: norms are directly action-guiding whereas values are not. Hence, suppose that we have a choice between three exhaustive and mutually exclusive action alternatives A, B, and C. The statement that A is better than each of B and C, and B better than C, is unproblematically compatible with each of the following three statements (1) A is obligatory whereas B and C are forbidden, (2) A and B are both permitted whereas C is forbidden, and (3) A, B, and C are all permitted. More generally speaking, even if we know what values someone assigns to a set of alternatives, we cannot infer from this what normative statements she endorses. The best alternative may be supererogatory (i.e. good, but not obligatory; cf. Chisholm 1963), or the normative appraisal may be based on a satisficing account of normativity (Slote 1984).

Of course, if we adopt the principle that maximal value-production is required of all agents, then the normative status of an action can be derived from its value status. However, this principle is implausible, since it rules out supererogatory acts and, even more importantly, limits the freedom of the agent to a choice among a few value-maximal alternatives (Hansson 2006).

Although norms are not in general derivable from preferences (or from other expressions of value), norms and values are not fully independent of each other. It would for instance not seem credible to claim that the action A is better than all alternative actions and at the same time maintain that A is forbidden whereas all its alternatives are permitted.[2] Criteria of coherence can be applied to combinations of norms and preferences, but these criteria cannot be assumed to be sufficiently specified to make norms derivable from preferences.

8.2.3 *Choices*

There is a strong tradition, particularly in economics, to equate preference with choice. Preference is considered to be hypothetical choice, and choice to be revealed preference. Hence, the Arrovian framework in social choice theory "conflates 'choice' and 'preference'", and treats these "as essentially synonymous concepts.... A preference is a potential choice, whereas a choice is an actualized preference" (Reynolds and Paris 1979). Arrow himself has defined preference as "choice from two-member sets" (Arrow 1977, p. 220). The same approach dominates in other areas of economics.

However, the conflation of choices and preferences is a rather far-stretched idealization that is not adequate for all purposes. In fact, choices and preferences are entities of quite different categories. Preferences are parts of *states of mind*. That a person prefers A to B means that she considers A to be better than B. Choices are

[2] This intuition is supported by a plausible deontic principle, namely contranegativity ($OX \,\&\, (\neg X \geq \neg Y) \rightarrow OY$, where O is the ought operator). To see this, let B be any of the alternatives to A. Then according to the assumptions of the example, $A > B$ and $O \neg A$. Contranegativity yields $O \neg B$ so that $\neg O \neg B$ does not hold, i.e. B is not permitted.

actions. That someone has chosen A means that she has actually selected A (irrespectively of whether she judges A to be better than its alternatives). Even in market behaviour, the primary subject-matter of economic theory, there are several types of situations in which choice and preference clearly do not coincide (Sen 1973). In particular, in markets and elsewhere, a person can select from alternatives that she is indifferent between or considers to be incomparable (Ullmann-Margalit and Morgenbesser 1977). It is also possible to have preferences over alternatives that one cannot choose between. Suppose that there are two prizes in a lottery: a luxury cruise worth € 10,000 and an account of € 10,000 at your local grocery store that you can use to buy food in the years to come. You may then very well prefer winning the luxury cruise to winning the account at the grocery, but since you cannot choose what to win – if you could it would not be a lottery – this preference is not directly connected to choice. (As this example may illustrate, you may prefer *winning* A to *winning* B even though you would, in a direct choice between A and B, choose B rather than A.)

Hence, although preferences and choices are often conflated, they are very different in nature and should be treated as phenomena on different levels in a model of mind and action. Preferences influence choices, just as beliefs do, but they should be carefully kept apart from choices in a formal model of mental changes.

8.2.4 Beliefs

Both beliefs and preferences are parts of a person's state of mind. Beliefs refer to the realm of facts and preferences to the realm of value. We have learned to keep these realms apart, and of course the distinction should not be blurred. However, this does not mean that preferences and beliefs are independent of each other, so that any combination of preferences is compatible with any combination of beliefs. Instead, their relation is similar to that between values and norms: neither is derivable from the other but a change in one of the two categories can have impacts on the other.

The influence of beliefs on preferences is largely uncontroversial since it is generally accepted that preferences should be factually well-informed. New information often leads us to modify our preferences. Some preferences would be considered irrational if we have certain beliefs. Hence, if a person believes (correctly) that Le Corbusier and Charles-Edouard Jeanneret-Gris were one and the same person, then it would be incoherent of her to prefer houses built by Le Corbusier to houses built by Charles-Edouard Jeanneret-Gris.

The reverse influence, from preferences to beliefs, is more controversial. Wishful thinking, i.e. believing that things are as we wish them to be, often leads us astray (Hansson 2006). However, other, more sophisticated influences of preferences on beliefs seem to be justified. In particular, our values can influence the standards of evidence, or burdens of proof, that we assign to different potential beliefs. Hence, our strong preference for safety in a medical context makes us put high demands on the evidence before we allow ourselves to believe that a new drug is safe for

humans. In comparison, we tend to require less stringent evidence before we come to believe that a drug has a serious side-effect. Although this asymmetry in standards of evidence is not uncontroversial, it should not be excluded in a general framework for studies of mental change.

8.2.5 Summary

We can summarize these considerations as follows: By changes in "preferences" we actually mean changes in value comparisons, i.e. in those values that we express with dyadic value statements. This also includes for instance a change from regarding two objects as incomparable to considering them to be of equal value. Preferences (in this wide sense) are closely related with monadic value concepts, and at least some monadic value statements are derivable from preferences.

Preferences, norms, and beliefs are all parts of the mental state, or state of the mind. These three categories are distinctly different, and a statement belonging to one of them cannot be synonymous with a statement belonging to one of the others. Nevertheless there are relationships among the three categories. Even if all beliefs in a state of the mind are internally coherent, and the same applies to the preferences and the norms, the combination of the three components may be incoherent.

Finally, choices are not parts of the state of the mind. In a model representing both actions and the mind, choices should be represented among the actions and preferences as parts of the state of mind. Just like beliefs and norms, preferences can influence choices. Yet choices and preferences are not identical.

8.3 The Representation of Preferences

The choice of a preference representation model has direct influence on the framework for preference change. We propose a representation of preferences as binary relations over sentences. This distinguishes our framework from expected utility functions over goods bundles on the one hand (as standardly used in microeconomics) and from modal logic frameworks on the other (for examples, see van Benthem, Chapter 3, this volume). We justify our choice by simplicity and convenience. It is convenient, because a representation based on sentences immediately places our framework close to standard models of belief revision, thus allowing easy comparison. It is further convenient because a large part of the social science literature naturally relates to this formal framework. It is simple because it leaves out probabilistic information. Deliberately forgoing this extra information forces us to model preference change without recourse to these parameters. Of course, this limits the application of our framework to many real-world situations, in which probability change often plays an important role (for a model including probabilistic change, see Bradley, Chapter 11, this volume). Yet it should also be pointed out that at the basis

of any expected utility theory lies a simple preference ordering of the sort modelled here. Therefore our framework can be seen as a necessary basis of expected utility theory, and thus as compatible with it.

