TAL SCRIVEN

PLATO'S "DEMOCRATIC MAN" AND THE IMPLAUSIBILITY OF PREFERENCE UTILITARIANISM

For some time J. C. Harsanyi has defended a theory called "preference utilitarianism". He poses his theory, against the classical hedonistic utilitarianism of Bentham and the ideal utilitarianism of Moore which, he claims, face the following objections:

> The hedonistic definition was based on a now completely obsolete hedonistic psychology, which assumed that human actions were always motivated by seeking pleasure and avoiding pain, as if people could not be motivated by a desire for money, social status, success, knowledge, or by a genuine concern for the interests of other people – regardless of the possible pleasures they may or may not expect to derive from attainment of their objectives. On the other hand, Moore's ideal utilitarianism assumed that "mental states of intrinsic worth" differed from other mental states in having some special "nonnatural qualities" – a metaphysical theory most of us find hard to accept (and would find even harder to support by credible arguments even if we were willing to accept it).[1]

Preference utilitarianism avoids this problem by allowing agents to determine their own "fundamental values" which may be radically different than the fundamental values of Bentham and Moore. What Harsanyi insists upon is "the familiar *principle of consumers' sovereignty*" which holds that the "interests of each individual must be defined fundamentally in terms of his *own* personal preferences and not in terms of what somebody else thinks is 'good for him'".[2] Correction of preferences which are based on factual misinformation or on miscalculation is allowable, according to Harsanyi, but for an individual *i* to *censor* the preferences of *j* "because *j*'s preferences conflict with *i's own fundamental value* judgements" and *i* "could not satisfy *j*'s preferences 'with good conscience'" is wrong.[3]

Two problems surface immediately for this theory. First of all, how far may we go in allowing various "fundamental values" to count as legitimate values for rational people to pursue? Do we allow malice or sadism to be legitimate fundamental values to count as equals with benevolence

or knowledge for its own sake? Harsanyi's mechanism for eliminating malice, sadism and the like is profoundly unsatisfactory. He simply claims that where preferences "are based on clearly antisocial attitudes, e.g., on sheer hostility, malice, envy and sadism" the preferences can be "censored".[4] The problem with this line is pretty transparent. It simply begs all the central questions of ethics. If we can determine in advance which preferences are "social" and which are "antisocial" what do we need preference utilitarianism or any other ethical theory for? We would know what the theory was going to prescribe before we applied it.

The second immediate problem with preference utilitarianism, however, is the concern of this paper. It is a problem anticipated in Book VIII of Plato's *Republic*. Having seen the lack of any clear vision of the Good as a central characteristic of democracy, he offers the following impression of the "democratic man":

> He does not welcome true reasoning or allow it into the guardhouse; if someone tells him that some pleasures belong to good and beautiful desires, but others belong to evil ones, that one should prize and pursue the former while the latter must be restrained and mastered, he denies all this and declares that all pleasures are equal and must be equally prized...
>
> And he lives on, yielding day by day to the desire at hand. At one time he drinks heavily to the accompaniment of the flute, at another he drinks only water and is wasting away; at one time he goes in for physical exercise, then again he does nothing and cares for nothing; at times he pretends to spend his time on philosophy; often he takes part in public affairs; he then leaps up from his seat and says and does whatever comes into his mind; if he happens to admire military men, he is carried in that direction, if moneyed men, he turns to making money; there is no plan or discipline in his life but he calls it pleasant, free, and blessed, and he follows it throughout his time.[5]

This impression of the democratic man brings us to a troublesome question about the preference utilitarian. Given even that certain fundamental values could satisfactorily be eliminated as "antisocial", many legitimate fundamental values would remain and one's preferences over specific alternatives at specific times could vary radically as one's fundamental values change over time. The question is, how often can one change one's mind about fundamental values and still remain rational in any straightforward sense? To require that any rational person never change his mind about fundamental values is, pretty clearly, to require far too much. At the other extreme, to allow one to change fundamental values every ten minutes is, pretty clearly, to require far too little of a rational individual. But at what point in between does it make any sense

at all to mark out a minimum and a maximum number of times a rational person could change his mind over a given amount of time? What kind of "discipline", as Plato, as well as A. K. Sen,[6] put it, can be required.

