A WELFARE THEORY FOR NON-MALEVOLENT AGENTS

Edward J. Green

Econometric Research Program
Research Memorandum No. 249

July 1979

Abstract: An abstract description of social institutions is given which generalizes the representation of an exchange economy as a cooperative game. Two conditions are stated which have the interpretation that, although agents may disagree about the relative importance of allocating resources to different uses, no agent desires injury to others as an end in itself. It is shown that, when these conditions are satisfied and when preferences and technology are smooth and convex, the set of Pareto optima and the set of equilibrium social states essentially coincide.

Econometric Research Program
Princeton University
207 Dickinson Hall
Princeton, New Jersey

1. Introduction

The first and second optimality theorems of welfare economics state respectively that, in "neoclassical environments," (a) every market equilibrium is Pareto optimal, and (b) every Pareto optimal allocation is a market equilibrium relative to some initial endowment, except in certain cases when the allocation situates some agent at the boundary of his consumption set. Several attempts have been made to extend these theorems by relaxing the assumption (part of the definition of a "neoclassical environment") that each agent's preference among social states depends only on his own consumption. In particular, S. Winter [6] has generalized (b), T. Bergstrom [1] has adapted the theory of Lindahl prices to prove a generalization of (a) and (b), for agents whose preferences satisfy a condition of non-malevolence. However, Bergstrom suggested that some Pareto optima could not be supported if pairs of individuals were allowed to make mutually beneficial gift transactions. S. Goldman [4] has constructed an example of such an unsupportable optimum. In order for the optimum of his example to be maintained as an equilibrium, coalitions must be subjected to a constraint which differs markedly from those imposed by market institutions.

In a neoclassical economy, a market equilibrium is both (c) an allocation decentralizable by a price vector, and (d) an allocation reached by the actions of coalitions operating in a market institution. Bergstrom's results are the generalization of the welfare theorems (a) and (b) which is obtained by considering allocations which satisfy (c). In the present paper, an alternative generalization will be obtained by considering allocations (or more generally, social states) which satisfy (d). Corresponding to Bergstrom's use of Lindahl prices as a generalization of market prices, the representation of an exchange economy as a cooperative game will be used here to generalize the notion of a market institution. Examples like Goldman's show that Bergstrom's criterion for benevolence must be tightened if this generalization is to be successful. This will be done by adding a condition which closely resembles one introduced by Bergstrom [2] in work closely related to that just discussed.

2. Societies and environments

In this section, the theory of neoclassical exchange economies will be generalized to a formal theory of societies. A society will be specified by a vector $S = \langle N,X,P,M \rangle$. $N = \{1,\dots,n\}$ is a set of agents. X is a set of social states. A social state is the generalization of an allocation or state of an economy (cf.[3 ,p.75]). $P = \langle P_1,\dots P_n \rangle$ is an assignment of preferences over X to agents. Each $P_i \subseteq X^2$ is an irreflexive relation. For $x \in X$, $y \in X$, $i \in N$ and $C \subseteq N$, xP_iy will be interpreted to mean that agent i strictly prefers state x to state y , and xP_Cy will abbreviate $\exists i \in C\ xP_iy$ and $\forall j \in C[xP_jy$ or not $yP_jx]$.

If S is an exchange economy, then every state x is an assignment $\langle x_1,\dots,x_n \rangle$ of consumption bundles to agents. Agent i is called individualistic if, whenever $x_i = x_i'$ and $y_i = y_i'$, xP_iy and $x'P_iy'$ are equivalent.

A coalition is a non-empty subset of N . An institution M is a prescription of authority to various coalitions to make specific changes of social state. Formally, $M \subseteq (2^N - \{\phi\}) \times X^2$, and $\langle C,y,x \rangle \in M$ will be interpreted to mean that coalition C is authorized to change the social state from x to y . In particular, if S is an exchange economy, a coalition is allowed to make any transaction which does not alter the consumption of non-members. Formally, the elements of M are all triples $\langle C,y,x \rangle$ such that

$\forall i \notin C \;\; x_i = y_i$. Such an institution will be called a market.

Besides the institutional constraints on coalitions, there are constraints of technical feasibility on the society as a whole. These constraints are specified by the environment $E = \langle A, \bar{x} \rangle$. $A \subseteq X$ is the set of attainable states, and $\bar{x}$ is the initial state. In an exchange economy, $\bar{x}$ is the endowment and $A = \{x \epsilon X | \sum_{i \epsilon N} x_i = \sum_{i \epsilon N} \bar{x}_i\}$.

If A is a set of attainable states for a society S , a state x is Pareto-optimal in A if (a) $x \epsilon A$, and (b) there is no $y \epsilon A$ for which $y P_n x$.

3. Finality and the first welfare theorem for non-imposed societies

In a society S , the institution M and an environment $E = \langle A, \bar{x} \rangle$ naturally determine a cooperative game without side-payments. Specifically, coalition C can obtain state x if $\langle C, x, \bar{x} \rangle \; \epsilon \; M$ and $x \; \epsilon \; A$. Coalition C can improve x if, for some y , C can obtain y and $y P_C x$. State x is in the core of E if (a) some coalition can obtain x , and (b) no coalition can improve x .

