PRISONERS' DILEMMAS AND RECIPROCAL ALTRUISTS

JOHN J. TILLEY

Several authors have shown (or correctly implied) that altruists can face a Prisoner's Dilemma (PD); hence altruism itself is no sure remedy for the problem posed by that game.[1] This might suggest that *no* form of altruism provides such a remedy — i.e., that for any type of altruism, we can find situations that would be PD's for the sort of altruists in question. In this paper I forestall that conclusion by discussing a brand of altruism (not to be confused with utilitarianism) that prevents PD's from occurring. (I also discuss some slightly disheartening facts about it.) My point of departure is a remark I made in an earlier essay, discussion of which provides a clue to finding the requisite brand of altruism.

I

The remark is this: "Given any particular *type* of preference structure (distinguished according to the *objects* of the preferences), we can find games that would be PD's for players with preference structures of that type."[2] Unfortunately, the parenthetical phrase is ambiguous. There is a way of reading it which makes the quoted remark true,[3] but a second natural reading is this: "distinguished according to the desired objects — that is, according to precisely *what* is desired." This reading makes the quoted remark false. To see this, suppose that Hester and Lester (*H* and *L*) each want nothing but to maximize the money collected between the two of them. To keep things simple, suppose also (not just for now, but from here on) that anytime *H* and *L* interact, money is the only commodity at stake, which means we can represent possible outcomes in terms of monetary awards *per* player. (Figure 5 shows a matrix of this sort.) Suppose also that *H* and *L* are always fully informed about their situation — neither is ignorant about, say, the monetary prizes or the strategies available in the game.

PHILOSOPHIA 27, 1-2: March 1999

Such players cannot face a PD. A PD is an interactive situation which, when represented using utility payoffs, has the ordinal structure of Figure 1.[4] In Figure 1 the players have different preferences over *CD* and *DC*. Cell *CD*, for instance (the cell produced if player *H* chooses *C* and *L* chooses *D*), is *H*'s *least* preferred cell, but *L*'s *most* preferred. Hence an essential feature of any PD is that the players have *different* rankings of the outcomes. If there is something about the players' desires which rules out such rankings, no PD can arise.

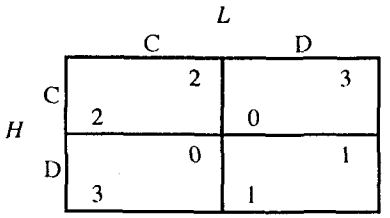

Fig. 1

So it is easy to see that *H* and *L* cannot face PD's. Whenever they interact, each has the exclusive goal of maximizing the money collected between the two of them; hence they rank the outcomes in exactly the same way. Since their rankings cannot diverge, their situation cannot have the structure of Figure 1 if depicted using utility payoffs.

II

Now that we know that the above desires can prevent PD's, we have a clue to finding a form of *altruism* (as defined in note 1) that prevents PD's. The trick is to find a brand of altruism which ensures non-divergent rankings of the relevant outcomes. Our efforts will be frustrated if we think simply in terms of *pure vs. partial* altruists, or in terms of *ordinary* altruists *vs.* those with a mixture of ordinary and *second order* altruism. For each of these classifications, we can find situations that would be PD's for players thus classified.[5]

Our efforts will not be frustrated, however, if we distinguish forms of altruism using not only the categories just mentioned, but some specific assumptions about the players' utility curves and the weights

they assign to their objectives. I will explain this with an example. Suppose that *H* and *L* are partial altruists: each wants to maximize not only his own monetary gain, but his co-player's preference-satisfaction (utility) from money. Suppose the following as well:

(i) The interval utility curves in Figures 2 and 2′ represent the players' "selfish" preferences, meaning their preferences regarding personal monetary gain.[6] (The symbol "$u_H(\$_H)$" represents *H*'s utility payoffs relative to money to *H*; "$u_L(\$_L)$" has a similar meaning with regard to *L*.)

(ii) *H* and *L* have the altruistic preferences represented, respectively, by Figures 3 and 3′.

