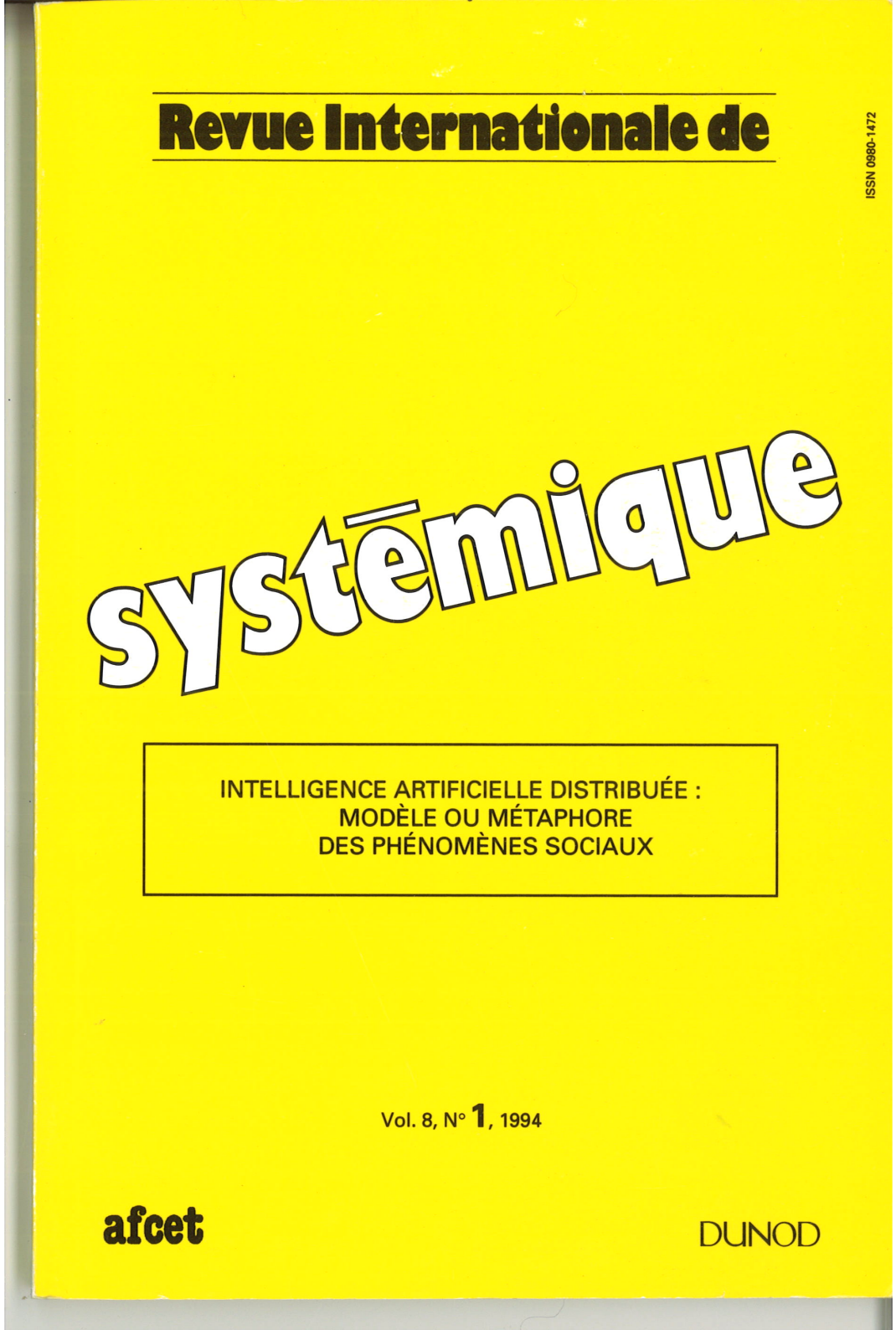

AFSCET

Revue Internationale de systēmique

Revue
Internationale
de Sytémique

volume 08, numéro 1, pages 91 - 103, 1994

Preference and Rationality

Samuel Guttenplan

Numérisation Afscet, janvier 2016.

