11 Why Don't You Want to Be Rich? Preference Explanation on the Basis of Causal Structure

Till Grüne-Yanoff

To explain people's behavior, we often cite their preferences. It is commonly accepted that to be explanatory, a preference—in combination with other mental states—must have brought about the behavior in question in the appropriate way. One condition for being an appropriate preference for this purpose is that the preferred alternative stands in a relevant relation to the behavior in question. This restricts the explanatory use of many preferences. For example, an agent's preference for coffee over brandy at this moment (at 8 a.m., after waking up) does not explain her choice of coffee over brandy at the end of the dinner party yesterday night. Instead, to explain yesterday's choice requires a preference over alternatives that stand in some abstracting relationship to yesterday evening's choices—maybe a preference for coffee over brandy after dinners, or a preference for nonalcoholic beverages. In order to be explanatorily useful, the most preferred alternative has to exhibit a degree of abstraction, so that the behavior in question can be related to it.

This condition easily comes into conflict with the need to empirically justify preference ascriptions. Preferences are mental states. Given that introspection does not provide a reliable epistemic basis, preferences cannot be directly observed, but can only be derived indirectly from observed behavior. From choices, however, one can only infer preferences over specific states of the world. For example, observing an agent choosing coffee over brandy for breakfast does not justify attributing to her a general preference for coffee over brandy. Even attributing to her a general preference to have coffee for breakfast, rather than brandy, would require observing her breakfast behavior under many different circumstances. Choice observations only justify attributing preferences over the very specific circumstances in which the choice was observed.

This presents a problem for explaining behavior with preferences. Only the most specific preferences can be derived from an agent's observable behavior. But if the outcomes over which preferences are defined are very specific, then they cannot be employed in explaining any behavior except for the very choices that justified their attribution in the first place. Such explanations would be trivial. So how can one ascribe preferences that are sufficiently abstract for explanatory purposes, in an empirically justified manner?

What is needed is a way to construct abstract preferences on the basis of specific preferences, such that the empirical justification of the specific preferences, based on observed choices, is preserved in the derived abstract ones. Indefinitely many degrees of abstraction can be distinguished. For simplicity, only two levels of abstraction are distinguished here—specific preferences over *worlds*, and abstract preferences over *prospects*. This paper develops a *principle of preference abstraction* that connects the world-preferences with prospect-preferences. The basis of this link, I will argue, is a model of causal belief.

It has been claimed that world preferences are not basic in decision making, and that instead we make decisions from prospect preferences to world preferences.[1] Some have argued from this claim against the methodological construction of abstract preferences from more specific ones.

> For a disposition to choose to count as a preference, it must be a disposition to choose with a reason—a disposition to choose on the basis of the properties displayed by the alternatives. . . . The equation of preferences with such brute [mere behavioral] dispositions is bound to seem inappropriate under the assumption of desiderative structure. And rightly so. After all, even if a person is disposed to choose one unconsidered prospect rather than another, he will be equally disposed, if possible, to consider the properties before making his choice. (Pettit 2002, p. 209; my italics)

This may be an argument against a kind of Methodological Behaviorism, but not against the methodological identification of prospect preferences from behavioral data. What needs to be distinguished is a metaphysical from a methodological meaning of "basic." While the atomism-holism debate remains undecided, it is not methodologically controversial that the only empirical justification of preferences can be obtained from choice observations. The principle of preference abstraction presented here, therefore, does not take a stance on the former debate, but is only constructed to clarify the role of preferences in explanation of behavior.

Section 1 presents the principle of abstraction as a selection problem of the worlds relevant for defining a preference relation between two prospects. Various restrictions on that function are proposed—first for conjointly exhaustive prospects only, and then for prospect pairs that are not exhaustive. Last, the specific problem of actions is discussed. Section 2 presents a model of the agent's causal beliefs. On the basis of this model, the selection function is specified. The resulting definition of prospect preferences allows characterizing the conditions under which this relation is reflexive, transitive, and complete. Section 3 addresses two problems for further research, and concludes the essay.

1 Prospect Preferences

To formally present this principle, I will make a number of assumptions. First, I assume that there exists a level of maximally specific states of the worlds, denoted $w_1, \ldots, w_n$. Second, a weak preference pre-order (i.e., a binary relation over worlds that is reflexive and transitive) is defined over these worlds, based on the agent's choices. For simplicity reasons, it will be assumed that all choices are made over certain outcomes.[2] Choices are made over certain, most specific outcomes—over worlds. Preferences over worlds are derived from these choices as follows: An agent (weakly) prefers w_i to w_j ($w_i \geq w_j$) if she chooses w_i over the available w_j. She is said to be indifferent between w_i and w_j ($w_i \approx w_j$) iff both $w_i \geq w_j$ and $w_j \geq w_i$. She is said to (strictly) prefer w_i to w_j iff $w_i \geq w_j$ and not $w_i \geq w_j$.[3]

Third, I assume that worlds are fully analyzable into conjunctions of certain prospects. A prospect can be the particular realization of a property, or a conjunction thereof, or the fact that a property is realized at all. Trivially, worlds are prospects as well. A further restrictive assumption I make is that of determinism. Ultimately, there is no uncertainty in any world; hence every world is fully analyzable into certain prospects. Prospects are denoted p, q, r and worlds are sets of the prospects into which they are analyzable—for example, $p \in w_i$.

Last, I assume deterministic causal relations to be defined over certain prospects. This relation is irreflexive, asymmetric, and acyclical, but not complete. It is interpreted as the beliefs an agent holds about the causal dependence of particular prospects.

The principle of abstraction that I propose comes in the guise of a definition of the preference relation $\succeq$ over prospects $p, q, \ldots$ in terms of the preference relation $\geq$ over worlds $w_1, w_2, \ldots$ It employs a representation function f that picks out pairs of worlds $\langle w^p, w^q \rangle$ for each pair of propositions $\langle p, q \rangle$:

Definition 1. $p \succeq q \Leftrightarrow w_i^p \geq w_i^q$ for all $\langle w_i^p, w_i^q \rangle \in f(\langle p, q \rangle)$

Definition 1 is trivial if the propositions p and q are worlds themselves. It becomes interesting if p and q are more abstract than the respective worlds. Then, from left to right, f picks out all those worlds that are specifications of the preference between p and q. Conversely, the preferences between all worlds picked out by f determine the preference between the prospects p and q.[4]

I will discuss the form of f in two separate installments. In the first step, I will focus on the special case where prospect preferences are only defined over a prospect p and its negation $\neg p$. In such a preference, mutually exclusive and conjointly exhaustive prospects are compared.[5] In the second step, I will discuss prospect preferences defined between mutually exclusive, but conjointly not exhaustive prospects.[6] This distinction is important, because the latter prospects feature in preference orderings beyond the pairwise level, while the former do not. Thus preferences over mutually exclusive, but conjointly not exhaustive prospects are subject to the transitivity property, and I will present an interesting result here.