Traditional theories which uncompromisingly lay down some fixed vision of "The Good", be it Plato's or Bentham's vision, of course, don't have this problem. Under such theories a legitimate change of mind about a specific set of alternatives would entail that the individual either had previously had the wrong fundamental value or the wrong factual information or, possibly, that the factual circumstances of the decision had changed.

This problem points to a kind of decision problem that is quite separate from the problems which have been the central concern of decision theory for centuries. From Bayes on, the complications of formally specifying what it means for an individual decision made *at a specific time* to be rational have been examined. From Condorcet on, the complications of formally specifying what it means for a social decision made on the basis of rational individual preferences to be, itself, rational have been examined. But the problem of what it means, formally, for preference orderings collected over time from one individual to be *rational over time* given that each ordering was rational *at a time* has not been subjected to this kind of scrutiny.

The pessimistic conclusion argued for below is that the search for a formal set of constraints on rationality over time is hopeless for the preference utilitarian. Unfortunately for preference utilitarianism, the problem of specifying rationality-over-time constraints for preference orderings given by one individual at various times is open to a modification of the argument used by Kenneth Arrow to show the impossibility of specifying social rationality constraints for preference orderings given by various individuals at one time. What is worse is that some of the major problems with Arrows's Impossibility Theorem do not arise for the modification of his argument to be given here.

Four basic elements are involved in rationality-over-time for the preference utilitarian. First, there are the various specific preference orderings that an individual would have at various times. Second, there will be the

relevant factual information which constrains the orderings. Third, also constraining the orderings will be whatever fundamental values the individual was trying to maximize at the various times. Finally, theories like preference utilitarianism must specify some set of normative constraints either on the orderings or directly on the fundamental values underlying the orderings which yield some intuitive boundaries for rationality-over-time.[7] There is, of course, no point in looking for normative constraints on facts. Facts are the way they are whether they ought to be that way or not. For this reason, rationality-over-time is just not a problem for theories like Bentham's because the only variable impinging on orderings over time is factual information which is not subject to normative constraint.

As already indicated, coming up with a serious set of normative constraints for rationality-over-time is not, even on the surface, going to be easy. Thus, pointing to the fact that none have been established wouldn't be very interesting. My intention, however, is to prove that none *can* be established which would be compatible with preference utilitarianism. The strategy in proving this will parallel Arrow's strategy in proving the impossibility of social decision rationality constraints. A set of very minimal constraints which the preference utilitarian would have to accept as the barest necessary conditions of rationality-over-time will be shown to be jointly inconsistent.

Some rephrasing of the standard notation of decision theory is necessary before proceeding to the proof. Numeral subscripts will represents times in the individual's life with 1 being the earliest time slice in his life, n being the latest and i, j and k being variables. The lower case letters x, y, z and w denote specific alternatives. P denotes the strong preference relation and so xP_1y means the individual preferred x to y at time 1. R denotes the weak preference relation and so xR_1y means the individual preferred x at least as much as y at time 1. When xPy appears without a subscript it means the individual must (or is rationally required to) prefer x to y and, likewise, xRy means the individual must prefer x at least as much as y. The relation P is definable in terms of the relation R as follows: xPy if and only if xRy and $y\bar{R}x$ (i.e. not yRx). The relations R, P,, $\bar{R}$ and $\bar{P}$ are transitive, R and $\bar{P}$ are reflexive, and P and $\bar{R}$ are irreflexive. R_1, R_2 and so on, when standing alone, will denote preference orderings of the individual at times 1, 2, and so on. $C(S)$ denotes the

individual's choice set from a set S of alternatives, that is, the set of alternatives in S that the individual prefers at least as much as any other alternatives in S. $C(s,R)$ denotes the individual's choice set from S given his preferences over time $(R_i,\ldots,R_j)$ and $C(S,R')$ denotes his choice set from S given his preferences over time $(R'_i,\ldots,R'_j)$.

