Preferences vs. Desires:

Debating the Fundamental Structure of Conative States

Armin W. Schulz

Department of Philosophy

University of Kansas

3083 Wescoe Hall

Lawrence, KS 66045

awschulz@ku.edu

785-864-3976

Abstract

In this paper, I address a major, but quite overlooked, question about the structure of the cognitive / conative model of the mind that (arguably) underlies much of the work in economics, psychology, and philosophy: namely, whether conative states are fundamentally monistic (desire-like) or comparative (preference-like). I begin by making clear that deciding this issue is important, both for its inherent interest and for its implications for other questions. I then argue that two seemingly promising sets of theoretical considerations are unable to resolve this debate: in particular, neither the structure of Rational Choice Theory, nor considerations of computational efficiency (as has been suggested by John Pollock) can be used to answer the question about the fundamentality of desires or preferences. Given this, I suggest that a consideration that speaks in favor of the preference-based view is the fact that it makes it easier to explain certain empirically observed patterns in decision making: namely, violations of choice transitivity.

Preferences vs. Desires:

Debating the Fundamental Structure of Conative States

I. Introduction

Much work in economics, psychology, cognitive science, and philosophy is based on the assumption that at least some of our (higher-level) decisions are determined by the interplay of two different kinds of representational mental states: cognitive ones and conative ones. Specifically, it is commonly assumed that (a) some of our mental states are responsible for informing us about what the world is like (cognitive ones), and others for what the world ought to be like (conative ones), and (b) we put these two types of mental states together when making at least some decisions about what to do (Davidson 1963; Goldman 1970, 2006; Fodor 1981; Nichols & Stich 2003; Sterelny 2003; Pollock 2006; Hausman 2012).

Given the importance of this model of decision making, it is surprising that there is a major issue surrounding it that has not been discussed in any kind of detail. This issue concerns the question of what the logical structure of our conative states is: are they fundamentally monistic or comparative – that is, other than by our beliefs, are our actions ultimately determined by what we want or by what we prefer? In this paper, I try to come closer to answering this question.

To do this, I start, in section II, by making the problem to be discussed more precise and bringing out its importance. In sections III and IV, I then show why two seemingly promising sets of considerations cannot be used to solve this problem: in particular, I make clear that neither the structure of Rational Choice Theory nor John Pollock's recent efficiency-based account can be seen to favor either the fundamentality of desires or preferences. In section V, I present a novel attempt towards addressing this issue. I conclude in section VI.

II. Desires vs. Preferences: The Question in More Detail

Conative states, as they are commonly understood, are mental states that tell an organism what it is that it ought to try to achieve: they represent its goals (what it would be good for the world to be like, given its needs) or what 'the thing to do' is in the situation it is in (or something of this sort – Schroeder 2009). Now, there are many interesting questions that could be asked about these kinds of states: for example, one might wonder about their neurological foundations (Glimcher et al. 2005; Schroeder 2004; Morillo 1990), the exact way in which they figure in an organism's action-generating mechanisms (Bratman 1987; Sterelny 2003), or how they are related to other mental states (Lewis 1988; Morillo 1990).

However, there is also another question about these states that could be asked: what is their logical structure? Specifically, we might want to know whether conative states are fundamentally monistic (one-place) or comparative (multi-place): at bottom, do we seek to achieve certain goals *simpliciter*, or do we always seek to achieve certain goals rather than other ones? Interestingly however, while many of the other issues concerning the nature of conative states have received their fair share of attention in the literature, this question has been almost completely overlooked: there is hardly any work that even raises it – not to mention trying to answer it (though see Pollock 2006; Schroeder 2010; Allais 1953; Ross 2005; Bermudez 2009). However, answering this question is important. This is so for at least three reasons.

First, it has possible consequences for what our minds (and those of other animals) are really like. For example, if our fundamental conative states are relational, then, despite common assumptions to the contrary, we cannot have conative attitudes towards single propositions (Nichols & Stich 2003; Carruthers 2006). Second, it has possible implications for our understanding of the relationships between cognitive and conative states. For example, if it turns

out that cognitive states are fundamentally monistic and conative states fundamentally comparative, then cognitive and conative states differ in logical structure, and not just in content (Lewis 1988, 1996; Smith 1987; Bradley & List 2009; see also Stampe 1986). Third, it has possible implications for what we should think about various other (not purely cognitive) mental states. For example, if it turns out that there are close connections between conative states on the one hand, and pleasures and rewards on the other (as has been suggested e.g. by Morillo 1990; Damasio 1994; Schroeder 2004), then understanding the logical structure of the former should also tell us more about how the latter function in an agent's cognitive architecture (Morillo 1990; Damasio 1994; Schroeder 2004).

To make the discussion of the question about the structure of our fundamental conative states easier, I shall phrase it as a dispute between those who think conative states are fundamentally 'desire-like' and those who think they are 'preference-like'. The reason for this terminological choice is that desires are commonly described as monistic states: they are thought to have some *one* intentional object *p* (see e.g. Fodor 1981; Nichols & Stich 2003). By contrast, preferences are commonly thought to be comparative: they are thought to have *two* ordered intentional objects $\langle p, q \rangle$ (see e.g. Luce & Raiffa 1958; Hausman 2012). Hence, instead of asking whether conative states are fundamentally comparative or monistic, we can also ask whether they are fundamentally *desires* or *preferences*.

Before addressing this question, it is important to point out that conative states (whether as desires or preferences) are typically thought to be graded – i.e. to come in degrees (Jeffrey 1983; Hausman 2012). Note also that it is often thought that the measurement of these gradations is somewhat arbitrary: while we might want (prefer) something more or less (to something else),

how much more or less we want (prefer) it (to something else) might not permit of a fully precise answer (see e.g. Hausman 1995, 2012; Jeffrey 1983, 1992).