1.1 Mutually Exclusive and Conjointly Exhaustive Prospects

In this subsection I will restrict myself to cases where definition 1 defines preferences over prospects and their negations only; preferences of the sort $p \succeq \neg p$. The way f picks out worlds is of central importance for the preference relation between prospects; definition 1 says nothing about it. There are at least three different doctrines about how to specify f.

The *absolute* preference approach stipulates that all worlds that are logically compatible with a prospect have to be taken into account. That is, any world w^p that contains a prospect p has to be preferred to any other world $w^{\neg p}$ that does not contain the prospect p. This very quickly leads to enormous numbers of world-comparisons necessary for the derivation of a prospect preference. For example, imagine worlds differentiated were by only four prospects, p, q, r, s. Then there would be $2^3 = 8$ different worlds that contain p, and 8 that do not. In the absolute preference approach, all possible $8^2 = 64$ comparisons between p-worlds and $\neg p$-worlds have to show a preference for p worlds, in order to derive the prospect preference $p \succeq \neg p$ from it.

In such a universe, let p be the agent's consumption of Marmite, q and r prospects irrelevant at the moment, and s the case that the agent is allergic to Marmite. Now, whether q and r are realized or not, as long as s is not, the agent prefers the world in which she consumes Marmite to the one where she does not. But, quite understandably, she does prefer the world where she is allergic to the stuff and does not consume it to worlds where she does consume it and suffers the allergic consequences of her actions. Should her preference between those last two worlds determine her prospect preference over Marmite consumption? I don't think so. The scenario is *counterfactual*; she does not actually suffer from the allergy. This does not mean that counterfactual scenarios do not have any influence on prospect preferences; I will show further down that they do. But in this case, the counterfactual scenario is *causally independent* of the prospect in question; Marmite consumption does not cause Marmite allergy. The absolute account does not allow this abstraction and thus should be discarded.

The *ceteris paribus* preference approach stipulates that only those worlds are taken into account, which are as similar as possible to each other, while realizing and not realizing the prospect in question respectively. That is, any world w^p that contains a prospect p has to be preferred to that other world $w^{\neg p}$ which is as similar to w^p in as many aspects as possible.[7]

For illustration, let's imagine that the four aspects of our four-aspect worlds are logically independent. Then, clearly, there is exactly one w^p-world that is most similar to one $w_{\neg p}$-world: namely that world that shares with w^p the realization or nonrealization of all aspects but p. According to the ceteris paribus approach, then, there are only eight comparisons between the four-aspect-worlds necessary to establish prospect preferences. This can be illustrated in table 11.1, where the numerals in the columns signify the realization or nonrealization of an aspect in the respective world.

Table 11.1 shows the sufficient conditions for $p \succeq \neg p$ according to the ceteris paribus approach. Each world in which p is realized is compared with the world in which p is not realized, but which is otherwise as similar as logically possible. If all aspects are logically independent—that is, no aspect is implied by any other aspect nor implies any other aspect—then the

Table 11.1
Ceteris paribus comparisons

w_i^p	p	q	r	s		$w_i^{\neg p}$	p	q	r	s
(1)	1	0	0	0	$\geq$	(1)	0	0	0	0
(2)	1	1	0	0	$\geq$	(2)	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
(8)	1	1	1	1	$\geq$	(8)	0	1	1	1

two worlds compared differ only in the realization of p. As we are free to choose how to partition the worlds into prospects, we can avoid partitions with logically dependent prospects. Thus the comparisons will always look like the one illustrated in table 11.1.

There are two fundamental problems with the ceteris paribus account. First, it rests on a concept of logical possibility, which is too wide for the purpose at hand. Second, it disregards the world the agent is in when making the comparison. The following example will illustrate both of these shortcomings.

Diogenes Laertius, the ancient chatterbox, tells of an incident where Alexander the Great puts Diogenes of Sinope to the touch. ''Ask of me any boon you like,'' the Macedonian is reported to have offered; to which the reply came, ''Stand out of my light.''[8] The anecdote is quite popular, and rightly so. At first sight, Diogenes seems to act contrary to a knee-jerk reaction of most of us. You are offered wealth or power for free—then take it! In this version of the story, Alexander embodies the ancient idea of Kairos, Machiavelli's Fortuna or, if you will, one of the brothers Grimm's good fairies. When Diogenes declines the seemingly irresistible offer, he must have good reasons for it.

As revealed in his choice, Diogenes prefers a world w^u undisturbed by any patron, however powerful, to a world w^o which promises all the wealth and influence Alexander has to offer. If we now think that the two worlds differed in only one relevant aspect, wealth, we could derive Diogenes' preference for poverty over wealth. But even though we do not know much about them, we can suspect that Diogenes' other choices could not have been subsumed under such a simple prospect preference. Even that most hardened despiser of material wealth, we suspect, must have seen that wealth and power were desirable for him too—he could have survived without panhandling, he could have bought his freedom, or he could have convinced the elders of Sinope to remove the ban and let him return to his homeland. So had Alexander asked (his immediate reaction is not reported), in slight astonishment, ''But don't you want to be rich?'' Diogenes' answer, if for once straightforward, would have also been complex. ''On the one hand,'' he would have retorted, ''there is a sense in which I want to be rich. But on the other hand look at the world I live in—if I took a significant boon from you, I would be obliged to show my gratitude. Further, my lifestyle would be considered implausible; and people would envy me for my easily achieved wealth. Under these conditions, I do not want to be rich.''

