From: Edmund S. Phelps (ed.)
Altruism, Morality, and Economic Theory
(New York: Russell Sage, 1975)

# *Charity: Altruism or Cooperative Egoism?**

Peter Hammond

If a Covenant be made, wherein neither of the parties performe presently, but trust one another; in the condition of meer Nature, (which is a condition of Warre of every man against every man,) upon any reasonable suspition, it is Voyd: But if there be a common Power set over them both, with right and force sufficient to compell performance; it is not Voyd. For he that performeth first, has no assurance the other will performe after; because the bonds of words are too weak to bridle mens ambition, avarice, anger, and other Passions, without the feare of some coerceive Power; which in the condition of meer Nature, where all men are equall, and judges of the justnesse of their own fears cannot possibly be supposed. And therefore he which performeth first, does but betray himselfe to his enemy; contrary to the Right (he can never abandon) of defending his life, and means of living. [Hobbes's "Leviathan," chapter XIV]

## INTRODUCTION

It is evident that altruism *can* be invoked to explain any charitable behavior we may observe. But it is not quite obvious that altruism *must* be invoked to explain *all* charitable behavior. May not a person be charitable because he

*I am especially conscious of and grateful for the influence upon this paper of discussions with and comments by Kenneth Arrow, Christopher Bliss, Avinash Dixit, James Mirrlees, Edmund Phelps and Amartya Sen. But none of these is responsible for the contents.

believes that his present charity increases the likelihood that charity will also occur in the future, when the person may himself be in need? If so, charity is compatible with complete egoism, together with certain beliefs about the future.

The beliefs that an egoist needs to have, in order for charity to be rational, can be of two somewhat different kinds. He may believe that, even though he is not altruistic, charitable behavior may encourage altruism in the future, when he may be in need. He may, for example, believe that, if his children see him being charitable, they are more likely to grow up to be altruistic. Of course, he may encourage this by exhortation as well, but such words usually need the backing of deeds.

Alternatively, the egoist may believe that he is involved in a dynamic game. In this game, intertemporal cooperation is possible, and takes the form of charitable behavior.

The first alternative—in which future tastes are influenced—is no more than a temporary departure from altruistic charity. The egoist is only charitable because he hopes that this will bring about altruistic charity if he needs it. This explanation of charity is not very convincing. Moreover, it is hard to say very much more about it. Accordingly, I shall not discuss it further.

The second alternative—in which charity is an outcome of a game—is much more interesting.[1] It is also closely related to the theory of social contracts. Charitable behavior could be regarded as complying with a social contract. The egoist is worried that, if he breaks the contract, then many others may also decide to break the contract later on, with the result that the egoist's needs are not adequately met if he should ever require help in the future.

It is evident that charity is only likely to be an outcome of a dynamic game. In a static game, the rich are in no real danger of becoming poor. They may bribe the poor in order to gain their cooperation, but bribes are fairly easily distinguished from charitable gifts, it would seem. It is only when the return on a charitable gift is uncertain or fairly remote in time that we are likely to regard it as charity rather than a bribe.

It should also be observed that games in normal form miss the whole point. A cooperative solution—e.g., a core solution—may well involve egoists making transfers in different periods. An egoist will be prepared to exchange income in periods when he is rich for income when he is poor. But these are clearly not charitable transfers; they are merely transactions in a simple financial market—or intertemporal exchange economy. But this is not the main objection to the normal form I have in mind. Rather, the problem is that which Hobbes discusses. If the egoist makes a charitable transfer now, what guarantee has he that charitable transfers take place later? After all, by the time the rich egoist is poor, there is a new game. Why should other egoists take account of past generosity?

[1] It is also, I believe, a generalization of the first alternative.



The answer to this seems obvious. Egoist $A$ can only expect egoist $B$ to be charitable later, if egoist $B$ also expects egoist $C$ (or $A$) to be charitable later still. And even then, if egoist $B$ is charitable because he expects $C$ to be charitable later, what relevance does the charity of $A$ have to $B$'s decision? At first, it seems, the answer must be none. In which case there was no reason for $A$ to be charitable in the first place. But I believe this reasoning to be false—even for completely rational egoists. Indeed, I hope to demonstrate that egoists *can* rationally be charitable, under certain conditions.

Nevertheless, it should be clear that the normal form of the game is not the appropriate one for considering charity. Intertemporal agreements, which one looks for in the normal form, may be broken in practice. Only the extensive form of the game allows us to deal with this.

There seems to be, at present, two approaches to games in extensive, as opposed to normal, form. The first approach considers "supergames"—i.e. repeated plays of a static game.[2] This is a restricted type of dynamic game. In fact, it is not directly applicable to the charity problem, because that involves plays of games in which at least one player is rich in some periods, and poor in others—i.e., it is not the *same* game being repeated in different periods.[3] A second approach looks at slightly more general games, but considers only noncooperative equilibria.[4] This, too, misses the point of the charity game. For the noncooperative solution which is proposed, $B$ takes no account of whether $A$ was charitable before, when $B$ is deciding whether to be charitable. As we saw above, this gives no incentive for $A$ to be charitable.

For this reason, I have found it necessary to experiment with a new type of solution for dynamic games. It does, however, bear a certain resemblance to both of the above approaches.

The rest of the paper consists of an extended discussion of two particular dynamic games—the "Poverty Game" and the "Pension Game"—together with suggestions for treating more general games of this kind.

## THE "POVERTY GAME"

There are two players, $P_1$ and $P_2$. Time is discrete. In each odd-numbered period, $P_1$ is endowed with one chocolate, and $P_2$ has nothing.[5] In each

[2] See Luce and Raiffa (1957) (section 5.5.), Aumann (1959 and 1967), Friedman (1971).

