Center for Research on Economic and Social Theory
CREST Working Paper

# Social Contract I: Harsanyi and Rawls

*Ken Binmore*

June, 1988
89-03

DEPARTMENT OF ECONOMICS
**University of Michigan**
Ann Arbor, Michigan 48109

# SOCIAL CONTRACT I:

# HARSANYI AND RAWLS

by Ken Binmore
Economics Department
University of Michigan
Ann Arbor, MI 48109.

# SOCIAL CONTRACT I: HARSANYI AND RAWLS

by Ken Binmore*

> And don't kid yourselves that there's any law in Poisonville except what you make for yourself.
>
> Dashiell Hammett, Red Harvest.

**1. Introduction.** This is the first of several papers whose beginnings lie in Rawls' [1958,1968,*1972*] theory of the social contract. The aim of the sequence of papers is to defend a version of Rawls' "egalitarian"[1] conclusion for a world in which agents are assumed to be constrained only by rational self-interest. No foundational issues are taken for granted. This is partly because I hope to make the work accessible to a wider audience; but mostly because I believe that much confusion in the literature derives from straightforward misunderstandings on matters which ought not to be controversial.

It should be emphasized that the program entails a very substantial reevaluation of Rawls' approach. The current paper follows numerous others in arguing that he is on shaky ground in rejecting orthodox decision theory. But, if orthodox decision theory is retained, then his line of reasoning leads inexorably to a species of *utilitarianism*, as demonstrated by Harsanyi [1953,1955,1958,*1977*]. Since much of the apparatus is required for later purposes, the paper describes an elaborated version

* The material of this paper is extracted from a long ST/ICERD discussion paper, "Game Theory and the Social Contract" (88/170). This supersedes an earlier ST/ICERD discussion paper (84/108) with the same title. I am grateful to Reinhart Selten and to John Harsanyi for their comments. It should be noted that Harsanyi is not comfortable with the contractarian reinterpretation of his model proposed here.

the model that Harsanyi uses in this defense of utilitarianism. For simplicity, only two agents are ever considered.

The paper continues with a criticism of the assumptions underlying this defense. Three difficulties are distinguished:

(1) The first difficulty concerns the use of the *Harsanyi doctrine*: namely that rational agents in identical circumstances necessarily have identical characteristics.

(2) The second difficulty is to be found in the argument's reliance on exogenously determined and unexplained *interpersonal utility comparisons.*

(3) The third difficulty lies in the fact that agents are assumed to be able to make binding *commitments* about their future conduct.

The third difficulty seems to me an insurmountable obstacle for both Rawls and Harsanyi. Rawls [1972, p167] faces the difficulty squarely in his "slave-holder's argument" which is examined in Section 7. It is essential for my purposes that Rawls' conclusions on this subject be rejected but, in doing so, it seems to me that one is only taking the views he expresses elsewhere on the "strains of commitment" [1972, p176] to their logical conclusion. Harsanyi's [1977] position on commitment is formally watertight[2], but it seems to me to face the same difficulties as Rawls' in so far as practical appplication is concerned.

The remainder of the paper is concerned with the manner in which it is proposed to deal with the first and second difficulty in later papers. The procedure is valid whether or not Harsanyi's commitment assumption is thought appropriate. For the purposes of illustrating the technique, it is therefore convenient to grant Harsanyi his commitment assumptions in the last part of the paper. The resulting analysis can then be seen as an attempt to strengthen Harsanyi's defense of utilitarianism by easing difficulties one and two. But the author's skepticism about difficulty three should always be born in mind. Of the three sections devoted to this analysis, section 8 surveys the simplifying properties of the model bought by taking commitment for granted. Sections 9 and 10 contain the substantive contribution. A method for *constructing* the "extended sympathy preferences" required in Harsanyi's methodology is described and an evolutionary justification for this construction is proposed. This provides an "idealized explanation" for both the origin of inter-personal comparisons of utility and the use of the Harsanyi doctrine in this context. The argument requires the contemplation of asymmetries in Rawls' "original position". In anticipation of this need, section 4 offers an elaboration of Harsanyi's model within which such asymmetries are expressible.

Five further papers in the current sequence are planned. Part II will be an attempt to clear up some misunderstandings about game theory and its applications to bargaining. Parts III and IV will be asides on evolutionary issues relevant to what comes later. Part V will contain a Humean reinterpretation of Rawls' social contract theory (as opposed to his Kantian view). Within this reinterpretation, "egalitarian" conclusions can be defended without recourse to hypotheses that need distress any conservative, no matter how red his neck. (This will seem less surprising when one learns how what it is that gets split equally is defined.) The ideas are closely related to those of Buchanan [1975,1976] and Sugden [1986]. Part VI will relate this work to the literature on cooperative bargaining theory. Finally, there will be a paper with the title "A Liberal Leviathan" which offers a philosophical overview.

**2. The original position.** Harsanyi [1977] is not a contractarian. Nevertheless, this section contains a reconstruction of his defense of utilitarianism within a contractual setting. It seems to me that his argument is then at its most powerful. The immediate discussion will be a *naive* one. In particular, the difficulties mentioned in the introduction are suppressed until later. Other more fundamental difficulties are not considered at all in the current paper. Why, for example, do some people find the device of the original position morally satisfying? What is the source of such moral intuitions? And why should they be thought relevant to the conduct of rational men of affairs? The program of which this paper is a part requires that answers be proposed to such questions, but it would be premature to offer them here.

To quote Rawls [1972, p17]: "... the original position is the initial status quo[3] which insures that the fundamental agreements reached in it are fair". The *original position* is called upon to give expression to what Rawls terms "the principle of redress". To quote again from Rawls [1972, p100]: "This is the principle that undeserved inequalities call for redress; and since inequalities of birth and natural endowment are undeserved these inequalities are to be somehow compensated for."

As a very simple example, which nevertheless gives the flavor of the general situation, consider the problem faced by Adam and Eve in the Garden of Eden. Fig. 1A illustrates a set[4] $X$ whose members consist of utility pairs corresponding to feasible *social contracts*—i.e. various ways of life they can jointly adopt. The point d represents a *state-of-nature* point—i.e. it represents the consequences of a final and complete failure to come to joint arrangements. Asymmetries in the configuration correspond to Rawls' "inequalities of birth and natural endowment". The problem for Adam and Eve is to agree on a point $x$ in $X$ on which to coordinate.

Figures 1A and 1B here

If they were to bargain without further constraints, then there is no particular reason why the point $x$ on which they finally settle should compensate the party disadvantaged by the shape of the set $X$ and the location of the point d. Even if one postulates that the choice of $x$ is to be delegated to an "ideal observer", who will arbitrate in a "fair" manner, one is still left to provide a definition of what "fair" should mean. Harsanyi deals with this difficulty by symmetrizing the situation. The symmetrization is achieved by requiring that negotiation is carried on behind a "veil of ignorance". Beyond the veil of ignorance, the roles that the players occupy in society become unknown. The players therefore have to negotiate a deal without knowing who is Adam and who is Eve. This ought to benefit the naturally disadvantaged party. The question is: by how much?

Fig. 1B is a representation of the problem faced by two persons, labeled 1 and 2, in such an "original position". For definiteness, it will always be assumed that player 1 is *actually* Adam and player 2 is *actually* Eve, but that this information is withheld behind the veil of ignorance. Their ignorance of their actual roles is reflected in the fact that both player 1 and player 2, in Harsanyi's formulation, attach probability 1/2 to each of the possible *role assignments*, $AE$ and $EA$. With role-assignment $AE$, player 1 is Adam and player 2 is Eve: the feasible set $X_{AE}$ is therefore $X$ and the state-of-nature point $d_{AE}$ is $d$. With role-assignment $EA$, player 1 is Eve and player 2 is Adam: the feasible set $X_{EA}$ and the state-of-nature point $d_{EA}$ are then the reflections of $X$ and $d$ in the line $x_1 = x_2$.

Players 1 and 2 in the original position will be assumed to be able to make agreements which are *contingent* on who turns out to be who[5]. They may agree, for example, to implement the utility pair y in set $X_{AE}$ of fig. 1B if player 1 turns out to be Adam and player 2 turns out to be Eve, but to implement the utility pair $z$ in set $X_{EA}$ if the roles turn out to be reversed. Since they attach equal probabilities to the two eventualities, they will assign the utility pair t of fig. 1B to this "contract". The utility pair $t$ lies at the midpoint of the line segment $yz$. The set[6] $T$ of all such points $t$, with $y$ in $X_{AE}$ and $z$ in $X_{EA}$, is shaded in fig. 1B. It is the set of all utility pairs from which it is feasible for players 1 and 2 to make a joint choice.

Since $T$ is convex and everything is symmetric, it does not matter greatly what theory of bargaining or arbitration is now employed. Any sensible theory will select, as the bargaining outcome, the point $n$ in fig. 1B that is the unique, Pareto efficient, symmetric point in $T$. To obtain this outcome, the players have to agree on implementing the utility pair $u$ in $X_{AE}$ if player 1 turns out to be Adam and player 2 to be Eve, and to implement $v$ in $X_{EA}$ if the roles are reversed.

An interpretive comment may be helpful. The *actual* situation is as in fig. 1A and it is the "fairness" of $u$ in this situation which is the *substantive* issue. The bargaining in the original position is only *hypothetical.* In defending the point $u$, Adam tells Eve that, if they had been placed in the original position under the specified conditions, then they surely would have agreed to make commitments to abide by the "contract" agreed to, whoever turned out to occupy whichever role. Eve may not like the consequent outcome, especially if it assigns her less than her state-of-nature utility, but she has an obligation to honor the commitment she would have made in the original position. In summary, if Eve is agreed that it is appropriate to use the device of the original position and that Adam is right in analyzing what her behavior in the original position would be, then Eve ought to act *as though* she were committed to maintaining the outcome $u$.

