Soc Choice Welfare (1985) 2:207–219

Social Choice and Welfare

Anarchy, Altruism and Cooperation*

A Review

N. Schofield

Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, USA

Received June 10, 1985 / Accepted June 13, 1985

1. Collective Action

The problem of collective action is concerned with the analysis of whether common goals of a group of individuals may be attained when the individuals behave rationally with respect to their own varied preferences. Olson's book on the logic of Collective Action [14] published in 1965, is very probably the most influential exposition of this problem. Olson argues that rational, self interested members of a group are unlikely to take onto themselves the costs that, in general, they must accept if the group is to succeed in its purpose. More precisely Olson considers the case where a collective good can be created for the "consumption" of a group. A collective good for some group is nothing more than some aspect of the state of the world that all members of the group wish to see brough about, with the property that once the good is created, no member may be excluded from its enjoyment or consumption. However, the good is costly to create. If the costs are shared equally then for each individual the benefit associated with the good exceeds the cost. But if the group is unable to exert coercion an individual may decline to accept his share of the cost. He nonetheless receives benefit from provision of the good even though it might be produced perhaps at a lower level. When his marginal loss of the good (induced by his failure to contribute) is less than his fair share of the cost, then it is not in his interests to contribute. Since every member may perform the same calculation, it is unlikely that anyone will contribute. Consequently, the good will not be provided, and the members will be worse off than they might otherwise have been.

It was very early realized that the logic of this problem was parallel to, maybe even identical with, the problem of cooperation in the *prisoner's dilemma* (PD). In this two-person game each player may choose either to cooperate or to defect.

* I am indebted to Brian Barry, Toby Page, Jim Woodward and especially Will Jones for helpful discussion in the study group at Caltech where this paper was presented. Howard Margolis kindly offered a number of useful comments. All errors in interpretation are mine alone

Depending on the strategies they adopt, their payoffs (with Row's payoffs first) are as in the array:

Table 1.

		Column	
		Cooperate	Defect
Row	Cooperate	$(R=3, R=3)$	$(S=0, Q=5)$
	Defect	$(Q=5, S=0)$	$(P=1, P=1)$
			$Q>R>P>S$

In a general PD any numbers may be used for Q, R, P, S as long as $Q>R>P>S$. Whether Column cooperates or defects, Row prefers defection (since $Q>R$ and $P>S$ respectively). Indeed, both Row and Column defect to produce (P, P). On the other hand both players prefer the state of the world (R, R) brought about when both cooperate (to ensure this it is usual to assume $2R>S+Q$). That is to say, rational behavior does not lead to the "optimal" outcome (R, R), and thus we obtain a paradox: that cooperation is unlikely in the prisoners' dilemma.

In this essay I shall discuss the general conclusions that were drawn from analysis of the prisoners' dilemma up until about 1980 and then focus on four books (Axelrod [3], Hardin [7], Margolis [9] and Taylor [19]), all of which appeared between 1982 and 1984 and deal, in one way or another, with the problem of cooperation.

2. The n-Person Prisoners' Dilemma

For the purpose of exposition it is useful to take a version of the n-person prisoners' dilemma (nPD) due to Hardin [6]. Let $N=\{1,\ldots,n\}$ be some society and suppose that each individual, i, has available a strategy $d_i \in [0,1]$ where we may think of d_i as the contribution of i to the provision of the good. If the individuals adopt a list of strategies $\mathbf{d}=(d_1,\ldots,d_n)$ then the net benefit to i is taken to be

$$a_i(\mathbf{d})=r/n(\sum_N d_j)-d_i. \tag{1}$$

Here r may be regarded as the ratio of social benefits to social costs (and we assume $r>1$). We suppose that i only takes into account the consequences of his own choice when determining his strategy, and to indicate this we shall write $a_i(d_i, d_{-i})$ for $a_i(\mathbf{d})$. Clearly

$$\begin{aligned} a_i(d_i, d_{-i}) &= r/n(d_i+\sum_{j\neq i} d_j)-d_i \\ &= d_i(r/n-1)+r/n\sum_{j\neq i} d_j. \end{aligned}$$

Suppose further that $r<n$. Then it is evident that $a_i(d_i, d_{-i})$ is maximized, with respect to $d_i \in [0,1]$, when $d_i=0$. If each individual adopts this *dominant strategy* $(d_i=0)$ then $\mathbf{d}=0$ and $a_i(\mathbf{d})=0$. On the other hand if $d_i=1$ for all i then $a_i(\mathbf{1})=r-1$.

