PREFERENCE STRENGTH AND TWO KINDS OF ORDINALISM*

ALLAN F. GIBBARD

In his paper "Extended Sympathy and the Possibility of Social Choice", Professor Arrow first considers an approach to individual utilities which is "strictly ordinal" and rules out interpersonal comparisons. This is the approach he took when he formulated his original impossibility theorem [1]. He then takes up an approach which allows interpersonal comparisons, but only of an ordinal kind. The strictly ordinal approach without interpersonal comparisons turns out to lead to a dead end, and the second approach – the one that allows interpersonal comparisons of an ordinal kind alone – leads either to lexical maximax or to lexical maximin. What are we to make of all this?

The approach without interpersonal comparisons leads to a dead end in the sense that it is inconsistent with certain other conditions: in the version of the Arrow theorem in the paper I am discussing, with the conditions of Binary Relevancy, Anonymity, and the Weak Pareto Principle. A further condition is built into the formalism Arrow uses: that the just or optimal alternatives from among the feasible alternatives are determined by a weak ordering which is independent of the feasibility of alternatives. The ordinal approach without interpersonal comparisons, then, is discredited by the Arrow theorem only in so far as these other conditions are plausible.

Just what, more precisely, is the approach that is supposed to lead to trouble? I can think of two positions the term 'a strictly ordinal approach' might designate; I shall call them 'Basic Ordinalism' and 'Pairwise Ordinalism'. One is this.

Basic Ordinalism: The only meaningful statements that can be made expressing a person's preference are those that can be derived from his preference ordering of the alternatives.

This seems to be the view that motivates the condition of Ordinal Invariance. According to Basic Ordinalism, any two individual utility scales that give the same ordering of the alternatives express the same information. Now a mere difference in the form in which information is expressed should not affect the conclusions that are drawn from that information. In particular, the form in which information about individual preferences is expressed should not affect conclusions about which states are optimal or just. The condition of Ordinal Invariance ought therefore to obtain.

An ordinalist might want to make a further claim.

> *Pairwise Ordinalism:* For any pair of alternatives x and y, the only meaningful statements that can be made expressing a person's preferences between x and y are those that can be derived from the statement that he prefers x to y, the statement that he is indifferent between x and y, or the statement that he prefers y to x.

This claim is not a logical consequence of Basic Ordinalism, and a proponent of Basic Ordinalism might have good reason for rejecting Pairwise Ordinalism.

What information, for instance, did Patrick Henry convey when he said "Give me liberty or give me death"? The datum he evinced was ordinal: He said in effect that he preferred death (D) to life under tyranny (T). Context made it clear that he preferred life with liberty (L) to death, and so we have his preference ordering LDT. Now on the basis of this ordinal information, in the absence of any indications to the contrary, it would be natural to conclude that Henry's preference for life with liberty over life under tyranny was extremely strong. That, indeed, must have been what Henry meant to convey. A proponent of Basic Ordinalism, then, might plausibly reject Pairwise Ordinalism; he will do so if he thinks that Patrick Henry's preference ordering indicates a strong preference for life with liberty over life under tyranny.

In one sense, a basic ordinalist rejects the possibility of interpersonal comparisons of utility: he thinks that once each person's preference ordering of the alternatives has been given, there is nothing further to be said about peoples' comparative strengths of preference or the comparative levels of satisfaction which a given alternative would bring them. In another sense, a basic ordinalist may think interpersonal comparisons possible. He may think that a person's entire preference ordering bears on the question of how

seriously a preference of his between a given pair of alternatives should be taken, and in particular, on how strong that preference should be regarded as being.