With this extra bit of information, we may try to apply the ceteris paribus framework for an analysis of Diogenes' preferences. According to the account I have put into his mouth, Diogenes identifies four aspects of w^u and w^o to be relevant—wealth (r), independence from donors (i), personal credibility (c), and the envy of others (e). Clearly, all these aspects are logically independent. Thus the specification of f in the table 11.1 applies. According to it, Diogenes compares $w_1^u = \{\neg r, \neg i, \neg c, \neg e\}$ with $w_1^o = \{r, \neg i, \neg c, \neg e\}$; $w_2^u = \{\neg r, i, \neg c, \neg e\}$ with $w_2^o = \{r, i, \neg c, \neg e\}$, and so on. Whatever his preferences between those worlds are, and whatever the resulting prospect preferences are, this specification of f does not capture his story at all if it goes: ''On the one hand, I want to be rich. But on the other hand, look at the world I live in . . .'' There, he compares $w_i^u = \{\neg r, i, c, \neg e\}$ with $w_i^o = \{r, \neg i, \neg c, e\}$. According to the ceteris paribus account and the logical independence of the aspects, such a comparison is not admissible, because the worlds are too far apart. So does Diogenes tell us an incoherent story, or is the ceteris paribus approach wrong?

I propose that it is the ceteris paribus approach that is flawed. Diogenes does not employ logical but *causal* possibility when assessing the independence of the worlds' aspects. He envisages a particular way in which he can achieve wealth—through his submission to a donor. As he tells us, he believes in the causal dependence of the other relevant aspects on this genesis of wealth. His wealth would cause the envy of others; his submission to a donor would cause the loss of his independence, which in turn would cause the loss of his credibility. Given the causal dependence Diogenes believes in, worlds that are most similar to w^u except for the realization of wealth are not those the ceteris paribus account suggests. It is causally impossible for Diogenes to be wealthy without being envied; it is equally impossible for him to be wealthy through the benefits of a donor without becoming dependent on him, and hence losing his credibility.

Even though these worlds are logically possible, what matters for a principle of abstraction is *causal* possibility. Logical possibility only forbids what is inconsistent, while causal possibility allows only what can be produced. An agent takes only those p-worlds as possible that are producible according to her causal beliefs. This epistemic notion of causality will restrict the selection function in the following way:

Restriction 1. f picks out only those worlds that are causally compatible with p and $\neg p$, respectively.

But this restriction alone is not sufficient for the right choice of f. The causal structure an agent believes in restricts the worlds she will deem possible; but she will not compare all possible worlds, as some of them are too far removed from her actual situation. Thus, facts believed to be true play a role too.

To stay with the above example, Diogenes might reasonably believe that secretly inheriting from a distant relative causes one to be wealthy without any strings attached. Thus, such a causal story would allow him to introduce into definition 1 the world w^o where he is wealthy, independent, credible, and not envied, due to the secrecy of the inheritance. So it might seem that because of the possibility that this belief opens, Diogenes does not prefer poverty to wealth unconditionally. It seems he only prefers it conditional on other aspects, in this case the absence of any living patron.

This appearance is wrong. Diogenes does not have any wealthy relatives from whom to inherit (or at least, we, as the interpreters of his behavior, do not know of any). To define his prospect preferences, we do not take into account *all* the causally possible worlds that realize the relevant prospects; we only take into account the causal possibilities that can be realized in the world the agent is in.

The above preference expression should therefore be interpreted as taking the relevant causal background conditions to be the same as in the actual world. Of course, not all background conditions can be the same—otherwise no counterfactual world could be constructed that adheres to the causal structure. For Diogenes to imagine a world in which he is wealthy—seen from his actual predicament of poverty—a counter*factual* assumption is necessary; but changing the facts does not mean changing the causal dependency structure. The change of facts, under stable causal dependencies, will require certain causally prior prospects to change as well—somehow, his wealth has to be caused in this possible world. But there are facts in the actual world that offer themselves as ready causes; there are donors offering their support, but there are not any wealthy, distant relatives ready to bequeath estates to Diogenes. Those facts that do not have to be changed in order to accommodate the counterfactual—either because there is no causal link to them at all, or because there are other causes closer to the actual situation—remain as they are in the actual world. Hence,

Restriction 2. f picks out only those worlds that realize p and q but maximally comply with the background conditions pertaining to the actual world.

Under the two restrictions on f for which I argued here, we can indeed say that Diogenes preferred poverty to wealth unconditionally; f identifies the necessary preferences over worlds, and definition 1 determines a prospect preference on that basis. In this sense, definition 1 is a principle of preference abstraction for preferences over mutually exclusive and conjointly exhaustive prospects.

1.2 Mutually Exclusive and Conjointly Nonexhaustive Prospects

Prospect preferences are not only used in the sense that one prefers the realization over the nonrealization of a prospect, as Diogenes prefers poverty over wealth, according to the scheme $p \succeq \neg p$. Preferences also occur in contexts where the two relata do not exhaust the alternatives. For example, over breakfast I prefer reading an English paper to a German one; and I prefer a German to a Russian newspaper. These three types of newspapers certainly do not exhaust the possibilities of breakfast reading, nor do they exhaust my ordering of breakfast readings. However, it is perfectly intelligible to hold preferences between conjointly nonexhaustive prospects; the problem only is that such preferences cannot be represented as cases of the scheme $p \succeq \neg p$.

Conjointly nonexhaustive relata occurring in preference types $p \succeq q$ are not necessarily mutually exclusive. For example, one can meaningfully hold a preference like "I prefer an apartment in New York to a house in

Tuscany," even though it is clearly possible to own an apartment in New York and a house in Tuscany at the same time.[9] However, to express a preference $p \succeq q$ without meaning to express a preference for $p \wedge \neg q$ over $q \wedge \neg p$ is likely to violate Grice's Cooperative Principle (Grice 1989). In particular, if uttered in a situation of choice between either p or q, the conversational contribution made does not satisfy the pragmatic convention of relevance—preferences over relata involving $p \wedge q$ do not help making such a choice.

If uttered in a situation where information about the speaker's evaluations is sought, it does not satisfy the pragmatic convention of informativeness; $p \wedge q \succeq q \wedge p$ is tautological, and thus empirically empty. By conversational implication, then, a preference between mutually nonexclusive relata is interpreted as a preference between the corresponding mutually exclusive relata (Halldén 1957, p. 28). This conventional translation procedure has to be amended for cases where at least one relatum logically implies the other or causally requires the presence of the other. Thus $p \succeq q$ is translated to $p \wedge \neg q \succeq q \wedge \neg p$ only if it is possible that $p \wedge \neg q$ and $q \wedge \neg p$. In cases where it is not, the original relatum remains untranslated (cf. Hansson 2001, pp. 68–70). Thus restriction 1 needs to be reformulated for conjointly nonexhaustive prospects in the following way.