If we further assume that $r > 1$, then as in the two person dilemma, it is the case that the consequence of the joint cooperative strategy ($d_i = 1$ for all i) is preferred by all i to the consequence of the joint defect strategy ($d_i = 0$ for all i). A further point worth making is that the n-person dilemma contains the two person dilemma. Consider two individuals $\{1,2\}$ and let $x = \sum_{i \neq 1,2} d_i$. Then we obtain

Table 2.

		Column	
		$d_2 = 1$	$d_2 = 0$
Row	$d_1 = 1$	$\left(\frac{2r}{n} - 1 + \frac{rx}{n}, \frac{2r}{n} - 1 + \frac{rx}{n}\right)$	$\left(\frac{r}{n} - 1 + \frac{rx}{n}, \frac{r}{n} + \frac{rx}{n}\right)$
	$d_1 = 0$	$\left(\frac{r}{n} + \frac{rx}{n}, \frac{r}{n} - 1 + \frac{rx}{n}\right)$	$\left(\frac{rx}{n}, \frac{rx}{n}\right)$

For both players there is a cost of $1 - r/n > 0$ of switching from defection to cooperation. The framework within which this situation is embedded is that of the theory of non-cooperative games. In essence there are no bounds on the choice of individuals (other than the formal restriction that $d_i \in [0,1]$). Moreover, if an individual has a dominant strategy available then there is no alternative to choosing that strategy. As far as I can tell within that framework there is no resolution to the paradox – indeed I would hardly call it a paradox. The real question, however, is whether the framework within which the game is embedded is appropriate for devising thought experiments relevant to the analysis of cooperation. One way out of the paradox is to embed the particular prisoners' dilemma in a larger game with different properties.

For example, one could think of the game as played out in continuous time and could allow the players, while still acting as individuals, to explore the consequences of their choice. As I see it, this frames the game as a bargaining situation. In this frame, if one individual moves in the direction of defection, then so may the other player without any cost to himself. Since punishment is costless, the threat of punishment can readily be used to induce cooperation the defector. In other words, if time is introduced or if the game is repeated (i.e. iterated), then at least in the two-person case, defection can be punished or policed. However, there is a difficulty with policing that results from an asymmetry. Using the notation of Table 1, the sequence when Column defects and Row punishes is:

$$(3,3) \rightarrow (0,5) \rightarrow (1,1)$$

Recognizing this, Column might say to Row "Look I don't really want to defect, and after all my defection is a cost to you. Why don't you give me a little something to keep me happy". While this is not likely to work in the two-person case, it may in the n-person case.

This will be clear if we change the frame to be that of the theory of cooperative games. In this frame coalitions of members of the group are allowed to form and to coordinate strategies. Let M be such a coalition of size m. The problem for M is to

choose $(d_i : i \in M)$ so as to maximize, in some sense, the vector $(a_i(\mathbf{d}_M, \mathbf{d}_{N-M}) : i \in M)$ where again $a_i(\mathbf{d}_M, \mathbf{d}_{N-M}) = a_i(\mathbf{d})$. For purposes of exposition suppose M chooses to maximize

$$\sum_{i \in M} a_i(\mathbf{d}_M, \mathbf{d}_{N-M}) = S_M(\mathbf{d}_M, \mathbf{d}_{N-M}) = S_M .$$

Then evidently

$$S_M = \sum_{i \in M} [d_i(r/n - 1) + r/n \sum_{j \neq i} d_j]$$
$$= \left[\frac{mr - n}{n}\right] \sum_{M} d_i + mr/n \sum_{N-M} d_j .$$

Suppose that $m > n/r$. Then clearly $(mr - n)/n > 0$, and so the "best" strategy for M is to choose $d_i = 1$ for all $i \in M$. In other words, in the frame of cooperative game theory, and in the context of this example, if M is large enough then it is sensible for it to choose cooperation for its members [16]. Note, of course, since it is assumed that $r < n$, that a coalition of one member cannot be "cooperative". More generally, whatever the "maximization rule" used by coalitions, since N is assumed to be cooperative (in the sense that $\mathbf{a}(\mathbf{1})$ is best) then there do exist cooperative coalitions. In a sense Hardin's [6] resolution of the paradox through voting made use of this observation by focussing on majority coalitions of size at least $n/2$.

