Too Much (And Not Enough) of a Good Thing: How Agent Neutral Principles Fail in Prisoner's Dilemmas

Author(s): Michael J. Almeida

Source: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, Jun., 1999, Vol. 94, No. 3 (Jun., 1999), pp. 309-328

Published by: Springer

Stable URL: https://www.jstor.org/stable/4320940

MICHAEL J. ALMEIDA*

TOO MUCH (AND NOT ENOUGH) OF A GOOD THING: HOW AGENT NEUTRAL PRINCIPLES FAIL IN PRISONER'S DILEMMAS

(Received in revised form 29 December 1997)

1. INTRODUCTION

The principle of ethical egoism is paradigmatic among agent relative theories. Ethical egoism requires each agent to rank the outcomes of her alternative actions, best to worst, by appeal to her own self-interest, or the maximization of her own utility. In its purest form, it requires that we give no weight to the interests of others, but only to our own interests. Since the criteria for evaluating outcomes varies from agent to agent, the ranking of outcomes, even when restricted to a unique object of evaluation, will typically vary from agent to agent.

Less familiar among agent relative theories is the principle of ethical altruism. The principle of ethical altruism requires each agent to rank the outcomes of her alternative actions, best to worst, by appeal to the interests of others, or to the maximization of the utility of others. In its purest form, it demands that we give no weight to our own interests, but only to the interests of others. Since, again, the criteria for evaluating outcomes varies from agent to agent, the ranking of outcomes, even when restricted to a unique object of evaluation, will vary from agent to agent.

The principle of ethical egoism and the principle of ethical altruism share a notorious problem common to all agent relative principles.[1] In some familiar prisoner's dilemmas each of the principles does poorly. It is precisely in these prisoner's dilemmas that each of the principles displays its collective irrationality as determined by CI.

CI. A principle *P* is collectively irrational when it is certain that if every member of some group *G* were to successfully follow *P*,

Philosophical Studies **94:** 309–328, 1999.

then *each* member would be worse off, in *P*'s terms, than they would be were no member of *G* to successfully follow *P*.[2]

Assuming that CI correctly specifies sufficient conditions on collective irrationality, it appears irrational for any group facing a prisoner's dilemma whose objective is to maximize individual utility at the collective level to select either ethical egoism or ethical altruism as its governing principle.[3]

In the familiar prisoner's dilemma situations (henceforth PD's), the agent neutral pursuit of individual utility maximization is not collectively irrational. Agent neutral principles specify criteria for ranking the outcomes of alternative actions, best to worst, from an *impersonal* point of view.[4] Since the ranking of outcomes is not indexed to particular agents, it will not vary relative to each agent. In addition to specifying criteria neutral with respect to moral agents, agent neutral theories specify criteria which are temporally and geographically neutral. Relative to a unique object of evaluation, say, Smith's studying philosophy in Boston at t1 and Smith's studying science at Austin at t2, temporal neutrality entails that the assessment of the sequence of actions from temporal point t1 will not differ from the assessment of the sequence from point t2.[5] The criteria for evaluating the sequence are neutral over time. And geographic neutrality entails that the evaluation of the sequence in Boston will not differ from the evaluation in Austin, at either temporal point. The criteria are, in short, neutral over location.

Agent neutral principles are never certain to yield a deficient outcome for agent neutral players. But I show in Section 2 that there is a large class of PD's in which agent neutral principles *cannot* yield more individual utility than agent relative principles. If a group G of agent neutral players is acting in the context of at least one other player who is not agent neutral then, in a large class of PD's, G cannot do better than a group of agent relative players. In such contexts, groups of ethical egoists and ethical altruists have an advantage over agent neutral players in the pursuit of individual utility. Even a very large group of ethical egoists can be better off, in terms of individual utility, collectively defecting in contexts of even a few others who are not agent neutral. And it is a virtual certainty that we act in the context of others who are not agent neutral.[6]

If we confine ourselves to the comparatively small class of complete, absolute PD's, we find that a group G of agent neutral players will collectively maximize individual utility, independently of the behavior of other, non-members of G.[7] In Section 3, I consider first complete, absolute PD's (henceforth, CAPD's).[8] I show that, for every CAPD, there are many groups of agent relative players who will, for certain, achieve the cooperative outcome. It follows that, even in the small class of CAPD'S, agent neutrality is *not necessary* for cooperation. I consider next the broader class of total, absolute PD's (henceforth, TAPD's). In TAPD'S, *only* agent relative players are certain to reach the cooperative outcome. In egoistic, TAPD'S, for instance, only a group of ethical altruists will, for certain, reach the cooperative outcome. And in altruistic, TAPD'S, only a group of ethical egoists will, for certain, reach the cooperative outcome. Agent relative principles are collectively irrational according to CI, nonetheless they guarantee cooperation in many PD contexts where agent neutral principles cannot.

