PART V

Normative public choice

CHAPTER 23

Social welfare functions

> The interest of the community then is – what? The sum of the interests of the several members who compose it.
>
> Jeremy Bentham

Whereas one can speak of *the* positive theory of public choice, based upon economic man assumptions, one must think of normative *theories* of public choice, for there are many views of what the goals of the state should be and how to achieve them. This potential multiplicity has been the focus of much criticism by positivists, who have argued for a "value-free" discipline. For the bulk of economics, it might be legitimate to focus on explanation and prediction, and leave to politics the explication of the goals of society. For the study of politics itself, in toto, to take this position is less legitimate; thus the interest in how the basic values of society are or can be expressed through the political process. The challenge that normative theory faces is to develop theorems about the expression and realization of values, based on generally accepted postulates, in the same way that positive theory has developed explanatory and predictive theorems from the postulates of rational egoistic behavior. Part V reviews some efforts to take up this challenge.

23.1 The Bergson-Samuelson social welfare function

The traditional means for representing the values of the community in economics is to use a social welfare function (SWF). The seminal paper on SWFs is by Bergson (1938), with the most significant further explication by Samuelson (1947, ch. 8). The SWF can be written as follows:

$$W = W(z_1, z_2, \ldots, z_n),$$

where W is a real valued function of all variables, and the z_is and W are chosen to represent the ethical values of the society or of the individuals in it (Samuelson, 1947, p. 221). The objective is to define a W and set of z_is, and the constraints thereon, to yield meaningful first- and second-order conditions for a maximum W. Although in principle any variables that are related to a society's well-being (e.g., crime statistics, weather data, years of schooling) might be included in the SWF, economists have focused on economic variables. Thus, the SWF literature has adopted the same assumptions about consumers, production functions, and so on, that underlie the bulk of economics and public choice and has made these the focal point of its analysis.

The only value postulate upon which general agreement has been possible has been the Pareto postulate. This postulate suffices to bring about a set of *necessary* conditions for the maximization of W, which limit social choices to points along the generalized Pareto frontier. The proof is analogous to the demonstration that movement from off the contract curve to points on it can be Pareto improvements, and the necessary conditions are also analogous. With respect to production, these conditions are

$$\frac{\partial X_i/\partial V_{1i}}{\partial X_k/\partial V_{1k}} = \cdots = \frac{\partial X_i/\partial V_{mi}}{\partial X_k/\partial V_{mk}} = \frac{T_{xk}}{T_{xi}}, \tag{23.1}$$

where $\partial X_i/\partial V_{mi}$ is the marginal product of factor V_m in the production of output X_i, and T is the transformation function defined over all products and inputs (Samuelson, 1947, pp. 230–3).

> In words this takes the form: *productive factors are correctly allocated if the marginal productivity of a given factor in one line is to the marginal productivity of the same factor in a second line as the marginal productivity of any other factor in the first line is to its marginal productivity in the second line. The value of the common factor of proportionality can be shown to be equal to the marginal cost of the first good in terms of the (displaced amount of the) second good.* (Samuelson, 1947, p. 233; italics in original)

These conditions ensure that the economy is operating on the production possibility frontier. If these conditions were not met, it would be possible to transfer factors of production from one process to another and obtain more of one product without giving up any amounts of another. Such possibilities are ruled out by the Pareto principle.

The necessary conditions for consumption require that the marginal rate of substitution between any two private goods, i and j, be the same for all individuals consuming both goods:

$$\frac{\partial U_1/\partial X_i}{\partial U_1/\partial X_j} = \frac{\partial U_2/\partial X_i}{\partial U_2/\partial X_j} = \cdots = \frac{\partial U_s/\partial X_i}{\partial U_s/\partial X_j}, \tag{23.2}$$

where $(\partial U_k/\partial X_i)/(\partial U_k/\partial X_j)$ is voter k's marginal rate of substitution between i and j (Samuelson, 1947, pp. 236–8). If (23.2) were not fulfilled, gains from trade would exist, again violating the Pareto postulate. Thus, choice is limited to points along the production possibility frontier – distributions of final products that bring about equality between the marginal rate of transformation of one product into another, and individual marginal rates of substitution (Samuelson, 1947, pp. 238–40).

Through the appropriate set of lump-sum taxes and transfers it is possible to sustain any point along the Pareto-possibility frontier as a competitive equilibrium. Thus, the normative issue to be resolved with the help of the SWF is which point along the generalized Pareto-possibility frontier should be chosen; what set of lump-sum taxes and subsidies is optimal. Both Bergson and Samuelson speak of solving this question with the help of a variant of the SWF in which the utility indexes of each individual are direct arguments in the welfare function

$$W = W(U_1, U_2, \ldots, U_s). \tag{23.3}$$

The issue then arises as to what form W takes, and what the characteristics of the individual utility functions are. In particular, one wants to know whether ordinal utility functions are sufficient, or whether cardinal utility indexes are required, and if the latter, whether interpersonal comparability is required as well. Since the evolution of utility theory over the last century has led to an almost unanimous rejection of cardinal, interpersonally comparable utility functions throughout much of economics, the hope is, of course, that they will not be needed here. But, alas, that hope is in vain.

To see why this is so consider the following simple example: six apples are to be divided between two individuals. On the basis of knowledge of the positions of the two individuals, their tastes for apples, and the ethical values and norms of the community, we believe that social welfare will be maximized with an even division of the apples. The question then is whether an ordinal representation of individuals 1 and 2's preferences can be constructed that always yields this result. Consider first the additive welfare function

$$W = U_1 + U_2. \tag{23.4}$$

We wish to select U_1 and U_2 such that

$$U_1(3) + U_2(3) > U_1(4) + U_2(2). \tag{23.5}$$

Inequality (23.5) implies

$$U_2(3) - U_2(2) > U_1(4) - U_1(3). \tag{23.6}$$

If U_1 is an ordinal utility function, it can be transformed into an equivalent ordinal function by multiplying it by k. This transformation multiplies the right-hand side of (23.6) by k, however, and given any choice of U_2 that is bounded, a k can always be found that will reverse the inequality in (23.6), assuming $U_1(4) - U_1(3) > 0$.