8.3.1 *Relata*

In order to represent preferences we need a representation of that which they refer to, the relata. A manageable formal representation of the relata will require some degree of simplification, since in non-regimented language, all sorts of abstract and concrete entities can serve as the relata of preference relations. Thus, one may prefer coffee to tea, logic to postmodern literature theory, or novels to poetry. In spite of this, preference theory has been almost exclusively restricted to two representations of relata: Either relata are taken as primitives, or they are taken to be sentences representing states of affairs.

The use of states of affairs to represent relata can be defended with reference to the ease with which we can translate talk about other types of relata into talk about states of affairs. This translation has often been taken as unproblematic. Hence R. Lee (1984, pp. 129–130) claimed that "all preferences can be understood in terms of preference among states of affairs or possible circumstances. A preference for bourbon, for example, may be a general preference that one drink bourbon instead of drinking scotch" (Cf. von Wright 1963, p. 12; Trapp 1985, p. 303).

Arguably, it is not quite as simple as that. A person's preference for one musical piece over another, for example, cannot be translated into a single preference for one state of affairs over another. Instead, it can be represented by a conglomeration of preferences referring to these pieces of music: she may prefer states of affairs in which she *plays* the first rather than the second piece, but she may also prefer a state of affairs in which she *listens* to the first rather than the second, etc. To dissolve this ambiguity, one needs to investigate the context in which a preference over the primitives was expressed.

In spite of not being a perfect representation, sentences expressing states of affairs are the best general-purpose representation of the relata of preferences. This is a welcome conclusion, since sentences are also the best general-purpose representation of beliefs, and the best general-purpose representation of the objects of norms. If we are interested in building general models of mental states that include several of these elements, then the use of sentential representation is highly advisable since it provides unity and thereby simplifies investigations of connections between the categories.

8.3.2 *The Comparative Predicates*

There are two fundamental dyadic (comparative) value concepts, namely 'better' (strict preference) and 'equal in value to' (indifference) (Halldén 1957, p. 10).

We will use the symbols $>$ and $\equiv$, respectively, to denote them.[3] Furthermore, in accordance with a long-standing philosophical tradition, we will take $A > B$ to represent "B is worse than A" as well as "A is better than B", thus abstracting from whatever psychological or linguistic asymmetries that may persist between betterness and worseness.

The following two properties of the two comparative relations will be taken to be part of the meaning of the concepts of (strict) preference and of indifference:

$$A > B,\ B > A,\ \text{and}\ A \equiv B \text{ are pairwise mutually exclusive} \tag{8.1}$$

$$\text{If } A \equiv B \text{ then } B \equiv A \tag{8.2}$$

These two conditions combine to ensure that $>$ and $\equiv$ give rise to a fourfold classification of all pairs of objects of comparison:

$$A \text{ is equal in value to } B (A \equiv B) \tag{8.3}$$

$$A \text{ is strictly preferred to } B\ (A > B) \tag{8.4}$$

$$B \text{ is strictly preferred to } A\ (B > A) \tag{8.5}$$

$$\text{The comparison between } A \text{ and } B \text{ is undetermined } (A <> B)^4 \tag{8.6}$$

We will also assume that indifference is reflexive:

$$A \equiv A \tag{8.7}$$

Preference logic is usually not performed with $>$ and $\equiv$ as primitives. Instead, it is common to use 'at least as good as' (or more precisely: 'better than or equal in value to'), denoted $\geq$, as the sole primitive. With our three basic assumptions, the two sets of primitives are interdefinable with the definition $A \geq B \leftrightarrow (A > B) \vee (A \equiv B)$ in one direction and the two definitions $A > B \leftrightarrow (A \geq B) \,\&\, \neg(B \geq A)$ and $A \equiv B \leftrightarrow (A \geq B) \,\&\, (B \geq A)$ in the other (Hansson 2001a, p. 19).

The choice of primitives (either $\geq$ alone or both $>$ and $\equiv$) is a choice between formal simplicity ($\geq$) and conceptual clarity ($>$ and $\equiv$) (Hansson 2001b, pp. 321–322). For most purposes, the choice is not important, but for some basic conceptual purposes it is necessary to choose the option with $>$ and $\equiv$. (This will be exemplified in Section 8.5.3.) We therefore propose that $>$ and $\equiv$ be used as the primitive comparative value terms.

---

[3] For simplicity, we will leave out from explicit discussion two important topics in a formal account of preferences and related concepts: (1) The set of alternatives or options that $>$ and $\equiv$ range over. (2) The standard of evaluation, such as moral value, aesthetic value, or value *tout court*, that they refer to.

[4] The fourth category consists in the absence of any of the other three. This has consequences for its representation. Thus whereas $A > B$ will hold in a preference set $\mathbf{S}$ if and only if $A > B \in \mathbf{S}$, and similarly $A \equiv B$ holds in $\mathbf{S}$ if and only if $A \equiv B \in \mathbf{S}$, $A <> B$ holds in $\mathbf{S}$ if and only if $\mathbf{S} \cap \{A > B,\ B > A,\ A \equiv B\} = \emptyset$.

8.3.3 *Preference States*

As we have already emphasized, preference states are excised parts of the mental state, or state of the mind. A preference state cannot exist in isolation; therefore when we choose to treat it as a self-sufficient entity this is an idealization.

The most obvious way to represent a preference state is probably to let it be represented by the set of all sentences in the preference language that it endorses. This construction is similar to that of a belief set (corpus) that consists of all the sentences that the subject believes in, or is committed to believe in. In analogy to belief sets, a set consisting of all the sentences (in a given language) that represent preferences held by the subject will be called a *preference set.*

Just like belief sets, preference sets as defined here are closed under logical consequence. Hence, a person who subscribes to the preference sentence $A \geq B$ is assumed to also subscribe to the disjunctive sentence $(A \geq B) \vee (A \geq C)$. Furthermore, if transitivity is one of the background conditions, and she subscribes to both $A \geq B$ and $B \geq C$, then we assume that she also subscribes to $A \geq C$. The logical closure of the preference set may be seen as the outcome of a reflective equilibrium. Somewhat more modestly, we may interpret a preference set as representing, not the set of actually endorsed preference sentences, but the set of preference sentences that the agent is committed to endorse.[5]

This construction has the advantage of conforming with representations that we have reasons to choose for other parts of the mental state. If we have both a belief set and a preference set, then they can be combined in ways that facilitate the formal treatment of connections between the two.

However, the logical closure of belief sets and preference sets also has drawbacks. Many distinctions are lost in the process of logical closure. This problem has been highlighted in previous studies of belief change (Hansson 1999, pp. 17–24). One major problem with belief set models is that they allow for only one inconsistent state. The reason for this is that there is only one set that is both inconsistent and logically closed (namely the whole language). This is an unsatisfactory property of belief set models, since intuitively speaking there are many ways to hold inconsistent beliefs. This is a problem that belief modelling has in common with preference modelling. For example, it may be important for the understanding of possible preference state changes whether the agent violated transitivity (where consistency could for instance be restored by removing any one of the three relations $A > B$, $B > C$, $C > A$), or whether she violated asymmetry (where consistency could be restored by removing either $A > B$ or $B > A$).