As will become clearer below, there is more to be said about the plausibility of this 'gradedness' assumption – for now, though, it can simply be taken for granted. What is worthwhile pointing out at this point is just that making this assumption does not cloud the distinction between monistic and comparative conative states. Whether or not conative states are (somewhat arbitrarily) graded, there is a difference between their being monistic and their being comparative: for example, an explanation of why an agent is more inclined to bring about A than B would cite different numbers of conative states in the two cases (at least two vs. at least one).

With all of this in mind, consider now three attempts at determining the fundamentality of desires or preferences. The first of these is based on the structure of Rational Choice Theory, the second on considerations of cognitive efficiency, and the third on empirically observed patterns in people's choice behavior.

III. Rational Choice Theory and the Structure of Conative States

One promising strategy to find out about the fundamentality of desires or preferences would seem to lie in considering the structure of those well corroborated scientific theories that make most appeal to these sorts of states. After all, since it is mostly these theories that are responsible for our scientifically informed thinking about the existence and features of conative states, their implications about the fundamental structure of our conative states should be taken seriously.

Furthermore, it seems clear that, among the scientific theories that make most appeal to these kinds of states, Rational Choice Theory (RCT) holds a pivotal place. RCT is one of the most widely used theories codifying the cognitive / conative model of the mind, and a major part of a

number of scientific disciplines from economics to behavioral ecology (Hausman 2012; Rosenberg, 2012; Ross 2005). Finally, many other decision theories are quite obviously unable to help decide the question about the fundamentality of desires or preferences.[1] Hence, RCT will be the main focus in what follows.

However, perhaps despite initial appearances to the contrary, the structure of RCT cannot settle the dispute over the fundamentality of desires and preferences. To see this, it is best to start with a brief overview of the key elements of the canonical formulations of that theory (Savage 1954; Luce & Raiffa 1958; Jeffrey 1983; Hausman 2012).

Assume some agent has to make a decision among a set *A* of *m* different courses of action (A_1 to A_m), that she is aware of all the consequences of these actions, and that she divides the world into a set *S* of *n* exhaustive and mutually exclusive states (S_1 to S_n). Also, assume that she can rank all the available actions in terms of the degree to which they are 'choiceworthy' (i.e. the extent to which she evaluates them as something worth doing), and that this ranking satisfies a number of conditions – e.g. completeness, transitivity, independence, etc.. Given these assumptions, RCT shows that this ranking can be represented by two functions – a probability function P(x) ranging over *S* and a real-valued and only minimally cardinal (set of) utility function(s) *u*(x) ranging over (*A* x *S*) – that combine in such a way that the agent ends up maximizing her expected utility by choosing the option that is highest in her 'choiceworthiness' ranking. In particular, any action A_k out of the set *A* of feasible actions is assigned an expected utility value by setting:[2]

[1] For example, simple heuristics-based theories (Gigerenzer & Selten 2001; Gigerenzer et al. 1999; Simon 1957) are quite clearly not structured in a way that would allow us to answer the question about the fundamentality of desires or preferences. This also goes for theories that hold that many actions are best seen as direct – i.e. non-representationally mediated – responses to the presence of certain environmental factors (Reed 1996; van Gelder 1996; Clark 1997, chap. 9).

[2] This glosses over the debate between causal and evidential forms of RCT, for which these formalisms would be slightly different (see e.g. Eells 1982). Note also that the argument to follow lends itself more to versions of RCT

(1) $EU(A_k) = \sum_{i=1}^{n} P(S_i)\, u(A_k \,\&\, S_i)$.

Further, it then holds that

(2) For any two actions A_k and A_j in *A*, $EU(A_k) > EU(A_j)$ if and only if A_k is ranked higher in the agent's choiceworthiness ranking than A_j.

Interpreted psychologically – the only interpretation that is even remotely useful in the present context –[3] the idea is that the agent combines her conative states (represented by *u*(x)) and her cognitive states (represented by P(x)) in such a way that they form a well-ordered choiceworthiness ranking over the available actions. It is this ranking that is then used as the basis of the choice she actually makes.

At this point, it is important to be clear about a confusion that might arise. In many discussions surrounding RCT, the choiceworthiness rankings that are the output of the decision making mechanism are referred to as 'preference rankings' (Luce & Raiffa 1958; Jeffrey 1983; Joyce 1999; Hausman 2012). However, it is important to realize that this is not the sense of 'preference' that matters in the present context. To see this, note that choiceworthiness rankings are made up from the agent's more fundamental cognitive and conative states – they express how the agent has evaluated the options open to her, taking into account both what she *fundamentally*

---

that allow utilities to be state-dependent (as, for example, that of Jeffrey 1983) than to those that require them to be state-independent (as, for example, that of Savage 1954). However, it would be easy to reformulate the argument in terms of theories that require state-independence.

[3]Other interpretations of RCT are normative (they see the theory as setting out which decisions an agent ought to make) and behavioral (they see it as merely representing the outcome of the agent's decisions – she acts as if she were maximizing her expected utilities). These are not useful here, though, as they have no direct implications for what our minds are actually like (which is what we are trying to determine).

wants / prefers the world to be like *and* what she thinks the world is likely to, in fact, be like (Hausman 2012). In other words: choiceworthiness rankings are combinations of the agent's more fundamental cognitive and conative states – and it is the structure of the latter that we are interested in here.

Now, for different reasons, it might appear that RCT favors *both* the fundamentality of desires *and* preferences. On the one hand, one might think that RCT can be used to argue in favor of a preference-based view of our fundamental conative states. This is due to the fact that, typically, in RCT only differences in utility values are meaningful, while their absolute values are irrelevant – as noted above, $u(x)$ is typically assumed to only have a fairly low degree of cardinality (see e.g. Luce & Raiffa 1958). This may be thought to suggest that RCT speaks in favor of the preference-based view, since the theory really only seems to be committed to a relation among utility values – not to the values themselves. However, this reasoning would be fallacious, as it would fall prey to precisely the point made at the end of section II: the fact that there is a certain arbitrariness in how conative states are measured in RCT does not imply that the latter is committed to either the fundamentality of desires or preferences.