The same holds if W is multiplicative. We then seek a U_1 and U_2 such that

$$U_1(3) \cdot U_2(3) > U_1(4) \cdot U_2(2), \tag{23.7}$$

which is equivalent to

$$\frac{U_2(3)}{U_2(2)} > \frac{U_1(4)}{U_1(3)}. \tag{23.8}$$

However, the ordinality of U_2 is not affected by adding a constant to it, so that (23.8) should also hold for

$$\frac{U_2(3)+k}{U_2(2)+k} > \frac{U_1(4)}{U_1(3)}. \tag{23.9}$$

But the left-hand side of (23.9) tends toward one as k becomes larger, and the inequality will thus reverse for some sufficiently large k if individual 1 experiences some positive utility from consuming the fourth apple.

Other algebraic forms of W are possible, but it should be obvious that the pliability of an ordinal utility function is such that these, too, will be incapable of yielding a maximum at (3,3) under every possible transformation that preserves the ordinality of the Us. The same arguments could be repeated with respect to a comparison of the distribution (4,2) with (5,1), and the distribution (5,1) and (6,0). The only way we will get a determinant outcome from an SWF whose arguments are ordinal utility indicators is to define it lexicographically, that is, to state that society prefers any increase in 1's utility, however small, to any increase in 2's utility, however large, and have this hold independently of the initial utility levels (distribution of income and goods); which is to say, an SWF defined over ordinal utility indexes must be dictatorial if it is to select a single outcome consistently. This result was first established by Kemp and Ng (1976) and Parks (1976) with proofs that follow the Arrow impossibility proofs discussed in Chapter 24 (see also Hammond, 1976; Roberts, 1980c).

The very generality of the ordinal utility function, which makes it attractive for the analysis of *individual* decisions, makes it unsuitable for the analysis of *social* decisions, where trade-offs *across individuals* are envisaged. To make these trade-offs, *either* the relative positions of individuals must be compared directly in terms of the bundles of commodities or command over these commodities they enjoy using the ethical norms of the community, or, if utility indexes are employed, these must be defined in such a way as to make cardinal, interpersonal comparisons possible.

All of this would appear to have been known for some time. Although Bergson's initial exposition of the SWF seems to have led to some confusion over the need for cardinal utilities and interpersonal comparisons,[1] this need was emphasized by Lerner (1944, ch. 3) and clearly addressed by Samuelson (1947, p. 244) in his initial

[1] At several places Bergson emphasizes that only ordinal utility indexes are required when deriving the optimality conditions for the SWF and he states directly, "In my opinion the utility calculus introduced by the Cambridge economists [i.e., cardinality] is not a useful tool for welfare economics" (1938, p. 20). From these statements undoubtedly arises the view that Bergson claimed that welfare judgments could be based on ordinal utility indicators. Thus, for example, we have Arrow (1963, p. 110) stating, "It is the great merit of Bergson's 1938 paper to have carried the same principle [Leibnitz's principle of the identity of indiscernibles] into the analysis of social welfare. The social welfare function was to depend only on indifference maps; in other words, welfare judgments were to be based only on interpersonally observable behavior." But the clauses preceding and following "in other words" are not equivalent. And, in fact, Bergson goes on following his attack on the Cambridge economists' use of cardinal utility to argue not for the use of ordinal utility indexes or "interpersonally observable *behavior*," but for interpersonal comparisons of "relative economic positions" and "different commodities." Thus, in rejecting cardinal utility, Bergson opts not for a W defined over ordinal Us but for W defined over the actual physical units, that is, $W(z_1, z_2, \ldots, z_n)$. This leaves the status of W defined over individual, ordinal utility indexes indeterminate, at best.

In his discussion of Arrow's theorem in 1954, Bergson states quite clearly, to my mind, that interpersonal cardinal utility comparisons are required (see, in particular, his discussion of the distribution of wine and bread on pp. 244–5, and n. 8), but Arrow (1963, pp. 111–12) would not agree.

exploration of the SWF:

> An infinity of such positions [points along the generalized contract locus] exists ranging from a situation in which all of the advantage is enjoyed by one individual, through some sort of compromise position, to one in which another individual has all the advantage. Without a well-defined W function, i.e., without assumptions concerning interpersonal comparisons of utility, it is impossible to decide which of these points is best. In terms of a given set of ethical notions which define a *Welfare function* the best point on the generalized contract locus can be determined, and only then. (Italics in original)

And we have Samuelson's (1967) subsequent proof that cardinality alone will not suffice; that is, cardinality *and* interpersonal comparability are required. The issue of whether the arguments of the SWF can be ordinal utility indexes would seem to be finally closed with the appearance of the papers by Kemp and Ng and Parks, were it not that these articles sparked a controversy over precisely the cardinality-ordinality issue involving, perhaps surprisingly, Samuelson (and indirectly Bergson, also). Given the personages involved and the issues at debate, it is perhaps useful to pause and examine their arguments.

The main purpose of Samuelson's (1977) attack on the Kemp-Ng and Parks theorems is, as the title of his note states, to reaffirm the existence of "reasonable" Bergson-Samuelson SWFs. And the note is clearly provoked by the claims by Kemp and Ng and Parks of having established nonexistence or impossibility theorems. In criticizing their theorems, Samuelson focuses on the particular form of axiom Kemp and Ng use to capture ordinality in a Bergson-Samuelson SWF, an axiom that implies that the SWF must be lexicographic. Samuelson is obviously correct in deriding an axiom that makes one individual an "ethical dictator," but his criticism of the theorems of Kemp-Ng, and Parks is misplaced. As Parks's proof most clearly shows, all Bergson-Samuelson SWFs based on ordinal preferences make one individual an ethical dictator.

