The Southern Journal of Philosophy (1991) Volume XXIX, No. 3

McCLENNEN'S EARLY CO-OPERATIVE SOLUTION TO THE PRISONER'S DILEMMA

Duncan MacIntosh
Dalhousie University

I. Introduction

In the Prisoner's Dilemma (PD), there are two agents, each of whom cares only to minimize his individual jail time. The time each will do depends on how they both next choose among the actions of Co-operating and Defecting. Thus each prefers the following outcomes in the following orders. (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. Numbers in round brackets indicate the utility the agent will receive for each outcome; numbers in squiggly brackets, the number of years in jail): DC {1} (4), CC {2} (3), DD {3} (2), CD {4} (1). If a rational agent chooses so as to maximize his individual expected utility by each choice, since he maximizes whatever the other does if he Defects (i.e., Defecting 'dominates' Co-operating), each will Defect and get 2 utiles.[1] But since if both would Co-operate each would get 3, many philosophers think Co-operation must somehow be rational. If proven, this may show that while to be rational is to choose in ways one thinks will advance satisfaction of one's preferences, that is not necessarily to choose as a maximizer in every choice.[2]

In this paper, I distinguish and review six major attempts to give a Co-operative solution to the PD.[3] I then detail and criticize those of Ned McClennen and David Gauthier (whose solution McClennen tries to augment). I conclude with some

Duncan MacIntosh (Ph.D. Toronto, 1986) is an Associate Professor of Philosophy. He has published on philosophy of science and language, epistemology, meta-ethics and decision theory in Analysis, Dialogue:1, Canadian Philosophical Review, Pacific Philosophical Quarterly, The British Journal for the Philosophy of Science *(forthcoming),* The Southern Journal of Philosophy *and elsewhere.*

observations about what the failure of their solutions shows must be the parameters of any correct Co-operative solution.

I pick on these two attempts because I think each contains one third of the final truth on the matter (the last third remaining to be developed; more on this below). Gauthier has argued that it is rational to unilaterally adopt a conditional disposition to Co-operate (or a conception of rationality favoring compliance with joint optimizing strategies with other optimizers), and thereafter rational to Co-operate provided others share it. (Call this last clause "the proviso.") The dispositions of such agents would supposedly induce them to Co-operate. But I argue that his solution cannot rationalize Co-operating from the disposition, for after one agent has chosen among actions, it is rational for the other agent, even using Gauthier's standards of rationality, to adopt a disposition to Defect, and then to Defect. I think, however, that he is right about the functional properties of a disposition it is rational to unilaterally adopt; it must induce one to Co-operate just with those who would be made to reciprocate by one's having the disposition; it must embed the proviso. But to make it rational to comply with the disposition, we must use a proposal which Ned McClennen seems to have made, if in a somewhat impure form.

McClennen seems to think each PD agent should simply *resolve* to Co-operate with the other, or perhaps simply prefer to do so. Each thereby gives the other reason not to fear Defection, supposedly encouraging him to Co-operate. This is supposed to redound to the benefit of both, each thereafter Co-operating with the other. But I argue, first, that McClennen's proposal has the Defect Gauthier's proposal avoids; McClennen fails to justify the agents unilaterally resolving or preferring to Co-operate because he leaves out the proviso; any agent who makes choices following McClennen's recommendations will be susceptible to exploitation. For if the other agent has not so resolved, he will find it rational to defect against the first agent.

But even incorporating the proviso, there remains a problem. If McClennen means that the agents should simply resolve to Co-operate, it is unclear how resolves can commit rational agents to, and rationalize their performance of, non-maximizing actions. Thus resolving fails to justify Co-operating. But McClennen may really have meant that the agents should so revise their preferences as to find Co-operation maximizing. Since Co-operating would then maximize, it would be rational by the classical standard.

If we then combine Gauthier and McClennen, we have the following solution: rational agents should adopt maximizing dispositions which will induce them to Co-operate with just those similarly disposed, but adopting the disposition must *consist* in adopting revised preferences, ones favoring Co-operating under certain conditions, and able to rationalize it as straightforwardly maximizing on the new preferences. Co-operation would then be rationalized by the new preferences, and we would then have in place the form of a successful Co-operative solution to the PD. (Still, we will not know exactly which preference-function PD agents should adopt, only that it must maximize to adopt it, and maximize to Co-operate from it with similar agents. The details are complicated, and must await further study.)