If A is a set of attainable states, a state x will be called final in A if $x \epsilon A$ and x is in the core of $\langle A, x \rangle$. Finality is the equilibrium concept to be studied in this paper. That a state is final means that, once it has been reached by the activity of coalitions, no coalition can benefit (asuming that its members do not act strategically to manipulate other coalitions) from further recourse to the institution. The close connection between the core and the Walras equilibria of an exchange economy with individualistic agents makes finality an appropriate equilibrium concept to use in a reformulation of the welfare theorems.

In this setting, the first welfare theorem holds for every society having an institution which allows any change of social state to be made by unanimous consent. A society with such an institution is called non-imposed. Formally, S is non-imposed if $\{N\} \times X^2 \subseteq M$ ..

Theorem 1: If S is non-imposed, and $A \subseteq X$ is any set of attainable states, then every final state in A is Pareto-optimal in A .

Proof: Suppose $x \in A$, $y \in A$, and $yP_N x$. Since S is non-imposed, $\langle N,y,x\rangle \in M$. Thus y improves x for N , so x is not in the core of the environment $\langle A,x\rangle$. I.e., x is not final in A . Q.E.D.

4. Societies for which the second welfare theorem holds

In this section, instances of the second welfare theorem for two types of society will be proven. The first instance of the theorem will apply to generalized exchange economies with individualistic agents. The second instance will apply to non-imposed societies in which any departure from the status quo must be made by unanimous consent. A striking distinction between the results of this section and the usual welfare theorem for Walras equilibrium is that, because the use of finality as an equilibrium concept abstracts from the issue of decentralizability, no convexity assumptions are required here.

The notions of exchange economy and individualistic agent have already been defined in section 2. A generalized exchange economy will satisfy all the conditions defining an exchange economy, except that the attainable states may be any subset of X . Intuitively, one may think of a generalized exchange economy as being a description of those aspects of a production economy which do not pertain to individual firms.

<u>Proposition 1</u>: If S is a generalized exchange economy with individualistic agents and $A \subseteq X$, then every Pareto-optimal state in A is final in A.

<u>Proof</u>: Suppose that $x \in A$ and x is not final in A. Specifically, let y improve x for some coalition C. I.e., $y \in A$, yP_Cx, and $\langle C,y,x\rangle \in M$. Since M is a market, $y_i = x_i$ for all $i \notin C$. Since agents are individualistic and P_i is irreflexive for all i, not xP_iy for any $i \notin C$. Therefore yP_Nx, so x is not Pareto-optimal in A. Q.E.D.

Many economists and political theorists have observed that, if an institution permitted a change of state proposed by any agent and not vetoed by any other agent, then a Pareto-optimal state unanimously weakly preferred to the initial state would ultimately be reached in any environment. This insight may be formalized conveniently within the theory presented here. Let $D = \{\langle x,x\rangle | x \in X\}$, and define M to be a <u>unanimous consent institution</u> if $M = (\{N\} \times X^2) \cup ((2^N - \{\phi,N\}) \times D)$. Note that, in a society having such an institution, the initial state is the only state which a proper sub-coalition of N can obtain in any environment.

<u>Proposition 2</u>: If S is a society having a unanimous consent institution M, and $A \subseteq X$, then every Pareto-optimal state in A is final in A.

<u>Proof</u>: Suppose that $x \in A$ and x is not final in A. Specifically, let $y \in A$, $y P_C x$, and $\langle C,y,x\rangle \in M$. Since P_C is irreflexive, $\langle y,x\rangle \notin D$. Therefore, since M is a unanimous consent institution, $C = N$. Since yP_Nx, x is not Pareto-optimal in A. Q.E.D.

There is a great difference between markets and unanimous consent institutions. Markets are introduced in a formal theory of societies for two reasons. First, a market as defined here is an idealization of a concrete institution which plays an important role in determining the equilibrium states of actual societies. Second, a market is defined here in terms of a characteristic which is shared by many other social institutions, that each coalition is allowed to determine those aspects of the social state which are of direct concern only to its members. The fact that a market is characterized in terms of this feature, rather than in terms of features which distinguish actual markets from other actual institutions, makes it plausible that theoretical results about markets may be valid with respect to social organization in general.

In contrast, although unanimous consent institutions satisfy the formal definition of an institution, it is not suggested that any actual society could operate without granting considerable autonomy to individuals and to small coalitions. Thus proposition 2 does not convincingly refute the conjecture that the second welfare theorem fails for societies in which agents are not individualistic. For a convincing refutation, it is necessary that the theorem be established for a class of non-individualistic societies having a market or a closely related institution. To do this is the goal of the remaining sections of the paper.

5. Exchange economies for which the second welfare theorem fails

It was shown in the proof of proposition 1 that, whenever any coalition in an exchange economy with individualistic agents can obtain an improvement, the change of state is in fact a Pareto-improvement. It is evident that this condition will not hold in general for societies in which some agents are malevolently disposed toward others. Surprisingly, the condition need not hold either for societies in which agents' preferences might seem to be non-malevolent. Two characterizations of non-malevolence which have been used in the economics literature will be stated in this section, and for each characterization a counterexample to the second welfare theorem will be constructed. Each of the counterexamples is an exchange economy with a single good (so that the only exchanges are voluntary transfers) which has a non-final Pareto-optimal allocation in which every agent has strictly positive consumption. Thus, the allocation is not a boundary optimum which the minimum-wealth constraint might prevent from being a Walras equilibrium if the agents were individualistic(cf. G. Debreu [3, p.96]).