On the other hand, one might think that, since, in standard RCT, utilities are monistic ($u(x)$ evaluates only one action at a time), they are most naturally interpreted as representing the agent's desires – and thus, that RCT can be used to argue in favor of the desire-based view of our fundamental conative states. However, this, too, would be mistaken. To see this, note that it is possible to reinterpret the basic framework of RCT so that it does not appeal to (monistic) utilities at all.

In particular, one can just define a new relational theoretical notion – that of a 'primary preference'– and treat *it* (instead of utilities) as representing our basic conative states in RCT.

Primary preferences could then be formalized with a real-valued and only minimally cardinal (set of) function(s) PP(x, y | z), where x and y range over *A*, and z ranges over *S*. (Note that PP(x, Y | z) is relational in that, unlike *u*(x), it evaluates two actions A_i and A_j at a time.) In order to recover all of the structure of RCT, we then just need to require that:

(Primary Preference) $PP(A_i, A_j \mid S_k) = u(A_i \,\&\, S_k) - u(A_j \,\&\, S_k)$

(with *u*(x) as defined above). To see this, note that taking the expectation of (Primary Preference) (with P(x) defined as above) yields:

(3) $EPP(A_k, A_j) = \sum_{i=1}^{n} P(S_i)\, PP(A_k, A_j \mid S_i)$.

Given (Primary Preference) and (3), we can replace (2) with

(4) For any two actions A_k and A_j in *A*, $EPP(A_k, A_j) > 0$ if and only if A_k is ranked higher in the agent's choiceworthiness ranking than A_j.[4]

Interpreted psychologically, the idea here is (as before) that the agent combines her conative states and her cognitive states in such a way that they form a well-ordered choiceworthiness ranking over the available actions. However, this time, the agent's conative states are represented by a (set of) relational 'degree of primary preference' function(s) (the agent's beliefs are still

[4] It is easy to see that (4) is equivalent to (2): $EPP(A_k, A_j) > 0$ if and only if $\sum_{i=1}^{n} P(S_i)\, [u(A_k \,\&\, S_i) - u(A_j \,\&\, S_i)] > 0$ – i.e. if and only if $EU(A_k) > EU(A_j)$. Hence all the results that can be derived on the basis of (2) can also – with the appropriate rewriting – be derived from (4).

represented by the monistic probability function). In this way, it becomes clear that, as far as the essence of the formal framework of RCT is concerned, an agent can be seen to choose A_k over A_j (for all *j*) *either* because doing A_k maximizes her expected utilities *or* because her expected primary preference for A_k over A_j is positive for all *j*. In turn, what this shows is that, as far as the essence of the formal framework of RCT is concerned, fundamental conative states can be seen to be either desire-like (as in (1) and (2)) or preference-like (as in (3) and (4)). Hence, the key question to be addressed here remains open.

IV. Pollock's Account: Desires, Preferences, and Computational Complexity

In a recent – widely discussed and well received (see e.g. Schroeder 2010) – book, Pollock claims that it is possible to argue in favor of a desire-like structure of fundamental conative states by pointing to the kind of mind designs that it is physically feasible to implement (Pollock 2006, 23-35; see also Schroeder 2010).[5] However, while raising some important points, his argument ultimately fails to be successful.

Pollock's argument for the fundamentality of desires proceeds as follows (2006, chap. 2). Assuming that desires are graded, but preferences are not, he notes that it takes many more binary preference relations to encode a body of conative information than it takes a set of graded desires to encode this body of conative information (Pollock 2006, 24, 30-35). In fact, it takes so

[5] Technically, Pollock (2006, 24-25) frames his account in terms of the question of whether we should think that our actions are driven by our desires or our choiceworthiness rankings (which he refers to as 'preferences' – see above). While he does not specify explicitly why he takes this to be the relevant comparison, it seems that the main reason for it lies in his interpretation of the classic representation theorems of RCT – i.e. (2) above (Pollock 2006, 24-25). Looking at these theorems, he wonders in what direction they are to be read: are they meant to show that only choiceworthiness rankings are psychologically real (and can merely be *expressed* in terms of more fundamental cognitive and conative states), or do they show that only fundamental cognitive and conative states are psychologically real (and can merely be *expressed* in terms of choiceworthiness rankings)? However, as noted in the previous section, this is not the question asked here (which concerns the structure of our fundamental conative states); fortunately, as made clear in the text, it is easy to transpose his arguments to the present debate. For this reason, it might be more accurate to describe the discussion to follow as concerning a 'Pollockian' position, rather than his actual view.

many more preferences to encode this information that it would break the bounds of what is physically realizable (Pollock 2006, 28). Hence, he concludes that we should expect conative states to be desire-like, not preference-like (Pollock 2006, 35).