One possible remedy is to replace belief sets by belief bases, sets that are not closed under logical consequence. Hence, instead of the belief set $\mathrm{Cn}(\{A, B\})$ we can use one of the belief bases $\{A, B\}$, $\{A\ A \rightarrow B\}$, $\{A \vee B,\ A \leftrightarrow B\}$ etc., all of which have the same logical closure and therefore correspond to the same belief set. For each belief set there are many belief bases. To the extent that we can give the distinction between belief bases a meaningful interpretation, much more

[5] This distinction was introduced for belief sets by Isaac Levi (1974, 1977, 1991).

information can be conveyed in this representation. Arguably there is a meaningful such representation, namely that the belief base consists of those beliefs that have an independent justification.

The same argumentation applies to preference states. In the same way, we can replace the preference set by a preference base that contains those preferences that the agent has actively accepted and that have survived subsequent changes. Elements of the preference base thus contrast with the merely derived preference statements that form the rest of the preference set. This allows us to distinguish between different ways of holding inconsistent preferences. For example, if someone prefers drinking tap water to drinking mineral water ($T > M$) then it follows that she either prefers drinking tap water to drinking mineral water or prefers drinking sewage water to drinking tap water, $(T > M) \vee (S > T)$. However, this latter element in her state of preferences is merely derived and will disappear as soon as $T > M$, from which it was derived, disappears. If she adopts the new preference $M > T$, to replace $T > M$, the option of retaining $(T > M) \vee (S > T)$ (which with $M > T$ yields $S > T$) does not even arise. In contrast, in a framework with preference sets, $(T > M) \vee (S > T)$ does not disappear automatically when $T > M$ is given up. Its elimination will have to be ensured with some priority-setting mechanism (see further Section 8.6.3).

8.3.4 Summary

In accordance with tradition, we propose the use of sentences denoting states of affairs as a general representation of the relata of value comparisons. This choice will facilitate combinations of preference states with other parts of mental states, since such sentences are also the best general-purpose representations of beliefs and of the objects of norms. We propose the use of strict preference ($>$) and indifference ($\equiv$) as the primitive comparative predicates, since they are conceptually more fundamental than the alternative primitive predicate ($\geq$).

There are two major alternative ways to combine these elements into preference states, namely (logically closed) preference sets and (logically open) preference bases. Both types of models are worth further investigations. In combined models including other mental entities such as beliefs it will mostly be advisable to use either closed or open representation throughout.

8.4 Integrity Constraints

8.4.1 Integrity Constraints Versus Priorities

With integrity constraints we mean requirements that a preference state has to satisfy in order to be an adequate representation of preferences according to the standards of the chosen model. Integrity constraints are exceptionless, i.e. they apply to all

preference states that are the outcome of some operation of change. (We can leave it open whether such an operation can be applied to a preference state not satisfying the constraints; the essential criterion is that the posterior states satisfy them.) Typical examples of integrity constraints are logical closure in preference set models and transitive closure in models of rational preferences that require transitivity.

Integrity constrains should be distinguished from input constraints that come with the specific input, and therefore do not apply to all preference states. A typical example of an input constraint is the requirement that the outcome of contracting some preference state by some non-tautological sentence (such as $A > B$) should be a new preference state not containing or endorsing that sentence.

Integrity constraints should also be distinguished from priorities and priority-setting mechanisms. A priority-setting mechanism, such as a selection function, incision function, or entrenchment relation, tells us for instance which of two elements in a preference set we should retain when the combination of integrity and input constraints prevents us from retaining both of them.

We propose, as a general strategy, that if there is a choice between expressing a condition as an integrity constraint or as a priority-setting principle, then the former option should be chosen. A major reason for this is that many integrity constraints can be included in the logic, which allows for a more unified formal treatment. In the next section we will proceed to show how this is done. Integrity constraints can often perform the function of priority-setting criteria, e.g. they can contribute to determining the choice between alternative ways to restore consistency when a new preference is added that is consistent with the set of previous preferences. In this way, priorities can to some extent be *endogenised* through incorporation into the logic (see further Section 8.6.3).

8.4.2 *Formalizing Integrity Constraints*

Many integrity constraints take the form of rationality postulates such as transitivity, $(X > Y)\ \&\ (Y > Z) \rightarrow (X > Z)$. In order to see how such postulates can be incorporated into the logic it is useful to focus on the consequence operator that is associated with the logic. The minimal consequence operator for our purposes is the classical truth-functional consequence operator Cn_0, such that for each set S of sentences, $\mathrm{Cn}_0(S)$ is the set of its truth-functional consequences. Rationality postulates will be represented by stronger consequence operators. Let T be a set of preference postulates. Then Cn_T is the operator such that $\mathrm{Cn}_T(S)$ consists of the logical consequences that can be obtained from S, using both truth-functional logic and the postulates in T. In other words,

$$\mathrm{Cn}_T(S) = \mathrm{Cn}_0(s(T) \cup S), \tag{8.8}$$

where $s(T)$ is the set of substitution instances of elements of T. As an example, if $(X \geq Y) \vee (Y \geq X) \in T$ and $\neg(A \geq B) \in S$, then $B \geq A \in \mathrm{Cn}_T(S)$.

(Clearly, $(A \geq B) \vee (B \geq A) \in s(T)$, and the rest follows truth-functionally.) The important observation that makes this construction work is that whenever Cn_0 is Tarskian consequence operator then so is Cn_T (Hansson 2001a, pp. 35–36).

For any model $\mathcal{M}$ of a preference state, let $|\mathcal{M}|$ be the set of preference sentences that it endorses. Then we can define $\mathcal{M}$ as consistent if and only if $\mathrm{Cn}_T(|\mathcal{M}|)$ is consistent. In this way, integrity constraints (such as transitivity) are expressible in terms of logical consistency. This makes the formal treatment much more unified than if we had to treat each integrity constraint separately.

8.4.3 *Internal Integrity Constraints*

We can divide the integrity constraints concerning preferences into two major categories, namely those that refer to relations among preferences and those that refer to relations between preferences and other mental objects. Of course only the former are directly relevant in a model that contains only preferences and no other mental entities.

One major class of such internal integrity constraints are properties such as completeness, acyclicity, transitivity, IP-, PI-, II-, and PP-transitivity that facilitate the use of the preference state for action-guiding purposes. The decision which of these potential constraints to include in a preference model will depend largely on the purpose of the model. In a descriptive model most of these properties will probably not be satisfied for the simple reason that people tend to violate them. Two major potential reasons have been given why one should honour these constraints. First, the standard *meaning* of preferences is held to be partly constituted by these constraints (Davidson 1980, p. 273). Secondly, it may be argued that preferences should have such a structure that they can be used to guide our choice among the alternatives that they cover. To make consistent choice-guidance possible, some integrity constraints will have to be satisfied.[6]

There is also another class of constraints, namely those that refer to logical relations among relata. Such relations are often excluded by the simplifying assumption that all objects of preferences are mutually exclusive, i.e. none of them is compatible with, or included in, any of the others. However, actual agents often have preferences that refer to compatible relata. One can prefer ice cream to fruit cake although it is quite feasible to have both. We should expect there to be logical connections among preferences that refer to logically related relata.

The following are two plausible such integrity constraints:

$$(X \geq Y) \rightarrow (X \geq (X \vee Y)) \,\&\, ((X \vee Y) \geq Y) \text{ (\textit{disjunctive interpolation})} \quad (8.9)$$

and its weaker variant

$$(X \equiv Y) \rightarrow (X \equiv (X \vee Y)) \quad (8.10)$$

[6] See Hansson (2001a, pp. 23–26) for a detailed discussion of what integrity constraints a preference relation has to satisfy in order to be useful for action-guidance.