II. Attempts at Co-operative Solutions

All of the main previous attempts to give a Co-operative solution to the PD have had problems. The *Symmetry* solution argues that since rational agents will choose the same in same situations, and since it is better for each if both Co-operate than if both Defect, each should Co-operate.[4] But while rational agents will choose the same way if there is *one* rational way to choose, and while both agents do better if both Co-operate, neither fact gives either an individual reason to do so; for each still does better individually (whatever the other does) if he Defects. (Their choices remain causally independent, and neither *cares* how the other does.) So *Defecting* still seems rational.[5]

Mechanism solutions argue that it maximizes for each agent to bring in a force which will make him choose non-maximizing actions, i.e., choose against his preferences (if choosing so as to express them consists in maximizing). E.g., each delegates the choice to a machine which will choose Co-operation for him (provided the other also delegated); or each takes a pill (or acquires a socially conditioned reflex) which would make him Co-operate (provided the other made similar arrangements).[6] But while it may be rational to adopt mechanisms, that may not make it rational to Co-operate at their behest in the PD. Indeed, that Co-operative behaviors are selected by the mechanisms, not directly by the agents, may moot whether the resulting behavior is rational, for it is not *action*.[7]

Inducement solutions involve altering the circumstances of choice so that there are advantages to doing something previously dispreferred. New inducements or penalties are added. E.g., each agent allies with a state which (if both ally with it) penalizes Defecting, thus making Co-operating preferable—i.e., maximizing—for each.[8] But that it is rational to make side-bets before facing a PD, does not make it rational to Co-operate while *in* one. For given the side-bet, there is no longer a PD. A PD is defined by the agents' values and choice options giving them a partial conflict: neither can get his best result without the other getting his worst, yet both can get their second-best if they Co-operate. But with the new inducements factored into the pay-off structure, each agent can get his all things considered *best* result, for there are new consequences of choice to consider.

In *Resolution* solutions, one makes a resolution or plan to choose against one's preferences—to Co-operate—and then sticks to it (provided the other also so resolves).[9] But while it may seem rational to resolve to Co-operate so as to make the other Co-operate, one has no reason to keep to the resolve; deviating from it is individually maximizing since one still does best, whether or not the other will Co-operate, if one Defects. Since both know this, each sees that the resolve has no force and that there is no point to making it, perhaps no way of rationally and sincerely making it.

Alternative Principle solutions conceive the principles of rational choice differently: it is not rational always to maximize, but rather, to choose by the disposition it maximizes to adopt to determine further choices—here, the disposition to choose by the strategy which, if both agents choose by it, makes both better off than both choosing by any other, i.e., the jointly optimizing strategy of Co-operating (provided the other adopted a similar disposition).[10] But this solution may show not that it is rational to *Co-operate*, only to *persuade* oneself that it is.[11] Besides, if what motivates revising one's conception of rationality is that it would maximize to have a conception which would allow one to credibly promise to Co-operate, why should one not revise one's conception of rationality yet again after the other has chosen among actions, when one would then find it maximizing to have a conception which would allow one to Defect?

Finally, in *Preference-Revision* solutions one revises those of one's individual preferences responsible for Defecting being dominant, then chooses from the revised preferences (ones which argue Co-operating provided the other adopted preferences arguing same).[12] But this solution may be incoherent: if one prefers outcomes with minimal jail time, and if Defection will minimize it whatever the other chooses, how can one rationally prefer to Co-operate?[13] (I just gave each solution in what I take to be its ideal form. Each tells agents to do something or to think a certain way before making the final choice; but neither could be correct without the proviso clauses in the parentheses. To save oneself from exploitation, one should do or think something which will make one Co-operate only if the other does or thinks it too. Many of these proposals were first published without provisos. We will see the significance of this shortly.)

The *Mechanism* and *Inducement* solutions are methodologically disappointing. They only show that it is rational to take steps to avoid Defecting, not that it is rational to directly choose Co-operating in the original scenario. For exogenous factors had to be introduced, whether other values (ones not derived from those agents have going in to the PD) and other consequences of choice (ones not found in the original scenario) as in the *Inducement* solution, or non-rational forces guiding behavior, as in the *Mechanism* solution. The philosophical interest in the other four solutions is that they claim something surprising about rationality: it is not always rational to choose by maximizing on the preferences one initially brings to the game. In these two solutions, however, rationality proceeds as usual, but in situations so modified as either not to be PDs—as in the *Inducement* solution—or as to make the agents' behaviors non-actions and so non-rational—as, in the *Mechanism* solution, where they Co-operate non-voluntarily.