To lay out this argument out in more detail, consider the following example (following Pollock, I here consider a situation of perfect certainty, but the extension to cases of uncertainty is straightforward). Assume an agent can bring about four different 'simple outcomes' (O_1, O_2, O_3, and O_4) – where a 'simple outcome' is the most basic result of the agent's actions that she considers to be relevant to the case at hand. Assume further that these simple outcomes can (at least sometimes) be brought about together – i.e. that the agent can also bring about more complex outcomes like (O_1 & O_2) and (O_3 & O_4).[6] Next, note that a functional desire-driven organism can be assumed to associate a 'desirableness' (or 'utility') value with every simple outcome: this will give her all she needs in order to decide which of these simple outcomes to bring about – she can simply maximize her explicitly assigned utilities over these outcomes. Importantly, though, it will also give her all she needs to determine which complex outcomes to bring about – for the utility values of these complex outcomes can be constructed out of the utility values of their components. The following table of utility assignments makes this clearer (see table 1):

[Table 1]

[6] Pollock here thus assumes a sort of 'value atomism'. However, it is important to note this assumption is actually quite weak. In particular, Pollock does not need to assume that all the outcomes the agent considers as relevant can be constructed out of a small set of simple outcomes, or that all combinations of simple outcomes are genuine outcomes the agent can deliberate over. His basic point is just that, to the extent that the agent has to make *some* decisions among outcomes that are constructed out of simpler outcomes, a fundamentally desire-based organism will be much more efficient than a fundamentally preference-based one. Note also that the value atomism here is purely subjective: the claim is just that *this particular agent* takes some possible outcomes and their conative values to be basic. This is consistent with these outcomes being non-basic for other agents. At any rate, questioning Pollock's assumption of a value atomism – however minimal – would only help my case.

Using the utility assignments in table 1, it becomes possible to determine the utility of a complex outcome – like (O_1 & O_4) – by simply summing the utilities of its more basic components (formally, this is equivalent to representing these utilities with a real valued function $\underline{u}(x)$ that ranges over all the possible outcomes – i.e. the set of the simple outcomes plus all the complex outcomes that can be constructed out of them – and that is constrained so that, for any complex outcome O made up from more simple outcomes A and B, $\underline{u}(O)=\underline{u}(A)+\underline{u}(B)$).[7] For example, it is now possible to set $\underline{u}(O_2 \& O_3) = \underline{u}(O_2) + \underline{u}(O_3) = 17$ and $\underline{u}(O_1 \& O_4) = \underline{u}(O_1) + \underline{u}(O_4) = 11$. Hence, with just four utility assignments, the agent is able to determine which complex outcomes to bring about as well.

However (according to Pollock) the situation is very different for an agent that operates only with non-graded preferences. To see this, return to table 1, and note that it entails that the agent prefers O_1 to O_2, O_2 to O_3, and O_3 to O_4 (as $\underline{u}(O_1) > \underline{u}(O_2) > \underline{u}(O_3) > \underline{u}(O_4)$). However, these non-graded preference relations over the simple outcomes alone will never tell the agent which complex outcome to bring about. In particular, by themselves, these preferences leave it completely open whether the agent prefers (O_1 & O_4) to (O_2 & O_3) or the other way around (and similarly for other complex outcomes). Therefore, the agent needs to be assumed to have a further set of non-graded preferences to be able to decide which of these complex outcomes to bring about – and so on for all other complex outcomes (Pollock 2006, 32-35).

In this way, Pollock concludes, (graded) desires can be seen to be more efficient in encoding conative information than (non-graded) preferences (Pollock 2006, 28-29). In fact, he goes

[7] This assumes that $\underline{u}(x)$ is at least cardinal in first differences. Note also that (depending on the cardinality of $\underline{u}(x)$), many mappings of the utility values of the simple outcomes into utility values of the complex outcomes will do. The one thing that does need to be excluded is that bringing about two simple outcomes together always generates *non-systematic* (i.e. varying) interaction effects – in other words, some kind of separability assumption needs to be made here. However, it seems plausible that, in many cases, such non-systematic interactions are indeed absent; hence, I shall grant Pollock this assumption (noting again that dropping it would only help my argument).

further than this: he notes that the difference in cognitive efficiency between these two types of conative states is of an extraordinarily high order of magnitude. For example, for agents that consider merely 60 different simple outcomes – a fairly small number – the necessary number of non-graded preferences would go beyond what any physical system (such as a biological brain) could handle: there are about 2^{60} ways of generating (consistent) complex outcomes out of 60 simple ones, which is far beyond the amount of information a brain can store in this (explicit) way (Pollock 2006, 27-28). For this reason, Pollock concludes that we should think that our actions are fundamentally driven by desires, not (non-graded) preferences: there is simply not enough 'room' in our brain to save all the required information if just (non-graded) preferences are used (Pollock 2006, 25-28).

As it turns out, though, Pollock's argument only works if we grant the assumption that desires, but not preferences, are graded. However, as noted above, there is no need to do so – it is entirely possible that our fundamental conative states are both relational and graded. If this is so, though, then Pollock's argument no longer goes through. To see this, return to the utility assignments in table 1, and re-express them in terms of the 'degrees of primary preferences' defined in the previous section, but this time replacing $u(x)$ with $\underline{u}(x)$ (and, as we are now in framework of certainty, suppressing the reference to *S*). This yields (see table 2):

[Table 2]

A look at this table makes clear that the primary preference for $(O_1 \,\&\, O_4)$ over $(O_2 \,\&\, O_3)$ can be determined by summing the primary preference for O_1 over O_2 (which is 1) and the primary

preference for O_4 over O_3 (which is -7), which yields a value of -6. Specifically, it can be put down as a theorem that, for all *i*, *j*, *k*, and *l*,

(5) $PP[(O_i \,\&\, O_j), (O_k \,\&\, O_l)] = PP(O_i, O_k) + PP(O_j, O_l) = PP(O_i, O_l) + PP(O_j, O_k)$.[8]

Realizing this is key, for it means that there is no reason to think that computational explosion will ensue if it is assumed that the fundamental conative states are preference-like: in fact, the number of basic preferences that would be needed for an organism to be a successful decision maker will be about the same as the number of basic desires needed (the exact relation will depend on how transitive the basic preferences are – see below for more on this). All that Pollock has shown is that fundamental conative states are likely to be graded – non-graded, purely binary conative states run into computational problems. This is an important insight that shores up the commonly made assumption that conative states are graded. However – and this is key for present purposes – it does nothing to settle the debate about the fundamentality of desires and preferences.

