Co-operative Solutions to the Prisoner's Dilemma

Author(s): Duncan Macintosh

Source: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, Dec., 1991, Vol. 64, No. 3 (Dec., 1991), pp. 309-321

Published by: Springer

Stable URL: https://www.jstor.org/stable/4320264

DUNCAN MACINTOSH

CO-OPERATIVE SOLUTIONS TO THE PRISONER'S DILEMMA*

(Received in revised form 14 March, 1991)

1. INTRODUCTION: THE PRISONER'S DILEMMA

In the Prisoner's Dilemma (PD), two agents will get certain jail sentences depending on how both independently next choose. They hate jail, so each will get a utility inversely proportional to his jail sentence. If one Co-operates and the other Defects, the former gets 1 utile, the latter 4. If both Co-operate, both get 3; if both Defect, 2. If a rational agent chooses so as to maximize his individual expected utility, since he maximizes whatever the other does if he Defects (i.e., Defecting 'dominates' Co-operating), each will Defect and get 2 utiles. But since if both would Co-operate each would get 3, many philosophers think Co-operation must somehow be rational. It is often thought that if proven, this would show two things. First, while to be rational is to advance one's preferences, that is not necessarily to maximize in every choice. Second, it is rational to be moral, to refrain from exploiting others.[1] Here, I develop a new Co-operative solution to the PD in the course of criticizing the solutions of David Gauthier and Amartya Sen.

2. PROBLEMS WITH GAUTHIER

One might think the agents should just agree to Co-operate, then act on the agreement. But all either wants is the shortest possible jail time. So in spite of the agreement, each should Defect for the shorter time. Thus, given their preferences and given that to be rational is to maximize, they cannot rationally keep the agreement. So it is pointless to make it, rationally impossible to sincerely make it.

But suppose people can so dispose themselves that if they genuinely make an agreement, they will keep it. Were you about to face a PD, it would be rational (because maximizing) for you unilaterally to dispose

Philosophical Studies **64**: 309—321, 1991.

yourself to Co-operate with just those like-disposed. For if you meet such a person in a PD, he will Co-operate with you to your advantage, seeing you have a disposition with which he is disposed to Co-operate. Of course since you are disposed to Co-operate with such people, you will Co-operate to his advantage. You each do less well than by unilateral Defection, but better than had you both Defected, as you would without your dispositions. You are safe from victimization by habitual Defectors, for since they lack the disposition which makes you Co-operate, you may Defect. Yet you can exploit unconditional Co-operators, since they are not disposed to Co-operate *only* with those like-disposed, but with everybody. So you may Defect against them, to your advantage.[2]

But this PD is really two problems. First, suppose an agent will Co-operate if one gives him a credible guarantee that one will reciprocate: Is it rational to give it? Second, is it then rational to Co-operate? Call giving the guarantee (i.e., acquiring the conditional Co-operator's disposition, the "CCD"), "Intending"; call Co-operating after having given it (i.e., after having acquired the CCD, and having noted the other player has one too), "Acting". Assume the game is played sequentially: First, you manifest your Intentions. Then the other manifests his. Then he chooses to Act or not. Then you choose. It does not matter whether you know how he chose, but both your earlier Intentions are common knowledge when you each later choose whether to Act. We now have two problems. First, is it rational to Intend — to have a maximizing intention to do a non-maximizing action?[3] Second, the compliance problem: is it rational to Act — to comply with a maximizing agreement to do a non-maximizing action?