There are two natural criteria for non-malevolence. First, a non-malevolent agent ought to weakly prefer an allocation which provides better consumption bundles both for himself and for others. Second, a non-malevolent agent ought to weakly prefer an allocation which provides him with a better consumption bundle and which is weakly preferred by all other agents. These two criteria will be called hedonic non-malevolence and libertarian non-malevolence, respectively.

Note that these criteria would be trivial if the term "better consumption bundle" were defined in terms of agents' preferences. A concept of individual hedonic welfare similar to that used by the classical utilitarians is needed. Formally, an <u>ascription of welfare</u> to a society S is an assignment $W = \langle W_1, \ldots, W_n \rangle$ of irreflexive orderings $W_i \subseteq X^2$ to agents in N . Agent i displays <u>hedonic non-malevolence</u> relative to W if, whenever yP_ix , yW_jx for some $j \in N$. Agent i displays <u>libertarian non-malevolence</u> relative to W if, whenever yP_ix , either yW_ix or yP_jx for some $j \neq i$.

The informal descriptions of non-malevolence in terms of consumption bundles imply a restriction on W .An ascription of welfare to a generalized exchange economy will be called individualistic if each agent's welfare depends only on his own consumption (i.e., if the agents of the generalized exchange economy $<N,X,W,M>$ are individualistic). Note that, if an ascription of welfare is individualistic, the market institution does not permit members of a coalition to reduce the welfare of non-members. Formally, if $<C.y,x> \varepsilon M$, then not $xW_i y$ for any $i \notin C$. Whenever this formal condition is satisfied by an arbitrary society S (not necessarily an economy) and welfare ascription W , it will be said that S respects W . The main theorem of this paper will state that if a society is welfare-respecting and if agents display both hedonic and libertarian non-malevolence relative to the welfare ascription, then the society satisfies the second welfare theorem if standard convexity and non-satiation conditions are met.

Examples are now presented to show that neither of the non-malevolence criteria is sufficient without the other to guarantee the second welfare theorem. To make these examples easier to follow, preference and welfare relations will be represented by real-valued utility functions. For each agent i , functions $p^i:X \to R$ and $w^i:X \to R$ will be specified, and P_i and W_i will be taken to be $\{<y,x>|p^i(x)<p^i(y)\}$ and $\{<y,x>|w^i(x)<w^i(y)\}$, respectively. For agent i to display hedonic non-malevolence relative to W , it is sufficient that a differentiable real-valued function F^i be defined on a convex region of R^n containing all points $<w^1(x),\ldots,w^n(x)>$ for $x \varepsilon X$, and that this function and its partial derivatives F^i_j satisfy

(1) $$p^i(x) = F^i(w^1(x),\ldots,w^n(x))$$

for all $x \varepsilon X$, and

(2) $$0 \leq F^i_j$$

for $1 \leq j \leq n$, everywhere in the domain of F^i . For agent i to display libertarian non-malevolence relative to W, it is sufficient that there exist a differentiable real-valued function G^i defined on a convex region of R^n containing all points $\langle w^i(x), p^1(x), \ldots, p^{i-1}(x), p^{i+1}(x), \ldots, p^n(x)\rangle$ for $x \in X$, and that this function and its partial derivatives G^i_j satisfy

(3) $$p^i(x) = G^i(w^i(x), p^1(x), \ldots, p^{i-1}(x), p^{i+1}(x), \ldots, p^n(x))$$

for all $x \in X$, and

(4) $$0 \leq G^i_j$$

for $1 \leq j \leq n$, everywhere in the domain of G^i.

<u>Example 1</u>: Let S be an exchange economy with three agents and a single good. I.e., $N = \{1,2,3\}$, X is the non-negative orthant of R^3. Let agents' preferences be represented by the functions

(5) $$p^1(x) = \ln(x_1 + \tfrac{1}{2}) + 2\ln(x_2 + \tfrac{1}{2}) + \ln(x_3 + \tfrac{1}{2})$$

and

(6) $$p^2(x) = p^3(x) = \ln(x_2 + \tfrac{1}{2}) + \ln(x_3 + \tfrac{1}{2}) .$$

Consider the ascription of individualistic welfare represented by the functions

(7) $$w^i(x) = \ln(x_i + \tfrac{1}{2}) \qquad i = 1,2,3 .$$

That all agents display hedonic non-malevolence relative to this welfare ascription is seen by verifying (1) and (2) for the functions

(8) $$F^1(z) = z_1 + 2z_2 + z_3$$

and

(9) $$F^2(z) = F^3(z) = z_2 + z_3 .$$

Consider the set of attainable states $A = \{x \epsilon X | x_1 + x_2 + x_3 = 1\}$. Define states x and y by $x_1 = x_3 = 1/8$, $x_2 = 3/4$, $y_1 = 1/8$, $y_2 = y_3 = 7/16$. State x is the unique state at which p^1 achieves its maximum in A, so x is Pareto-optimal in A. State y improves x for the coalition $\{2,3\}$, though, so x is not final in A.

