FRANZ VON KUTSCHERA

SUBJECTIVE PREFERENCES, RATIONALITY, AND JUSTICE

ABSTRACT. Three main types of subjectivist ethics are distinguished and specified by the use of elementary game-theoretical notions. It will be argued that all these theories run into difficulties that cannot be overcome within the self-imposed limits of subjectivism.

1. TYPES OF SUBJECTIVE ETHICS

It is the fundamental thesis of subjectivism in ethics that moral concepts are definable on the basis of subjective preferences. This basis is strictly empirical. The individual preferences may be determined by observations and experiments in principle. They are taken over into the definition of moral concepts as they are given and in their totality. No moral yardstick is applied to them; no attempt is made to cancel out morally irrelevant or illegitimate interests. Subjectivism, then, is a naturalistic theory.

With respect to their ideas about the relation between individual preferences and the moral preference ordering four main types of subjectivism may be distinguished:

1.1. *Individualistic subjectivism* holds that every statement about values contains a (hidden) parameter referring to a specific individual. If we say, something is good, what we mean in fact is that it is good for somebody, for the speaker, e.g. 'Moral' preferences, then, are relative to persons and coincide with their individual preferences, so that there really are no moral values beyond subjective values for individuals. Morality in this way is relativized and there is no attempt to construct a transsubjective preference ordering.

1.2. *Rationalistic subjectivism* maintains that what is morally good coincides with what is good for everybody in the long run. The maximes of morality, therefore, are nothing else than the maximes of rationality, and these, properly conceived, are the same for all individuals. He acts morally good, who lets himself be governed by "l'amour éclairé de

Erkenntnis **11** (1977) 97–111.

nous-mêmes", as d'Alembert put it. Here again morality coincides with self-interest, but the difference against individual subjectivism is, that the self-interests are conceived as being identical in the long run.

1.3. *Altruistic subjectivism* takes the view that individual preferences are aggregates of an egoistic and an altruistic component, where the components may have different weights for different persons. The first derives from our self-interest, the second from our benevolence toward others, from a concern for their self-interests. The altruistic component, therefore, is conceived as an aggregation of the individual self-interests. The moral preference ordering is identical with the benevolent component in the individual preferences, which is the same for all persons and therefore coincides, not wholly as in rationalistic subjectivism, but at least partly with the individual orderings.

1.4. *Social subjectivism*, finally, determines the moral preference ordering by aggregating individual interests, without claiming, however, that this ordering coincides fully or partly with the individual orderings.

Of these four types individualistic subjectivism is the least interesting one. It is closely related to emotivism, and only distinguished from it by treating moral statements as descriptive ones, i.e. as being true or false. It is clearly inadequate, since there is a substantial difference of meaning between the two types of statements 'This is *morally good*' and 'This is *good for me*'. They are not even extensionally equivalent. A lie may be good (advantageous) for me without being morally good, even in my own judgment. There are a number of other arguments against individualism, but we shall not go into them here but be concerned exclusively with the more interesting other types of subjectivism in the following.

The primary goal of these approaches is to develop a theory of *justice* as a theory of fair and reasonable compromise between conflicting individual interests in situations where all participants are interested in cooperating with the others, because the success of their actions depends on what the others do. Moral principles are not limited to the maximes of justice, of course, but justice, certainly, is the central social virtue, and social ethics again is the core of any moral philosophy. A justification of principles of justice, therefore, would be a decisive step towards establishing a useful moral theory.

2. RATIONALISM

The thesis of rationalism that an enlightened egoism is a sufficient basis for moral behavior seems, on first glance, scarcely more interesting than that of individualism. We are all acquainted only too well with situations in which the postulates of morality conflict with our self-interests. But rationalism is in fact one of the oldest ethical doctrines. One of the first moral maximes we find in the history of Greek thought is Chilon's advice: "ὅρα τέλος" ("Think of the consequences"), and the idea that the good is what is truly advantageous for man, is prominent throughout antiquity.[1] It was taken up with new emphasis by Spinoza and Hobbes[2] at the beginning of the enlightenment, which laid the foundations of modern subjectivism.

