Chapter III

THE SOCIAL WELFARE FUNCTION

1. Formal Statement of the Problem of Social Choice

I will largely restate Professor Bergson's formulation of the problem of making welfare judgments [1] in the terminology here adopted. The various arguments of his social welfare function are the components of what I have here termed the social state, so that essentially he is describing the process of assigning a numerical social utility to each social state, the aim of society then being described by saying that it seeks to maximize the social utility or social welfare subject to whatever technological or resource constraints are relevant or, put otherwise, that it chooses the social state yielding the highest possible social welfare within the environment. As with any type of behavior described by maximization, the measurability of social welfare need not be assumed; all that matters is the existence of a social ordering satisfying Axioms I and II. As before, all that is needed to define such an ordering is to know the relative ranking of each pair of alternatives.

The relative ranking of a fixed pair of alternative social states will vary, in general, with changes in the values of at least some individuals; to assume that the ranking does not change with any changes in individual values is to assume, with traditional social philosophy of the Platonic realist variety, that there exists an objective social good defined independently of individual desires. This social good, it was frequently held, could best be apprehended by the methods of philosophic inquiry. Such a philosophy could be and was used to justify government by the elite, secular or religious, although we shall see below that the connection is not a necessary one.

To the nominalist temperament of the modern period, the assumption of the existence of the social ideal in some Platonic realm of being was meaningless. The utilitarian philosophy of Jeremy Bentham and his followers sought instead to ground the social good on the good of individuals. The hedonist psychology associated with utilitarian philosophy was further used to imply that each individual's good was identical with his desires. Hence, the social good was in some sense to be a

[1] Bergson, "A Reformulation . . . ," *op. cit.*, *passim*.

composite of the desires of individuals. A viewpoint of this type serves as a justification of both political democracy and laissez-faire economics or at least an economic system involving free choice of goods by consumers and of occupations by workers.

The hedonist psychology finds its expression here in the assumption that individuals' behavior is expressed by individual ordering relations R_i. Utilitarian philosophy is expressed by saying that for each pair of social states the choice depends on the ordering relations of all individuals, i.e., depends on $R_1, \cdots, R_n$, where n is the number of individuals in the community. Put otherwise, the whole social ordering relation R is to be determined by the individual ordering relations for social states, $R_1, \cdots, R_n$. We do not exclude here the possibility that some or all of the choices between pairs of social states made by society might be independent of the preferences of certain particular individuals, just as a function of several variables might be independent of some of them.

DEFINITION 4: *By a* social welfare function *will be meant a process or rule which, for each set of individual orderings* $R_1, \cdots, R_n$ *for alternative social states (one ordering for each individual), states a corresponding social ordering of alternative social states,* R.

As a matter of notation, we will let R be the social ordering corresponding to the set of individual orderings $R_1, \cdots, R_n$, the correspondence being that established by a given social welfare function; if primes or seconds are added to the symbols for the individual orderings, primes or seconds will be added to the symbol for the corresponding social ordering.

There is some difference between the concept of social welfare function used here and that employed by Bergson. The individual orderings which enter as arguments into the social welfare function as defined here refer to the values of individuals rather than to their tastes. Bergson supposes individual values to be such as to yield a social value judgment leading to a particular rule for determining the allocation of productive resources and the distribution of leisure and final products in accordance with individual tastes. In effect, the social welfare function described here is a method of choosing which social welfare function of the Bergson type will be applicable, though, of course, I do not exclude the possibility that the social choice actually arrived at will not be consistent with the particular value judgments formulated by Bergson. But in the formal aspect the difference between the two definitions of social welfare function is not too important. In Bergson's treatment, the tastes of individuals (each for his own consumption) are represented by utility functions, i.e., essentially by ordering relations; hence the Bergson social welfare function is also a rule for assigning to each set of individual

orderings a social ordering of social states. Furthermore, as already indicated, no sharp line can be drawn between tastes and values.