Restriction 3. f picks out only those worlds that are causally compatible with $p \wedge \neg q$ and $q \wedge \neg p$, respectively.[10]

Concerning the actual causal background, the similar restriction holds as for the conjointly exhaustive case. Trapp (1985) gives an example for preferences over different diseases (that are not conjointly exhaustive). A man who prefers contracting cholera to being ill with cancer should not be interpreted as preferring a situation where there is no cure for cholera (e.g., in a country where there are no antibiotics available). The belief in the existence of a cure has significance consequences if one has either cholera or cancer, and hence naturally plays a crucial role in the evaluation of both situations. Thus, the agent prefers cholera to cancer iff he prefers worlds where he has cholera and all the contemporary cures are available, to worlds where he has cancer and all the contemporary cures are available. The restriction is thus reformulated as follows:

Restriction 4. f picks out only those worlds that realize $p \wedge \neg q$ and $q \wedge \neg p$ but maximally comply with those background conditions pertaining to the actual world.

A particularly interesting feature of preferences over conjointly nonexhaustive prospects is that the pairwise comparisons *may* give rise to a preference ordering—but not necessarily. Under particular conditions, the preference pairs $p \succeq q$ and $q \succeq r$ imply the additional preference pair $p \succeq r$. This transitivity property of preferences need not be fulfilled by prospect preferences, even though it is (by assumption) satisfied by the preferences over worlds underlying it. All that needs to be established is that the worlds $w^{p \wedge \neg q}$ (compared in $p \succeq q$ with the worlds $w^{q \wedge \neg p}$) and the worlds $w^{r \wedge \neg q}$ (compared in $q \succeq r$ with the worlds $w^{q \wedge \neg r}$) are not the same as the worlds $w^{p \wedge \neg r}$ and $w^{r \wedge \neg p}$ compared in $p \succeq r$. Thus, if $w^{p \wedge \neg q} \neq w^{p \wedge \neg r}$ and $w^{r \wedge \neg q} \neq w^{r \wedge \neg p}$, it does not follow from $w^{p \wedge \neg q} \geq w^{q \wedge \neg p}$ and $w^{q \wedge \neg r} \geq w^{r \wedge \neg q}$ that $w^{p \wedge \neg q} \geq w^{r \wedge \neg q}$. Hence, by definition 1, it does not follow from $p \succeq q$ and $q \succeq r$ that $p \succeq r$.

1.3 Actions

An action is a particular kind of prospect. Distinguishing actions from other prospects here is important because agents evaluate their own actions, and sometimes those of others, in a different way than they evaluate other prospects. While the evaluation of a prospect takes into account all causal antecedents of that prospect, the evaluation of an action only takes into account the action itself and all its consequences, while disregarding any causal history that led to the action.

Take Diogenes' example again. The only way for him to achieve wealth would have been to submit to a donor, which in turn would have had consequences for his independence and credibility. All in all, he preferred a world without those consequences to a world with them; thus, he preferred poverty to wealth. But if he took those indirect consequences of wealth and poverty into account, shouldn't he cast the net even wider? Should he not also take into account the causes of his own action, and other effects that those causes brought about?

Let us push the Diogenes story one step further. Imagine that his propensity to reject a potential sponsor is based on his contempt for authority. This character trait *causes* Diogenes (he believes) to be so disposed. It also *caused* him (he suspects) to rebel against paternal authority, shaming his father and bringing disgrace to his family—consequences he found utterly undesirable. To have to choose between being rich and being dependent upon Alexander's offer will reminds him of his character trait and its consequences. Insofar as the prospect that Diogenes stays poor is realized only in a world in which he rejects Alexander's offer, that world then also brings with it his rebellious character trait, his father's shame, and his family's disgrace.

Should he derive his prospect preferences from a comparison between worlds that honor these causal dependencies? Some claim so: Actions should not be treated differently from other prospects.

> [T]o the extent that acts can realistically be identified with propositions, the present notion of preference is active as well as passive: it relates to acts as well as to news items ... From this viewpoint, the notion of preference is neutral, regarding the active passive distinction. If the agent is deliberating about performing act A or act B, and if AB is impossible, there is no effective difference between asking whether he prefers A to B as a news item or as an act, for he makes the news. (Jeffrey 1983, p. 84)

On Jeffrey's account, Diogenes takes his rejection of a donor as the news of his character trait and its consequences. Presumably, what Jeffrey means by "he makes the news" is that there is no further causal history to an action that carries news characteristics. But the above example shows that this assumption is not generally true. By observing his own choice, the news that he makes includes information about the cause of his choice—his character trait—and further effects of that cause. If Diogenes evaluated his action just as any other prospect, he would take those effects into account, comparing worlds in which his contempt for authority disposes him to reject his sponsor and shame his father, with worlds in which he had not humiliated his father, accepts Alexander's offer, becomes wealthy, dependent, and loses his credibility. If he preferred the latter to the former, given the causal dependencies, he may indeed have prefer being wealthy to being poor.[11]

I think this model of evaluation is flawed. Neither Diogenes nor any other responsible actor takes into account the causes of their actions, and the effects of these causes, when evaluating their actions. An agent who evaluates a non-action state of the world takes a passive outlook—he takes into account what consequences this state has, and how this state came about, with the other consequences which that cause witnessed. An agent who performs an action exhibits an active outlook—she chooses between various options according to the benefit of their consequences; but she takes the world as it is, disregarding any influences that might have caused her action.

> Statements that describe acts are different in kind from other sorts of propositions simply because the actor has the power to make them true. With this power comes a kind of responsibility. An agent must, if rational, do what she can to change things for the better . . . rational decision makers should choose actions on the basis of their efficacy in bringing about desirable results rather than their auspiciousness as harbingers of these results. Efficacy and auspiciousness often go together, of course, since most actions get to be good or bad news only by causally promoting good or bad things. In cases where causing and indicating come apart, however, the causal decision theorist maintains that it is the causal properties of the act, rather than its pure evidential features, that should serve as the guide to rational conduct. (Joyce 1999, p. 150)

Acts must be considered exogenous. Thus Diogenes should disregard the causes of his choices and their respective effects, when evaluating the prospects of wealth and poverty respectively. Instead, he compares worlds that replace the effects of the cause of his action with what he actually believes happened, irrespective of what action he chose. The principle of abstraction is therefore amended for the case of actions.

Restriction 5. If p and q are actions, f picks out all those worlds that are compatible with $p \wedge \neg q$ and $q \wedge \neg p$, respectively, and their respective causal consequences, while disregarding their causal histories.