But let us turn from ancestry to arguments. The schema of rationalistic argumentation, bared from psychological and sociological embellishments, is this:

As social beings we depend on cooperation with our fellows in almost all domains of life. Cooperation yields a better life for all than the war of each against all. Therefore it is in the common interest to agree on rules or establish conventions for cooperative action in recurrent situations of specific types so that everybody, when such a situation arises, may rely on the others doing their part. The social roles of the people may change in the different situations, and nobody knows beforehand in what role he will find himself when such a situation arises, whether he will function as debtor or creditor, seller or buyer, giving or receiving orders etc. Hence it is rational for everybody to consent to rules that keep the possible disadvantage as small as possible, i.e. to rules that maximize the advantage of the most disadvantaged social roles. For it might be himself to whom one of the least attractive positions is assigned.

This principle for the choice of rules for cooperation, however, which is a postulate of individual *rationality*, is also a maxime of *justice*. It takes into account the interests of all concerned in an impartial way, and even in a strongly egalitarian manner.

The choice of such rules is not only rational in some Rawlesian 'original position', i.e. before the situations to which the rule applies actually come up, but also when the 'veil of ignorance' has dropped and when, in some specific situation, we find ourselves in one of the more fortunate social roles in which the rule demands some sacrifice from us. We know that

such situations will come up again, and since we may then be in the worst position we should not endanger our prospects by breaking the agreement, which lowers our risk in future contingencies.

This argument can be stated in game-theoretical terms and we shall do so now to make it, and what we have to say to it, more precise.

Let $I = \{1, \ldots, n\}$ $(n \geq 2)$ be a set of persons who in some situation can choose between acts from a set $F = \{f_1, \ldots, f_m\}$ $(M \geq 2)$[3]. $R = F^n$ is to be the set of possible results of actions of the $i \in I$ in S. On R n utility-functions u_i are to be defined such that $u_i(x)$ *for* $x \in R$ is the subjective, individual utility of result x for person i. We shall assume an interpersonal comparability of the utilities, so that the u_i are determined uniquely up to common positive linear transformations.

The game $G = \langle I, F, u \rangle$, where u is the *n-tuple* $(u_1, \ldots, u_n)$, is to be *cooperative*, i.e. the players can communicate with each other and form agreements as to which result is to be realized.

If the game is played only once a cooperative result should be in the *negotiation set* $N = P \cap \{x: \Lambda i(u_i^0 \leq u_i(x))\}$, where $P = \{x: \Lambda y(Vi(u_i(x) < u_i(y)) \supset Vk(u_k(y) < u_k(x)))\}$ is the set of Pareto-optimal results and $u_i^0 = \max_f \min_{x:(x)_i = f} u_i(x)$ is the *security level* of the game for player i. $(x)_i$ is to be the ith member of the n-tuple x.[4]

We need not try to narrow down the set of rational cooperative solutions by further criteria, which would be highly controversial anyhow. The sole reason for mentioning the above condition for such solutions is to point out, that rational compromises depend on the natural advantages, here: on the security levels, of the players.

This is also true if the game is played repeatedly and joint mixed strategies are employed. Then the security level u_i^* of i is what player i can guarantee himself by adopting a separate mixed strategy. But the situation changes completely if we assume that before each match starts it is decided by lot which role each player has to take. In this case we have to distinguish the set I, which now is to be the set of roles, and the set $P = \{P_1, \ldots, P_n\}$ of players. We assume that in role i every player has utilities expressed by u_i. This means that all players have utilities, which, on condition that they are assigned role i, are determined by the utility-function u_i for that role. The players then need not be without preferences before the play starts, as in Rawls' original position, but their preferences change with the roles they have to take.