[3] This is a little too sweeping. In some cases it may be possible to fit a charity problem into a supergame framework. See the "poverty game" below.

[4] The precise definition of noncooperative equilibrium is left until section II of this chapter. It has been common to identify it with Nash equilibrium. Here, I shall use noncooperative equilibrium in a somewhat different sense, as in Phelps and Pollak (1968) and Phelps (this volume). I am grateful to Menahem Yaari for drawing my attention to the distinction.

[5] "Chocolates" are suggested by Shell (1971).

even-numbered period, $P_2$ is endowed with one chocolate, and $P_1$ has nothing. The game lasts for $T$ periods, where $T$ is even.[6] Chocolate cannot be stored, but it can be costlessly and instantaneously transferred between the players. Each player has the same utility function:

$$U(c_1, \ldots, c_T) = \Sigma_{t=1}^{T} u(c_t)$$

where $c_t$ is the amount of chocolate he eats in period $t$. Here, $u$ is strictly concave, strictly increasing, differentiable, and $u'(c)$ tends to infinity as $c$ tends to zero.

In the absence of transfers, the players alternate between extreme hunger and comfortable over-eating. If transfers take place because of agreement, then we are not inclined to call them "charitable." But, any agreement can be broken, and so transfers might be regarded as charitable. In any case, we can ask whether they will take place.

Consider $P_2$ (who is the last to receive chocolate) in period $T$. He has a chocolate, but $P_1$ does not. There is no future. Why, then, would $P_2$ share his chocolate? Even if he promised $P_1$ a share, there is no incentive for him to keep his promise. So we may presume that an egoist would keep all his chocolate.[7] Now consider $P_1$ in period $T-1$. $P_1$ knows that $P_2$ is an egoist, and so, that $P_2$ will not give any chocolate in period $T$ whatever $P_1$ does in period $T-1$. So he also has nothing to gain, and his chocolate to lose, if he shares his chocolate with $P_2$.

Once again, if $P_1$ is an egoist, we may presume that he does not share his chocolate with $P_2$, even if he promised to do so. The same argument applies to $P_2$ in period $T-2$, and then to $P_1$ in period $T-3$, to $P_2$ in period $T-4$,..., and so on, right back to $P_1$ in period 1. Therefore, if $P_1$ and $P_2$ are both egoists, we may presume that chocolate transfers never take place. This is the noncooperative equilibrium to the poverty game (as is easily checked). It is also disastrously inefficient, in the Pareto sense.

Of course, this argument is not new. It is precisely that of Luce and Raiffa (*Games and Decisions*, pp. 97–102), who took exception to the conclusion. It is, of course, also a generalization of Hobbes. But, old as the argument is, there is no satisfactory suggestion for getting around it, if the horizon really is finite and known.[8]

If, however, the horizon is infinite, the game is totally different. There is no last period to start the argument going. Then it is possible to look for a cooperative equilibrium. Indeed, it is almost too easy to find one; a continuum of equilibria giving efficient outcomes will now be found.

[6] This makes the analysis a little easier without really affecting anything.

[7] We assume that $P_2$ is deaf to the rumblings of discontent from $P_1$ and $P_1$'s stomach, and that $P_2$ would beat $P_1$ in any chocolate fight.



Two streams of chocolate consumption—one for each player—are Pareto efficient if and only if each player's consumption is constant for all time, and no chocolate is discarded. To see the need for constancy, suppose that $P_1$ consumes $c_s$ in period $s$ and $c_t$ in period $t$, where $s \neq t$. $P_2$ consumes $1-c_s$ and $1-c_t$ in these two periods. A Pareto improving move occurs if $P_1$ consumes $\frac{1}{2}(c_s + c_t)$ in both the periods $s$ and $t$, and $P_2$ consumes the rest in both periods.

Let $\bar{x}$ denote $P_1$'s constant level of consumption, on some Pareto efficient path. Then $1-\bar{x}$ is $P_2$'s constant level of consumption. Let $x_t$ denote the chocolate transfer in period $t$ from the player who is rich that period to the player who has nothing that period. To maintain the levels of consumption, we have:

$$x_t = \begin{cases} 1-\bar{x} & (t \text{ odd}) \\ \bar{x} & (t \text{ even}) \end{cases}$$

It is easy to show that any such Pareto efficient sequence of transfers can be sustained by noncooperative equilibrium strategies. Indeed, consider the following pair of strategies, which are reminiscent of the supergame strategies studied by Aumann (1959 and 1967) and also Friedman (1971).
For $t$ odd, $P_1$'s strategy is:

$$x_t = \begin{cases} \bar{x} & (\text{if } x_k \geqslant 1 - \bar{x} \text{ for all odd } k \text{ such that } k < t) \\ 0 & (\text{otherwise}). \end{cases}$$

For $t$ even, $P_2$'s strategy is:

$$x_t = \begin{cases} 1 - \bar{x} & (\text{if } x_k \geqslant \bar{x} \text{ for all even } k \text{ such that } k < t, \text{ or if } t = 1) \\ 0 & (\text{otherwise}) \end{cases}$$

Each player cooperates as long as the other does—but as soon as one player breaks the agreement, it collapses completely. There is consequently a tremendous incentive to keep to the agreement. Of course, the proposed strategies are in strong equilibrium (see Aumann, 1967).