Example 2: Let S be an exchange economy with four agents and a single good. I.e., $N = \{1,2,3,4\}$, X is the non-negative orthant of R^4. Consider the ascription of individualistic welfare represented by the function

(10) $$w^i(x) = x_i^2 \qquad i = 1,2,3,4 .$$

Libertarian non-malevolent preferences will be constructed by specifying functions G^i which satisfy (4), and by solving for utility functions p^i which will satisfy (3). In particular, consider the functions

(11) $$G^1(z) = z_1 + z_2 + z_3$$

(12) $$G^2(z) = z_1 + 2z_2$$

(13) $$G^3(z) = z_1 + 2z_4$$

(14) $$G^4(z) = z_1$$

By (3), (11), and (12),

(15) $$p^1(x) = - [x_1^2 + x_2^2 + p^3(x)]$$

By (3), (13), and (14),

(16) $$p^3(x) = x_3^2 + 2x_4^2$$

and

(17) $$p^4(x) = x_4^2$$

Therefore, by (15) and (16),

(18) $$p^1(x) = -\ [x_1^2 + x_2^2 + x_3^2 + 2\ x_4^2]$$

and by (18), (3), and (12),

(19) $$p^2(x) = -\ [2x_1^2 + x_2^2 + 2x_3^2 + 4x_4^2]$$

Note that, although all agents display libertarian non-malevolence relative to the welfare ascription, agents 1 and 2 spectacularly fail to display hedonic non-malevolence. With such perverse agents, it is not surprising to find a Pareto-optimal state which is not final.

Consider, in particular, the set of attainable states $A = \{x \epsilon X | x_1 + x_2 + x_3 + x_4 = 1\}$. Define states x and y by $x_1 = x_2 = x_3 = 2/7$, $x_4 = 1/7$, $y_1 = y_2 = 2/7$, $y_3 = 0$, $y_4 = 3/7$. State x is the unique state at which p^1 achieves its maximum value in A , so x is Pareto-optimal in A . State y improves x for the coalition $\{3,4\}$, though, so x is not final in A .

6. <u>Non-malevolence with linear preferences and welfare</u>

Since both the hedonic and libertarian criteria are reasonable to impose as necessary conditions for an agent to be considered non-malevolent, it makes sense to impose both criteria jointly. An agent will be said to be <u>non-malevolent</u> relative to a welfare ascription W if he displays both hedonic and libertarian non-malevolence relative to W . Under some technical assumptions, the second welfare theorem holds of a society which respects a welfare ascription relative to which its agents are non-malevolent. This will be proven as was proposition 1, by showing that a state which some coalition can improve is not Pareto-optimal.

Heuristically, let S respect W, let preferences and welfare be represented by utility functions as in the last section, and suppose that y improves x for C, a proper sub-coalition of N. Without loss of generality, let $C = \{k+1,\ldots n\}$. If S respects W, then y improves x for C implies that

(20) $$w^i(x) \leq w^i(y)$$

for $i \leq k$ (because $\langle C,y,x\rangle \varepsilon M$ implies not xW_iy for $i \notin C$), and

(21) $$p^i(x) \leq p^i(y)$$

for $k < i$ (because yP_Cx). For $i \notin C$, define i to display C-non-malevolence relative to W if there exists a differentiable function H^{iC} defined on a convex region of R^n containing $\{\langle w^1(x),\ldots,w^k(x),p^{k+1}(x),\ldots p^n(x)\rangle | x\varepsilon X\}$ which satisfies

(22) $$p^i(x) = H^{iC}(w^1(x),\ldots,w^k(x),p^{k+1}(x),\ldots p^n(x))$$

and

(23) $$0 \leq H_j^{iC}(x)$$

for $1 \leq j \leq n$. Equations (20) - (23) imply that $p^i(x) \leq p^i(y)$. Thus if all agents outside C are C-non-malevolent, yP_Cx and x is not Pareto-optimal. If i is C-non-malevolent for all i and C with $i \notin C$, then, the second welfare theorem holds of S. Note that, when $i \notin C$, H^{iC} (which takes as arguments w^i, w^j for some other agents, and p^j for the remaining agents) is in a sense intermediate between F^i (which takes as arguments w^i and w^j for all other agents) and G^i (which takes as arguments w^i and p^j for all other agents).

A reasonable conjecture is that if agents are non-malevolent, the non-negativity conditions (2) and (4) will imply the non-negativity conditions (23) for the agents in N . This conjecture is easy to prove if the functions p^i and w^i are linear, and it implies a preliminary version of the second welfare theorem in that case:

Theorem 2: Let a non-imposed society S and a welfare ascription W satisfy the following conditions:

(a) $X \subseteq R^q$ for some natural number q , and X is open .