If we assume now, that the players cannot assign probabilities to the events that they draw one or another of the roles, then the maximin-criterion for decision under uncertainty tells them:

(K1) Agree on a result x for which $\min u_i(x)$ is maximal.

This criterion applies both to single and repeated performances of the game. The important difference between these two cases arises only, if the lots have been drawn for some match and the roles distributed. Then a player P_i, for instance, may find that he is in a role k for which $u_k(x) < u_k^0$, where x is the result agreed upon in accordance with (K1). If the game is played only once then it might be profitable for P_i – even if possible sanctions by the others are considered – to break the agreement and adopt this security strategy which guarantees him at least u_k^0. But if the game is repeated afterwards sufficiently often, then his possible loss will outweigh his momentary gain $u_k^0 - u_k(x)$, since if in the future he always gets the worst part k' (for which we have $u_{k'}(x) > u_{k'}^0$ for all non-trivial games) his possible loss after n repetitions is $n \cdot (u_{k'}(x) - u_{k'}^0)$.

If probabilities p_{ik} of P_i getting part k are known ($\sum_k p_{ik} = 1$ for all i), then we have a decision under risk and (expected) natural advantages come in again, since each player can determine his future prospects if he acts alone or in coalitions. If each player, however, has the same chance of getting every part, i.e. if we have $p_{ik} = 1/n$ for all i and k, then the criterion of maximizing their expected utilities tells them:

(K2) Agree on a result x for which $\sum_i u_i(x)$ is maximal.

And here, as in (K1), natural advantages are irrelevant. The derivation of this *utilitarian principle of justice* from considerations of rationality, however, is based on the rather implausible assumption of equal known probabilities for all roles. Therefore we have referred only to (K1) in our intuitive argument above and will mainly refer to it in what follows. But, given the prerequisites of its derivation, (K2) is a true principle of justice also, since, even if somebody has to pay a big price for the greater well-being of others in some situation, he will find comfort in the knowledge that in the long run he will be equally well off as the others with this arrangement.

After stating the rationalistic argument in more formal terms let us now point out its difficulties. The most important objections against it may be summed up as follows:

1. *The Problem of Realism*

The rationalistic argument rests on assumptions about the type of social situations to which norms of justice apply. It is questionable whether these assumptions are realistic.

1a. *Uncertainty of the Social Roles.* There are many roles which alternate in the course of life for a person, and we have named some above. But there are also many roles that accompany a person throughout his life (sex, talents, training, e.g.) or roles, a change in which is improbable (profession, membership in social groups, e.g.). But this implies that for such roles the maximes of rationality (K1) (or (K2)) are superseded by criteria oriented to people's natural advantages, and rational behaviour does not anymore coincide with what justice prescribes.

1b. *Repetition of the Situations.* There are situations, in which we find ourselves only very seldomly and which are unlikely to repeat themselves. Take a situation where someone is drowning and you can save him only with considerable risk to your own life. Since the roles in this situation are known and the small chance that one day they will be reversed is even further diminished if you do what morality prescribes, the postulates of rationality will be markedly different from those of morality. Generally speaking, in the long run we will all be dead, and therefore criteria of rationality which presuppose that the same situations will be repeated again and again are not very realistic.

1c. *Identity of Conditional Preferences.* It is highly probable that there are basically different individual preference-structures; that different people often have different preferences even if they act in the same roles.[5] But if the conditional preferences are not identical, (K1) (or (K2)) are not anymore generally acceptable criteria for rational behavior.