Now suppose I do think so. Then although I think that any two utility scales that are ordinally the same give the same information, I may think that some scales give their information more perspicuously than others: that on some scales, the difference between the utility numbers of two alternatives reflects the strength of preference which the ordering given by the scale indicates, whereas on other scales, differences in utility numbers have no such significance. More to the point, I may think that the interpersonal comparisons indicated by some pairs of utility scales are misleading. Call an assignment of a utility scale to each person a *utility profile;* a utility profile that assigns each person i a scale u, will be written u. We can talk of a utility profile as being *perspicuous* or *imperspicuous.* I shall not define these notions precisely, but the rough idea is this: a utility profile u is perspicuous if for any two people i and j and states of the world w, x, y, and z such that $u_i(w)>u_i(x)$ and $u_j(y)>u_j(z)$, the ratio of $u_i(w)-u_i(x)$ to $u_j(y)-u_j(z)$ is the ratio of the strength of i's preference for w over x, as indicated by the preference ordering given by scale u, to the strength of j's preference for y over z as indicated in a like way by scale u.

When I talk of the strength of a preference "as indicated by" a preference ordering, I have in mind two views a basic ordinalist might hold. One is that from a person's preference ordering of all conceivable alternatives, one can precisely determine the strength of any of his preferences between pairs of alternatives. A second possible basic ordinalist position is this. First, consider how a naive account of preference strength might go. On this naive account, peoples' preferences have definite strengths, and, as in the case of Patrick Henry, a person's preference ordering of all conceivable alternatives gives indications of the strengths of his various preferences. Indications of strength, though, are all that his ordering gives; one cannot use the ordering to deduce precise strengths of preference. We must therefore distinguish between a person's genuine strength of preference and his indicated strength of preference. Turn now to the basic ordinalist position I want to sketch. On this view, the naive notion of genuine strength of preference is not, strictly speaking, empirically meaningful. The notion of indicated strength of preference, on the other hand, has an empirical content, since claims about indicated strength of preference can be tested by observations that give a person's preference ordering. On this view, a

perspicuous utility profile is a profile such that ratios of utility differences, as given by that profile, are ratios of indicated strengths of preference – not ratios of genuine strengths of preference, for those ratios are empirically meaningless.

How, now, does the distinction between Basic Ordinalism and Pairwise Ordinalism bear on the Arrow theorem, and on the results by Hammond and Strasnick which Arrow reports? A pairwise ordinalist may well fall prey to the Arrow theorem. He denies that there is any distinction to be made between perspicuous and imperspicuous utility profiles. Suppose he also believes the following:

> *Determination by Pairwise Preferences:* Whether a state x is more just than a state y depends only on individuals' preferences between x and y.

Since he thinks that preferences are equally well displayed by any utility profile, he will accept *Binary Relevancy:*

> For any $\mathbf{u}$, $\mathbf{u}'$, x, and y, if $u_i(x) = u_i(x)$ and $u_i(y) = u_i(y)$ for all i, then $x\ f(\mathbf{u}')\ y$ if $x\ f(\mathbf{u})\ y$.

(Here and in what follows, variables w, x, y, and z will take alternatives as values; variables i, j, and k, the n people in question. I am interpreting $x\ f(\mathbf{u})\ y$ as saying that if each person's preferences were correctly given by the scale which utility profile $\mathbf{u}$ assigns him, then x would be more just than y.) Once he accepts Binary Relevancy, he is most of the way to the Arrow contradiction. As an ordinalist, he accepts Ordinal Invariance. Anonymity and the Weak Pareto Condition are innocuous (and in any case could be weakened considerably without losing the impossibility result. See, for instance, Hansson, [2: pp. 27-9.]). One further assumption is enough to yield the contradiction: that the relation "is more just than" is an ordering.

A proponent of Basic Ordinalism, on the other hand, may hold that some utility profiles are perspicuous whereas others are not. He may, on that account, accept Determination by Pairwise Preferences, but accept Binary Relevancy only with the restriction that the profiles in question, $\mathbf{u}$ and $\mathbf{u}'$, be perspicuous. For suppose profiles $\mathbf{u}$ and $\mathbf{u}'$, are imperspicuous, and that $u_i(x) > u_i(y)$. Even if $u_i'(x) = u_i(x)$ and $u_i'(y) = u_i(y)$, that in no way shows that the strength of preference for x over y which the entire scale u_i' indicates is the same as the strength of preference for x over y which the entire scale

u_i indicates. If a utility profile is imperspicuous, then by ignoring a person's rankings of alternatives other than x and y and looking only at the utility numbers of x and y as given by that profile, one loses information about the strength of preference for x over y which his entire ranking indicates.

