The Southern Journal of Philosophy (1988) Vol. XXVI, No. 4

LIBERTARIAN AGENCY AND RATIONAL MORALITY: ACTION-THEORETIC OBJECTIONS TO GAUTHIER'S DISPOSITIONAL SOLUTION OF THE COMPLIANCE PROBLEM[1]

Duncan MacIntosh
Dalhousie University

Introduction

Moral rationalists think perfectly informed, free, and rational agents will always act morally.[2] David Gauthier claims, for example, that in the Prisoner's Dilemma (or PD), co-operation is rational iff both parties are known by each other to be disposed to co-operate with just those similarly disposed.[3] He thinks the dispositions give each agent reason to think that co-operative acts of his (but not defections) will be met with co-operative acts from others. This, Gauthier thinks, ensures "those conditions under which individuals may rationally expect the degree of compliance from their fellows needed to elicit their own voluntary compliance."[4] Moreover, since one's expected utility is higher if one has such a disposition, it is rational to acquire it.[5] Richmond Campbell generally agrees. But he thinks Gauthier's conclusions only hold if people's dispositions to co-operate are sufficiently deterministic and enduring as to make them resist the temptations of dominance reasoning, which argues defection. Otherwise, one cannot guarantee reciprocation of one's co-operative actions, making them irrational. The Libertarian view that people lack dispositions altogether, or can alter or step out of such co-operative dispositions as they may have and defect, must thus be false if co-operation is truly rational in PDs.[6]

Now Gauthier says,

> [a] constrained maximizer is conditionally disposed to co-operate in ways that, followed by all, would yield nearly optimal and fair outcomes, and *does* co-operate in such ways when she may actually expect to benefit.[7]

But an immediate problem with Gauthier's argument is that he does not make clear *why* someone with a disposition to co-operate may be expected to co-operate with someone else with a similar disposition.

Duncan MacIntosh is an assistant professor of philosophy at Dalhousie University in Halifax, Nova Scotia, Canada. He received his Ph.D. from the University of Toronto in 1986. He is interested in the philosophies of language, science, and mind, epistemology, metaphysics, and meta-ethics, and his recent writings are on the relationship between morality and rationality.

MacIntosh, Duncan, ***Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem*** , Southern Journal of Philosophy, 26:4 (1988:Winter) p.499

Standard dominance reasoning in the PD shows that defecting has the higher individual expected utility. It is thus unclear why, if, say, I have such a disposition and see that you have one, I will co-operate with you, and why, if you have such a disposition and see that I do too, you will co-operate with me. The fact is that my preferences would be best expected to be served by me defecting, similarly for you. And rational agents as understood both by Gauthier and in the tradition, will, if free to act, act so as to maximize expected satisfaction of their preferences. If we are rational then, we should each defect, not co-operate. So how can the fact that we each have such a disposition make a difference to how we will each behave, and to how we may be reasonably predicted to behave?

Gauthier seems to think that while normally, agents may be predicted to behave in whatever way they would expect to maximize satisfaction of their preferences, his agents, once having made a second-order choice among dispositions, may be predicted to act from their dispositions, not from their preferences. *Prima facie* though, this means that they do not act rationally in the sense of maximizing their individual expected utility. How then can he maintain that co-operation would be rational for them?

In preserving their rationality, he seems to me to have the following choices:

a) Perhaps he could argue that to acquire a co-operative disposition is to acquire a preference to co-operate with anyone who has acquired a preference to co-operate. Thereafter, co-operation with those with a similar disposition could be predicted because, given a sufficiently strong co-operative preference, co-operation would be straightforward maximizing behavior. It would be predictable on the same rationale as was defection before preference revision. Each agent would *want* to co-operate, provided each knew the other so wanted, and provided both knew that each other, being rational and free, would do what they wanted. But he explicitly disavows this:

> Duty *over-rides* advantage[8]; we shall recognize the need for *restraining* each person's pursuit of her own utility[9]; the rational principles for making choices *constrain* the actor pursuing his own interest in an impartial way.[10]; we shall . . . undermine the force of the demand that rational choice reveal preference by showing that its scope may be restricted by . . . a meta-choice, a choice about how to make choices.[11]; [A] rational utility maximizer . . . chooses, on utility maximizing grounds, *not to make further choices on those grounds.*[12]; These principles require a person to *refrain* from the direct pursuit of her maximum utility[13]; [A] constrained maximizer may find herself required to act in such a way that she would have been better off had she *not* entered into co-operation.[14]; constrained maximization is *not* straightforward maximization in its most effective disguise.[15]

Apparently a rational agent does not revise his preferences, thus changing his interests, and so changing what he may be predicted to do so far as he is assumed to rationally advance his interests. Rather, he behaves against the dictates of his preferences in co-operating.

