Preferences, Game theory, and the "Harmony of the Soul"

Jan Narveson

Preferences have come to be the preferred item in utility theory. It's not hard to see why. Old-fashioned, hedonistic-type utility is a strictly subjective magnitude. Prattle tho we may about maximizing it, getting enough of a grip on the sucker to measure it even roughly is a pretty well-nigh insurmountable problem, leaving one stuck at the starting block (or, tripped over the first hurdle, to update the metaphor...) But with preferences, we get to say: if Jones prefers x to y, then Jones will damn well *choose* x over y, and this we get to watch.

Or do we? Of course, what we see is Jones picking up the chocolate cookie. This may show that he chose it, and that may show that he prefers it. But then again, maybe not. Can't we sometimes do what we don't prefer? Or do we have to say that if we *do* x, then we *prefer* x - whether we *think* we do or not?

It is certainly tempting to stamp one's dialectical foot at this point and insist that of course we must, "in some sense" have preferred to do it if we up and actually *did* it. (As opposed to finding oneself skidding down the slope, contrary to what one would *now* like to be doing..) The question then is, what is the "some sense" in which this "must" be so?

Let's see. First, it seems to me that a great deal of my action is pretty close to automatic. Suppose I do indeed prefer to be sitting here typing this; in the last sentence or so there are a few 't's' for instance. Did I prefer each finger-stroke resulting in a 't'? While I presume that "in some sense" I did, it would be hard to locate the preference operating during that very small fraction of a second in which I complete a given keystroke for 't.' So perhaps the sense that I might do things I don't explicitly recognize that I prefer might be spelled out thusly: a general train of action got initiated, and midway through it I am on autopilot, and realize that I'm not sure this is where I want to be - but the momentum (say) is too great for me to stop.

Second: If there is an alternative account of what preference is, it presumably is what we might call the "phenomenological" account, according to which preference is felt from the inside, and is what *causes*[1] us to act, but still, is not simply identical with the tendency to do whatever it is a preference for.

Of course 'simply' is misleading. Nothing is simple, as they say. My preference for p might, for one thing, be a complex of preferences for various components or side effects of p's being the case. And for a very big other, there is the evident fact that we need always to say that "other things being equal" we will choose what our preference is a preference for. Other things have a way of being unequal, and moreover, it is not always obvious where "other things" begin and the direct object of our preference ends. Perhaps characteristically, indeed.

Well, so what? Well, this: How do I "know" what my overall preference *is*? Here, suppose, are several distinct intensional magnitudes whose integrated vectorial sum is what we - and *I* - call my 'preference.' How do I do this problematic arithmetic? A suggestion that seems to me to have a lot of merit is that *I* look to see what *I* do and then infer from the result what my preference *must have been*. This is "revealed preference from the inside." If this account is right, then the forces pushing us toward reveal preference are very strong indeed.

Taming the Passions?

Let us now remind ourselves what moral philosophy is trying to do. (Or anyway, what it *should* be trying to do, and *would* be if it knew what it was doing....) Here we have The Individual with his jostling crowd of impulses, interests, emotions, desires. Over in some perhaps prominent corner of the soul is the Individual's

[1] We'll assume for present purposes that 'causes' is the right word here, though some would dispute it.

set of Moral Principles. What we want to know is how the latter are related to the former. What we want to get, actually, is Morals: that is to say, where it came from and what it does and why.

What it does (or rather, is supposed to do) is clear enough in general: it harnesses the passions, that's wot. What's interesting about this classic description is that there is some reason to suppose that it's impossible to do this. I mean, reason is the Slave of the Passions, right? So, how do we go about, *on the basis of reason*, doing what amounts to altering one's "passions"? That has been a problem plaguing this approach for some time.

And it is an interesting question, but we need to be fairly precise about which question it is. Some in the trade have thought that "reason" is underbilled in the post-Humean area and can do more than we credit it with. I am not sure whether this is true. On the other hand, I'm quite sure that we won't have much luck trying to "be rational" if that means to deduce what your actions are to be from axioms of mathematics or logic alone. I don't think it's surprising that people like Kant who allegedly say such things also drag in things like "respect" to help matters out. But I think that there is, in a way, a simpler solution. Or maybe there's a clarification of just how "reason" has to do with practice in general, and "passion" in particular.

Every desire, every interest, every evaluation we make has, on the general view we are inclined toward in the CARP biz, a bit of "push" to it. But as noted, these are extremely various, and, as Plato noted, they push us in different directions, often enough. The question we need answering about action, then, may be put like this: when does Push translate into Shove? That is to say, among the n conflicting pushes we are subject to, how does the one that actually gets us into action come about?