So far, so good, but even in the cooperative framework there is no permitted procedure by which coalitions can forbid exit. In other words a member of a cooperative coalition may always choose to be a "renegade" and defect. The difficulty is that within the rules of this framework, the only way the members of the coalition can punish the renegade is to defect as well. In the case that $m + 1 > r/n$, then in the example punishment by defection is a costly act, since the rest of the coalition would prefer to cooperate.

To illustrate the points just made, suppose the coalition to be a town in the wild west, and the defector to be a bank robber. A posse could be formed to hunt down the defector but this could also be very costly (in terms of lives, time, etc.) for the members. But they will calculate that if they don't punish the defector, more defectors will occur. Actually, their best strategy might therefore be to commit themselves to hunting the renegade for some fixed period of time.

The underlying theoretical point that is worth drawing out from this story is that it is extremely difficult to use the natural cooperative game theoretic equilibrium notion, namely the "core", because it is not obvious what it is that coalitions, whether cooperative ones or defectors, may "guarantee" for themselves. Thus, the potential bankrobber who is deciding whether or not to renegade must focus on the decision calculus of the remaining coalition. That, in turn, depends on the costs and benefits as projected by the coalition members. It seems to me that this "game" need not have an equilibrium. More formally, I should say that the appropriate equilibrium notion would have to encode information at a deeper level than is probably possible at the present. It is this difficulty which makes me doubt the robustness of Nozick's [13] use of the notion of the core to argue that a family of associations, each providing certain kinds of collective goods for themselves, can be in equilibrium.

One solution to this problem of information is, as we mentioned, for the coalition or association to commit itself to hunting down any renegades. From this solution it might not be a large step to the formation of a state, "the notion of a concentration of force and the attempt by those in whose hands it is (incompletely) concentrated to determine who else shall be permitted to employ force and on what occasions", (Taylor [19]).

In an earlier and fascinating book, Taylor [18] had inquired whether it is indeed the case that the only resolution to the most important collective good problem, namely the formation and maintenance of social order, depends on the existence of the state. Taylor analyzed the iterated n-person prisoners and argued that cooperative strategies could arise. Briefly, an individual strategy for i is a choice d_{it} at each time t which is conditional on the vector $(\mathbf{d}_{t-1})$ of choices at the previous time (or more generally on the set of all past choice vectors). From the initial point of view, the discounted value of the payoff $a_{it}(\mathbf{d}_t)$ is $w^{t-1}a_{it}(\mathbf{d}t)$, so the total value is

$$\sum_{i=1}^{\infty} w^{t-1}a_{it}(\mathbf{d}_t) \quad \text{where} \quad 0<w<1.$$

To defect or to cooperate for all time are non-conditional. In the two person case the simplest conditional strategy is Tit for Tat – "next time I shall do what you do now, but first of all I shall cooperate". More severe punishment strategies involve defecting an increasing number of times in response to defection.

The point of the iterated n-person prisoners' dilemma (inPD) is that in general there is no dominant strategy as in the single move nPD. As a consequence strategies other than constant defection may be in equilibrium. As Taylor showed, even when there are defectors, conditional cooperative strategies may be in equilibrium. Roughly speaking, in a conditional cooperative strategy one cooperates at each time, as long as a certain number of other players also cooperate. If there are enough conditional cooperators then no one of them will have an incentive to switch to another strategy.

Taylor's analysis clearly indicated that, formally speaking, it was possible for cooperation to emerge within small groups when the collective action problem could in general terms be characterized as an inPD.

3. The Emergence of Cooperation

In 1980, Robert Axelrod [1] published the results of a tournament between fifteen programs (or conditional strategies) for playing the iterated two person PD. In order to win, a program had to do well against all the others. For example, if Tit for Tat plays C (cooperate always) or plays against another Tit for Tat then (assuming the matrix is given by Table 1) they both receive a stream of the joint cooperation payoff (R). If Tit for Tat plays with D (defect always) then the payoff stream is:

Tit for Tat : $SPP\ldots$

D : $QPP\ldots$

Surprisingly, the strategy Tit for Tat (submitted by Anatol Rapoport) won the first tournament, and then later a second tournament against sixty-two other programs.