[8] As Christopher Bliss has pointed out to me, the argument is uncomfortably close to that in the "Unexpected Examination" Paradox—see, for example, Nerlich (1961). In fact, $P_1$ can make the following statement to $P_2$:

> For some $t$, provided that in each previous period when you had chocolate you shared it with me, I shall share my chocolate with you on the first $t$ occasions when I have chocolate. Immediately after you stop sharing chocolate, I shall stop. After $t$ occasions, I shall stop sharing anyway. You will not succeed in anticipating the period in which I stop sharing chocolate—if I am the first to stop sharing.

This is just like the Unexpected Examination. $P_2$ cannot disprove this statement. Does he share his chocolate, in the hope that $P_1$'s chocolate sharing will continue?

Consider what form cooperation takes in this game. Somehow, a value of $\bar{x}$ is hit upon, through bargaining, threats, etc. Maybe $\bar{x} = ½$ is agreed upon because it is generally seen to be fair. Then, each player cooperates in the sense that he holds the right beliefs, and acts upon them. It is rational for $P_1$ to choose his equilibrium strategy if and only if he believes that $P_2$ will choose his equilibrium strategy, and vice versa. If each player believes that the other will be uncooperative, no transfers occur.

## THE "PENSION GAME" – GENERAL DISCUSSION

This game is based on an exchange economy first discussed by Samuelson (1958) and, more recently, by Shell (1971). There are an infinite number of players, $P_t$ ($t$=0,1,2,...).

In period $t$ ($t$=1,2,3,...) $P_t$ is endowed with one chocolate, and no other player has any endowment. Each player (including $P_0$) has utility function:

$$U_t \equiv u(c^t_t) + u(c^t_{t+1})$$

where $c^t_k$ is $P_t$'s consumption of chocolate in period $k$. The function $u$, and the physical properties of chocolate, are exactly the same as in the poverty game.

This model has a fairly obvious interpretation. $P_t$ lives in period $t$, when he is young and earns a chocolate—and in period $t+1$, when he is old and earns no chocolate, but wants chocolate to eat. In period $t+1$, $P_t$ relies on a chocolate pension, which only $P_{t+1}$ can provide for him.

If the game stops in period $T$, the solution is as miserable as in the poverty game with a finite horizon. Every old man goes without chocolate. What happens if the game never stops?

It seems that we run straight into Phelps's indeterminacy problem.[9] $P_t$'s action is determined by his beliefs about $P_{t+1}$'s reaction. There is an infinite chain of successive expectations, and there is no way of breaking it, or getting to the end. Let us assume, for simplicity, that each player $P_t$ considers only two possible actions—giving $P_{t-1}$ half a chocolate, or giving $P_{t-1}$ nothing. Then, the typical link of the chain can be described as follows. Should $P_t$ believe that he will receive a pension from $P_{t+1}$ *if and only if* he pays $P_{t-1}$ a pension, $P_t$ will surely pay a pension. Should, however, $P_t$ believe that $P_{t+1}$ will pay $P_t$ a pension, whatever $P_t$ does (because of $P_{t+1}$'s expectations about $P_{t+2}$'s reactions), then $P_t$ has no incentive to pay a pension. Similarly, if $P_t$ believes that $P_{t+1}$ will not pay $P_t$ a pension, whatever $P_t$ does (because of $P_{t+1}$'s

[9] See Phelps (this volume).



expectations about $P_{t+2}$'s expectations concerning $P_{t+3}$'s reactions to $P_{t+2}$'s decision), then $P_t$ again has no incentive to pay a pension.

Clearly, then, in order to determine $P_t$'s beliefs, we need to determine $P_{t+1}$'s strategy. Suppose first that the game is a stationary game—i.e. one in which each player's problem at each moment of time looks precisely the same. Then we expect any equilibrium to be stationary. In other words, $P_t$ and $P_{t+1}$ will have effectively identical strategies. Then, we can look for some kind of fixed point $s^*$ in strategy space. The fixed point has the property that, if $P_t$ expects $P_{t+1}$ to adopt $s^*$, then it is optimal for $P_t$ to adopt $s^*$. But the multiplicity of these fixed points is *precisely* the Phelps problem.[10]

Even in a nonstationary game, we can define equilibrium strategy sequences. A strategy sequence $s = (s_1, s_2, ...)$ is a list of strategies, one for each player. The sequence $s^*$ is a noncooperative equilibrium if and only if, for all $t$, $s^*_t$ is $P_t$'s optimal strategy, given that he expects the players who follow him—$P_k$ ($k>t$)—to adopt the strategies $s^*_k$ ($k>t$). Of course, there are just as likely to be multiple equilibria in this case as in the stationary case.

It seems that this problem comes about because there are not enough restrictions on an equilibrium. The only restriction put on an equilibrium has been that of consistency—if $s^*$ is an equilibrium, then, *given that each player expects* $s^*$, there should be no incentive for any player to depart from $s^*$. But, even if $s^*$ *is* an equilibrium in this sense, why should each player expect $s^*$, if other equilibria are possible? In particular, if $\hat{s}$ is also a solution, and each player is better off with $\hat{s}$ than with $s^*$, why should $s^*$ be expected? There is every incentive for each player to expect $\hat{s}$ rather than $s^*$.

In the pension game, for example, there is obviously, in one sense, an "equilibrium" in which pensions are never paid. But, if, as will be seen later, there is also an equilibrium in which pensions *are* paid, and then every player is better off, surely it is rational to pay pensions.

