Proposition-Preferences and World-Preferences: Connecting the Two Levels

Till Grüne-Yanoff

This paper discusses the meaning of expressed preference statements. A holistic explanation of preferences is proposed: preference relations between propositions are explained by preference relations over worlds. Only those world-preferences function as *explanans* which are maximally similar to the actual world, and which are maximally similar to each other. The concept of similarity as intuitive is rejected, and is interpreted instead with reference to causal structure: 'closest to the actual world' is interpreted as compatible with the causal structure of the actual world, and 'most similar to each other' as sharing the same causal background conditions.

It is generally agreed that preferences are propositional attitudes.[1] On the other hand, a typical deliberation goes as in the following example: 'I prefer eating seafood to meat, because each time I can choose between an instantiation of seafood and an instantiation of meat, I prefer the former'. Let's take 'eating seafood' as a proposition and 'eating an instantiation of seafood' as a maximally specified state description, and let's remind ourselves that propositions are often characterised as sets of possible worlds. Hence, if we mould those maximally specific states as possible worlds, we can take the seafood/meat example as a rough explanatory model: proposition-preferences are explained in terms of world-preferences.[2]

This may sound like preference holism.[3] If it is, it is of a very weak kind: preferences over worlds are used as explanans, but their primary status is not defended. In fact, I deem this position compatible with claims for the priority of preferences over propositions.[4] I believe that both preference rankings interact, providing support and consistency checks for the other level. Nevertheless, I will restrict myself to a one-directional explanation here.

The question is which world-preferences are to explain the proposition-preference. The propositions A and B are characterised by the sets of worlds $W^A = \{w_1{}^A,\ldots,w_m{}^A\}$ and $W^B = \{w_1{}^B,\ldots,w_n{}^B\}$, respectively. Which of the preference relations P between $w_i{}^A$ *and* $w_j{}^B$ are to explain the preference relation A>B?

Eight different approaches can be distinguished, according to three questions. How is the relation P to explain the relation >? Which members of W^A and W^B should feature in P? Which of those worlds should be compared?

First, the relation P can explain > in a strong and in a weak sense. In the strong version, A is preferred to B only if all reference worlds of A are preferred to all reference worlds of B. In the weak version, A is preferred to B only if no reference world of B is preferred to a reference world of A, and at least one reference world of A is preferred to a reference world of B.

Second, either all members of W^A and W^B feature in P, or only a proper subset of one or both of the sets is taken into consideration. This choice determines the *reference sets* for A and B, $W_R{}^A$ and $W_R{}^B$, with the *reference points* as its members.

Third, either all members of the reference set for A are compared with all members of the reference set for B, or certain comparison pairs $\{w_i{}^A, w_i{}^B\}$ are singled out.

The following table shows all eight approaches and the respective definitions of the preference A>B:[5]

		Strong	Weak
Reference sets: W^A and W^B	Comparison of all members of W^A and W^B	(1) $\forall i,j$: $w_i{}^A P w_j{}^B$	(2) $\forall i,j$: $\neg(w_i{}^B P w_j{}^A)$ & $\exists$ i,j: $w_i{}^A P w_j{}^B$
	Comparison of designated pairs of W^A and W^B	(3) $\forall i,\{w_i{}^A,w_i{}^B\}$: $w_i{}^A P w_i{}^B$	(4) $\forall i,\{w_i{}^A,w_i{}^B\}$: $\neg(w_i{}^B P w_i{}^A)$ & $\exists$ $i,\{w_i{}^A,w_i{}^B\}$: $w_i{}^A P w_i{}^B$
Reference sets: subsets of W^A and W^B, $W_R{}^A$ and $W_R{}^B$	Comparison of all members of $W_R{}^A$ and $W_R{}^B$	(5) $\forall i,j$, $w_i{}^A \in W_R{}^A$, $w_j{}^B \in W_R{}^B$: $w_i{}^A P w_j{}^B$	(6) $\forall i,j$: $\neg(w_i{}^B P w_j{}^A)$ & $\exists$ i,j: $w_i{}^A P w_j{}^B$, *with* $w_i{}^A \in W_R{}^A$, $w_j{}^B \in W_R{}^B$
	Comparison of designated pairs of $W_R{}^A$ and $W_R{}^B$	(7) $\forall i,\{w_i{}^A,w_i{}^B\},w_i{}^A \in W_R{}^A$, $w_i{}^B \in W_R{}^B$: $w_i{}^A P w_i{}^B$	(8) $\forall i$, $\{w_i{}^A,w_i{}^B\}$: $\neg(w_i{}^B P w_i{}^A)$ & $\exists i$, $\{w_i{}^A,w_i{}^B\}$: $w_i{}^A P w_i{}^B$, *with* $w_i{}^A \in W_R{}^A$, $w_j{}^B \in W_R{}^B$

But how are we supposed to make sense of definitions (3) to (8)? How can we identify the proper subsets $W_R{}^A$ and $W_R{}^B$? And how can we identify the pairs $\{w_i{}^A,w_i{}^B\}$? A common answer to these two questions has been given with the help of the concept of *similarity:* (i) A reference set for a proposition A is a set of those worlds most similar to the actual world. (ii) Those members of the reference sets that are most similar to each other constitute a pair.