I conclude that the familiar, two-person PD's are extremely misleading concerning the success of agent neutral players in reaching cooperation, and the failure of agent relative players in doing so. In some PD's, some agent relative players do worse than agent neutral players, but in some PD's they do considerably better. How well agents facing a PD do in the pursuit of individual utility is determined not by their neutrality, but by the largely contingent matter of the types of PD's they encounter.

2. AGENT NEUTRAL PLAYERS AND NON-ABSOLUTE PRISONER'S DILEMMAS

In egoistic PD's, the self-interested pursuit of individual utility by each member of a group G results, for certain, in a deficient outcome for G.[9] This claim is familiar, and seems to follow directly from the minimal defining conditions of PD's. Consider utility-structure in Figure 1.

Row and Column each have two options: cooperate or defect. The rewards for cooperation to Row and Column are specified in R_R ands R_C respectively. The temptation payoffs for each, T_R and T_C, are what each would receive were he to succeed in unilateral

		C	D
		C	
R	C	R_R, R_C	S_R, T_C
	D	T_R, S_C	P_R, P_C

Figure 1.

defection. The sucker payoffs, S_R and S_C, are what each receives for cooperating with a defector. Finally, the punishment for universal defection is specified in P_R and P_C.[10] Assuming causal independence, the defining conditions on a *basic* PD are as follows.

C1. $T_R > R_R$ and $P_R > S_R$
C2. $T_C > R_C$ and $P_C > S_C$
C3. $R_R > P_R$ and $R_C > P_C$

Conditions C1 and C2 ensure that defection is the strictly dominant pure strategy for each player, and condition C3 ensures that the uncooperative outcome is deficient.[11] All players prefer the cooperative outcome to the uncooperative outcome.

Let the group be G = {Row, Column}. Assume that Row and Column are purely self-interested, or are pure egoists. Each is interested in maximizing his own utility, or preference-satisfaction. Neither receives any utility from the preference-satisfaction of the other. We can express this fact more precisely by saying that each of the players places a weight of 1 on his reception of what he prefers, and a weight of 0 on the preference-satisfaction of the other.[12] Finally, let's assume that concrete prizes, whether they are years in prison or monetary rewards or whatever, are always linear with utility.[13] It is easy to see that the distribution of utilities in Figure 2 meets conditions C1–C3. The first and second numbers in each cell are the

		C	D
		C	
R	C	3, 3	1, 4
	D	4, 1	2, 2

Figure 2.

utilities afforded Row and Column respectively in that outcome. It is apparent that pure egoists using their dominant pure strategy would each defect in Figure 2, and the result would be the deficient outcome (D, D). Pure egoists are agent relative players, and *G* realizes a deficient outcome if all members of *G* are pure egoists.

Consider whether G would do better, in terms of individual utility, were all members of G agent neutral players. Agent neutrality demands that no player place any greater weight on her own preference-satisfaction than on the preference-satisfaction of any other player. In the two-person game, each person is required, then, to place a weight of 0.5 on her own preference-satisfaction and 0.5 on the preference-satisfaction of the other player.[14] Let 'U_R' symbolize Row's utility and 'U_C' symbolize Column's utility. The utility payoffs to Row in Figure 2 are transformed for agent neutral players in the following way.[15]

$$\begin{aligned}
U_R(C,C) &= 0.5[U_R(3)] + 0.5[U_R(U_C(3))] = 3\\
U_R(C,D) &= 0.5[U_R(1)] + 0.5[U_R(U_C(4))] = 2.5\\
U_R(D,C) &= 0.5[U_R(4)] + 0.5[U_R(U_R(1))] = 2.5\\
U_R(D,D) &= 0.5[U_R(2)] + 0.5[U_R(U_C(2))] = 2
\end{aligned}$$

Column's situation is perfectly symmetrical and her utilities mirror Row's utilities in each cell. Figure 2 is transformed for agent neutral players into Figure 3.

The conclusion typically drawn from Figures 2 and 3 is that in PD's a group composed of agent neutral players will, in general, do *better than* a group of agent relative, egoistic players, in the collective maximization of individual utility.[16] Were this claim true, then any group concerned with the maximization of individual utility and facing a PD would do well to have agent neutral players.

		C	
		C	D
R	C	3, 3	2.5, 2.5
	D	2.5, 2.5	2, 2

Figure 3.

But the claim is not true. How well a group G of agent neutral players does in a PD depends on what everyone else outside of the group, E–G, is doing. A group of agent neutral players facing a PD in the presence of purely egoistic non-members will not in general do better than a group of agent relative players. And, even in the presence of altruistic non-members, a group of agent neutral players will not in general do better than agent relative players.[17]

It is easy to verify that the utility structure in Figure 4 satisfies the conditions specified in C1 through C3, and so is a basic PD. In this case we have a three-person PD. Suppose that the group G = {Row, Column} were composed of agent neutral players. If Side is assumed to be a pure egoist, then the group of agent neutral players cannot do better than a group of agent relative, pure egoists.[18]

S:C

R \ *C*	C	D
C	5, 5, 5	3, 6, 3
D	6, 3, 3	5, 5, 0

S:D

R \ *C*	C	D
C	3, 3, 6	0, 5, 5
D	5, 0, 5	4, 4, 4

Figure 4.