Gauthier argues that (i) it is maximizing and so rational to Intend, (ii) it is rational to act on a rational intention, and so (iii) it is rational to Act. His Defense of (i) and (iii) depends on his reading of (ii). This reading either changes classical rationality or deduces an unexpected consequence from it. Classically, an action is rational just if it maximizes, an intention, just if an intention to do a maximizing action. But for Gauthier, an action is rational just if dictated by an intention it maximizes to adopt; an intention, just if it maximizes to adopt it. Classical maximizers normally understood can rationally neither intend

nor do non-maximizing actions. But Gauthier thinks his "constrained maximizers" (whose dispositions sometimes stop them from doing individually maximizing actions) can do both.[4]

But can they rationally do both? Many think not.[5] Their argument: Gauthier's agents find an action rational just if dictated by a disposition it maximizes to adopt. Now in the sequential PD, it initially maximizes to adopt a CCD, for when similar agents see it in one, they will Co-operate, to one's advantage. But after they have chosen among actions, it no longer maximizes to have a CCD. It now maximizes to adopt a Defector's disposition (since one always does best by Defection). So by Gauthier's standards, a rational agent should now dispose himself to Defect. He should then Defect. Informed PD agents would see this and so would not Co-operate given a choice. The CCD "divides through," and free and rational Gauthier agents will behave as classical maximizers.

Now it would advantage each to have a disposition, irrevocable in the circumstances, that would *force* him to Co-operate with anyone like-disposed, for it would then genuinely guarantee Co-operation to a similar agent, making him Co-operate, to the first agent's advantage. Thus, Gauthier has an argument for (i) if the CCD is an irrevocable causal mechanism which forces its agent to Co-operate. But in being forced to do so when the standard of rationality says a choice is rational only if dictated by a maximizing disposition, one will not be Co-operating freely and rationally. Rather, one is caused to behave irrationally by a disposition it was rational to adopt, but which is no longer the rational one to have and act upon. It is now rational to adopt a different one (though one cannot if the first is irrevocable, as it must be to have been advantageous). So Gauthier has not solved the compliance problem. His agents may behave compliantly from a disposition forcing them to comply, but they will not be acting rationally, not even by *his* standards.[6]

This also threatens his solution to the intention problem (intentions are rational if maximizing and if the acts intended are rational, which *they* are if intending them is; it initially maximizes to Intend, so Intending is rational). He agrees that it is only rational to intend a rational action, which he defines as one from an intention it maximizes

to adopt. But after the other chooses his action, it maximizes to intend to Defect. So at that time, intending to Co-operate is irrational, and Co-operating then would thus be an action from an irrational intention, and so irrational. Since it would be irrational to Co-operate then, one cannot rationally intend to do so earlier.

What went wrong? Gauthier thought that classical maximizers could not find Co-operation rational given their preferences. But the aim of rational choice is maximization; since it would maximize were a rational agent able to make and keep Co-operative commitments, both must be rational. Since, given his preferences, he cannot do either in classical rationality, Gauthier thought it must be false; a choice is not rational if maximizing, but if dictated by a maximizing disposition (commitment, intention). But the preferences making Co-operation non-maximizing also guarantee that when one is to comply with the CCD it will not be a maximizing disposition. Thus we might conclude that his theory of rationality must also be false.

But maybe neither theory is really false. Both accounts stumbled because of the agent's preferences. It was because Co-operating was ultimately non-maximizing that it proved irrational, and this seemed to prove the falsity of the two theories. But since it is the agents' initial preferences which prevent them from making and keeping advantageous commitments, perhaps the PD is not a *reductio* of these conceptions of rationality, but of the rationality of continuing to have the PD preferences which make rationally Co-operating impossible. For surely a choice is rational not just if it maximizes on present preferences; they must also be ones it is rational to have. And it seems to some philosophers[7] that it is irrational for PD agents to keep their initial preferences. Rather, to maximize on these, they must adopt ones which would rationalize Co-operating.

But *which* preferences should they adopt? We know it must maximize to adopt them given one's PD preferences (else one has no reason to adopt them), maximize on them for one to Co-operate with those with like preferences (or one will not find Co-operation rational). But we do not know what to prefer, nor in what order. Amartya Sen's is the classic study of what preferences would rationalize PD Co-operation. Maybe rational agents should adopt these . . .