2. *The Problem of Different Rationality Concepts*

It has often been pointed out that the maximin-criterion is not adequate in all situations of decision under uncertainty. If I have to choose between two acts f and g, e.g., in a situation where the outcomes depend on whether p or not-p obtains, and if my utilities for the outcomes are 0 for f and p, 100 for f and non-p, and 0.1 for both g and p, and g and non-p, then I certainly would not do g as advocated by the maximin-rule. This rule seems adequate only if great risks are involved, i.e. if the maximin strategy is much safer than other ones. Even if, for the sake of the argument, we concede that social choices generally involve great risks, it still should be pointed out, that there is no purely 'rational' criterion for preferring pessimistic maximin procedures, e.g., against an optimistic maximax strategy. Rather it is the other way round: such criteria define different concepts of rationality. So if rationalism speaks of justice coinciding with rationality it has to face the question: 'Rationality in what sense?' Different notions, or principles of rationality determine different notions and principles of justice in the rationalistic argument, and these are not any better justified than the former.

3. *The Problem of Adequacy*

3a. *Morally Legitimate Interests.* (K1) and (K2) as principles of justice are adequate only, if they are not based just on any subjective preferences, but on *morally legitimate* ones. First, a theory of justice has to determine the range of application for its principles. (K1) certainly is not to be applied to parlour games or to all business transactions, e.g., but only to cases where morally legitimate interests of people are involved. Second, sadistic interests or excessive greeds should not be honored by (K1). So to exclude applications of (K1) that yield strongly counter-intuitive results, we would have to base (K1) on a distinction of morally legitimate and morally illegitimate interests. This, however, would not only be contrary to the intentions of subjectivism, but it is simply impossible within the framework of this theory. To justify moral criteria by reference to individual preferences, and then say that some of the preferences are unjustified by these criteria would be grossly circular.[6]

Furthermore, while principles of justice call for a restriction to morally legitimate interests, considerations of rationality exclude it. For a sadist it is not moral, but certainly rational to insist on his perverse interests to be honoured by society. So even if the distinction could be drawn on subjectivist grounds, it would, by marking the point where the principles of justice deviate from those of rationality, still show the untenability of rationalism.

3b. *Inclusiveness.* If the group P of players is taken to be humanity as a whole, including present and future generations, then the presuppositions of the rationalistic argument (unknown or equal chances for all social roles, equal preferences in equal roles, cooperative action) become extremely unrealistic. But if, on the other hand, we want to make them more realistic by taking P to be a relatively small and homogenous group, cohering by social interaction, then the postulates (K1) or (K2) lose their moral character since they disregard the interests of outsiders. Thus there is a conflict between realism and adequacy which cannot be resolved.

That the extremely conservative character of the maximin rule (K1) also poses a problem of adequacy has already been pointed out above.

All these objections, with the exception of (2), seem to be decisive. They show that, if the maximes of justice are to be anything like (K1) or (K2), they are rational only in some very specific cases.

3. SOCIAL SUBJECTIVISM

The theory of justice in social subjectivism today coincides largely with the theory of social welfare. Although economists usually try to refrain from explicit moral considerations, there is frequent reference to fairness and like moral concepts. And this, of course, is to be expected since the theory is not to be descriptive and criteria of individual rationality do not suffice to single out certain social welfare functions (SWF).

If $I = \{1, \ldots, n\}$ is a group of individuals again, Z a set of social states, $u = (u_1, \ldots, u_n)$ a n-tuple of individual utility functions on Z, as specified above, or $r = (r_1, \ldots, r_n)$ a n-tuple of individual preference orderings on Z, then a SWF $R(r)$ (or $R(u)$) is a function which assigns to every possible r (or u) on Z a social preference ordering $R(r)$ (or $R(u)$) on Z.[7] If Z is finite, as we shall assume here, the set $O_R(r) := \{x: \Lambda y(yR(r)x)\}$ of

optimal results is not empty, and the maxime of social subjectivism is

(K3) Choose a result of $O_R(r)$.

Among the many SWF's that have been proposed in the literature there is not one that has been generally accepted, so that it has been doubted whether there exists a single SWF that is appropriate for all cases, i.e. an appropriate SWF for all sets Z and all r (or u) on Z.