Now if a basic ordinalist accepts Binary Relevancy only with a perspicuity rider, he can avoid impossibility results. That is not to say that any view whatsoever of what makes a utility profile perspicuous would, when formulated as a rider on the condition of Binary Relevancy, block impossibility results. It is to say that some perspicuity riders would do the job.

The easiest example is this: Suppose the alternatives are finitely many. Call a utility profile *perspicuous* if f for any i and x, $u_i(x)$ is the number of alternatives i ranks below x. A basic ordinalist might then accept the following principle of *Restricted Binary Relevancy:*

> For any x and y, and any two perspicuous utility profiles $\mathbf{u}$ and $\mathbf{u}'$, if $u_i'(x) = u_i(x)$ *and* $u_i'(y) = u_i(y)$ for all i then $xf(\mathbf{u}')y$ if f $xf(\mathbf{u})y$.

He can then consistently also accept the conditions of Ordinal Invariance, Anonymity, and Weak Pareto Principle, and he can accept that the relation *is more just than* is a weak ordering independent of the feasibility of alternatives. One social welfare function which satisfies all these conditions might be called *Utilitarianism:*

> For all $\mathbf{u}$, x, and y, x $f(\mathbf{u})$ y if Σ_i $u_i(x) > \Sigma_i$ $u_i(y)$ for the perspicuous utility profile $\mathbf{u}'$ which is *ordinally the same as* $\mathbf{u}$ – that is, such that for all i, w, and z, $u_i(w) > u_i(z)$ if f $u_i'(w) > u_i'(z)$.

With 'perspicuous' defined as it is here, Utilitarianism is equivalent to the Borda Rule: For each person, assign each alternative as many points as there are alternatives below it on his utility scale, add up the points each alternative gets, and order the alternatives by their point totals. I leave it to the reader to check that with 'perspicuous' defined as it has been, Utilitarianism as characterized here does satisfy the conditions I claim it to satisfy.

A slightly more plausible view that a minimal ordinalist might take is this. He might hold that what is just should be decided on the assumptions that each person's degrees of liking are distributed normally and that everyone has the same standard deviation for this distribution. He could then consider a utility profile perspicuous if,

on each person's utility scale, the utilities of the alternatives are distributed normally and all these distributions have the same standard deviation. With a suitable definition of the term 'distributed normally', this would presumably entail that if two perspicuous utility profiles **u** and **u**′ are ordinally the same, then **u**′ differs from **u** at most by everyone's scale being stretched or compressed by a uniform positive factor, and by individuals' scales being moved uniformly up or down. We will, in other words, have the following principle of *Comparative Interval Invariance:*

> For any two perspicuous utility profiles **u** and **u**′ that are ordinally the same, and there are real numbers $a>0$ and $b_1, \ldots, b_n$ such that for each person i and alternative x, $u_i'(x) = au_i(x)+b_i$.

Therefore, for any two alternatives x and y and any two perspicuous utility profiles **u** and **u**′ that are ordinally the same, we will have

$$\Sigma_i u_i'(x)>\Sigma_i u_i'(y) \text{ if f } \Sigma_i u_i(x)>\Sigma_i u_i(y).$$

This new definition of the term 'perspicuous' will give a new content to the definitions of Restricted Binary Relevancy and Utilitarianism which were formulated earlier. I leave it to the reader to check that Utilitarianism, in this new sense, satisfies Ordinal Invariance, Anonymity, the Weak Pareto Principle, and Restricted Binary Relevancy as now understood.