The disagreement between the two positions sketched out here remains, however, insofar as prospects often cannot be clearly identified as actions or non-actions. Thus the allies of Jeffrey might be right in insisting that some apparent actions are evaluated as news items. This does not touch on the basis of the argument, and is of no further relevance here. With these amendments added to the specification of f, definition 1 is a principle of abstraction for all prospect preferences.

2 Constructing the Selection Function

The concepts of causal compatibility, maximal compliance with the actual world, and causal history so far have been given only intuitive meaning. This section seeks to specify their meaning more formally, by reference to a formal concept of causal models.

A causal model is defined by Pearl (2000, p. 203) as a triple

$M = \langle U, V, G \rangle$

where:

1. U is a set of background variables, determined by factors outside the model.

2. V is a set of endogenous variables, determined by variables of the model—that is, variables in $U \cup V$.

3. G is a set of functions $\{g_1, g_2, \ldots g_n\}$ such that each g_i is a mapping from $U \cup (V \backslash V_i)$ to V_i and such that the entire set G forms a mapping from U to V. In other words, each g_i tells us the value of V_i given the values of all other variables in $U \cup V$, and the entire set G has a unique solution $V(u)$. Symbolically, G is represented by

$V_i = g_i(V_j, U_i), \quad i = 1, \ldots, n.$

$U_i \subseteq U$ stands for the unique minimal set of variables in U sufficient to determine V_i on the basis of G.

The variables in Pearl's model are random variables. I take the individual realization of a random variable to be equivalent to a prospect, for example, $p \equiv (V_i = v_i^1)$. Given a particular constellation of background variables, $U_1 = u_1, \ldots, U_n = u_n$, the model has the unique solution $V(u_1, \ldots, u_n)$. Prospects can be directly deduced from this solution: $V(u_1, \ldots, u_n) \vdash p$, where $\vdash$ is the classical inference relation.

M can be represented as an acyclical directed graph, with the arrows representing the function g. Forked arrows show that g has more than one argument. Figure 11.1 is an illustrative example of such a representation of $M^* = \langle U^*, V^*, G^* \rangle$, with all variables in $U^* = \{U_1, \ldots, U_4\}$ and $V^* = \{V_1, \ldots, V_4\}$ having only two realizations each, and $G^* =$

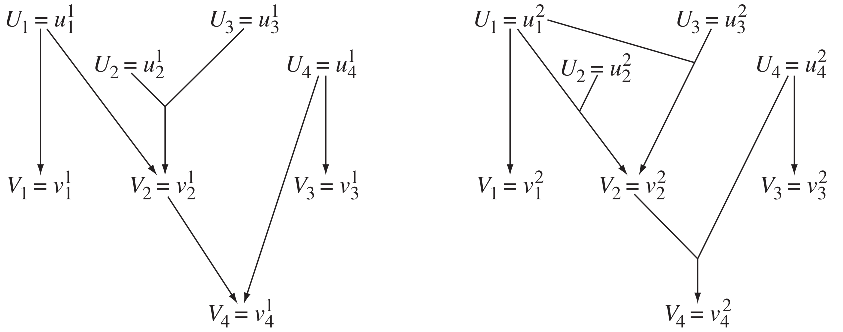

Figure 11.1
An example of a causal graph.

$\{V_1 = g_1(u_1), \dots, V_4 = g_4(v_2, u_4)\}$. Each realization is equivalent to a proposition or its negation. Let the first realization of a background variable be expressed by $a, b, c \dots$, that is, $a \equiv (U_i = u_i^1)$, etc.; the realization of a endogenous variable by $p, q, r, \dots$, that is, $p \equiv (V_i = v_i^1)$, etc.; and the respective second realization by a negation of that proposition, $\neg p \equiv (V_i = v_i^2)$, etc.

Each world w specifies the values for all U_i and for all V_i of every M. Because the functional relationships g_i of M restricts the endogenous V's given the exogenous U's, many worlds are inconsistent with a specific causal model.

Definition 2. A world w is *consistent* with a causal model $M = \langle U, V, G \rangle$ iff there is a set of realizations u^*, such that $U_1 = u_1, \dots, U_n = u_n$, for which $w \vdash u^*$ and $w \vdash V(u^*)$.

For example, the world $w_1 = \{a, \neg b, c, \neg d, p, q, r, \neg s\}$ is consistent with the model M represented in figure 11.1, while the world $w_2 = \{\neg a, b, \neg c, d, p, q, r, s\}$ is not. Having specified the relations between prospects and worlds on the one hand and the causal model and its variables on the other, we can now define causal compatibility:

Definition 3. w is *causally compatible* with p with respect to M iff there is a causal model $M = \langle U, V, G \rangle$ such that w is consistent with M, and $w \vdash p$.

For example, the world $w_1 = \{a, \neg b, c, \neg d, p, q, r, \neg s\}$ is compatible with p with respect to M. Worlds which are causally compatible with p represent the possible causal histories of p. In such a world there is at least one "chain" that leads from background conditions to p in the following way.

Definition 4. A prospect p is dependent with respect to the background conditions in $U^* \subseteq U$ iff there is a functional chain: $V_1 = g_1(u^*)$, $V_2 = g_2(V_1, u^*), \dots, V_n = g_n(V_1, \dots, V_{n-1}, u^*)$ with $g_1, \dots, g_n \in G$ and p being equivalent to $V_n = g_n(v_1, \dots, v_{n-1}, u^*)$.

According to M represented in figure 11.1, for example, q is dependent on a and (b, c), while r is dependent on a, (b, c), and d. Now if a prospect p is not realized in the actual world $w^{@}$, all it takes for p to be realized is that one background condition on which p is dependent is realized. Of course p is realized as well in worlds where more than one background condition on which p is dependent is realized, but in those cases the ensuing worlds are not as similar as possible to the actual world.

Definition 5. A world w^* is maximally similar to the actual world $w^@$ iff for w^* out of the set of all worlds: $\max(\#(w^* \cap w^@))$.

"#" here signifies the cardinality of the intersection of the respective world with the actual world. By maximizing the cardinality of this set, those worlds are chosen which have the highest overlap with the actual world.

Restrictions 1 and 2 (or 3 and 4, respectively) are satisfied if f selects worlds w^p and w^q, which are compatible with p and q with respect to M, respectively, such that both w^p and w^q are most similar to $w^@$ by the above similarity measure. With the concepts discussed in this section, we can therefore specify definition 1:

Definition 1*. $p \Leftrightarrow q \Leftrightarrow w_i^p \geq w_i^q$ for all $\langle w_i^p, w_i^q \rangle$ that are compatible with $p \wedge \neg q$ and $q \wedge \neg p$ with respect to M, respectively, such that both w_i^p, w_i^q are most similar to $w^@$.