A special type of social welfare function would be one which assigns the same social ordering for every set of individual orderings. In this case, of course, social choices are completely independent of individual tastes, and we are back in the Platonic case.

If we do not wish to require any prior knowledge of the tastes of individuals before specifying our social welfare function, that function will have to be defined for every logically possible set of individual orderings. Such a social welfare function would be universal in the sense that it would be applicable to any community. This ideal seems to be implicit in Benthamite social ethics and in its latter-day descendant, welfare economics.

However, we need not ask ourselves if such a universal social welfare function can be defined. Let an *admissible* set of individual ordering relations be a set for which the social welfare function defines a corresponding social ordering, i.e., a relation satisfying Axioms I and II. A universal social welfare function would be one for which every set of individual orderings was admissible. However, we may feel on some sort of a priori grounds that certain types of individual orderings need not be admissible. For example, it has frequently been assumed or implied in welfare economics that each individual values different social states solely according to his consumption under them. If this be the case, we should only require that our social welfare function be defined for those sets of individual orderings which are of the type described; only such should be admissible.

We will, however, suppose that our a priori knowledge about the occurrence of individual orderings is incomplete, to the extent that there are at least three among all the alternatives under consideration for which the ordering by any given individual is completely unknown in advance. That is, every logically possible set of individual orderings of a certain set S of three alternatives can be obtained from some admissible set of individual orderings of all alternatives. More formally, we have

CONDITION 1: *Among all the alternatives there is a set S of three alternatives such that, for any set of individual orderings $T_1, \cdots, T_n$ of the alternatives in S, there is an admissible set of individual orderings $R_1, \cdots, R_n$ of all the alternatives such that, for each individual i, $x R_i y$ if and only if $x T_i y$ for x and y in S.*

Condition 1, it should be emphasized, is a restriction on the form of the social welfare function since, by definition of an admissible set of

individual orderings, we are requiring that, for some sufficiently wide range of sets of individual orderings, the social welfare function give rise to a true social ordering.

We also wish to impose several other apparently reasonable conditions on the social welfare function.

2. Positive Association of Social and Individual Values

Since we are trying to describe social welfare and not some sort of illfare, we must assume that the social welfare function is such that the social ordering responds positively to alterations in individual values, or at least not negatively. Hence, if one alternative social state rises or remains still in the ordering of every individual without any other change in those orderings, we expect that it rises, or at least does not fall, in the social ordering.

This condition can be reformulated as follows: Suppose, in the initial position, that individual values are given by a set of individual orderings $R_1, \cdots, R_n$, and suppose that the corresponding social ordering R is such that $x\,P\,y$, where x and y are two given alternatives and P is the preference relation corresponding to R, i.e., defined in terms of R in accordance with Definition 1. Suppose values subsequently change in such a way that for each individual the only change in relative rankings, if any, is that x is higher in the scale than before. If we call the new individual orderings (those expressing the new set of values) $R_1', \cdots, R_n'$ and the social ordering corresponding to them R', then we would certainly expect that $x\,P'\,y$, where P' is the preference relation corresponding to R'. This is a natural requirement since no individual ranks x lower than he formerly did; if society formerly ranked x above y, we should certainly expect that it still does.

We have still to express formally the condition that x be not lower on each individual's scale while all other comparisons remain unchanged. The last part of the condition can be expressed by saying that, among pairs of alternatives neither of which is x, the relation R_i' will obtain for those pairs for which the relation R_i holds and only such; in symbols, for all $x' \neq x$ and $y' \neq x$, $x'\,R_i'\,y'$ if and only if $x'\,R_i\,y'$. The condition that x be not lower on the R_i' scale than x was on the R_i scale means that x is preferred on the R_i' scale to any alternative to which it was preferred on the old (R_i) scale and also that x is preferred or indifferent to any alternative to which it was formerly indifferent. The two conditions of the last sentence, taken together, are equivalent to the following two conditions: (1) x is preferred on the new scale to any alternative to which it was formerly preferred; (2) x is preferred or indifferent on

the new scale to any alternative to which it was formerly preferred or indifferent. In symbols, for all y', $x\ R_i\ y'$ implies $x\ R_i'\ y'$, and $x\ P_i\ y'$ implies $x\ P_i'\ y'$. We can now state the second condition which our social welfare function must satisfy.