What I am saying in essence is this: One may consistently believe that interpersonal comparisons of strength of preference – or perhaps of indicated strength of preference – are possible, and still accept Basic Ordinalism. To do so, one must believe that the information needed for making these interpersonal comparisons is contained in the preference orderings of the people involved. A person who holds this view will reject Pairwise Ordinalism. He may accept Determination by Pairwise Preferences, but he will not on that account accept the principle of Binary Relevancy unless it is restricted to "perspicuous" utility profiles. He can then consistently accept Ordinal Invariance, Anonymity, the Weak Pareto Principle, and the view that the relation *is more just than* is an ordering (with ties allowed), and still take a utilitarian position. I have sketched two ways of doing all this, and considerations like those in the Patrick Henry example may suggest more reasonable ways of gleaning indications of comparative strength of preference from ordinal information.

Suppose now a basic ordinalist of the kind I have pictured is persuaded to relax his basic Ordinalism, and accept interpersonal comparisons of utility levels at the outset. Suppose, in other words, he now thinks possible interpersonal comparisons of utility levels that are not derived from the preference orderings of the people involved. That will not drive him to accept a maximax or maximin position. His conversion will amount to a liberalization of the constraints he sets on a social welfare function: in place of Ordinal Invariance, he will accept the weaker condition of Coordinal Invariance. That leaves open to him all the social welfare functions he before found acceptable, and so in particular, if he previously he before found acceptable, and so in particular, if he previously took the sort of utilitarian position that I have sketched, he can continue to do so.

* * *

I have not so far been evaluating Basic Ordinalism; I have only been sketching various positions a basic ordinalist might take. How well-founded is Basic Ordinalism?

One possible objection to it will not work. It is clear, an objector might say, that we do make interpersonal comparisons of strength of preference, and that we sometimes do so with good justification. Any theory that says that we cannot make such comparisons is at odds with a fact more evident than the theory itself. Now it should be clear from what I have said that the basic ordinalist can accept most of this. We do sometimes justifiably make interpersonal comparisons of strength of preferences, he can agree, or at least of indicated strength of preference. When we do so, though, the basis of our judgment still lies in the preference orderings of the people in question. When, for instance, we ascribe to Patrick Henry an extremely strong preference, for life under liberty to life under tyranny, we do so on the basis of his preference for death over life under tyranny. The issue between a basic ordinalist and his opponents is not whether we can make justified interpersonal comparisons of preference strength, but whether such comparisons give information that is not derivable from the preference orderings of the people concerned.

The appeal of Basic Ordinalism is epistemological. One way the argument for it might be put is this. The theory that ascribes a utility scale to a person (and hence allows comparisons of strength

of preference and the theory that utility scales are observationally equivalent. For what is explained by ascribing a utility scale to a person is the choices he is disposed to make and those same dispositions to choose can be equally well explained simply by ascribing a preference ordering to him. Now if two theories are observationally equivalent, the more austere one is to be preferred, and statements in the language of the less austere theory are meaningful only in so far as what they convey can be expressed in the language of the more austere theory. A utility scale, then, is meaningful only in so far as what it conveys can be expressed by a preference ordering; thus two utility scales that give the same ordering convey the same meaning.

How good is this argument? I can give almost nothing in the way of an answer here; I shall only touch on some ways the argument might be appraised. In the first place, I think that the vague theory of meaningfulness the argument invokes – or something like it – is correct. It is sometimes legitimate to reject a common sense concept as empirically meaningless; the concept of absolute distant simultaneity in physics is a prime example. When such conceptual pruning is legitimate, I think, it is because the concept rejected has no explanatory power: because we believe that everything observable can be at least as well explained without the concept as with it. The test of whether ordinally like utility scales convey the same information, then, is one of explanatory power. The question is whether everything that can be explained with utility scales can be equally well explained ordinally.