A metaphysically scary scenario?

The metaphor (if it is one) we are working on suggests something disconcerting - something like this: that the various "forces" inherent in the various "passions" each has a vectorial assignment composed of its direction and its force-measure (intensity), and when you get a bunch of them on the scene in a suitably relevant way, then what you *do* is the result of a vectorial integration. It's a matter of the integrated vectorial sum of these different forces.

On that picture, got from Leibnizian mechanics, I guess, it would seem to follow that there's no place for reasoning about what one does. Each force has its vectorial number and the integration of them all can be read off what we do. So: Revealed Preference rears its head! And what else , after all, can there be?

On more or less Knowing what we're Doing

Well, it seems to me that there is one thing missing from this picture, and it's an essential thing. At least, phenomenologically it is - but I don't think that the difference of viewpoint on action is relevant in the end. What's missing, or at least not obviously mentioned, is this: when we act, we have a *description* of what want, and of what we propose to doabout it. We have a description of the former because otherwise, as Aristotle put it, action would be "empty and vain"; which is an understatement, I think, since the point is that if you have no idea at all what you're trying to bring about, then you can have no idea what to do. And we need a description of the latter, our "plan of action" (or, in typical cases, just, our action), without which, again, it's hard to see how we could do *anything at all*. These things give us the sense of action's being "intelligent." What's intelligent about it is that it's guided by descriptions, not just pushes and pulls. This is not to say that there are no pushes and pulls or that they are ghostlier than we thought. It is merely to say that action is, obviously, intentional.

Now, that's why thinking about action is possible, as well as why it's necessary. It's possible because

formulating our ends is required if we are to have any idea where to look for what to do on their behalf, and formulating our means is required if we are to have any idea what to do.

Now, these descriptions are, it seems to me, an inherent part of the machinery of action. If someone says that we are prey to blind forces, my reply would be that if the force is genuinely blind, it's hard to see how we could be prey to it *in action*. Which way is a blind force going to push us - being blind, after all?

One model of course is entirely physical. You push me down stairs, and I fall: *now* the metaphor of vectors isn't a metaphor at all any more, but simply the straight account of the matter. But neither is falling down stairs an action (in the normal case; stuntmen are another matter.) When we have an *action* on our hands, we have intentionality, and then we have descriptions.

The "descriptions" we are talking about need not be and characteristically are not very well articulated. I take it that they usually aren't because they usually don't need to be. We're walking along the street headed for the corner store, and suddenly there comes into view a new store, not previously at that location, and now we may have to decide which one to patronize. Or, there's a question whether this one is open on Sunday and suddenly we remember that it is not, thus calling for some change in plan. The "pictures" we have of what we are trying to achieve and how we think to achieve it can be made more precise, more articulate - or not. What matters is that they are there.

Now, it is very clear that what we are going to do is affected by the content of these descriptions, articulate or no, and can or will very likely change when we know more. I take it too that the "descriptive part" of these descriptions is not what moves us to action, in and of itself - Hume has the field here. What makes a description functional is the desire, the interest *in* the item so described. Intellect, as Aristotle also notes, of itself moves nothing. But since "push" of itself pushes us nowhere in particular, the situation now changes regarding the plausibility or applicability of the vectorial idea. The *direction* in which we will be moved by given desires is a function of those descriptions. And so it is not true that our actions can be deduced, or more precisely could by deduced by some superpotent mad scientist type, on the basis purely of a calculation of all those pushes by themselves. Where we shall actually go in response to all this is a function of the more precise descriptions we assemble, and which we must assemble, it seems, if we are to act.

So consider any of the innumerable cases in which a decision needs to be made, owing to an inflow of information, or an indeterminacy of what we have so far. What's going on here? Evidently, a need to make one's information base more precise. (This last could be made to sound trivial by counting as 'information' statements about what one wants. But I don't think it really is. If we've made up our mind what we want, then the claim that we want it becomes information.) Anyway, the process of doing this can be identified as the reason why the mechanistic thesis about our desires can be rejected as either false or not interesting. It's not until we've made these decisions that we know what the sums are, and in that case there's little point in supposing that our decisions are a straight function of some antecedently existing forces, our desires.

But that, I think, provides the answer to our question about the Aristotelian project. We can tame our passions because we can extend and clarify the information base from which we act.