(b) For every $i \in N$, there is a non-zero $p^i \in R^q$ such that
$P_i = \{<y,x> | 0 < p^i \cdot (y-x)\}$,

(c) For every $i \in N$, there is a non-zero $w^i \in R^q$ such that
$W_i = \{<y,x> | 0 < w^i \cdot (y-x)\}$,

(d) The qxn matrix P^* , the i^{th} column of which is p^i [i.e.,
$P^* = (p^1 \ldots p^n)$] has rank n ,

(e) The $q \times n$ matrix W^* , the i^{th} column of which is w^i [i.e.,
$W^* = (w^1 \ldots w^n)$] has rank n ,

(f) S respects W , and

(g) Every agent in N is non-malevolent relative to W .

Then, for any set $A \subseteq X$, every Pareto-optimal state in A is final in A .

The only hypotheses of this theorem which need any comment are (a) (d) and (e). The openness requirement of (a) rules out, for instance, the specification of X as the non-negative orthant of commodity space in a generalized exchange economy. This restriction is not very serious, because the set A of attainable states is still allowed to be closed. The purpose of the openness requirement is to rule out pathological behavior of preference and welfare orderings at boundary points of A.

Conditions (d) and (e) are naturally viewed as assertions that agents are significantly different from one another, in terms of both their actual preferences and the welfare orderings ascribed to them. These conditions require that $n \leq q$ but they are innocuous when this inequality is satisfied (e.g., when S is a generalized exchange economy). Technically, the set of preference-welfare pairs (P^*, W^*) which satisfy (d) and (e) is open, dense, and of full measure in the Euclidean space of $q \times 2n$ matrices when $n \leq q$.

Four lemmas facilitate the proof of theorem 2. The first two describe algebraic consequences of non-malevolence, and the third and fourth, matrix-theoretic results.

Lemma 1: If X is open in R^q, P and W satisfy conditions (b), (c) and (e) of theorem 1, and agent j displays hedonic non-malevolence relative to W, then there exist non-negative coefficients b_{ij}, for $1 \leq i \leq n$, which satisfy

$$p^j = \sum_{i \in N} b_{ij} w^i . \tag{24}$$

Proof: There exist unique $v \in R^q$, $b_{ij} \in R$ such that $p^j = v + \sum_{i \in N} b_{ij} w^i$ and $v \perp w^i$ for all $i \in N$. Suppose first that $v \neq 0$. Choose a state $x \in X$. Since X is open, $x + tv \in X$ for some $t > 0$. $p^j \cdot ((x+tv)-x) = tp^j \cdot v = t(v + \sum_{i \in N} b_{ij} w^i) \cdot v = t\, v \cdot v > 0$, so $x + tvP_j x$ by (b). However, $w^i \cdot ((x+tv)-x) = tw^i \cdot v = 0$, so not $x + tvW_i x$ for any $i \in N$, by (c). These results contradict the assumption of hedonic non-malevolence for agent j, so $v = 0$ and $p^j = \sum_{i \in N} b_{ij} w^i$.

Now suppose that $b_{ij} < 0$ for some $i \in N$. There are unique $z \in R^q$, $a_k \in R$ for $k \neq i$, such that $w^i = z + \sum_{k \neq i} a_k w^k$ and $z \perp w^k$ for all $k \neq i$. By (e), $z \neq 0$. For some $t < 0$, $x + tz \in X$. $p^j((x+tz)-x) = tp^j \cdot z = t(b_{ij}(z + \sum_{k \neq i} a_k w^k) + \sum_{k \neq i} b_{kj} w^k) \cdot z = tb_{ij}\, z \cdot z > 0$, so $x + t\, z\, P_j\, x$ by (b). However, $w^i \cdot ((x+tz)-x) = t(z + \sum_{k \neq i} a_k w^k) \cdot z = tz \cdot z < 0$, and $w^k \cdot ((x+tz)-x) = tw^k \cdot z = 0$ for $k \neq i$, so not $x + t_z P_k x$ for any $k \in N$. Again the assumption of hedonic non-malevolence for agent j is violated, so $0 \leq b_{ij}$ for all $i \in N$. Q.E.D.

Lemma 2: If X is open in R^q , P and W satisfy conditions (b), (c) and (d) of theorem 2, and agent j displays libertarian non-malevolence relative to W , then there exist a strictly positive coefficient d_{jj} and non-negative coefficients e_{ij} for $1 \leq i \leq n$, $i \neq j$, which satisfy

$$(25) \qquad p^j = d_{jj} w^j + \sum_{i \neq j} e_{ij} p^i .$$

Proof: Closely analogous to the proof of lemma 1. Note in particular that, once it has been established that p^j lies in the subspace of R^q spanned by w^j and $\{p^i | i \neq j\}$, (d) implies that $d_{jj} \neq 0$ and that there exist $z \neq 0$ and a_k such that $w^j = z + \sum_{k \neq j} a_k p^k$, $z \perp p^k$ for $k \neq j$. Q.E.D.

Lemma 3: (L. McKenzie, [5,p.50]): If $A = (a_{ij})$ is an $n \times n$ matrix which satisfies the condition (a) $0 < a_{ij} \iff i = j$, then the conditions (b) for some $x \in R^n$, $0 < x$ and $0 < Ax$, and (c) A has a non-negative inverse (i.e., for some $n \times n$ matrix, C , $C = A^{-1}$ and $0 \leq C$) , are equivalent.