The agent neutral players place an equal weight on all members of G, and seek to maximize individual utility at the collective level for G.[19] Notice that the strongly dominant pure strategy of Side is defection. The group of agent neutral players gain nothing by their neutrality in the context of a non-member who is not agent neutral. Were both Row and Column agent relative, egoistic players, they would for certain realize (D, D, D), the best outcome possible for *each* member of G in the context of a single, egoistic non-member.[20] Since agent neutral players do not have a dominant strategy in this sub-game, they will reach (D, D, D) only if good luck has their strategies coincide to collective and individual benefit. In any case, unlike agent relative players, they are not certain to reach (D, D, D).[21]

Suppose instead that Row and Column are interacting in the context of impure altruists. Were G composed of agent neutral players in the PD to follow, they could not do better than a group of agent relative pure egoists. Assume that Side is an altruist who has some concern for her own well-being. We can imagine that Side places a

weight of 0.4 on the interests of Row and Column, and 0.2 on her own interests. We assume that each member of G is concerned with the collective maximization of individual utilities accruing to each agent neutral player, and that each places an equal weight on the interests of agent neutral players. Figure 5 displays the *unweighed* utilities accruing to Row, Column, and Side.

S:C

R \ *C*	C	D
C	4, 4, 4	2, 5, 2
D	5, 2, 2	3, 3, 0

S:Altr.

R \ *C*	C	D
C	2, 2, 20	0, 3, 20
C	3, 0, 20	3, 3, 20

Figure 5.

S:C

R \ *C*	C	D
C	4, 4, 4	2, 5, 3
D	5, 2, 3	3, 3, 2

S:Altr.

R \ *C*	C	D
C	2, 2, 6	0, 3, 5
D	3, 0, 5	3, 3, 6

Figure 5a.

It is clear in the transformed game depicted in Figure 5a that defection is strictly dominant pure strategy for the altruist.[22] The utilities accruing to each player are rounded off.[23] The group of agent neutral players are afforded no advantage by their neutrality in the context of a non-member who is altruistic. Were both Row and Column agent relative egoistic players, they would for certain realize (D, D, D), the best outcome possible for G in the context of a single, altruistic non-member. Agent neutral players are not certain to reach (D, D, D).

In simple two-person PD's, an agent neutral player who does not secure the cooperation of others will receive the sucker payoff. This much is obvious. What is not obvious is that, in many PD's containing a single player who is not agent neutral, groups of agent neutral players cannot do better in the (collective) maximization of individual utility than groups of agent relative players. It is widely conceded that sub-group cooperation among agent neutral players in PD's will not yield collective benefits rivaling those afforded by universal cooperation. But the equally common assumption that

sub-group cooperation is at least better than sub-group defection is false.[24] In the presence of one agent relative egoistic player, we found that a group of pure egoists would do at least as well as a group of agent neutral players. In the presence of an agent relative altruistic player, we found again that a group of pure egoists would do at least as well as a group of agent neutral players.

Figures 4 and 5a depict non-absolute PD's. In non-absolute PD's generally, sub-groups of agent relative players have an advantage over sub-groups of agent neutral players. In PD's which are absolute, every sub-game of the basic PD is also a basic PD. Partial cooperation in absolute PD's is always better for any sub-group G than universal defection. It is clear by inspection that Figures 4 and 5a depict PD's whose sub-games are not all basic PD's.[25] And the sub-group, G, of agent neutral players is ill-equipped for such games.

In Section 3, I show that a group of agent neutral players are certain to do as well as any group of agent relative players only in the relatively small class of CAPD's. However, for every CAPD, there are any number of agent relative players who will do as well as agent neutral players.[26] Agent neutrality is not necessary for cooperation, even in the small class of CAPD'S. I consider finally, total absolute PD's. In TAPD'S, *only* agent relative players will, for certain, reach the cooperative outcome. Agent neutrality is not sufficient for cooperation in the broader class of TAPD's.[27]

3. AGENT RELATIVE PLAYERS, EGOISTIC CAPD'S AND EGOISTIC TAPD'S

Egoistic prisoner's dilemmas, in general, are games in which at least some groups of egoistic players are certain to reach the uncooperative outcome.[28] Consider again the familiar, two-person PD depicted in (2). Figure 2 depicts an absolute PD of the sort in which agent neutral players are certain to reach the cooperative outcome, as became apparent in Figure 3. Pure egoists, on the other hand, are certain not to reach the cooperative outcome in Figure 2. But, as we'll see, this is no compelling reason for pure egoists to become agent neutral players.