CONDITION 2: *Let* $R_1, \cdots, R_n$ *and* $R_1', \cdots, R_n'$ *be two sets of individual ordering relations,* R *and* R' *the corresponding social orderings, and* P *and* P' *the corresponding social preference relations. Suppose that for each* i *the two individual ordering relations are connected in the following ways: for* x' *and* y' *distinct from a given alternative* x, $x'\ R_i'\ y'$ *if and only if* $x'\ R_i\ y'$; *for all* y', $x\ R_i\ y'$ *implies* $x\ R_i'\ y'$; *for all* y', $x\ P_i\ y'$ *implies* $x\ P_i'\ y'$. *Then, if* $x\ P\ y$, $x\ P'\ y$.

3. THE INDEPENDENCE OF IRRELEVANT ALTERNATIVES

If we consider $C(S)$, the choice function derived from the social ordering R, to be the choice which society would actually make if confronted with a set of alternatives S, then, just as for a single individual, the choice made from any fixed environment S should be independent of the very existence of alternatives outside of S. For example, suppose that an election system has been devised whereby each individual lists all the candidates in order of his preference and then, by a preassigned procedure, the winning candidate is derived from these lists. (All actual election procedures are of this type, although in most the entire list is not required for the choice.) Suppose that an election is held, with a certain number of candidates in the field, each individual filing his list of preferences, and then one of the candidates dies. Surely the social choice should be made by taking each of the individual's preference lists, blotting out completely the dead candidate's name, and considering only the orderings of the remaining names in going through the procedure of determining the winner. That is, the choice to be made among the set S of surviving candidates should be independent of the preferences of individuals for candidates not in S. To assume otherwise would be to make the result of the election dependent on the obviously accidental circumstance of whether a candidate died before or after the date of polling. Therefore, we may require of our social welfare function that the choice made by society from a given environment depend only on the orderings of individuals among the alternatives in that environment. Alternatively stated, if we consider two sets of individual orderings such that, for each individual, his ordering of those particular alternatives in a given environment is the same each time, then we require that the choice made by society from that environment be the same when indi-

vidual values are given by the first set of orderings as they are when given by the second.

CONDITION 3: *Let* $R_1, \cdots, R_n$ *and* $R_1', \cdots, R_n'$ *be two sets of individual orderings and let* $C(S)$ *and* $C'(S)$ *be the corresponding social choice functions. If, for all individuals* i *and all* x *and* y *in a given environment* S, $x R_i y$ *if and only if* $x R_i' y$, *then* $C(S)$ *and* $C'(S)$ *are the same* (independence of irrelevant alternatives).

The reasonableness of this condition can be seen by consideration of the possible results in a method of choice which does not satisfy Condition 3, the rank-order method of voting frequently used in clubs.[2] With a finite number of candidates, let each individual rank all the candidates, i.e., designate his first-choice candidate, second-choice candidate, etc. Let preassigned weights be given to the first, second, etc., choices, the higher weight to the higher choice, and then let the candidate with the highest weighted sum of votes be elected. In particular, suppose that there are three voters and four candidates, x, y, z, and w. Let the weights for the first, second, third, and fourth choices be 4, 3, 2, and 1, respectively. Suppose that individuals 1 and 2 rank the candidates in the order x, y, z, and w, while individual 3 ranks them in the order z, w, x, and y. Under the given electoral system, x is chosen. Then, certainly, if y is deleted from the ranks of the candidates, the system applied to the remaining candidates should yield the same result, especially since, in this case, y is inferior to x according to the tastes of every individual; but, if y is in fact deleted, the indicated electoral system would yield a tie between x and z.