REVUE INTERNATIONALE DE SYSTÉMIQUE
Vol. 8, N° 1, 1994, pp. 91 à 103

PREFERENCE AND RATIONALITY

Samuel GUTTENPLAN [1]

Résumé
Mon objectif est de suggérer que la théorie de la décision est fondée sur une idéalisation susceptible de distordre plutôt que de clarifier notre compréhension des choix humains. A la base de la théorie de la décision on demande que les préférences rationnelles soient transitives. On considère comme irrationnel pour un agent de préférer A à B, B à C et C à A. En effet, tant que cette figure de choix ne serait pas refusée comme irrationnelle, il se révèlerait impossible de construire une échelle d'utilité cardinale pour un agent. En utilisant une analogie particulièrement suggestive fondée sur une certaine sorte de jeu de dés – un jeu dans lequel les chances relatives de gagner avec un dé donné se révèlent non transitives. Je suggère qu'il y a beaucoup de points communs entre le choix dans un système multi-agents et la décision individuelle. Et, justement, on connaît depuis longtemps la présence de non transitivité *rationelle* dans les choix des systèmes multi-agents (au sens du paradoxe de Arrow).

Abstract
My aim is to suggest that decision theory is based on an idealization which may distort, rather than clarify, our understanding of human choice. Basic to decision theory is the demand that rational preference be transitive. That is, it is considered irrational for an agent to prefer A to B, B to C and C to A. For, unless this pattern of choices is condemned as irrational, it will prove impossible to construct cardinal scales of utility for an agent. Using a particularly suggestive analogy based on a certain kind of dice game – a game in which the relative chances of winning with a given die turn out to be intransitive – I suggest that there is a lot in common between multi-agent choice and individual decision. And, crucially, multi-agent choice has long been known to display *rational* intransitivity (in the sense of the so-called Arrow's paradox).

1. Birkbeck College, Philosophy Department, 14, Gower Street, WC1E 7HX-UK-London, Great Britain.

Rev. intern. systémique. 0980-1472 Vol. 8/94/01/ $ 4.00/

I

Idealisations play a prominent part in philosophy. For example, in understanding what it is for one theory to be better confirmed than some competing theory, one might find it helpful to see the first as the theory that would be accepted by a fully rational observer. Such an observer would be understood as equipped with the rules of inference, observations and background knowledge available to any of us, but he would be credited with the ability to apply those rules flawlessly and to keep in mind all the relevant observational and other knowledge. This ability is, of course, something human observers lack, though for apparently incidental reasons. We forget things, misapply rules of inference and cannot always keep relevant information in focus, but it seems a harmless idealisation to ignore these defects in the notion of a fully rational observer. Metaphorically one might almost think of this observer as locked up inside our less than rational selves.

There are dangers here though: there could be cases in which an idealisation distorts rather than aids our thinking. For example, it has been argued that the idealisation used by some writers in ethics of a perfectly impartial observer is unhelpful precisely because there can be no standpoint of such impartiality. If this is right – if all real ethical decisions require us to be partial in one way or another – then this particular idealisation does more to distort than to illuminate our conception of ethics.

II

The idealisation which I should like to consider in this paper is that which goes under the name of "decision theory". Since I cannot expect that familiarity with this subject, I shall devote a bit of time to expounding it. The basic aim of decision theory (on at least one way of understanding it) is to provide a plausible and revealing account of the explanation of human action. That is, it seeks to make explicit the reasoning processes which in fact, or at least in principle, go into our choices of actions when we either deliberate about our actions or, at least, act deliberately.