## EQUILIBRIA IN THE PENSION GAME

Let $x_t$ denote the amount of chocolate which $P_t$ donates to $P_{t-1}$ (at time $t$, of course). Let $h_t$ denote the historical sequence of pensions preceding time $t$—i.e. $h_t = (..., x_{t-2}, x_{t-1})$. Then we expect an equilibrium strategy to be a function of the form:

$$x^*_t = f(h_t).$$

[10] It is interesting that the problem does not arise in Phelps and Pollak (1968). But I believe that this is because they only considered constant saving ratio policies, which were quite independent of the past. In fact, if more general policies are considered—in particular, ones which do depend on the past—then many more solutions are possible, I believe.

Let us assume for the moment that the game had a definite starting date, which we may as well take to be one.[11] Then, for $h_t = (x_1, \ldots, x_{t-1})$, define

$$k^*(h_t) = \begin{cases} 0 \text{ (if } x_k < \tfrac{1}{2} \ (k = 1, 2, \ldots, t-1) \\ \max \{k \mid x_k \geq \tfrac{1}{2}\} \text{ (otherwise).} \end{cases}$$

$k^*(h_t)$ was the last date on which a fair pension was paid—or else is zero if a fair pension was never paid.

Define $n(h_t) = t\text{-}k^*(h_t)-1$—the number of periods in which unfair pensions have been paid, since a fair pension was last paid.

Now consider the strategy:

$$f(h_t) = \begin{cases} 0 \text{ (if } n(h_t) \text{ is odd)} \\ \tfrac{1}{2} \text{ (if } n(h_t) \text{ is even) (including 0)} \end{cases}$$

This demands an explanation. Suppose $x_{t-1} \geq \tfrac{1}{2}$. Then the strategy tells $P_t$ to pay $P_{t-1}$ a fair pension. $P_{t-1}$ has been fair—so $P_t$ is fair to $P_{t-1}$. Then, moreover, every succeeding player gives and receives a fair pension. Suppose $x_{t-2} \geq \tfrac{1}{2}$, $x_{t-1} < \tfrac{1}{2}$. Then $P_{t-1}$ has been "unfair" to $P_{t-2}$, and so $P_{t-1}$ is to be punished. Suppose $x_{t-3} \geq \tfrac{1}{2}$, $x_{t-2} < \tfrac{1}{2}$, $x_{t-1} = 0$. Then $P_{t-2}$ has been unfair to $P_{t-3}$, and $P_{t-1}$ has punished him. Now $P_{t-1}$ is not to be punished—rather, he is to be rewarded for punishing $P_{t-2}$'s unfairness. We can go on like this, working out what happens for various histories $h_t$. But the idea is that, if $n(h_t)$ is odd, then $P_{t-1}$ has been unfair to $P_{t-2}$, either because $x_{t-2} \geq \tfrac{1}{2}$, or because $P_{t-2}$ punished $P_{t-3}$, who had been unfair to $P_{t-4}$, either because $x_{t-4} \geq \tfrac{1}{2}$, or because $P_{t-4}$ punished $P_{t-5}, \ldots$. On the other hand, if $n(h_t)$ is even (including 0), then either $x_{t-1} \geq \tfrac{1}{2}$, or $P_{t-1}$ punished $P_{t-2}$, who had been unfair to $P_{t-3}$—either because $x_{t-3} \geq \tfrac{1}{2}$, or because $P_{t-3}$ punished $P_{t-4}, \ldots$.

Notice that, if $n(h_t)$ is odd, $P_t$ has nothing to gain by choosing $x_t > 0$—provided that all later players follow the strategy $f$—because he secures a pension of $\tfrac{1}{2}$ by choosing $x_t = 0$. Conversely, if $n(h_t)$ is even (or zero), $P_t$ has nothing to gain by choosing $x_t \neq \tfrac{1}{2}$; if $x_t > \tfrac{1}{2}$, he secures no larger pension than $\tfrac{1}{2}$, and if $x_t < \tfrac{1}{2}$, then he is severely punished by $P_{t+1}$. Thus, we have proved:

*Lemma* 1: *The strategy sequence* $x_t^* = f(h_t)$ $(t=1,2, \ldots)$ *is a noncooperative equilibrium for the pension game.*

[11] To maintain stationarity, we had better assume that at time 1 there was a player $P_0$ who never experienced youth.



In other words, if all players after $P_t$ adopt $f$ (the players before $P_t$ do not affect him), then $P_t$ can do no better than he does by adopting $f$. So $f$ is a noncooperative equilibrium in the pension game. Moreover, the outcome of adopting $f$, after history $h_t$, is

$$(0,\tfrac{1}{2},\tfrac{1}{2},\tfrac{1}{2}, \ldots) \quad (\text{if } n(h_t) \text{ is odd})$$

$$(\tfrac{1}{2},\tfrac{1}{2},\tfrac{1}{2},\tfrac{1}{2}, \ldots) \quad (\text{if } n(h_t) \text{ is even}).$$

This outcome—at least after period $t$—is clearly all that can be desired. It is both Pareto efficient and perfectly egalitarian. But starving the old is another "equilibrium" outcome. Is there no way of persuading rational men that they *must* choose $f$, without appealing to ethical arguments?

Let $f'$ denote the strategy of starving the old. It is clear that $f$ dominates $f'$, in the sense that everybody (except perhaps $P_{t-1}$, who is at worst indifferent) prefers the outcome of $f$, given $h_t$, to the outcome of $f'$. This alone seems enough to ensure that $f'$ will never be chosen—in fact, we should not regard it as an "equilibrium" at all.

Since few would quibble at the desirability of $f$ for the pension game, it seems proper to enquire precisely what properties $f$ has which make it an "equilibrium."