Major assumptions upon which proof depends are as follows:

CONDITION U (*Unrestricted Domain*): No logically possible configurations of individual preference orderings given over time can be rejected unless they are a result of factual misinformation or of miscalculation.[8]

CONDITION P (*Temporal Pareto Principle*): For any pair of alternatives x and y, if for every time i, xP_iy then xPy.

CONDITION I (*Independence of Irrelevant Alternatives*): Let R and R' be the binary relations determined by a rational individual's preferences corresponding, respectively, to two sets of the individual's preference orderings over time, $(R_1,\ldots,R_j)$ and $(R'_i,\ldots,R'_j)$ for a set S of alternatives. If for all pairs of alternatives x and y in a subset S' of S and any time k, xR_ky if and only if xR'_ky then $C(S',R)$ and $C(S',R')$ must be the same.

CONDITION D (*Non-decisiveness of Single Orderings*): There is no single preference ordering at any time i such that for every logically possible combination of preference orderings over time xP_iy implies xPy.

This statement of the major assumptions involved in the proof of the impossibility of preference utilitarianism makes it clear yet again why theories like Bentham's will not face a similar problem. The classical utilitarian, or anyone else whose theory imposes a fixed fundamental value, would clearly deny conditions U, P and D. U would be rejected since logical incoherence, factual misinformation and miscalculation are not the only reasons for rejecting preference orderings. They could also be rejected for being based on the wrong fundamental values. P would be rejected for the same reason. The fact that an individual has always preferred x to y doesn't mean he was ever right even if he was never factually misinformed and he never miscalculated. Similarly, D would be rejected by Bentham on the grounds that some single preference ordering

R_i *should* be decisive if it was only at time i that the individual held the correct hedonistic fundamental value.

However, given the background assumption that an individual's preference orderings over time are individually logically coherent and not based on misinformation or miscalculation, none of the above conditions can be rejected by the preference utilitarian. In fact, much more would be required in order for the preference utilitarian to give a reasonable account of individual rationality. Unfortunately, these minimal constraints are enough to show that the theory can't possibly give even a minimally adequate account of individual rationality over time.

DEFINITION. $D_V(x,y)$ if and only if xP_Vy implies xPy and yP_Vx implies yPx, where $D_V(x,y)$ means that the set V of preference orderings over time is decisive in determining whether the individual must (or is rationally required to) prefer x to y or y to x and xP_Vy means that x is ranked over y in all of the orderings in V.

THEOREM. *Conditions U, P, I and D are jointly inconsistent given transitivity, irreflexivity and the assumption that the individual's preference orderings over time are individually non-contradictory.*

Proof.[9] (1) From Condition P it follows that for any pair of alternatives x and y there exists a set of personal preference orderings given over time which is decisive for $\{x,y\}$ (namely, the set of all of the individual's orderings over time). Thus, for any pair of alternatives there is a minimal decisive subset of this set of all orderings which can be constructed by removing orderings one at a time until we reach a set which will not be decisive if any one more of its orderings is removed. Let V be such a minimal decisive set for $\{x,y\}$, R_i be a member of V and $\overline{V}$ be the set of all orderings not in V. In addition, suppose that no proper subset of V is decisive for any pair of alternatives.

(2) Let z be a third alternative and the orderings for R_i, $V-R_i$ and $\overline{V}$ be as follows:

R_i	$V-R_i$	$\overline{V}$
x	z	y
y	x	z
z	y	x

(3) By definition, since $D_v(x,y)$ and xP_Vy, xPy.

(4) Since $V-R_i$ is not decisive for any pair of alternatives and only orderings in $V-R_i$ have z ranked over y, $z\bar{P}y$.

(5) By transitivity and irreflexivity, xPy and $z\bar{P}y$ imply xPz.

(6) But R_i is the only ordering which has x ranked over z, so $DR_i(x,z)$.

