PARTHA DASGUPTA

ON SOME PROBLEMS ARISING FROM PROFESSOR RAWLS' CONCEPTION OF DISTRIBUTIVE JUSTICE*

ABSTRACT. This paper is concerned with two specific aspects of Professor Rawls' conception of distributive justice. In Section I Rawls' controversial difference principle is discussed in the context of individual decision under uncertainty, as well as the problem of interpersonal comparison of utilities. It is noted that, contrary to some recent accounts, Rawls' conception of the original position is considerably different from that to be found in the works of Harsanyi. In Section II an attempt is made to articulate in a precise way Rawls' intergenerational savings principle. While it is noted that the Rawlsian savings rule possesses a number of attractive properties, it is demonstrated that it conflicts with the principle of intertemporal Pareto efficiency.

0. INTRODUCTION

In this paper I shall point out a few difficulties that arise in attempting to formalise certain aspects of a conception of justice put forward recently by Professor John Rawls.[1] In doing so I shall conduct the arguments in the context of a really rather simple model of economic behavior. In point of fact the arguments can relatively easily be generalised in various directions. But that would complicate the presentation somewhat. I shall refrain, therefore, from doing so.

Even at the expense of repetition, it is as well at the outset to recall briefly the framework within which Professor Rawls has developed his particular conception. In Rawls' view:

> A conception of social justice... is to be regarded as providing... a standard whereby the distributive aspects of the basic structure of society are to be assessed.... The guiding idea is that the principles of justice for the basic structure of society are the object of the original agreement. They are the principles that free and rational persons concerned to further their own interests would accept in an initial position of equality as defining the fundamental terms of their association.... Thus we are to imagine that those who engage in social cooperation choose together, in one joint act, the principles which are to assign basic rights and duties and to determine the division of social benefits. Men are to decide in advance how they are to regulate their claims against one another and what is the foundation charter of their society. Just as each person must decide by rational reflection what constitutes his good, so a group of persons must decide once and for all what is to count among them as just or unjust. The choice which rational men would make in this hypothetical situation of equal liberty... determines the principles of justice (Rawls ([11], pp. 11–12).

Theory and Decision 4 (1974) 325–344. *All Rights Reserved*

Rawls then goes on to say,

> In justice as fairness the original position of equality corresponds to the state of nature in the traditional theory of social contract... (The original position) is understood as a purely hypothetical situation characterised so as to lead to a certain conception of justice. Among the essential features of this situation is that no one knows his place in society, his class position or social status, nor does anyone know his fortune in the distribution of natural assets and abilities, his intelligence, strength, and the like. I shall even assume that the parties do not know their conception of the good or their special psychological propensities. The principles of justice are chosen behind a veil of ignorance. This ensures that no one is advantaged or disadvantaged in the choice of principles by the outcome of natural chance or the contingency of social circumstances. Since all are similarly situated and no one is able to design principles to favour his particular condition, the principles of justice are the result of a fair agreement or bargain (Rawls [11], p. 12).

This, in Rawls' view, is the framework within which the principles of justice are to be conceived. A good part of the arguments of the book is then devoted to making plausible the claim that rational and disinterested persons behind the veil of ignorance would opt for the following two principles of justice, ranked in a lexicographic order:

> *First Principle*
>
> Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all.
>
> *Second Principle*
>
> Social and economic inequalities are to be arranged so that they are both:
> (a) to the greatest benefit to the least advantaged, consistent with the just savings principle, and
> (b) attached to offices and positions open to all under conditions of fair equality of opportunity.
>
> (Rawls [11], p. 302)

In this paper I shall not wish to question the framework. Rather, the purpose of this paper is to discuss the first part (part (a)) of the second principle of justice. In Section I, I shall touch briefly on the controversial maxi-min principle and relate it to the theory of decision under uncertainty, as well as to the problems connected with interpersonal comparison of well-being. In Section II I shall attempt to elaborate Rawls' savings principle. There is a further point in attempting such an elaboration here, since his savings principle is most likely to attract the least attention on the part of philosophers.

I. THE MAX-MIN PRINCIPLE

I.1. In a classic paper of 1954, John Milnor presented sets of axioms that are both necessary and sufficient for different decision rules under uncertainty.[2] The rules that he considered included the Wald-Rawls maxi-min criterion, the Mill-Sidgwick-Harsanyi utilitarian rule, Hurwicz's optimism-pessimism criterion, and Savage's min-max regret rule. Unfortunately the axiom sets do not lend themselves to interpretations that are any more transparent than the criteria they are equivalent to. But each axiom in itself looks, prima-facie, reasonable enough. The point in my referring to Milnor's contribution is that I think it would be a mistake to interpret Rawls as suggesting that a rational individual behind the veil of ignorance would *necessarily* be supremely risk-averse and opt for the difference principle. What Rawls does for a good part of his book is to attempt to make *plausible* the claim that *given* the defining characteristics of the basic structure of society, *given* the nature of the primary goods that are at stake in the basic structure, and *given* a particular cultural background and psychological motivation, a rational individual would, *in fact*, opt for the maxi-min rule.[3] In fact both the framework and the two principles reflect a particular conception of justice.[4] It cannot have been without deliberation that Professor Rawls has called his book '*A Theory of Justice*'. It would be an error to argue against Rawls by pointing out that he does not *prove* that the maxi-min principle would be chosen in the original position. He cannot be expected to provide a proof, for the simple reason that such a 'proof' cannot be constructed. It is as well to note that an individual's behaviour under uncertainty would generally be expected to depend on the objects of choice. There is nothing inherently contradictory in an individual being greatly risk-averse in one choice situation and at the same time being a gambler in another. The fact that men have a flutter in Monte-Carlo would, it would seem, not really constitute an empirical refutation of the second principle of justice.