Shortly afterwards Axelrod [2] published a theoretical analysis of the iterated 2PD, with further discussion in a later book [3]. The interesting point of the analysis was that it was focused on what is known as an evolutionary stable strategy. Suppose N is some collectivity and each member, i, of N adopts a strategy $\mathbf{d}_i$ for the iterated 2PD. Then if a further individual j invades N and adopts a strategy $\mathbf{d}_j$ say, then is it possible for the expected value $E(a_j(\mathbf{d}_j, \mathbf{d}_i))$ to exceed $E(a_i(\mathbf{d}_i, \mathbf{d}_i))$? If not then say $\mathbf{d}_i$ is stable under attack from $\mathbf{d}_j$. If $\mathbf{d}_i$ is stable under attack from every possible $\mathbf{d}_j$, then $\mathbf{d}_i$ is evolutionary stable. For example, consider the case where D (defect always) attempts to invade T (Tit for Tat). As in Taylor's analysis, the future at time t is discounted by w^{t-1}. Then

$$a_D(D, T) = Q + wP + w^2P \ldots$$
$$= Q + \frac{wP}{1-w}.$$

On the other hand when Tit for Tat plays Tit for Tat then

$$a_T(T, T) = R + wR + \ldots$$
$$= \frac{R}{1-w}.$$

As Axelrod shows, if $w \geqq \frac{Q-R}{Q-P}$ then $a_T(T, T) \geqq a_D(D, T)$ so T is stable under attack from D. Note that $a_T(T, T)$ is used as the expectation, since N is taken to be sufficiently large so that average T players effectively never meet D. Let DC be the strategy that alternates defection and cooperation. Evidently

$$a_{DC}(DC, T) = (Q + wS) + w^2(Q + wS) \ldots$$

It follows that if $w \geqq \frac{Q-R}{R-S}$ then T is stable under attack from DC. Moreover, if T is stable from both D and DC then it is stable under attack from any strategy and is thus evolutionary stable. However, D is stable under attack from T since

$$a_D(D, D) = P + wP \ldots$$
$$> a_T(T, D) = S + wP \ldots$$

Although a society of "nice" T-players (playing Tit for Tat) can be maintained, single nice players cannot get cooperation started.

Axelrod supposes that a group of T-players now invades the non-cooperative society. Let p be the probability that two T-players meet and let $1-p$ be the probability that T meets D. Then

$$E(a_T(T)) = pa_T(T, T) + (1-p)a_T(T, D).$$

Supposing that the probability that D meets T is much smaller than the probability that D meets D, gives

$$E(a_D(D)) = a_D(D, D).$$

Then the group of T-players can invade whenever

$$E(a_T(T)) \geqq E(a_D(D))$$

or

$$p(a_T(T,T) - a_T(T,D)) \geqq a_D(D,D) - a_T(T,D)$$

Just to relate this to the earlier analysis of Hardin's version of the PD, let $Q=r/n$, $R=2r/n-1$, $P=0$, $S=r/n-1$. In the single shot PD, p must satisfy

$$p \geqq p^* = \frac{1-r/n}{r/n} = \frac{n-r}{r}$$

In the fully iterated 2PD with discount rate w, the required probability is

$$p_w^* = (1-\varepsilon)p^*$$

where

$$\varepsilon = \frac{w - wp^*}{1 - wp^*} \text{ and so } 0 < \varepsilon < 1.$$

If w is close to 1 then so is ε and thus p_w^* is close to 0. In other words when the future matters, very small coalitions of T-players may invade a population of non-cooperators.

Having discussed Axelrod's model I shall raise some problems with interpretation and relevance.