Lemma 4: If A is an $n \times n$ matrix which satisfies (a) and (c) of lemma 3,

$1 \leq k_1 < \ldots < k_m \leq n$, and B is the $m \times m$ matrix $(a_{k_i k_j})$, then B has a non-negative inverse.

Proof: By lemma 3, $0 < x$ and $0 < Ax$ for some $x \in R^n$. Define $y \in R^m$ by $y_i = x_{k_i}$. $0 < y$, and $0 < (Ax)_h = \sum_{j \leq n} a_{hj} x_j \leq \sum_{j \leq m} b_{ij} y_j = (By)_i$ for $h = k_i$ so $0 < By$. B satisfies (a) of lemma 3, so it satisfies (c) as well. Q.E.D.

Note that (24) and (25) express (1) and (3) for linear preferences, and that lemmas 1 and 2 assert (2) and (4), respectively. Lemma 4 will be used to derive the linear version of (22) and (23).

Proof of theorem 2: In addition to the coefficients defined in lemmas 1 and 3 let $d_{ij} = 0$ for $i \neq j$ and $e_{ij} = 0$ for $i = j$. Let $B = (b_{ij})$, $D = (d_{ij})$, $E = (e_{ij})$. By lemmas 1 and 2, respectively,

$$P^* = W^* B \tag{26}$$

and

$$P^* = W^* D + P^* E . \tag{27}$$

Equation (27) is equivalent to

$$P^* (I-E) = W^* D . \tag{28}$$

W^* is assumed to have rank n , and D has rank n by lemma 2. Therefore, I-E has rank n and is invertable. By (26) and (28)

$$W^* B (I-E) = W^* D . \tag{29}$$

I-E satisfies assumption (a) of lemma 3, by lemma 2. To show that (c) of lemma 3 holds also, premultiply (29) by $D^{-1} W^{*-1}$ and postmultiply by $(I-E)^{-1}$.

This yields

$$D^{-1}B = (I-E)^{-1} . \tag{30}$$

$0 \leq D^{-1}$ by lemma 2 since D is diagonal, and $0 \leq B$ by lemma 1, so $0 \leq (I-E)^{-1}$. Therefore, lemma 4 applies to $I-E$.

Consider a state x which is not final in A . It must be shown that x is not Pareto optimal in A . Since x is not final and S is non-imposed, there exist a state y and a coalition C such that $y \in A$, $yP_C x$ and $\langle C,y,x \rangle \in M$. If $C = N$, then x is not Pareto optimal. If $C \neq N$, assume without loss of generality that $C = \{k+1,\ldots,n\}$. Let $P^* = (P_1^*,P_2^*)$ and $W^* = (W_1^*,W_2^*)$, where P_1^* and W_1^* correspond to agents in $N-C$ and P_2^* and W_2^* correspond to agents in C . Let $D = \begin{bmatrix} D_{11} & 0 \\ 0 & D_{22} \end{bmatrix}$ and $E = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}$ where D_{11} and E_{11} are $k \times k$ matrices. Then by (27) ,

$$P_1^* = W^* \begin{bmatrix} D_{11} \\ 0 \end{bmatrix} + P^* \begin{bmatrix} E_{11} \\ E_{21} \end{bmatrix} = W_1^* D_{11} + P_1^* E_{11} + P_2^* E_{21} . \tag{31}$$

Equivalently,

$$P_1^* = W_1^* D_{11}(I-E_{11})^{-1} + P_2^* E_{21}(I-E_{11})^{-1} . \tag{32}$$

$0 \leq D_{11}$ and $0 \leq E_{21}$ by lemma 2, and $0 \leq (I-E_{11})^{-1}$ by lemma 4, so (32) states that there exist non-negative h_{ij} for $i \in N$, $j \notin C$, such that

$$p^j = \sum_{i \notin C} h_{ij} w^i + \sum_{i \in C} h_{ij} p^i . \tag{34}$$

$yP_C x$ implies that $0 \leq p^i \cdot (y-x)$ for $i \in C$. $\langle C,y,x \rangle \in M$ implies that $0 \leq w^i \cdot (y-x)$, since S respects W . Therefore, because $0 \leq \sum_{i \notin C} h_{ij} w^i \cdot (y-x) + \sum_{i \in C} h_{ij} p^i (y-x) = p^j \cdot (y-x)$ for $j \notin C$, $yP_N x$. Q.E.D.

7. Non-malevolence with smooth convex preferences and welfare

Hypotheses (b) and (c) of theorem 2, which require that the preference and welfare orderings of all agents be representable by linear utility functions, can be relaxed substantially. In this section, the theorem will be extended to cover a class of societies including those for which these orderings are representable by differentiable quasi-concave utility functions.

The linearity assumptions were used twice in the proof of theorem 2. First, in proving lemmas 1 and 2, linearity was used to justify several assertions such as:

$$p^j \cdot ((x+tv)-x) > 0 \text{ , so } x + tvP_j x \text{ ,}$$

and

$$w^i \cdot ((x+tv)-x) = 0 \text{ , so not } x + tvW_i x \text{ .}$$

Second, at the end of the proof of the theorem itself, linearity was used to infer not $xP_j y$ from $0 \leq p^j \cdot (y-x)$, and conversely.