My own view is that this argument begs the questions which really matter. However, comment on this issue is postponed until section 7. Until then, the discussion will proceed without challenging the commitment assumptions built into Adam's defense of the point $u$.

In what sense can Harsanyi's proposed resolution of the problem be said to be utilitarian? The answer is simple, although it is not perhaps an answer that a dyed-in-the-wool utilitarian would find entirely satisfying. The point $u$ in fig. 1A is utilitarian in that it is the point in $X$ at which the arithmetic sum

$$u_A + u_E$$

of Adam and Eve's utilities is maximized.

It is immediately apparent that some sleight-of-hand has been visited upon the unwary. The sum of the utilities can only be meaningful if the utilities are comparable. Harsanyi does not neglect this issue. But, before discussing his approach to the inter-personal comparison of utilities, it is necessary to give some attention to Rawls' wholesale rejection, in the current context, of utility theory itself.

**3. Maximin criterion.** If a naive view is taken and problems of utility comparison are ignored, then the Rawlsian "difference principle" selects the point $r$ in fig. 1A—i.e. the equal-split point. However, Rawls does not take a naive view on utility questions. Indeed, so sophisticated is his view that he will have no truck with utility theory at all. Instead, he works in terms of what he calls "primary goods". But it does not seem to me that he can be allowed to dispense with utility theory and then proceed as though all the problems for which utility theory was created do not exist.

For example, if the axes in fig. 1B are re-interpreted in terms of primary goods (however defined), then orthodox decision theory would require the players to be heavily risk-averse if the point $r$ is not to be Pareto-inferior to a lottery attaching equal probabilities to u and v as advocated by Harsanyi. Rawls therefore needs to deny the validity of orthodox Bayesian decision theory in this context. Consequently, he insists that those in the original position use only objective probabilities (i.e. probabilities obtained by the observation of long-run frequencies) and eschew the use of subjective probabilities altogether. The aim is to rule out Harsanyi's lottery on the grounds, that since no *actual* coin is tossed, any evaluations made in the original position about prospective role-assignments will necessarily be subjective in character. He buttresses this position by recalling the well-known difficulties with Laplace's "principle of insufficient reason" and goes on to advocate the *maximin criterion* (rather than the maximization of expected Von Neumann and Morgenstern utility) with a reference to Milnor's axiomatization of "complete ignorance" (Luce and Raiffa [1957, p297]. Since the "difference principle" *is*, in essence, the maximin criterion, little further analysis is then required.[7]

It is not denied that Rawls' views on this subject have considerable force when applied to the arcane practices of naive Bayesians in general. Indeed, I have written at length elsewhere [Binmore,1987,section 6] about the difficulties in the foundations of game theory that arise from a naive use of Bayesian principles. I am therefore not entirely in sympathy with the position from which Harsanyi [1975] attacks Rawls' use of the maximin criterion. (See also Rawls' [1974] reply.) But these problems in the foundations of game theory arise from a thoughtless application of a decision theory designed for use with "closed universe" problems to "open universe problems". In brief, for a "closed universe" problem, it is necessary that there be no unresolved doubts about the nature of the fundamental domain of uncertainty. In Milnor's axiomatization of "complete ignorance" such doubts are embodied in his "column duplication" axiom. His theory, as its title suggests, is therefore very much an "open universe" one.

But should we follow Rawls in classifying the informational situation in the original position as "open"? I think not. After all, the fundamental domain of uncertainty is just a two-element set $\{AE, EA\}$ whose members represent the two possible role assignments. No doubts can exist about what this set is, because it is given *a priori* as part of the definition of the original position.

One cannot then simply deny that individuals will make decisions as though maximizing the expected value of a utility function relative to subjective probabilities for the two role assignments $AE$ and $EA$. Such a denial requires rejecting one or more of Savage's [1951] axioms for Bayesian decision theory. But if they are to be rejected in this simplest of all possible cases, then they have to be rejected *always*. Perhaps they ought to be. But, if so, their rejection should be based on their failings as rationality principles rather than because a rather distant and doubtful consequence is found displeasing.

The Savage theory only implies that rational individuals will make decisions as though $AE$ and $EA$ occur with certain probabilities. It does not require that these subjective probabilities be equal. However, in requiring that those in the original position do attach equal subjective probabilities to the role assignments, one is not claiming universal validity for the Laplace principle. Indeed, the whole idea that the informational circumstances of those beyond the veil of ignorance has somehow to be deduced from primitive ur-assumptions strikes me as misguided. In particular, it should be a matter of ***definition*** that the subjective probabilities assigned to the role assignments be equal. Everything is then up-front. Fairness is being defined as "symmetry in the original position": nothing more and nothing less.

These remarks are perhaps over-emphatic. But I agree very much with Rawls [1972, p121] that:

> We should strive for a kind of moral geometry with all the rigor
> that this name connotes.

**4. Extended sympathy.** How would one implement Harsanyi's conclusions in the real world? First it would be necessary, by observation or experiment, to discover Adam's and Eve's *personal preferences.* If Adam and Eve are rational in their attitudes to risk in the sense axiomatized by Von Neumann and Morgenstern, these preferences can be represented by utility functions $\varphi_A : S \to \mathbf{R}$ and $\varphi_E : S \to \mathbf{R}$ in such a way that Adam and Eve always behave as though maximizing *expected* utility. For example, if

$$\varphi_A(\tau) = (\varphi_A(\rho) + \varphi_A(\sigma))/2 \,,$$

then Adam will necessarily be indifferent between getting the state $\tau$ for certain, and the lottery which assigns probability 1/2 to each of the states $\rho$ and $\sigma$.

The rationality principles do not determine a Von Neumann and Morgenstern utility function $\varphi : S \to \mathbf{R}$ *uniquely* on the set $S$ of states over which the preferences extend. A second function $\psi : S \to \mathbb{R}$ describes the *same* preferences if and only if constants $B > 0$ and $C$ exist for which $\psi = B\varphi + C$. This serves to emphasize the fact that the rationality principles, in themselves, provide ***no basis at all*** for comparing the utility scales of different individuals.

One may fix a unique Von Neumann and Morgenstern utility function for an individual by insisting that it take the values 0 and 1 at two appropriate benchmark states. It is convenient to take a state $h$ (hell) which is regarded as worse than any immediately relevant state by both Adam and Eve, and a state $H$ (heaven) which is regarded by both as better than any relevant state. Thus, in what follows:

$$\varphi_A(h) = \varphi_E(h) = 0$$
$$\varphi_A(H) = \varphi_E(H) = 1.$$

The set of *feasible* states will be denoted by $S_0$. The states $s$ in $S_0$ are those which can result from the implementation of a social contract. The set $X$ in fig. 1A is defined by

$$X = \{(\varphi_A(s), \varphi_E(s)) : s \in S_0\}. \tag{4.1}$$

The state-of-nature is denoted by $s_0$. Also, $a = \varphi_A(s_0)$ and $1 - b = \varphi_E(s_0)$. Thus $d = (a, 1-b)$ in fig. 1A.

Sometimes, an *arbitrary* calibration of utility scales, as introduced above, is treated as though it established a sensible basis for the inter-personal comparison of utilities, but nothing of the kind is proposed here. The line taken on this issue will be an orthodox one: namely that use must be made of the theory of extended sympathy as introduced by Arrow [1978] and Suppes [1966].

In order for fig. 1B to be meaningful for players 1 and 2, it is necessary for them to be able to compare in their own minds, how it would be. for example, to be Adam drinking a cup of tea as opposed to Eve drinking a cup of coffee. To be more precise, they need to have preferences over the set $\{A, E\} \times S$. These preferences will be called *extended sympathy preferences.* As with personal preferences, if rationality is assumed, Von Neumann and Morgenstern utility functions $\Phi_1 : \{A, E\} \times S \to \mathbf{R}$ and $\Phi_2 : \{A, E\} \times S \to \mathbf{R}$ will exist. For example, $\Phi_1(E, s)$ is the utility assigned by player 1 to being Eve in state $s$.

If player 1 takes seriously what it would mean to be Eve, then he must surely admit that, if he were Eve, then his preferences over states would be the same as Eve's personal preferences. This will be taken as a basic assumption. The utility functions $\Phi_1(E, \cdot\,)$ and $\varphi_E$ then represent the *same* preferences over $S$, and hence, for each $s \in S$,

$$\Phi_1(E, s) = B_{1E}\varphi_E(s) + C_{1E},$$

where $B_{1E} > 0$ and $C_{1E}$ are constants. Similar equations will also hold with $E$ replaced by $A$ and with 1 replaced by 2.

The extended sympathy utility functions will be normalized so that

$$\Phi_i(A, h) = 0 \;\; ; \;\; \Phi_i(E, H) = 1 \qquad (i = 1,2). \tag{4.2}$$

The eight constants $B_{iA}$, $C_{iA}$, $B_{iE}$, $C_{iE}$ $(i = 1, 2)$ can then be expressed in terms of the quantities

$$U_i = \Phi_i(A, H) > 0; 1 - V_i = \Phi_i(E, h) < 1 \qquad (i = 1,2).$$

(To insist that $U_i \leq 1$ and $V_i \leq 1(i = 1,2)$ means that it is commonly agreed that Adam would suffer more in hell then Eve, and Eve would enjoy greater bliss in heaven than Adam.) A trivial calculation shows that, for each $s \in S$,

$$\begin{aligned} \Phi_i(A, s) &= U_i\varphi_A(s) \\ \Phi_i(E, s) &= 1 - V_i(1 - \varphi_E(s)) \end{aligned} \tag{4.3}$$

The sets $X_{AE}$ and $X_{EA}$ of fig. 1B are defined as

$$X_{AE} = \{(\Phi_1(A,s), \Phi_2(E,s)) : s \in S_0\}$$

$$X_{EA} = \{(\Phi_1(E,s), \Phi_2(A,s)) : s \in S_0\}.$$

Similarly, $d_{AE} = (\Phi_1(A,s_0), \Phi_2(E,s_0))$ and $d_{EA} = (\Phi_1(E,s_0), \Phi_2(A,s_0))$. These definitions expose one of the difficulties in the exposition of Harsanyi's argument as given in section 2. The difficulty arises from the fact that nothing in the argument offered so far justifies the *symmetry* of fig. 1B.