One way of answering this question will not help us. Choices among gambles, it might be thought, are better explained by expected utility maximization than they can be by means of any purely ordinal theory. Since two utility scales that are ordinally alike will sometimes, on tthe hypothesis of expected utility maximization, yield different choices among gambles, the difference between two such scales may be empirically significant. Even should all this be so, though, it will not dissolve the Arrow paradox. Choices among gambles will not distinguish between scales derived from each other by a positive affine transformation (that is, by a uniform expansion or contraction, a uniform raising or lowering, or a combination of the two). If such scales are not distinguished, the Arrow paradox remains. (See [4: pp. 129-30]).

What else, then, might be better explained by utility scales showing strength of preference than by mere preference orderings? People, it seems clear, evince preferences in ways other than by

making choices: on receiving news, for instance, they often show joy or disappointment. The intensity of these involuntary reactions to news might best be explained by the strengths of certain preferences: strong joy, for instance, might best be explained by the subject's having a strong preference for the news he has just received over what he previously thought true. If so, different utility scales that were ordinally alike might differ in their power to explain observations.

Another kind of observation that might better be explained by utility scales than by mere preference orderings is suggested by Harsanyi [3]. Take a characteristic C which a person can acquire and lose, and observe the choices of a person familiar both with having it and with lacking it. Suppose that, with other personal characteristics held fixed in some way B, a person with characteristic C prefers state w to state x, and a person lacking C prefers state y to state z. Take someone familiar both with having B and C in states w and x and with having B and lacking C in states y and z. Tell him he will have characteristics B, and that he will have an even chance of having C or lacking it. Then offer him a choice between the following two gambles: (1) State w if he has C, state z if he lacks C. (2) State x if he has C, state y if he lacks C. If he prefers (1), we can explain his choice by saying that people maximize expected preference satisfaction, and that among people with other characteristics B, the preference of a person with characteristic C for w over x is stronger than the preference of a person without C for y over z. This comparison of preference strength goes beyond what can be said merely with individual preference orderings.

Note that interpersonal comparisons of levels of utility might be explanatory in a similar way: To find out whether, other characteristics held fixed in some way, it is better to be in state x with characteristic C or in state y without C, offer someone familiar with both situations the choice. If he prefers x with C that can be explained by saying, other characteristics held fixed in the way in question, a person in state x with C is better off than a person in y without C. It is presumably tests of this kind that would give empirical significance to the theory Arrow gives in Section 7 of his paper.

There are, I think, three main questions to ask about the experiments proposed by Harsanyi. In the first place, would the experiments give consistent results? Would the results of the experiments, in other words, all be explicable by the hypothesis that subjects maximize expected utility on a single cardinal scale, which

gives level of utility as a function of a person's characteristics and external situation? In particular, will a person who has a characteristic C but knows what it would be like to lack it choose the same gambles as an otherwise similar person who lacks C but knows what it would be like to have it? In the second place, if the results of Harsanyi experiments indeed can be explained as expected utility maximization on an interpersonally applicable scale, is that the best explanation of the results? In the third place, if that is the best explanation, does the utility scale that figures in the explanation have ethical significance?

These are not questions I know how to answer. What I have tried to do is to take Harsanyi's challenge to Basic Ordinalism and ask the right questions about it. The appeal of Basic Ordinalism is, as I have said, epistemological. I have tried to put the argument for Basic Ordinalism in terms of explanatory power, or more specifically, in terms of the relation *better explains*. The main epistemological questions about a program such as Harsanyi's are, if I am right, questions about explanation, and good answers to those questions would require a good theory of explanation.

UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48109
U.S.A.

* Work on this paper was supported by a grant from the National Endowment for the Humanities.

REFERENCES

[1] Arrow, Kenneth, *Social Choice and Individual Values*, 1963.
[2] Hansson, Bengt, "The Independence Condition and the Theory of Social Choice", *Theory and Decision* 4: 25-49 (1973).
[3] Harsanyi, John C., "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility", *Journal of Political Economy* 63: 309-21 (1955).
[4] Sen, Amartya K., *Collective Choice and Social Welfare*, 1970.