A central issue, e.g., is the *conflict between equality and efficiency*. If the inequality in state x is measured by the Gini-coefficient, for instance:

$$G(x, u) = \frac{n}{2 \cdot U(x)} \left(\frac{1}{n^2} \sum_i \sum_k |u_i(x) - u_k(x)| \right)$$

$$= 1 + \frac{1}{n} - \frac{2}{n \cdot U(x)} (u_1(x) + 2 \cdot u_2(x) + \ldots + n \cdot u_n(x))$$

(for $u_1(x) \geq u_2(x) \geq \ldots \geq u_n(x)$, and for $U(x) := \sum_i u_i(x)$) then a SWF exclusively oriented towards equality would be $xR_1(u)y := G(y, u) \leq G(x, u)$.

But this function does not satisfy the Pareto-condition

(P1) $\Lambda rxy(\Lambda i(xr_iy) \wedge Vk(xp_iy) \supset xP(r)y)$

(where $xp_iy := \neg(yr_ix)$ and $xP(r)y := \neg(yR(r)x)$) which is an extremely plausible condition for SWF's.

On the other hand the utilitarian function

$$xR_2(u)y := U(x) \leq U(y),$$

which satisfies (P1) and is oriented exclusively toward efficiency ('the greatest happiness of greatest number') admits of gross inequalities.[8] In cases where equality is only to be had at the price of a considerable disadvantage for all or most people, R_1 is clearly inadequate, while R_2 is inappropriate if some individuals have to pay the price for a big total profit.

If we try to combine the merits of the two functions we run into trouble, however. The general idea would be something like this: The social utility $U_s(x, u)$ is to be a weighted mean of the $u_i(x)$, the weights depending on x in such a way that the utilities of persons worse off in x (for u) count more

heavily than those of the persons better situated in x. We then arrive at the following definitions

(D1a) $U_s(x, u) := \dfrac{\sum a_i(x, u) \cdot u_i(x)}{\sum a_i(x, u)}$ and

(D1b) $xR(u)y := U_s(x, u) \leqslant U_s(y, u)$,

where

$$u_i(x) < u_k(x) \supset a_i(x, u) > a_k(x, u) .$$

If we take only the *order* of positions of the individuals in x into account, we may set $a_i(x, u) = r$ for $i \in G_r$, where G_r is the group of persons in the rth best position in x for u. In case we have $u_1(x) > \ldots > u_n(x)$ we then obtain

(D2) $$U_s(x, u) = \frac{\sum i \cdot u_i(x)}{\sum i} = \frac{U(x)}{n+1}\left(1 + \frac{1}{n} - G(x, u)\right) .$$

$U_s(x, u)$ therefore increases with $U(x)$ (efficiency) and $1 + (1/n) - G(x, u)$ (equality). But $R(u)$ still does not satisfy (P1), and $R(u)$, furthermore, is unsatisfactory since it only takes the orders of the individual welfares into account and not their utility differences.

Therefore it seems to be better to determine the weights $a_i(x, u)$ as differences between $u_i(x)$ and some fixed parameter like $U(x)$. This gives us

(D3) $$U_s(x, u) = \frac{U(x)^2 - \sum u_i(x)^2}{U(x)(n-1)} .$$

The function $R(r)$ then satisfies (P1), but $R(u)$ is not invariant with respect to (common) linear transformations of the u_i, since $U_s(x, u)$ is not a linear function of the u_i.[9]

The disagreements in the theory of SWF's are not confined to specific SWF's, however, they also relate to conditions of adequacy for SWF's, with the possible exception of (P1) and the postulate of anonymity (or impartiality). The apparent plausibility of many postulates that have been proposed, is shaken if one looks at the consequences they have in connection with other equally plausible candidates.[10] Therefore we shall not try to criticize special SWF's or special conditions of adequacy here on

the grounds that, with respect to the SWF's they propose, (K3) is not an intuitive adequate principle of justice. Most of what may be said in this respect has already been said in the literature. We shall instead raise an objection that applies to all possible SWF's.