Write $\alpha = U_2/U_1$ and $\beta = V_1/V_2$. If $p_{AE} = (u, 1-v)$ is the utility pair arising from a state $s$ when player 1 is Adam and player 2 is Eve, then $p_{EA} = (1-\beta v, \alpha u)$ is the utility pair resulting from the *same* state $s$ when the roles are reversed. Thus, for example,

$$X_{EA} = \{(1-\beta(1-x_2), \alpha x_1) : (x_1, x_2) \in X_{AE}\}.$$

Fig. 2A illustrates the situation. In this figure, $\lambda = \sup\{\varphi_A(s) : s \in S_0\}$ and $1-\mu = \sup\{\varphi_E(s) : s \in S_0\}$.

Note in particular that symmetry arises if and only if $U_1 = U_2$ and $V_1 = V_2$ —i.e., $\Phi_1 = \Phi_2$. Without symmetry, it is no longer entirely trivial to resolve the bargaining problem in the original position. Section 8(iii) contains some justifying remarks for using the *Nash bargaining solution* for this purpose. The Nash bargaining solution for a set $T$ of feasible "deals", when the result of disagreement will be the point $w$, is the point $n$ at which the "Nash product" $(n_1 - w_1)(n_2 - w_2)$ is maximized subject to the constraints that $n \in T$ and $n \geq w$. If matters were as in fig. 2B, there would be nothing obviously utilitarian about the point $u$ that Adam would then have to defend to Eve as "fair".

Figures 2A and 2B here

It will be necessary to return to this asymmetric set-up in section 10. A more careful analysis will then be offered.

**5. The Harsanyi doctrine.** Equations (4.3) provide *intra*-personal comparisons of utilities. Player 1 regards one Adam-util as equivalent to $U_1/V_1$ Eve-utils. Player 2 regards one Adam-util as equivalent to $U_2/V_2$ Eve-utils. Harsanyi converts these intra-personal comparisons into a single *inter*-personal comparison using a form of what game theorists, following Aumann [1987], call the "Harsanyi doctrine". To caricature the doctrine, it is that human beings share a common psychological inheritance and hence, *if* they shared a precisely identical history of experience, *then* they would arrive at precisely the same judgements.

In the circumstances of the original position, Harsanyi argues that players 1 and 2 have effectively the *same* history of experience and hence will form the *same* extended sympathy judgements. This means that $\Phi_2 = B\Phi_1 + C$ for suitable constants $B > 0$ and $C$. In view of the normalization (4.2), it follows that $U_1 = U_2, V_1 = V_2$ and hence $\Phi_1 = \Phi_2$. Thus symmetry is secured. Notice that, when player 1's and player 2's *intra*-personal comparisons coincide, their common view constitutes an *inter*-personal comparison of Adam's and Eves personal utility scales.

I am uncomfortable with this piece of legerdemain. I agree with Hume [1793, p576] that we actually *do* have extended sympathy preferences. The reason is that we need to be able to imagine ourselves in the shoes of our fellows in order to be able to operate successfully as social (but autonomous) animals. Since player 1 is actually Adam and Adam has extended sympathy preferences, why then are player 1's extended sympathy preferences not simply those of Adam? Of course, if Adam and Eve have *different* extended sympathy preferences, then the veil of ignorance would need to be extended so that they forget whose extended sympathy preferences are whose. Otherwise they would be able to deduce their real-world identities from their information. However, I believe that it can be successfully argued that, in an ideal world, evolution would supply Adam and Eve with the *same* extended sympathy preferences. One would then have what one might call a Humean defense of the Harsanyi doctrine in this context.

If, on the other hand, one insists on maintaining a Kantian perspective within which evolutionary arguments have no place, then the prognosis seems to me to be gloomy. The fiction of hypothetical bargaining in the original position is powerful and arresting, but each new fiction which is introduced to facilitate its analysis attenuates its appeal. Obviously, one wants to *minimize* on what is assumed about those in the original position. Once this principle is abandoned, who is to say whose fiction is to be preferred? Rawls' fiction, that bargainers in the original positon will treat $\{AE, EA\}$ as an "open universe", was examined in section 3. Harsanyi offers the fiction that Adam and Eve are able to predict what their extended sympathy preferences would be if they were called upon to construct these *de novo* beyond the veil of ignorance—i.e. without reference to their own current extended sympathy preferences or the experiences which led to these preferences being held. Given this hypothesis, doubtless Adam and Eve would indeed make the same prediction. Equally, if the Riemann conjecture were decided, all mathematicians would agree on its truth value.

**6. Inter-personal comparison.** Harsanyi [1977, p55] asserts that there is "an unavoidable need for inter-personal comparison of utility in ethics". (See also, for example, Hammond [1976].) To see why he says this, observe that four constants, $U_1, U_2, V_1$ and $V_2$ characterize the normalized extended sympathy utility functions of section 4. The Harsanyi doctrine reduces these to two: namely, $U_1 = U_2 = U$ and $V_1 = V_2 = V$. Fig. 3A makes it obvious that the values of $U$ and $V$ matter. Two situations are considered. The difference is only that, in the second, $V$ is replaced by $V^*$. But it will be clear that the state corresponding to the utilitarian point u is very different to that corresponding to $u^*$.

Figures 3A and 3B here

A further difficulty with the exposition of Harsanyi's argument as given in section 2 has therefore been isolated. The identification of $X_{AE}$ with $X$ in section 2 is not innocent. Without this identification, only a form of *weighted utilitarianism* is obtained. That is to say, the point u selected from $X$ is *not* that illustrated in fig. 1A. Instead, it is the point at which

$$Uu_A + Vu_E$$

is maximized. Harsanyi deals with this issue by replacing the personal utility functions $\varphi_A$ and $\varphi_E$ by *renormalized* alternatives $\psi_A = U\varphi_A$ and $\psi_E = V\varphi_E$. This provides a *fiat* identification of $X$ and $X_{AE}$. But the problem, of course, does not go away. For one cannot tell at what state $\psi_A(s) + \psi_E(s)$ is maximized without knowing the value of $U/V$. Thus Harsanyi's insistence on the necessity for inter-personal comparison in ethics.

I do not doubt that this necessity exists. My complaint is that I do not see how such comparisons are to be "distilled from the air". To proceed as though such comparisons can be assumed to be *exogenously* determined, is to beg many of the questions which matter.

**7. Commitment and slave-holding.** Section 9 contains some evolutionary considerations which I believe can be used to defend Harsanyi's analysis from the difficulties raised in sections 5 and 6. But the next difficulty seems to me insuperable for a utilitarian conclusion.

A commitment is a *binding* promise. Since Schelling [1960], it has been understood how difficult it is, in real life, to make *genuine* commitments. Ellsberg's [1975] kidnapping paradigm is often quoted in this context. The victim would dearly love to make a commitment not to reveal the kidnapper's identity if released. The kidnapper, not wishing to murder his victim and having already received the ransom, would dearly love to be able to believe that a commitment had been made. But no *mechanism* exists for making an attempted commitment stick in such situations. In consequence of such examples, most game theorists nowadays treat commitment very gingerly. Each commitment opportunity, in so far as these exist, is modeled as a formal move within the game itself. The formal game is then analyzed without further commitment assumptions of any kind being incorporated into the analysis. Harsanyi has been a leading advocate of this approach in other contexts, but does not believe it appropiate to take the same line in his ethical writings. Nor does Rawls, as the following story indicates.

Rawls [1972, p167] considers a slave who complains about her condition of servitude. The slave-holder justifies his position to the slave with the followng version of a social contract argument. "If you had been asked to agree to a structure for society without knowing what your role in that society was to be, then you would have agreed to a slave-holding society because the prospective benefits of being a slave-holder would have outweighed in your mind the prospective costs of being a slave. Finding yourself a slave, you therefore have no just case for complaint." Rawls, of course, denies that a slave-holding society would be agreed to in the original position. But he takes pains to argue that the slave-holder's argument would be correct if it were true that a slave-holding society would have been agreed to in the original position, and explicitly rejects the slave's objection that she sees no grounds for honoring a hypothetical contract to which she never actually assented and which postulates a lottery for the random distribution of advantage which never actually took place. The slave-holder may continue by saying, "I agree that you never actually signed a contract. But this is not the issue. We can compute the terms of the contract that you surely would have wished to have signed if the circumstances of the original position had arisen. Either you must admit that you have a commitment to honor this contract or you have to deny that the device of the original position is appropriate in determining what is or is not fair.[2] As for the lottery, it is not hypothetical. Nature runs the lottery when she chooses who will be born into what station in life."

The weakness of this argument lies in its quasi-legal use of the word "contract". It takes for granted that, because one would have *wished* to have made a commitment and perhaps therefore have uttered appropiate words or signed a piece of paper, *therefore* a commitment would have been made. But, without a mechanism for making commitments stick, such gestures would be empty. For a person to have *claimed*, whether hypothetically or actually, that he or she is committed to a course of action is not the same as that person *being* committed to the course of action. As Hume [1739, p306] observes

> . . . we are not surely bound to keep our word because we have given our word to keep it.

The slave should point out that, although she and the slave-holder might have *wished* to be able to make commitments in the original position, no magic wand could have been at hand to make such wishes come true.

Essentially the same criticism may be applied to Harsanyi, although he differs from Rawls in seeing himself only as describing certain "ethical preferences". It is therefore enough for his *formal* conclusions that people should *wish* that society were such that certain hypothetical obligations were binding. The criticism therefore has to be directed, not at his theory as such, but at the possibility of its finding any useful application.