A similar problem arises in ranking teams in a contest which is essentially individual, e.g., a foot race in which there are several runners from each college, and where it is desired to rank the institutions on the basis of the rankings of the individual runners. This problem has been studied by Professor E. V. Huntington,[3] who showed by means of an example that the usual method of team scoring in those circumstances, a method analogous to the rank-order method of voting, was inconsistent with a condition analogous to Condition 3, which Huntington termed the postulate of relevancy.

The condition of the independence of irrelevant alternatives implies that in a generalized sense all methods of social choice are of the type of

[2] This example was suggested by a discussion with G. E. Forsythe, National Bureau of Standards.

[3] E. V. Huntington, "A Paradox in the Scoring of Competing Teams," *Science*, Vol. 88, September 23, 1938, pp. 287–288. I am indebted for this reference to J. Marschak.

voting. If S is the set $[x, y]$ consisting of the two alternatives x and y, Condition 3 tells us that the choice between x and y is determined solely by the preferences of the members of the community as between x and y. That is, if we know which members of the community prefer x to y, which are indifferent, and which prefer y to x, then we know what choice the community makes. Knowing the social choices made in pairwise comparisons in turn determines the entire social ordering and therewith the social choice function $C(S)$ for all possible environments. Condition 2 guarantees that voting for a certain alternative has the usual effect of making surer that that alternative will be adopted.

Condition 1 says, in effect, that, as the environment varies and individual orderings remain fixed, the different choices made shall bear a certain type of consistent relation to each other. Conditions 2 and 3, on the other hand, suppose a fixed environment and say that, for certain particular types of variation in individual values, the various choices made have a certain type of consistency.

4. The Condition of Citizens' Sovereignty

We certainly wish to assume that the individuals in our society are free to choose, by varying their values, among the alternatives available. That is, we do not wish our social welfare function to be such as to prevent us, by its very definition, from expressing a preference for some given alternative over another.

Definition 5: *A social welfare function will be said to be* imposed *if, for some pair of distinct alternatives x and y, $x\,R\,y$ for any set of individual orderings $R_1, \cdots, R_n$, where R is the social ordering corresponding to $R_1, \cdots, R_n$.*

In other words, when the social welfare function is imposed, there is some pair of alternatives x and y such that the community can never express a preference for y over x no matter what the tastes of all individuals are, even if all individuals prefer y to x; some preferences are taboo. (Note that, by Definition 1, asserting that $x\,R\,y$ holds for all sets of individual orderings is equivalent to asserting that $y\,P\,x$ never holds.)

At the beginning of this study, allusion was made to the type of social choice in which decisions are made in accordance with a customary code. It is arguable whether or not Definition 5 catches the essence of the intuitive idea of conventional choice. In the true case of customary

restraints on social choice, presumably the restraints are not felt as such but really are part of the tastes of the individuals. The problems here involve psychological subtleties; can we speak, in the given situation, of true desires of the individual members of the society which are in conflict with the custom of the group?

If the answer to the last question is yes, then Definition 5 is indeed a correct formalization of the concept of conventionality. But we need not give a definite answer, and this is especially fortunate since an examination of the question would take us very far afield indeed. For certainly we wish to impose on our social welfare function the condition that it not be imposed in the sense of Definition 5; we certainly wish all choices to be possible if unanimously desired by the group. If Definition 5 is not a model of customary choice, it is at least a model of external control, such as obtains in a colony or an occupied country.

CONDITION 4: *The social welfare function is not to be imposed.*

Condition 4 is stronger than need be for the present argument. Some decisions as between given pairs of alternatives may be assumed to be imposed. All that is required really is that there be a set S of three alternatives such that the choice between any pair is not constrained in advance by the social welfare function. This set S must also have the properties indicated in Condition 1.