Logicians have tried to construct these connections in two ways. The most common of these is the *holistic* approach that takes preferences over wholes for basic and uses them to derive the other preferences. The wholes chosen for this purpose have usually been possible worlds. Preferences over sentences can then be derived in various ways from preferences over the worlds in which these sentences are true (Hansson 2001a, pp. 57–113). Unfortunately this construction has the disadvantage of blatant cognitive unrealism. In practice, we are not capable of deliberating on anything approaching the size of completely determinate possible worlds.[7] The use of smaller holistic objects ("myopic holism") has been proposed as a means to overcome these difficulties (Hansson 2001a, p. 59).

The other approach has been called *aggregative*. It takes small units to be the fundamental bearers of value, and the values of larger entities are obtained by aggregating the values of these units. This means that preferences over states of affairs that describe only a small part of the state of the world are the fundamental bearers of value, and preferences over truth-functional combinations of these states are derived from them. A precise numerical aggregative model was developed by Warren Quinn on the basis of a proposal by Gilbert Harman. In that model (intrinsic) values are assigned to certain basic propositions, and precise rules are given for deriving the values of truth-functional combinations of these basic units (Harman 1967; Quinn 1974; Oldfield 1977; Carlson 1997; Danielsson 1997). This construction is based on the assumption that there are completely separable and evaluatively independent bearers of value, and that a numerical representation is available in which aggregate value is obtainable through addition of the values of these isolable units (Spohn 1978, pp. 122–129). Needless to say, these are strong and implausible assumptions (Moore 1903, p. 28).

We will leave open the issue how the relations between preference statements with logically related relata should be constructed. It should at any rate be clear that such connections belong among the integrity constraints in a model of preference change.

8.4.4 External Integrity Constraints

In models containing other mental elements in addition to preferences, integrity constraints should be included that refer to these combinations. Without going into details, we would like to mention a few examples.

[7] In studies of concepts or phenomena that one considers to be independent of cognition, it may be reasonable to abstract from cognitive limitations and use models with completely determinate possible worlds. This applies to some concepts of possibility; R.M. Adams has indeed claimed that "possibility is holistic rather than atomistic, in the sense that what is possible is possible only as part of a possible completely determinate world" (Adams 1974, p. 225). However, this argument for possible world modelling is not applicable to evaluative and normative concepts.

One plausible such constraint concerning preferences and *norms* is the following, for the ought operator O:

$$OX \ \& \ (\neg X \geq \neg Y) \rightarrow OY \ (\textit{contranegativity}) \tag{8.11}$$

If preferences are complete, this constraint will be equivalent with $OX \ \& \ \neg OY \rightarrow (\neg Y > \neg X)$. Hence, if you ought to pay your debt to your destitute neighbour, but you are not obliged to pay your debt to the car-dealer, then it would be worse of you not to pay your neighbour than not to pay your car-dealer.

Concerning the relationship between preferences and *beliefs*, one obvious and fairly plausible principle is *intersubstitutivity of relata believed to be logically equivalent*. In other words, if the agent believes two relata to be logically equivalent, then they should be exchangeable with no impact on truth or endorsement. For example, if a persons prefers ingesting 10 g of Vitamin C a day to not doing so, and believes that Vitamin C is ascorbic acid, then she is also committed to prefer ingesting 10 g of ascorbic acid a day to not doing so.

The topic of integrity constraints that connect preferences and beliefs is closely connected with central issues both in epistemology and value theory. As one example, we may ask: How and under what conditions should a change in factual beliefs give rise to a change in the values held by the subject? That such changes take place is obvious. Suppose that you prefer spending the evening with one person rather than another. Unless these are very entrenched preferences, they can be reversed if you acquire relevant new information about one of the persons. A common reaction to examples like this is that such a preference change is superficial or perhaps even illusory, and that your underlying preferences for what kind of person you spend the evening with are unchanged (cf. Stigler and Becker 1977). But can these underlying preferences be clearly delineated and in either case, what are the implications for preference change? These are issues worth a careful investigation.

8.4.5 Summary

By integrity constraints we mean requirements to be satisfied by all preference states, or at least by all preference states that are arrived at by an operation of change. Integrity constraints have the major advantage of being incorporable into the logic, which allows for a unified formal treatment. They should be distinguished from (1) input constraints that apply only to preference states that result from a specific operation of change, and (2) priority criteria that instruct us on the choice between different ways to satisfy the integrity and input constraints.

A model of preference change should include integrity constraints that mirror the logic of preferences. In addition, a model that includes other mental entities than preferences should include connecting constraints, such as constraints connecting preferences to beliefs and norms.

8.5 The Representation of Change

8.5.1 Input-Assimilation

The major models of belief change are *input-assimilating*: the belief state (either a belief set or a belief base) is exposed to an input, which imposes input constraints. As a result, the belief state is changed. The outcome of these changes is determined by the combination of integrity constraints, input constraints, and priority criteria, as outlined above.

We will adopt this general approach, but we will be very open concerning the nature of the inputs.

8.5.2 The Types of Belief Change

As a starting-point, let us consider the types of change that have been developed in belief change theory. There are three dominating kinds of belief change. In *expansion*, a specified sentence is added to the belief set. In *revision*, a specified sentence is added, and if needed, other sentences are removed in order to retain consistency. In *contraction*, a specified sentence is removed.

In addition to these, several other types of belief change have been proposed:

Consolidation: A belief base is made consistent by removing enough of its more dispensable elements. (Hansson 1994)[8]

Semi-revision: An input sentence that contradicts previous beliefs is accepted if it has more epistemic value than the original beliefs that contradict it. Otherwise the original belief state is retained. (Hansson 1997; Olsson 1997)

Selective revision: This is a generalization of semi-revision in which it is possible for only a part of the input information to be accepted. (Fermé and Hansson 1999; Gabbay 1999)

Shielded contraction: This is a version of contraction in which some non-tautological beliefs cannot be given up; they are shielded from contraction. (Makinson 1997; Fermé and Hansson 2001)

Replacement: One sentence is replaced by another in a belief set. Hence $K\left|_{q}^{p}\right.$ is a belief set that contains q but not p. Replacement can be used as a "Sheffer stroke" for belief revision. Contraction by p can be defined as $K\left|_{\mathrm{T}}^{p}\right.$, revision by p as $K\left|_{p}^{\perp}\right.$, and expansion by p as $K\left|_{p}^{\mathrm{T}}\right.$, where $\perp$ is falsum and T is tautology. (Hansson 2009)

Multiple contraction and revision: The simultaneous contraction (revision) of more than one sentence. (Fuhrmann and Hansson 1994)

[8] This operation cannot be meaningfully applied to belief sets, since there is only one inconsistent belief set. Once inconsistency has been reached in a belief set system, all distinctions have been lost, and they cannot be regained in an operation of consolidation.

An important feature in all but one of these operations is that they are defined in terms of one or several input sentences, although they differ in what is done with that input sentence (remove it, add it, add it if it has a sufficiently high position in the priority ordering, etc.). The exception is of course consolidation.