Notice that agent relative, impure egoistic players are also certain to reach the cooperative outcome in Figure 2. Suppose each player

places a weight of 0.65 on his own interests, and 0.35 on the interests of others. The distribution of utilities in the transformed game for these impure egoists are calculated as follows.

$$\begin{aligned} U_R(C,C) &= 0.65[U_R(3)] + 0.35[U_R(U_C(3))] = 3 \\ U_R(C,D) &= 0.65[U_R(1)] + 0.35[U_R(U_C(4))] = 2.05 \\ U_R(D,C) &= 0.65[U_R(4)] + 0.35[U_R(U_R(1))] = 2.95 \\ U_R(D,D) &= 0.65[U_R(2)] + 0.35[U_R(U_C(2))] = 2 \end{aligned}$$

Column's utilities are reversed in outcomes (C, D) and (D, C), and are otherwise the same. The transformed game is displayed in Figure 6. The cooperative outcome in Figure 6 is the unique optimum outcome, and agent relative players will, for certain, reach the cooperative outcome.

		C	
		C	D
R	C	3, 3	2.05, 2.95
	D	2.95, 2.05	2, 2

Figure 6.

Figure 2 is so familiar that it has nearly become the standard formulation of the prisoner's dilemma. The PD depicted in Figure 2, however, is one of a relatively small class of complete and absolute PD's, and no general conclusions about the behavior of any group of players should be drawn from Figure 2 alone. Complete PD's are basic PD's which meet additionally the *total condition* T, and the *group defection condition*, D. In two-person PD's, the total condition ensures that the cooperative outcome, (C, C), has more total utility than either partial defection outcome, (C, D) or (D, C).

T. $R_R + R_C > \max(T_R + S_C, S_R + T_C)$

In PD's that violate condition T agent neutral players may lack a dominant strategy and do considerably worse than a group of agent relative players.[29] Let the partial defection outcomes in Figure 2 be modified to (6, 1) and (1, 6). Agent neutral players are not certain to do as well as impure altruists who weight the interests of others at

0.7, and their own interests at 0.3. Compare the transformed games displayed below.

		C	D
		C	
R	C	3, 3,	4.5, 2.5
	D	2.5, 4.5	2, 2

Impure Altruists

		C	D
		C	
R	C	3, 3	3.5, 3.5
	D	3.5, 3.5	2, 2

Agent Neutral

The group defection condition on complete PD's ensures that universal defection for any sub-group of players is worse than partial defection. In two-person PD's, the condition is specified as follows.

D. $P_R + P_C < \min(T_R + S_C, T_C + S_R)$

In PD's that violate the group defection condition, such as those discussed below, only agent relative players, egoists or altruists, are certain to reach the cooperative outcome.[30] PD's that meet conditions T and C are complete PD's. If in addition a PD meets condition A, it is a complete and absolute PD.[31]

A. For each sub-game i of a PD: $Pi_1 < Ri_1, \ldots, Pi_n < Ri_n$ $(n \geq 2)$

Condition A ensures that for all sub-games i of a PD, the punishment to each member of sub-group G for universal defection in i is worse than the reward for sub-group cooperation in i. Figures 4 and 5a displayed the difficulties for agent neutral players in PD's violating condition A.

Agent neutral players are certain to do at least as well as agent relative players in the relatively small class of CAPD's. The converse is also true. Agent relative players are certain to do at least as well as agent neutral players in CAPD'S. In general, a group G of agent relative players facing an egoistic CAPD will, for certain, reach the cooperative outcome if each member of G is an impure egoist and $w < 0.5$ in principle P below. As above 'Pi', 'Si', 'Ti', and 'Ri' in principles P and P* below represent respectively the utilities afforded player i in the punishment, sucker, temptation and reward outcomes.

P. $$\frac{Pi - Si}{(Tj - Pi) + (Pi - Si)} = w \quad (i, j : i \neq j)$$

We note that in the egoistic PD depicted in Figure 2, principle P yields,

$$\frac{2-1}{(4-2)+(2-1)} = 0.33 < 0.5$$

Each member of G must place a weight w ($0.33 < w < 0.5$) on the interests of others, and G will certainly reach the cooperative outcome.[32] In general, in every case where P yields an w ($0 < w < 0.5$), agent neutrality *will not be necessary* for the achievement of the cooperative outcome. A group of agent relative egoistic players will also reach the cooperative outcome in egoistic CAPD's.

In basic egoistic PD's meeting conditions T and A, but not D, TAPD'S, we find that P yields an w ($0.5 < w < 1$), as in the following. Applying P to Figure 7, we find that w = 0.60. To ensure that a group reach the cooperative outcome in this TAPD, they *must* all be agent relative players. In fact all members of G must be altruistic players who place a weight w ($0.6 < w < 1$) on the interests of others. A group of agent neutral players would not for certain realize the cooperative outcome in this TAPD. Compare the transformed games of a group of impure altruists and a group of agent neutral players, respectively.

		C	
		C	D
R	C	4, 4	0, 5
	D	5, 0	3, 3

Figure 7.