(7) Since no proper subset of V is decisive for any pair of alternatives and $DR_i(x,z)$, it must be that $V=R_i$ and thus, by hypothesis, $DR_i(x,y)$.

(8) Having shown $DR_i(x,y)$ and $DR_i(x,z)$ it can easily be shown that (a) $DR_i(w,z)$ for any $w\neq z$ and (b) $DR_i(w,x)$ for any $w\neq x$.

(a) Suppose

R_i	$\bar{R}_i$
w	z
x	w
z	x

By condition P, wPx and since $DR_i(x,z)$ and xP_iz, xPz. By transitivity, wPz. So, $DR_i(w,z)$.

(b) Suppose

R_i	$\bar{R}_i$
w	z
z	x
x	w

By condition P, zPx and since $DR_i(w,z)$ and wP_iz, wPz. By transitivity, wPx. So $DR_i(w,x)$. Thus, some single preference ordering is decisive over every pair of alternatives. Hence, conditions U, P, I and D are jointly inconsistent given transitivity, irreflexivity and the assumption that the individual's preference orderings over time are individually non-contradictory and not based on factual misinformation or miscalculation.

One question arises immediately concerning the above proof and it centers on the use of Condition D. This condition rules out the possibility of an individual's last preference ordering, R_n, being decisive since it cannot be stipulated that $i\neq n$. But why should a preference utilitarian accept this restriction? Why can't we just take an individual's last preference ordering as his real preference ordering? We can't do this because it would lead us directly to the problem of the "Democratic

Man''. Nothing would prevent an individual from rearranging his fundamental values every fifteen seconds since xP_ny would always entail xPy as n moves through time. On the other hand if $i \neq n$ then no change of mind is allowed at all. The individual must always hold R_i and, thus, the theory simply contradicts its own central claim that there are many legitimate values that individuals may pursue. So, either preference utilitarianism allows for a radical change of mind every millisecond or it allows for no change of mind at all. There is nothing in between.

Although the basic logic of the above argument is the same as Arrow's, some of the standard objections against Arrow's Theorem will not apply to my argument. One objection made by J. M. Buchanan [10] and also by Kurt Baier concerns to the whole business of applying individual rationality constraints like transitivity to social decisions. Baier claims that the failure of transitivity of the social preference relation

> ... would seem to be no more surprising or paradoxical than the fact that an equal division of voters on some issue should show itself in a "contradictory social decision". In such cases there is then no genuine (transitive) social will or preference, however genuine (transitive) the individual wills or preferences may be.[11]

What makes the proposal that transitivity be dropped as a constraint on the social preference relation especially intriguing is A. K. Sen's work on social decision. In his *Collective Choice and Social Welfare* [12] he devotes considerable attention to the examination of social decision procedures which yield intransitivities but, nevertheless, manage to map any logically possible combination of individual orderings onto a social choice set without contradiction and within Arrow's four social rationality constraints.

However important all of this may be concerning Arrow's Theorem none of it causes any trouble for the argument above. Nowhere does my argument involve any concept of *social* rationality. The problem I'm concerned with is the relationship between *individual* rationality at any given time and *individual* rationality over time, and it seems pretty safe

to me to assume that transitivity is a legitimate rationality constraint in either case.

Arrow's formulation of the condition on social decisions which corresponds to the above Condition I has probably raised more objections than any other part of his proof. I have argued myself[13] that, given the strategic factors involved when various individuals come together to work out social decisions, a basic intuition behind the independence of irrelevant alternatives condition used by Arrow is dubious. The centrality of strategic considerations and their relationship to Arrow's Condition *I* is made formally clear by what has come to be known as The Gibbard-Satterthwaite Theorem.[14] What Gibbard and Satterthwaite have shown is that Arrow's Condition *I* can be replaced by a condition that social decision mechanisms be non-manipulable or "strategy-proof" and the same pessimistic theorem follows.