Definition 1* yields a preference relation $\succeq$ over propositions with the following properties.

Theorem 1. If the causal model is non-cyclical, $\succeq$ is reflexive.

Proof: For each world w_i compatible with p, there is a realization of the background variables u_i such that the proposition equivalent to $V(u_i) \cup u_i$ is contains in w_i. u can be distinguished into the independent and the dependent background conditions, u^*. If there is only one set u^* for p, the proof is trivial, because there is only one world that is compatible with p. If there is more than one u^*, then the similarity relation ensures that only identical u_i^*'s are paired. Hence, for all $\langle w_i, w_j \rangle \in f(\langle p, q \rangle)$: $w_i = w_j$. Given that $\geq$ is reflexive, the relation $\succeq$ defined thus is equally reflexive.

Theorem 2. If for all prospects $p, q, r \ldots$, all causally possible conjunctions $p \wedge \neg q$, $p \wedge \neg r$ are dependent on the same background variable u^{p^*} (and similarly for $q \wedge \neg p$, $r \wedge \neg q, \ldots$), then a prospect preference ordering over $p, q, r, \ldots$ is transitive.

Proof: Without loss of generality, we take the case where $p \succeq q$ and $q \succeq r$. If all $p \wedge \neg q$, $p \wedge \neg r$ are causally possible, then there are causally compatible worlds such that $p \wedge \neg q \in w^{p \wedge \neg q}$ and $p \wedge \neg r \in w^{p \wedge \neg r}$. If for all p, q, r, $p \wedge \neg q$ and $p \wedge \neg r$ are dependent on the same variable u^{p^*}, then there is at least one world $w^p = w^{p \wedge \neg q} = w^{p \wedge \neg r}$ which is causally compatible with both $p \wedge \neg q$ and $p \wedge \neg r$. If $p \wedge \neg q$ and $p \wedge \neg r$ depend only on u^*, then $w^p = w^{p \wedge \neg q} = w^{p \wedge \neg r}$ is the world causally compatible with $p \wedge \neg q$ and $p \wedge \neg r$

which by definition 5 is most similar to $w^{@}$ (for the same reasons, mutatis mutandis, $w^{q} = w^{q\wedge\neg p} = w^{q\wedge\neg r}$ is the world causally compatible with $q\wedge\neg p$ and $q\wedge\neg r$ which is most similar to $w^{@}$). By definition 1*, and $p\succeq q$ and $q\succeq r$, $w^{p\wedge\neg q}\geq w^{q\wedge\neg p}$ and $w^{q\wedge\neg r}\geq w^{r\wedge\neg q}$. By the argument above, $w^{p\wedge\neg q} = w^{p\wedge\neg r}$ and $w^{q\wedge\neg p} = w^{q\wedge\neg r}$, hence $w^{p\wedge\neg r}\geq w^{q\wedge\neg r}$ and $w^{q\wedge\neg r}\geq w^{r\wedge\neg p}$, and thus by transitivity of $\geq$ $w^{p\wedge\neg r}\geq w^{r\wedge\neg p}$. Then by definition 1*, $p\succeq r$.[12]

It is further noteworthy that $\succeq$ is not complete, even if $\geq$ is. This can easily be seen by the following counterexample. Take a $\langle p,q\rangle$ such that $\langle w_1^p, w_1^q\rangle \in f(\langle p,q\rangle)$ and $\langle w_2^p, w_2^q\rangle \in f(\langle p,q\rangle)$, such that $w_1^p > w_1^q$ and $w_2^q > w_2^p$. Then $\succeq$ is not defined over $\langle p,q\rangle$.

These results are quite weak, but they represent genuine properties of pairwise preferences. The antecedent of theorem 2 is of course often not fulfilled, which explains the manifold existence of intransitive preference comparisons. That preferences are not complete over the set of all prospects should not be surprising at all.

The formal apparatus developed in this section can now be applied to the case of Diogenes, discussed in section 1. Diogenes lives in world where he is without a donor, and therefore poor and unenvied, but independent and credible in his ideology: $w^{@} = \{\neg s, \neg r, \neg e, i, c\}$. The causal model M that Diogenes believes in is represented in figure 11.2.

The actual world is thus causally compatible with the prospect of poverty ($\neg r$) with respect to M, and it is obviously maximally similar to itself. The world $w^{o} = \{s, r, e, \neg i, \neg c\}$ on the other hand, is compatible with the prospect of wealth (r) with respect to M. Because wealth is dependent on only one background variable in model M, there is no other world compatible with the prospect of wealth with respect to M. Thus even though

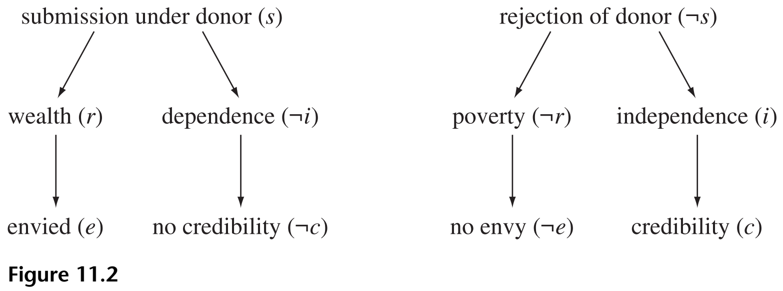

Figure 11.2
Diogenes' causal beliefs.

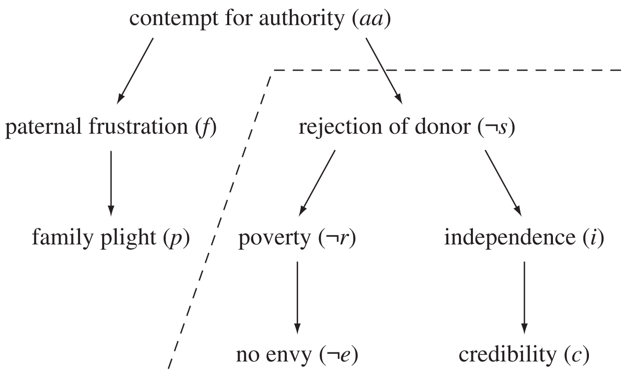

Figure 11.3
Truncating the causes of Diogenes' action.