4. THE PROBLEM OF CONSISTENCY

Subjective preferences are subject to moral evaluations. This is true even if social states are the primary objects of such evaluations as in subjectivism. For if some social state is good then an action which brings it about is good, and therefore the subjective preference is good from which this action derives by rational decision – given correct and sufficient information about relevant circumstances. We cannot say that only a person's actions are morally relevant, not his subjective preferences. Since it certainly is morally alright to act rationally, it follows that if we allow certain preferences we also allow the actions based on them by rational decision.

Let us say, that the subjective preference ordering $\leqslant_i$ of person i is *legitimate* relatively to the moral preference ordering $\leqslant$. iff for all states x and y we have $x <. y \supset x <_i y$.

Now every subjectivistic theory which does not claim, as individualism and rationalism do, that moral preferences coincide with the subjective ones, admitting therefore conflicts between the two (in the sense that we have $x <. y$ and $y <_i x$ for some person i) faces the following dilemma: As a subjective theory it bases its definition of moral preference on subjective preferences as they are. It can either say that these subjective preferences are no object of moral evaluation, or it can maintain, that their moral values derive from the preference ordering defined from them. Now after what was said above the first position can not be interpreted to mean that it is impossible or irrelevant to apply moral criteria to subjective preferences. Therefore it must be taken to imply that all subjective preferences are morally indifferent, i.e. permitted. But this is a postulate for the moral preference ordering which is independent and therefore possibly inconsistent with its explicit definition. And such an inconsistency obtains in all cases of conflict in which actual subjective preferences are not legitimate with respect to the moral valuation derived from them.

According to the second approach there is no independent stipulation about the moral legitimacy of the given subjective preferences. But some of them may in fact turn out to be illegitimate, i.e. forbidden, and then it is a postulate of morality that they should be changed. But doing so might upset the moral preference ordering which is based not on legitimate but on given preferences. So there is a feedback effect here and the whole system of subjective preferences determining the moral standards by definition and being in turn redetermined by them may be unstable.

This problem can also be illustrated in a formal way by the following impossibility result.

If some or all of the individuals in I accept the social preference ordering $R(r)$ and make it their own, this ordering should not be disturbed thereby. If some people accept a moral maxime this should not change its content. Otherwise the maxime would have to be spelled out like this: Prefer x to y, but if you (or somebody else) should actually do so, stop doing so and prefer y to x instead, and so on. Therefore it seems adequate to postulate that every SWF be *consistent* in the following sense:

(P2) $\quad \Lambda rxyi(xR(r)y \equiv xR(r/_i^{R(r)})y)$,

where $r/_i^{R(r)}$ is to be the n-tuple $(r_1, \ldots, r_{i-1}, R(r), r_{i+1}, \ldots, r_n)$.[11]

This is a very strong condition, which is not satisfied by the utilitarian SWF, e.g. It is satisfied by the maximin-relation, but this does not satisfy (P1). In fact it can easily be seen that there is no SWF $R(r)$ which satisfies (P1), (P2) and the extremely weak indifference postulate

(P3) $\quad Vrixy(xE(r)y \wedge xp_iy)$.[12]

This condition is very plausible, for otherwise we should have for all r, x and y: $xE(r)y \supset \Lambda i(xe_iy)$, i.e. two states would only be socially indifferent if they are indifferent for all individuals, but not if the contrary interests of people balance each other.

Finally, social subjectivism again faces the difficulties of morally legitimate interests and inclusiveness. Inclusivity is here in conflict with the fact that to determine the SWF for a group I of individuals, their preferences have to be known. Short of prophecy this condition cannot be fulfilled for future generations.