There are an embarrassingly large number of different types of "equilibrium" in cooperative games. But all of them require, by definition, some form of agreement between the members of a coalition. The orthodox view of coalitions is that they are bodies of agents who make binding agreements.[12] These agreements specify a coordinated strategy for the coalition as a whole—and so, a strategy for each member of the coalition. The agreements are reached because a coordinated strategy is better for each member of the coalition than the alternative uncoordinated strategy.

But there is no mention of how agreements are to be enforced. It seems that cooperative game theory is either a theory of honorable agents, or else, implicit in any agreement is the possibility of drastically punishing any member of a coalition who fails to keep his word. It seems clear therefore that such a theory is not applicable to the pension game. The agents are not necessarily honorable, and the possibilities of punishment are completely explicit. Does this imply that cooperation is impossible?

I have already suggested that, in the poverty game, cooperation takes the form of holding appropriate beliefs about the other player's strategy.[13] This kind of cooperation is perfectly possible in the pension game as well. $P_t$ can be

[12] This "orthodox view" is stated, for example, in Von Neumann and Morgenstern (1953), section 21.2, and in Luce and Raiffa (1957), chapter 6.

[13] Von Neumann and Morgenstern's "understanding" seems an appropriate term here.

cooperative by believing that $P_{t+1}$ will choose the strategy $f$ (because $P_{t+1}$ believes that $P_{t+2}$ will choose $f$, because $P_{t+2}$ believes that $P_{t+3}$ will choose $f$, because . . . .); then $P_t$ will choose $f$. On the other hand, $P_t$ can be uncooperative by believing that $P_{t+1}$ will choose $f'$ (because $P_{t+1}$ believes that $P_{t+2}$ will choose $f'$, because $P_{t+2}$ believes that $P_{t+3}$ will choose $f'$, because . . . .); then $P_t$ will choose $f'$.

I shall assume that all players are willing to cooperate by holding appropriate beliefs. This is, of course, an additional demand on the rationality of "rational egoists," and one that is at least as much open to dispute as the other demands.[14] But it seems no stronger than the traditional assumption of cooperative game theory, that egoists join coalitions and keep agreements.

It remains to decide what are "appropriate beliefs." This will be attempted by defining "dynamic equilibrium." It is easier to treat a generalized form of pension game first. So the solution of the pension game will be left until section 7.

## GAMES WITH SEQUENTIAL PLAYERS

Consider a game $\Gamma$ in which no player has more than one move. There may be a finite or an infinite number of players.

The assumption that each player has no more than one move is less restrictive than it first seems. If a player $i$ has more than one move, he can be regarded as a number of players, or "egos," each of whom has the same preferences, and has no more than one move. By regarding a player with more than one move in this way, we may overlook possibilities for cooperation between the "egos" of a single player. But I hope to be able to show, in later work, that even this is not a major problem.

We shall also assume perfect information and no uncertainty. At each moment of time $t$, there will have been a sequence of moves up to time $t$; $t$ and this sequence together determine

(i) The set of feasible move sequences which include the game.
(ii) The player who is to move at time $t$.
(iii) The preferences of the player who is to move at time $t$.

The time and earlier move sequence can be summarized in a history vector $h$—the dimension of $h$ may depend upon $t$. Given $h$, let the feasible set of concluding move sequences be $X(h)$, and let the preferences of the player who is to move be described by the weak preference relation $R(h)$ and the strict preference relation $P(h)$. It will be assumed that $R(h)$ is transitive, as well as reflexive and connected.

[14] The additional demand lies at the heart of Schelling's discussion of coordination and tacit bargaining. See Schelling (1960).



Suppose that, given the history $h$, the player to move makes a move $m$. This results in a new history vector $h'$ which depends only on $h$ and $m$. Write $h' = \psi(m, h)$. $\psi$ is the *history updating function.*

Now, a game $\Gamma$ which is completely described by the sets $X(h)$, the preorders $R(h)$, and the history updating function $\psi$, will be called a *game with sequential players.* Examples are the pension game, and the game studied in Phelps and Pollak (1968). The game in Phelps (this volume) is a similar type of game in continuous time—such games will not be considered here, although much of the analysis is probably applicable to them. Note also that a finite game with sequential players is a particular type of the extensive games studied by Kuhn.

For each history $h$, there is a subgame $\Gamma(h)$ involving move sequences which conclude $\Gamma$. We expect the solution to $\Gamma$, and to each of its subgames, to involve *strategy sequences* $s(.)$ specifying the move which is to be made after history $h$, for each possible $h$. Associated with each strategy sequence is a *history generating function* $\phi$. Given the strategy sequence $s(.)$, $\phi(s, h)$ specifies the history which results when $s$ is followed for a single period. In fact, given the history updating function $\psi$, $\phi$ must satisfy

$$\phi(s, h) = \psi(s(h), h)$$

for all possible $s(.)$ and $h$. It is convenient to have an expression for the history which results from following $s(.)$ for $k$ periods, starting from $h$. This will be written as $\phi^k(s, h)$.

By convention, $\phi^0(s, h) = (\text{all } s(.), h)$
and $\phi^1(s, h) = \phi(s, h)$ $(.,,)$
Of course:- $\phi^2(s, h) = \phi(s, \phi(s, h))$
$\phi^3(s, h) = \phi(s, \phi(s, \phi(s, h)))$, etc.

There is also an *outcome function* $x(s, .)$ associated with each $s(.)$. The outcome function specifies the entire move sequence which results from following $s(.)$ for ever, starting with $h$. In fact, $x(s, h)$ can be regarded as $\phi^\infty(s, h)$ (or as $\phi^T(s, h)$, if the outcome has only a finite number of moves before the end of the game).