I.2. The difference principle by itself is not, of course, necessarily a guide to action. We may all, behind the veil of ignorance, agree to opt for this principle of justice, but by itself it does not necessarily prescribe a basic structure. We have to contend with the problem of interpersonal comparison of well-being. This is not a problem unique to Rawls' scheme, but is in fact common to any conception of justice that is not outlandish. Professor Rawls is, of course, fully aware of this, and he meets it with what seems to me to be a set of splendid considerations which are, nevertheless, a bit of a cheat. In effect, Rawls assumes that individuals are *identical* with regard to their needs and preferences for primary social goods, the distributions of which among individuals are the defining characteristics of the basic structure of society.

> 'The difference principle also avoids difficulties by introducing a simplification for the basis of interpersonal comparisons. These comparisons are made in terms of expectations of primary goods. In fact I define these expectations simply as the index of these goods which a representative individual can look forward to. One man's expectations are greater than another's if this index for someone in his position is greater (Rawls [11], p. 92).

Furthermore,

> ... While the persons in the original position do not know their conception of the good, they do know, I assume, that they prefer more rather than less primary goods.... This interpretation of expectations represents, in effect, an agreement to compare mens' situations solely by reference to things which it is assumed they all prefer more of (Rawls [11], pp. 93–95).

Now even though there is much to commend in this rather ingenious device, it must be admitted that it does not actually boil down to any simplification of the problem of interpersonal comparison of well-being, except in the trivial sense of the device explicitly assuming the problem away. There would be many who would find this route not very compelling. In view of this it may be of interest to look briefly at the implications of the difference principle from the point of view of a characterisation of the original position that, in terms of *informational* requirement, is substantially more demanding than the one in *A Theory of Justice*. The particular characterisation that I refer to was offered by Harsanyi [3].[5] With this characterisation it would seem that the conditions of the original position tackle the problem of interpersonal comparison of well-being at the level of individual choice. But there does not appear to be any mechanism in the framework to tackle the problem at the level of collective choice. To see this, it is helpful to formalise the framework a bit.

Let us suppose that there are n members of society and that there are m basic structures to choose from. I denote by $N=\{1, 2,..., n\}$, and by $M=\{1, 2,..., m\}$. Let A_{kj} denote the vector of economic goods that individual j receives in basic structure k. I denote by R_j a reflexive, transitive and complete binary preference relation that individual j has over the set $(A_{1j}, A_{2j}, \ldots, A_{mj})$.[6] R_j is, therefore, nothing other than individual j's preference ordering over the various social structures in terms of the vectors of economic goods that *he* receives in them. But the veil of ignorance within the Harsanyi characterisation calls for a more pervasive ordering on individual j's part. For in the original position, an individual is invited to place himself in any other individual's circumstances in a given basic structure. I denote by $\tilde{R}_j$ a reflexive, transitive, and complete binary relation that individual j has over the entire $m \times n$ matrix of vectors $A_{ki} (k \in M, i \in N)$.[7] Following Sen [13] I shall refer to $\tilde{R}_j$ as individual j's *extended* ordering.[8] But in fact with this more stringent characterisation the veil of ignorance calls for more. It calls for one to take on other people's actual preferences as well. This would imply that the extended orderings must satisfy the following regularity condition:[9]

(A.1.1) For $i, j \in N$ and $k, l \in M$

$$A_{ki}\tilde{R}_jA_{li} \rightleftarrows A_{ki}R_iA_{li}.$$

It follows, then, that (A.1.1) resolves the problem of interpersonal comparison of well-being that any given individual confronts when *he* tries to rank the various basic structures. But even if all individuals abide by the difference principle, (A.1.1) is not sufficient to identify the 'just' basic structure, since the extended orderings may still differ from individual to individual. This can be seen from the following:

EXAMPLE: Let $m = n = 2$, and consider the situation where $A_{11}R_1A_{21}$ and $A_{22}R_2A_{12}$. Suppose that $A_{11}\tilde{R}_1A_{22}\tilde{R}_1A_{12}\tilde{R}_1A_{21}$ and that $A_{11}\tilde{R}_2A_{21}$ $\tilde{R}_2A_{22}\tilde{R}_2A_{12}$. It is plain that R_1, $\tilde{R}_1$, R_2, $\tilde{R}_2$ satisfy (A.1.1). However even if both individuals follow the difference principle, one would find individual 1 choosing basic structure 1 and individual 2 choosing basic structure 2.

It would seem then that one needs something more than merely (A.1.1) to identify the just basic structure. But it is worth noting that interpersonal comparisons at the level of the extended orderings is substantially less problematic with the maxi-min principle than it is with, say, the utilitarian scheme. Different individuals may well have different extended orderings. But if, for instance, all agree on which individual they would least like being in each basic structure, and of these least preferred positions they all agree on which individual they would most prefer to be, the difference principle would identify the just basic structure.[10]

I.3. But suppose that interpersonal comparison of well-being were to pose no problems. In particular, suppose that all individuals at all time have identical tastes and abilities. Suppose as well that the population remains constant over time and that the world's fund of technical knowledge does not change one bit. Earlier generations can bestow benefits on later generations by refraining from excessive consumption and by accumulating capital. Later generations cannot, of course, reciprocate. Invoking the maxi-min principle to handle the problem of intergenerational savings poses serious difficulties, since it leads readily to the policy that each generation invests only to the extent of making good the wear and tear of machinary that result naturally through their being in use. Or to put it in economists' jargon, the difference principle leads to zero net savings, a principle that one finds rather awkward to accept, since it would imprison the economy to perpetual poverty if it begins in poverty. It is at this point (Rawls [11], pp. 128–129) that Rawls alters the motivation assumption and suggests that generations are not likely to be totally selfish. Concern extends at least to one's offspring. But this again is not a guide to action unless one knows what one's offspring's intentions are regarding the amounts that they will save in the future. This line of argument leads Rawls to his savings principle. In the next section I take this up in detail.