David Gauthier has argued for either the *Mechanism* or *Alternative Principle* solutions (depending on how we read him); Ned McClennen finds problems with Gauthier's solution, but attempts to defend its conclusion that Co-operation is rational by arguing for either the *Resolution* or *Preference-Revision* solution (again, depending on how we read him).

III. Gauthier[14]

The PD seems to have a natural way out. The agents should make a mutually beneficial agreement to mutual Co-operation and then act on it. 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 choose rationally 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 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 similarly disposed, but with everybody. So you may Defect against them, to your advantage.[15]

But this PD is really two problems. First, suppose some agents will Co-operate if one gives them a credible guarantee that one will reciprocate: Is it rational to give it? Second, is it rational to Co-operate if one gave it? Call giving the guarantee (i.e., acquiring the conditional Co-operator's disposition), "Intending"; call Co-operating after having given it (i.e., after having acquired the disposition, 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 one problem in whether it is rational to Intend, another in whether it is rational to Act. The former is a problem in the rationality of intentions: is it rational to have a maximizing intention to do a non-maximizing action?[16] The latter is the compliance problem: is it rational 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 is either an innovation to classical rationality, or it 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" (so called because their dispositions sometimes constrain them from doing individually maximizing actions) can do both.[17]

IV. Problems with Gauthier's Solution

But can they rationally do both? Mark Vorobej, myself and others think not.[18] Our 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 conditionally Co-operative disposition, one to Co-operate with agents inclined to Co-operate with those so disposed; for when such 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 disposition to Co-operate. It now maximizes to adopt one that would make one Defect (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 disposition "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 one genuinely inclined to Co-operate with those so disposed, making him Co-operate, to the first agent's advantage. Thus, Gauthier has an argument for (i) if the disposition 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 conferred an advantage). 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.[19]

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 it maximizes to adopt an intention not to Co-operate after the other chooses his action. 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 itself irrational. Since it would be irrational to Co-operate then, one cannot rationally intend to do so earlier.

V. Lessons

What went wrong? Gauthier thought that classical maximizers could not find Co-operation rational given their preferences. But the aim of all rational choice is maximization, and 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 concluded that 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 Co-operative disposition it will not be a maximizing one. Thus we might conclude that his theory of rationality must also be false.

But is either theory really false? Maybe not. Both accounts stumbled because of the agent's preferences. In both cases it was because Co-operating was ultimately non-maximizing that it proved irrational, and this seemed to prove the falsity of the two theories of rationality. 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 PD preferences, ones which make rationaly 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 has seemed to some philosophers (e.g., Ned McClennen, on one reading) 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.

VI. McClennen

McClennen defends Gauthier's conclusions about the rationality of Co-operation with the idea that rational agents should adapt their preferences to situations.[20] He thinks classical decision theory takes rational choices in a PD to be ones which maximize on the preferences each agent brings to the game.[21] The theory has only seen Co-operation as rational on two conditions. First, if both agents prefer to Co-operate (making the game's payoff structure different) both will do so rationally. Second, if they believe their choices are not independent, so that whatever one chooses, so will the other, both will rationally Co-operate, since that is better for each than mutual Defection.[22] But he thinks classically rational agents can also be "sophisticated choosers." These will delegate their choices to some automated process guaranteed to Co-operate for them; or they may somehow bind themselves, pre-commit to Co-operating.[23]

He claims, however, that these cost each agent something in utility, each getting only a payoff in between that of their second and third-best outcomes.[24] It would be cheaper for them each to simply pre-*resolve* to Co-operate, then act on that resolve (i.e., comply). Each will then get the utility of the mutually Co-operative outcome.[25] But he notes that on the classical theory, agents cannot rationally comply; for we assume they have preferences from which dominance argues Defection.[26] If they *do* comply, since rational choice must reveal preferences, they must really prefer to Co-operate, violating the assumption.[27] Now, McClennen thinks the theory allows that the agents want to Co-operate going in, for that would get them their second rather than third-best outcome, but that when they choose, each must Defect to protect himself from exploitation by the other.[28] So they are doomed to their third-best payoff.