Game Theory and the Harmony of the Soul

Plato is famous for drawing an analogy between the individual soul and the state. The comparison was not entirely satisfactory, of course, but this is not the place for a general discussion of that point. But I think we can put the question we pursue in moral philosophy with some precision by reflecting on Plato's idea. The general thesis is that *justice* is *harmony*. Plato supposes that this is an idea with quite a lot of zap. Plato saw that in order for morals to work, we need in some sense to *want* to be just. But this seems risky: if we are just

only if we want something, and it's something we might *not* want, then morals seems to be a risky business. Modern commentators suppose that Plato's project runs aground on this. It seems that justice doesn't have the sort of necessity Plato wants it to have, since it seems to be something that turns out to be a matter of taste.

A related problem was that it isn't clear how Plato is going to connect the idea with our usual understanding of justice, an understanding that Plato shows he both knows and shares in various parts of The Republic. Or rather, let us say that it isn't clear how Plato *could* connect these, for what he does in the book, as we know, is to cheat: he assumes that people who are just in the sense developed in his theory will also be just by ordinary lights. *Will* the person whose soul is in harmony also be one you can rely on to keep his agreements? Will he knife you when your back is turned? If not, why not?

The Place of Preference(s)

The reader may at this point wonder what all this has to do with preferences. That will come out in what follows.

Question: suppose that we think we prefer r to s and s to t and also p to q; but it turns out that if I choose p I shall wind up with s or even t. It may well be thought that uncovers this - that is, more information. But it readily happens, as we know. In the new situation, *what do I prefer?* My suggestion is that until you do the homework, *you don't know*.

Revealed preference says that whatever you do is what you prefer. But in the case where I scratch my head and go home to think about it, what I "do" about the actual zero-level choices in question is, nothing, yet. This doesn't mean that I prefer to do nothing, except in the sense that I prefer at the moment to put it off until I figure out what I want. What it means is that I don't know what the hell I want!

The Anatomy of Harmony

I take it that the hallmark of the Harmony of the Soul is, Absence of Conflict. Desire D1 is in conflict with D2 if, if D1 is fulfilled, then D2 is *not*.

Admittedly, some of these conflicts will be due to external circumstances, and others due to internal confusion, including perhaps bad logic. External circumstances can make it the case that p, though logically compatible with q, is not extensionally compatible with it, alas. Internal confusion could mean that we've got major work to do, of the more or less psychiatric variety. In all these cases, we are now faced with problems whose general form is that *we don't know what we want and must somehow decide.* The idea that our preferences are "given" doesn't have much going for it when we realize all this. Try talking about revealed preference theory to the consumer trying to choose among umpty-three brands of toothpaste or cookies, etc. - or to the lover trying to decide between Al and Bob. We can guarantee that revealed preference talk is going to do him no good whatever. But wasn't game theory supposed to be, possibly, of *some use* for *making* decisions, as distinct from mapping the decisions people do make?

Consider our old friend, the Prisoner's Dilemma. The official theory has it that in a one-shot game, defection is rational, certain - indeed, it's analytic that the rational player will defect, even when that rational player is so rational as to be aware that the other player will do the same and that both in consequence will do worse than if they had both played the cooperative strategy. But the claim is that this won't make any difference, because each "knows" that if the one player does try to cooperate, then the other should defect, thus doing best of all.

In real life, however, people in situations that look quite a lot like PDs often do not defect. Why not? We

all know some answers that are pretty good - good enough, anyway. Sometimes we know the other party and don't want to do him down. Sometimes we think that this just may not be the last time we play and maybe a habit of reciprocation might get started. Indeed, we often know perfectly well that it has got started, has been going for quite awhile in fact - and this counts with us. Some will have read David Gauthier and been persuaded that it's irrational to adopt the disposition to defect in all cases. And maybe more.

Now, one way to put the moral of this is that PDs have certain abstractly defined properties, but that real life does not. It insists on being a lot messier. And so we should always say that we are addressing what *might otherwise be* a PD (or chicken, or whatever it is) - *if it weren't for* factors f, g, and h. Or to put it another way, in real life, we have to add clauses like "other things being equal" or something of the sort - enough to fudge things.

The (a) use of Game Theory

I have been insisting all these years, and will boringly do so once again, that what we *can* learn from game theory even in the real world is *what a cooperative outcome would look like*; and we can conclude that efforts to bring about a habit of acting in such a way as to realize those outcomes are worth making, and even have some prospect of success.