II. THE SAVINGS PRINCIPLE

II.1. In arguing for the savings principle, we are to imagine, then, that

> .... The parties do not know which generation they belong to .... They have no way of telling whether it is poor or relatively wealthy, largely agricultural or already industrialized .... The veil of ignorance is complete in these respects. *Thus the persons in the original position are to ask themselves how much they would be willing to save at each stage of advance on the assumption that all other generations are to save at the same rates.* That is, they are to consider their willingness to save at any given phase of civilization with the understanding that the rates they propose are to regulate the whole span of accumulation (Rawls [11] p. 287, italics mine).

Rawls then goes on to say,

> Since no one knows to which generation he belongs, the question is viewed from the standpoint of each and a fair accommodation is expressed by the principle adopted. All

> generations are virtually represented in the original position, since the same principle would always be represented.... Only those in the first generation do not benefit... for while they begin the whole process, they do not share in the fruits of their provision. *Nevertheless, since it is assumed that a generation cares for its immediate descendents, as fathers care for their sons, a just savings principle... would be adopted.* (Rawls [11], p. 288, italics mine).

The savings principle reads a little vague and, as we shall see, leads to two possible interpretations, each of which has problems of its own. But first it would be best to articulate a model economy in which we are to imagine the Rawlsian scheme to be set into motion.

II.2. We are to imagine an economy that contains a single, non-deteriorating, homogeneous commodity, the stock of which at time t is K_t. Time is discrete and takes on the non-negative integer values. I suppose that each generation lives for precisely one period and is replaced by an equal number of direct descendents the instant they die. I suppose further that people of the same generation have the same tastes. We may as well normalise, then, and suppose that there is, at any one period, precisely one individual. Generation t consumes C_t of the commodity and is, naturally, constrained to choose so that $0 \leqslant C_t \leqslant K_t$. The quantity $K_t - C_t$ of the commodity that is carried over to the next period converts itself into $\lambda(K_t - C_t)$ units of the commodity (where $\lambda > 1$), and the whole process starts again.[11] We then have the basic accumulation equation

$$(1) \qquad \left.\begin{array}{l} t = 0, 1, 2, \ldots \\ \text{with } 0 \leqslant C_t \leqslant K_t \\ K_0 \text{ is given} \end{array}\right\} \quad K_{t+1} = \lambda (K_t - C_t).$$

We have therefore, started the economy off at $t = 0$ and have assumed, quite naturally, that K_0 is given. I shall denote by (C_t) a non-negative infinite sequence $(C_0, C_1, \ldots, C_t, \ldots)$ and by F the set of all infinite sequences (C_t) that satisfy (1). F is, therefore, the set of all *feasible* consumption sequences.

In fact it will be convenient some of the time to work instead with the savings ratio which I define at t by

$$(2) \qquad S_t \equiv \frac{K_t - C_t}{K_t} \quad t = 0, 1, 2, \ldots$$

It is plainly the case that $0 \leqslant S_t \leqslant 1$. I shall denote by (S_t) an infinite sequence $(S_0, S_1, \ldots, S_t, \ldots)$ and by F^1 the set of all infinite sequences that are bounded between 0 and 1. It is plain that to all intents and purposes there is a one-one correspondence between the sets F and F^1.

Now one can verify readily enough that in order to keep the stock, K_t, of the commodity at t precisely intact for the next generation, the generation at t must save at the rate $\bar{S}$, where

$$(3) \qquad \bar{S} = \frac{1}{\lambda} < 1$$

which would imply that $C_t = (\lambda - 1)K_t/\lambda$. A continued savings ratio of less than $1/\lambda$ leads to a depletion of the stock and thereby to an eventual deterioration of the standard of living of future generations. Furthermore, unless $S_t > 1/\lambda$, generation t would not be able to pass on to the next generation a capital stock larger than the one t had inherited.

From the definitional Equation (2) it is plain that consumption at t can be represented as

$$(4) \qquad C_t \equiv (1 - S_t)\, K_t .$$

Using (4) in the basic accumulation Equation (1), one notes that the stock, K_{t+1}, of the commodity at $t+1$ is given by

$$(5) \qquad K_{t+1} = \lambda S_t K_t \quad t = 0, 1, 2, \ldots$$

Turning to the preferences of different generations, I assume that they are identical. I shall incorporate Rawls' altered motivational assumption and suppose in particular that generation t is concerned solely with its own consumption level, C_t, and the consumption level, C_{t+1}, of the next generation. In fact I shall suppose

(A.2.1) The preferences of generation t are defined over C_t and C_{t+1} and can be represented by a utility function $U(C_t, C_{t+1})$ that is increasing in both C_t and C_{t+1}.

Now it may be immediately objected to that, even though I have allowed for Rawls' alteration in the motivational assumption, I have done so in a rather spurious way. Why must one's concern extend only to one's children? Ought not one's concern extend directly to one's grandchildren, to their children and so on? It can indeed be argued that it ought. But the function U reflects a generation's *tastes* and not its ethical values.[12] Whether or not a generation's interests extend into the affairs of distant future generations is an empirical question. As such (A.2.1) merits consideration.[13]

To keep the analysis simple I shall specify the function U more explicitly and suppose in particular

(A.2.2) The utility function U of generation t is of the form

$$U(C_t, C_{t+1}) = V(C_t) + \beta V(C_{t+1}),$$

where $0 < \beta \leqslant 1$, where $V(.)$ is twice differentiable, and where

$$\frac{\mathrm{d}V}{\mathrm{d}C} \equiv V'(C) > 0, \qquad \frac{\mathrm{d}^2V}{\mathrm{d}C^2} < 0 \quad \text{and} \quad \operatorname*{lt}_{C \to 0} \frac{\mathrm{d}V}{\mathrm{d}C} = \infty .$$

The parameter β, assumed constant, reflects the degree of natural concern that generation t bears for the next generation. If concern is complete, we shall have $\beta = 1$. But it may well be plausible to set $\beta < 1$.