But he thinks the theory wrongly supposes it rational to maximize on fixed, exogenously specified preferences originally brought to the game. In fact, one should maximize on ones selected in response to the predicament's logic: PD agents should resolve to Co-operate; they will then prefer and choose to Co-operate.[29]

VII. Problems with McClennen's Solution

McClennen seems to think that, in the classical analysis, each PD agent wants to Co-operate, but must Defect to Defend against the chance that the other will Defect.[30] But it is false that each initially wants to Co-operate; rather, each only wants to minimize his jail time. And McClennen surely knows this, for he admits that to make his point, he must use a variant, sequential choice game where the agents have different preferences.[31] Here, agent A prefers: DC, CC, DD, CD; agent B: CC, DC, DD, CD. If B goes first B would love to Co-operate if only A would pre-commit to Co-operating. But A does not want to Co-operate, and A's preferences Define the standard PD: each agent ranks outcomes by how little jail time he gets in them. Each gets less by Defecting, whatever the other does. So each prefers to Defect. At no point does either prefer to Co-operate. Each may wish that both would Co-operate. But that is different from an individual preference for any outcome where he Co-operates, and from an individual preference *to* Co-operate. True, in the variant game, B prefers outcome CC. But even *that* is not the same as preferring an individual Co-operative *choice*. He prefers that *both* players Co-operate, not that he himself does *no matter what the other* chooses.

Still, even standard PD agents see the advantage in efficacious arrangements for *mutual* Co-operation; both prefer CC to DD. So what of McClennen's claim that given this, each rationally *ought* to prefer to Co-operate? This is false. For if I unilaterally so preferred, you could exploit me. I will Co-operate (because I want to, wrongly thinking I have good reason to so prefer), and you know I will (because you know I am rational and so will do what I want), and so you will Defect (for your best payoff). Of course, if I *start out* preferring always to Co-operate, and so caring more about that than about my jail time, I should Co-operate. If you preferred minimum jail time, you should Defect. But if *these* are our preferences, we have no dilemma. Our maximizing choices optimize. But I am only justified in revising my preferences if that will *serve* them. And in the standard PD, my preferring to Co-operate regardless of your preferences will not do this; it will only get me exploited.

But in McClennen's *alternate* game, B will Co-operate if A prefers to Co-operate. A can make B Co-operate by coming to want to do so himself, so it *would* be rational for him to so want. B could then safely Co-operate, sure of CC. For he can expect that A, being rational, will do what he (now) wants, namely Co-operate. But what of the standard PD? Well, availing ourselves of Gauthier's wisdom, while it is not individually rational for either agent to prefer to Co-operate *simpliciter*, it may yet be rational for each to individually prefer to Co-operate with just those whom this would make Co-operate. McClennen should have required each agent to prefer conditional Co-operation.[32]

But we have still another puzzle in McClennen: If both agents begin wanting to Co-operate, as he says, why must they *resolve* to do so? Likewise, in the variant game, if it is rational for A to prefer to Co-operate so as to make B Co-operate, why must A also *resolve* to Co-operate? If Co-operating is our first choice we do not *need* resolutions, for each of us would find it *maximizing*.[33] Indeed, what, in general, is the relation between resolves and preferences for McClennen? Do I resolve to Co-operate because I prefer to Co-operate? Then why must I resolve? What could tempt me to Defect that I must resolve not to? Or do I prefer to Co-operate because I resolve to? Then how does resolving make me so prefer? Is it because I *begin* with a preference to keep to resolutions? But where does one find *that* in the original PD preferences? Or does resolving *cause* one to prefer to Co-operate? But why believe that? And even if true, how does that show it is *rational* to so prefer? Or is so resolving the coming to prefer to Co-operate? But then it does not give a reason to prefer Co-operation; it is just the acquiring of that preference. And what justifies that?

Finally, what happened to the preference to minimize jail time? How do one's preferences remain well-ordered? The preference for minimum jail time recommends preferring to Defect (for it maximizes whatever the other does), so preferring to Co-operate would cause value conflict.[34]

So there are four infelicities in McClennen. First, he thinks PD agents each want to Co-operate. But they only want to minimize their individual jail times. Second, he thinks each should want to Co-operate. But if either does, the other may Defect with impunity. Third, he thinks each should resolve to Co-operate and that this will involve them coming to prefer to Co-operate, or that it will rationalize their Co-operating, or both. But if either efficaciously resolves or prefers to Co-operate, the other can exploit him. Finally, if dominance reasoning gives one a preference to Defect, but McClennen's, one to Co-operate, how can one hold both preferences? And given the first one, how can one acquire the second? The first mistake is just a misreading of the PD; we can ignore it. The second yields a valuable lesson: though it is irrational to individually prefer to Co-operate *simpliciter*, it is rational to individually prefer to Co-operate *with those who would Co-operate if one so preferred*. The third involves either a redundancy or a mis-placed optimism about the power of resolves to make non-maximizing behavior rational. If the agents have altered their preferences, they do not need resolutions in order to Co-operate; they need only maximize on their new preferences to Co-operate with each other provided both so prefer. If they will not alter their preferences, resolutions are no help; neither has a reason in his preferences to keep to them.[35]