It is convenient to have notations for the set of feasible strategy sequences, and the set of feasible histories. So define

$$S(h) = \{ s(.) \mid x(s,h) \in X(h) \}$$

the set of feasible strategy sequences, given history $h$. Define

$$H(h) = \{ h' \mid \exists k \geq 0 \text{ and } s \in S(h) \text{ s.t. } h' = \phi^k(s,h) \}$$

the set of histories—including $h$ itself—which can follow $h$. Define

$$\tilde{H}(h) = \{ h' \mid \exists\, k > 0 \text{ and } s \in S(h) \text{ s.t. } h' = \phi^k(s,h) \}$$

the set of histories—not including $h$ itself—which can follow $h$.

## DYNAMIC EQUILIBRIA

This paper has already moved far from its real subject, and a full discussion of dynamic equilibrium would take us even further astray.

Nevertheless, in connection with the poverty game, it was suggested that the noncooperative equilibrium is an appropriate type of solution for a finite horizon. The definition of "dynamic equilibrium" which follows is therefore designed to reduce to the noncooperative equilibrium when the horizon is finite. When the horizon is infinite, cooperative beliefs are embodied in the definition.

With this in mind, consider what a dynamic equilibrium must look like. For each subgame $\Gamma(h)$, dynamic equilibria will be strategy sequences in $S(h)$. If $s^*$ is an equilibrium, for each $h' \epsilon H(h)$, the player at $h'$ must be willing to make his prescribed move $s^*(h')$, given his expectations about the strategies of his successors. Moreover, as there is perfect information and no uncertainty, these expectations must be fulfilled, in an equilibrium.

These considerations give us two properties of dynamic equilibria immediately. The first is that, if $s^* \epsilon E(h)$, then the player at $h$ must be expecting his successors to follow $s^*$. Such expectations only make sense if $s^*$ is an equilibrium from his successors' points of view as well. So, if $s^* \epsilon E(h)$, then, for all $h' \epsilon H(h)$, $s^* \epsilon E(h')$. Secondly, $s^*(h)$ must be the optimal reaction of the player at $h$, if he expects his successors to follow $s^*$. As this is true for all $h$, it follows that any dynamic equilibrium is also a noncooperative equilibrium.

These two considerations alone are sufficient to determine the dynamic equilibrium for a finite game with sequential players. It is any noncooperative equilibrium. But, in infinite games, as Phelps has clearly shown, indeterminacy is a major problem.

The definition of dynamic equilibrium will make use of a "dynamic dominance" relation $D(h)$ defined on the set $S(h)$ of strategy sequences as follows:

$s_1 D(h) s_2$ if and only if

(i) $x(s_1,h)P(h)x(s_2,h)$, and $s_1(h) \neq s_2(h)$

(ii) For all $h' \epsilon \tilde{H}(h)$, either (a) $x(s_1,h')P(h')x(s_2,h')$
or (b) $s_1(h') = s_2(h')$
(or both).



Notice that $D(h)$ is an asymmetric relation—i.e. if $s_1 D(h) s_2$, then not $s_2 D(h) s_1$. Also:

*Lemma* 2: *If* $s_1 D(h) s_2$, $h' \epsilon \tilde{H}(h)$, *and* $s_1(h') \neq s_2(h')$ *then* $s_1 D(h') s_2$.

*Proof*: Immediate.

The relationship of $D(h)$ to noncooperative equilibria is shown in the following two theorems.

*Theorem* 1: *Suppose that* $s^*$ *is a strategy sequence for which there is no* $h$ *and no* $s \epsilon S(h)$ *such that* $s\, D(h)\, s^*$. *Then* $s^*$ *is a noncooperative equilibrium.*

*Proof*: Suppose $s^*$ is not a noncooperative equilibrium. Then there exists an $h_0$ and an $\hat{s} \epsilon S(h_0)$ such that $x(\hat{s}^*, \phi(\hat{s}, h_0))\, P(h_0)\, x(s^*, h_0)$ and $s^*(h_0) \neq \hat{s}(h_0)$.

Define $s(h) = \begin{cases} \hat{s}(h) & (h = h_0) \\ s^*(h) & (\text{otherwise}) \end{cases}$

Then it is easy to check that $s\, D(h_0)\, s^*$.

*Theorem* 2: *Suppose that there is a finite upper bound,* $N$, *on the possible number of moves in the game. Then, if there exists* $h$ *and* $s \epsilon S(h)$ *such that* $s\, D(h)\, s^*$, $s^*$ *cannot be a noncooperative equilibrium.*

*Proof*: The proof will be by induction on $N(h)$, the maximum possible number of moves in the subgame $\Gamma(h)$.

If $N(h) = 1$, then $s\, D(h)\, s^*$ implies $x(s,h)\, P(h)\, x(s^*, h)$. Since the player at $h$ has the last move, or no move at all, it is obvious that $s^*$ is not a noncooperative equilibrium.

Suppose that the result is true whenever $h$ satisfies $N(h) \leqslant n$. Consider any $h$ such that $N(h) = n+1$. Suppose that, although $s\, D(h)\, s^*$, $s^*$ is a noncooperative equilibrium strategy sequence for the subgame $\Gamma(h)$. Then, for all $h' \epsilon \tilde{H}(h)$, $s^*$ is a noncooperative equilibrium for the subgame $\Gamma(h')$. But $N(h') \leqslant n$, so, by the induction hypothesis, $s\, D(h')\, s^*$ is impossible. Together with $s\, D(h)\, s^*$, this implies that, for all $h' \epsilon \tilde{H}(h)$, $s(h') = s^*(h')$. But $x(s, h)\, P(h)\, x(s^*, h)$. Therefore $s(h) \neq s^*(h)$, and so $s^*$ is not a noncooperative equilibrium for the subgame $\Gamma(h)$. Therefore the result is true for $N(h) \leqslant n+1$. This is what remained to be proved.