$\#(w^o \cap w^@) = 0$, w^o is picked by f. By definition 1*, $\neg r \succeq r$ iff $w^@ \geq w^o$. Diogenes' behavior before Alexander, as reported by Diogenes Laertius, does indicate his preference for $w^@$ over w^o; and hence—through his causal beliefs—his preference for poverty over wealth.

But what if the causal model gets extended to include causes of Diogenes' choice between accepting and rejecting the donor? The background intuitions of such an extended model were discussed in section 1.3. In figure 11.3, the corresponding causal model M' is represented.

If definition 1* operated with M' instead of M, the conclusions of the above example would no longer be valid. Diogenes would prefer poverty over wealth if and only if he preferred the world $w^{o\prime} = \{\neg aa, \neg f, \neg p, s, r, e, \neg i, \neg c\}$ over the world $w^{aa} = \{aa, f, p, \neg s, \neg r, \neg e, i, c\}$; which is a completely different condition from preferring w^o to the actual world.

However, restriction 5 tells us to neglect all causal antecedents of a prospect if that prospect is an action. When evaluating non-action prospects, we assumed the truth of a prospect counterfactually and investigated how the causal dependencies and effects of that counterfactual assumption would determine the worlds compatible with that prospect. When evaluating an action, we assume the truth of that action-prospect not counterfactually, but as an intervention. An intervention, in contrast to a counterfactual assumption, does not have a retrospective influence on the past.[13] An intervention is represented as a truncation of the causal graph—

all direct ancestors of the model are removed from a causal model M, the model thus transformed into a truncated model M^T.

Definition 6. A causal model M is transformed into a truncated causal model $M^T = \langle U, V, G \rangle$ by eliminating all $g_i \in G$ which have an action prospect in their range.

The thick dotted line in figure 11.3 shows such a truncation. The function that connects *aa* with $\neg s$ is eliminated, thus cutting the causal connection between *aa* and $\neg s$ in M^T. By replacing M in definition 1 with M^T, restriction 5 is always satisfied.

Definition 1**. $p \succeq q \Leftrightarrow w_i^p \geq w_i^q$ for all $\langle w_i^p, w_i^q \rangle$ that are compatible with $p \wedge \neg q$ and $q \wedge \neg p$ with respect to M^T, respectively, such that both w_i^p, w_i^q are most similar to $w^@$.

In cases where M does not include any action prospects, definition 1** is obviously identical with definition 1*. In all other cases, definition 1** still satisfies theorems 1 and 2, as they were proven for all causal models, including truncated ones.

Thus, despite Diogenes' belief in the extended causal model M', definition 1** secures that his preference for poverty over wealth is still derived on the basis of the truncated model M^T, which in this case coincides with the original model M.

3 Conclusion and Remarks

I have offered a principle of abstraction between an agent's preferences over prospects and her preferences over worlds. More specifically, I represented the agent's beliefs as a causal model, and argued with the help of this model which of the agent's preferences over worlds serve as definiens for her preferences over propositions.

I have argued why such a principle of abstraction is necessary for the explanation and prediction of behavior with preferences; however, the model presented here leaves open many important questions. I will finish with two remarks on how to develop the discussion further.

3.1 Possibility or Probability

The criterion of the causal possibility of a world might be too rough a distinction to be viable. Instead, it has been suggested that prospects should be evaluated according to a weighted average of value of those worlds in which they are realized. The weighing can be determined as a probability index, which measures the likelihood of a world occurring given the actual world. Ideally, such a measure combines the criteria of causal possibility and actuality.

A first step was made by Rescher (1967). He constructed a ranking over worlds by assigning to them a numerical index of merit. From this ranking he derived an index over states: The index number of a state $\#(a)$ is the arithmetic mean over the index numbers of all possible worlds in which a is true. These index numbers over states give rise to a semantic definition of preferences over states: a is preferred to b iff $\#(a) > \#(b)$.

Trapp (1985) picked up Rescher's idea; but unlike him, Trapp suggested a probabilistic weighing of the index of possible worlds. Such a weighing can be interpreted as a continuous similarity metric: An agent assigns higher probability to those worlds that he thinks are closer to actuality. Jeffrey gave a similar account. He derived the desirability index over propositions from the desirability index over worlds: "the desirabilities of a proposition is a weighted average of the desirabilities of the cases [worlds] in which it is true, where the weights are proportional to the probabilities of the cases" (Jeffrey 1983, p. 78).

The most pressing problem of these accounts is their uniform treatment of actions and non-action prospects, as discussed in section 1.3. More generally, the probabilistic weighing of the worlds does not necessarily coincide with the concept of causal compatibility presented here. Causal decision theory has tried to remedy this problem by recasting the probability measure as a specification of objective chances or a measure of counterfactual dependency. Instead of trying to import all relevant information into the probability measure, the natural expansion of the account presented here suggests employing a subjective probability measure *conditional* on other relevant causal factors held fixed. The notion of relevant causal factors, of course, needs to be provided independently and prior to the probability measure; a task fulfilled by the causal graph discussed in this paper. The structure needed for a probabilistic weighing of worlds to determine the preferences (expressed as a utility index) over prospects then requires a *Bayesian network* which consists of a causal model and a probability function defined over it, satisfying certain conditional dependencies. To construct a utility function on the basis of Bayesian networks will be the task of a future paper.

3.2 Prospect Preference Aggregation

The model presented here provides a definition of more abstract prospect preferences in terms of world preferences. Once the prospect preferences are specified for a particular agent, the question arises: How are they employed in the explanation of the agent's behavior in novel situations? In the simplest case, the new situation is analyzed into its aspects, and the prospect preferences of the agent may provide us with clues of what the agent will do or why she did what she did. For example, a guest's preference for coffee over brandy at dinnertime is applicable to all new dinner situations. Such an application is easy in cases where the available prospect preferences are unanimous—that is, where all aspects of one situation are either preferred or noncomparable to the aspects of another situation. But what can we say if conflicts arise? For example, our guest may have a preference for nonalcoholic beverages, but also a preference for caffeine-free ones. How can one explain her choice between brandy and coffee with these preferences?

One suggestion for such a case is to employ the framework of ordinal preference aggregation as found in the social choice literature. But instead of using it for questions of how the rational preferences of a group of individuals can be aggregated into a coherent ranking of the group, this strategy proposes "to apply interpersonal economic theory to intrapersonal problems" (Elster 1985, p. 232)—that is, the different prospect preferences are aggregated back into one world preference.