There are two further problems with McClennen's claims about the character of the preference-revision solution. First, he notes that, standardly, a rational agent, "on each occasion calling for a decision, chooses so as to maximize with respect to an antecedently and exogenously specified preference function [hereafter, an ESPF] . . . ."[36] But he argues: "what it is rational for a player to choose . . . is in part a function of a potential in the structure of the game itself for achieving optimality if only there is coordination and not just a function of what would maximize some [ESPF]." He continues: "the very preferences a rational agent has in such a situation need to be understood as shaped by the [situation's] logic . . . . [C]ooperation can be understood as arising from the logic of the interaction situation itself."[37] So to be rational is not to maximize on an ESPF. But presumably he thinks individually preferring to Co-operate is (somehow) justified as helping minimize jail time, the ESPF for PD agents. Thus surely one chooses from the original ESPF in choosing new preferences. (He may just mean that thereafter, one acts on the new preferences, ones not given in the ESPF. Fine. But we could not have gotten them without the originals.) And how can the situation's "logic" be understood apart from the ESPF of preferring minimum jail time? Unless the agents initially so prefer, there is no dilemma, only two criminals in jail. The circumstances do not by themselves dictate preference-revision. Rather, one must begin with certain preferences, and then, given the circumstances, one's preferences motivate their own revision as a way of maximizing on them.

The second problem: He says "the resolute chooser . . . [maximizes] utility at the level of the particular choice [not at the level of dispositions to choose, as in Gauthier], but . . . this utility is contextually dependent on the nature of the interaction situation."[38] But I think there must be some maximization other than at the level of choice of action, namely, at the level of choice of preferences for action with certain kinds of agents, given one's PD preferences. But the *principles of rationality* justifying choice of preferences and of action do not compete: always maximize. (Compare Gauthier: one rationally maximizes in choosing dispositions, but rationally *constrains* oneself from maximizing in complying with them.)

So, two more lessons: First, the agents' preferences to minimize jail time define the choice problem. If they will solve it by choosing new preferences, they will rationalize this by its advancing their original ones. Second, they are then making a second-order choice, are choosing how to choose by choosing preferences about how to choose given what they prefer. Since they can best advance their PD preferences by individually acquiring conditional Co-operator's preferences, they should do so. So they maximize at the level of preferences in choosing among *them*, and then again at the level of choice of actions in Co-operating with those with appropriate preferences.

VIII. Summary and Prolegomenon

McClennen suggested (with the benefit of a bit of selective reading)[39] that one should so revise one's preferences as to find Co-operation maximizing. One would not then face a compliance problem. But while it would be rational to Co-operate were it rational to prefer to Co-operate, he did not manage to specify Co-operation-rationalizing preferences it would be individually maximizing and so rational to adopt. The ones he suggested would leave the agent vulnerable to exploitation. Nor did he explain how to keep one's preferences well-ordered; if one prefers minimal jail time, how can one simultaneously prefer to Co-operate? Meanwhile, Gauthier invented a Co-operation-causing disposition it individually maximized to adopt. For it would induce others to Co-operate with you, yet would not cause you to Co-operate unless it would cause the other agent to reciprocate, thus protecting you from exploitation. But he failed to prove the rationality of Co-operating from it.

The truth, I think, is hybrid, and any successful attempt to Co-operatively solve the PD will have to fit into the following mold: From Gauthier, we learned that it would individually maximize for one to adopt a disposition which would make one Co-operate with just those like-disposed, for so adopting would make suitable others Co-operate, to one's advantage, while protecting one from exploitation by those without a suitable disposition to reciprocate, and while yet allowing one to exploit those unselectively Co-operative. But from our critique of Gauthier and from McClennen we learned that for the disposition to rationalize Co-operation, it must be a revised set of individual preferences, ones which, like Gauthier's dispositions, it maximizes to adopt, but which, unlike his dispositions, make Co-operation rational because maximizing with those so-preferring. McClennen's move solves the compliance problem, since such agents would find Co-operation maximizing. Gauthier's move solves the intention problem (given McClennen's solution to the compliance problem), since it maximizes to so prefer.