So, for any *finite* game with sequential players—i.e. one in which the possible number of moves has a finite upper bound—a strategy sequence $s^*$ is a

noncooperative equilibrium if and only if there is no $h$ and no $s \in S(h)$ such that $s\ D(h)\ s^*$. But this has not been shown for an infinite game. Indeed, it is false. It is not hard to see that, for the pension game, the strategy $f$ dominates the strategy $f'$ of always starving the old. In fact, for all histories $h$, $x(f, h)\ P(h)\ x(f', h)$. But $f'$ is a noncooperative equilibrium.

This suggests that the dominance relation may give what is required for cooperative expectations.

Say that $s^*$ is a *dynamic equilibrium* for the subgame $\Gamma(h)$ if and only if there exists no $h' \in H(h)$ and no $s \in S(h)$ such that $s\ D(h')\ s^*$. Write $E(h)$ for the set of dynamic equilibria for $\Gamma(h)$.

Is this definition plausible?

Suppose that the player to move at $h$ is trying to convince himself that $s^*$ is not an equilibrium strategy. He must show that, for some $h' \in H(h)$, $s^* \notin E(h')$. He might argue that $s^* \notin E(h')$ because there is some $h'' \in \tilde{H}(h')$ such that $s^* \notin E(h'')$, but this is just delaying the finding of a proper reason. Really, the player at $h$ must have in mind an alternative strategy sequence $s \in S(h)$. Ideally, this should be an equilibrium, but such a restriction leads to an infinite regress. So let us see what happens if $s$ is not required to be an equilibrium.

Suppose that the player at $h$ can find an $s \in S(h)$ such that $s\ D(h)\ s^*$. Then $x(s, h)\ P(h)\ x(s^*, h)$, and so he would prefer to regard $s$ rather than $s^*$ as an equilibrium. As $s(h) \neq s^*(h)$, the player at $h$ is tempted to follow $s$ rather than $s^*$. He can safely yield to this temptation, it seems, unless a later player will be subject to the reverse temptation—to follow $s^*$ instead of $s$. But if $s\ D(h)\ s^*$, no later player will be subject to this reverse temptation.

In fact, no later player will be subject to this reverse temptation if (in addition to $s(h) \neq s^*(h)$, $x(s, h)\ P(h)\ x(s^*, h)$), for all $k > 0$, either

(a) $x(s, \phi^k(s, h))\ P(\phi^k(s, h))\ x(s^*, \phi^k(s, h))$ or

(b) $s(\phi^k(s, h)) = s^*(\phi^k(s, h))$

(or both)

so the condition $s\ D(h)\ s^*$ may appear too strong. However, the extra strength is illusory:

*Lemma* 3: *Suppose that there exists* $s \in S(h)$ *such that*

*(i)* $s(h) \neq s^*(h)\ x(s, h)\ P(h)\ x(s^*, h)$

*(ii) For all* $k > 0$, *either* (a) $x(s, \phi^k(s, h))\ P(\phi^k(s, h))\ x(s^*, \phi^k(s, h))$

*or* (b) $s(\phi^k(s, h)) = s^*(\phi^k(s, h))$

*(or both)*



*Then there exists* $\hat{s} \in S(h)$ *such that* $\hat{s}\ D(h)\ s^*$.

*Proof*: Define $\hat{s}(h') = \begin{cases} s(h') & \text{(if } h' = \phi^k(s, h), \text{ for some } k \geq 0) \\ s^*(h') & \text{(otherwise).} \end{cases}$

Now, if $\hat{s}(h') \neq s^*(h')$, $x(s, h')\ P(h')\ x(s^*, h')$. Then it is easy to confirm that $\hat{s}\ D(h)\ s^*$.

For $s^* \in E(h)$, it is necessary and sufficient that no player in the subgame $\Gamma(h)$ be tempted to change his strategy. So, the argument above supports the definition of dynamic equilibrium.

Finally, what are "appropriate beliefs"? They are simply that each player believes that his successors will not follow dominated strategies. Since it is certainly against the relevant players' interests to follow a dominated strategy, such beliefs seem sensible.

## DYNAMIC EQUILIBRIA IN THE PENSION GAME

For the pension game, Lemma 1 shows that the strategy $f$ is a noncooperative equilibrium. Here, it will be shown that $f$ is a dynamic equilibrium.

*Lemma* 4: *In the pension game, there is no history* $h_t$ *and no strategy* $g$ *such that* $g\ D(h_t)\ f$.

*Proof*:

(1) Suppose $g\ D(h_t)\ f$.

Write $g_k = g(\phi^{k-1}(g, h_t))$

$f_k = f(\phi^{k-1}(g, h_t)) \quad (k = 1,2,3, \ldots)$

If history is $\phi^{k-1}(g, h_t)$, and $g$ is followed, the player at $h_t$ has consumption $(1-g_k, g_{k+1})$ But if $f$ is followed, he has $(1-f_k, \tfrac{1}{2})$.

Then $u(1-g_1) + u(g_2) > u(1-f_1) + u(\tfrac{1}{2})$, and $g_1 \neq f_1$.