3. SEN'S MORALITY AND THE RATIONALITY OF CO-OPERATION

Sen[8] says each PD agent has the following preferences (where D = Defect, C = Co-operate; the agent's action is the left-most in each couple, his partner's, the right; each letter is an action, each pair of letters an outcome, their ordering from left to right an ordering of outcomes from most to least preferred): DC, CC, DD, CD.[9] Defection dominates so both agents do it, each getting only his third-best outcome. This, Sen thinks, shows a conflict between individual rationality and both individual and social optimality. If each agent rationally maximizes, each will fail to make the choice which, if both made it, would make the first agent better off than if both choose any other way, and both agents as well off as possible without at least one doing worse.[10] He suggests that a moral person would have one of two other preference orderings. First are those of an "Assurance Game" (AG): CC, DC, DD, CD.[11] Second are "Other Regarding" (OR) preferences: CC, CD, DC, DD.[12] Sen thinks that in the AG, each would rationally Co-operate if assured the other would, an assurance given by their common knowledge of each other's preferences and rationality. In the OR, each would rationally Co-operate regardless since each prefers most to have Co-operated whatever the other does. Sen then claims that if each PD agent acted *as if* he had one of these orderings, both would Co-operate, each getting his second rather than third-best outcome by his PD preferences.[13]

Sen's conclusions: First, if agents have moral preferences, moral and rational recommendations for behavior are consistent. Second, if people act *as if* they had moral preferences they do better by their non-moral ones.[14] Third, people should therefore socialize each other to have or act as if they had moral preferences.[15] Fourth, individual welfare orderings (orderings of choices by the levels of welfare they would cause for the agent) needn't be identical with revealed preferences (orderings of choices by whether they maximize). Fifth, we can 'define a moral ordering not directly on the space of outcomes (or actions) but on that of the orderings of outcomes (or actions)':[16] One is moral not only if one prefers such and such outcomes or actions, but also if one prefers (or is, somehow, prepared to so act) that those preferences get

satisfied in a certain order of priority. You are moral if you prefer not to hurt people (here, not to Defect), but also if you prefer (or are prepared) not to act on evil preferences. It is unclear whether Sen thinks people ought rationally unilaterally to revise (or act as if they had revised) their preference orderings into those of AG or OR agents. He *does* think they should be encouraged to have such orderings, or to act *as if* they did.[17]

Now, if Sen's preferences will afford a rational, Co-operative solution to the PD, it must be rational to adopt and C-operate from them, or to act as if one had adopted them and Co-operate. Would it be rational to revise one's PD preferences into those of OR, or to act as if one had? No. It would not be better for me to so act, for I always do better by my PD preferences if I Defect, but acting as if I had OR preferences asks me to Co-operate. Nor would it be better for me to actually adopt OR preferences, for if you do not, you can exploit me; if you do, better that I not have changed so that I could exploit you. Forseeing this, I can't rationally supplant my PD preferences with OR ones.[18]

We might try fixing this with Gauthier's insight: one should only be ready to Co-operate with those whom this readiness would make Co-operate.[19] And when Sen said that we should socialize each other into changing our preference orderings or into being ready to act as if we had, he may have meant that we should each try to make *both of us* do this and not Co-operate until we succeed.[20] But if I adopt OR preferences I will Co-operate whatever yours are; if you have not revised yourself, you could exploit me. So I should not revise; same for you. We can't get started here.

Perhaps I should *dispose* myself to choose *as if* I had OR preferences with *just* those like-disposed; same for you. I can do that unilaterally without fear of exploitation, since I will not be made to Co-operate unless you change too. But now we have the problem which did in Gauthier. I retain my PD preferences, but have a disposition to act as if I had OR ones with those who also have them or have a disposition to act as if they did. But I only adopted the disposition because having it was supposed to be maximizing when I faced those like-disposed — it was supposed to make them Co-operate, to my advantage. But since I still have PD preferences, after you have seen my disposition and

chosen your actions, it would maximize for me to abandon my disposition to choose as if I had OR preferences and to adopt one to choose as if I had PD ones, then Defect; same for you. So it would be irrational to comply with an OR disposition. If we guarantee compliance by making the disposition permanent and irrevocable, it will only force Co-operation, not rationalize it.