(iii) Each player gives an *equal* weight to the two commodities involved. More specifically, each player's preferences are such that we can assign a weight of .5 to each of the things she is concerned with: money in her own pocket; and her co-player's utility from money in *his* pocket.[7] (For instance, in *H*'s case we can assign a weight of .5 both to $\$_H$ and to $u_L(\$_L)$.)

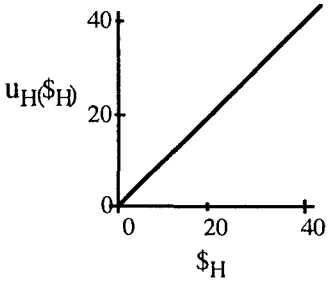

Fig. 2

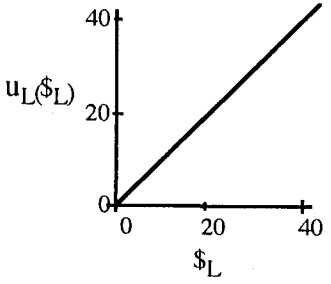

Fig. 2′

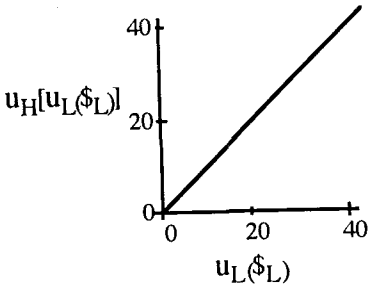

Fig. 3

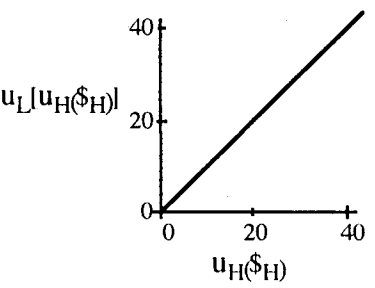

Fig. 3′

Given the above assumptions, *H* and *L* cannot face a PD. Suppose they face the situation in Figure 4. For a given cell, suppose that *H* stands to gain $\$x$ (i.e., $\$_H = \$x$) and *L* stands to gain $\$y$. Player *H*'s utility payoff for that cell is $\{.5xu_H(\$x)\}+\{.5xu_H[u_L(\$y)]\}$, and *L*'s is $\{.5xu_L(\$y)\}+ \{.5xu_L[u_H(\$x)]\}$.[8] Given Figures 2 and 3′, we can see that $u_H(\$x)$ and $u_L[u_H(\$x)]$ have the same value, namely x.[9] Similar remarks go for $u_L(\$y)$ and $u_H[u_L(\$y)]$, only this time the shared value is y, and the figures which show this are 2′ and 3. So *H*'s payoff, being $.5(x + y)$, is the same as *L*'s. The upshot is that *H* and *L* receive the same payoff *per* cell, which is to say that they rank the cells in the same order. Their situation cannot be a PD.

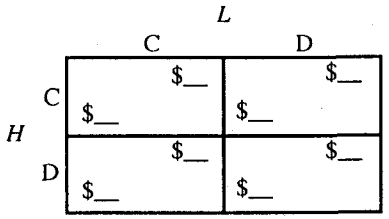

Fig. 4

We have succeeded, therefore, in finding a pair of altruists who, when interacting solely with each other, cannot encounter PD's. As some would put it, we have found a form of altruism that "remedies"

the PD. Three things are worth noting about this altruism, one of them heartening; the others disheartening. The first is that there may be plausible *moral* arguments — arguments, moreover, that are independent of any considerations about the PD — for the claim that we ought to cultivate the altruism in question. Players *H* and *L* seem to display the impartiality prescribed by many ethical theories.[10]