Much of time I shall parametrize even further, and suppose that

$$(6) \qquad V(C_t) = -C_t^{-v} \quad (v > 0).$$

It is readily checked that the special form, as represented in (6) satisfies the conditions stipulated in (A.2.2).

Now it is plain that for any $(C_t) \in F$ (or equivalently $(S_t) \in F^1$), there is an associated infinite sequence of utility *levels* $(U_t) = (U_0, U_1, \ldots, U_t, \ldots)$. The question is: on what basis will a savings sequence be chosen in the original position? It is in fact rather striking that even after altering the motivation of generations, Professor Rawls does not once refer explicitly to an inter-generational difference principle as a basis for a just savings programme. But as the difference principle is so pervasive in *A Theory of Justice*, it is worth investigating the nature of the solution implied by this principle in the inter-temporal context.[14]

II.3. We suppose then that the economy is at $t = 0$. It has inherited a capital stock K_0. We are concerned with determining, if possible, that consumption sequence $(\hat{C}_t) \in F$ such that

$$(7) \qquad \sup_{(C_t) \in F} \inf_{t \geqslant 0} (U_t) = \inf_{t \geqslant 0} (U(\hat{C}_t, \hat{C}_{t+1})) .$$

To describe (7) *very* loosely, one looks at a feasible consumption sequence (C_t) and evaluates the resulting utility sequence (U_t). One then looks at the 'lowest' value of this sequence. Finally one searches within F until one locates that sequence $(\hat{C}_t)$ for which the lowest value in the corresponding utility sequence is greatest. $(\hat{C}_t)$, therefore, satisfies the intergenerational difference principle.

One can, then, establish readily enough

PROPOSITION 1: *If $\beta\lambda \leqslant 1$, then the consumption sequence $(\hat{C}_t)$ which satisfies the intergenerational difference principle is of the form*

$$\forall t: \quad \hat{C}_t = \frac{\lambda - 1}{\lambda} K_0 \equiv \bar{C}$$

(implying that $\forall t$: $S_t = 1/\lambda$).

It would appear, then, that so long as $\beta\lambda \leqslant 1$, altering the motivation condition of generations does not lead to any alteration in the implication of the intertemporal difference principle.[15] The economy preserves precisely the capital stock that it inherited at $t=0$. Proposition 1 is discouraging and, presumably, the considerations that led Rawls to shy away from the intergenerational difference principle to begin with would prevent him from arguing for the principle in the present case. It can, however, be argued that $\beta\lambda \leqslant 1$ is not really the empirically interesting case. What happens when $\beta\lambda > 1$? Assuming that $V(C_t)$ satisfies (6) one can, in fact, establish[16]

PROPOSITION 2: *If $\beta\lambda > 1$ and $V(C_t)$ satisfies (6), the consumption sequence $(\hat{C}_t)$ that satisfies the intergenerational difference principle is of the form*

$$\begin{aligned} \hat{C}_{2t+1} &= (\beta\lambda)^{1/(1+v)} \; \hat{C}_{2t} \\ \hat{C}_{2t+2} &= (\beta\lambda)^{-1/(1+v)} \; \hat{C}_{2t+1} \end{aligned} \qquad t = 0, 1, 2, \ldots$$

and where

$$\hat{C}_0 = \frac{(\lambda^2 - 1) K_0}{\lambda^2 + \lambda(\beta\lambda)^{1/(1+v)}}.$$

In Figure 1, the consumption sequence $(\hat{C}_t)$ of Propositions 1 and 2 are presented.[17] Alas, the difference principle is none too promising, even for the case $\beta\lambda > 1$. Granted that the sequence $(\hat{C}_t)$ is no longer the constant sequence $(\bar{C})$, it is nonetheless not much rosier to look at than $(\bar{C})$. Over cycles of two periods, for instance, the economy returns to the asset level

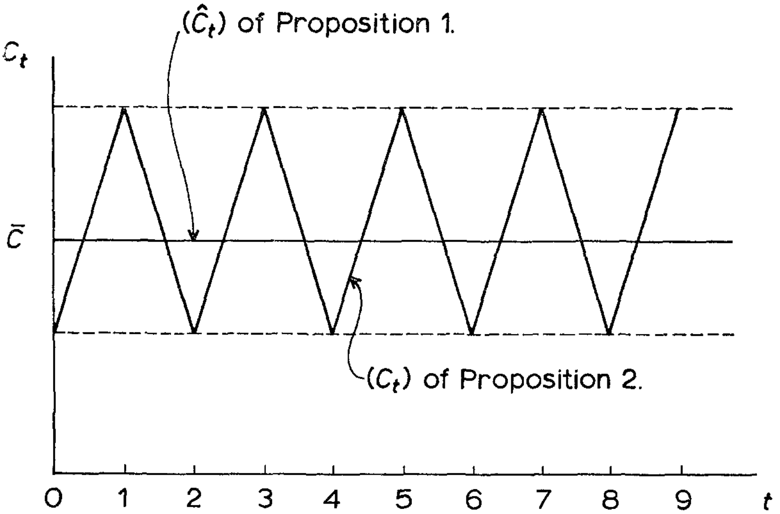

Fig. 1.