		C	
		C	D
R	C	4, 4	3.1, 0
	D	0, 3.1	3, 3

Impure Altruists

		C	
		C	D
R	C	4, 4	2.5, 2.5
	D	2.5, 2.5	3, 3

Agent Neutral

These impure altruists place a weight of 0.61 on the interests of others and a weight of 0.39 on their own interests. It is clear by inspection that the impure altruistic players will reach cooperation for certain, and that agent neutral players might not reach cooperation. In fact, agent neutrality is not sufficient for cooperation in all egoistic TAPD'S.

4. AGENT RELATIVE PLAYERS AND ALTRUISTIC TAPD'S

Altruistic prisoner's dilemmas, in general, are prisoner's dilemmas in which at least some groups of altruistic players are certain to reach the uncooperative outcome. Altruistic TAPD's are PD's which meet, additionally, the total and absolute conditions.[33] We found, in Section 3, that groups of altruistic players are able to reach cooperation, for certain, in all egoistic TAPD'S, where groups of impure egoists and agent neutral players are not sure to succeed. In many altruistic TAPD's, by contrast, only groups of egoists are certain to reach cooperation.

Figure 8 displays a game that presents no problems at all for egoists, even pure egoists, who derive no utility from the preference-satisfaction afforded other players. Suppose, now, that Row and Column are impure altruists who place a weight of 0.8 on the interests of others and 0.2 on their own interest. Such altruistic players facing a distribution of utilities such as displayed in Figure 8 would find themselves certain to reach a strongly deficient outcome. Figure 8 is transformed for altruistic players as follows.

		C	
		C	D
R	C	3.5, 3.5	5, 0
	D	0, 5	3, 3

Figure 8.

$$U_R(C,D) = 0.2U_R(5) + 0.8U_R[U_C(0)] = 1 \quad U_R(C,C) = 3.5$$
$$U_R(D,C) = 0.2U_R(0) + 0.8U_R[U_C(5)] = 4 \quad U_R(D,D) = 3$$
$$U_C(C,D) = 0.2U_C(0) + 0.8U_C[U_R(5)] = 4 \quad U_C(C,C) = 3.5$$
$$U_C(D,C) = 0.2U_C(5) + 0.8U_C[U_R(0)] = 1 \quad U_C(D,D) = 3$$

Impure altruistic players face the transformed game displayed in Figure 9.

		C	
		C	D
R	C	3.5, 3.5	1, 4
	D	4, 1	3, 3

Figure 9.

Row and Column can escape the uncooperative outcome in Figure 9 only if each of the agents acts more egoistically. To determine, in general, the minimum weight Row and Column must place on their own interests in order for certain to reach cooperation, we use principle P*.

$$\text{P.*} \quad \frac{Pj - Ti}{(Sj - Pj) + (Pj - Ti)} = w \quad (i,j : i \neq j)$$

Applying P* to Figure 8, we find the minimum weight for Row and Column as follows.[34]

$$\frac{3 - 0}{(5 - 3) + (3 - 0)} = 0.60$$

Each player must place a weight of w ($0.6 < w \leq 1$) on her own interests if the group is to reach the cooperative outcome in the game displayed in Figure 9. Each player must behave not only more egoistically, but each must behave as *an egoist* to ensure the realization of the cooperative outcome in this altruistic TAPD.

Compare a group of purely egoistic players and a group of agent neutral players facing a game such as is displayed in Figure 9.

		C	D
R	C	3.5, 3.5	5, 0
	D	0, 5	3, 3

Pure Egoists

		C	D
R	C	3.5, 3.5	2.5, 2.5
	D	2.5, 2.5	3, 3

Agent Neutral

It is evident that agent neutral players might not reach the cooperative outcome. In general, only egoistic players are sure to reach cooperation when facing an altruistic TAPD, such as the one displayed in Figure 9.

5. CONCLUSION

Agent neutral theories, such as act consequentialism, cannot guarantee that agent neutral players will always cooperate in PD's. We found that, in altruistic TAPD'S, only a group of ethical egoists would, for certain, cooperate. In egoistic TAPD'S, only a group of ethical altruists would for certain cooperate. In any event, agent neutrality is never necessary for cooperation in PD's, and is sometimes not sufficient for cooperation.

Neutrality is notoriously demanding on moral agents, and we have shown that *a priori* it offers no greater assurance of maximizing benefits than does ethical egoism or ethical altruism. How well a group of agents does in the collective pursuit of individual utility is determined not by their neutrality, but by the contingent matter of the types of PD's they encounter.

NOTES

* I would like to thank David Gauthier, John Tilley, Wlodek Rabinowicz and an anonymous referee of this Journal for their insights and comments on earlier versions of this paper.

[1] Any principle according to which the criteria of evaluation vary over agents, groups, times, locations, or in any other respect, is an agent relative principle. I use the term 'agent relative' to cover various sorts of relativity, including temporal or geographical relativity, of ethical and rational principles. See Kryster Bykvist, 'Utilitarian Deontologies?' in Wlodek Rabinowicz (ed.) *Preference and Value:*

Preferentialism in Ethics, Studies in Philosophy, Department of Philosophy, Lund University, 1996, 1–16.