V. Decision Intransitivities, Preferences, and Desires

In order to take one step closer towards answering the question about the structure of our fundamental conative states, I want to suggest that it is useful to consider a well known and very

[8] This follows from the fact that $PP[(O_i \,\&\, O_j), (O_k \,\&\, O_l)] = \underline{u}(O_i \,\&\, O_j) - \underline{u}(O_k \,\&\, O_l) = \underline{u}(O_i) + \underline{u}(O_j) - \underline{u}(O_k) - \underline{u}(O_l) = \underline{u}(O_i) - \underline{u}(O_k) + \underline{u}(O_j) - \underline{u}(O_l) = PP(O_i, O_k) + PP(O_j, O_l)$ (and similarly for the last equality in the text). It is straightforward to generalize this to complex actions consisting of any number of simple actions. Note also that (5) compares complex outcomes with the same number of conjuncts. For comparing complex outcomes with different numbers of conjuncts, one needs to assume agents have a further primary preference of one outcome O_i over a 'no decision' outside option ND: $PP(O_i, ND)$. (For consistency with RCT, one would also need to define an appropriate value for $\underline{u}(ND)$). Given this, one can simply require that $PP[(O_i \,\&\, O_j), O_k] = PP[(O_i \,\&\, O_j), (O_k, ND)]$, and the above holds as before. Again, it is straightforward to generalize this for any number of simple outcomes, simply by adding enough ND alternatives to create equally structured complex outcomes.

recalcitrant empirical phenomenon: the fact that people often make intransitive choices. In particular, many people seem to choose as follows: in what appears to be a choice between two options A and B, they choose A, and in what appears to be a choice between B and C, they choose B, and, in what appears to be a choice between A and C, they choose C. Furthermore, it turns out that these sorts of 'decision cycles' are quite common: they have been found in many different contexts, and appear to be quite robust to changes in the particular decision problems with which the agents are presented (Broome 1991, 1999; Sopher & Gigliotti 1993; Guala 2005, pp. 98-105; Johnson & Busemeyer 2005; Tsai & Bockenholt 2006; Rieskamp et al. 2006; Waite 2001; Houston et al. 2007).

For present purposes, the key point about these cycles is that they pose an explanatory challenge: can we make sense of the way people choose options if those choices exhibit intransitivities? What makes this question so important here is that it can be used to drive a wedge between desire-based and preference-based views of fundamental conative states: in particular, as I try to make clear in what follows, it can be easier to answer this question using preferences as the fundamental conative states than using desires as the fundamental conative states.

To see this, note that organisms that make decisions using graded desires as their fundamental conative states are, at least to some extent, forced to at least *evaluate* the world transitively. If the organism desires A to degree d_1, B to degree d_2, and C to degree d_3, and if $d_1 > d_2$ and $d_2 > d_3$, then it must be that $d_1 > d_3$ just by the principles of basic arithmetic (and the assumption of a well-defined desirability ranking). In other words, if you want A more than you want B, and you want B more than C, then you *must* also want A more than you want C, just in virtue of what it means to have graded desires. Given this, in order to explain the above choice intransitivities,

desire-based theories have to appeal to something *beyond* the agent's conative evaluation of the options open to her. There are several options that could be and are being explored in this context (see e.g. Hagen et al. 2012; Johnson & Busemeyer 2005; Broome 1991, 1999; Bratman 1987; Davidson 1969).

For example, a desire-focused theorist can note that wanting A more than C need not translate into choosing A over C. Because of this, some cases of intransitive choices may be able to be explained as decision making failures or as the result of the agent having acquired new information. Equally, it might be that several seeming intransitivities are no such things if analyzed properly. In particular, it might be that people represent the choices they are facing in a way different from what the relevant researchers assume: instead of conceiving their choices as A vs. B, B vs. C, and A vs. C, the two As may in fact refer to different alternatives as far as the agent is concerned. Another possibility for the desire theorist is to appeal to changes in the agent's desirability rankings: the agent might start out desiring A to degree d_1, B to degree d_2, and C to degree d_3, with $d_1 > d_2$ and $d_2 > d_3$, but then come change her mind, and set $d_3 > d_1$. Various other explanations are conceivable as well.

There is much that could be said about these strategies for dealing with choice intransitivities. For present purposes, though, it is enough to make the following two remarks. On the one hand, together, these strategies really do seem to explain some aspects of the observed choice behavior. From classic work on the 'Allais Paradox' (Kahneman & Tversky 1979) and so-called 'Preference-Reversals' (Grether & Plott 1979; Johnson & Busemeyer 2005) to more recent work on this topic (Houston et al. 2007; Kalenscher et al. 2010), these different schemes do seem to sometimes capture what makes people choose intransitively. However, on the other hand, it also seems clear that, at least up to now, these schemes do not fully succeed in resolving every

question surrounding choice intransitivities (Johnson & Busemeyer 2005). In particular, given the intransigence of many of these intransitivities, it must be concluded that there is no reason yet to think that these schemes can be considered to give a complete account of why people choose the way they do (Cox & Epstein 1989; Loomes et al. 1991; Johnson & Busemeyer 2005).

It is here where assuming our minds are based on preference-like fundamental conative states gains plausibility. For if it is true that people fundamentally rely on preferences, rather than desires, to make decisions, it is possible for there to be 'brute intransitivities' in the way they evaluate the world. To see this, note that preferring A to B to degree d_1, and B to C to degree d_2, does not force one to prefer A to C to any degree whatsoever. So, in a choice between different foodstuffs, people might just fundamentally prefer A (bananas) to B (apples), B (apples) to C (oranges), and C (oranges) to A (bananas) – and all of these preferences might differ in their degree of strength. In turn, this is important, as it gives us a new way of making sense of people's intransitive actions: *apart* from the possibilities sketched above, it is now *also* possible to say that these actions stem from people's intransitive fundamental preferences. Put differently: a preference-based account has an *additional* free parameter (the transitivity of the agent's fundamental conative states) and can thus explain and predict decision intransitivities more easily than a desire-based theory can (and that without introducing major novel elements into the way the agent's decision procedure is conceived). Importantly, moreover, the current empirical evidence clearly favors adding this parameter: as matters stand, there is significantly more choice intransitivity than what would be predicted if a desire-based theory were true (Johnson & Busemeyer 2005).