The work that forms the centre of decision theory can seem quite technical but the principles on which it is based arise from two fairly obvious features of deliberate action. The first is this: when we decide to do something, we usually weigh up our wants and act so as to maximise the satisfaction of our desires. If you like "foreign" films more than Hollywood films then, given a choice, you will

almost certainly choose the former. The second consideration can be seen as to some extent a restriction on the first. It is this: We do not make our decisions *solely* on the basis of our wants since this would lead to actions that most of us would think of as mad. For example, if you are betting on horseraces it would be insane always to choose the rank outsider merely because what you want in betting is the most money. Our decisions are certainly based on our wants but they are no less based on our beliefs about the way the world will turn out. If you think it *very* unlikely that a horse will run, then even if you stand to gain a lot if it does, you may well think it unwise to back it. Combining these two common-sense thoughts we get the unsurprising result that our actions guided both by our beliefs – in particular our beliefs about the future course of the world – and our wants. What is perhaps more contentious however is the way that decision theory transforms these thoughts into a principle of rational action.

Utilities	Subjective probabilities
I most want_____	I am convinced that_____
I quite strongly want_____	I strongly believe that_____
I want_____	I am fairly sure that_____
I would like_____	I think it likely that_____
I wouldn't mind_____	I am tempted to believe that_____
I am indifferent to_____	I am uncertain whether_____
I don't care for_____	It is possible that_____
I wouldn't like_____	It is unlikely that_____
I strongly dislike_____	I would be surprised if_____
I most dislike_____	It cannot be that_____

Figure 1.

Imagine that you know someone so well (perhaps yourself) that you can completely rank his wants or desires from strongest to weakest. In the trade this is, perhaps confusingly, called a "utility ranking". Also imagine that you can similarly rank his beliefs at any time about how the world will turn out. That is, you can say of him that he thinks such-and-such more likely than... and etc. This is his subjective probability ranking. The common-sense thoughts just discussed support

the idea that these two rankings (*see* figure 1 below) will, in a given context of decision, combine to yield some most favoured action; the problem is that we have as yet no precise principle of combination.

Suppose, however, that we had more than the mere rankings of utility and probability. Suppose we had proper scales of utility and probability for our subject. That is, imagine that we can not only say where for example two items come in the utility ranking, but we can say that for instance one is three times as much preferred as the other. And instead of saying merely that our subject thinks rain more likely than snow, we can say that he thinks rain twice as likely as snow.

If we can give the rankings this much precision then it is a simple further step to assigning numbers to each item in the ranking – the numbers can be thought of as indices of utility and probability or even as measures of utility and probability. Thus suppose for example we knew that our subject ranked three films A, B and C in the *order* A, B, C and that he desired to see A two times as much as B and B three times as much as C. Then if we think of C as giving him *one* unit of utility or satisfaction then B will have 3 units and A six units. The choice of numbers for one item is of course arbitrary but, having fixed it, all the other items in the scale will have non-arbitrary numbers assigned to them. A similar supposition can be used in respect of the subject's probability rankings, though for reasons I think fairly obvious the numbers should be thought of as falling within an interval and as obeying the usual axioms of probability. Using the conventional interval 0-1 we can say that if our subject is fully *certain* that p is true, then his assignment to p will be 1 and not-p 0. If he thinks p is as likely as not-p then both will be assigned 1/2 etc.

Having first thought of utility and probability for our subject as ranked or ordered and now as indexed or characterised by what is called a "cardinal" scale, we are finally in a position to say something about how the two scales combine to give us a principle of rational action. (Remember, though that the provision of scales is as yet only a supposition.) The central clain of decision theory is that in deciding which action or choice is best we multiply the units of utility for each action by the subject's estimate of the probability that events will turn out appropriately. This sum is called *expected utility* and rational action is said to consist in the *maximisation* of expected utility. Here is an example.

Our sujbect has a great desire to meet up with a certain person on Friday evening but he doesn't know which of several social events she will attend. In particular, there is to be a disco at the student union, a wine and cheese party at someone's home and a sherry party in his department. We shall assume he can attend only one of these events. What he has to decide is which event to attend. The table

below gives his utility ranking of the events with and without the object of his desire (X) present as well as the probabilities he estimates of X's being at one or other of these events. For example, he would get most satisfaction out of running into her at a disco, but he would hate to be at the disco if she weren't there. This explains the choice of 10 and 1 in the first row of the utility ranking.

Table 1.