(i) The model appears remote from real world situations that one would like to analyse. Suppose a society of D-players is invaded by one or a number of T-players. In computing $a_T(T, D)$ it is assumed that the T- and D-players lock onto one another and play out the iterated PD for an infinite duration. It is not assumed in the model that the T-players range throughout the D-society engaging in 2PD against each D-player a random number of times. This peculiar aspect of the model is unlikely to correspond in any way at all with the kinds of situations one would like to examine. If engagement occurs a random number of times then presumably the required probability would lie between p^* and $(1-\varepsilon)p^*$, so some of the force of Axelrod's model is lost.

Conversely, if a group of T-players is invaded by D-players it is not evident that they can be stable. Clearly, if the engagement period is just one unit then $a_D(D, T)$ is a string of Q's, which beats a string of R's resulting from $T-T$ engagements. Effectively this means that for T to be evolutionary stable, the discount parameter, w, has to be even higher than the value deduced by Axelrod. This difficulty can be avoided, however, if it is assumed that each group has a "collective memory". Alternatively, one could assume that members of each group could be unambiguously identified. For example, in the invasion of the D-society by T's, once one T had found out that D's actually played D, then all T's would know this as well as knowing who were D's. However, this escape requires, it seems to me, that this datum be "common knowledge" within the group of T's.

This problem is partially addressed in the chapter jointly written with Hamilton, where in fact the discount parameter, w, is referred to as the probability "of the same two individuals meeting again" ([3], Chap. 5, p. 100).

In this chapter the authors note also that "mutualism" (which I take to be close cooperation, other than symbiosis, between different species) tends to occur in restricted milieux and not in the 'free mixing circumstances of the open sea". This line of though naturally leads to a discussion of the possibility of discrimination or perception on the part of one player that the other uses a different strategy.

To some extent the possibility of common knowledge or discrimination weakens my point about infinitely long periods of interaction. It is still true, however, that for discrimination to be practiced some interaction is necessary.

(ii) The second point is related to the fact that the practice of discrimination depends not only on there being some period of interaction but also on the existence of some asymmetry in the players. To illustrate what I mean consider the coordination game known as "hawk-dove".

Table 3.

		Column	
		$d_2=1$	$d_2=0$
Row	$d_1=1$	$\left(\frac{V}{2}, \frac{V}{2}\right)$	$(0, V)$
	$d_1=0$	$(V, 0)$	$(\frac{1}{2}(V-W), \frac{1}{2}(V-W))$
		$W>V>0$	

The story is that two animals meet over some spoil, of value V. Both may display ($d_i=1$) and have equal probability of winning. If one escalates the conflict ($d_i=0$) it gains the spoil. If both escalate they both have the same probability of being wounded (with cost W). Clearly the equilibria are $(d_1, d_2)=(0,1)$ or $(1,0)$. The stable strategy is a mixed strategy with probability of escalating set at V/W [17].

In this symmetric game the solution is hardly satisfactory, and in general, some asymmetry exists. For example, with some territorial animals, if the conflict occurs in the territory of one then the other will perhaps display and then retire [10, 11]. This "convention" is often reinforced by certain aspects of behavior (leaving scent in one's territory). Axelrod and Hamilton actually give as an example a male territorial bird who reacts aggressively to the song of an unfamiliar male in his territory. In this case residence defines the asymmetry.

The models developed by the evolutionary biologists have been directed at the generation of complex social orders which have the effect of regulating conflict. The key to these is "common knowledge" and this is attained through the use of cues based generally on asymmetries. Axelrod's model unfortunately leaves unclear the manner by which common knowledge and discrimination can occur. I shall come back to this point later.

(iii) Thirdly, Axelrod is ambiguous as to whether we are to view the model as being appropriate to a single generation of players, or whether the game is being played by genes (in the manner outlined by Dawkins [5]). For example, in Chap. 5, Axelrod and Hamilton refer to genetic kinship theory to give an account of the formation of

the cooperative group. Now it seems to me that in general a group of genetically linked cooperators would tend to react very aggressively indeed towards those unrelated to themselves. This would invalidate the entire 2PD model under examination. There also seems to me to be a difficulty in extending the genetic link model of altruism to encompass conditional cooperation.

4. Cooperation and Convention

Axelrod's intention is to provide a plausible account of the emergence of cooperation, while the books by Hardin, Margolis and Taylor, in very different ways, concentrate on the situations within which cooperation can be maintained.