A careful reading of the Kemp and Ng and Parks papers indicates that they do not claim the nonexistence of *all* reasonable Bergson-Samuelson SWFs, but only of those whose arguments are ordinal, individual utility indicators. Interestingly enough, Kemp and Ng (1976, p. 65) cite Samuelson himself as one of those holding "the apparently widely held belief that Bergson-Samuelson SWFs can be derived from individual ordinal utilities." They cite page 228 of the *Foundations*, the same page, incidentally, that Arrow (1963, pp. 10, 110, n. 49) cites, to indicate that the SWF *is* based on ordinal utilities. On this page appears the following:

> Of course, if utilities are to be added, one would have to catch hold of them first, but there is no need to add utilities. The cardinal utilities enter into the W function as independent variables if assumption (5) [individuals' preferences are to "count"] is made. But the W function is itself only ordinally determinable so that there are an infinity of equally good indicators of it which can be used. Thus, if one of these is written as
>
> $$W = F(U_1, U_2, \ldots),$$

> and if we were to change from one set of cardinal indexes of individual utility to another set $(V_1, V_2, \ldots)$, we should simply change the form of the function F so as to leave all social decisions invariant.

This passage clearly states that W is ordinal and seems to imply that the individual utility arguments need not be interpersonally comparable. But the passage appears in the section in which the necessary conditions defining points *along the generalized Pareto-possibility frontier* are derived, and is obviously superseded, or amplified by the passage appearing later in the book on p. 244 and quoted above, where Samuelson makes clear that one *must* "catch hold" of the individual utilities and compare them if a single point out of the Pareto set is to be chosen. However, subsequent statements by Samuelson and his vigorous attack on the Kemp-Ng-Parks theorems would seem to imply that he believes that Bergson-Samuelson SWFs are well defined even when they have the ordinal utility functions of individuals as arguments.[2] The theorems of Kemp and Ng (1976), Parks (1976), Hammond (1976), Roberts (1980c), and still others deny this interpretation. Rather, one must conclude (1) that ordinal utility functions are sufficient as arguments of W when deriving the necessary conditions for a Pareto optimum, but (2) that cardinal, interpersonally comparable arguments are required to select a single, best point from among the infinity of Pareto optima.

23.2 Axiomatic social welfare functions

Kemp and Ng (1976) and Parks (1976) prove their impossibility theorems by demonstrating that it is impossible to have an SWF that satisfies a particular set of axioms, which among other things imply that the arguments of the function are ordinal utility functions. Their theorems naturally raise the question of the sorts of axioms we need to impose to obtain a *reasonable* SWF. In this section we review some of the answers that have been given to this question.

23.2.1 *Fleming's social welfare function*

The pioneering axiomatic treatment of SWFs was by Fleming (1952). Fleming proved that any SWF satisfying the Pareto principle and the elimination of indifferent individuals axiom (EII) must be of the following form:

$$W = f_1(U_1) + f_2(U_2) + \cdots + f_s(U_s). \tag{23.10}$$

Elimination of indifferent individuals axiom: *Given at least three individuals, suppose that i and j are indifferent between x and x′, and between y and y′, but i prefers x to y, and j prefers y to x. Suppose that all other individuals*

[2] See Samuelson (1967, 1977, 1981). Samuelson also attributes this position to Bergson (Samuelson, 1967, pp. 44–5, 48–9), but see my discussion in n. 1.

are indifferent between x and y, and x' and y' (but not necessarily between x and x', and y and y'). Then social preferences must always go in the same way between x and y as they do between x' and y'. (Name and statement follow Ng's (1981b) simpler presentation.)

EII has two important properties. First, as its name implies, it does eliminate individuals who are indifferent between x and y. Second, it requires that whatever convention is used to decide whether i's preferences regarding x and y override j's, it must also decide the pair (x', y') given i and j's indifference between x and x', and y and y'. One sort of convention for deciding whose preferences are overriding would, of course, be to make one person a dictator. An alternative convention would be to posit interpersonally comparable cardinal utility functions for i and j.

The value of W in (23.10) is obviously independent of the ordering of individuals in the $1, s$ sequence, and so the theorem satisfies the anonymity axiom. But the theorem does not tell us much about the functional form of W. In particular, if

$$f_i(U_i) = a_i U_i, \tag{23.11}$$

then (23.10) becomes an additive W. If

$$f_i(U_i) = log(U_i), \tag{23.12}$$

we have essentially a multiplicative W.[3] To specify the SWF more precisely we need additional axioms.

23.2.2 *Harsanyi's social welfare function*

Harsanyi (1953, 1955, 1977) derives an SWF from the following three assumptions:

1. Individual personal preferences satisfy the von Neumann–Morgenstern–Marschak axioms of choice involving risk.
2. Individual ethical preferences satisfy the same axioms.
3. If two prospects P and Q are indifferent from the standpoint of every individual, they are indifferent from a social standpoint.

An individual's personal preferences are those he uses when making his day-to-day decisions. His ethical preferences are used on those more seldom occasions when he makes moral or ethical choices. In making the latter decisions, the individual must weigh the consequences of a given decision on other individuals, and thus must engage in interpersonal utility comparisons.

[3] The summation of the logs of the U_is equals the log of their product. Thus the transformation given in (23.12) makes W equal to the log of the product of the individuals' utilities. Since log(x) obtains a maximum when x does, both a W defined as the product of s individuals' utility functions and a W defined as the log of this product will carry the same implications for the optimal values of the arguments of the individual utility functions.

From these three postulates Harsanyi proves the following theorem concerning the form of the SWF, W:

Theorem: *W is a weighted sum of the individual utilities of the form*

$$W = a_1 U_1 + a_2 U_2 + \cdots + a_s U_s, \tag{23.13}$$

where a_i stands for the value that W takes when $U_j = 0$, for all $j \neq i$ (Harsanyi, 1955, p.52).