But maybe things could go like this. I retain my PD ranking of outcomes; I still prefer DC to CC, CC to DD, and DD to CD. But I add a preference to choose as if I preferred CC to CD, CD to DC, and DC to DD provided you so prefer too. The first we might describe as a ranking of outcomes, the second, a ranking of choices given the other agent's ranking of choices. I still prefer minimal jail time, but prefer to act as if I preferred not to let you down if you prefer likewise. The hope is that this will save me from exploitation, yet give me preferences which rationalize my Co-operating if you have appropriate preferences. But this gives me ill-ordered preferences. E.g., I prefer DC to CC by my PD preferences, but CC to DC by my OR ones.

Maybe we can we fix this by saying that I should prefer *more strongly* to choose like an OR with suitable other agents than to get less jail. So I prefer first *to make the choice that would contribute to* CC, second, the one that would contribute to CD, third . . . DC, fourth . . . DD; I prefer fifth, *outcome* DC, sixth, outcome CC, seventh, DD, eighth, CD. But my preferences for actions here amount to preferences for outcomes, since they are not merely the preference to Co-operate or the preference to Defect, but the preference first to Co-operate where the other Co-operates, second, to Co-operate where he Defects, third to Defect where he Co-operates, fourth, to Defect where he Defects. This, combined with my PD preferences for outcomes, again gives me ill-ordered preferences, for I in effect prefer CC both most (by my OR ranking of choices given the other's choice-ranking) and sixth-most (by my PD ranking of outcomes), CD both second and eighth-most, etc.

There is, then, no rational way individually to use OR preferences, whether by supplanting or supplementing PD with OR ones, or by being conditionally disposed to choose as if by OR ones while retaining PD ones. Supplanting is not maximizing, supplementing causes ill-ordered preferences and disposing causes a compliance problem. Ordering problems will also be caused by supplementing PD outcome

orderings with AG action orderings where they require one to act differently than someone with only PD preferences, as where, if actions are public, PD preferences dictate Defection with those who Co-operate, but AG preferences, Co-operation. For one will then prefer DC to CC by the first, CC to DC by the second. Thus, contra Sen, we cannot understand a moral agent as someone who has PD orderings on outcomes, but AG or OR orderings on actions. He is conflicted, not moral. In general then, one cannot acquire a preference ordering on actions which would require one to Co-operate, and yet keep one's preferences well-ordered if one retains a preference ordering on outcomes requiring Defection (by dominance). One's ordered preferences for actions deriving from one's ordered preferences for outcomes will collide with the ordered preferences for actions Sen pushes.[21]

But might it not be individually rational to *replace* one's PD preferences with *AG* ones? Let us see. With them, if you have PD preferences, I know you will Defect. Since I cannot then get my first or second-best AG outcomes, CC or DC, I will Defect for my third-best, DD. Were we both AGers, were our choices public, we could each make the other Co-operate by doing so ourselves. For if either of us does, so must the other for his best outcome, CC. So if our actions are public, adopting AG preferences gives an advantage over keeping PD ones. But it will make one Co-operate with ORers; better by one's PD preferences to have adopted ones that would let one Defect against them. So AG preferences do not maximize compared to ones that make one Co-operate with AGers, but Defect against PDers *and* ORers. Further, if our actions are secret, as in the normal PD, that we each most prefer CC will not make it rational for either of us to Co-operate; for preferring that we both Co-operate is not preferring to individually Co-operate whatever the other does. So AGers will not Co-operate in private choices, nor will PDers find becoming AGers rational for private action games.