The role of strategic manipulation in social decisions is easy to see. Suppose we have three people *A*, *B* and *C* whose preferences over three alternatives *x*, *y* and *z* are representable as follows:

A	*B*	*C*
x	*z*	*y*
y	*x*	*z*
z	*y*	*x*

Suppose also that the social decision will be made by having two majoritarian elections, the first one between *x* and *y* and the second between the winner of the first election and *z*. If each person honestly expresses his preferences then *x* will beat *y* and *z* will beat *x* and so *z* will be the winner. If *A* sees this coming and realizes there is no way his first choice, *x*, is going to be chosen he will see that by misrepresenting his preferences in the first election he can at least get his second choice, *y*, chosen rather than his last choice, *z*. So he votes for *y* in the first election, it beats *x* and in the second election it beats *z* and is thus the social choice. If *B* and *C* are as smart as *A* they will start strategically manipulating their preferences also and the situation will be bogged down in preference misrepresentation.

Arrow always realized that strategic factors could not be properly taken into account by the method involved in his proof.[15] Of course, the problem for Arrow's proof is the adequacy of any approach to social

decision theory which methodologically rules out such factors. This problem is overcome by Gibbard and Satterthwaite inasmuch as they rule out strategic manipulation at the normative level rather than at the methodological one. Manipulability is directly stated as an undesirable property of social decision mechanisms and non-manipulability is shown to be inconsistent with other intuitive conditions on social decisions. At any rate, reservations about the Condition *I* used in my proof which are based on strategic considerations don't carry the same weight they do against Arrow's Condition *I*. Again, the subject of my theorem is individual rationality, not social rationality. What sense, then, can be made of the possibility of engaging in strategic manipulation of one's own preferences over time against each other? What sense is there in misrepresenting to yourself your own preference ordering on Monday so that it prevails over your own preferences on Tuesday and Wednesday?

Preference utilitarianism has become one of the most widely discussed variants of utilitarianism in recent years. It has, in fact, found its way onto the standard list of different utilitarian theories offered in standard undergraduate ethics textbooks.[16] There is good reason for this. Harsanyi and others have made the theory painstakingly rigorous. It has blended smoothly into a vision Harsanyi shares with many theorists of a "general theory of rational behavior" that would encompass ethics, social decision theory and game theory. The principle of consumer sovereignty mixes well with the general relativism of the day while still leashing ethics to formal constraints that don't allow just anything to pass as moral. This principle also mixes well with the political liberalism with which most utilitarians have chosen to associate themselves. The value of self-determination of ends at the individual and social level certainly ranks high for Mill. The theory also has interesting and distinct consequences when applied to practical moral problems.[17]

But there has always been reason to be suspicious as well. When applied to certain practical problems where its consequences are clearly distinct from those of classical hedonistic utilitarianism it seems to say the wrong thing.[18] Furthermore, it is not at all clear that the hedonism of Bentham

and Mill implies anything like the paternalism Harsanyi associates with it. Mill, after all, went quite out of his way in *On Liberty* to develop classical utilitarian arguments directly against paternalism. As noted earlier on in this paper, it is not at all clear that preference utilitarianism really can avoid an unbridled relativism without a lot of question begging along the way.

Finally, we have the problem Plato saw with theories that fail to enforce some clear vision of the Good upon individuals and societies. One might argue that this problem is too theoretical to be of serious consequence even in decision theory. The fact of the matter, it might be asserted, is that, by and large, individuals' values just aren't as chaotic as Plato's anti-democratic rhetoric has it. If they were, the suggestion that social decisions be based on them would never have been taken seriously in the first place. On the face of it, however, there is something wrong with this objection. Would it lessen the importance of Arrow's Theorem if it could be shown that, as a matter of fact, voting cycles don't arise very often in the real world? After all, if such oddities did occur regularly, the suggestion that social decisions should be based on amalgamations of individual preferences would never have been taken seriously in the first place!

In fact, theoretical problems like the one presented here are important when weighing theory against theory, especially when the problems arise for one theory (in this case, preference utilitarianism) and not for others (like classical hedonistic utilitarianism). And it is the contention of this paper that the problem of Plato's "Democratic Man" is not only serious but unsolvable.