This is clearly a rather powerful result given the three postulates. As always, when powerful results follow from seemingly weak premises one must reexamine these premises to see whether they perhaps contain a wolf in disguise.

The first assumption simply guarantees a form of individual rationality in the face of risk and seems innocuous as such. When deciding whether to go to the beach or stay home, the rational individual first computes his expected utility from being at the beach. If π_r is the probability that it will rain and π_s is the probability that the sun shines, and U_r and U_s are her utilities in these two states of the world, then her expected utility from being at the beach is $U_B = \pi_r U_r + \pi_s U_s$. The rational individual goes to the beach if this expected utility exceeds the (let us assume) certain utility from staying at home.

The second assumption extends the concept of rationality in the face of risk from the individual's personal preferences to her ethical ones. When making a decision about whether to give \$100 to a poor person, the rational, ethical individual envisages the utility that she would experience if she were a poor person and received \$100, and the utility she would experience if she had \$100 less, and places the appropriate probabilities on each state of the world. The assumption that an individual's preferences satisfy the von Neumann-Morgenstern-Marschak axioms of choice induces the ethical person to add individual utilities when making ethical choices.

Harsanyi's second assumption can be criticized as an illegitimate extension of the notion of individual rationality to social choices. Pattanaik (1968) made this criticism of Harsanyi's SWF and Buchanan (1954a) made a similar criticism of Arrow's SWF, which we shall take up in the next chapter. But this objection seems to carry less weight against Harsanyi than against Arrow. Harsanyi is assuming *individual* evaluations of different social states in both cases; no aggregate will or organic being is even implicitly involved as arguably is the case with Arrow's SWF. The W in Harsanyi's theory is a subjective W in the mind of the individual. If individuals differ in their subjective evaluations, there will be different Ws for different individuals. A collective W need not exist.

Under the assumption that individuals make decisions involving risk by maximizing the expected value of their subjective utilities, Ng (1984a) has established an equivalence between von Neumann-Morgenstern utility indexes and subjective utility indexes. Thus, Harsanyi's first two assumptions effectively introduce interpersonally comparable, cardinal utilities into the SWF.[4]

[4] See also Binmore (1994, ch. 4).

The third postulate introduces the individualistic values that underlie Harsanyi's SWF. What is remarkable about Harsanyi's theorem is that he has been able to derive the intuitively plausible additive SWF from these three rather modest looking sets of assumptions.

Knowing that the SWF is additive is only the first, even though large, step in determining the optimal social outcome, however. The weights to be placed on each individual's utility index must be decided, and the utility indexes themselves must be evaluated. It is here that Harsanyi derives the ethical foundation for his SWF. He suggests that each individual evaluate the SWF at each possible state of the world by placing himself in the position of every other individual and mentally adopting their preferences. To make a selection of a state of the world impartial, each individual is to assume that he has an equal probability of being any other person in society (Harsanyi, 1955, p. 54).

The selection of a state of the world is to be a lottery with each individual's utility – evaluated using her own preferences – having an equal probability. "This implies, however, without any additional ethical postulates that an individual's impersonal preferences, if they are rational, must satisfy Marschak's axioms and consequently must define a cardinal social welfare function equal to the arithmetic mean of the utilities of all individuals in society" (Harsanyi, 1955, p. 55). Thus, the *Gedankenexperiment* of assuming that one has an equal probability of possessing both the tastes and position of every other person solves both of our problems. The utility functions are evaluated using each individual's own subjective preferences, and the weights assigned to each, the a_i, are all equal. The SWF can be written simply as the sum of all individual utilities:

$$W = U_1 + U_2 + \cdots + U_s. \tag{23.14}$$

Of course, there are serious practical problems of getting people to engage in this form of mental experiment of evaluating states of the world using other individuals' subjective preferences, and Harsanyi (1955, pp. 55–9; 1977, pp. 57–60) is aware of them. Nevertheless, he holds the view that with enough knowledge of other individuals, people could mentally adopt the preferences of others, and the U_i terms in each individual's evaluation of social welfare would converge. The mental experiment of adopting other individuals' preferences combined with the equiprobability assumption would lead all individuals to arrive at the same, impartial SWF (Harsanyi, 1955, p. 59). Later both Rawls (1971) and Buchanan and Tullock (1962) would introduce uncertainty over future position to bring about unanimous agreement over a social contract and a constitution, respectively. Their work is discussed in Chapters 25 and 26.

23.2.3 *Two criticisms of Harsanyi's social welfare function*

23.2.3.1 *Should individual attitudes toward risk count?* Writing before Harsanyi derived his SWF, but in clear anticipation that the then newly invented von Neumann-

Morgenstern utility indexes would be used to create an SWF, Arrow (1951, 2nd ed. 1963, pp. 8–11) raised the following objection against this use of them:

> This [the von Neumann-Morgenstern theorem] is a very useful matter from the point of view of developing descriptive economic theory of behavior in the presence of random events, but it has nothing to do with welfare considerations, particularly if we are interested primarily in making a social choice among alternative policies in which no random elements enter. To say otherwise would be to assert that the distribution of social income is to be governed by the tastes of individuals for gambling. (Arrow, 1963, p. 10)

More generally, as Sen (1970a, p. 97) notes, the use of the von Neumann-Morgenstern axioms introduces a degree of arbitrariness that is inherent in all cardinalization of utilities.

Whether social choices *should* depend on individual attitudes toward risk is a knotty question. If Jane's attitude toward risk affects her decision about whether to go to the beach or not, then presumably her attitude toward risk may also affect her willingness to give to the poor, assuming that she makes this choice after engaging in the kind of mental experiment that Harsanyi described. Conceivably her attitude toward risk might also affect how she votes on redistribution legislation. To say "that the distribution of the social income" *should not* be governed by such tastes would assume that the preferences of individuals formed in this way should not count. The specter of a "social planer" deciding what the distribution of the social income should be using the "proper" preferences as given in *his or her* SWF arises.