	Utilities table	
	X present	X absent
Disco	10	1
W&C	8	4
Sherry	6	0

	Probabilities table	
	X present	X absent
Disco	2/5	3/5
W&C	1/4	3/4
Sherry	2/3	1/3

What the principle of rational decision demands is that one multiply the utilities by the probabilities for each box in each row and then add them across rows. The row with the highest total is the rational choice. In the specific example, row 1 gets 4.6 ($10\times2/5 + 1\times3/5$); row 2 gets 5; row 3 gets 4. So our subject should elect to go to the wine and cheese party, since this is the course of action with the highest expected utility. Remember, the product/sum of each row of the two matrices gives the expected utility of choosing that option.

The very possibility of dealing with the example in the way decision theorists do depends of course on the assumption that cardinal scales of probability and utility can be drawn up. Expected utility, the notion at the heart of the decision theoretic account of rationality, is the product of these two scales. Moreover, basic to the possibility of cardinal scales is the assumption that preferences and probabilities can be ordrered or ranked in a way that is fine-grained enough to be represented by numbers. Both of these are, of course, meant as idealisations of

what goes on in decision making, though the time has come to do something on behalf of the decision theorist to discharge the weight which the above assumptions have been made to carry. For as I have explained it, we only get the principle of maximisation of expected utility on the assumption that we can construct cardinal scales and, whilst this may be helpful as a way of *understanding* that principle, we do not need to see the construction of cardinal scales as an *assumption* on the decision theorist's part.

A more accurate picture of how the decision developed is this: if we make the not implausible assumption that our actions are chosen in part by an underlying merely *ordinal* ranking of preferences and by our assessments of *some* probabilities then we can use this data to actually *construct* complete cardinal scales of utility and probability. The way this is done need not concern us here however and would in any case take too much time to explain. My main interest is in the expected utility principle as an idealisation of ordinary decision making and the thing to keep in mind is this: the model of rational decision making given here is one that requires us to make the following idealisations:

1. We must see the agent as having an underlying *ordering* of all his preferences or desires.

2. We must see him as being able consistently to assess evidence so as to be in a position to estimate probabilities for his judgements about the future.

3. We must see his decisions as the joint product (in any given case) of the scales of utilities and probabilities which can be constructed using 1 and 2.

Now there are no doubt many ways in which this picture of the rational decider is misleading; many ways in which it is untrue to features of ordinary decision making. For example, this model doesn't take account of changes of mind or of built in biases against taking chances – both features of human decision making. Still, what I have presented is merely the starting point of decision theory and the model can be made more realistic by being made more sophisticated; though it will of course always be an idealisation of that often haphazard procedure we call "deciding".

The objection that I have is of a different sort. I think that these are reasons for regarding one of the basic assumptions of the model as misguided and hence that no amount of tinkering with it will allow it to capture what we intuitively think of as rational decision making. In particular, I should like to look more closely at this idea that our decisions can be seen as in part based on an underlying ordinal ranking of preferences. In order for this ranking to serve the needs of the expected utility principle it must be transitive. That is, it must be the case that if

A is preferred to B and B to C then A is preferred to C. This is because unless the ranking is transitive no cardinal scale can be constructed and the expected utility principle will be inapplicable.

Transitivity does seem a sensible feature of one's preference ranking. We all know that through inattention or forgetfulness we can lapse into intransitively ordered choices, but we do usually regard them as *lapses*. If you would prefer a holiday in the sun to a skiing holiday and a skiing holiday to camping in Wales then it would seem obvious that you must prefer a holiday in the sun to camping in Wales. Were you to dissent from this then it would appear that either you had changed your mind or had somehow made a mistake. In spite, however, of the *prima facie* evidence in support of transitivity, I shall argue that, in at least some cases, it is fully rational to order preferences *intransitively*. The argument will proceed by stages and I begin by inviting you to consider what may at first seem both paradoxical and irrelevant to decision theory.