My own view has already been expressed in section 5. It is that it is necessary to *minimize* the number of fictions to be promulgated about the original position. The slave-holder may respond to the slave's objection that a magic wand is necessary to make wishes come true by ***hypothesizing*** the existence of a magic wand, but the slave is not likely to be convinced. Of course, proponents of the opposing view are more subtle about the *deus ex machina* brought onto the scene. They assert that ethical considerations can be relied upon to secure the honoring of agreements. In speaking of the "strains of commitment", for example, Rawls [1972, p178] remarks that "self-respect"[8] may be counted upon to take up the strain. But do questions of personal integrity take precedence over questions of social justice? Is it not precisely to elucidate such questions that the device of the original position is posited? What value does it have as a rhetorical device if there is no *a priori* agreement over what ethical ground-rules apply in the original position? Surely ***all*** ethical considerations should be "factored out" and built into the structure over which those in the original position negotiate.

This is to take a very severe line on what is to be regarded as "natural law". *No* ethical conventions are to be carried into the original position because these are *all* seen as *artificial* and hence needing to be *constructed* in the original position. In particular, to quote Hume [1739, p516] again:

> . . . a promise wou'd not be intelligible, before human conventions had establish'd it. . . even if it were intelligible, it wou'd not be attended with any obligation.

To follow Hume in seeking to identify moral behavior with enlightened self-interest is certainly to cut more than one Gordian knot. But it is a strategy which is not without its own problems. However, these are left for later papers in this program.

## 8. Commitment in Harsanyi's model.

The severe line on "natural law" advocated above is *not* pursued in the remainder of the current paper. Instead, Harsanyi is granted his commitment assumptions. The discussion is therefore about the ethical preferences of those who wish that society were organized in the manner that would be agreed to in the original position by individuals behind the veil of ignorance with an unlimited capacity for making commitments. This section examines, with more care than hitherto in the paper, what such a position on commitment buys Harsanyi in respect of the subsidiary assumptions of his model.

**(i) Feasibility.** There has been no discussion of the circumstances under which a social contract to operate a state $s$ is viable. The implicit assumption has been that, if a state is physically achievable. then it is available as the end product of an agreement. With no restrictions on commitment, such an assumption makes good sense. Without commitment, the question arises of whether agents will or will not honor the terms of the social contract. If this difficulty is taken seriously, attention has to be restricted to social contracts which only specify behavior which is *in equilibrium.* Then no party to the contract will ever have an incentive to be the first to dishonor its terms. But, once one has started to think of $S_0$ as the set of equilibrium outcomes of the "game of life", problems begin to accumulate. In Harsanyi's model, however, these can be neglected.

**(ii) The shape of $X$.** It has been implicitly assumed that $X$ is *convex* and *comprehensive* (as well as closed and bounded above, although the latter are assumptions that will be treated as harmless). To say that $X$ is comprehensive means that "free disposal" is assumed—i.e. if $x \in X$ and $0 \leq \xi \leq x$, then $\xi \in X$. These assumptions are standard in axiomatic bargaining theory. The defense is that, if the personal utility pairs in $Y$ are available, then so must be those in its "convex-comprehensive hull" $X$, as shaded in fig. 3B. It is argued that, if $x \in Y$ and $0 \leq \xi \leq x$, then the agents can make commitments to first achieve $x$ and then to "burn" sufficient utility to bring about $\xi$. This is the argument for $X$ to be comprehensive. The argument for $X$ to be convex is as follows. If $x \in Y$, $y \in Y$, and $z = px + (1-p)y$ $(0 \leq p \leq 1)$ lies on the line segment joining $x$ and $y$, then commitments can be made to the lottery which yields the "prize" $x$ with probability $p$ and the "prize" $y$ with probability $1-p$. When the agents can sign *legally enforceable* contracts, as it is often reasonable to assume in an economic context, the arguments seem solid. But no *a priori* legal system exists in the original position.

These issues *matter*. For example, fig. 3B shows the point $r$ selected by the maximin criterion[9] when applied to $Y$ and the point $q$ when it is applied to $X$. Can we rely on a (hypothetical) commitment reached in the original position to remove society from $r$ to $q$? In Rawls' phrase, this would certainly impose a substantial "strain of commitment".

The question of lotteries needs a little more attention, since the discussion of section 7 (and of part V) hinges on whether "coin-tossing" can be seen as a legitimate coordinating device in the original position. With commitment, there is clearly no problem. No problem exists either, without commitment, provided that the agents have no opportunity for communication after the fall of the coin and that the states from which a random selection is made represent *equilibria.* It will then be optimal for each agent to abide by the agreement to use the lottery, provided the other does as well.

But usually agents will have the opportunity to communicate after the fall of the coin. Imagine, for example, a dispute in a new nation about which side of the road on which to drive. Leftists argue that driving on the left leaves the right hand on the wheel when changing gear. Rightists have some other argument. A conference decides to settle the issue by tossing a coin. But why should the losers honor the agreement? They can produce another coin and call for a further trial. If they do not do so, it is because the act of honoring the agreement *is itself in equilibrium.* This implies that there must be some system of checks and balances in the background providing the necessary incentives for honest behavior. One may look to some *a priori* code of "natural law" to fulfil this function, but this is not a strategem which I favor for the reasons given in section 7.

*Without commitment,* I therefore believe it to be important to stress that there is *not* a trivial argument for the convexity of $X$ as often proposed. But perhaps $X$ is convex anyway. Perhaps the back-up system of checks and balances necessary to sustain randomizing agreements is in place. If not, there are other ways in which options can be mixed, as in taking turns or sharing physical goods (although such mixing does not necessarily have the same pleasant consequences for Von Neuman and Morgenstern utilities). In any case, when $X$ is assumed to be convex and comprehensive when commitment is not possible, the assumption is ***substantive*** and requires defense. Often, no such defense will be available.

**(iii) Bargaining.** Nash [1950] proposed a bargaining model, his "demand game", in which each player $i$ simultaneously makes a once-and-for-all, take-it-or-leave-it utility demand $x_i$. If the demands are compatible (i.e. $(x_1, x_2) \in X$), then each player receives his demand. Otherwise a disagreement outcome $d$ is implemented.

This model captures the essence of bargaining with an unlimited power to make commitments. Such a situation reduces to a race to confront the opponent with a take-it-or-leave-it problem. Hence one should expect all the action to be telescoped into the first instant after which the players will have left themselves no room for maneuver.

Nash showed that, provided there is some vestigial uncertainty about the precise location of the frontier of $X$, then any non-trivial Nash equilibrium of the demand game requires that the pair n of equilibrium demands is approximately the Nash bargaining solution. To say n constitutes a ***Nash equilibrium*** means that $n_1$ is an optimal choice for player 1 given that player 2 has chosen $n_2$, and that $n_2$ is an optimal choice for player 2 given that player 1 has chosen $n_1$.

With full commitment, a sound defense is therefore available for using the Nash bargaining solution to resolve the bargaining problem in the original position described at the end of section 4. (Further discussion appears in Part II of this sequence of papers).

**(iv) The state of nature.** How is the disagreement point *d* determined? With commitment, Nash [1953] has the answer to this also. If the disagreement point is not fixed in advance, each player will accompany his or her take-it-or-leave-it demand with a binding *threat* as to the action to be taken should the demand be left. The disagreement point is then the outcome which *would* result if the *optimal* threats in this "threat game" *were* to be used. (As is often the case, what actually happens in equilibrium depends on what *would* happen *were* there to be a deviation from equilibrium.)

Hobbes is nowadays credited with modeling the human condition as a species of Prisoners' Dilemma game (e.g. Kavka [1986]). If this were indeed the underlying situation (which seems to me an absurdly simplistic hypothesis), then what would be the optimal threats for those bargaining in the original position? One may deduce from Nash's theory that each bargainer should threaten to use his or her dominating strategy whoever turn out to be who. But the technical details are irrelevant since, if this were not the case, it would simply mean that there was something wrong with Nash's theory. The conclusion is that the consequent disagreement point would be Hobbes' state-of-nature, the "war of all against all", in which life is "solitary, poor, nasty, brutish, and short". Of course, without commitment, one is *compelled* to take a more sophisticated view of the situation.

With commitment, Harsanyi's model therefore hangs together very well provided that one takes a relaxed view about the difficulties concerning inter-personal comparison and the Harsanyi doctrine raised in sections 5 and 6. In what follows, an attempt is made to provide some underpinning for his assumptions on these latter topics.

**9. Evolution and extended sympathy.** Why, unlike ants, do we have *personal* preferences? This is presumably because agents whose behavior is not rigidly predetermined and who seek to optimize, given the beliefs experience has equipped them with, are better able to adapt to changing environments. Why do we have *extended sympathy* preferences? Because we are social animals and need to coordinate our behavior. For this purpose, we need to be able to put ourselves into the shoes of others (and into the shoes of future persons that we might ourselves become).

What determines the preferences we have? I claim that it is ultimately questions of evolutionary fitness[10] (social and economic, as well as biological). Extended sympathy between *unrelated*[11] individuals exists to facilitate coordination and for no other reason. If the result of the coordination affects the fitness of the individuals involved, then the possession of certain extended sympathy preferences will favor those that hold them at the expense of others who find themselves with different extended sympathy preferences in the same situation. The favored individuals will then be selected by evolution. Thus the extended sympathy preferences of an individual will be determined *strategically* in the long-run, *given* the coordinating device being operated. Individuals may well truthfully report that they are not misrepresenting what lies deep within their hearts. But, if an individual did not have strategically optimal extended sympathy views, then he or she would not be around to do any reporting.