Hardin focuses on the nPD and gives a very readable account of Olson's [14] analysis and the debate over the difficulty of collective action in large groups. The main point of this part of his discussion has been discussed earlier – that in the frame of cooperative game theory if a coalition is above a certain minimal size m^s, (which in the Hardin game is n/r) then that coalition will be cooperative if it behaves rationally. The difficulty of collective action depends not just on the size of the group but also on the ratio (r) of costs to benefits.

Hardin also makes a number of valuable comments on improving the possibility of collective action in a political milieu. For example a group, such as the Sierra Club, might by itself be able to provide very little in the way of collective goods for its members, but in acting as a political interest group it might be able to extract extensive benefits. As Hardin notes, pollution abatement in the U.S. (which may be regarded as a public good from the point of view of the Sierra Club), costs about $ 23 billion. Accordingly "each dollar's worth of politics may buy over $ 2000".

This analysis of politics suggests an extension of the standard nPD to a model where there are a number of potential groups in society say $N_1 \ldots, N_k$ each one of which stands to gain from the provision of a particular kind of collective good $g_1, \ldots, g_k$. This model is not a pure nPD since it is quite reasonable to suppose that provision of good g_i for N_i hurts the other groups. The existence of an equilibrium provision of these goods is doubtful because of the problems of the existence of the core to which I alluded before. Some other points are worth making about this generalized form of the nPD. It is very reasonable to suppose that in a political milieu if N_i is below a certain size then the probability of good g_i being provided is very low, where the critical size would depend on political calculations. As Sam Beer has pointed out once a government goes seriously into the business of providing local or group collective goods, then the degree of conflict between $N_1, \ldots, N_k$ may well increase significantly. Beer's analysis is conceptually very close to Olson's [15] recent discussion of the way in which interest group politics of this kind might lead to outcomes which are disastrous for the society. Olson's suggestion is that in a political context the critical size m_i^s for a politically significant group N_i might be relatively low. Thus a number of such groups may be able to resolve their own internal PD or collective action problems. The groups then become players in a higher level PD within which success by group N_i in providing its own g_i can be seen as a non-cooperative act vis à vis the other groups. The natural question to ask is whether enough of these groups could get together to form a higher level coalition

able to maintain cooperation between themselves. Olson's conclusion about this possibility is pessimistic. It seems to me, however, that the behavior of this higher order PD will depend on certain common knowledge properties of the society. I shall come back to this point below.

For me, however, the most interesting section of Hardin's book is in the last five chapters on convention. Consider the symmetric coordination game:

Table 4.

		Column	
		$d_2=1$	$d_2=0$
Row	$d_1=1$	(W, W)	(V, V)
	$d_1=0$	(V, V)	(W, W)

$W>V>0$

The equilibria are, of course, $(d_1, d_2)=(0,0)$ or $(1,1)$. Since it is immaterial whether 0 or 1 is chosen, as long as $d_1=d_2$, it is relatively easy to imagine a convention arising wherein both players would choose 1. The information requirement underlying this convention is not prohibitively costly since just a few iterations of the game would make it clear to the players how to cooperate. As the discussion of territoriality and so on indicates, conventions (which lead to some form of "social order") may arise particularly when there are asymmetries through which the participants may explore each other's motivations and capabilities. Hardin argues that conventions may also come into being in the iterated nPD when the participants adopt conditional strategies of the kind considered by Taylor [18]. However, the cooperative convention in the nPD, unlike the coordination game, requires policing or enforcement (by threats). Although Hardin suggests the policing problem is generally easier to solve than the collective action problem, the examples I gave earlier suggest that policing does have certain informational or knowledge requirements. Hardin acknowledges this but argues that in small groups the knowledge problem is solvable. It is clear however that the possibility of solution by convention becomes remote when the group is large. Suppose we say that A can trust B if A *knows* there will be an opportunity to punish B whenever it turns out that A's *expectation* of B's cooperation proves misplaced. The knowledge problem concerns those conditions for a society $N=\{..i,j..\}$ under which, for all i, j, i trusts j, j knows i trusts j etc. I shall call this the *common knowledge basis of cooperation*. In his discussion of contract by convention in social theory, Hardin observes that since both economic and social exchange require information about other parties' past behavior, both are backward looking. But it seems to me that in most collective action problems, other than those for very small groups, it is impossible for each player to have access to direct knowledge about the past behavior of every other player, or even about a group of players capable of forming a conditionally cooperative coalition. This does not mean that there cannot be a common knowledge basis of cooperation. As Hardin observes, sporadic dyadic interactions might reinforce agents' knowledge and provide the basis of cooperation.