Note that I am here not trying to argue for a specific way of incorporating intransitive fundamental preferences into a decision theory. In particular, for all that I have said, it may be

that the best way of modeling our fundamental preferences is via Prospect Theory (Kahneman & Tversky 1979), Regret Theory (see e.g. Loomes & Sugden 1982), Decision Field Theory (see e.g. Busemeyer & Townsend 1992), or some other theory (including, as made clearer below, even RCT). All that I am here trying to show is that appealing to preferences as the fundamental conative states gives us more degrees of freedom in explaining choices – and that these extra degrees of freedom are in fact useful for making sense of the data. Exactly how this appeal to fundamental preferences is to be formalized in a decision theory can be determined at a different occasion.

At this point, it is useful to consider two key objections that might be raised to this account. The first objection is based on the thought that the argument of this section seems incompatible with that of the previous two. The preferences underlying (Primary Preference) are strictly transitive: since they are representable by differences in utilities that assign every action only one value, they cannot fail to be transitive as a matter of basic arithmetic. However, in this section, I argued that our fundamental conative states are to be seen as preference-like and *in*transitive. How does this fit together? It seems that we have to give up either the claim that both RCT and Pollockian cognitive efficiency-based considerations are neutral when it comes to the dispute over the fundamentality of desires or preferences, or the claim that choice intransitivities speak in favor of the fundamentality of preferences.

Fortunately, it is possible to respond to this objection – the three sets of arguments are in fact consistent with each other. However, to bring this out, the account defended here needs to be a little bit further developed. To do this, it is best to begin by noting that (full) transitivity of our fundamental preferences is not required to block Pollock's argument. This is due to the fact that, even if our fundamental preferences are intransitive, this does not mean that we need to define a

separate degree of primary preference for *every* action outcome that we might want to consider, thus leading to computational explosion. Rather, we just need to follow the structure of many natural languages, and combine following a rule with explicitly noting exceptions to this rule. In particular, we can allow our fundamental preferences to be transitive *unless* a different degree of primary preference is encoded in the system. Note that the reasons for this separate encoding can differ widely – from a specific learning history to an on-the-fly evaluation of the options that is independent of the agent's other conative commitments. While an in-depth treatment of these reasons would be instructive, for present purposes, this can be left open.

To make this more explicit, take over all of the basic assumptions of my response to Pollock earlier. That is, continue to assume that our fundamental conative states are (i) relational and graded in such a way that they can be represented by a minimally cardinal real-valued (set of) primary preference function(s) $PP'(O_i, O_j)$ ranging over all possible outcomes (constrained so that for any two complex outcomes $O_a=(O_x\&O_y)$ and $O_b=(O_r\&O_w)$, $PP'(O_a, O_b)=PP'(O_x, O_r)+PP'(O_w, O_b)$; though see also note 10 below), (ii) internally consistent (so that $PP'(O_i, O_j) = -PP'(O_j, O_i)$), and (iii) complete (such that it is *in principle* possible to connect every simple outcome to every other simple outcome using either an explicitly encoded degree of primary preference or transitive summations of the latter). To allow for intransitivity, we then simply add to these assumptions the claim that (iv) some transitively implied primary preferences can be overridden, in a specific way, by explicit assignments of fundamental conative states. For example, we could say:

(6) $PP'(O_i, O_j) = PP'(O_i, O_k) + PP'(O_k, O_j)$ *unless* it is separately defined as x,

where O_i and O_j are two arbitrary simple outcomes and O_k is the simple outcome connected to O_i that is closest to – i.e. separated by the smallest number of links in the agent's chain of primary preferences from – O_j.[9] Given this, we can construct the degree of primary preference between two arbitrarily chosen outcomes using a combination of explicitly defined primary preferences and their summative consequences.

An example might make this clearer. Consider table 3.

[Table 3]

Here, the agent overall has an intransitive set of primary preferences: in particular, her degree of primary preference of O_1 over O_3 is not derived from her degree of primary preference of O_1 over O_2 and that of O_2 over O_3 – it is separately defined. However, all of her other degrees of primary preference *are* transitive: for example, the agent's degree of primary preference of O_1 over O_4 is – by equation (6) – simply her degree of primary preference of O_1 over O_3 (i.e. -5) plus that of O_3 over O_4 (i.e. 1), which yields a value of (-4). Thus, despite these preferences being intransitive overall, we only need *one* extra fundamental conative state compared to what was true in the transitive case. Importantly, together with equation (5) above, all *complex* outcomes (e.g. (O_1 & O_4) vs. (O_2 & O_3)) can be evaluated in this way as well.[10]

[9] Note that the fact that O_k is defined as above is merely one possible option; for example, we could also define it as the outcome connected to O_i that is furthest away from O_j. The key point is just that (6) needs to contain a *unique* way of creating some transitivity in the face of overall intransitivity.

[10] Note that, as such, my account does not require preferences to compose at all – i.e. my account does not need to assume the kind of value atomism assumed by Pollock. The point is just that it is *possible* to respond to the Pollockian argument of the previous section using the expanded notion of primary preferences of this section. Of course, if this Pollockian argument is not found cogent for other reasons – e.g. because of the value atomism it is based on – then there is no need to provide any kind of response to it. Note also that, if we do want to respond to Pollock, we might also want to allow for exceptions to (5) on the model of (6). If so, we could say (7) $PP'[(O_i \& O_j), (O_k \& O_l)] = PP'(O_i, O_k) + PP'(O_j, O_l) = PP'(O_i, O_l) + PP'(O_j, O_k)$, *unless it is separately defined as x*. Doing this would add further complexity to the argument, but not alter the point in the text in any way.