We need to introduce these restrictions to prevent extreme and outlandish realisations of *A* or *B* from influencing the preference between the two. Deliberators evaluate propositions by how they are likely to be realised and what they are likely to affect, and they do not take into account all possibilities. Neither do deliberators compare all realisations of two propositions with each other: they simply do not have the resources for all the necessary relations involved, and this inability would effectively prevent explaining preferences over propositions at all. Hence from all the definitions, only (7) and (8) possess any plausibility. For reasons of simplicity, I will restrict myself to discussing (7).

But definition (7), based on similarity, turns out to be inadequate as well. I will briefly present and discuss two concepts of similarity to make this problem clearer.

von Wright (1972, pp. 146-147) conceived of possible worlds as being exhaustively representable as sets of logically independent atomic states of affairs. Two worlds can be compared in terms of the atomic states which represent them. Similarity then reduces to an arithmetical count of differences in these states. Such a conception of similarity, disregarding all formal problems,[6] leads to counterintuitive results when applied to definition (7).[7] It treats laws as being just as variable as facts, and it allows a violation of a small number of laws in order to maintain a larger number of facts. In contrast to this, a deliberating agent takes laws largely to be fixed and determines her preference for propositions according to her beliefs about how these propositions came about and what they will affect.

It seems that we do not get sufficient information from the logic itself to administer an appropriate notion of similarity. Hence we need to resort to concepts of similarity external to logic. An obvious starting point would be the work of Lewis, who takes similarity between worlds as intuitively primitive.[8] But his approach is contested exactly because of this approach to similarity, and two problems about it have been raised. First, the approach has problems getting the trade-off between laws and facts right. It remains questionable on this account why a world in which a law is violated (for the sake of maintaining particular facts) is less similar to the actual world than one in which the law is maintained and the facts are changed.[9] Second, this account takes divergences as 'small miracles' and hence disallows causal backtracking, i.e. inferences from the present into the past. Causal backtracking, nevertheless, is important for the explanation of preference: when deliberating about two propositions, it matters *how* these propositions came about.

I will not discuss these issues any further. These brief remarks must suffice to show that intuitive similarity is too flimsy a base for explanation.

Instead, I propose to base the explanation of preferences over propositions on a concept of causal structure. Causal structure will tell us (i) that those A-worlds which are compatible with the causal structure of the actual world are reference points for A and B and (ii) those reference points which share the same causal background conditions will constitute a pair.

A causal structure informs us that some variables are causes for others. I will employ Mackie's[10] concept of causes as *inus*-conditions. An *inus*-condition is the insufficient but non-redundant part of an unnecessary but sufficient condition for the existence of an effect. On the one hand, this notion is much simpler in its conceptual framework as well as in its technical representation than probabilistic models, in particular than those that are explicitly based on causal structure -- Bayes Nets. On the other hand, the *inus* concept allows a distinction that seems to be disregarded in most Bayes Nets frameworks: the distinction between necessary and sufficient causes.[11]

A causal graph $\langle V, \rightarrow \rangle$ consists of two components: a finite and consistent set V of states and an n-ary relation $\rightarrow$ defined over V. The first n-1 places of $\rightarrow$ then represent the individually necessary and jointly sufficient causes of the variable e which appears in the last place of $\rightarrow$. A causal graph $\langle V^A, \rightarrow \rangle$ is a graph with $A \in V$. Causal graphs are connected, i.e. $c \in V \Leftrightarrow c \in \rightarrow$. There might be more than one sufficient cause for e, so there might be more than one relation $\rightarrow$ with e in its last place. Different sufficient causes for A are represented in different causal graphs, which are labelled $\langle V^A_i, \rightarrow \rangle$.

A causal structure can be actual or possible. A causal relation is actual if it captures causal tendencies in the actual world. A causal relation is possible if it does not contradict the concept of a causal relation. But this modal distinction is different from the modality of possible worlds: a world may be non-actual but still comply with the actual causal structure. In fact, those are the worlds, I claim, which function as reference points for propositions which are not true in the actual world.

To construct the reference points for a proposition A, select all A-worlds which are compatible with the actual causal structure as represented in the graphs $\langle V^A_i, \rightarrow \rangle$. That is, in each of these worlds, at least one sufficient cause for A must be active. Often, many worlds will pass this test: A might have more than one sufficient cause, and A itself will be a necessary but insufficient cause for a host of effects. Quite a lot therefore depends on the causal background of each

world: those other states which function as sufficient causes in conjunction with A or with other necessary causes for A. To weed out some of those many worlds, a further selection takes place: some causes are considered factually impossible, their falsity asserted with certainty.