8.5.3 *Three Basic Types of Preference Change*

We will not take it for granted that the inputs of preference change should be determined by sentences in the same way as in belief change. Therefore, we will begin by classifying preference changes in terms of the relationships between the prior and the posterior preference state, for the moment making no assumption about the nature of the input.

Given two relata A and B, there are three fully specific comparative statements that can be made about them: $A > B$, $B > A$ and $A \equiv B$. There are also some other, less specified types of statements that we can make. Three of these are practically relevant, namely

$$A \geq B, \text{ that is equivalent to } (A > B) \vee (A \equiv B), \tag{8.12}$$

$$B \geq A, \text{ that is equivalent to } (B > A) \vee (A \equiv B), \text{ and} \tag{8.13}$$

$$A <> B \text{ that holds if and only if neither } A > B, B > A, \text{ nor } A \equiv B \text{ holds.} \tag{8.14}$$

We will leave out other, practically less relevant combinations such as $(A > B) \vee (B > A)$. This leaves us with six substates of the preference state with respect to the two relata A and B, namely $A > B$, $B > A$, $A \equiv B$, $A \geq B$, $B \geq A$, and $A <> B$. A change of preference concerning the two alternatives A and B consists in a move from one to another of these six states. Such changes group into three different kinds. (We leave aside the trivial type of "change" in which the prior and the posterior states coincide.)

The first type of change concerning A and B occurs when the part of the preference set that refers only to A and B is replaced by one of its proper supersets. This happens when an undetermined state is changed either into a weak preference, a strict preference or a value-equality, or when a weak preference is changed either into a strict preference or a value-equality. We call this type of change an *expansion* with respect to A and B. Figure 8.1 captures the intuitive notion of expansion.

The second type of change occurs when the preference set is replaced by one of its proper subsets. This happens when a value-equality or a strict preference is changed into a weak preference, or when a value-equality, strict preference or weak preference is transformed into an undetermined state. We call this type of change a *removal* with respect to A and B. Figure 8.2 clarifies the inverse relation between expansion and removal.

The third type of change occurs when the preference set is replaced by another, such that neither of them is a proper subset of the other. This happens when a strict or weak preference is changed into a strict or weak preference in the opposite direction,

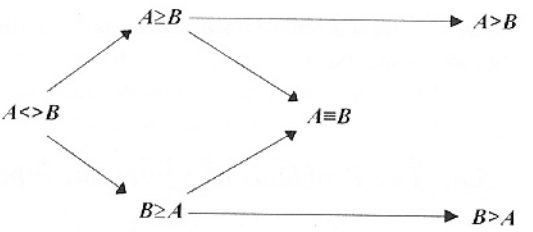

Fig. 8.1 Expansion of a preference state with respect to A and B

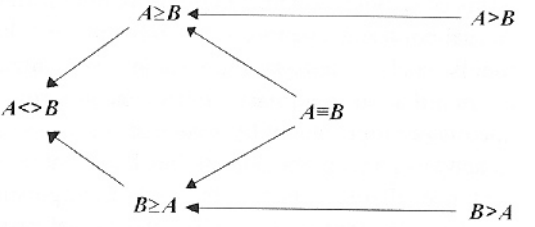

Fig. 8.2 Removal on a preference state with respect to A and B

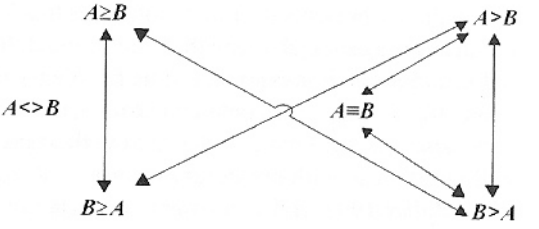

Fig. 8.3 Exchange in a preference state with respect to A and B

or when a value-equality is replaced by a strict preference, or vice versa. We call such a transformation an *exchange* with respect to A and B, since it consists in removing some relations from the preference set and adding others to it. See Fig. 8.3.

Of course, preference changes normally refer to more than just a single pair of relata. More generally, we will call a preference change

a *removal* if it is a removal with respect to all pairs that are affected by the change,
an *expansion* if it is an expansion with respect to all pairs that are affected by the change, and
an *exchange* if it is neither a removal nor an expansion.

This is an exhaustive categorization, i.e. all non-vacuous changes belong to exactly one of these categories. It is important to note that nothing has yet been said here about the input that gives rise to the change.

There is an obvious way to "sentencify" each of these three types of changes (Hanssson 1995):

A *removal* can be constructed as a contraction by some sentence(s).
An *expansion* can be constructed as an expansion by some sentence(s).
An *exchange* can be constructed as a revision by some sentence(s), or as a replacement by some pair of sentences.

However, even if such constructions are feasible it remains to determine whether they are plausible.

8.5.4 The Problems of Sentential Input Representation

As already mentioned, in standard accounts of belief change all inputs are defined in terms of sentences. This feature of belief change theory is far from unproblematic. Actual epistemic agents are moved to change their beliefs largely by non-linguistic inputs, such as sensory impressions. Sentential models of belief change (tacitly) assume that such primary inputs can, in terms of their effects on belief states, be adequately represented by sentences. Thus, when a person sees a hen on the roof (a sensory input), she adjusts her belief state *as if* she modified it to include the sentence "there is a hen on the roof" (a linguistic input).

There are many cases when the causal processes underlying belief formation take the form of accepting sentential information in the way that standard belief revision theory presents it. (Education relies to a large part on that mechanism.) There are also preference changes that can be modelled as caused by accepting sentential information. For example, some people try hard to prefer what they believe has value. If they succeed in their endeavours, the sentences that identify the preferences they aspired to will have contributed to their preference change. Something similar will be the case with preference changes induced by accepting an ideological doctrine (Zaller 1992). Further, when agents adopt preferences from their information about what people of high social standing consume (Leibenstein 1950), the sentential representation of these 'celebrity' preferences will play a causal role. And last, preference sentences will be causally efficacious in those cases where parental example and teaching form children's preferences (Cavalli-Sforza and Feldman 1981).

However, in many important classes of preference change, the actual causes of preference formation are decidedly non-sentential. This is most obvious with respect to *visceral preferences*. An increase in the concentration of the hormone leptin in the blood stream, for example, leads to a reduced desire to eat (Zhang et al. 1994). Other hormonal variations affect human sexual libido (Bullivant et al. 2004). Many new preferences are formed (and older ones lost) with increasing age. Thus was the experience of Shakespeare's Benedick: "but doth not the appetite alter? A man loves the meat in his youth that he cannot endure in his age" [*Much ado about nothing* act II, scene III]. Preferences for a romantic partner can slowly and unnoticeably erode, until the sudden realisation that one 'fell out of love' – as Bertrand Russell describes at his own example (1967, p. 195). Preferences can even be lost or formed through physical damage done to the brain tissue. For example, Oliver Sacks reports a case where an outbreak of neurosyphilis awakened in a shy elderly lady a preference for telling jokes and flirting. In another case, a carcinoma in the brain apparently transformed a reserved research chemist into an impulsive and facetious punster (Sacks 1987, pp. 97–111). In all these cases it is quite obvious that preference sentences are not part of the cause of the described changes. Philosophers of

all ages have acknowledged the importance of these kinds of preference change. "A man often believes himself leader when he is led; as his mind endeavours to reach one goal, his heart insensibly drags him towards another" (La Rochefoucauld 1871, maxim 43). "The heart has its reasons, which reason does not know" (Pascal 1958, Section 277). It seems at least possible that affects sometimes have a direct effect on action (compare the German 'im Affekt handeln'); momentary emotions dominate a person so strongly that they appear to be the only causes of her action.