But we still do not know exactly *which* preferences PD agents should adopt. We only know it must maximize to adopt them given one's PD preferences (else the agent has no reason to adopt them in his original preferences), maximize on them for one to Co-operate with those with like preferences (else the agent will not find it rational to comply with Co-operative agreements). We also know that it must involve a complete reordering of one's preferences over outcomes, else one's old preference for minimum jail time (which recommends Defecting) will conflict with one's new preference to Co-operate. But we do not know what to prefer, nor in what order. Until we have an answer to this question, there remains no fully satisfactory Co-operative solution to the PD.[40]

Now I said the *Mechanism* and *Inducement* solutions were methodologically disappointing. They so changed the circumstances of choice, the values from which agents first choose, or even whether they really *choose*, that they did not solve the PD as such and showed nothing new about rationality. But won't my form of "solution" (assuming it can be completed) be equally bogus, since, by proposing that the agents have different preferences, it in effect changes the game? In a sense. If the PD is a situation in which someone with standard PD values is arrested, briefed in isolation from his partner, and then chooses by himself, everyone believes[41] that he rationally must maximize and Defect. If his preferences made him Co-operate as a maximizer here he would not be in a PD proper.

But Gauthier retells the story: you have PD preferences, you have been briefed, you are about to confer with your partner; you must then choose an action. The question is now, "Is it rational to so alter yourself now that, if the other is a certain kind of agent, you would Co-operate with him after conferring?" Gauthier says "yes": it is rational to acquire a disposition to Co-operate with the like-disposed. Those finding such dispositions in each other while conferring would Co-operate after. I argued that this would be irrational unless the agents have changed their preferences to find Co-operation maximizing. Thus I give the preference-revision proposal as a solution to the problem, "is there a rational way to revise one's nature in a sequential PD with a pre-interactive opportunity for character amendment that would make it rational to Co-operate?" I think this genuinely solves the problem of whether it is rational to Co-operate in *this* PD. For *it* has two parts, the first posing the question whether one should self-revise, the second, whether one should act on the revision. One's original values justify acquiring new ones; *they* justify Co-operating. Unlike in the *Inducement* solution, we introduce no new exogenous values; all are either initially given in the original PD, or are derived from them as a rational and individually available response to it. And we introduce no new circumstances; there are no new enemies to fear. Nor do we introduce mechanistic determinants which would make the "choice" a non-action and so non-rational. So this is a genuine solution to a sequential PD.

But the result could be put in these terms: If you and your partner are *about to face* a standard (pre-Gauthier) PD—are about to be arrested, separated, and forced to make a choice—rationally, you should each change yourself now and verify the other's having changed so that you will not *get into* a PD; you should pre-arrange to Co-operate. Thus we do not Co-operate in "solving" a PD, but in *avoiding* one. But note the differences between *this* way of avoiding a PD, and those recommended in the *Inducement* and *Mechanism* "solutions." It is individually available to the agents given their current preferences; we do not alter the circumstances of choice by introducing new threats or exogenous values; nor do we introduce mechanisms that would make compliance a non-choice. We use only the individual rationality of those about to face the standard scenario.

This changes the subject from whether it is rational to Co-operate in the standard PD. But the new (or the new question about the old) PD reveals new things about rationality and the rationality of Co-operation. E.g., agents *can* rationally make and keep commitments to do initially non-maximizing actions, for it is rational for them to acquire preferences that make so acting maximizing. It also raises new questions for the study of the paradoxes of instrumental rationality. For as McClennen notes, decision and game theory have only concerned themselves with rational choices given current well-ordered preferences. Little has been done on the rational choosability of preferences, nor on its implications for the rationality of actions. It seems inevitable then, that our views on the correct rational solutions to a variety of decision problems will have to change to reflect the fact that it appears, sometimes, to be instrumentally rational to revise the preferences which have hitherto been thought to form the basis of one's rational choices, and then to make one's final choices on the basis of one's revised preferences.[42]

NOTES

[1] For standard expositions of the PD see Richmond Campbell, "Background for the Uninitiated," in Richmond Campbell and Lanning Sowden, eds., *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem* (Vancouver: The University of British Columbia Press 1985), pp. 3-41, and David Gauthier, "Morality and Advantage," *The Philosophical Review*, 76 (1967), pp. 460-75.

[2] E.g., see Gauthier, "Morality and Advantage," and David Gauthier, *Morals By Agreement* (Oxford: Clarendon Press 1986), Chs. V, VI.