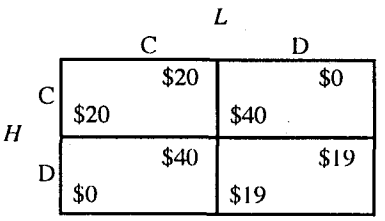

Fig. 5

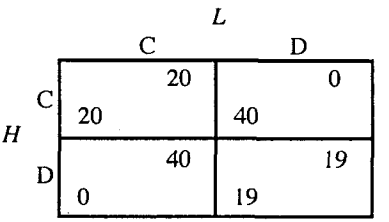

Fig. 6

Now for the disheartening facts. The first is that the "remedy" furnished by the above form of altruism is very tenuous in this sense: with only a slight modification of the players' preferences we can change it to a brand of altruism that does *not* rule out PD's. Suppose we make a minor change in assumption (iii). Each player, let's assume, assigns a weight of .45 to his selfish goal and .55 to his altruistic one. Now suppose that *H* and *L* are facing the situation in Figure 5. From Figures 5, 2 and 2′ we can construct figure 6, which shows each player's payoff relative to monetary gain (i.e., it shows the relevant values of $u_H(\$_H)$ and $u_L(\$_L)$). Using that figure along with Figures 3 and 3′, we can determine the players' ultimate payoffs as follows:[11]

Cell	*H's utility payoffs* $\{.45xu_H(\$_H)\} + \{.55xu_H[u_L(\$_L)]\}$:	*L's utility payoffs* $\{.45xu_L(\$_L)\}+ \{.55xu_L[u_H(\$_H)]\}$:
CC	.45x20+.55x20=20	.45x20+.55x20=20
CD	.45x40+.55x0=18	.45x0+.55x40=22
DC	.45x0+.55x40=22	.45x40+.55x0=18
DD	.45x19+.55x19=19	.45x19+.55x19=19

Given our assumptions, *H* and *L* are facing the game in Figure 7, which has the ordinal structure of Figure 1. In short, they are facing a PD.

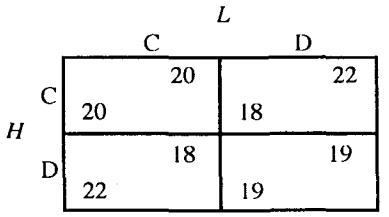

Fig. 7

The second disheartening fact is this: Although the altruism we have identified — the one that involves the unrevised version of assumption (iii) — prevents PD's, this is not to say that it always ensures an optimal outcome.[12] Consider Figure 8. If each player cares only about increasing his own wealth, the players are sure to achieve an optimal cell, namely *CC*. But if the players are altruists of the kind we have been discussing — that is, if assumptions (i)-(iii) are true — they face the game in Figure 9. This game has two equilibria, *CC* and *DD*, where neither *C* nor *D* is dominant for either player. It's arguable that this game is unsolvable, on the following grounds. If the players are rational in the game-theoretical sense, neither player can choose a strategy that she knows to be utility-maximizing relative to the strategy of her co-player.[13] To do so she must form a determinant expectation about the strategy her co-player will choose, but the structure of the game prevents this. So the players are not ensured an optimal outcome.

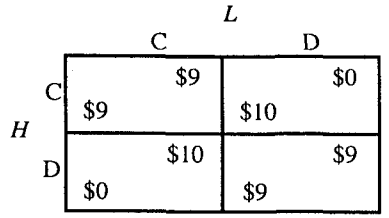

Fig. 8

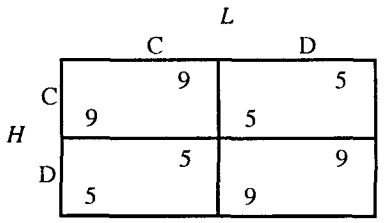

Fig. 9

Let's sum up by distinguishing two questions. The first receives a fair amount of attention (see the authors in note 1); the second does not.

(A) Is it true that PD's arise specifically for selfish, or at least non-altruistic, agents, meaning that if everyone were altruistic, no PD's could arise?

(B) Can we define a very specific form of reciprocal altruism that rules out PD's?

The answer to (A) is no; the answer to (B) is yes. (The answer to (A) would be no even if we revised it as follows: Is it true that PD's arise only for selfish and minimally altruistic agents, meaning that if everyone were intensely altruistic, no PD's could arise?) The answer to (B) does not undermine the answer to (A), but it does reveal an important fact about altruism, and about PD's, that usually goes unmentioned when (A) is addressed.