[2] Compare the principle of direct, collective self-defeat in Derek Parfit, *Reasons and Persons* (Oxford: Oxford University Press, 1984). Parfit maintains that self-interest theory (ethical egoism, in my case) is not directly, *individually*, self-defeating, but is directly *collectively* self-defeating. A theory T is

directly individually self-defeating when it is certain that if someone successfully follows T, he will thereby cause his own T-given aims to be worse achieved than they would have been if he had no successfully followed T. (section 22, p. 55)

In what follows I assume that ethical egoism is not directly individually self-defeating. Contrast CT above with an alternative principle of collective self-defeat. A theory T is

directly collective self-defeating when it is certain that, if we all successfully follow T, *we* will thereby cause *our T-given* goals to be worse achieved than they would have been if none of us had successfully followed T . . . (section 21, p. 54, underlining added).

Agent relative principles such as ethical egoism and ethical altruism give different aims or a goals to different agents. In evaluating agent relative principles, we can apply CI, but not a principle (such as the one above) which assumes common goals. Principles which assume common goals make all agent relative principles *trivially* rational. Parfit of course recognizes this, and applies a principle similar to CI in evaluating the collective rationality of agent relative principles.

[3] It is crucial here that we are considering the collective (not individual) maximization of individual (not collective) utility. We are asking whether, if *we* universally conform to principle P, will *each* be worse off in P's terms. We are not asking whether, if we universally conform to P, will *we* as a group (though perhaps not each) be worse off in P's terms.

[4] See Samuel Sheffler, *The Rejection of Consequentialism* (Oxford: Oxford University Press, 1987).

[5] Wlodek Rabinowicz has shown that a future-oriented, temporally relative version of act utilitarianism is directly collectively self-defeating. See his, 'Act-Utilitarian Prisoner's Dilemmas' *Theoria* 55 (1989) 1–44. A geographically relative analogue of the principle yields the same result.

[6] We cannot draw the additional conclusion that agent neutral theories are therefore collectively irrational in the absence of an additional principle supplementing CI above. Instead, I draw the more cautious conclusion that, if the goal is the collective maximization of individual utility in PD's, then we need not adopt so demanding an approach as that endorsed by agent neutral theorists. Agent relative players can often better achieve that goal. For suggested clarifications of this point I thank an anonymous referee for this Journal.

[7] The name and notion of an absolute PD were suggested to me by Wlodek Rabinowicz in correspondence. An absolute PD, in brief, is a PD all of whose sub-games are PD's as well. Alternatively, an absolute PD is a PD which meets condition A specified and discussed below, pp. 11–13.

[8] The conditions on complete and absolute PD's (CAPD's) and total and absolute PD's (TAPD's) are presented and discussed below, pp. 11–13. I want to note here only that the success of agent neutral players in reaching cooperation for certain is restricted to a certain class of prisoner's dilemmas. And even within that class, there are many groups of agent relative players who also reach cooperation.

[9] 'Deficient' is sometimes used in a weak sense meaning Pareto deficient. An outcome is Pareto deficient if and only if there is some other outcome preferred by someone and not dispreferred by anyone. I will use 'deficient' in the strong sense. An outcome is strongly deficient if and only if there is some other outcome that is preferred by *everyone*.

[10] A similar method for modeling the basic prisoner's dilemma is found in Steven T. Kuhn and Sergei Moresi, *op. cit.*

[11] A strategy S_i is strictly dominant if and only if for any strategy chosen by other players, S_i yields a utility payoff to player i greater than that yielded by any other strategy S^*_i. We could have defined the PD by replacing one of the strict inequalities with a weak inequality. The structure would remain a PD, but instead of the strong dominance of defection we would, in some cases, have weak dominance. A strategy S_i is weakly dominant if and only if for any strategy chosen by other players, S_i yields a utility payoff to i at least as great as any other strategy S^*_i, and for some strategies chosen by other players, S_i yields a utility payoff to i greater than any other strategy S^*_i. For our purposes, the more restrictive defining conditions do not affect the argument.

[12] For a discussion of altruistic players modeled in this way see John Tilley, 'Altruism and the Prisoner's Dilemma', *Australasian Journal of Philosophy*, Vol. 69 (1991) 264–287. For an earlier and similar method of modeling altruistic and egoistic players, see Nicholas Rescher, Unselfishness (Pittsburgh: University of Pittsburgh Press, 1978) and John A. Weymark, 'Unselfishness and Prisoner's Dilemmas', *Philosophical Studies* 34 (1978) 417–425.