[3] Authors often mix and match, not always clearly.

[4] Lawrence Davis, "Prisoners, Paradox, and Rationality," in Campbell and Sowden, eds., *Paradoxes*, pp. 45-59.

[5] See Campbell, "Background," p. 15.

[6] Some of David Gauthier's writings seem to suggest this, though he disdained it in correspondence.

[7] I argue this in "Two Gauthiers?," *Dialogue*, XXVIII (1989), pp. 43-61.

[8] Thomas Hobbes, *Leviathan*, C. B. MacPherson, ed., (Penguin Books: Harmondsworth, Middlesex, England 1968), pp. 228-239, 251-257.

[9] Edward F. McClennen, "Prisoner's Dilemma and Resolute Choice," in Campbell and Sowden, eds., *Paradoxes*, pp. 94-104.

[10] One can so read Gauthier, *Morals By Agreement*, confirmed in David Gauthier, "In the Neighborhood of the Newcomb-Predictor (Reflections on Rationality)," (1985), "Economic Man and the Rational Reasoner," (1987), and in correspondence. Peter Danielson may defend this view in his *Artificial Morality: How Morality is Rational* (draft 0.4, 1988).

[11] See Howard Sobel's critique—"Maximizing, Optimizing, and Prospering," *Dialogue*, XXVII (1988), pp. 233-262, part two—of Gauthier's arguments for an alternative conception of rationality.

[12] Such a view may be in Amartya Sen, "Choice, Orderings and Morality," in Stephan Körner, ed., *Practical Reasoning*, (Oxford: Basil Blackwell 1974), pp. 54-67, in Edward F. McClennen, "Prisoner's Dilemma," and "Constrained Maximization and Resolute Choice," *Social Philosophy and Policy*, 5 (1988), pp. 95-118. I identify the problem to which this is a solution in my "Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem," *The Southern Journal of Philosophy*, XXVI (1988), pp. 399-425, propose it as a defense of Gauthier in "Two Gauthiers?," and defend the intelligibility of revising one's basic preferences in "Persons and the Satisfaction of Preferences: Problems in the Rational Kinematics of Values" (Dalhousie University, 1991), and "Preference-Revision and the Paradoxes of Instrumental Rationality," conditionally forthcoming, *The Canadian Journal of Philosophy*. The latter, tracing from my "Retaliation Rationalized: Gauthier's Solution to the Deterrence Dilemma," *Pacific Philosophical Quarterly*, Vl. 72, No. 1, March (1991), pp. 9-32, and my "Kavka Revisited: Some Paradoxes of Deterence Dissolved," (Dalhousie University, 1990), asks, can a harm-hater rationally threaten nuclear retaliation to deter attack?; can he rationally retaliate if deterrence fails? The Deterrence Dilemma (DD) is congruent with the PD, since both involve the rationality of intending and performing non-maximizing actions it

MacIntosh, Duncan, *McClennen's Early Cooperative Solution to the Prisoner's Dilemma* ,
Southern Journal of Philosophy, 29:3 (1991:Fall) p.341

maximizes to intend. E.g., see David Gauthier, "Deterrence, Maximization, and Rationality," *Ethics*, 94 (1984), pp. 474-495, and "Afterthoughts," in Douglas Maclean, ed., *The Security Gamble: Deterrence Dilemmas in the Nuclear Age* (Totowa, NJ: Rowan and Allenheld 1984), pp. 159-161; also Gregory Kavka, "Some Paradoxes of Deterrence," *The Journal of Philosophy*, 75 (1978), pp. 285-302, "The Toxin Puzzle," *Analysis*, 43 (1983), pp. 33-36, "Responses to the Paradox of Deterrence," in Maclean, ed., *The Security Gamble*, pp. 155-159; also David Lewis, "Devil's Bargains and the Real World," in Maclean, ed., *The Security Gamble*, pp. 141-154, and Mark Vorobej, "Gauthier on Deterrence," *Dialogue*, XXV (1986), pp. 471-476.

[13] David Gauthier, "Critical Notice of Stephan Körner, ed., *Practical Reasoning* (Oxford: Basil Blackwell 1974)," *Dialogue*, XVI (1977), pp. 510-518.

[14] I have borrowed much of the summation and critique of Gauthier in this and the next section from my "Co-operative Solutions to the Prisoner's Dilemma," *Philosophical Studies* (conditionally forthcoming).

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

[16] See Kavka, "Some Paradoxes."