5. Community, Anarchy and Altruism

Taylor [19] in his recent book further explores those conditions under which social order can be maintained in anarchy – in the absence of a state. He accepts first of all that promises or *offers* of rewards as well as threats of penalties (what Taylor cails "throffers") may be made in anarchy. As I understand his argument, it is that only within a community are offers (and throffers) intelligible. The key features of community are: *(i)* shared common beliefs or norms, *(ii)* direct and complex relations between members, *(iii)* reciprocity. Communication is not a key aspect since people may communicate while sharing hardly any norms. On the other hand, of course, shared norms may be reinforced through communication.

Social order as defined by Taylor is characterized both by predictability of social life and by general conformity to social norms. For agents who prefer to live in a predictable and morally intelligible social universe, social order is a generalized public good. Taylor takes Hardin's observation about the importance of trust further (although he does not use the term). Besides sanctions of the threat of retaliation, agents may sanction by disapproval (which has force because of shared norms) or through supernatural agents (important because of shared beliefs). Moreover, when reciprocity or sharing is habitual then the threat of revoking promises may be made.

Taylor's argument is that community is necessary for anarchic social order. To interpret Taylor (I hope not incorrectly), it is only within community that the common knowledge basis of social order, in the absence of a state, can be satisfied. Taylor draws the further conclusion that since small size is a necessary feature of community, population growth would tend to fission or fragment communities. Under some conditions, such as limited available land or other resources, fission becomes impossible, so that anarchic social order collapses.

It is evident that the concentration of force in the state to some extent replaces the necessity of trust, on the part of individuals, with the exercise of sanctions by the delegates of the state. Since this weakens the common knowledge requirement, social order may be more readily attained. On the other hand, in the presence of the state the importance of reciprocity and shared beliefs is reduced. In a sense, as Taylor argues, the state destroys community. Since a degree of altruism might be expected in community, it may equally well be said that the state destroys altruism.

In this context it is interesting to note that Margolis [8] bases his model of individual behavior on the existence of altruism. More specifically Margolis argues that we may posit for each individual a group oriented preference or G-utility, and a selfish preference, or S-utility. (See also [4] for a discussion of G-utility or "extended sympathy".) Optimal behavior by the individual involves making the best trade off between G-utility and S-utility. Margolis uses this to good effect in accounting for why people vote when it is irrational to do so on the basis of selfish calculations. It certainly makes sense for an individual to have broad preferences over social states (G-utility) that may conflict in many ways with his prefernces over those aspects of the world that affect him personally. The former preferences might very well be "altruistic" in some sense or another.

There are two points which arise in connection with this concept of G-utility. Suppose we assume, for purposes of illustration, that the group utility of individual i

in state x may be written in the form

$$u_i^G(x) = \sum \lambda_{ij} u_{ij}(x)$$

Here u_{ij} is i's estimate of the S-utility of state x for another person, namely j.

As Margolis notes, there is no reason for each individual to give equal weight to everyone. We should therefore expect two different individuals (i and i') to give different weights to the same individual ($\lambda_{ij} \neq \lambda_{i'j}$), and to be given different weights by j ($\lambda_{ji} \neq \lambda_{ji'}$). Indeed the weights might well reflect group loyalities held by the individual (so that λ_{ij} is large for all j in a group to which i is loyal).