More generally, once we decide that an individual's attitudes toward risk should not count, we must inquire what other preferences of hers should not count – for pornography, for education? Here the conflict between the elitist view of social choice as represented by the social planer choosing social outcomes using an SWF, and the individualistic view of social choice as the outcome of a voting process in which each individual's preferences are counted becomes apparent.

The knowledge that Jane would pay X for a p probability of winning Y tells us something about her preferences for X and Y, just as the knowledge that she prefers Y to X does. The former knowledge actually contains more information than the latter, and this information does not seem a priori inherently inferior to knowledge of simple preference orderings. At least the inferiority of the former sort of information would seem to require further justification.[5]

23.2.3.2 *Can individuals agree on a value for W?* The dependence of the individually determined Ws on individual risk preferences has led both Pattanaik (1968) and Sen (1970a, pp. 141–6) to question whether individuals who engaged in Harsanyi's equiprobability experiment would unanimously agree on which state of the world maximizes W.

[5] For additional criticism and discussion of the role of risk preferences in the Harsanyi SWF, see Diamond (1967), Pattanaik (1968), and Sen (1970a, pp. 143–5). For a defense of the use of von Neumann-Morgenstern utilities in social choice analysis, see Binmore (1994, pp. 51–4, 259–99).

Table 23.1. *Outcomes in dollars*

State of the world	T	W
Person		
R	60	100
P	40	10

To see the problem, consider the following example. Let there be two individuals in the community, rich (R) and poor (P), and two possible states of the world, with a progressive tax (T) and without one (W). Table 23.1 gives the possible outcomes in dollar incomes.

In Table 23.2 the von Neumann-Morgenstern utilities for each outcome are presented, scaled in such a way as to make them interpersonally comparable. R is assumed to have constant marginal utility of income; P diminishing marginal utility. If each individual now assumes that he has an equal probability of being R or P in either state of the world, then the von Neumann-Morgenstern postulates of rationality dictate the following evaluation of the two possible states:

$$W_T = 0.5(0.6) + 0.5(0.4) = 0.5$$

$$W_W = 0.5(1.0) + 0.5(0.2) = 0.6.$$

The state of the world without the progressive tax provides the highest expected utility and would, according to Harsanyi, be selected by all impartial individuals. But, reply Pattanaik and Sen, P might easily object. He is clearly much worse off under W than T and experiences a doubling of utility in shifting to T, while R loses less than 1/2. The utility indexes in Table 23.2 reveal P to be risk averse. Given a choice, he might refuse to engage in a fair gamble of having R or P's utility levels under T and W, just as a risk-averse person refuses actuarially fair gambles with monetary prizes. Although the Harsanyi SWF incorporates each individual's risk aversion into the evaluations of the U_i, it does not allow for differences in risk aversion among the impartial observers who determine the SWF values. If they differ in their preferences toward risk, so too will their evaluations of social welfare under the possible states of the world, and unanimous agreement on the SWF will not be possible (Pattanaik, 1968).

The Pattanaik-Sen critique basically challenges Harsanyi's assumption that the von Neumann-Morgenstern-Marschak axioms can reasonably be assumed to hold for the ethical choices individuals make when uncertain of their future positions. In defense of postulating that these axioms hold at this stage of the analysis, one

Table 23.2. *Outcomes in utility units*

State of the world	T	W
Person		
R	0.6	1.0
P	0.4	0.2

Table 23.3. *Outcomes in utility units (second round of averaging)*

State of the world	T	W
Person		
N	0.5	0.6
A	0.44	0.42

can reiterate that the utilities that make up the arguments of W *already* reflect individual attitudes toward risk. To argue that special allowance for risk aversion must be made in determining W from the individual U_is is to insist that social outcomes be discounted twice for risk, a position that requires its own defense (Harsanyi, 1975b; Ng, 1984a).

An alternative response to the Pattanaik-Sen critique is to extend the logic of Harsanyi's mental experiment and assume that each individual uses not his own risk preferences, but that he assumes that he has an equal probability of having the risk preferences of every other individual. Suppose that in our example one individual was risk neutral (N) and the other risk averse (A). Their evaluations of the alternative states of the world might then look something like the figures in Table 23.3.

The elements of row N represent the simple expected values of states T and W occurring, assuming that an individual has the same probability of being R or P and is risk neutral. Row A presents the lower evaluations that a risk-averse person might place on the possible outcomes. The social welfare levels under these two states of the world, assuming that each individual has an equal probability of being rich and poor *and of being risk averse or risk neutral*, would then be

$$W_T = 0.5(0.5) + 0.5(0.44) = 0.47$$

$$W_W = 0.5(0.6) + 0.5(0.42) = 0.51.$$

The state of the world without the tax is again preferred, although by a narrower margin.

The same objection to this outcome can be raised, however, as was raised to the first. A risk-averse person will recognize that the tax alternative favoring the rich has a greater likelihood of being selected under risk-neutral preferences than under risk-averse preferences. He might then object to being forced to accept a gamble that gave him an equal chance of having risk-neutral or risk-averse preferences, in the same way that he would reject a fair gamble of experiencing the utility levels of the rich and poor. This objection can be met in the same way as the previous objection, however. Reevaluate the two states of the world assuming each individual has an equal probability of being risk neutral or risk averse using the utility levels from the previous round of averaging as this round's arguments for the utility functions. If the utility functions are smooth and convex, convergence on a single set of values for W_T and W_W can be expected.[6]

[6] Vickrey (1960, pp. 531–2) was the first to suggest repeated averaging of welfare functions to bring about consensus. Mueller (1973) and Mueller, Tollison, and Willett (1974a) have proposed using this technique as an

Here the reader may begin to feel his credulity stretching. Not only is an ethically minded citizen supposed to take on the subjective preferences of all other citizens, these preferences must be defined over both physical units (like apples and money) and the interpersonally comparable cardinal utility units of each individual, and he must be prepared to engage in a potentially infinite series of mental experiments to arrive at *the* social welfare evaluation to which all impartial citizens agree. The price of unanimity is high.