4. THE PROBLEM SOLVED: RATIONAL PREFERENCES FOR THE PD

So *which* preferences should one adopt? They must have the functional properties of Gauthier's dispositions, must make one Co-operate just

where their doing so would make the other agent do so too. They must allow one to Defect against PDers *and* ORers. And they must work for public action *and* secret games. Thus they must be free of what Richmond Campbell calls "the circularity problem":[22] each agent had better not prefer (or be disposed) to Co-operate just if the other prefers (or is disposed) to Co-operate, for then neither yet prefers (or is disposed) to Co-operate, and nothing will ever trigger Co-operation. AG preferences have this problem with secret actions. Each AG agent prefers to Co-operate if he thinks the other will Co-operate, but neither has reason to Co-operate simply in knowing both feel this way. Finally, there must not result an ill-ordering of preferences. We saw that it will not work to add to PD preferences a preference about how to act regarding them, e.g., to bring about first, the second-most preferred outcome (here, CC), for this made ill-ordered preferences, even if we call the meta-preference a *stronger* preference. Rather, the agent must become such that the choice maximizing given all his preferences (i.e., "on balance") would be to Co-operate with similar agents; he must *replace* the old preferences with the new.

Here is a preference ordering satisfying all of these requirements. (Assume the agents will know each others' preferences and that each other are rational, but that their choices among actions may or may not be secret.) Each agent should prefer to: (1) Defect against (i) anyone who he knows did, will, or likely will Defect, (ii) anyone unconditionally disposed to Co-operate, and (iii) anyone unconditionally disposed to Defect; (2) have outcome CC; (3) Co-operate with just (iv) those disposed to choose as if their first two preferences were, in order, (1) and (2) and who do not fit (i); (4) have outcome DC with agents who satisfy (iv); (5) have outcome CD. He will then Defect on those who fit (i), (ii) or (iii) for that directly maximizes given his strongest preference, (1). But even if actions are secret, with someone, B, who fits (iv) and not (i), he thinks: "I can't Defect on the rationale of (1) because B does not fit (i), (ii) or (iii). I can satisfy (3) by Co-operating because B fits (iv) and not (i), while if I Defect, I can only satisfy (4). So I have sufficient reason to Co-operate. But I have even more reason in that my doing so will likely help satisfy (2), since B also has those reasons to Co-operate. So, *a fortiori*, I should Co-operate." Since conditions with different rankings are different conditions, there is no ordering conflict here.

It is rational to individually adopt this ordering since that maximizes with *all* agents. It lets one Defect against PDers, suckers and those who are neither, but who are known to Defect through some error. Yet it guarantees that one will Co-operate to those for whom this is necessary and sufficient to make them Co-operate. It gives one a preference to Co-operate with similar agents, beating the circularity problem. And one has reason to Co-operate not depending on the other's actions, but only on his basis for action, so we preserve the independence of the agents' actions; one's choice does not determine the other's, only one's basis for choice.[23]

Sen thought that to get individual and social optimality in a PD, we must sometimes separate individual welfare orderings and revealed preferences, must distinguish actions which would make the agents better off from maximizing actions. They must then, in Co-operating, act against the preferences for action they would normally derive by dominance from their original outcome preferences, so action need not "reveal" those preferences. Sen envisioned agents being socialized into choosing in ways that do not reveal their PD preferences. But we can now identify welfare with revealed preference and still see rational Co-operation in a PD (no socialization required). For if one changes one's preferences so that one finds it maximizing to Co-operate with just those whom just this makes Co-operate, this would make them Co-operate; so modifying oneself maximizes, reveals one's initial PD preferences. If one then faces such an agent, Co-operating reveals one's modified preferences, since one now has ones on which that maximizes. At each juncture one's rational choices reveal one's preferences. So at no choice-point must individual welfare orderings conflict with revealed preferences, nor Co-operation not reveal them.[24]