How would such an effort ever work? Apparently by installing some more software in ourselves: a sort of override mechanism which goes "beep" when one finds oneself tempted to play the Defect strategy against people who don't obviously appear to be similarly inclined. The beep, of course, has to be wired to some software that has a genuine effect on one's decisions - otherwise, it would just be irrelevant noise, and there'd be no point in installing it. Game theorists might be read as having assembled reasons why no such thing is possible. They might: but they shouldn't - first because it obviously *is* possible, since it happens all the time; and second because to conclude this would be to confuse diagrams of aspects of the real world with the real world itself.

Morals and the Harmony of the Soul

So, what about the "harmony of the soul"? Socrates is notorious (among us contemporary philosophers) for holding that justice has a whole lot of *intrinsic value*. If it did - or rather, if we thought it did - then our behavior would be quite different from that of people who deny this, or so we think. But the trouble is that intrinsic value is a loose cannon: there's no accounting for it. Justice, however, is something that we *need* to have an account of. It's no good letting people loose with beliefs about intrinsic value as a way of helping to get justice done, for if it's intrinsic, then there's no way to be sure we won't get the wrong job done.

What game theory does is to help us see what the "right job" is. We organisms who live at close enough quarters so that our actions significantly affect each other, and live in ways that are amenable to description, communication, and appraisal, can see how things might go if this, that, or the other sort of strategy were typically utilized, and so can see point in making proposals to our fellows for helping matters out in those respects. There is "point" because we'd do well, together, if a habit of cooperation is successfully instilled, and we think we are organisms amenable to this kind of appeal. And we do badly if we do not.

Although there are various ways in which we might be wrong about the latter, I nevertheless think that cooperation as the social norm has the weight of reason behind it; and I think game theory analysis is useful in enabling us to see that.

Now, we can view the situation in perhaps two ways as regards Preference and the Harmony of the Soul. Preferences in the first instance are just that - in the plural, and carried around by the organism prior to

application to action. But because of all the aforementioned complexity, how to arrive at the ultimate preference that determines action in individual cases is no simple or obvious matter. What we'd like to say, and what accounts, I think, for the appeal of the CARP way of thinking about these things, is that *if* we've got all these various preferences, *and* the relevant information about what happens if this or that one is acted on, *then* we have also, or can get, the preference that determines action. That is to say, it is, we suppose, a matter of *reasoning and analysis* that we'll do best by the whole range of preferences we have if we do *this*, and *so* we do it. The preference to do *this, now,* in other words, is not one we need to defend by appeals to intuition or whatever, but rather one we can form by doing our cognitive homework, given the input preferences we have.

But, of course, those also, often, are unclear and shifting, and so we have to do whatever sort of homework is needed about them, too. Both because they are and because in any case they'll be highly variable among people, any thought of putting out general dicta or "rules" for getting us all functioning together must be oriented toward coping with the expected diversity. Again, I think that game theoretic analysis can be of great aid in showing how to do this.

The "harmony" of the individual soul, we could now propose, is the condition of finding a (or *the*?) set of strategies that is best, or at least well, suited for achieving an *optimal* balance and thus *maximal* satisfaction for one's whole utility schedule. Then our thesis about morals in particular is that it should *flow from the above operation when taking into account the facts of social life* - viz., that there are lots of others around who are different enough to create problems and like enough to enable solutions of the otimizing kind. One's *preference* for doing the right (sort of) thing will then be derived from the rest, and yet have as much independence as is needed to carry us across the sundry gaps of temptation inciting us to engage in defection.

I hope that this is neither too much of a mouthful nor guilty of naivete. What I have tried to do is identify reasons why no sort of "mechanics" of the passions will get us to satisfactory action theory even if Hume is right (as, it is assumed, he is). Ken Binmore doesn't like talk of being the "slaves" of our desires, and in a sense what I'm doing here is finding reason to believe him, even though one can see why Hume says this, and that he must be somehow right. He's right that in the absence of passion-like stuff, no action is possible. But he'd be wrong if it was supposed that we are then pushed and pulled by those passions in a way that crowds "reason" out of the picture. Passions (in the very broad sense needed to make Hume's idea plausible) in the nature of the case require information, including information about relations to all those other passions, and all that crowds mechanism off the stage.

Of course, it might be held that there is also a mechanism of reason, say in one's neurons which are what function as one's computer system. Oddly enough, I think it would make *no difference if this is true.* We know what thinking "feels like" - we do it all the time - and if there are computational processes whirriing away underneath all this, that is something that doeosn't matter in the least to us. For we still have to *do* our thinking, be there an account or not.