If the answer to the question asked earlier is that there is no sense in speaking of a conflict of wills between the individual and the sacred code, then we have a situation in which it is known in advance that the individual orderings of social alternatives conform to certain restrictions, i.e., that certain of the choices made by individuals are preassigned. In that case, we might desire that the social welfare function be defined only for sets of individual orderings compatible with the known socio-ethical norms of the community; this requirement may involve a weakening of Condition 1. This point will be discussed at greater length in Chapter VII.

It should also be noted that Condition 4 excludes the Platonic case discussed in Section 1 of this chapter. It expresses fully the idea that all social choices are determined by individual desires. In conjunction with Condition 2 (which insures that the determination is in the direction of agreeing with individual desires), Condition 4 expresses the same idea as Bergson's Fundamental Value Propositions of Individual Preference, which state that, between two alternatives between which all individuals but one are indifferent, the community will prefer one over the other or be indifferent between the two according as the one indi-

vidual prefers one over the other or is indifferent between the two.[4] Conditions 2 and 4 together correspond to the usual concept of consumer's sovereignty; since we are here referring to values rather than tastes, we might refer to them as expressing the idea of citizens' sovereignty.

5. The Condition of Nondictatorship

A second form of social choice not of a collective character is the choice by dictatorship. In its pure form, it means that social choices are to be based solely on the preferences of one man. That is, whenever the dictator prefers x to y, so does society. If the dictator is indifferent between x and y, presumably he will then leave the choice up to some or all of the other members of society.

Definition 6: *A social welfare function is said to be* dictatorial *if there exists an individual i such that, for all x and y, $x P_i y$ implies $x P y$ regardless of the orderings $R_1, \cdots, R_n$ of all individuals other than i, where P is the social preference relation corresponding to $R_1, \cdots, R_n$.*

Since we are interested in the construction of collective methods of social choice, we wish to exclude dictatorial social welfare functions.

Condition 5: *The social welfare function is not to be dictatorial* (nondictatorship).

Again, it cannot be claimed that Definition 6 is a true model of actual dictatorship. There is normally an element of consent by the members of the community or at least a good many of them. This may be expressed formally by saying that the desires of those individuals include a liking for having social decisions made by a dictator[5] or at least a liking for the particular social decisions which they expect the dictator to make. The idea of a taste for dictatorship on the part of individuals will be discussed in Chapter VII at somewhat greater length. However, in any case, Condition 5 is certainly a reasonable one to impose on the form of the social welfare function.

We have now imposed five apparently reasonable conditions on the construction of a social welfare function. These conditions are, of course,

[4] Bergson, "A Reformulation . . . ," *op. cit.*, pp. 318–320. The Fundamental Value Propositions of Individual Preference are not, strictly speaking, implied by Conditions 2 and 4 (in conjunction with Conditions 1 and 3), though something very similar to them is so implied; see Consequence 3 in Chapter V, Section 3. A slightly stronger form of Condition 2 than that stated here would suffice to yield the desired implication.

[5] See E. Fromm, *Escape from Freedom*, New York: Rinehart and Co., 1941, 305 pp.

value judgments and could be called into question; taken together they express the doctrines of citizens' sovereignty and rationality in a very general form, with the citizens being allowed to have a wide range of values. The question raised is that of constructing a social ordering of all conceivable alternative social states from any given set of individual orderings of those social states, the method of construction being in accordance with the value judgments of citizens' sovereignty and rationality as expressed in Conditions 1–5.