K_0 that it had inherited to begin with. The economy is not, therefore, allowed to lift itself permanently out of poverty if it begins with a low value of K_0. This feature alone is sufficient to enable one to mount a criticism of the intergenerational difference principle. The point about Propositions 1 and 2 is that they suggest that altering the motivation assumption does not make the intergenerational difference principle any more appealing.

But in fact there is a further disturbing feature of the difference principle that arises when $\beta\lambda > 1$. Suppose that at $t=0$, the existing generation consumes $\hat{C}_0$, as specified in Proposition 2. It follows directly from equation 1 that generation 1 would then inherit a capital stock given by

$$K_1 = \hat{K}_1 \equiv \frac{(\lambda(\beta\lambda)^{1/(1+v)} + 1)\,K_0}{(\beta\lambda)^{1/(1+v)} + \lambda} > K_0 .$$

Now one should recall that the consumption sequence specified in Proposition 2 was a consequence of a social contract undertaken at $t=0$. If this contract is to be regarded as being binding (as it clearly should,) generation 1 ought to consume precisely $(\beta\lambda)^{1/(1+v)}\hat{C}_0$. The question now is: can generation 1 enter the original position at $t=1$? If it can, and if it then appealed to the difference principle afresh from the vantage point

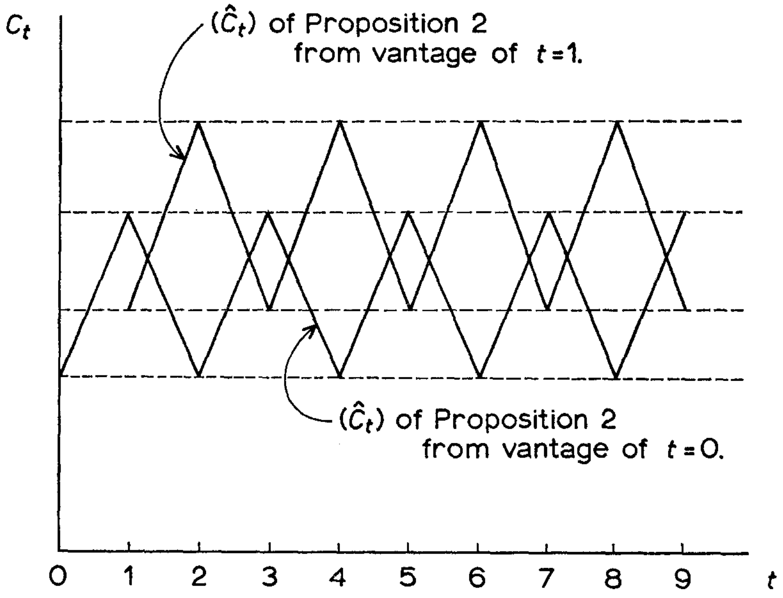

Fig. 2.

of $t=1$ it would be committed to consuming the amount

$$\frac{(\lambda^2-1)}{\lambda^2+\lambda(\beta\lambda)^{1/(1+v)}}\hat{K}_1.^{18}$$

But this amount is less than $(\beta\lambda)^{1/(1+v)}\hat{C}_0$, and the latter is the amount that generation 1 is committed to consuming by the contract at $t=0$. But unless the generation at $t=0$ is assured that the generation at $t=1$ will consume $(\beta\lambda)^{1/(1+v)}\hat{C}_0$, $\hat{C}_0$ would not be the appropriate quantity for it to consume. So some sort of an assurance is required for the generation at $t=0$ that generation 1 will consume the appropriate amount. But at $t=1$, the previous generation is dead and buried. Nothing that generation 1 does can affect it. So long as the generation at $t=1$ is allowed to enter the original position at $t=1$ and appeal to the difference principle afresh, there would be little reason for it to wish to consume $(\beta\lambda)^{1/(1+v)}\hat{C}_0$.[19]

It would seem that we have reached an impasse. But the critical question is: can the generation at $t=1$ enter the original position at $t=1$ and decide on the principles of justice afresh? It would appear that this would be allowed:

> ... The original position is not to be thought of as a general assembly which includes at one moment everyone who will live at some time; or, much less, as an assembly of everyone who could live at some time. It is not a gathering of all actual or possible persons... it is important that the original position be interpreted so that one can at any time adopt its perspective. It must make no difference when one takes up this viewpoint, or who does so: the restrictions must be such that the *same principles* are always chosen. (Rawls [11], p. 139 italics mine).

We sum up these considerations into

PROPOSITION 3: *The intergenerational difference principle leads* either *to a constant consumption programme (Proposition 1)* or *to inter-temporal inconsistency.*

In view of Proposition 3, I do not suppose that the intergenerational difference principle has much chance of being adopted within the framework of *A Theory of Justice.* It is certainly the case that even after altering the motivation assumption, Professor Rawls does not explicitly refer to the difference principle at any stage. But in fact there is a much more direct (and to me, compelling) interpretation of Professor Rawls' savings rule. In Section II.4 I take this up.

II.4. Recalling the admittedly somewhat vague savings principle, consider a feasible sequence (S_t^*) of savings ratios (i.e. $(S_t^*)\in F^1$). A member in the original position asks himself the following question: Assuming I were born at $t=0$ and were assured that all subsequent generations would save according to the sequence $(S_1^*, S_2^*,..., S_t^*,...)$, would S_0^* be the best savings ratio from my point of view? He then supposes that he is born at $t=1$ and asks himself if S_1^* is the best savings ratio from his point of view on the assurance that the generation at $t=0$ saved the ratio S_0^* and that subsequent generations are to save according to the sequence $(S_2^*, S_3^*,..., S_t^*,...)$. He continues asking this question by supposing in turn that he is born at every t. Roughly speaking, if the answers to this infinite sequence of questions is invariably a resounding 'yes', the sequence (S_t^*) would be deemed the just savings rule.[20] With this interpretation, the Rawlsian savings rule is most certainly intertemporally consistent. Each generation would be willing to save according to its obligation in the sequence (S_t^*) on the assumption that all other generations are to save at their respective obligatory rates.