Yet we want to resist the conclusion that there are two wholly distinct kinds of preference change. Instead, we argue for a unified account by pointing out that in most cases, preference affects are curbed by existing preference states. The ascete, no doubt, feels pangs of hunger, but resists eating and even the motivation to search for food. Some people in monogamous relationships feel sexually attracted to others, but resist the impetus to develop a preference for sex with others. Affects may be a source of motivation, but more often than not, they are filtered through the accepted preferences people hold.

Thus, affects can determine an action directly, or alternatively they can influence the accepted preference state and thereby the individual's actions. Figure 8.4 presents these two effects of preference affects.

This provides us with a cue to how a model of preference change can accommodate the wide variety of decidedly non-sentential causes of preference change, and yet allow for logical analysis. A distinction should be made between the formation of affects and the effects of affects on preference states. Affects can be taken as *primary inputs* that give rise to a *secondary input* in the form of a new preference pattern that has to be incorporated into the preference state.[9] This secondary input has to be expressible in the language of preferences. This gives rise to the structure of preference change theory presented in Fig. 8.5.

It is therefore possible to represent the input by a sentence or set of sentences, although this is of course an idealization. Models of belief change similarly idealize when modelling inputs as sentences, yet the difference between primary and secondary inputs appears to be more important in preference change. Thus, in this way, models of preference and belief change differ.

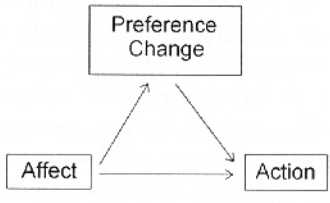

Fig. 8.4 Direct and preference-mediated effects of affects on actions

[9] We will continue to use the term "input" for "secondary input" when that can be done without causing confusion.

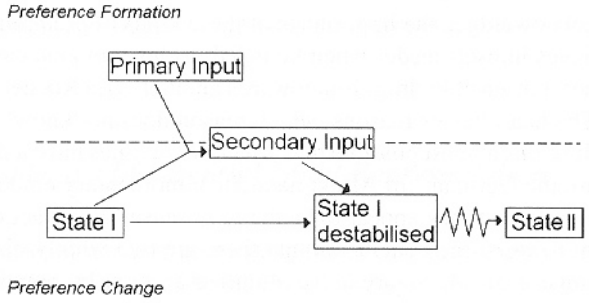

Fig. 8.5 The roles of primary and secondary inputs in preference change

8.5.5 Summary

Preference change, we argue, is exhaustively characterized by three kinds of relationships between prior and posterior preference states: removal, expansion and exchange. For many reasons, however, it is desirable to provide a more structured representation. We therefore propose an input-assimilating model of preference change, in which each change is seen as the reaction to a sentential input. Such a model requires the identification of a sentential input for each kind of preference change. Distinguishing between the formation of preference affects and the subsequent preference change proper allows such an idealized representation.

8.6 Priority-Setting

8.6.1 The Need for Prioritizing Mechanisms

In belief revision, the combination of the integrity constraints and the input constraints is not sufficient to determine the output. As an example of that, consider standard AGM belief contraction. Let the language be infinite. Let the belief set $\mathbf{K}$ be the closure of a single contingent sentence, $\mathbf{K} = \mathrm{Cn}(\{p\})$. Let q be any contingent sentence that follows logically from p. Our task is to contract $\mathbf{K}$ by q. Then there is an infinite number of belief sets that satisfy the integrity and input constraints for this operation, i.e. there is an infinite number of belief sets that satisfy the AGM postulates for being the result of contracting $\mathbf{K}$ by q (Hansson 2008). In view of this, it should be no surprise that belief revision theory makes abundant use of various formal methods to select among the contraction outcomes that are compatible with the integrity and input constraints. The most well-known such formal mechanisms are selection functions and entrenchment relations. In addition, belief bases can be seen as a means to set priorities. If we replace the belief set $\mathbf{K} = \mathrm{Cn}(\{p\})$ by any finite belief base B such that $\mathrm{Cn}(B) = \mathbf{K}$, then of course there are only a finite number

of contraction outcomes that satisfy the integrity and input constraints, given that being a subset of B is one of the input constraints.

Turning to preference change, the need for priority-setting mechanisms is more difficult to assess since we do not have a well-investigated canonical account of the integrity and input constraints as we have for belief change. However, it seems to be a realistic assumption that priority-setting mechanisms are needed here as well.

8.6.2 *Priority Information as a (Second-Order) Preference Ordering*

In any input-assimilating model, the posterior state is determined by the prior state and the input. Therefore, any priority information will have to be carried either by the prior state or by the input. In standard belief revision theory, all priority information comes with the prior belief state. An entrenchment ordering, for instance, is a prior ordering of the belief set that is one and the same for all inputs. In belief revision, a natural interpretation of the entrenchment ordering is that agents should give up beliefs that have as little explanatory power and overall informational value as possible. As an example of this, in the choice between giving up beliefs in natural laws and beliefs in single factual statements, beliefs in the natural laws, having much higher explanatory power, should in general be retained (Gärdenfors 1988).

For a theory of preference change, the technical correlate to such a priority ordering is an ordering of preferences. The notion of 'second-order preferences' (Sen 1977) or 'preferences amongst preferences' (Jeffrey 1974) has been used to investigate questions of morality, personhood, and akrasia. It can also be reinterpreted for the present purpose. Jeffrey offers the example of the 'good soldier', who prefers adopting his preferences to his orders, rather than following his appetites, fears, or moral judgments, and who thus has a second-order preference ranking for adopting certain sorts of first-order preferences on command (Jeffrey 1974, pp. 158–159). The good soldier thus has an ordering of his preference set that is the same for all inputs and identifies which preferences are to be excised first when in conflict with other preferences.

The usefulness of this notion, however, may be limited. While explanatory power and overall informational value provide an intersubjective criterion by which to interpret the priority ordering for beliefs, the values at the basis of second-order preferences are highly subjective and may shift at any moment. Jeffrey's 'good soldier' is more characteristic of a role that a person can take than of the person himself: at a moment's notice, the 'good soldier' may change into a 'conscientious citizen' or a 'self-preserving egoist'. Considering this instability raises uncomfortable questions about the applicability of second-order preferences to the regulation of (first-order) preference change.