6. The Summation of Utilities

It may be instructive to consider that proposed social welfare function which has the longest history, the Bentham-Edgeworth sum of individual utilities. As it stands, this form seems to be excluded by the entire nature of the present approach, since, in Chapter II, Section 1, we agreed to reject the idea of cardinal utility, and especially of interpersonally comparable utility. However, presumably the sum of utilities could be reformulated in a way which depends only on the individual orderings and not on the utility indicators. This seems to be implied by Bergson's discussion of this social welfare function;[6] though he presents a number of cogent arguments against the sum-of-utilities form, he does not find that it contradicts the Fundamental Value Propositions of Individual Preference (see Section 4 above), which he would have to if he did not consider that form to be determined by the individual orderings. The only way that I can see of making the sum of utilities depend only on the indifference loci is the following: Since to each individual ordering there corresponds an infinite number of utility indicators, set up an arbitrary rule which assigns to each indifference map one of its utility indicators; then the sum of the particular utility indicators chosen by the rule is a function of the individual orderings and can be used to establish a social ordering.

Obviously, this formation of the sum of utilities will lead to different decisions in a given situation with different choices of the rule. For any rule, Condition 1 is satisfied. However, Conditions 2 and 3 essentially prescribe that, for a given environment, the choice made shall vary in a particular way with certain variations in the orderings of individuals. This being so, it is clear that for the sum of utilities to satisfy Conditions 2 and 3, it would be necessary for the rule to be stringently limited; in fact, the general theorem, established in Chapter V, guarantees that the only rules which would make the sum of utilities satisfy Conditions 2 and 3, if any, lead it to violate either Condition 4 or Condi-

[6] Bergson, "A Reformulation . . . ," *op. cit.*, pp. 324, 327–328.

tion 5. Indeed, according to Theorem 3 in Chapter VI, Section 3, the same would be true even if it were assumed that the utility of each individual depended solely on his own consumption. I have not been able to construct a special proof of this fact for the sum of utilities which is essentially different from the proof of the general theorem.

It may be of interest, however, to consider a particular rule for assigning utility indicators to individual orderings.[7] Assume that the individual orderings for probability distributions over alternatives obey the axioms of von Neumann and Morgenstern;[8] then there is a method of assigning utilities to the alternatives, unique up to a linear transformation, which has the property that the probability distributions over alternatives are ordered by the expected value of utility. Assume that for each individual there is always one alternative which is preferred or indifferent to all other conceivable alternatives and one to which all other alternatives are preferred or indifferent. Then, for each individual, the utility indicator can be defined uniquely among the previously defined class, which is unique up to a linear transformation, by assigning the utility 1 to the best conceivable alternative and 0 to the worst conceivable alternative. This assignment of values is designed to make individual utilities interpersonally comparable.

It is not hard to see that the suggested assignment of utilities is extremely unsatisfactory. Suppose there are altogether three alternatives and three individuals. Let two of the individuals have the utility 1 for alternative x, .9 for y, and 0 for z; and let the third individual have the utility 1 for y, .5 for x and 0 for z. According to the above criterion, y is preferred to x. Clearly, z is a very undesirable alternative since each individual regards it as worst. If z were blotted out of existence, it should not make any difference to the final outcome; yet, under the proposed rule for assigning utilities to alternatives, doing so would cause the first two individuals to have utility 1 for x and 0 for y, while the third individual has utility 0 for x and 1 for y, so that the ordering by sum of utilities would cause x to be preferred to y.

A simple modification of the above argument shows that the proposed rule does not lead to a sum-of-utilities social welfare function consistent with Condition 3. Instead of blotting z out of existence, let the individual orderings change in such a way that the first two individuals find z indifferent to x and the third now finds z indifferent to y, while the relative positions of x and y are unchanged in all individual orderings. Then the assignment of utilities to x and y becomes the same as it

[7] This particular rule was suggested by A. Kaplan.

[8] See fn. 1, Chapter II.

became in the case of blotting out z entirely, so that again the choice between x and y is altered, contrary to Condition 3.

The above result appears to depend on the particular method of choosing the units of utility. But this is not true, although the paradox is not so obvious in other cases. The point is, in general, that the choice of two particular alternatives to produce given utilities (say 0 and 1) is an arbitrary act, and this arbitrariness is ultimately reflected in the failure of the implied social welfare function to satisfy one of the conditions laid down.