Two further points are worth making. First, despite the fact that (A) demands a negative answer, the opposite answer remains tempting to many minds. For instance, a recent author, having shown that reciprocal altruism can render unproblematic many situations that would be PD's for egoists, concludes that "this approach, naively simple and straightforward as it may appear to be, still serves the purpose of solving the prisoner's dilemma..."[14] By "solving" the PD he seems to mean *preventing* PD's, and he reaches his mistaken conclusion — his affirmative answer to (A) — by overlooking situations that would be PD's for altruists.

The next point is that our "yes" answer to (B) does not amount to the

claim that *utilitarians* can avoid PD's. To answer (B) in the affirmative we need not assume that players *H* and *L* accept utilitarianism or a similar consequentialist theory, nor that their preferences and choices always dovetail with the dictates of such a theory. Perhaps *H* cares for no one but himself and *L*, meaning that he becomes a pure egoist in games in which *L* is not his co-player. A "yes" answer to (B) is compatible with the assumption that *H* and *L* have sympathies and moral concerns that are highly limited, and that only toward each other do they display the kind of impartiality morality demands.

III

To put the results of section 2 in perspective, I will finish with a conjecture about why question (B) receives less attention than (A). It does so, I think, because most of those who suspect that altruism prevents PD's are not concerned with (B) — they would find it uninteresting regardless of the answer it received. They are looking for a "yes" answer to (A); anything short of that they will find unsatisfying. (So, understandably, their critics mainly address (A), not (B)). This is because their assumption that altruism prevents PD's is part of a wider assumption: that given the desires people typically have, few PD's really exist. This assumption is frequently voiced when the PD is used (by someone who does not share the assumption) to analyze an economic or political problem. On such occasions we commonly hear this complaint: "But what can we really conclude about the world from this abstract model! After all, people are not the purely selfish creatures your two 'players' seem to be — they have at least some concern for others!"[15] The idea here is that given the desires people actually have, many of which are unselfish, few human interactions have the structure of a PD; hence the PD sheds little light on actual problems.

We can easily see why the person who makes this (mistaken) complaint takes little interest in question (B). The answer to (B) provides no grounds for saying that few PD's arise in the actual world. Actual people are altruistic to some extent, but this is a far cry from saying that most people, most of the time, exhibit the highly *specific* brand of altruism which rules out PD's. So our answer to (B) does nothing to support the complaint in the preceding paragraph. What

would support it is an affirmative answer to (A), but such an answer would be mistaken.

We must conclude, then, that although we have revealed an interesting fact about PD's, namely that a form of altruism prevents them from arising, we have not produced the comforting result some people would like. In the actual world we can expect to find plenty of PD's.[16]

INDIANA UNIVERSITY/PURDUE UNIVERSITY INDIANAPOLIS
425 UNIVERSITY BLVD.
INDIANAPOLIS, IN 46202-5140
USA

NOTES

1 See, e.g., John Tilley, "Altruism and the Prisoner's Dilemma," *The Australasian Journal of Philosophy* 69 (1991): 264-287; Richmond Campbell, "Background for the Uninitiated," in R. Campbell and L. Sowden, eds., *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem* (Vancouver: University of British Columbia Press, 1985), p. 10; F. C. T. Moore, "The Martyr's Dilemma," *Analysis* 45 (1985): 29-33; Philip Pettit, "The Prisoner's Dilemma and Social Theory: An Overview of Some Issues," *Politics* 20 (1985): 2f; Sheldon Wein, "Prisoners' Dilemmas, Tuism, and Rationality," *Simulation & Games* 16 (1985): 23-31; Robert Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984), pp. 6f, 12; Neil Cooper, *The Diversity of Moral Thinking* (Oxford: Clarendon Press, 1981), pp. 274f; Nicholas Rescher, *Unselfishness* (Pittsburgh: University of Pittsburgh Press, 1975), p. 43; J. Howard Sobel, "The Need for Coercion," in J. R. Pennock and J. W. Chapman, eds., *Nomos XIV: Coercion* (Chicago: Aldine and Atherton, 1972), pp. 148-77; and Thomas Schelling, "Some Thoughts on the Relevance of Game Theory to the Analysis of Ethical Systems," in I. R. Buchler and H. G. Nutini, eds., *Game Theory in the Behavioral Sciences* (Pittsburgh: University of Pittsburgh Press, 1969), p. 50. The footnoted sentence contains two key terms. The first I will clarify using a standard definition; the second by quoting from Tilley, "Altruism and the Prisoner's Dilemma," p. 264 n. 1:

Prisoner's Dilemma: A game with this feature: each player has a dominant strategy (a strategy which, relative to anything the other players might do, is at least as good as any other strategy; and which, relative to at least some of the things the other players might do, is better than any

other strategy), but those strategies lead to an outcome that's deficient in this sense: there's another outcome to which it is dispreferred by every player. (Figure 1 shows a PD. The following assumptions apply to that figure and to the other games I discuss: first, the game is non-repeated; second, the players are confined to pure strategies; third, the players choose simultaneously; fourth, communication and binding agreements are impossible; and fifth, each player's choice is independent of the choice of his co-player.)

Altruism: "Imagine a 2x2 game in which the outcomes are defined in terms of a single commodity... and suppose that each player prefers more of the commodity to less... By an *altruist*, I mean a player who wants the *other* player's desires for the commodity to be satisfied. Less roughly, if a player *P* is an altruist, then: (i) the other player's utility payoffs *u* relative to the given commodity will serve as an item over which a utility function for *P* can be established; (ii) *P* will prefer more of that item to less; and (iii) *P*'s utility function relative to *u* will determine, in part anyway, her overall payoff for each of the possible outcomes. It will completely determine her payoffs if she is a *pure* altruist..., but only *partly* determine them if she is a *partial* altruist. In the latter case, a second determining factor will be her preferences regarding her own reception of the initial commodity." (Some partial altruists are discussed in section 2; their altruistic preferences are represented by the utility curves in Figures 3 and 3′.)

2 Tilley, "Altruism and the Prisoner's Dilemma," p. 282.

3 More fully, the quoted remark is true if we read the parenthetical phrase to mean "distinguished according to the kinds of objects over which the preferences range." (Some counterexamples may come to mind, but they can be handled by imagining players with unusual rankings of the outcomes.) To see the difference between this reading and the one following the footnoted sentence, assume that players *H* and *L* want to maximize the money collected between the two of them. The object of *H*'s desire is just that — *to maximize the money collected by the team made up of H and L.* But the objects over which his preferences range are simply *monetary sums collected by that team.* To say that *H* wants the highest of those sums — that he wants to *maximize* the money collected by the team — is to say something, not about the *type* of objects over which he has preferences, but about the way his preferences range over those objects, i.e., about how he *ranks* the monetary sums. In short, to talk of the object of a player's desire is one thing; to talk of the objects over which he has preferences is another. In doing the latter we imply nothing about how the player ranks those objects. A point illustrated by this is

that although there usually is no harm in using "desire" and "preference" interchangeably, there are differences between preferences and desires.

4 The "ordinal structure" of an interactive situation is found by depicting it using ordinal utility payoffs rather than concrete prizes. Such payoffs merely rank order the outcomes. (The higher the payoff, the higher the player ranks the outcome.) Ordinal payoffs contrast with *interval* payoffs, which not only rank the outcomes, but reflect the relative preference intervals between them. Such payoffs come into play in section 2.

5 As shown in Tilley, "Altruism and the Prisoner's Dilemma." *A second order* altruistic desire — of player *H*'s, let's say — is a desire for the satisfaction, not of *L*'s *egoistic* desires (in the present case *L*'s desire for money), but of *L*'s *altruistic* desires, which have as their object the satisfaction of *H*'s egoistic desires.

6 The assumption that the utility curves be exactly those shown in Figures 2 and 2′ is a feature of the example I have chosen; it is not required for my argument that a certain form of altruism can prevent PD's. What's required is that *if* the players' selfish preferences are those represented in Figures 2 and 2′, their altruistic preferences are those represented in 3 and 3'. This ensures that $u_H[u_L(\$_L)] = u_L(\$_L)$ and $u_L[u_H(\$_H)] = u_H(\$_H)$.