The second point is that there is no reason to expect i's estimate of j's S-utility (u_{ij}) to coincide with j's S-utility (u_{jj}). Indeed Margolis identifies $u_{ij}(x)$ not so much as the i's estimate of $u_{jj}(x)$ but rather as the level of welfare that i believes j has in state x. This means that for each x there can be wide variation in $\{u_{ij}(x)\}$ as i varies across the society. In turn this implies that the individuals' G-utilities could in theory be quite different, even among individuals who know each other well in a small group. Thus G-utilities, no less than S-utilities, might be incoherent with respect to one another. So even if individuals use G-utilities in determining how to behave in a generalized n-person prisoners' dilemma, the result might very well not be Pareto optimal (with respect to the S-utilities $(u_{11}, \ldots, u_{nn})$). On the other hand the more closely one individual's estimate of another's S-utility matches the true S-utility, and the more equal the λ weights became the more likely that generalized cooperative problems could be resolved and the more likely would the G-utilities be coherent. Margolis discusses the possibility that cognitive clues may be used to identify group interests, and suggests that individuals use general rules of thumb, such as "what would someone in my position usually do" to estimate prevailing group judgments. It is clear however that what underlies the process of arriving at coherent G-utilities is precisely a "common knowledge" phenomenon.

In an initial state where individuals know very little about each other, even the use of G-utilities is unlikely to lead to any resolution of the generalized cooperation problem. With reiteration, it might be reasonable to suppose that each individual learns something of the true preferences and beliefs of others. However even i's use of G-utilities and his *knowledge* that other individuals make use of G-utilities will not lead directly to a resolution of the problem. Indeed in a recent analysis of the nPD in situations where individuals do have partial beliefs that others will behave cooperatively, it is quite evident that behavior can become exceedingly complex [8].

6. Knowledge

Throughout this essay I have argued that the fundamental theoretical problem underlying the question of cooperation is the manner by which individuals attain knowledge of each others preferences and likely behavior. Moreover, the problem is one of common knowledge, since each individual, i, is required not only to have information about others preferences, but also to know that the others have knowledge about i's own preferences and strategies.

In the restricted nPD it might be possible to argue that this problem is partially resolvable, in the sense that certain types of actors might have good reason to believe

that others are of a particular type. In the restricted context of a community, Taylor's argument makes good sense: social norms will be well understood and will provide the basis for common knowledge, and this knowledge will be maintained by mechanisms designed to make acts intelligible. In more general social situations, however, individuals will be less able to make reasonable guesses about other individuals' beliefs. The theoretical problem underlying cooperation can be stated thus: what is the minimal amount that one agent must know in a given milieu about the beliefs and wants of other agents, to be able to form coherent notions about their behavior, and for this knowledge to be communicable to the others. It seems to me that this problem is at the heart of any analysis of community, convention and cooperation.

References

1. Axelrod R (1980) Effective choice in the prisoner's dilemma, J Conflict Resolution 24:3–25
2. Axelrod R (1981) The emergence of cooperation among egoists, Am Po Sci Rev 75:306–318
3. Axelrod R (1984) The evolution of cooperation. Basic Books, New York
4. Binmore KG (1984) Game theory and the social contract. Working Paper, London School of Economics and Political Science
5. Dawkins R (1976) The selfish gene. Oxford University Press, Oxford
6. Hardin R (1971) Collective action as an agreeable *n*-person prisoners' dilemma, Behav Sci 16: 472–481
7. Hardin R (1982) Collective action. Johns Hopkins University Press, Baltimore
8. Kreps DM, Milgrom P, Roberts J, Wilson R (1982) Rational cooperation in the finitely repeated prisoner's dilemma. J Econ Theory 27:245–252
9. Margolis H (1982) Selfishness, altruism and rationality: A theory of social choice. Cambridge University Press, Cambridge (Reprinted, Chicago University Press, Chicago, 1984)
10. Maynard Smith J (1982) Evolution and the theory of games. Cambridge University Press, Cambridge
11. Maynard Smith J, Parker GA (1976) The logic of asymmetric contests, Anim Behav 24:159–175
12. Maynard Smith J, Price GR (1973) The logic of animal conflict. Nature 246:15–18
13. Nozick R (1974) Anarchy, state, and utopia. Basic Books, New York
14. Olson M (1965) The logic of collective action. Harvard University Press, Cambridge, Mass
15. Olson M (1982) The rise and decline of nations. Yale University Press, New Haven
16. Schofield N (1975) A game theoretic analysis of Olson's game of collective action. J Conflict Resolution 19:441–461
17. Selten R (1983) Evolutionary stability in extensive two person games, Math Soc Sci 5:241–365
18. Taylor M (1976) Anarchy and cooperation. Wiley, London
19. Taylor M (1982) Community anarchy and liberty. Cambridge University Press, Cambridge