To capture this notion, the next section studies the consequences of supposing that extended sympathy preferences are "in equilibrium". The idea will be explained in the context of the Harsanyi model. It will therefore be necessary to imagine a world as-it-might-have-been in which "natural law" for the species *homo* is as a follower of Harsanyi would wish it to be. Coordination problems are therefore resolved by employing the device of the original position in the form described in section 4. However, Adam and Eve will not be left to distil extended sympathy preferences from the air behind the veil of ignorance. Instead, they take their real-world extended sympathies with them as proposed in section 5.

It will not be assumed that their real-world extended sympathy preferences are arbitrary. Evolution will be assumed to have tailored these to the pair $(X, d)$ of fig. 1A. *The precise requirement will be that, behind the veil of ignorance, neither player 1 nor 2 would wish his or her extended sympathy preferences to be other than they actually are.*

The mathematical implications are outlined in section 10. In brief, the equilibrium requirement on extended sympathy preferences has a *symmetric* resolution and this provides a justification of sorts for the Harsanyi doctrine[12]. Perhaps more significantly, it also determines the ratio of the hanging constants $U$ and $V$ of section 6 and thereby ties down the precise rate at which Adam's utils are to be compared with Eve's.

More on evolutionary issues will be found in parts III and IV of this program.

**10. Sympathy equilibrium.** Under simplifying assumptions, it will be shown that *a necessary condition for an "interior" equilibrium in extended sympathy preferences is that the point in $X$ implemented using Harsanyi's version of the original position is simply the Nash bargaining solution of $X$ relative to the disagreement point $d$.* The conclusions about symmetry and inter-personal comparison of utilities mentioned in section 9 follow immediately for the reasons given at the end of this section.

It is important to stress that this result does *not* mean that the device of the original position is redundant *nor* that the Nash bargaining solution has virtues as a scheme for fair arbitration. Adam and Eve have *fixed* extended sympathy preferences in this story[13]. The ratio $U/V$ built into these preferences is tailored only to *one* pair $(X, d)$, which should be thought of as a surrogate for those pairs which figured large in the evolutionary history of the population from which Adam and Eve are drawn.

If faced with a *new* pair $(X', d')$, they will use the *same* ratio $U/V$ as for $(X, d)$ and hence the outcome will *not* be the Nash bargaining solution for $(X', d')$.

The argument begins at the point where the discussion of section 4 ends. An asymmetric version of Harsanyi's model had been constructed and the outcome of the bargaining in the original position, as seen by players 1 and 2, identified with the Nash bargaining solution $n$ for the pair $(T, w)$ in fig. 2B. In what follows, it is convenient to work in terms of underlying states, rather than directly in terms of utilities. Recall that a deal reached in the original position can be identified with a pair $(s, t)$ of states, where $s$ is to be implemented if player 1 turns out to be Adam and player 2 to be Eve, and $t$ is to be implemented if the roles are reversed. In terms of $s$ and $t$, the appropriate Nash product for players 1 and 2 is proportional to

$$\Pi = \{\theta f(s) + g(t)\}\{\psi f(t) + g(s)\}, \tag{10.1}$$

where $f(s) = \varphi_A(s) - \varphi_A(s_0)$, $g(s) = \varphi_E(s) - \varphi_E(s_0)$, $\theta = U_1/V_1$ and $\psi = U_2/V_2$. To see this, recall that the payoffs to players 1 and 2 when the deal $(s, t)$ is agreed are

$$\{\Phi_1(A, s) + \Phi_1(E, t)\}/2 = \{U_1\varphi_A(s) + 1 - V_1(1 - \varphi_E(t)\}/2,$$
$$\{\Phi_2(A, t) + \Phi_2(E, s)\}/2 = \{U_2\varphi_A(t) + 1 - V_2(1 - \varphi_E(s)\}/2.$$

The disagreement payoffs are found by writing $s = t = s_0$ in these expressions.

The immediate aim is to characterize the deal $(s, t)$ at which the Nash product is maximized for fixed values of $\theta$ and $\psi$. To simplify matters[14], attention is confined to the case when the basic problem is "divide the dollar". The state-of-nature $s_0$ then corresponds to neither player receiving a monetary payment. The Pareto-efficient states $s$ are those in which Adam gets a payoff of $s$ and Eve gets a payoff of $1 - s$. The states $s$ and $t$ to be considered in (10.1) are therefore real numbers in the interval $[0, 1]$. In addition, the functions $f$ and $g$ will be assumed to be twice differentiable on $(0, 1)$.

Only necessary conditions for an "interior" maximum will be considered. For this purpose, the partial derivatives of (10.1) with respect to $s$ and $t$ need to be equated to zero. Thus,

$$\Pi_s = \theta f'(s)\{\psi f(t) + g(s)\} + g'(s)\{\theta f(s) + g(t)\} = 0, \tag{10.2}$$

$$\Pi_t = g'(t)\{\psi f(t) + g(s)\} + \psi f'(t)\{\theta f(s) + g(t)\} = 0. \tag{10.3}$$

For (10.2) and (10.3) to have a non-trivial solution in the bracketted variables, the requirement is that

$$\theta\psi f'(s)f'(t) = g'(s)g'(t). \tag{10.4}$$

This last equation has a geometric interpretation. Note first that none of the preceding analysis changes if $\Phi_i$ is replaced throughout by $\Phi_i^*$, where $V_i\Phi_i^*(I,s) \;=\; \Phi_i(I,s) - \Phi_i(I,s_0)$ $(i \;=\; 1,2)$. This substitution allows fig. 2B to be replaced by fig. 4A which is simpler in that both players 1 and 2 attach a utility of 0 to the state-of-nature whether they turn out to be Adam or Eve. The analogs of $X_{AE}$ and $X_{EA}$ have been relabeled $Y_{AE}$ and $Y_{EA}$ in fig. 4A, but the analogs of $n$, $T$, $u$, $v$ and $w$ are labeled as in fig. 2B. Note that $w = 0$ in fig. 4A. Equation 10.4 says that the slope of the supporting line to $Y_{AE}$ at $u$ is equal to the slope of the supporting line to $Y_{EA}$ at $v$. (This feature of the situation is neglected in fig. 2B.)

Figures 4A and 4B here

Suppose now that an equilibrium in extended sympathy preferences occurs when $\theta = \Theta$ and $\psi = \Psi$. In examining the consequences of this assumption, $s$ and $t$ will be taken to be differentiable functions of $\theta$ and $\psi$ satisfying (10.2), (10.3) and (10.4). Thus $(s,t)$ is the deal negotiated in the original position when the extended sympathy preferences are determined by $\theta$ and $\psi$. Observe that, if player 1's extended sympathy preferences are *actually* determined by $\Theta$ and those of player 2 by $\Psi$, but

$\Theta$ is misreported as $\theta$ while $\Psi$ is reported accurately, then player 1's expected payoff is proportional to $\Theta f(s(\theta, \Psi)) + g(t(\theta, \Psi))$. A necessary condition for an "internal" extended sympathy equilibrium is therefore that

$$\Theta f'(s)s_\theta + g'(t)t_\theta = 0. \tag{10.7}$$

In this and later expressions, the functions $s$ and $t$, together with their partial derivatives, are evaluated where $\theta = \Theta$ and $\psi = \Psi$. A corresponding necessary condition for player 2 is that

$$\Psi f'(t)t_\psi + g'(s)s_\psi = 0. \tag{10.6}$$

To make something of (10.5) and (10.6), it is necessary to have data on the partial derivatives of $s$ and $t$. This is obtained by differentiating (10.2) and (10.3) partially with respect to $\theta$ and $\psi$. After various algebraic manipulations, the details of which appear in an appendix, (10.5) and (10.6) yield the simple necessary condition that

$$\Theta\Psi f(s)f(t) = g(s)g(t). \tag{10.7}$$

Like the functionally similar (10.4), this has a geometric interpretation in fig. 4A. It is that the slope of the line joining 0 and $u$ is the same as that joining 0 and $v$. This can only happen if $u = v$ as illustrated in fig. 4B.

An immediate consequence is that both the state s and the state $t$ implement the Nash bargaining solution for the pair $(X.d)$ (as well as for $(T, w)$). This is the result announced at the beginning of this section.

It follows, again immediately, that $s = t$. Thus, $\Theta f(s) = g(s)$, because $u_1 = v_1$. Similarly, $\Psi f(s) = g(s)$. Hence $\Theta = \Psi$. This provides the justification of the symmetry[15] derived by Harsanyi from his doctrine. The actual value of $\Theta$, which determines how Adam's utils are rated against Eve's, may be found by solving the equations

$$\Theta = \frac{\varphi_E(s) - \varphi_E(s_0)}{\varphi_A(s) - \varphi_E(s_0)} = -\frac{\varphi_E'(s)}{\varphi_A'(s)}$$

obtained by writing $s = t$ and $\Theta = \Psi$ in (10.4) and (10.7).

**11. Conclusions.** Rawls' device of the original position would lead him to Harsanyi's utilitarian outcome in those places where he currently advocates the maximin principle if his decision theory were orthodox. Three difficulties with Harsanyi's theory are isolated. Under certain circumstances, the requirement that extended sympathy preferences be "in equilibrium" removes two of the difficulties. These concern the reasons that symmetry may be assumed in the original position and the origin of the necessary inter-personal comparison of utilities. The remaining difficulty, that of the basis for making commitment assumptions, is seen as insuperable and requiring a reconstruction of the model to be attempted in later papers.

## Footnotes

1. "Egalitarian" is not intended in a technical sense. Thus, Rawls' [1972] difference principle is deemed to be egalitarian.

2. To quote Harsanyi [1958]: "Similarly, it is not a matter of social expediency, but it is a self–evident analytical truth, that people are required to fulfil morally *binding* promises".

3. The term "status quo" is used very loosely in this literature. As in this quotation it does *not* necessarily refer to the state-of-nature point as would seem natural to a bargaining theorist. Indeed, it is often used to refer to the social contract which *results* from bargaining on the grounds that this is what has to be justified.