7 In order to make use of these (or other) assigned weights, we must stipulate that if a given player makes a legitimate change to either of the two diagrams which represent her preferences (meaning a change that does not alter the information conveyed by the diagram), she makes the same change on the other diagram, and her co-player, knowing that she has done so, makes the same change on his two diagrams. (E.g., if *H* makes a positive linear transformation of the units on the vertical axis of Figure 2, he does the same to Figure 3, and player *L* does the same to Figures 2′ and 3′.) To fail to honor this stipulation is to violate the spirit, if not the letter, of the footnoted sentence — i.e., in many interactive situations it can have the same effect as changing the weights assigned to the two objectives. It's reasonable to assume that the two players, each wanting to give a determinate meaning to the assigned weights, will agree to this stipulation.

8 If this is puzzling, note that the weight assigned to a given commodity is multiplied, not by any value or quantity of that commodity, but by the *utility payoff* corresponding to the player's reception of that commodity. For instance, the weight assigned to $\$_H$ is always multiplied by some value of $u_H(\$_H)$. For more on the formula employed here, see note 11.

9 Note 7 is relevant here.

10 See Cooper, *The Diversity of Moral Thinking*, p. 274f. (In note 29 of my "Altruism and the Prisoner's Dilemma" I criticized Cooper's assertion

that "sober" altruism prevents PD's. My criticism was out of place. If read sympathetically and in context, Cooper's assertion is about the altruism we have been discussing in this section.)

11 For example, to determine *H*'s utility payoff for the CD cell, we use Figure 6 to find $u_H(\$_H)$ (= 40). We use it again to find $u_L(\$_L)$ (= 0), and then use Figure 3 to find the corresponding value for $u_H[u_L(\$_L)]$ (= 0). Having found the values for $u_H(\$_H)$ and $u_H[u_L(\$_L)]$, we multiply the first of those values by .45; the second by .55. We then add the two products. (By proceeding in this way we are treating *H*'s situation as a *multiattribute decision problem* — one "attribute" being money to himself ($\$_H$), the other being *L*'s preference-satisfaction from money to *L* ($u_L(\$_L)$). For a thorough treatment of this subject see Ralph Keeney and Howard Raiffa, *Decisions with Multiple Objectives* (New York: Wiley, 1976).

12 I'm indebted here to Michael J. Almeida's "Too Much (and not enough) of a Good Thing: How Agent Neutral Principles Fail in Prisoner's Dilemmas" (unpublished manuscript, February 1997). The term "optimal," by the way, is short for "Pareto-optimal." An outcome *M* is Pareto-optimal just in case there is no other outcome in the game that would grant at least one player a higher payoff, and no player a lower payoff, than *M* grants. Another term in the paragraph is "equilibria." An outcome is an equilibrium just in case each of its component strategies is utility-maximizing relative to the others.

13 For more on this subject see John Tilley, "Accounting for the 'Tragedy' in the Prisoner's Dilemma," *Synthese* 99(2) (1994), notes 17 and 26, and the accompanying text.

14 C. L. Sheng, "A Note on the Prisoner's Dilemma," *Theory and Decision* 36 (1994): 239f; see also 234. For a discussion of Sheng's essay see John Tilley, "Prisoner's Dilemma from a Moral Point of View," *Theory and Decision* 41 (1996): 187-93.

15 Similar complaints are often made in print. The following is from Rescher's *Unselfishness*, p. 40: "Its shock effect [that of the PD] for students of political economy inheres solely in their ill-advised approach to *rationality* in terms of a prudential pursuit of selfish advantage.... This view quite unrealistically renders each man a self-centered island..." (This is a strange statement, given that on a later page Rescher demonstrates that *non-selfish* agents can face PD's.)

16 The thoughts in this paper and the intention to write it were prompted by comments from Michael Burke and Cliff Landesman on my "Altruism and the Prisoner's Dilemma." I'm grateful to both of them. A second thanks goes to Cliff Landesman for many helpful remarks on the first draft of this paper.