Also, $u(1-g_k) + u(g_{k+1}) > u(1-f_k) + u(\tfrac{1}{2})$, or $g_k = f_k \quad (k = 2,3,4, \ldots)$

(2) Suppose that, for some $k$, $g_k > \tfrac{1}{2}$ and $f_k = \tfrac{1}{2}$.

Then, for $u(1-g_k) + u(g_{k+1}) > 2u(\tfrac{1}{2})$ to be true, we must have $g_{k+1} > g_k$, because $\tfrac{1}{2}u(1-g_k) + \tfrac{1}{2}u(g_{k+1}) \leq u(\tfrac{1}{2}(1-g_k) + \tfrac{1}{2}g_{k+1})$. In particular $g_{k+1} > \tfrac{1}{2}$ and also $f_{k+1} = \tfrac{1}{2}$. Therefore, by induction, for all $m \geq k$,

$$g_{m+1} > g_m > \tfrac{1}{2} \text{ and } f_m = \tfrac{1}{2}.$$

Since the sequence $g_m$ is clearly bounded above by 1, it has a limit $\hat{g}$. As $u$ is continuous, we must have $u(1-\hat{g}) + u(\hat{g}) \geqslant 2u(\tfrac{1}{2})$. But, because $\hat{g} > \tfrac{1}{2}$ and $u$ is strictly concave, this is impossible.

It follows that, if $g\ D(h_t)\ f$, there can be no $k$ such that $g_k > \tfrac{1}{2}$ and $f_k = \tfrac{1}{2}$.

(3) Suppose that $n(h_t)$ is odd. Then $f_1 = 0$.

To get $u(1-g_1) + u(g_2) > u(1) + u(\tfrac{1}{2})$, we clearly need $g_1 < \tfrac{1}{2}$ and $g_2 > \tfrac{1}{2}$. But, since $g_1 < \tfrac{1}{2}$, $n(\phi(g, h_t))$ is even, and so $f_2 = \tfrac{1}{2}$. So $g_2 > \tfrac{1}{2}$ and $f_2 = \tfrac{1}{2}$. By (2), it follows that $g\ D(h_t)\ f$ is false.

(4) Suppose that $n(h_t)$ is even. Then $f_1 = \tfrac{1}{2}$.

(a) Suppose that $g_1 \geqslant \tfrac{1}{2}$. Then we need $g_2 > \tfrac{1}{2}$. But $f_2 = \tfrac{1}{2}$. So, by (2), $g\ D(h_t)\ f$ is false.

(b) Suppose that $g_1 < \tfrac{1}{2}$. Then $f_2 = 0$. Also, for $u(1-g_1) + u(g_2) > 2u(\tfrac{1}{2})$, we need $g_2 > 0$. So $f_2 \neq g_2$. We therefore need $g\ D(\phi(g, h_t))\ f$. Since $n(\phi(g, h_t))$ is odd, this is impossible, by (3).

## CONCLUSION

It seems that this paper has strayed far from a theory of charitable behavior. In addition, it is dangerous to draw conclusions from an analysis of very special cases. But it does seem that the initial hypothesis—that some charitable behavior could arise even in a world of total egoists, provided these egoists have appropriate expectations—has been confirmed for a special case. How generally this explanation might work is not yet known. Moreover, it obviously ignores the most important questions which charitable behavior raises. Do we really believe that charity is to be explained by cooperative egoism?

But, if it is shown that some kinds of behavior, usually regarded as altruistic, can in fact arise from egoism, the implications are in any case more normative than positive. In particular, it is a weapon to be used in arguments about certain issues of public policy. For example, the wealthy may object to paying taxes which are required to finance state benefits. But they may object less if they know that the level of benefits in the future, when they may need them, is related to the level today.



## REFERENCES

Aumann, R. J., "Acceptable Points in General Cooperative *n*-person Games," *Contributions to the Theory of Games IV* (Princeton: Princeton University Press, 1959), pp. 287–324.

Aumann, R. J., "A Survey of Cooperative Games Without Side Payments," M. Shubik (ed.), *Essays in Mathematical Economics in Honor of Oskar Morgenstern* (Princeton: Princeton University Press, 1967), pp. 3–27.

Friedman, J. W., "A Non-Cooperative Equilibrium for Supergames," *Review of Economic Studies* 38 (1), Jan. 1971, pp. 1–12.

Kuhn, H. W., "Extensive Games and the Problem of Information," *Contributions to the Theory of Games II* (Princeton: Princeton University Press, 1953), pp. 193–216.

Luce, R. D., and Raiffa, H., *Games and Decisions* (New York: John Wiley & Sons, Inc. 1957).

Nerlich, G. C., "Unexpected Examinations and Unprovable Statements," *Mind* 70, Oct. 1961, pp. 503–13.

Phelps, E. S., "The Indeterminacy of Game-Equilibrium Growth in the Absence of an Ethic" (this volume).

Phelps, E. S., and Pollak, R. A., "On Second-Best National Saving and Game-Equilibrium Growth," *Review of Economic Studies* 35 (2), April 1968, pp. 185–200.

Samuelson, P. A., "An Exact Consumption Loan Model of Interest with or without the Social Contrivance of Money," *Journal of Political Economy* 66 (6), Dec. 1958, pp. 467–82.

Schelling, T. C., *The Strategy of Conflict* (Cambridge: Harvard University Press, 1960).

Shell, K., "Notes on the Economics of Infinity," *Journal of Political Economy* 79 (5), Sept./Oct. 1971, pp. 1002–11.

Von Neumann, J., and Morgenstern, O., *Theory of Games and Economic Behavior*, third edition, (Princeton: Princeton University Press, 1953).