Let us now suppose that generation t knows that generation $t+1$ will save at the rate S_{t+1}. Given the stock, K_t, of the commodity that generation t has inherited from the past, and noting (4) and (5), its problem is plainly to choose S_t with a view to maximising

$$V((1-S_t)K_t)+\beta V((1-S_{t+1})\lambda S_t K_t). \tag{8}$$

It follows immediately that in order to maximise (8) by a suitable choice of S_t, one must satisfy the condition

$$V'((1-S_t)K_t)=\beta(1-S_{t+1})\lambda V'((1-S_{t+1})\lambda S_t K_t). \tag{9}$$

It is now plain that a Rawlsian savings rule is an infinite sequence (S_t^*) of savings ratios, the adjacent components of which must bear a relation to each other via Equation (9).

To obtain sharp answers one must specify one's functions further. For the remaining part of the analysis I shall suppose that $V(C_t)$ has the form given in (6). In which event it is simple to check that Equation (9) reduces to the more amiable form

$$S_t=\frac{1}{1+\beta^{-1/(1+v)}(1-S_{t+1})^{v/(1+v)}\lambda^{v/(1+v)}}, \quad t=0,1,2,\ldots \tag{10}$$

Equation (10) is fundamental, for the Rawlsian savings rule will be a solution to it. If we were to leave aside the welfare parameters β and v and the technological parameter λ, it is a striking feature of Equation (10) that the 'just' savings ratio S_t for generation t depends only on the savings ratio S_{t+1} of generation $t+1$. This simplicity of expression (10) is due really to the special form of the utility function we are working with. With more general utility functions, one would expect that S_t would as well depend on the stock of the commodity, K_t, at t.

Now, there are what mathematicians call two stationary points of the system of Equations (10). In Figure 3 the curve AB represents S_t as a function (as reflected by the form (10)) of S_{t+1}. The pair of points (S^*, S^*) and (1, 1) are the two stationary points of the system. What one needs to note, however, is that any sequence (S_t) of savings ratios with $S_0>S^*$ that satisfies the system of Equations (10) will have the property that $\lim_{t\to\infty} S_t=1$. Moreover, any sequence with $S_0<S^*$, that satisfies the system of Equations (10) will take the economy in *finite* time to the point A, so that in finite time the stock of the commodity will be exhausted, leaving nothing for subsequent generations to consume. This is plainly not very just. So let us ignore them. All this implies that the point (S^*, S^*) is an unstable stationary point, and that the point (1, 1) is stable. It also suggests that

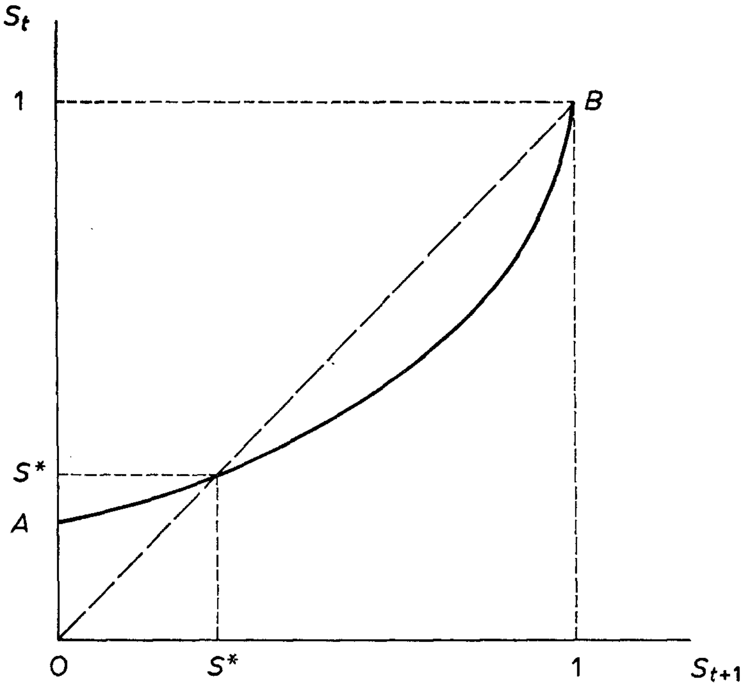

Fig. 3.

Rawls specification of the savings principle is not unique. Rawls would, of course, lose no sleep over this feature since he can, rightly, claim that the constant savings sequence (S^*) is Pareto superior to *any* sequence tending to the point (1, 1), in the sense that *all* generations would prefer to follow the savings rule (S^*) to *any* sequence satisfying (10) and tending to the point (1, 1).