5. ALTRUISTIC SUBJECTIVISM

Altruistic subjectivism was the alternative that J. Butler and, after him, D. Hume put against Hobbes' thesis, that egoism is the only true motive of human action. According to Butler and Hume many actions and attitudes can only be explained if we assume an altruistic component in our preferences besides the egoistic one. According to the basic idea the altruistic component is conceived of as an aggregation of the egoistic individual preferences. If $U(x, u)$ is a social utility function defined from the individual egoistic utilities u_i of the persons $i \in I$, for instance the utilitarian function $U(x, u) = \sum_i u_i(x)$, then the *effective* or *total* utility of i is given by

$$u_i^+(x) = e_i \cdot u_i(x) + (1 - e_i) \cdot U(x, u) .$$

Here e_i is i's coefficient of egoism. $u_i^+(x)$ therefore is a weighted mean of i's self-interest and benevolence, in Hume's terminology. We shall assume $O \leqslant e_i < 1$ for all $i \in I$. (For $e_i = 1$ P_i would have no altruistic component.)

While Hume seems to think of $U(x, u)$ as the utilitarian function, any function $U(x, u)$ determined by a satisfactory SWF $R(u)$ might be employed here. So the first problem for altruism is the same as that for social subjectivism: Which function $U(x, u)$ should we choose?

Although in this approach the effective individual preferences contain a component that coincides with the moral preference defined by $xR(u)y := U(x, u) \leqslant U(y, u)$, these preferences are not identical with it, i.e. conflicts between an effective individual preference and the moral preference may arise, and therefore this approach also faces the problem (4) of consistency.

Some further problems are:

(1) Why should $U(x, u)$ only take the egoistic utilities u_i of the individuals into account and not their effective utilities u_i^+? If somebody has strong moral interests why should only his egoistic and not his moral preferences be heeded by the others' benevolence? This indeed, seems inadequate, and even more so if we think of a restriction of $U(x, u)$ to morally legitimate interests, as advocated in objection (3a) (which may also be directed against the present theory). But if the function U would refer to the u_i^+ instead of the u_i its definition would be circular.

(2) How are we to distinguish egoistic and altruistic interests in a subjectivistic approach? This problem is mirrored in Hume's theory, since he originally also included a *limited generosity*, a concern for the well-being of close relatives or friends, into the egoistic component. A concern for the interests of others may also be based on egoistic considerations in some cases, cases, e.g., for which the rationalistic thesis holds. So it should be rather difficult to draw a sharp line between egoistic and altruistic interests.

(3) Finally, as in the cases of rationalism and of social subjectivism, there is a conflict between realism and intuitive acceptability: We can only respect the interests of the others if we know about them, and we shall know them only if I is a relatively small group. But then $R(u)$ is not adequate as a moral preference relation, since it pays no heed to the interests of people outside I.

Universität Regensburg

NOTES

[1] It is extremely interesting to see how moral problems slowly emerged from considerations of usefulness in Greek antiquity. This process is not only documented in philosophical and poetical writings, but also in the growing differentiation of the pertinent semantic fields of the language.

[2] Spinoza says, for instance: "Ex virtute absolute agere nihil aliud nobis est, quam ex dŭctu rationis agere, vivere, suum esse conservare (haec tria idem significant) ex fundamento proprium utile quaerere" (Eth. IV, prop. XX) ("To act virtuously is nothing else than to act and live rationally, to conserve one's own being (these three are the same), and to serve ones own real interests".)

[3] If the individuals have different sets F_i of acts to choose from ($i = 1, \ldots, n$), we can always redefine the acts so that they can choose from the same set of acts. So no loss of generality is involved in assuming this.

[4] The domains of the variables here and in what follows are: $i, k, \ldots \varepsilon I$, $x, y, z, \ldots \varepsilon R$, $j, l, \ldots \varepsilon\{1, \ldots, m\} f, g, \ldots \varepsilon F$.

[5] If one were to define the social roles so that they also determine the preference structures of the people playing these roles, it would be improbable that some people should play some roles, and the difficulties would then only be transfered back to (1a).