[13] When we come to the discussion of impure egoists and impure altruists, this assumption will simplify the transformation of games under various assumptions about individual weightings. The weightings noted are scaling constants. Scaling constants are used in situations described as "multiattribute decision problems", where the overall utility payoff to one or more agents is determined, in part, by the utility derived from the payoff to another. Under the assumption that prizes are linear with utility, games can be transformed by applying the scaling constants directly to the utilities received, and discussion of prizes can be left out altogether.

To avoid any confusion, let me make the assumption explicit. Supposing that Ui is the utility to i and $i is the prize to i, we are assuming that U($i) is linear with $i, for all agents in every game to follow. Further, we assume that if Uj is the utility to j (where j =/= i), then Ui[Uj($j)] is linear with Uj($j), for all agents in every game to follow. See R. Keeney and H. Raiffa, *Decisions with Multiple Objectives* (New York: Wiley, 1976).

[14] Since it is assumed throughout that there are no population changes, maximizing average utility is equivalent to maximizing total utility.

[15] We assume that the preferences of each agent are such that the unspecified attribute or prize whose utility to Row (Column) is specified on the left (right) in each cell of the matrix is additively independent of the other attribute. See, H. Raiffa and R. Keeney, *ibid.*

[16] Figure 3, of course, is not a PD. Agent neutral players reach cooperation for certain in some two-person games that frustrate ethical egoists precisely because, when faced with an egoistic PD, it is transformed for agent neutral players.

[17] In maintaining that a group of agent neutral players would do better, in certain PD's, were they agent relative players, I am not making the well-known point that each player would do better if he, alone, defected. Rather, in certain PD contexts, *all* agent relative players do better if *all* defect.

[18] The many-person PD depicted in Figure 4 is sufficiently familiar to have been cast as a *Foul-dealer and Benefactor Dilemma*. Cf. Wlodek Rabinowiez, 'Cooperating with Cooperators', *Erkenntnis* 38 (1993) 23–55, and Philip Pettit, 'Free Riding and Foul Dealing', *Journal of philosophy* (1986) 361–379.

[19] In Figures 4 and 5, we could instead assume that the agent neutral players place an equal weight on all players, including the defector, and not just members of G. The argument is unaffected by that assumption.

[20] It is worth noting that there are no situations in which a group of egoists in the sub-game for certain reach an outcome that is strongly superior to the outcome reached, for certain, by agent neutral players. Assume that the agent relative player has defected. Let 'C' represent sub-group cooperation, 'P' represent sub-group partial cooperation , and 'D' represent sub-group defection. In order to guarantee that agent neutral players reach D in this sub-game, it must be true that $C < P < D$, and so $C < D$ and $P < D$. If agent relative players reach C or P, then they have reached an outcome inferior to D. In short, agent relative players in such a sub-game cannot do better than agent neutral players. Suppose that we design the sub-game so that agent neutral players are certain to reach C. To do so we must assume that $C > P > D$, and so $C > D$ and $C > P$. Agent relative players in this sub-game can again do as well or worse than agent neutral players. In the game I describe in Figures 4 and 5, agent relative players might do *better* than agent neutral players. Since agent neutral players are acting independently and without a dominant strategy, they can only hope that good luck will have their strategies coincide to collective benefit. If so, then they will do as well as agent relative players; otherwise, they will do worse.

[21] I do not here explore the possibility of employing utilitarian metastrategies which might help agent neutral players overcome some of the difficulties raised here and below. I am focused here on what the precise advantages are of agent neutral principles as compared to agent relative principles. In this context, agent relative principles handle such difficulties more easily. It is worth noting that J. Howard Sobel discusses similar problem faced by constrained maximizers acting in the context of straightforward maximizers. See his 'Straight Versus Constrained Maximization', *Canadian Journal of Philosophy*, Vol. 23 (1993) 25–54.

[22] In Figure 5a, 'S: Altr.' represents 'Side acts atruistically'.

[23] The utilities in Figure 5 are, as noted, unweighed. It is the distribution of utilities accruing to each player assuming that no player is afforded any utility from the preference-satisfaction of any other player.

[24] I am not, however, maintaining that agent neutral principles are therefore collectively irrational. I am rather objecting to the familiar suggestion that increasing the number of cooperators in PD's yields corresponding benefits. The view is strongly suggested in simple, 2×2, PD's, but we find a similar idea expressed in

Thomas Hobbes' admonition that we form confederations as a way of ensuring more likely survival in the state of nature.

. . . [I]n a condition of war wherein every man to every man . . . is an enemy. There is no man can hope by his own strength or wit to defend himself from destruction without the help of confererates (where everyone expects the same defense by the confederation that anyone else does). *Leviathan*, (Inffianapohs: Hackett Publishing Company, 1994) Chapter 14, section 4.

Gregory Kavka and David Gauthier seem to make similar suggestions.

It seems evident, however, that a strategy of group formation and collective defense would offer the individual greater protection and security against 'forces united, to dispossess and deprive him'. *Hobbesian Moral and Political Theory*, (New Jersey: Princeton University Press, 1986) p. 126 ff.