These arguments establish

PROPOSITION 4: *The Rawlsian savings rule is the sequence* (S^*) *of savings ratios, where* S^* *is the solution* ($\neq 1$) *of the equation*

$$S = \frac{1}{1 + \beta^{-1/(1+v)}(1-S)^{v/(1+v)}\lambda^{v/(1+v)}}. \tag{11}$$

There are two immediate points that can be raised about the Rawlsian savings rule, one of which is trivial, the other possibly not so. First, the fact that the savings rule in Proposition 4 dictates a constant savings ratio over time is not really a general property of the Rawlsian savings principle, but is in fact a consequence of the number of parametric assumptions that I have made so far about the economy. Second, the assumption that I have made about all generations having the *same* tastes might well suppress the fact that in deriving the Rawlsian savings rule, one needs *no* assumption about the extent to which intergenerational comparisons of well-being can be made. To confirm this, one needs merely to note that one would still obtain Equation (9) if one were to replace the utility function $U(C_t, C_{t+1})$ of generation t by the function $\phi_t(U(C_t, C_{t+1}))$, where ϕ_t is an increasing and differentiable function of U. Since ϕ_t may vary with t any way we like, we would not be invoking any intergenerational comparison of utilities. This is in sharp contrast with the intergenerational difference principle, where intergenerational utility *levels* must be comparable.[21]

Utilitarian savings rules, based on models of economic growth similar to the one we are discussing here have tended often to yield somewhat excessively high savings ratios – at least during the initial stages. One would presumably wish to know what orders of magnitude the Rawlsian rule suggests. Suppose, for example, that $V(C_t) = \log C_t$. Then one confirms, by putting $v=0$ in Equation (11), that $S^* = \beta/1+\beta$. Assuming $\beta = 0.5$ (a not implausible figure), one obtains $S^* = \frac{1}{3}$. The figure is high, but not excessively so.

The Rawlsian savings principle, it would seem then, possesses a number of attractive properties. Alas, it has one unhappy feature that is seriously disturbing. This we can summarise in

PROPOSITION 5: *The Rawlsian savings rule is intertemporally strictly Pareto inefficient.*

What Proposition 5 reflects is the fact that there exists at least one sequence of savings ratios $(S_t) \in F^1$ which would yield a higher utility level for each generation than that attained via the Rawlsian savings rule.

Proof: One has to demonstrate the existence of an infinite sequence (S_t), bounded between 0 and 1, such that

$$\begin{aligned}(12)\qquad & (1 - S^*)^{-v} S^{*-vt} + \beta(1 - S^*)^{-v} \lambda^{-v} S^{*-v(t+1)} \\ & > (1 - S_t)^{-v} S_0^{-v} S_1^{-v} \dots S_{t-1}^{-v} \\ & + \beta(1 - S_{t+1})^{-v} \lambda^{-v} S_0^{-v} S_1^{-v} \dots S_{t-1}^{-v} S_t^{-v} \\ & \qquad\qquad\qquad \text{for} \quad t = 0, 1, 2, \dots\end{aligned}$$

For notational ease, write $f(x, y) = (1-x)^{-v} + \beta\lambda^{-v}(1-y)^{-v} x^{-v}$.

One can rewrite (12) in the more compact form

$$(13)\qquad f(S^*, S^*)\, S^{*-vt} > S_0^{-v}\, S_1^{-v} \dots S_{t-1}^{-v}\, f(S_t, S_{t+1}) \quad t = 0, 1, 2, \dots$$

Let $\varepsilon > 0$ and consider the following savings rule:

$$S_0' = S^*/1 - \varepsilon$$

$$(14)\qquad S_t' = S^*\left(1 - \frac{\varepsilon}{2^t}\right) \quad t = 1, 2, \dots$$

It follows that

$$S_0' S_1' \dots S_{t-1}' > S^{*t}(1-\varepsilon)^{-1} \prod_1^{\infty}\left(1 - \frac{\varepsilon}{2^t}\right) \geqslant S^{*t}.$$

The inequality in (13) would, then, be satisfied if there exists an $\varepsilon > 0$ such that $f(S^*, S^*) > f(S', S_{t+1}')$ for all t. But such an ε must surely exist, since $\partial f(S^*, S^*)/\partial y > 0$ and $\partial f(S^*, S^*)/\partial x = 0$ (Equation (10)), so that, for sufficiently small ε we have

$$f(S_0', S_1') = f\left(\frac{S^*}{1-\varepsilon}, S^* - \tfrac{1}{2}\varepsilon S^*\right) < f(S^*, S^*)$$

and

$$f(S_t', S_{t+1}') = f\left(S^* - \frac{\varepsilon S^*}{2^t}, S^* - \frac{\varepsilon S^*}{2^{t+1}}\right) < f(S^*, S^*)$$

$$(t \geqslant 1). \quad \|$$

The fact that in many intertemporal games no Nash equilibrium is Pareto efficient is not, of course, new. In the context of intergenerational savings, this has been noted by Sen [12], Marglin [6] and, much nearer to our purpose, in a striking paper by Phelps and Pollak [10].[22] But the earlier investigations were concerned with studying the consistency of certain savings processes. If a Nash equilibrium were found to be Pareto inefficient, as welfare economists the authors would point out to a need for cooperative action based on some ethical principle. The reason why Proposition 5 is rather disturbing is that it is the process of articulation of an ethical principle itself that has led to the dilemma. Members of the original position would, presumably, find it awkward to assent to the Rawlsian savings sequence (S^*), since there are many feasible sequences that are Pareto superior to it. But there does not appear to be any principle of rationality within *A Theory of Justice* which would predict how the members will choose from them.

III. CONCLUSIONS

In this paper I have discussed some aspects of a conception of justice as enunciated recently by Professor Rawls. In Section I Rawls' second principle of justice was discussed in the context of the problem of decision under uncertainty. In Section II an attempt was made to articulate Rawls' savings principle in the context of a simple model of accumulation. The intergenerational difference principle was first tried out as a possible interpretation. It was found to lead to either no accumulation or to intertemporal inconsistency (Proposition 3). In Section II.4 the savings principle was given the more plausible interpretation of an intergenerational Nash equilibrium. While it was noted that it had a number of attractive features, it was found (Proposition 5) to have the unfortunate property of being intertemporally Pareto inefficient.