4. Assumed closed, bounded above, convex and comprehensive. Some discussion appears in section 8.

5. In this the treatment differs from that of Harsanyi. The change in treatment is not significant in the current context, but is necessary later in situations for which it is not clear what Harsanyi's "ideal observer" would do. I am not, in any case, at all happy with the notion of an "ideal observer" as employed by Harsanyi and, although he denies the charge, also by Rawls.

6. Harsanyi restricts the choice set to be the intersection of T with the line $x_1 = x_2$. In the *immediate* context, this leaves the conclusion unaffected (see footnote 5).

7. Rawls would not be happy with this bowdlerization of his position!

8. He says that "self-respect" is, not so much part of any rational plan of life, as the sense that one's plan is worth carrying out". But this seems to beg the relevant questions. The later assertion that "self-respect is reciprocally self-supporting" is an equilibrium statement. But what sustains this equilibrium?

9. This is a convenient point to clarify the assertion of the introduction that Rawls' "egalitarian" conclusions were to be defended. This does not mean that

establishing the maximin criterion, as such, is the goal of later papers. Rawls [1958] earlier advocacy of *any* Pareto-efficient, Pareto-improvement of all equal-split points seems a more reasonable target. However, since, for the sake of simplicity, only strictly convex comprehensive $X$ are considered, the issue becomes moot. Only the Pareto-efficient, equal-split point is then a candidate.

10. Do I really prefer Mozart to Wagner for such reasons?

11. From the biological perspective, identical twins should be seen as two agents representing the *same* player.

12. Although, admittedly, it is still essentially a case of "symmetry in: symmetry out".

13. One might say that Adam and Eve are "committed" to these extended sympathy preferences.

14. The result generalizes very easily.

15. This is not full symmetry in that it has only been shown that $U_1/V_1 = U_2/V_2$, but not necessarily that $U_1 = U_2$ and $V_1 = V_2$. The latter conclusion can be achieved by making the inessential change of replacing $\Phi_2$ by $c\Phi_2$, where $c$ is a suitable positive constant. It will no longer be true that $\Phi_2(E, H) = 1$, but this does not affect the argument attributed to Harsanyi in section 4.

## References

K. Arrow [1978], "Extended sympathy and the possibility of social choice" *Philosophia* 7, 233-237.

K. Binmore [1984], "Game theory and the social contract II", ST/ICERD discussion paper 88/170, London School of Economics (supersedes part I).

K. Binmore and P. Dasgupta [1987], *Economics of Bargaining,* Basil Blackwell, Oxford.

J. Buchanan [1975], *The Limits of Liberty,* University of Chicago Press, Chicago.

J. Buchanan [1976], "A Hobbsian reinterpretation of the Rawlsian difference principle", *Kyklos* 29, 5-25.

D. Ellsberg [1975], *The Theory and Practice of Blackmail: Formal Theories of Negotiation,* University of Illinois Press, Urbana.

P. Hammond [1976], "Why ethical measures of inequality need interpersonal comparisons", *Theory and Decision* 7, 263-274.

J. Harsanyi [1953], "Cardinal utility in welfare economics and in the theory of risk-taking", *J. Political Economy* 61, 434-435.

J. Harsanyi [1955], "Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility", *J. Political Economy* 63, 309-321.

J. Harsanyi [1958], "Ethics in terms of hypothetical imperatives", *Mind* 47, 305-316.

J. Harsanyi [1975], "Can the maximin principle serve as a basis for morality?" A critique of John Rawls' theory", *American Political Science Review* 69, 594-606.

J. Harsanyi [1977], *Rational Behavior and Bargaining Equilibrium in Games and Social Situations,* C.U.P., Cambridge.

T. Hobbes [1651], *Leviathan,* ed. by C.B. Macpherson, Penguin Classics, London etc, 1986.

D. Hume [1739], *A Treatise of Human Nature,* ed. by L.A. Selby-Bigge, revised by P. Nidditch, Clarendon Press, Oxford, 2nd ed., 1978.

G. Kavka [1986], *Hobbesian Moral and Political Theory,* Princeton U.P., Princeton.

R. Luce and H. Raiffa [1957], *Games and Decisions,* Wiley, New York.

J. Nash [1953], "Two-person cooperative games", *Econometrica* 21, 128-140.

J. Rawls [1958], "Justice as fairness", *Philosophical Review* 57.

J. Rawls [1968], "Distributive justice:some addenda", *Natural Law Forum* 13.

J. Rawls [1972], *A Theory of Justice*, O.U.P., Oxford.

J. Rawls [1974], "Some reasons for the maximin criterion", *American Economic Review* 64 (papers and proceedings), 141-146.

L. Savage [1951], *The Foundations of Statistics*, Wiley, New York.

T. Schelling [1960], *The Strategy of Conflict*, Harvard U. Press, Cambridge, Mass.

R.Sugden [1986], *The Economics of Rights, Cooperation and Welfare*, Basil Blackwell, Oxford,

## Appendix

The purpose of this appendix ts to verify equation (10.7) which asserts that $\Theta\Psi f(s)f(t) \;=\; g(s)g(t)$.

The first step is to differentiate (10.2) and (10.3) partially with respect to $\theta$ and $\psi$. With the help of (10.2) to (10.6), it is possible to reduce the resulting expressions to the form

$$(a1) \qquad \Theta\Pi_{\theta s} = -g'(s)g(t) + \Theta A s_\theta = 0,$$

$$(a2) \qquad \Pi_{\psi s} = \Theta f'(s)f(t) + A s_\psi = 0,$$

$$(a3) \qquad \Pi_{\theta t} = \Psi f'(t)f(s) + B t_\theta = 0,$$

$$(a4) \qquad \Psi\Pi_{\psi t} = -g'(t)g(s) + \Psi B t_\psi = 0,$$

where,

$$A = \Theta\Psi f''(s)f(t) + \Theta f''(s)g(s) + \Theta g''(s)f(s) + g''(s)g(t),$$

$$B = \Theta\Psi f''(t)f(s) + \Psi f''(t)g(t) + \Psi g''(t)f(t) + g''(t)g(s).$$

To illustrate how this is done, consider (a1). From (10.2),

$$\Theta\Pi_{\theta s} = \Theta\{(fg)'(s) + \Psi f'(s)f(t)\} + \Theta s_\theta\{\Theta\Psi f''(s)f(t)\Theta\Psi f''(s)g(s) + 2\Theta f'(s)g'(s) +$$
$$\ldots + \Theta g''(s)f(s) + g''(s)g(t)\} + \Theta t_\theta\{\Theta\Psi f'(s)f'(t) + g'(s)g'(t)\}.$$

The terms in the first bracket are equal to $-g'(s)g(t)$ by (10.2). The terms in the final bracket cancel with $2\Theta f'(s)g'(s)$ in the middle bracket as a consequence of (10.4) and (10.5).

The elimination of $A$ and $B$ from $(a1) - (a4)$ yields that

$$(a5) \qquad g'(s)g(t)s_\psi = -\Theta^2 f'(s)f(t)s_\theta,$$

$$(a6) \qquad g'(t)g(s)t_\theta = -\Psi^2 f'(t)f(s)t_\psi.$$

It remains to note that these equations, together with (10.5) and (10.6), constitute a system of four homogeneous linear equations in the four unknowns $\Theta f'(s)s_\theta$, $g'(t)t_\theta$, $g'(s)s_\psi$ and $\Psi f'(t)t_\psi$. The condition for a non-trivial solution is that $\Theta\Psi f(s)f(t) = g(s)g(t)$ as required.