But how can one prefer to Co-operate where dominance argues preferring to Defect? PD agents prefer minimal jail time, so they prefer outcomes where they Defect and the other Co-operates. They do not prefer all outcomes where they Defect to all where they Co-operate. E.g., they prefer CC to DD. But where they can do nothing about the other agent's choice, each prefers to Defect since that maximizes, whatever he chooses. But what would they rationally prefer in a choice between being able to Defect whatever the other does (i.e., keeping their current preferences), and being inclined to Co-operate where this

inclination would make both themselves and their partners Co-operate (i.e., changing their preferences so Co-operation would maximize with such partners)? Since the second could get them CC, and since they prefer it to DD (all they could get in the *status quo*) they would prefer the second. So the preferences which normally, by dominance, justify their preferring to Defect, will instead justify their preferring to revise their preferences where they may affect each others' choices indirectly by individually altering their own preferences.

5. CONCLUSION AND PROLEGOMENON

When facing a PD, Gauthier's arguments rationalize adopting a disposition which will make one Co-operate just when it would make others do so. He was right on its functional properties: one is made to adopt it by that being maximizing, but it only is so if it makes one Co-operate just when its doing so will make others do so. But to rationalize Co-operating, the disposition must *constitute* a revised preference-ordering. It is maximizing and so rational to revise one's preferences, maximizing and so rational to Co-operate with the right kind of agent given one's new preferences. The intention problem is solved by its being maximizing to Intend with revised preferences, maximizing and so rational to comply given them; since it is rational to comply, there is no objection to intending from problems in the rationality of complying. And the compliance problem is solved in its being maximizing to comply given one's new preferences. Since the agent maximizes both in revising his preferences, and in Co-operating from his revised preferences, his act of Co-operation is *not* constrained; he straightforwardly maximizes throughout.

So we learn two things about rationality: Agents *can* rationally make and keep maximizing commitments to do initially non-maximizing actions, for it is rational for them to acquire preferences that make so acting maximizing. And a choice is rational just if it maximizes on preferences it is rational to have. It is rational to have current preferences, P, just if keeping them is no less maximizing by their measure than having some others, P^*. But if having P^* would maximize on P (because it would more likely cause the conditions preferred in P), one must supplant P with P^*. One must then maximize on P^*, one's new

basis for choice. This vindicates the maximization conception of rationality, except that it applies it not just to choice of means to ends, but also to the choice of ends (given current ends).[25]

NOTES

* For helpful comments, my thanks to Robert Bright, Richmond Campbell, Julia Colterjohn, Peter Danielson, Ish Haji, Keith Lehrer, Robert Martin, Victoria McGeer, Terrence Tomkow and an anonymous referee. I also thank the Canadian S.S.H.R.C. for a Doctoral Fellowship and Dalhousie University for a Killam Post-Doctoral Fellowship.

[1] E.g., see David Gauthier, 'Morality and Advantage,' *The Philosophical Review*, 76 (1967), pp. 460—475, and *Morals By Agreement* (Oxford: Clarendon Press, 1986), Chs. V, VI.

[2] Gauthier, *Morals By Agreement*, Chs. V, VI.

[3] See Gregory Kavka, 'Some Paradoxes of Deterrence,' *The Journal of Philosophy*, 75 (1978), pp. 285—302.

[4] Gauthier, 'Morality and Advantage,' and 'Deterrence, Maximization, and Rationality,' *Ethics*, 94 (1984), pp. 474—495.

[5] Mark Vorobej so objects re the Deterrence Dilemma (DD) in 'Gauthier on Deterrence,' *Dialogue*, XXV (1986), pp. 471—476. I object more fully re the PD in 'Libertarian Agency and Rational Morality,' *The Southern Journal of Philosophy*, XXVI (1988), pp. 399—425, 'Two Gauthiers?,' *Dialogue*, XXVIII (1989), pp. 43—61, and 'Preference's Progress,' *Dialogue*, XXX (1991), pp. 3—32; and are the DD in 'Retaliation Rationalized,' *Pacific Philosophical Quarterly* 72 (1991), pp. 9—32. See also Richmond Campbell, 'Moral Justification and Freedom,' *The Journal of Philosophy*, LXXXV (1988), pp. 192—213.