[6] J. Rawls' theory of justice in (72) is not a rationalistic theory. His original position is a purely fictitious situation and therefore a proof that the principles of justice are principles of rational choice in this situation does not entail, and is not supposed to entail, that in actual situations it is rational to act upon them. (Rawls' attempt to show – mainly by psychological considerations – that there is no conflict between rationality and justice has not been very successful.) But his theory has many parallels with rationalism, and Rawls, too, encounters something like the problem, of morally legitimate interests. His solution,

developed in the 'thin theory of the good', is not applicable to (3a), however. *Primary goods* are defined as such that the possession of them (to a certain measure) is a prerequisite for realizing any plan of life, i.e. any individual preferences, however different they may be. An interest in (this measure of) these goods is certainly legitimate, but it is also certain that not all legitimate interests are interests in primary goods. Therefore the principles of justice would be too weak if they were restricted to such interests.

[7] This characterization of SWF's implies the postulate of unrestricted domain and the postulates of ordering $xR(r)y \vee yR(r)x$ and $xR(r)y \wedge yR(r)z \supset xR(r)z$. '$xR(r)y$' is to be read as 'state x is socially not better than state y'. The r_i are to be interpreted likewise, so that we have $xr_iy \equiv u_i(x) \leq u_i(y)$. $R(u)$ is to be invariant with respect to common positive linear transformations of the u_i. These are rather strong postulates, but our critique of social subjectivism will not depend essentially on them. We make these assumptions only for the sake of simplicity. – Z may also be the set of results of a game. Then $R(r)$ does not just determine what the members of I should prefer, but what they should do. – For a comprehensive exposition of the theory of SWF's see A. Sen (1970).

[8] J. Harsanyi has pointed out that in some cases, for instance distribution problems with concave individual utility functions, we have $x \in O_{R_2}(r)$ for states x of equal distribution.

[9] J. Harsanyi has pointed out this defect in connection with Sen's weak equity axiom (cf. Sen (1973), p. 18). There are other, linear SWF's of course, combining the aspects of efficiency and equality, for instance $xR(u)y := xP_3(u)y \vee xE_3(u)y \vee xR_4(u)y$, where $xR_3(u)y := \Lambda i(u_i(x) \leq u_i(y)) \vee Vi(u_i(x) < u_i(y))$, $xE_3(u)y := xR_3(u)y \vee yR_3(u)x$ and $xR_4(u)y := \min_i u_i(x) \leq \min_i u_i(y)$. Cf. (76), pp. 71 seq.

But this relation is oriented towards efficiency only so long as unamimity obtains. So it refers the conflict between efficiency and equality to the Pareto-optimal set P of Z and there it is purely egalitarian in so far as it promotes the interests of those worst off at the cost of the others.

[10] This is witnessed especially by the impossibility results.

[11] (P2) is a consequence of the postulate $\Lambda rr'(\Lambda zz'(zr'z' \supset zR(r)z') \supset \Lambda ixy(xR(r)y \equiv xR(r/r'_i)y))$ which says that $R(r)$ is to be invariant with respect to substitutions of preference orderings r' for r_i which are legitimate with respect to $R(r)$.

[12] E is defined as in Note 9. – If r, i, x, y are as postulated in (P3) we can set $r'_k = R(r)$ for all $k \neq i$ and $r'_i = r_i$. Then (P1) tells us that $xP(r')y$, which violates (P2). For a SWF $R(u)$ defined by a function $U(u)$ and $xR(u)y \equiv U(x, u) \leq U(y, u)$, like the utilitarian or the maximin relation, (P2) would have to be formulated as $\Lambda uxyi(xR(u)y \equiv xR(u/_i^{U(u)})y)$.

REFERENCES

Harsanyi, J. C.: *Essays on Ethics, Social Behavior, and Scientific Explanation*, Dordrecht 1976.

Rawls, J.: *A Theory of Justice*, Oxford 1972.

Sen, A.: *Collective Choice and Social Welfare*, San Francisco 1970.

Sen, A.: *On Economic Inequality*, Oxford 1973.

Manuscript submitted 20 August 1976
Final manuscript received 22 November 1976