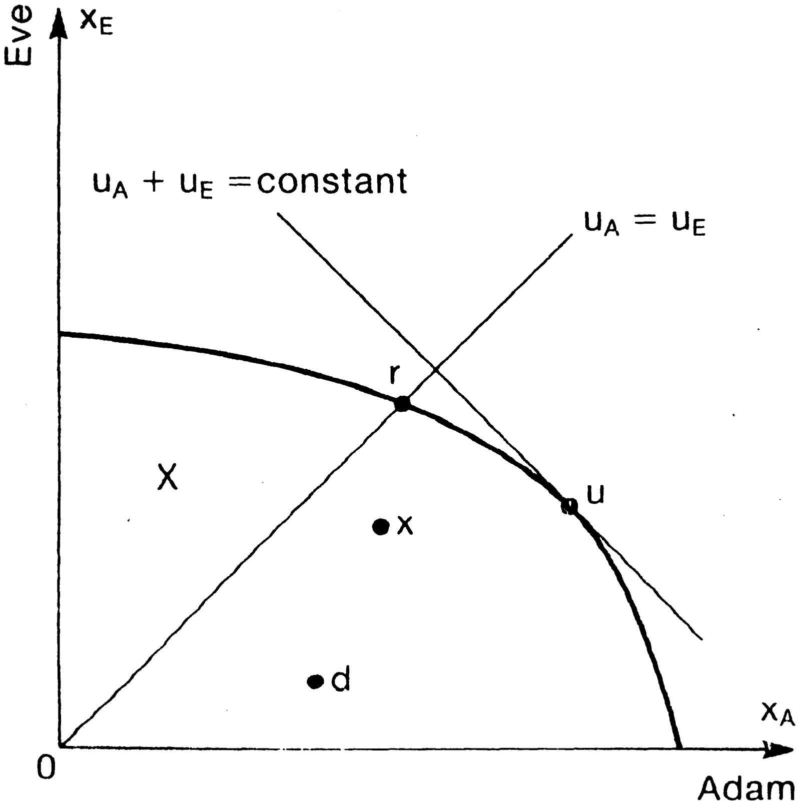

Figure 1A

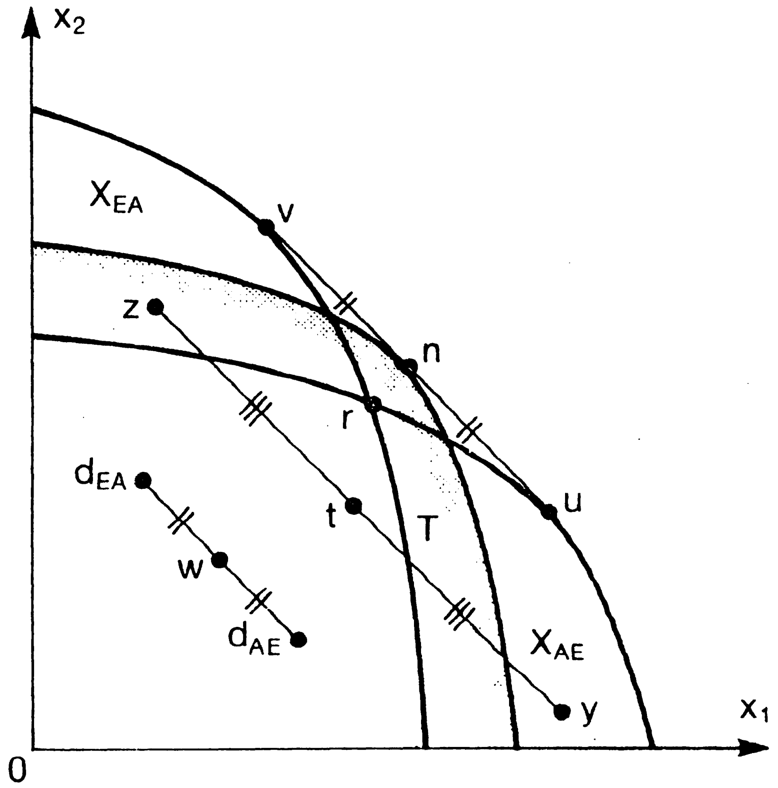

Figure 1B

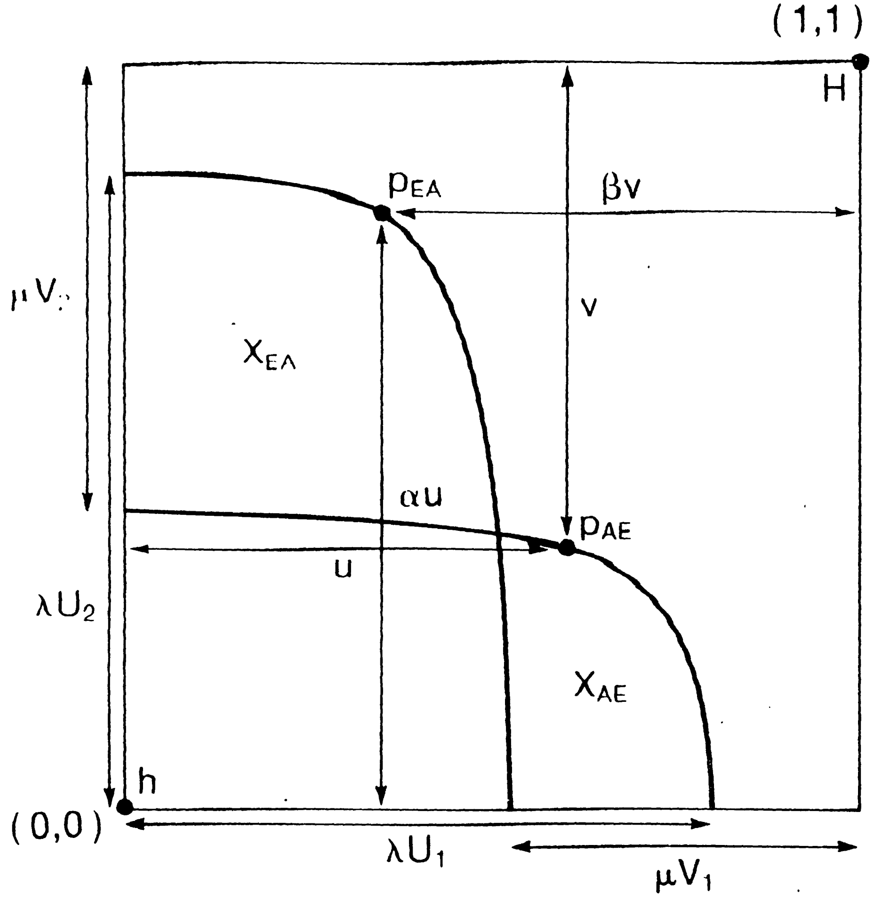

Figure 2A

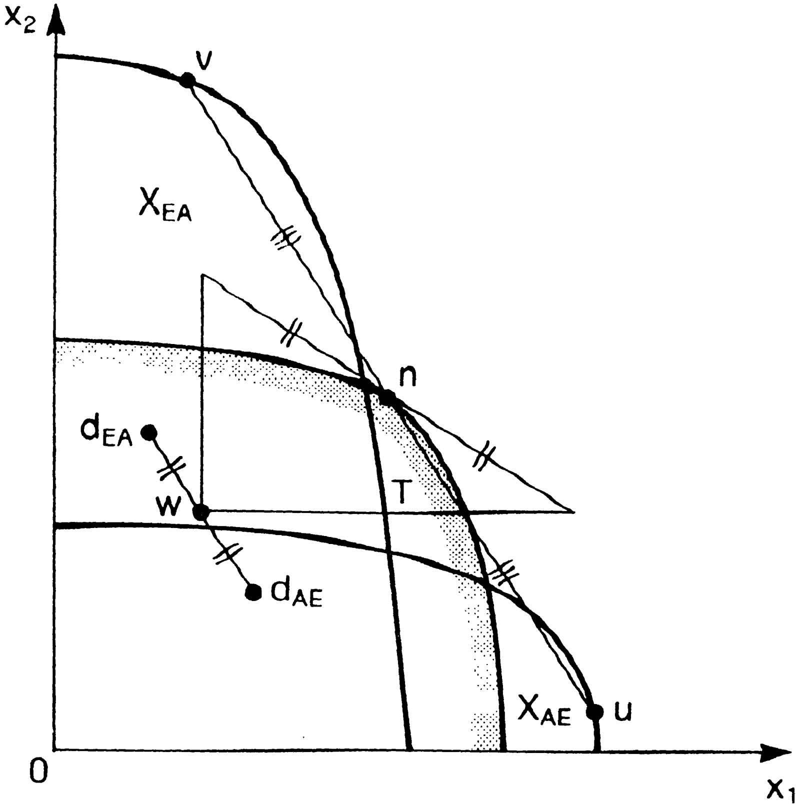

Figure 2B

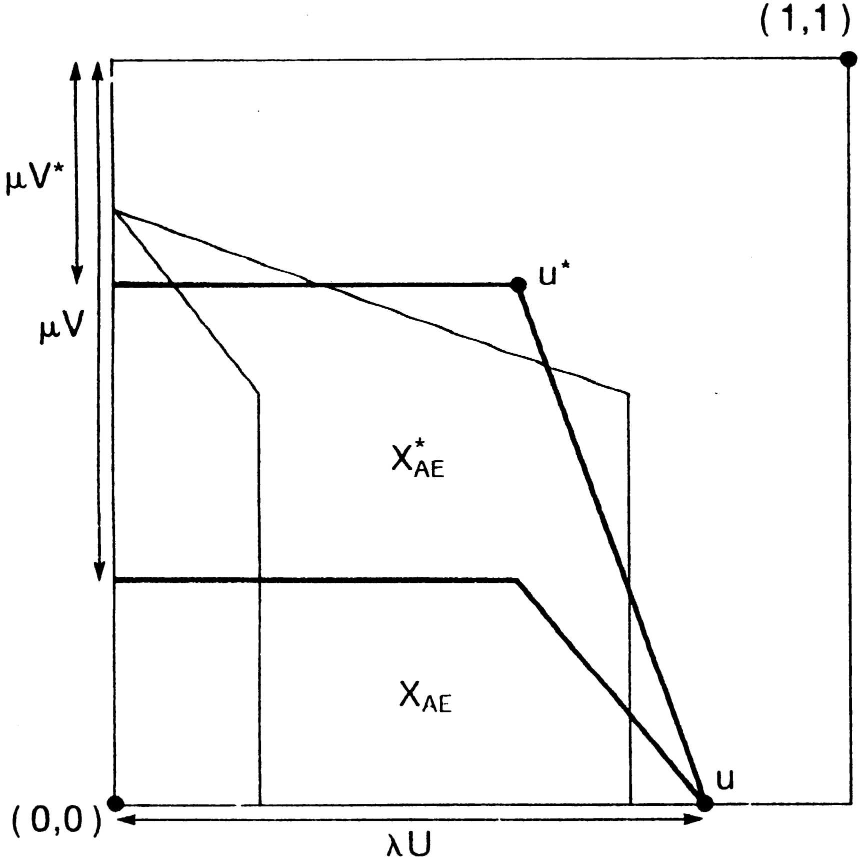

Figure 3A

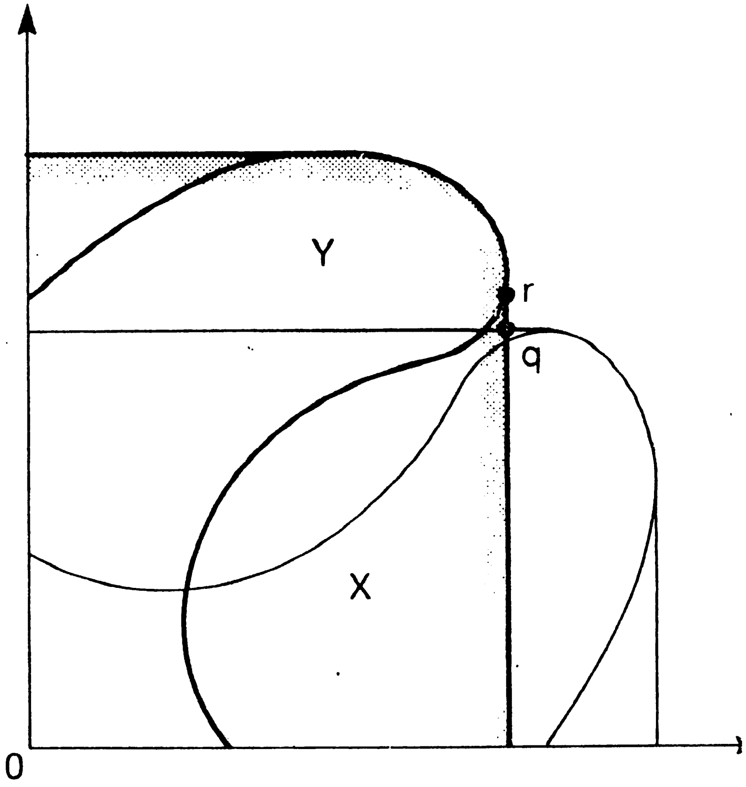

Figure 3B

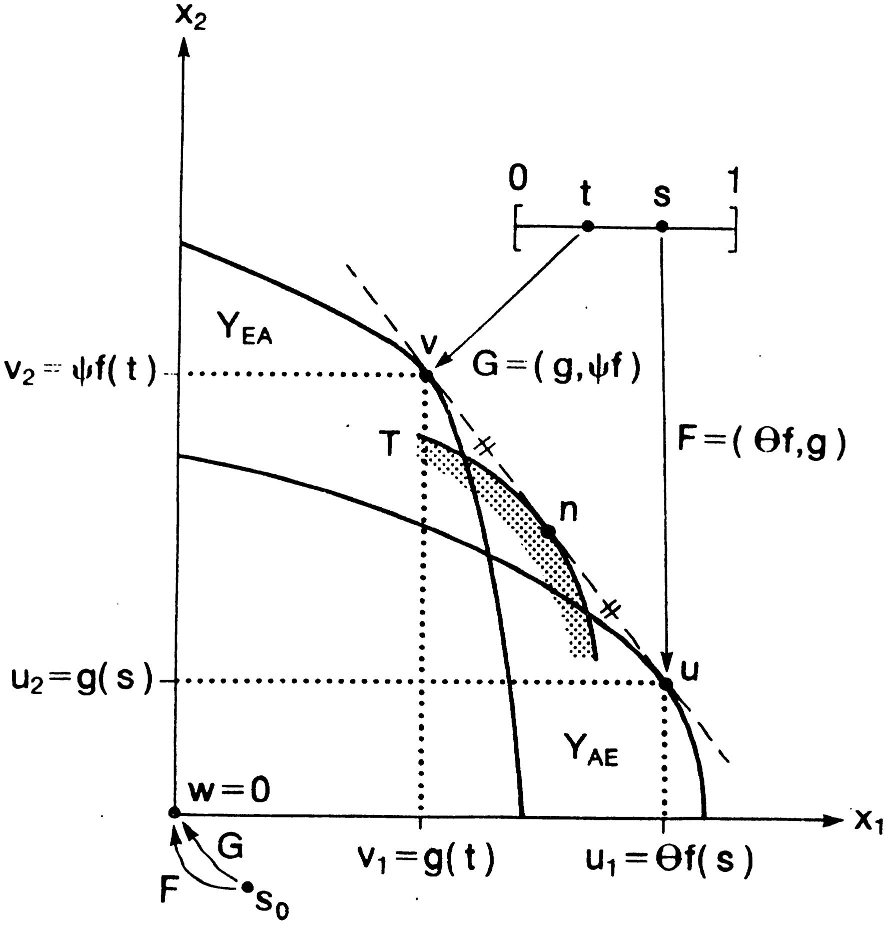

Figure 4A

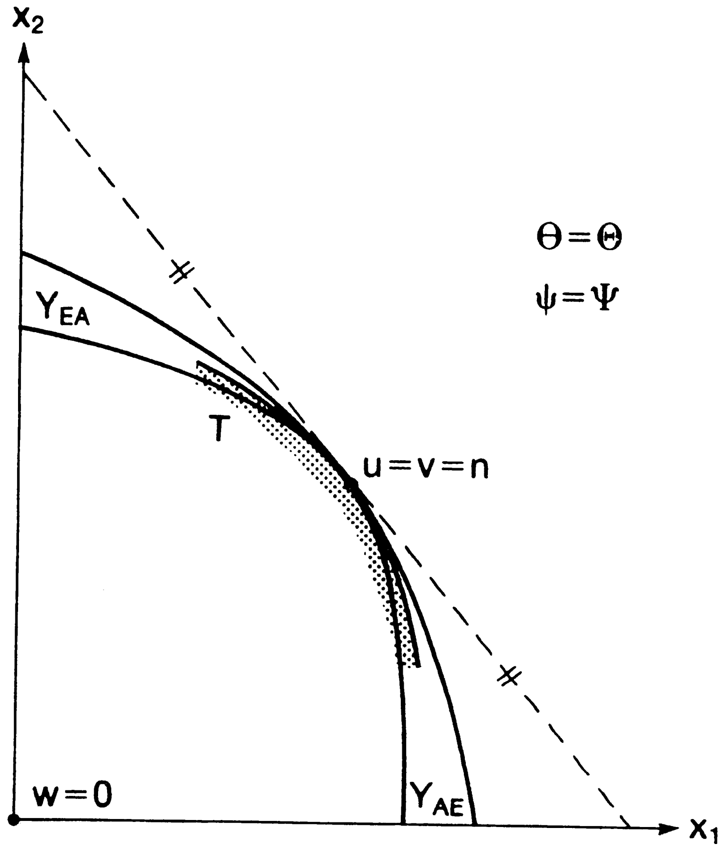

Figure 4B

**Recent Crest Working Papers**

88-1: Carol A. Jones, Suzanne Scotchmer, "The Social Cost of Uniform Regulatory Standards in a Hierarchical Government" December, 1987.

88-2: Ted Bergstrom, Judy Roberts, Dan Rubinfeld, Perry Shapiro, "A Test for Efficiency in the Supply of Public Education" December 12, 1987.

88-3: Mark Bagnoli, J. Bradley Barbeau, "Competition and Product Line Choice" February, 1988.

88-4: Severin Borenstein, Paul N. Courant, "How to Carve a Medical Degree: Human Capital Assets in Divorce Settlements" December, 1987.

88-5: Mark Bagnoli, Stephen W. Salant, Joseph E. Swierzbinski, "Pacman Refutes the Coase Conjecture: Durable-Goods Monopoly with Discrete Demand" January, 1988.

88-6: Jonathan Cave, Stephen W. Salant, "A Median Choice Theorem" December 29, 1987.

88-7: Mark Bagnoli, Naveen Khanna, "Why Are Buyers Represented by Seller's Agents When Buying a House?" December, 1987.

88-8: Mark Bagnoli, Roger Gordon, Barton L. Lipman, "Takeover Bids, Defensive Stock Repurchases, and the Efficient Allocation of Corporate Control" October, 1987.

88-9: Mark Bagnoli, Barton L. Lipman, "Private Provision of Public Goods can be Efficient" November, 1987.

88-10: Michelle J. White, "Urban Commuting Journeys are Not "Wasteful"" February, 1988.

88-11: Avery Katz, "A Note on Optimal Contract Damages When Litigation is Costly" February, 1988.

88-12: Ted Bergstrom, Jeffrey K. MacKie-Mason, "Notes on Peak Load Pricing" February, 1988.