[6] On possible replies, see my papers in note 5, above.

[7] E.g. Edward F. McClennen, (1985) 'Prisoner's Dilemma and Resolute Choice,' in Richmond Campbell and Lanning Sowden, eds., *Paradoxes of Rationality and Cooperation* (Vancouver: The University of British Columbia Press, 1985), pp. 94—104, and 'Constrained Maximization and Resolute Choice,' *Social Philosophy and Policy*, 5 (1988), pp. 95—118.

[8] Amartya Sen, 'Choice, Orderings and Morality,' in Stephan Korner, ed., *Practical Reasoning* (Oxford: Basil Blackwell, 1974), pp. 54—67.

[9] *Ibid.*, p. 56.

[10] *Ibid.*, p. 66.

[11] *Ibid.*, p. 59.

[12] *Ibid.*, p. 60.

[13] *Ibid.*, p. 61.

[14] *Ibid.*, p. 66.

[15] *Ibid.*

[16] *Ibid.*, p. 67.

[17] *Ibid.*, p. 66.

[18] Compare with Kurt Baier, 'Rationality and Morality,' *Erkenntnis*, 11 (1977), pp. 197—223.

[19] See David Gauthier, 'Moral Artifice,' *Canadian Journal of Philosophy*, 18 (1988), pp. 385—418.

[20] For he voices similar worries in 'Reply to Comments,' in Korner, ed., *Practical Reasoning*, pp. 78—82, and 'Rationality and Morality: A Reply,' *Erkenntnis*, 11 (1977), pp. 225—232.

[21] For similar worries, see David Gauthier, 'Critical Notice of Stephan Korner, ed., *Practical Reasoning* (Oxford: Basil Blackwell 1974),' *Dialogue*, XVI (1977), pp. 510–518, especially p. 514; Patrick Shaw, 'Preference, Choice and Paretian Liberals,' *Philosophy of Social Science*, 16 (1986), pp. 211–218; and Baier, 'Rationality and Morality.'

[22] See his 'Critical Study: Gauthier's Theory of Morals by Agreement,' *The Philosophical Quarterly*, 38 (1988), pp. 343–364; also J. H. Sobel, 'On Maximizers Who Would Co-operate,' (University of Toronto, 1989).

[23] For more on these preferences and their virtues, see my 'Preference's Progress.'

[24] Gauthier argued that *rational* actions *need* not reveal preferences. He sought to rationalize Co-operation by separating individual welfare orderings and revealed preferences; rational agents dispose themselves to optimize welfare, then choose so as to do so with the like-disposed. But I argued (above) that even agents who choose from dispositions it maximizes to adopt would still Defect if they could; their choices would (directly) reveal their preferences.

[25] Still, puzzles remain. E.g., I say that if having different wants would best satisfy one's current ones, one should change them. But are they really under direct rational control? If they are changeable, isn't it completely open what one should want? If one has abandoned one's old preferences as a means of causing their target conditions and if rational agents aim to satisfy their preferences, how will the obtaining of the originally desired target conditions satisfy preferences one no longer has? On the philosophical psychology and the rational kinematics of preferences needed to answer these questions, see my 'Preference Revision and the Paradoxes of Instrumental Rationality,' forthcoming in *Canadian Journal of Philosophy*, and my 'Persons and the Satisfaction of Preferences: Problems in the Rational Kinematics of Values' (Dalhousie University, 1990).

Department of Philosophy
Dalhousie University
Halifax, Nova Scotia
B3H 3J5
Canada