The general results of such an application are that there is no aggregation rule for prospect preferences that satisfies certain minimal constrains and results in a coherent, transitive world-preference order (cf. Steedman and Krause 1986). It does not preclude that in many situations, coherent world-preferences can be aggregated from prospect preferences, or that with the help of external information, the decisiveness of some prospect preferences can be justified.

Again, further discussions of these questions require another paper. With the current state of research in this area, however, there is hope that prospect preferences can, in many cases, be aggregated to coherent world preferences and can thus function in the prediction and explanation of action.

Acknowledgments

I thank Richard Bradley, Joseph Keim Campbell, Nancy Cartwright, Isaac Levi, Daniel M. Hausman, Rainer W. Trapp, Joanne Grüne-Yanoff, and two anonymous referees for helpful comments on earlier drafts of this essay.

Notes

1. Pettit's claim is that property preferences determine world-preferences. The ultimate determining preference is often called value (as does Pettit himself). Disagreement prevails between those who defend value atomism—that value has its origin in a few very abstract aspects of the world (cf. Harman 1967; Quinn 1974)—and those who defend value holism—that value has its origin in the most specific states of the world (cf. von Wright 1963, pp. 29–34; von Wright 1972; Rescher 1967; Trapp 1985; Hansson 2001). Pettit claims that it is a folk psychological platitude that "choosing on the basis of the properties displayed by the alternatives" captures "choosing for a reason." As there is considerable disagreement among philosophers about this claim, I am cautious granting it folk status.

2. This in effect assumes that all choices are, in Savage's terminology, constant acts (cf. Savage 1972, p. 25). I make no attempt to justify this problematic assumption. For the sake of simplicity, I have excluded all considerations of uncertainty. To elaborate in a probabilistic framework what is discussed here under certainty will be the task of another paper.

3. This account must not be identified with the revealed preference theory known from economics. Revealed preference theory *defines* preferences as consistent choices over options under a given budget. The present account discusses preferences as mental states, which are *indicated by* behavioral evidence. For a detailed discussion of the revealed preference approach, see Grüne 2004.

4. It now becomes clear why the paper is restricted to *certain* prospects. This definition does not work if p or q are gambles over worlds. Take the following example: p and q are gambles such that

p: if dice rolls 6 you receive \$100.

q: if dice rolls 4 or 5 you receive \$100.

According to the definition, one prefers q to p only if one *both* prefers worlds w^{q*}, in which the dice rolls 4 and you receive \$100 to w^p, as well as world w^{p**}, in which the dice rolls 5 and you receive \$100 to w^p. Given a fair dice, however, it is plausible to be indifferent between these worlds, even though it is very plausible to prefer q to p.

5. This is the case that comes closest to Pettit's discussion of property desires.

6. I will argue that the third case, preferences over mutually nonexclusive prospects, must be translated into preferences over mutually exclusive ones. A translation procedure will be discussed in section 2.2.

7. This approach was to my knowledge first discussed by von Wright (1972, p. 146). It is also defended by Hansson (2001, pp. 67–94).

8. Diogenes Laertius 1931, p. 38. I use the source as an inspiration, and hasten to add that the following is obviously not intended as a textual analysis.

9. Trapp claimed that "no two relata of a preference relation should be considered to be true in the same possible world," at least in those worlds that are chosen by the selection function (Trapp 1985, p. 301). For a rejection of this view, see Hansson 1989, p. 6.

10. Or, if one of the relata is not causally compatible with any world, f picks out worlds that are compatible with the untranslated relatum.

11. This situation is in many ways similar to the so-called Newcomb's Problems in probabilistic models of decision making.

12. The reverse claim does not hold: one cannot infer from the transitivity of a preference relation over $p, q, r, \ldots$ that all their causally possible conjunctions $p \wedge \neg q$, $p \wedge \neg r$ are dependent on the same background variable u^{p*}. For example, the evaluation of $w^{p \wedge \neg q}$ and $w^{p \wedge \neg r}$ might coincide without the two worlds being identical.

13. For a more extensive discussion of intervention, see Pearl 2000, pp. 85–89; and Spohn 2002, pp. 23–27.

References

Diogenes Laertius. 1931. *Lives, Teachings, and Sayings of Famous Philosophers*. Loeb Classical Library. Cambridge, Mass.: Harvard University Press.

Elster, J. 1985. "Weakness of Will and the Free-Rider Problem." *Economics and Philosophy* 1: 231–265.

Grice, H. 1989. "The William James Lectures." In *Studies in the Way of Words*. Cambridge, Mass. and London: Harvard University Press.

Grüne, T. 2004. "The Problems of Testing Preference Axioms with Revealed Preference Theory." *Analyse & Kritik* 26.2: 382–397.

Halldén, S. 1957. *On the Logic of Better*. Lund: Library of Theoria.

Hansson, S. 2001. *The Structure of Values and Norms*. Cambridge: Cambridge University Press.

Harman, G. 1967. "Towards a Theory of Intrinsic Value." *Journal of Philosophy* 64: 792–804.

Jeffrey, R. 1983. *The Logic of Decision*. Chicago: University of Chicago Press.

Joyce, J. 1999. *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Pettit, P. 2002. "Decision Theory and Folk Psychology." In *Rules, Reasons, and Norms*. Oxford and New York: Oxford University Press.

Quinn, W. 1974. "Theories of Intrinsic Value." *American Philosophical Quarterly* 11: 123–132.

Rescher, N. 1967. "Semantic Foundations for the Logic of Preference." In *The Logic of Decision and Action*, edited by N. Rescher. Pittsburgh: University of Pittsburgh Press.

Savage, L. 1972. *The Foundations of Statistics*. New York: Dover.

Spohn, W. 2002. "Dependency Equilibria and the Causal Structure of Decision and Game Situations." Unpublished ms., University of Konstanz.

Steedman, I., and U. Krause. 1986. "Goethe's Faust, Arrow's Possibility Theorem, and the Individual Decision Maker." In *The Multiple Self: Studies in Rationality and Social Change*, edited by J. Elster. Cambridge: Cambridge University Press.

Trapp, R. 1985. "Utility Theory and Preference Logic." *Erkenntnis* 22: 301–339.

von Wright, G. 1963. *The Logic of Preference*. Edinburgh: Edinburgh University Press.

von Wright, G. 1972. "The Logic of Preference Reconsidered." *Theory and Decision* 3: 140–169.