8.6.3 Priority Information in the Preference Base

As we argued in Section 8.3.3, if the preference state is constructed with a preference base instead of a preference set, then there is another way to convey priority information, namely as encoded in the choice of which preferences to include in the preference base. For illustration, imagine two gourmets with identical preference sets, Hans and Peter. Hans prefers Japanese cuisine to Italian, and Italian to French. Consistency also commits him to prefer Japanese to French – even though he never compared the two cuisines. Peter also prefers Japanese cuisine to Italian, and Italian to French. But he, in contrast to Hans, has compared Japanese to French, and found that he preferred the former to the latter. Their respective preference bases look as follows:

Hans:
{*Japanese>Italian, Italian>French*}
Peter:
{*Japanese>Italian, Italian>French, Japanese>French*}

Hence their preference bases are statically equivalent in the sense that the logical closures of their respective bases are identical. Then, after coming together for a dinner of Italian and French food, the two conclude that in fact the latter cuisine is better than the former, and they both accept a preference for French over Italian. Registering these changes leads to the following preference bases:

Hans:
{*Japanese>Italian, French>Italian*}
Peter:
{*Japanese>Italian, French>Italian, Japanese>French*}

These bases are no longer statically equivalent: Hans no longer prefers *Japanese* to *French*, which he was previously committed to for reasons of consistency. As soon as his preference base changed, that preference was no longer needed and dropped out of his preference set. Not so for Peter, who had accepted that preference in its own right.

When, after a second dinner, our friends conclude that Italian cuisine is in fact better than Japanese, Peter faces a difficult choice adjusting his preferences, while Hans has no such problem.

Hans:
{*Italian>Japanese, French>Italian*}
Peter:
{*Italian>Japanese, French>Italian*}
or
{*Italian>Japanese, Japanese>French*}
or
{*Italian>Japanese*}

Hans simply changes his preference over *Japanese* and *Italian* (accepting the derived preference for *French* over *Japanese* as a matter of consistency). Given his preference base, that is his unique reasonable choice. Not so for Peter. In order to accommodate his new preference for *Italian* over *Japanese*, he has to give up one or both of his two other preferences.

This example shows how information about the origin of identical preference sets can lead to differences in how these preference sets are adjusted in the face of new preferences. Thus, preference bases contain relevant priority information.

8.6.4 *Input-Carried Priority Information*

In preference change there are strong reasons to consider whether priority information can be carried by the inputs. The reason for this is that the context of the (secondary) input that gives rise to it (namely the primary input) often contains priority information of a special kind. There are, for instance, two major ways to change one's preferences in order to accommodate a new preference representable as $A \geq B$: either you change the position of A in the preference ordering or that of B. The primary input often tells us which of these to choose: You get tired of brand A, and start to like it less than brand B that was your previous second choice. You learn that the political party X has changed its policies on unemployment insurance, and start to like it more than party Y, etc. (Hansson 2001a, pp. 46–47). This positional information is special to changes pertaining to the position of individual items in an ordering. It is not treated in the standard belief revision models, which focus on changes of membership of individual items in a set. Thus, the need to deal with this special kind of priority information is another feature that sets models of preference change apart from models of belief change.

To exemplify this, consider Mr. Myer who orders four newspapers transitively as follows:

$$A > B > C > D \tag{8.15}$$

Case (i): He learns that newspaper A has participated in a cover-up of severe crimes committed by its principal owner. As a consequence of this, he changes his opinion about A, and now considers it to be worse than D.

Case (ii): He finds out that newspaper D has improved dramatically since he last read it, and now finds it to be even better than A.

Intuitively, in case (i), the outcome of his preference change should be $B > C > D > A$, whereas in (ii) it should be $D > A > B > C$. In case (i), it is A's position that should be moved relative to D and to all other options ranked relative to D. In case (ii), it is D's position that should be moved relative to A and to all other options ranked relative to A.

However, if the (secondary) input is specified only as revision by a preferred sentence, then this difference will be lost since that sentence will be $D > A$ in both cases. This problem is solved by using composite inputs that specify both the input

Hans simply changes his preference over *Japanese* and *Italian* (accepting the derived preference for *French* over *Japanese* as a matter of consistency). Given his preference base, that is his unique reasonable choice. Not so for Peter. In order to accommodate his new preference for *Italian* over *Japanese*, he has to give up one or both of his two other preferences.

This example shows how information about the origin of identical preference sets can lead to differences in how these preference sets are adjusted in the face of new preferences. Thus, preference bases contain relevant priority information.

8.6.4 *Input-Carried Priority Information*

In preference change there are strong reasons to consider whether priority information can be carried by the inputs. The reason for this is that the context of the (secondary) input that gives rise to it (namely the primary input) often contains priority information of a special kind. There are, for instance, two major ways to change one's preferences in order to accommodate a new preference representable as $A \geq B$: either you change the position of A in the preference ordering or that of B. The primary input often tells us which of these to choose: You get tired of brand A, and start to like it less than brand B that was your previous second choice. You learn that the political party X has changed its policies on unemployment insurance, and start to like it more than party Y, etc. (Hansson 2001a, pp. 46–47). This positional information is special to changes pertaining to the position of individual items in an ordering. It is not treated in the standard belief revision models, which focus on changes of membership of individual items in a set. Thus, the need to deal with this special kind of priority information is another feature that sets models of preference change apart from models of belief change.

To exemplify this, consider Mr. Myer who orders four newspapers transitively as follows:

$$A > B > C > D \tag{8.15}$$

Case (i): He learns that newspaper A has participated in a cover-up of severe crimes committed by its principal owner. As a consequence of this, he changes his opinion about A, and now considers it to be worse than D.

Case (ii): He finds out that newspaper D has improved dramatically since he last read it, and now finds it to be even better than A.

Intuitively, in case (i), the outcome of his preference change should be $B > C > D > A$, whereas in (ii) it should be $D > A > B > C$. In case (i), it is A's position that should be moved relative to D and to all other options ranked relative to D. In case (ii), it is D's position that should be moved relative to A and to all other options ranked relative to A.

However, if the (secondary) input is specified only as revision by a preferred sentence, then this difference will be lost since that sentence will be $D > A$ in both cases. This problem is solved by using composite inputs that specify both the input

preference sentence *and* that relatum whose position is to be changed to accommodate the input sentence. Hence a revision by the preference $A \geq B$ can have a dyadic input that specifies not only $A \geq B$ but also for instance that it is the sentence A that is going to be moved around in the ordering, rather than B.

8.6.5 Summary

As in belief revision, theories of preference change require further criteria that help selecting among removal and exchange outcomes compatible with integrity and input constraints. We consider three sources of such priority information: a second order preference ordering, a preference base, and the input. We caution expectations about information from second-order preferences, as the possible approaches tend to yield frameworks that are too unstable. Instead, we suggest the use of exogeneous information from inputs, and in particular the use of endogenous information from preference bases.

8.7 Conclusion

A reader who hoped for a simple translation of belief change methodology to preference change may be somewhat disappointed at this point. We have argued that the general input-assimilating framework from belief change can be transferred, but we have also indicated several modifications that seem to be necessary. The input model has to be complicated with the introduction of a distinction between primary (non-linguistic) and secondary (linguistic) inputs. The method of sentential representation has to be used with somewhat more caution for preferences than for beliefs. Not least, the priority-setting mechanism has to be adjusted, and it seems useful to include some priority-related information in the inputs.

In summary, preference change cannot be successfully pursued as a straightforward application of belief change. It can make use of many concepts and methods from belief change but it is, definitely, a research area with its own specific problems and potentials in need of investigation.

References

Adams, Robert M. 1974. Theories of actuality. *Noûs* 8: 211–231.

Arrow, Kenneth. 1977. Extended sympathy and the possibility of social choice. *American Economic Review* 67: 219–225.