Although this type of criticism cannot be readily dismissed, it must be kept in mind that what we seek here is not a formula for evaluating social outcomes that each individual can apply to come up with a unique number. What we seek is a way of conceptualizing the problem of social choice to which we all might agree, *and* which might help us arrive at an agreement over actual social choices were we to apply the principles emerging from this form of mental experiment. The difference between the straight application of the Harsanyi SWF to a social choice problem, and a version of it modified to take into account the criticisms of Pattanaik and Sen involves simply the question of how much weight should be placed on the preferences of risk-averse individuals. For example, if one individual in the community is maximin risk averse, repeated averaging will result in the selection of the state of the world that maximizes the welfare of the worst-off individual (Mueller, Tollison, and Willett, 1974a). This is essentially the *just* social outcome, which Rawls (1971) obtains from a similar starting position as that assumed by Harsanyi, but without the use of any utility calculations.

Thus, in evaluating the "realism" of the Harsanyi approach, the issues are these:

1. Can one envisage individuals obtaining sufficient information about the positions and psychology of other individuals to allow them to engage in the interpersonal comparisons inherent in the approach?
2. Can individuals assume an impartial attitude toward all individuals in the community, and from this impartial stance agree on a set of weights (a common attitude toward risk) to be attached to the positions of each individual when making the social choice?

If for some social choices it is reasonable to assume that the answers to these two questions are both "yes," then for these choices the Harsanyi SWF can be a useful analytic construct.

23.2.4 *Ng's social welfare function*

Ng (1975) has derived an additive SWF in which the utilities of each individual are measured in "finite sensibility" units. The concept of a finite sensibility unit is built on "the recognition of the fact that human beings are not finitely discriminative"

answer to Pattanaik and Sen's objections to the Harsanyi SWF. Vickrey set up the problem of maximizing social welfare as the choice of a set of rules for a community that one is about to enter not knowing one's position in it. The setting is obviously similar to that envisaged by Harsanyi, and not surprisingly we find Vickrey arguing for a weighted summation of von Neumann-Morgenstern (or "Bernoullian") utility functions. Vickrey resorts to repeated averaging in the event that there is disagreement over the values of the weighted sums.

(p. 545). Thus, for a small enough change in x to x' an individual is indifferent between x and x' even though $x \neq x'$. Individuals are capable of perceiving changes in x only for discrete intervals in x. These discrete steps in an individual's perceptions of changes in x become the building blocks for a cardinal utility index measured in finite sensibility units. To the finite sensibility postulate Ng adds the weak majority preference criterion.

Weak majority preference criterion: *If a majority prefers x to y, and all members of the minority are indifferent between x and y, then society prefers x to y.*

The weak majority preference criterion incorporates the ethical values built into the SWF. It is obviously a combination of both the Pareto principle and the majority rule principle that is at once significantly weaker than both. In contrast to the Pareto criterion, it requires a majority to be better off, not just one person, to justify a move. And, in contrast to majority rule, it allows the majority to be decisive only against an indifferent minority. In spite of this apparent weakness, the postulate nevertheless proves strong enough to support a Benthamite SWF whose arguments are unweighted individual utilities measured in finite sensibility units, that is, equation (23.14). For those who dismiss Harsanyi's theorem justifying (23.14), because it introduces attitudes toward risk through the use of von Neumann-Morgenstern utility indexes, Ng's theorem offers a powerful alternative justification for the Benthamite SWF, which does not introduce risk in any way.

From the perspective of public choice, the theorems of Harsanyi and Ng are the most important justifications for the additive SWF, since their basic axioms are easily interpretable as conditions one might wish to incorporate into a set of constitutional rules, and in Harsanyi's case, the whole context in which the SWF is derived resembles the settings from which Rawls and Buchanan and Tullock derive their social contract and constitution, respectively. In Chapter 26 we shall analyze the differences and similarities in the three approaches.

23.2.5 *Nash's and other multiplicative social welfare functions*

Where the additive SWF is most often associated with the name Jeremy Bentham, the multiplicative SWF is most often associated with the name John Nash. Nash's (1950) objective, however, was not to derive an SWF, but rather to come up with a solution to a two-person "bargaining problem." When generalized to s persons, however, Nash's solution to bargaining problems can be regarded as a multiplicative SWF (Luce and Raiffa, 1957, pp. 349–50).

$$W = (U_1 - U_1^*)(U_2 - U_2^*) \cdots (U_s - U_s^*). \tag{23.15}$$

The utilities that go into the welfare function are defined relative to a status quo point at which $U_i = U_i^*$ for all i. This formulation is natural for the bargaining problem that Nash first addressed. Should a bargain not be reached, the status quo is the outcome of the game. All gains from the bargain are measured relative to this status quo starting point.

The axioms needed to derive the Nash SWF are few and rather innocuous. The utility functions must, of course, be cardinal, and the Pareto principle, an α-contraction property and a symmetry condition, must also be satisfied.

Property α: *If x is a member of the choice set defined over the full set of alternatives S, then x is a member of the choice set of any proper subset of S of which it is a member* (Sen, 1969).

Symmetry: *If an abstract version of a bargaining game places the players in completely symmetric roles, the arbitrated value shall yield them equal utility payoffs, where utility is measured in units which make the game symmetric* (Luce and Raiffa, 1957, p. 127).

Nash's solution to the bargaining problem was put forward more as a description of the outcome of a game than as a prescription as to what the outcome ought to be. On the other hand, Nash does argue that the outcome is fair, and it is because of the inherent fairness of the outcome, which should be apparent to both sides, that one expects the solution satisfying (23.15) to emerge (Luce and Raiffa, 1957, pp. 128–32).