Once the A-worlds have passed both tests, the question remains how to compare them to the reference points of proposition B. It emerged from the above discussion that not all reference points are to be compared with each other, but only those which form pairs. With the help of the causal graph and the concept of causal background conditions, it is now possible to identify those pairs. Causal background conditions are those variables which vary freely across V^A_i and V^B_i.. Pairs, I claim, are constituted by those worlds with the most similar background conditions. As a similarity measure, I suggest an arithmetic count of differences between V^A_i and V^B_i, just as discussed by von Wright. I rejected this approach earlier on, but now, after distinguishing causal and factual modality, with the variables in V restricted to facts, this measure is appropriate.

This completes my discussion. I have shown why concepts of intuitive similarity are not sufficient for a definition of preferences between propositions in terms of preferences between worlds. I discussed a method to single out the reference points for definitions of type (7), by referring to causal structure. The advantage of this method lies in its straightforward interpretability. One the one hand, it allows a metaphysical perspective: by referring to true causes and facts. On the other hand, this framework can easily be reinterpreted as a representation of subjective beliefs in causes and facts, turning it into a type of non-quantitative decision theory.

E-mail: till.grune@infra.kth.se; website: www.infra.kth.se/~gryne.

NOTES

1 Or, at the least, desires are considered as propositional attitudes, and preferences can be derived from desires: if A is desired, then A is preferred to all B which are not desired. Or, presupposing that desires vary in strength: A is preferred to B if A is more strongly desired than B.

2 Maybe such an explanation appears trivial, or too shallow. According to the Humean perspective, this is all we can hope for: to explain preference with respect to another motivational concept.

3 For Preference Holism see Georg Henrik von Wright (1963), *The Logic of Preference* (Edinburgh: Edinburgh University Press), pp. 29-34; and Sven Ove Hansson (2001), *The Structure of Value and Norms* (Cambridge: Cambridge University Press), pp. 57-60. For the opposing view, see Gilbert Harman (1967), 'Towards a theory of intrinsic value', *Journal of Philosophy* 64: 792-804; Warren S. Quinn (1974), 'Theories of Intrinsic Value', *American Philosophical Quarterly* 11: 123-132; and

Erik Carlson (1997), 'The Intrinsic Value of Non-Basic States of Affairs', Philosophical Studies 85: 95-107.

4 Pettit claims that property desires are more fundamental than desires for an option. A property desire is a disposition to choose an option with the property rather than an option without. Property desires can be recast as preferences over propositions represented by existential sentences. Compare Philip Pettit (1991), 'Decision Theory and Folk Psychology' in *Essays in the Foundations of Decision Theory*, ed. Michael Bacharach and Susan Hurley (Oxford: Basil Blackwell).

5 von Wright (1963, p. 29) discusses and rejects definition (1) under the heading of *absolute preferences*. In a subsequent paper (Georg Henrik von Wright (1972), 'The Logic of Preference Reconsidered.' *Theory and Decision* 3: 140-169, p. 146) he defends definition (4). Rainer W. Trapp (1985), 'Utility Theory and Preference Logic', *Erkenntnis* 22: 301-339 seems to defend definition (6). Sven Ove Hansson (1989), 'A New Semantical Approach to the Logic of Preference', *Erkenntnis* 31: 1-42 defends definition (7) while he returns to (3) in his (2001, p. 77).

6 For a formal concern, see Hansson (2001, p. 76).

7 An agent has a hereditary disease which will eventually kill him. The actual world is represented by the fact D of him having the disease, the law L that the disease is fatal and that it is handed down to his children, the fact C that his children will have the disease and the fact M that he will die within a month's time. L, then, can be recast as D↔M&C. The agent prefers not having the disease, ¬D>D. We explain this by identifying {D,L,C,M} as the actual world, and as the reference set for D. The reference set for ¬D, according to von Wright's proposal, should be as close as possible to the actual world: the obvious candidate would be {¬D,¬L,C,M}, which scores lower in terms of difference in atomic states than {¬D,L,¬C,¬M}. Hence the agent prefers ¬D to D because he prefers world {¬D,¬L,C,M} to world {D,L,C,M}. That is a counterintuitive explanation: the plausible reference set for ¬D should be {¬D, L, ¬C, ¬M}.

8 David Lewis (1973), *Counterfactuals* (Cambridge, MA: Harvard University Press); (1979), 'Counterfactual Dependence and Time's Arrow', Nous 13: 455-476, reprinted in Lewis (1986a), *Philosophical Papers*, Volume II (Oxford: Oxford University Press): 32-51; (1986b), 'Postscripts to "Counterfactual Dependence and Time's Arrow"', in Lewis (1986a): 52-66.

9 For a discussion of the asymmetry between divergence and convergence, see Kit Fine (1975), 'Critical Notice', *Mind* 84: 451-458.

10 John L. Mackie (1980), *The Cement of the Universe* (Oxford: Clarendon Press).

11 The representation of *inus*-conditions as disjuncts in propositional logic produces counterintuitive results, as the rule of substitution allows modeling effects of a common cause as inus-conditions for each other (Nancy Cartwright (1989), *Nature's Capacities and their Measurement* (Oxford: Clarendon Press), p. 26). Hence I will not formulate the causal structure with the help of propositional logic. Instead, I will represent causal structure based on the *inus* concept as a directed, acyclical graph.