Brogan, Albert P. 1919. The fundamental value universal. *Journal of Philosophy, Psychology, and Scientific Methods* 16: 96–104.

Bullivant, Susan B., Suma Jacob, Martha K. McClintock, Julie A. Mennella, Sarah A. Sellergren, Natasha A. Spencer and Kathleen Stern. 2004. Women's sexual experience during the menstrual cycle: identification of the sexual phase by noninvasive measurement of luteinizing hormone. *Journal of Sex Research* 41 (1): 82–93.

Carlson, Erik. 1997. A note on Moore's organic unities. *Journal of Value Inquiry* 31: 55–59.

Cavalli-Sforza, Luigi Luca and Marcus W. Feldman. 1981. *Cultural transmission and evolution*. Princeton, NJ: Princeton University Press.

Chisholm, Roderick M. 1963. Supererogation and offence: a conceptual scheme for ethics. *Ratio* 5: 1–14.

Chisholm, Roderick M. and Ernest Sosa. 1966. On the logic of "intrinsically better". *American Philosophical Quarterly* 3: 244–249.

Danielsson, Sven. 1997. Harman's equation and the additivity of intrinsic value. *Uppsala Philosophical Studies* 46: 23–34.

Davidson, Donald. 1980. Hempel on Explaining Action. *Essays on Actions and Events*. Oxford: Oxford University Press.

Fermé, Eduardo and Sven Ove Hansson. 1999. Selective Revision. *Studia Logica* 63: 331–342.

Fermé, Eduardo and Sven Ove Hansson. 2001. Shielded contraction. In *Frontiers of belief revision*, eds. Hans Rott and Mary-Anne Williams, 85–107. Dordrecht, The Netherlands: Kluwer.

Fuhrmann, André and Sven Ove Hansson. 1994. A survey of multiple contractions. *Journal of Logic, Language, and Information* 3: 39–76.

Gabbay, Dov M. 1999. Compromise, update and revision. A position paper. In *Dynamic worlds*, eds. B. Fronhoffer and R. Pareschi, 111–148. Dordrecht, The Netherlands: Kluwer.

Gärdenfors, Peter. 1988. *Knowledge in flux. Modeling the dynamics of epistemic states*. Cambridge, MA: MIT Press.

Halldén, Sören. 1957. *On the logic of 'better'*. Lund, Sweden: Library of Theoria.

Hansson, Sven O. 1994. Taking belief bases seriously. In *Logic and philosophy of science in Uppsala*, eds. Dag Prawitz and Dag Westerståhl, 13–28. Dordrecht, The Netherlands: Kluwer.

Hansson, Sven O. 1995. Changes in preference. *Theory and Decision* 38: 1–28.

Hansson, Sven O. 1997. Semi-revision. *Journal of Applied Non-Classical Logic* 7:151–175.

Hansson, Sven O. 1999. *A textbook of belief dynamics, theory change and database updating*. Dordrecht, The Netherlands: Kluwer.

Hansson, Sven O. 2001a. *The structure of values and norms*. Cambridge: Cambridge University Press.

Hansson, Sven O. 2001b. Preference logic. In *Handbook of philosophical logic*, ed. Dov Gabbay and Franz Guenthner, 2nd ed, vol 4, 319–394. Dordrecht, The Netherlands: Reidel.

Hansson, Sven O. 2006. Ideal worlds – wishful thinking in deontic logic. *Studia Logica* 82: 329–336.

Hansson, Sven O. 2008. Specified meet contraction. *Erkenntnis* 69: 31–54.

Hansson, Sven O. 2009. Replacement – a Sheffer stroke for belief revision. *Journal of Philosophical Logic*, 38:127–149.

Harman, Gilbert H. 1967. Toward a theory of intrinsic value. *The Journal of Philosophy* 64: 792–805.

Jeffrey, Richard. 1974. Preferences among preferences. *Journal of Philosophy* 71(13): 377–391. Reprinted in *Probability and the art of judgment*, 154–169. Cambridge: Cambridge University Press, 1992.

Lee, Richard. 1984. Preference and transitivity. *Analysis* 44: 120–134.

Leibenstein, Harvey. 1950. Bandwagon, snob and veblen effects in the theory of consumer's demand. *Quarterly Journal of Economics* 64: 183–207.

Levi, Isaac. 1974. On indeterminate probabilities. *Journal of Philosophy* 71: 391–418.

Levi, Isaac. 1977. Subjunctives, dispositions and chances. *Synthese* 34: 423–455.

Levi, Isaac. 1991. *The fixation of belief and its undoing: changing beliefs through inquiry*. Cambridge: Cambridge University Press.

Makinson, David. 1997. Screened revision. *Theoria* 63: 14–23.

Moore, George. E. 1903. *Principia ethica*. Cambridge: Cambridge University Press.

Oldfield, Edward. 1977. An approach to a theory of intrinsic value. *Philosophical Studies* 32: 233–249.

Olsson, Eric J. 1997. A coherence interpretation of semi-revision. *Theoria* 63: 105–134.

Packard, Dennis J. 1987. Difference logic for preferences. *Theory and Decision* 22: 71–76.

Pascal, Blaise. 1958. *Pensées*. New York: Dutton.

Quinn, Warren S. 1974. Theories of intrinsic value. *American Philosophical Quarterly* 11: 123–132.

Reynolds, James and David Paris. 1979. The concept of choice and arrow's theorem. *Ethics* 89: 354–371.

Rochefoucauld Duc De La, Francois. 1871. *Reflections; or sentences and moral maxims*. Ed. J. W. Willis Bund and J. Hain Friswell. London: Simpson Low, Son, & Marston.

Russell, Bertrand. 1967–1969. *The autobiography of Bertrand Russell*, 3 vols. London: Allen & Unwin.

Sacks, Oliver. 1987. *The man who mistook his wife for a hat*. New York: Harper & Row, Perennial Library Edition.

Sen, Amartya. 1973. Behaviour and the concept of preference. *Economica* 40: 241–259.

Sen, Amartya. 1977. Rational fools. In *Choice welfare and measurement*, 84–106. Cambridge, MA: Harvard University Press, 1982.

Slote, Michael. 1984. Satisficing consequentialism. *Proceedings of the Aristotelian Society*, Supplementary volume 58: 139–164.

Spohn, Wolfgang. 1978. *Grundlagen der Entscheidungstheorie*, Monographien Wissenschaftstheorie und Grundlagenforschung, No. 8. Kronberg/Ts: Scriptor.

Stigler, George J. and Gary S. Becker. 1977. De gustibus non est disputandum. *American Economic Review* 67: 76–90.

Trapp, Rainer W. 1985. Utility theory and preference logic. *Erkenntnis* 22: 301–339.

Ullmann-Margalit, Edna and Sidney Morgenbesser. 1977. Picking and choosing. *Social Research* 44: 757–785.

von Wright, Georg H. 1963. *The logic of preference*. Edinburgh: Edinburgh University Press.

Zaller, John R. 1992. *The nature and origins of mass opinion*. New York: Cambridge University Press

Zhang, Yiying, Ricardo Proenca, Margherita Maffei, Marisa Barone, Lori Leopold and Jeffrey M. Friedman. 1994. Positional cloning of the mouse obese gene and its human homologue. *Nature* 372: 425–432.