In this way, we can simultaneously satisfy the twin demands of, on the one hand, appealing to intransitive fundamental preferences to account for the observed choice intransitivities, and, on the other, ensuring that our fundamental conative states are not excessively numerous. It is true that, in doing this, we might need a few more fundamental conative states than we did in section IV, but this increase will not in general be so massive as to break the bounds of what can be implemented in physical systems. In short: with intransitive primary preferences, the picture gets more complex, but the main idea behind the above response to Pollock's argument remains intact – as long as our fundamental preferences are allowed to be graded, computational considerations on their own cannot determine whether they are the fundamental conative states or not.

Further, note that the account here can also make sense of the fact that RCT often seems to be a reasonably accurate picture of the way we make decisions (Glimcher et al. 2005; Hausman 2012). While it might not be *universally* possible to represent our fundamental preferences in terms of consistent utility differences (i.e. transitively), it might still be possible to *sometimes* do so (i.e. when the preferences in question do not involve 'irregular', separately encoded ones). Put differently, the present account can be used to underwrite the idea that the transitivity assumption of RCT is an idealization that will often, but not always, work. In this way, we can go some way towards justifying the central status of RCT in economics and many other scientific disciplines – while still maintaining that our fundamental conative states are intransitive and relational.

The second objection that might be raised to the account defended here concerns the (seeming) fact that, if our fundamental preferences are assumed to be (somewhat) intransitive, people seem to be highly vulnerable to exploitation by money pumps of one kind or another. After all, a reasonably smart and knowledgeable bookie could use the agent's proneness to decision cycles to make money by buying her initial endowment and then selling it back to her.

This might seem to be biologically surprising, in that we might expect agents that are this pragmatically irrational to have a relatively hard time surviving and reproducing (and thus, for mind designs like this to be able to evolve at all). It may also seem empirically surprising, as it may appear that we do not commonly observe this kind of radical practical irrationality.

To respond to this worry, two points can be noted. First, the fact that, despite their intransitive preferences, people are not constantly subjected to money pumps may simply be due to the world not being set up to exploit these features of our minds. Systematically money pumping another agent requires quite detailed knowledge of that agent's mind – and it is not unreasonable to think that this knowledge is rarely had by other agents (and especially not by agents with the motivation and ability to use that knowledge to their advantage). Secondly, it is plausible that people will strive to change their (intransitive) preferences when they realize that they lead to them being money pumped. This is because it is plausible that people prefer to avoid unnecessary losses of goods. So, when an agent finds that her preference for avoiding sure losses conflicts with a number of her other (intransitive) preferences, it is plausible that she will revise some of the latter. In this way, radical money pumping might also turn out to be quite rare due to its dynamical instability: while people often have intransitive preferences to begin with, they will attempt to make them transitive when they realize it matters.

VI. Conclusion

I hope to have shown that there is an important, but largely overlooked question concerning the structure of our minds: namely, whether our fundamental conative states are desire-like or preference-like. I also hope to have made clearer that neither the structure of RCT, nor Pollock's efficiency-based account, can be used to answer this question. I further hope to have shown that

the intransigence of the many empirically ascertained choice intransitivities supports the truth of the preference-based view: the latter allows for these intransitivities to be the result of the agent's fundamentally intransitive conative states.

Bibliography

Allais, Maurice (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine. Econometrica, 21, 503-546.

Bermudez, Jose (2009). Decision Theory and Rationality. Oxford: Oxford University Press.

Bradley, Richard, and List, Christian (2009). Desire as Belief Revisited. Analysis, 69, 31-37.

Bratman, Michael (1987). Intentions, Plans, and Practical Reason. Cambridge, MA: Harvard University Press.

Broome, John (1991). Weighing Goods. Oxford: Blackwell.

Broome, John (1999). Ethics out of Economics. Cambridge: Cambridge University Press.

Busemeyer, Jerome and Townsend, James (1992). Fundamental Derivations from Decision Field Theory. Mathematical Social Sciences, 23, 255-282.

Busemeyer, Jerome, & Townsend, James (1993). Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment. Psychological Review, 100, 432-459.

Carruthers, Peter (2006). The Architecture of the Mind. Oxford: Oxford University Press.

Clark, Andy (1997). Being There. Cambridge, MA: MIT Press.

Cox, James, and Epstein, Seth (1989). Preference Reversals without the Independence Axiom. The American Economic Review 79, 408-426.

Damasio, Antonio (1994). Descartes' Error. New York: Grosset / Putnam.

Davidson, Donald (1963). Actions, Reasons, and Causes. In Essays on Actions and Events (pp. 3-20). Oxford: Oxford University Press.

Davidson, Donald (1969). How is Weakness of the Will Possible? In Essays on Actions and Events (pp. 21-42). Oxford: Oxford University Press.

Eells, Ellery (1982). Rational Decision and Causality. Cambridge: Cambridge University Press.

Fodor, Jerry (1981). RePresentations. Cambridge, MA: MIT Press.

Gigerenzer, Gerd; Todd, Peter, and the ABC Research Group (1999). Simple Heuristics that Make Us Smart. Oxford: Oxford University Press.

Gigerenzer, Gerd, and Selten, Reinhard (eds.) (2001). Bounded Rationality: The Adaptive Toolbox. Cambridge, MA: MIT.

Glimcher, Paul; Dorris, Michael, and Bayer, Hannah (2005). Physiological Utility Theory and the Neuroeconomics of Choice. Games and Economic Behavior, 52, 213–256.