. . . [S]hould we expect both groups of reciprocal altruists and groups of egoists to exist stably in the world? Not necessarily. The benefits of cooperation ensure that, in any given circumstances, each member of a group of reciprocal altruists should do better than a corresponding member of a group of egoists. Each reciprocal altruist should have a reproductive advantage. Groups of reciprocal altruists should therefore increase relative to groups of egoists in environments sin which the two come into contact." Moral By Aueement (Oxford: Oxford University Press, 1984) p. 188 ff.

Contrary to agent neutral theorists, Hobbes, Kavka and Gauthier, sub-group defection, wholesale, might be consistently better for the sub-group than sub-group cooperation. It is a purely empirical question, depending upon the types of PD's groups encounter. For clarifications of this particular point, I thank an anonymous referee for this Journal.

[25] When the behavior of Side is treated as background information, and we restrict our attention to Row and Column exclusively, we can drop out the utilities accruing to Side and represent the sub-group game faced by Row and Column (given Side's behavior) in Figures 4 and 5a as follows.

		C	D
		C	
R	C	2, 2	0, 3
	D	3, 0	3, 3

Fig. Sub-4

		C	D
		C	
R	C	3, 3	0, 5
	D	5, 0	4, 4

Fig. Sub-5a

Neither of these sub-games is a PD, since each of the sub-games violates condition C3 above: the uncooperative outcome in the sub-game is preferred to the cooperative outcome. That kind of result for sub-games does not occur in absolute PD's.

[26] To illustrate, take the standard, simple, 2 × 2 PD's. As noted, these are CAPD's. On p. 13 above, we found that any group of egoistic, agent relative players who placed a weight greater than 0.33 and less than 0.5 on the interests of others would

reach cooperation, and do as well as any group of agent neutral players. But there are infinitely many such groups of agent relative players. If we included altruistic, agent relative players, cooperation would be reached by any group who places a weight greater than 0.5 on the interests of others. Again, there is an infinite number of such groups.

[27] It is important to note here that I am searching for the entire class of PD's in which agent neural players do as well as agent relative players. In claiming that agent neutral players do as well as agent relative players only in the class of CAPD'S, I am claiming that it is only in the class of CAPD's are they *certain* to do as well. I am not claiming that in *every* PD that violates conditions T, D, or A, there is a group of agent relative players who will do better than the group of agent neutral players, though this may also be the case. Rather, my view is that in some PD's violating conditions T, D, or A, agent neutral players do not do as well as agent relative players, and so in the classes violating T, D, or A, agent neutral players are not certain to do as well as agent relative players. I use the examples to follow (and above) to illustrate PD's violating T, D, or A. In these agent neutral players are not guaranteed to do as well as agent relative players. I thank an anonymous referee for this clarification.

[28] Groups of pure egoists always do poorly in egoistic prisoner's dilemmas. However, as we'll see, in egoistic CAPD's, impure egoists can do very well, despite the fact that these games frustrate groups pure egoists.

[29] Condition T is a strengthened version of condition U, in which inequality is replaced by weak inequality. Condition U is introduced in Steven T. Kuhn and Sergei Moresi, *op. cit.*. The version of T generalized to many-person PD's, and their sub-games, is as follows, letting D range over partial defection outcomes.

T^*. $R, + \ldots + R. > \max(D_1, \ldots, D_{2^n-2})$ (for $n \geq 2$).

[30] The version of D generalized to many-person PD's, and their sub-games, is as follows, letting D range over the partial defection outcomes.

D^*. $P, + \ldots + P. < \min(D_1, \ldots, D_{2^n-2})$ (for $n \geq 2$).

[31] Note that condition A is just a generalization of condition C3 on basic PD's. In a basic PD, C3 ensures us that the Reward payoffs to each player exceed the Punishment payoffs to each player. Condition A ensures us that, for every sub-game in a many-person PD, sub-group cooperation (or, sub-group Reward) payoffs for each player exceed sub-group defection (or, sub-group Punishment) payoffs for each player. In each sub-game then, under condition A, the remaining cooperators never do better by collectively defecting.

[32] Of course, each member of G could reach the cooperative outcome by placing a weight of n ($n \geq 0.5$), but then no member of G would be an impure egoist, contrary to our assumption.

[33] For interesting discussions of altruistic prisoner's dilemmas see John Tilley, *op. cit.*, John A. Weymark, *op. cit.*, Nicholas Rescher, *op. cit.*, and F. Schick, *Having Reasons* (Princeton: Princeton University Press, 1984).

[34] P^* determines the minimum weight that altruistic players must place on their own interests in order, for certain to reach cooperation. I assume that altruistic

players to be reluctant to place more than the minimum weight on their own interests. P determines the minimum weight that egoists must place on the interests of others in order, for certain, to reach cooperation. I assume that egoists are reluctant to place any more than the minimum weight on the interests of others.

Division of English, Classics, Philosophy, & Communication
University of Texas at San Antonio
San Antonio, TX 78249-6643
U.S.A.