[17] Gauthier, "Morality and Advantage," and "Deterrence."

[18] Vorobej makes the objection re the DD in his "Gauthier." I argue for it more fully both re the DD and the PD in my articles cited in note 12, above. Richmond Campbell similarly objects in, "Moral Justification and Freedom," *The Journal of Philosophy*, LXXXV (1988), pp. 192-213.

[19] I cannot answer possible replies here; for a more complete discussion, see my papers in note 12, above.

[20] McClennen, "Prisoner's Dilemma."

[21] *Ibid.*, p. 94.

[22] *Ibid.*, pp. 94-95.

[23] *Ibid.*, pp. 98, 100-101.

[24] *Ibid.*, pp. 98-103. It is unclear why these options must be costly since the agents only care about minimizing their individual jail times; they do not value not delegating or not precomitting. Is the idea that they would have to do a little time for someone else to pay the delegatee, or to finance the precommittment process? But this is a mere contingency. For they might instead be able to take a Parfit pill, itself costless by their values, that would simply induce them to Co-operate. Maybe McClennen is thinking of the lost opportunity for further minimization of individual jail time that is the price of truly assuring mutual Co-operation; each agent would prefer to be Defecting when their assurance arrangements induce Co-operation. But this misrepresents the payoff. Them sacrificing this opportunity gets them their second-best result, not one between their second and third.

[25] McClennen, "Prisoner's Dilemma," pp. 98-103.

[26] *Ibid.*, p. 102.

[27] *Ibid.*, pp. 102-103.

[28] *Ibid.*, p. 102.

[29] *Ibid.*, pp. 94-95, 102-103.

[30] *Ibid.*, p. 102.

[31] *Ibid.*, pp. 96-98.

[32] There are problems with specifying preferences such that, if each agent acquires them, both will be moved to Co-operate. But I must leave this for another paper. See my "Co-operative Solutions."

[33] Compare with Amartya Sen's "OR game" in his article, "Choice." There, since our preferences for outcomes are to have Co-operated if the other Co-operates or Defects, our preference function amounts to a preference *to* Co-operate, since our preferences can be satisfied by our Co-

operating independently of what the other does. For more on this see my "Co-operative Solutions."

[34] This problem must wait. See my "Co-operative Solutions."

[35] There is not space to consider the fourth problem here.

[36] McClennen, "Prisoner's Dilemma," pp. 94.

[37] *Ibid.*, pp. 95-96.

[38] *Ibid.*, p. 95.

[39] McClennen comes closer in "Constrained Maximization and Resolute Choice," and *Rationality and Dynamic Choice* (Cambridge: Cambridge University Press 1990). Gauthier thinks agents need not revise their preferences for compliance with a maximizing agreement to be rational. See his "Morality, Rational Choice, and Semantic Representation: A Reply to My Critics," *Social Philosophy and Policy*, 5 (1988), pp. 173-221, the reply to McClennen, "Constrained Maximization." But I think this is wrong, as I argued, above.

[40] For a critical review of theories of which preferences agents would have to have to find Co-operation rational in a PD, see my "Co-operative Solutions," in which I also take a shot at specifying such preferences. I there take as my point of departure Amartya Sen's proposals in "Choice." Sen appears to think it rational for each agent either to act as if he had different preferences than those he has in the PD, or to acquire higher-order preferences prioritizing the orders in which his PD preferences for outcomes are to be satisfied. With such counter-preferential inclinations or prioritizing preferences, Sen thinks each could then rationally comply. But I argue that Sen gives no rationale for unilateral preference simulation, nor for unilateral reorderings of the priorities attaching to satisfaction of one's individual outcome preferences, nor for action on those orderings. (For he fails to incorporate the above proviso.) Moreover, prioritizing attainment of outcomes in some order different from that in which one prefers them may result in a conflicted, and so ill-ordered and irrational preference function (mapping from outcomes to utilities). I then try to specify a set of preferences which it is maximizing to adopt (and which leaves the agent free from the danger of exploitation), maximizing to Co-operate from, and which is internally coherently ordered.

[41] Everyone but friends of *Symmetry* "solutions."

[42] For helpful comments, my thanks to Robert Bright, Richmond Campbell, Julia Colterjohn, Peter Danielson, Ish Haji, Robert Martin, Victoria McGeer, and Terrence Tomkow. For financial support during early work on this project, I thank the Social Sciences and Humanities Research Council of Canada for a Doctoral Fellowship and Dalhousie University for a Killam Post-Doctoral Fellowship.