However, the delimitation of the gains to be shared is sensitive to the choice of the status quo point. The important role played by the status quo in the Nash SWF has led to its criticism as a normative construct by Sen (1970a, pp. 118–21). If bargaining on social choices takes place, given market-determined income and wealth and presently defined property rights, then the scope for alleviating current inequities through collective action will be greatly restricted.

On the other hand, conceptualizing the problem of selecting a set of rules to govern the political game as a "bargaining problem" does seem to be a reasonable way to view the writing of a constitution or a social contract by individuals *who are not uncertain about their future preferences and/or positions*. Were one to think of the social contract as being the set of rules selected from a hypothetical or real state of anarchy, then the status quo point would be the "natural distribution" of property that would exist under anarchy (Bush, 1972; Buchanan, 1975a). The gains from cooperation would then be enormous and a rather egalitarian sharing of these gains as implied by the Nash SWF might indeed be deemed fair, as Nash thought it would be.

Viewing the status quo as the starting position from a state of anarchy resembles the setup in the Kaneko and Nakamura (1979) theorem. They derive conditions for an SWF of the Nash form as in (23.15), but $(U_1^*, U_2^*, \dots, U_s^*)$ is defined not as the status quo, but as the worst possible state for each individual that we can imagine. It is doubtful if modern man were thrust into true anarchy that his utility would be much higher than that envisaged by Kaneko and Nakamura. As with all of the SWFs that we have been considering, the Kaneko/Nakamura SWF satisfies anonymity and the Pareto postulate. They also assume a form of the independence of irrelevant alternatives axiom, which we shall examine at length in the next chapter, and make the "fundamental assumption that we evaluate the social welfare by considering relative increases of individuals' welfare from the origin" (p. 426). This assumption, combined with their use of von Neumann-Morgenstern utility indexes, forces one to compare ratios of utilities across individuals rather than absolute differences and obviously goes most of the way toward requiring an SWF in multiplicative form.

The most general characterization of a multiplicative SWF is by DeMeyer and Plott (1971). They measure intensity differences as ratios of utilities (relative utilities) and go on to derive an SWF of the form

$$W = U_1^K \cdot U_2^K \cdots U_s^K, \tag{23.16}$$

where K is a real number.

23.3 What form of social welfare function is best?

We have now seen that it is possible to derive either an additive or a multiplicative SWF from a few basic axioms. In both cases we need to assume that their arguments are some form of cardinal, interpersonally comparable utility indexes if we are going to use them to select optimal states of the world, or optimal political institutions. Both types of SWF satisfy the Pareto postulate; both also satisfy an anonymity axiom. Each differs from the other, however, in some important ways with respect to their other axiomatic properties. Rather than analyze each axiom in detail, we shall close this chapter by considering some simple examples that illustrate the properties of these two different types of SWFs. We shall confine our attention to the simplest form of each.

$$W = U_1 + U_2 + \cdots + U_s \tag{23.17}$$

$$W = U_1 \cdot U_2 \cdots U_s. \tag{23.18}$$

Consider now Table 23.4. Each entry represents the cardinal, interpersonally comparable utility level of either individual i or j in the two possible states of the world G and M. These utility levels allow for any diminishing marginal utility of income, and thus i's income in state G might be 3, 4, or 10 times her income in M, even though her utility level in G is only double her utility in M. If a social choice had to be made between G and M, which state should be chosen? An additive W selects M – a multiplicative G.

Regardless of which choice the reader makes, it should be obvious that it is possible that other readers will make the opposite choice. To see this point more clearly, assume that i and j are really the same person at two different times in her life, and G and M are two alternative career paths. Path G is a job in government with somewhat lower income and utility at the start than later. Path M is a career in medicine with lower utility at the start than the government job, but much higher utility later. Given full knowledge of the utility payoffs to each career choice, it is conceivable that some rational, self-interested individuals favor the career in government,

Table 23.4.

	i	j
G	2	3
M	1	5

Table 23.5.

		Individuals				
		1	2	3	4	5
States	A	1	1	1	1	1
	B	0.0001	10,000	1	1	1
	C	0.0001	10	10	10	10

others medicine; if this is true, then some will probably prefer a multiplicative welfare function, others an additive.

As this example suggests, the choice of the multiplicative welfare function is likely to hinge on one's values with regard to how egalitarian the distribution of *utilities* ought to be. Recall that the entries in Table 23.4 are in utilities, not incomes. If the marginal utility of income declines, the differences in the utility levels i and j experience are smaller than the differences in their incomes. A choice of G over M as a state of the world (career) indicates a strong preference for egalitarian outcomes.

With a multiplicative SWF, a doubling of i's utility is offset by a halving of j's. An increase in i's utility from 100 to 200 is fully offset by a decline in j's from 100 to 50. Requiring that such trade-offs be made in the SWF has been criticized by Ng (1981b) on the grounds that it can lead some individuals to make very large sacrifices to avoid very small *absolute* declines in utility for others. Suppose, for example, that a society of five faces the choice among the three states of the world A, B, and C as in Table 23.5. In state A, all five experience a relatively modest level of welfare. In B, one is utterly miserable (almost to the point of suicide), two are ecstatic, and the other three individuals are as in state A. In C, one is again miserable, but all four of the other individuals are 10 times better off than in A. An additive welfare function ranks B above C, and places both above A. The multiplicative regards A, B, and C as socially indifferent.

Those who object to the choice of B over A argue that the use of the additive welfare function in this situation allows individual 1 to be used as a *means* to 2's gain in violation of Kant's fundamental dictum.[7] Indeed, with an additive W, a maximum could arise at which some individuals have zero or negative utilities. Killing a wealthy invalid and redistributing her property to the healthy poor could easily raise an additive W. If j were a sadist, then j's torture of i so that i has negative utility (wishes he were dead) could raise W. With a multiplicative W, no state with any $U_i \leq 0$ could ever be chosen as long as some states are feasible for which all $U_i > 0$.