Goldman, Alvin (1970). A Theory of Human Action. Princeton: Prentice-Hall.

Goldman, Alvin (2006). Simulating Minds. Oxford: Oxford University Press.

Grether, D., and Plott C. (1979). Economic Theory of Choice and the Preference Reversal Phenomenon. The American Economic Review, 69, 623-638.

Guala, Francesco (2005). The Methodology of Experimental Economics. Cambridge: Cambridge University Press.

Hagen, Edward; Chater, Nick; Gallistel, Charles; Houston; Alistair; Kacelnik, Alex; Kalenscher, Tobias; Nettle, Daniel; Oppenheimer, Danny; and Stephens, David (2012). ‘Decision Making: What Can Evolution Do For Us?’. In Peter Hammerstein and Jeffrey Stevens (eds.). Evolution and the Mechanisms of Decision Making (pp. 97-126). Cambridge, MA: MIT Press.

Hausman, Daniel (1995). The Impossibility of Interpersonal Utility Comparisons. Mind, 104, 473-490.

Hausman, Daniel (2012). Preference, Value, Choice, and Welfare. Cambridge: Cambridge University Press.

Houston, Alistair; McNamarra, John, and Steer, M. (2007). Violations of Transitivity under Fitness Maximization. Biology Letters 3: 365-367.

Jeffrey, Richard (1983). The Logic of Choice. Second Edition. Chicago: University of Chicago Press.

Jeffrey, Richard (1992). Probability and the Art of Judgment. Cambridge: Cambridge University Press.

Johnson, Joseph, and Busemeyer, Jerome (2005). A Dynamic, Stochastic, Computational Model of Preference Reversal Phenomena. Psychological Review 112, 841–861

Joyce, James (1999). The Foundations of Causal Decision Theory. Cambridge: Cambridge University Press.

Kahneman, Daniel, and Tversky, Amos (1979). Prospect Theory. Econometrica, 47, 263–291.

Kalenscher, Tobias; Tobler, Philippe; Huijbers, Willem; Daselaar, Sander, and Pennartz, Cyriel (2010). Neural Signatures of Intransitive Preferences. Frontiers in Human Neuroscience, 4, 1-14.

Lewis, David (1988). Desire as Belief. Mind, 97, 323-332.

Lewis, David (1996). Desire as Belief II. Mind, 105, 303-313.

Loomes, Graham; Starmer, Chris, and Sugden, Robert (1991). Observing Violations of Transitivity by Experimental Methods. Econometrica, 59, 425-439.

Loomes, Graham, and Sugden, Robert (1982). Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. The Economic Journal, 92, 805-824.

Luce, R. Duncan & Raiffa, H. (1957). Games and Decisions: Introduction and Critical Survey. New York: John Wiley.

Morillo, C. (1990). The Reward Event and Motivation. Journal of Philosophy, 87, 169-186.

Nichols, Shaun & Stich, Stephen (2003). Mindreading. Oxford: Oxford University Press.

Pollock, John (2006). Thinking about Acting. Oxford: Oxford University Press.

Reed, E. (1996). Encountering the World. Oxford: Oxford University Press.

Rieskamp, J.; Busemeyer, Jerome, and Mellers, B. A. (2006). Extending the Bounds of Rationality: Evidence and Theories of Preferential Choice. Journal of Economic Literature 44: 631-661.

Rosenberg, Alexander (2012). Philosophy of Social Science. Fourth Edition. Boulder: Westview Press.

Ross, Don (2005). Economic Theory and Cognitive Science: Microexplanation. Cambridge, MA: MIT Press.

Savage, Leonard (1954). The Foundations of Statistics. New York: John Wiley.

Schroeder, Timothy (2004). Three Faces of Desire. Cambridge: Cambridge University Press.

Schroeder, Timothy. (2009). Desire. In E. Zalta (ed.). The Stanford Encyclopedia of Philosophy Winter 2009 Edition, URL = http://plato.stanford.edu/archives/win2009/entries/desire/.

Schroeder, Timothy (2010). Desire and Pleasure in John Pollock's "Thinking in Acting". Philosophical Studies, 148, 447-454.

Simon, Herbert (1957). Models of Bounded Rationality. Cambridge, MA: MIT Press.

Smith, Michael (1987). The Humean Theory of Motivation. Mind, 96, 36-61.

Sopher, Barry, and Gigliotti, Gary (1993). Intransitive Cycles: Rational Choice or Random Error? Theory and Decision, 35, 311-336.

Stampe, Dennis (1986). Defining Desire. In J. Marks (Ed.). The Ways of Desire (pp. 149-174). Chicago: Precedent Publishing.

Sterelny, Kim (2003). Thought in a Hostile World. Oxford: Blackwell.

Tsai, Rung-Ching, and Bockenholt, Ulf (2006). Modeling Intransitive Preferences: A Random-Effects Approach. Journal of Mathematical Psychology, 50, 1-14.

van Gelder, Timothy (1996). Dynamics and Cognition. In J. Haugeland (ed.). Mind Design II (pp. 421-450). Cambridge, MA: MIT Press.

Waite, Thomas (2001). Intransitive Preferences in Hoarding Gray Jays (Perisoreus canadensis). Behavioral Ecology and Sociobiology, 50, 116–121.

Outcome	Utility
O_1	10
O_2	9
O_3	8
O_4	1

[Table 1: Utility Assignments to Simple Outcomes]

Outcomes	Primary Preference
(O_1, O_2)	1
(O_1, O_3)	2
(O_2, O_3)	1
(O_3, O_4)	7

[Table 2: Primary Preference Assignments to Simple Outcomes]

Outcomes	Primary Preference
(O_1, O_2)	9
(O_2, O_3)	1
(O_1, O_3)	-5
(O_3, O_4)	1

[Table 3: Intransitive Primary Preference Assignments to Simple Outcomes]