III

Suppose I show you four dice (A, B, C, D) which have non-standard numbers of spots on their surfaces, and offer to play the following game: you first choose one of them, then I choose one of the remaining three. We then "shoot" the chosen dice as many times as you would like, the winner of each round being the die with the higher number face up and each round paying the winner £1. Most would think that this game favours you since you are in a position to examine the dice and chose the one with the best chance of winning. Surprisingly, however, it is possible to construct the dice in such a way that this isn't so. When the dice are designed as in figure 2 below, it turns out that A is a more probable winner than B, B than C, C than D and D than A. Not only that but the advantage in each case is significant. So the person who chooses second stands to win £1 much more often.

At first these results might seem to be paradoxical for probability theory (and I came across it in this connection a number of years ago). For what we appear to have is a case of intransitive probabilities, *viz*: A's winning is more probable than B's, B's than C's, C's than D's and not A's than D's but the converse. This, if so, would of course undermine the whole of probability theory since that theory regards various events as having numerical probability indices, and numbers in their very nature are transitive. But on reflection it is not difficult to see why, surprising as the dice example is, it is not a problem for probability theory.

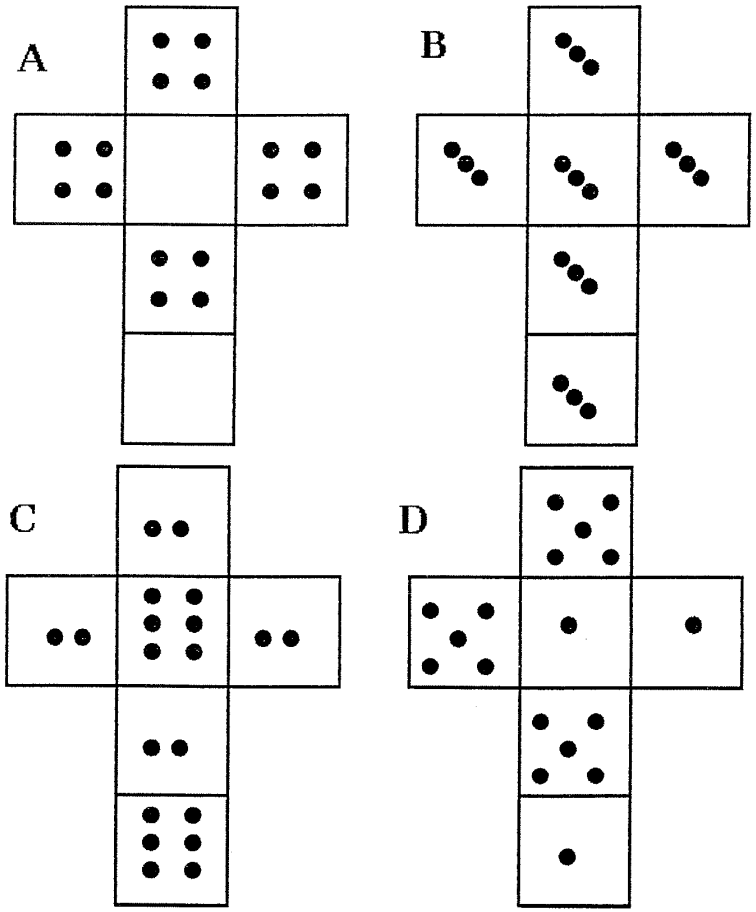

Figure 2

What has a probability assignment is an event. For example, of any of the dice we can say what the probability is that a certain number will appear face up – namely 1/6. And given any two of our dice we can assign a probability that one will have a higher number face up when *both* are tossed. But this latter event is one which depends on the properties of both dice; properties I shall describe more fully shortly. There is no number assigned to each die that is a measure of its

underlying superiority over *any other die.* In this it differs from the assignment of probability to events. These can be seen as indices of the events likelihood, and we can make direct numerical comparison between them. The only probabilities we get in the dice example when comparing any two of them are probabilities of a die's beating another. That is, we can assign probabilities of each of the following:

A beating B
B beating C
C beating D

but these are all independent and there is no reason to infer from them that A will or will not beat D to any particular degree of probability. To discover that we need to look at the faces of the dice and do a seperate calculation. (Actually, the probabilities of the above three are all 2/3. In other words, A beats B in two out of three tosses, B kbeats C in two out of three and C beats D in two put of three. So it is somewhat strange to most who come across the example to find that, when you do the calculation, D beats A also in two out of three tosses.)