NOTES

* I would like to thank Stephen Ball and Charles Hagen for helpful discussion of the topics in this paper.

[1] J. Harsanyi, 'Rule Utilitarianism and Decision Theory', in *Decision Theory and Social Ethics*, eds. H. Gottinger and W. Leinfellner (Boston: D. Reidel, 1978), p. 5. Harsanyi, of course, is not the only defender of preference utilitarianism nor is his specific version of the theory the only one available. A good analysis of various theories that share the fundamental assumptions of preference utilitarianism is given by A. K. Sen in 'Well-being, Agency and Freedom: The Dewey Lectures 1984', The *Journal of Philosophy* (April 1985):

169–221. It does not seem to me that multiplying variations of the theory will help at all against the result to be argued for in this paper.

[2] J. Harsanyi, *Rational Behavior and Bargaining Equilibrium in Games and Social Situations* (New York: Cambridge, 1977), p. 52.

[3] *Ibid.*, p. 62.

[4] *Ibid.*

[5] Plato, *The Republic*, trans. G. Grube (Indianapolis: Hackett, 1974), pp. 209–210 (561 c-d).

[6] A. K. Sen, *op. cit.*, p. 204.

[7] It is not all clear what the difference between constraints on specific preference orderings and constraints on basic values could amount to given that the distinction between fundamental and non-fundamental values is not at all clear. If some individual preferred drinking whiskey to reading Plato, common sense would suggest that the desire for whiskey was a function of some more fundamental value (maybe physical pleasure). But if the individual, when challenged, replied that drinking whiskey had its own intrinsic value for him, it is not at all clear how the preference utilitarian could argue against him. What this entails, in turn, is that it is not at all clear that individuals can even be required to have anything that corresponds intuitively to a set of fundamental values.

[8] It is important to note that the assumption of correct information is of an entirely different order than the other assumptions of the proof. These other assumptions are purely formal rationality constraints whereas the requirement of correct information is neither a formal one nor is it one that can be sensibly required as a rationality constraint, for if it could then hardly anyone would be considered rational. The presumption of total enlightenment is simply entered here, as it is, for example, in Arrow's Theorem, with the intent of specifying that even under conditions of total enlightenment things can't possibly go right. Similar considerations would apply if some non-question-begging mechanism for eliminating "anti-social" preferences could be designed. The assumption that all preferences are "social" would be entered as a background assumption along with the assumption of enlightenment.

[9] The proof to be given is essentially a modification of the version of the proof of Arrow's Theorem given by R. Luce and H. Raiffa in *Games and Decisions* (New York: John Wiley & Sons, 1957), pp. 339–340.

[10] J. Buchanan, 'Individual Choices in Voting and the Market', *Journal of Political Economy* **62** (1954): 334–343.

[11] K. Baier, 'Welfare and Preference', in *Human Values and Economic Policy*, ed. S. Hook (New York: NYU, 1967), p. 121.

[12] A. Sen, *Collective Choice and Social Welfare* (San Francisco: Holden-Day, 1970).

[13] 'Preference, Rational Choice and Arrow's Theorem', *Journal of Philosophy* **LXXVI** (December 1981), 778–781.

[14] A. Gibbard, 'Manipulation of Voting Schemes: A General Result', *Econometrica* **41** (1973): 587–601. M. Satterthwaite, 'Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions', *Journal of Economic Theory* **10** (1975): 187–217.

[15] K. Arrow, *Social Choice and Individual Values*, 2nd ed. (New Haven: Yale, 1963), pp. 20–21.

[16] See, for example, T. Beauchamp's *Philosophical Ethics* (New York: McGraw-Hill, 1982), pp. 84–85.

[17] See my 'Utility, Autonomy and Drug Regulation', *International Journal of Applied Philosophy* **2** (Fall 1984): 27–42.
[18] See *ibid.*

Philosophy Department,
California Polytechnic State University,
San Luis Obispo, CA 93401,
U.S.A.