In the lemmas, it was sufficient to find one non-zero t for which $x + tv$ and x would be related exactly as prescribed. Since t could be taken arbitrarily close to 0 , the lemmas could have been stated for preference and welfare orderings representable by differentiable quasi-concave utility functions. The gradient vectors of these functions at x would be used instead of the vectors p^j and w^i , and Taylor's theorem would be used to show that, for small t , the proof for the linear case remains valid. Note that, if $u: X \to R$ is differentiable and quasi-concave, $V = \{y|u(x) < u(y)\}$ for some $x \in X$, and $\nabla u(x) \neq 0$, then V is non-empty, open, and convex, x is a boundary point of V , and $\nabla u(x)/|\nabla u(x)|$ is the unique unit vector v such that $0 < v \cdot (y-x)$ for all $y \in V$. Furthermore, for any $z \in R^q$, if $0 < \nabla u(x) \cdot z$ then for some positive t , $x + rz \in V$ for all $0 < r < t$. In will now be shown that this equivalence follows directly from the properties of $\nabla u(x)$ and V just cited, without appeal to u itself. This result leads immediately to generalizations of lemmas 1 and 2.

Lemma 5: Suppose that $V \subseteq R^q$ is a non-empty, open, convex set which has x as a boundary point, and that v is the unique unit vector in R^q such that $0 < v \cdot (y-x)$ for all $y \in V$. Then, if z is any vector in R^q such that $0 < v \cdot z$, there is a strictly positive real t which satisfies $\forall r \in (0,t)\ x + rz \in V$.

Proof: Since V is open and has x as a boundary point, $x + tz \in V$ implies that $\forall r \in (0,t)\ x + rz \in V$. Thus, if the conclusion of the lemma does not hold for a vector z, V is disjoint from the convex set $Ƶ = \{x+tz \mid 0 \leq t\}$. By the separating hyperplane theorem, there is a unit vector $w \in R^q$ such that $\forall y \in V \forall u \in Ƶ\ 0 < w \cdot (y-u)$. In particular $\forall y \in V\ 0 < w \cdot (y-x)$, so $w = v$. Also, since x is a boundary point of V, $0 \leq w \cdot (x-z)$. I.e., $v \cdot (z-x) \leq 0$. Q.E.D.

A society S with welfare ascription W will be called smooth and convex at a state $x \in X$ if X is open in R^q and, for each $i \in N$, $\{y \in X \mid yP_ix\}$ and $\{y \in X \mid yW_ix\}$ are non-empty, open, and convex, and have x as a boundary point, and there are unique unit vectors p^i and w^i in R^q such that $0 < p^i \cdot (y-x)$ [resp. $0 < w^i \cdot (y-x)$] for all y such that yP_ix [resp. yW_ix]. S and W will be called regular at x if they are smooth and convex there, and if furthermore the $q \times n$ matrices $P^* = (p_1 \ldots p_n)$ and $W^* = (w_1 \ldots w_n)$ are both of rank n. Smoothness and convexity guarantee the applicability of lemma 5 to upper contour sets of orderings P_i and W_i. The additional rank conditions for regularity were used to prove lemmas 1 and 2. Generalizations of these lemmas will now be given.

Lemma 6: If S and W are regular at x, and if agent j displays hedonic non-malevolence relative to W, then there are non-negative coefficients b_{ij} such that $p^j = \sum_{i=1}^{n} b_{ij} w^i$.

Proof: As in the proof of lemma 1, if such coefficients do not exist, there exists a vector $y \varepsilon R^q$ such that $0 < p^j \cdot y$ but $w^i \cdot y \leq 0$ for all $i \varepsilon N$. By lemma 5, $x + ryP_j x$ for some $r > 0$. However, since $w^i \cdot ry \leq 0$ not $x + ryW_i x$ for any $i \varepsilon N$. This contradicts the hedonic non-malevolence of j. Q.E.D.

Lemma 7: If S and W are regular at x, and if agent j displays libertarian non-malevolence relative to W, then there exist a strictly positive coefficient d_{jj} and non-negative coefficients e_{ij} for $i \neq j$ which satisfy $p_j = d_{jj} w^j + \sum_{i \neq j} e_{ij} p^i$.

Proof: Analogous to proof of lemma 6.

The second use of linearity in the proof of theorem 2, to guarantee the equivalence of not $xP_j y$ and $0 \leq p^j \cdot (y-x)$, is less simple to replace by a convexity argument. That $0 \leq p^j \cdot (y-x)$ if not $xP_j y$ may be guaranteed in a natural way by asserting that j's indifference set through x is thin, i.e. that not $xP_j y$ implies $y \varepsilon cl(\{z | zP_j x\})$. The implication follows from continuity of the inner product, since $0 < p^j z$ if $zP_j x$. However, the converse implication, that not $xP_j y$ if $0 \leq p_j \cdot (y-x)$, may fail in a serious way. For example, let $X = R^2$ and define P_j by the utlity function $p^j(x) = x_2 - x_1^2$. Let $x = \langle 0,0 \rangle$ and $y = \langle 1,0 \rangle$. Then $\nabla p^j(x) = \langle 0,1 \rangle$, so $0 \leq \nabla p^j(x) \cdot (y-x)$, but $xP_j (1-r)x + ry$ for all $r > 0$.