Having then seen that there is no paradox here for probability theory, what about the transitivity of preference? Surely it wouldt appear that an ideally rational agent would choose A over B, B over C, C over D and, intransitively, D over A. Don't we then, in this case, have at least one bed-rock example in which it is positively rational to prefer intransitively? The dice example does, as I shall claim, argue against transitivity as a feature of rational decision, but this rather direct use of the example is too quick. To see why this is so it is necessary to be clear about the context of the preferences for one die over another. Strictly, the player who goes first, and who is aware of the relationships between the dice, will not have a preference for one over another. Since we can assume that his overall preference is for winning and since he realises that he cannot win whichever die he chooses, he ought to be indifferent in his choice of a die.

The player who goes second certainly does have preferences but, and this mirrors the probability case, they are not straightforward preferences for each die. His preferences will take a conditional form; *i. e.* if the first player chooses A, he prefers D, if B, then he prefers A etc. When seen this way, no interesting conclusions follow about intransitivity. The second player doesn't have an underlying strength of preference for each die any more than each die has an underlying probability of beating other die. In playing the game he exhibits preferences which depend on his knowledge of the dice and on the choice of the first player. So, for example, even if we known that:

if B is chosen, he prefers A
If C is chosen, he prefers B

we cannot infer that the hase any stronger preference for A if C is chosen. In fact, the strenght of preference for A given that C is chosen is a matter for a separate calculation – one that I haven't in fact made, though anyone can do so by examining the faces of the dice. So whilst it is appealing to think of the dice example as making trouble for probability and prefernce, it does not do so directly and, interestingly, for the same sort of reason in each case. Even more interesting, though, is that a bit more reflection on that reason will, I think, provide us with very strong *indirect* reason to reject the requirement of transitivity of preference in general.

To see why this is so, ask yourself why there can be no underlying probability assigned to each die which is its index of superiority over any other. It is because each die has six faces containing numbers which, depending upon the die it is played against, give it an advantage (or not) against the other one. The four fours in A are enough to give it a significant edge over B. But this tells us nothing about how it would do against C even when we know how B does against C. In each case the grounds of superiority vary, so it was possible to desing D so as to beat A. If one thinks of each of the six faces on a die as a *respect* in which it can be potentially a winner or loser against another, then it is clear that each pairwise comparaison depends on the relative strengths of the different respects. Since in each case *different* sets of respects are responsible for superiority in the game, it is not too surprising that we cannot be sure that superiority behaves transitively.

Crucially, however, there is no reason why many of our preferences should not be seen in the same way. When I have to make a decision I may well (even if informally) compare items by means of various respects in which they differ. And it may turn out that there are enough respects in which they differ. And it may turn out that there are enough respects to make it likely that my preferences will behave intransitively. for example, suppose I have to choose a person to fill a certain academic job. I might find Jones is to be prefered to Green because Jones is a better lecturer and has more research potential, and that these outweigh the fact that he is less valuable as a colleague. Green, however, is definitely to be preferred to Smith because whilst Smith is the best researcher of the three, he comes behind Green in lecturing ability and as a colleague. Of course it doesn't follow from this that Jones is the best candidate since, if you look at the table 2 below, you will realise that Smith ought to be preferred to Jones. This is because Smith is the best researcher and is also a better colleague than Jones. He "wins", so to speak, in two of the three respects.

Here then is a case in which preferences for the candidates behave intransitively and for precisely the same formal reason as illustrated by the dice example.