London School of Economics and
Trinity Hall, Cambridge

NOTES

[*] This is a revised and expanded version of the text of a lecture delivered at the Moral Sciences Club of the University of Cambridge in February, 1973. I would like to acknowledge the many instructions that I have received on the matters discussed in this paper from Kenneth Arrow, Simon Blackburn, Frank Hahn, Philip Pettit, John Rawls, Abhijit Sen, Amartya Sen, and Robert Solow. While I doubt very much if any of these gentlemen would agree with all that I assert in this paper, I hope that each agrees with some of the propositions madehere.

[1] Rawls [11].

[2] See Milnor [7]. For a thorough discussion of Milnor's axioms, see Luce and Raiffa [5], Chapter XIII.

[3] For a similar interpretation, see Pettit [9].

[4] I say 'two principles', though in fact there are *three*; the elucidation of the third principle being the main purpose of this paper.

[5] See also the discussion in Pattanaik [8] and Sen [13].

[6] R_j does not necessarily have to be a *complete* ordering for Professor Rawls' scheme, but it seems to be a pointless direction to generalise in here.

[7] Here I am following the notation in Sen [13], Chapter 9 *.

[8] See also Kolm [4], who calls this individual j's *fundamental* ordering.

[9] This is referred to as the 'identity axiom' in Sen [13], p. 156, though in fact the notation is slightly different there.

[10] The reader would recognise that this is by no means the weakest sufficient condition required to generate the 'just' basic structure in the Rawlsian scheme. Utilitarianism via the contract doctrine (as in Harsanyi [3]) would, of course, require a much more

stringent form of the assumption that all individuals have identical extended orderings. On this last, see Pattanaik [8].

[11] It is supposed that $(\lambda - 1 > 0)$, the rate of return to investment (or of the rate interest), remains constant over time.

[12] "It is partly to preserve... clarity that I have avoided attributing to the parties any ethical motivation. They decide solely on the basis of what seems best calculated to further their interests so far as they can ascertain them" (Rawls [11], p. 584). Thus, the function U is a representation of the ordering R of Section II and not of $\tilde{R}$.

[13] At any event, the problems that we shall note in Professor Rawls' savings principle will persist even if concern extends to some more descendents.

[14] I am grateful to Kenneth Arrow for pointing out to me the need to look into the intergenerational 'maxi-min' problem, as well as for providing a rigorous argument establishing a solution of this problem of which Proposition 2 below is a special case. For an interesting exploration of the 'maxi-min' principle (*without* the altered motivation assumption) in the context of planning with exhaustible resources, see Solow [15]),

[15] One should note that in the *absence* of any alteration in the motivation condition, one would typically have $\beta = 0$.

[16] See note 14.

[17] The adjacent points of each of the sequences have been joined by straight lines merely to make the figure more transparent.

[18] The reader can readily verify that this conclusion must be so since generation 1 would draw up Proposition 2 afresh from its vantage point with the then initial capital stock $\hat{K}_1$. Figure 2 brings together the solutions of the maxi-min problem from the two vantage points in time.

[19] The source of the problem that I have raised here is the fact that different generations have preferences that are inconsistent with one another. The problem itself was first discussed by Strotz [16] in the context of an individual planning over his lifetime. In his discussion of deliberative rationality Rawls himself touches on the question of consistent plans for a rational individual, (see Rawls [11], pp. 421–422). Among the more recent formal discussions of Strotz' problem see the papers by Blackorby *et al.* [1], and Hammond [2].

[20] One would note readily that I am here interpreting the Rawlsian savings rule as an intergenerational Nash equilibrium. See Luce and Raiffa [5], for a discussion of this concept.

[21] See Sen [14] in this Symposium for a thorough discussion of this.

[22] Actually I overstate somewhat. Phelps and Pollak considered only those Nash equilibria which yielded a constant sequence of savings ratios and showed that they were Pareto inefficient. Their model, possessing a more complicated preference structure on the part of generations, allowed for the possibility of Pareto efficient Nash equilibria, though one would doubt it.

BIBLIOGRAPHY

[1] Blackorby, C., Nissen, D., Primont, D., and Russell, R. R., 'Consistent Intertemporal Decision Making', *Review of Economic Studies* (1973).

[2] Hammond, P. J., 'Altruism and Consistent Dynamic Planning', mimeo., University of Oxford, 1970.

[3] Harsanyi, J. C., 'Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility', *Journal of Political Economy* (1955).

[4] Kolm, S.-C., *Justice et équité*, Monographie d'econometrie, No. 8, CNRS, Paris 1972.

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

[6] Marglin, S. A., 'The Social Rate of Discount and the Optimal Rate of Investment', *Quarterly Journal of Economics* (1964).

[7] Milnor, J. W., 'Games against Nature', in *Decision Processes* (ed. by R. M. Thrall *et al.*), John Wiley, New York, 1954.

[8] Pattanaik, P. K., 'Risk, Impersonality and the Social Welfare Function', *Journal of Political Economy* (1968).

[9] Pettit, P., 'A Theory of Justice?', mimeo., Cambridge University, 1973.

[10] Phelps, E. S. and Pollak, R. A., 'On Second Best National Savings and Game Equilibrium Growth', *Review of Economic Studies* (1968).

[11] Rawls, J., *A Theory of Justice*, Clarendon Press, Oxford, 1972.

[12] Sen, A. K., 'Isolation, Assurance, and the Social Rate of Discount', *Quarterly Journal of Economics* (1967).

[13] Sen, A. K., *Collective Choice and Social Welfare*, Oliver and Boyd, 1971.

[14] Sen, A. K., 'Rawls versus Bentham: An Axiomatic Examination of the Pure Distribution Problem', this issue, p. 301.

[15] Solow, R. M., 'Intergenerational Equity and Exhaustible Resources', MIT Discussion Paper, 1973.

[16] Strotz, R., 'Myopia & Inconsistency in Dynamic Utility Maximization', *Review of Economic Studies* (1955–1956).