Thus, when S and W are regular but not linear, $yP_C x$ and $\forall j \notin C$ not

xW_jy may not entail, as in the proof of theorem 2, that some convex combination of x and y is Pareto-preferred to x. What will be shown is that, if there is some z which all agents $j \notin C$ prefer to x, then some convex combination of x, y, and z is Pareto-preferred to x. That is, it is required that members of C be able at x to make a concession to non-members. If this concession is thought of as acceptance of a trade on unfavorable terms, the ability of an agent to make a concession seems intuitively related to the minimum wealth constraint in neoclassical theory. Although this analogy is not exact, it can be shown that there is a "large"class of Pareto-optima at which every agent can make a concession.

Formally, if $E = \langle A,x \rangle$, define y to be a concession from agent j at E if $y \in A$ and $\forall i \neq j \; yP_ix$. It is well known that, if A is convex and every agent's preference ordering P_i is represented by a quasi-concave function p^i, then (a) for every Pareto optimum x of A, there exists a unit vector $a \in R^n_+$ such that x maximizes $\sum_{i \in N} a_ip^i$ on A, and (b) if x maximizes $\sum_{i \in N} a_ip^i$ on A and $\forall i \; 0 < a_i$, then x is Pareto-optimal in A. If $a_j = 0$, clearly j cannot make a concession. It will now be shown that, except in this special case, local smoothness of the set of attainable utility profiles guarantees that every agent can make a concession.

Proposition 3: Suppose that $A \subseteq X$ is convex, that agents' preferences are representable by quasi-concave utility functions $p^i: X \to R$. Define $B = \{r \in R^n | \exists y \in A \;\; \forall i \in N r \leq p^i(y)\}$. Suppose that $x \in A$ is a Pareto-optimum, that $\forall i \; r^*_i = p^i(x)$, that there is a unique unit vector $v \in R^n$ such that $\forall r \in B \; v \cdot (r - r^*) \leq 0$, and that $\forall_i \; 0 < v_i$. Then every agent can make a concession at $\langle A,x \rangle$.

Proof: Suppose that agent j cannot make a concession. Define

$C = \{s\varepsilon R^n | \forall i \neq j, r^*_i < s_i\}$. Both B and C are convex, C is open, r^* is on the boundary of C , and $B \cap C = \phi$. By the separating hyperplane theorem, there is a unit vector $w \varepsilon R^n$ such that $\forall r \varepsilon B \forall s \varepsilon C\ w \cdot (r-s) \leq 0$. By construction of C , $\forall r \varepsilon B\ w \cdot (r-r^*) < 0$ and $w_j = 0$. This contradicts the uniqueness of v. Q.E.D.

This paper concludes with the second welfare theorem for regular non-malevolent societies.

Theorem 3: Consider a society S , a welfare ascription W , and an environment $E = \langle A,x \rangle$. Suppose that:

(a) S and W are regular.

(b) If not xP_iy , then $y \varepsilon cl\{z | zP_ix\}$.

(c) If not xW_iy , then $y \varepsilon cl\{z | zW_ix\}$.

(d) S respects W , and is non-imposed

(e) Every agent in N is non-malevolent relative to W ,

(f) A is convex, and x is Pareto-optimal in A , and

(g) every agent can make a concession at E .

Then, x is final in A .

Proof: Suppose to the contrary, that $y \varepsilon A$, yP_Cx , and $\langle C,y,x \rangle \varepsilon M$. By essentially the same argument as was used to prove theorem 2, $\forall i 0 \leq p^i \cdot (y-x)$. For some $j \varepsilon C$, yP_jx let z be a concession from j at E. Then $0<p^j.(y-x)$, efficient d_{jj} and $\forall i \neq j\ 0<p^i \cdot (z-w)$. For some small $t \varepsilon (0,1)$, $\forall i \varepsilon N\ 0 < p^i \cdot ((1-t)y + tz)-x)$

Thus, by lemma 5, for some small $r \in (0,1)$, $\forall i \in N ((1-r)((1-t)y+tz) + rx) P_i x$.
This contradicts the Pareto-optimality of x in Λ . Q.E.D.

References

1. Bergstrom, T. C., "A'Scandinavian consensus' solution for efficient income distribution among non-malevolent consumers," Journal of Economic Theory 2 (1970), 383-398.

2. Bergstrom, T. C., "Interrelated consumer preference and voluntary exchange," in Papers in Quantitative Economics, vol. 2, (A. M. Zarley, ed.), University of Kansas Press, 1971, 79-94.

3. Debreu, G., Theory of Value, Yale University Press, 1959.

4. Goldman, A. M., "Gift equilibria and pareto optimality," Journal of Economic Theory 18 (1978) 368-370.

5. McKenzie, L.,"Matrices with dominant diagonals and economic theory," in Mathematical Methods in the Social Sciences, 1959, (K. Arrow, S. Karlin, P. Suppes, ed.), Stanford University Press, 1960, 47-62.

6. Winter, S. G., Jr., "A simple remark on the second optimality theorem of welfare economics," Journal of Economic Theory 1 (1969) 99-103.