Moreover, like that example, this means that we cannot assign an underlying index to each candidate which is the index of our preference. So the dice example gives us a way of conceiving of our decisions as rational even when they turn out to be intransitive. But, of course, if this is accepted then the idealisation embodied in the expected utility model of rational decision will be grossly distorting. There can be no assignment of cardinal utility indices since the transitive ordering required for the assignment isn't something we should expect. For even a idealised rational decider will choose intransitively at times. I shall end this paper with an observation about the dice example and a brief consideration of two ways in which someone might take issue with what I have said so far.

Table 2.

	Lecturing ability	Research Potential	Value as a Colleague
Jones	First	Second	Third
Green	Second	Third	First
Smith	Third	First	Second

IV

The dice example is formally parrallel to what has been called the "Arrow paradox". This is the apparent paradox which, it has been argued, affects democratic choice. The idea is that, in an election with three candidates, it could happen that even through each voter had a transitive ordering of his preferences for the different candidates, the majority might end up favouring A to B, B to C and C to A. I could, of course, have used the Arrow paradox to construct a case in which the *rational* choices of an individual behave intransitevely as if they were the choices of three different individuals. However, I think the dice example is more persuasive since it does not require us to think of the preferences as those of three different individuals. The faces of the dice are respects in which a single die is compared to others. And the intransitivity arises from these respects.

I can think of at least two ways in which someone might challenge what I have said.

(A) It might be argued that, given my intransitive ranking of Jones, smith and Brown, I should see myself as like the first player in the dice game – as indifferent

among the alternatives. It certainly does seem that in the job-filling example I don't know which of the three to choose. However, this move doesn't really help. Suppose someone said of me that I was indifferent with respect to Jones, Smith and Brown. This would lead the decision theorist to assign the same index of utility to my preference for each of the candidates. But, of course, if one of the candidates were to withdraw then I would certainly be a position to choose decisively between the remaining to and this would conflict with the assignment of equal utilities to each choice of candidate. The problem here is that my inability to rationally choose from among the three candidates (just as my inability to rationally choose a die when I am the first player) is not a sign that I rank the canditates equally. To think that it is, is to beg the question against the argument of this paper. What is at issue is whether my preferences can be ranked so as to allow the eventual construction of a cardinal utility scale. It cannot, therefore, be assumed from the start that there is such a scale and that indifference or inability to choose is a sign of equality on the scale.

(B) The second objection seems to me much more interesting. In outline it goes as follows: it is true enough that when we rank certain of our preferences in complex cases we can begin with intransitive rankings and for just the reason described. However, we ought not to accept this situation as rational; we ought to go on to *weight* the respects in which we compare things so as to remove intransitivity. So, for example, in the academic appointment case we ought to come to some decision about the relative importance of teaching, research and quality as a colleague.

There are two things to note here in replying to this objection. First, whilst it will in some cases be possible to remove the intransitivity in this way, it will not always be effective. If the ranking depends on enough different respects then intransitivity will remain if all we do is *order* the importance of the respects in which the options differ. This can be seen without describing a case merely by consideration of the dice example: if the sides were numbered (in addition to having numbers of spots) then even if we said such things as winning on face 1 is more important than winning on face 2, and perhaps worth £1.50, we could construct the dice so that intransitivity remains. In order to fully guarantee that intransitivity disappears we have to be able to assign a *cardinal* ordering to the respects in which things differ. We must, that is, be able to say such things as research is twice as important as teaching, etc. This would come out as equivalent in the dice example to counting the number of spots on the faces as like a score. A six on one face of a die and a five on another die would not be counted simply

as a win for the first but as the first's scoring a one point victory. If you go through the example, you will see that this will guarantee that one of the four dice will have the best chance of obtaining the highest *score* against any of the others.

My worry is this: I can just about imagine that in some cases it is possible and even desirable to order the respects in which our comparisons are made. But I can imagine cases in which we would be more inclined to think of respects as incommensurable. Can we seriously think, just to take one example, that it is rational to say such things as that a developed plot in a novel is more important than the way individual characters are drawn? In any case, even if we do think it rational to say such things, intransitivities will remain. They can only be eliminated by our thinking it in every case rational to say things like: plot is three times as important as characterisation, and this seems to me something we would say only if our aim was the desparate one of eliminating intransitives at all costs.