88-13: Jerry A. Hausman, Jeffrey K. MacKie-Mason, "Price Discrimination and Patent Policy" February, 1988.

89-01: Mark Bagnoli, Severin Borenstein, "Carrot and Yardstick Regulation: Enhancing Market Performance with Output Prizes" October, 1988.

89-02: Ted Bergstrom, Jeffrey K. MacKie-Mason, "Some Simple Analytics of Peak-Load Pricing" October, 1988.

89-03: Ken Binmore, "Social Contract I: Harsanyi and Rawls" June, 1988.

89-04: Ken Binmore, "Social Contract II: Gauthier and Nash" June, 1988.

89-05: Ken Binmore, "Social Contract III: Evolution and Utilitarianism" June, 1988.

89-06: Ken Binmore, Adam Brandenburger, "Common Knowledge and Game Theory" July, 1988.

89-07: Jeffrey A. Miron, "A Cross Country Comparison of Seasonal Cycles and Business Cycles" November, 1988.

89-08: Jeffrey A. Miron, "The Founding of the Fed and the Destabilization of the Post-1914 Economy" August, 1988.

89-09: Gerard Gaudet, Stephen W. Salant, "The Profitability of Exogenous Output Contractions: A Comparative-Static Analysis with Application to Strikes, Mergers and Export Subsidies" July, 1988.

89-10: Gerard Gaudet, Stephen W. Salant, "Uniqueness of Cournot Equilibrium: New Results from Old Methods" August, 1988.

89-11: Hal R. Varian, "Goodness-of-Fit in Demand Analysis" September, 1988.

89-12: Michelle J. White, "Legal Complexity" October, 1988.

89-13: Michelle J. White, "An Empirical Test of the Efficiency of Liability Rules in Accident Law" November, 1988.

89-14: Carl P. Simon, "Some Fine—Tuning for Dominant Diagonal Matrices" July, 1988.

DATE DUE

MichU
DeptE
CenREST
W
89-03

Binmore, Ken
"Social Contract I:
Harsanyi and Rawls"

AUTHOR

TITLE

| DATE DUE | BORROWER'S NAME | DATE RETURNED |
|---|---|---|
| ~~Oct. 4, 93~~ | ~~Youngnam Choi~~ | |
| | ~~Kyle Logue~~ | |
| 8/15 | ~~Law prof.~~ | |