As a counterargument to these examples, note that although increases in W can easily be envisaged as involving murder and torture, that maximum W would occur at these points is less plausible. If i is not a masochist, then a less costly (in terms of the interpersonally comparable U's) way of increasing U_j is probably available, than by letting j torture i.

[7] See, in particular, Rawls (1971). Sen's (1979) critique of welfarism is also relevant here. Rawls does not argue for a multiplicative welfare function, but rather a lexicographic one (setting aside his objections to utilitarianism). Rawls's theory is discussed in Chapter 25.

The same logic and arithmetic that make A and B equal with multiplicative W, make C not better than A, although here the exchange of making four people considerably better off for making one modestly worse off in absolute terms may strike some as reasonable. Note again that one could well imagine individuals making such a trade-off for themselves. If at the age of 21 the reader were given a choice between living the next 50 years at, say, the poverty line, versus living 10 of those years at the margin of existence and 40 in the affluence of the upper middle class, it is more than conceivable that the reader would make the Faustian choice for the second alternative. If these options are represented reasonably by the utility numbers in rows A and C in Table 23.5, then the reader has made the choice using a criterion that is closer to the additive than to the multiplicative SWF. If the reader would make choices such as these by implicitly adding the different utility levels, why would it be wrong for society to use the same criterion?

One possible reply to this question is to argue that, although it is perfectly acceptable for an individual to make choices by adding her utility levels at different points in time, since she is making choices for herself and may compare her utilities at different points in time any way she wants, when the welfare levels of *different* persons are to be compared, the trade-offs inherent in the additive W are unacceptable for the means-ends reason given above. A different criterion, one more protective of individual rights as in the multiplicative W, is required when one makes interpersonal welfare choices, from that which may be reasonable or acceptable for making intrapersonal choices.

This reply raises indirectly the issue of the context in which the SWF is used. Many observers seem to think of an SWF as an analytic tool to be used by a policymaker, who plugs in the U_is and then maximizes; that is, some unknown third party is making social choices *for* society. In this setting, the issues of how the U_is are measured and what trade-offs in utility are allowed across individuals are salient. Constraints on the choices that protect individuals from having their welfare lowered for the benefit of others in the community, as introduced through a multiplicative W, have much appeal.

An alternative way to view W, however, is to see it as a guide to writing the constitution, the set of rules by which the society makes collective decisions. If one views these rules as being chosen by self-interested individuals who are uncertain of the future positions they will hold when the rules are in effect, then in choosing an SWF (that is, a set of rules to implement an SWF), one is not making an interpersonal choice but rather an *intra*personal one. One is choosing a set of rules to maximize one's own welfare, given that one is uncertain about what position and utility function one will possess. In this context, an additive W would seem appropriate as a social welfare function if individual choices tend to be made by comparing differences in utility levels at different points in time.

The context in which the SWF is to be used is also relevant to the issue of whether and how cardinal utilities are to be measured. The abhorrence of economists for the concept of cardinal utility would seem to stem from a fear that some bureaucrat would go about metering and somehow combining individual utilities to reach decisions on social policies. Evidence from the psychological literature and sensitivity studies that indexes of cardinal utility can be constructed might from this perspective be viewed more with alarm than enthusiasm.

But if one views the SWF as a construct to guide an individual's choice in selecting a constitution, a choice made from behind a veil of ignorance concerning one's future position and utility function, then the issue is whether people can conceptualize being a slave and a slave owner, and compare their utilities in both roles. If they can, then choosing a set of rules to implement a W of whatever functional form is at least a hypothetical possibility. This is the setting in which Rawls (1971) and Buchanan and Tullock (1962) envisage a social contract and set of constitutional rules being chosen, and Harsanyi an SWF. It is the context in which the concept of an SWF seems most useful to the study of collective decision making. We return to these issues in Chapters 25 and 26.

Bibliographical notes

Following the pioneering papers by Parks (1976) and Kemp and Ng (1976), several papers appeared reestablishing the impossibility of a Bergson-Samuelson SWF with ordinal utility arguments, or the necessity of using cardinal, interpersonally comparable utility indexes (D'Aspremont and Gevers, 1977; Pollack, 1979; Roberts, 1980a,b,c); for a survey see Sen (1977b).

I have been of the opinion, ever since I read Bergson (1938) and Samuelson (1947, ch. 8) on SWFs, that cardinal, interpersonal utility comparisons were necessary to select a single allocation as best among those in the Pareto set. Moreover, I believe this opinion was commonly shared among welfare-public choice theorists. The papers of Kemp and Ng (1976) and Parks (1976) appeared to me to be important not so much because they brought startling new results to light, but because they proved formally what had been known or suspected for some time. I thus confess to some befuddlement at the nature and tone of the Samuelson (1977, 1981) and Kemp and Ng (1977, 1987) debate.

The seminal contributions of Harsanyi appeared in 1953 and 1955. The argument has been reviewed and alternative proofs of the theorem presented in Harsanyi (1977, ch. 4).

Sugden and Weale (1979) link their SWF theorem directly to the constitutional-contracting setting. Their theorem resembles Fleming's (1952).

Ng's (1975) original theorem reviewed here, and his subsequent elaborations thereon (1981b, 1982, 1983, 1984b, 1985a, 2000), constitute a most forceful defense of the additive SWF.

The literature on experimentally measuring utilities is reviewed in Vickrey (1960) and Ng (1975).

For axiomatic derivations of the Nash SWF, besides Nash's (1950) own, see Luce and Raiffa (1957, pp. 124–32, 349–50) and Sen (1970a, pp. 118–21, 126–8).

Section 23.3 draws heavily on Ng (1981b). See also Bergson (1938), Samuelson (1947), Little (1957), Sen (1979), and Ng (1981a).

Binmore's (1994, 1998) two-volume treatise contains a broad-ranging discussion of utility indexes, cardinal and ordinal, and their use in normative analysis.