MacIntosh, Duncan, ***Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem*** , Southern Journal of Philosophy, 26:4 (1988:Winter) p.499

Presumably then, in acquiring a disposition to co-operate, those of his preferences which would make him, in turn, instrumentally prefer not to co-operate in the circumstances, remain unchanged. If they *had* changed, his co-operating would be maximizing of expected utility and so would issue without need of a constraint on the expression of his preferences.

b) Perhaps Gauthier construes the disposition as something with the psychological power to over-ride the agent's preferences, effectively forcing him to act co-operatively when (but only when) he detects a similar disposition in someone else. It forces him to co-operate against his will, at least so far as his will is defined as the normal issue of his considered preferences. He may now be predicted to co-operate because the disposition is predicted to force him to co-operate, whatever his preferences. This, indeed, appears to be how Campbell construes these dispositions, and Campbell uses this interpretation in his explanation of how dispositions can afford guarantees of mutual co-operation. This interpretation is, again, supported by Gauthier's words:

> . . . we do not purport to give a utility-maximizing justification for specific choices of adherence to a joint [co-operative] strategy. Rather we explain those choices by a general disposition to choose fair, optimizing actions whenever possible, and this tendency is then given a utility-maximizing justification.[16]

It appears, then, that Gauthier thinks the disposition does not rationalize co-operation given extant preferences, but that it will determine a person's behavior in such a way that we can, nonetheless, cite its functioning in explaining why he co-operates. Presumably the agent co-operates because the disposition *causes* him to do so whether he then *wants* to co-operate or not.

But can this count as rational, free, voluntary action, action chosen and performed at the time of co-operation? One co-operates against one's will, and co-operation is no longer chosen on the occasion of performance, but is brought about independently of the preferences which are classically thought, in directly causing a behavior, to make it a chosen action. Does Gauthier think, then, that co-operation is unfree, involuntary, unchosen—in short, not action? No:

> [internal moral constraints operate to ensure] . . . those conditions under which individuals may rationally expect the degree of compliance from their fellows needed to elicit their own *voluntary* compliance.[17]

How, then, can Gauthier count co-operation under these conditions as rational, voluntary, free, and as chosen and performed at the time of co-operation?

c) Perhaps Gauthier really conceives rational action as something other than *direct* maximization of individual preferences, and instead as behavior from a *disposition* which it maximizes preferences to adopt. Or perhaps he conceives it as direct maximization generally, but thinks

that in this case, one ought to rationally revise one's conception of rational action to incorporate the advantage of having a preference-overriding, deterministic disposition. But again, he seems to explicitly disavow this too:

> [rational] choice maximizes preference fulfillment given belief.[18]; rational choice must be directed to the maximal fulfillment of our present considered preferences.[19] Utility is . . . the measure of present preference[s] . . . of the self at a particular time; practical rationality is the maximization of utility and so the maximization of the satisfaction of present preferences.[20]

It appears then that Gauthier retains the conception of rational action as behavior chosen for its maximization of the satisfaction of individual preferences—as action which maximizes expected utility. But *that* conception of rationality, again, should dictate defection, and it remains unclear how co-operation can count as rational, voluntary, and free.

d) Perhaps Gauthier thinks that behavior from a rationally adopted co-operative disposition can still count as free, rational, voluntary action, action chosen and performed at the time of co-operation, in spite of the fact that one co-operates merely because caused to do so by one's disposition, and in spite of the fact that behaving that way goes against one's preferences. Perhaps co-operation inherits both rationality and its status as freely chosen action from the fact that it issues from a *rationally chosen* disposition. Here is Gauthier again:

> Our argument identifies practical rationality with utility-maximization at the level of dispositions to choose, and carries through the implications of that identification in assessing the rationality of particular choices.[21]

But I think this does not work either, and showing why not will be the burden of this paper. I will here argue on action-theoretic principles that if the dispositions justify expectation of mutual co-operation through such causal powers as they may have to prevent individual utility maximization, the behaviors they cause cannot count as rational actions. Indeed, under some construals, they are not actions at all. Yet if the dispositions do not have such powers it appears that to act rationally one must defect.[22] Thus, neither Gauthier nor Campbell has demonstrated the rationality of co-operative action where the payoff structure for choices (hereafter the POS) remains that of a standard PD.

They apparently think it sufficient for an actor's dispositions to choose being rational that it was rational for him to acquire them. Moreover, they think that if his dispositions are rational, choices issuing from them are rational.[23] Since they could demonstrate the rationality of acquiring a co-operative disposition, they assumed they had a solution demonstrating the rationality of co-operative choice and action.[24] Gauthier assumed this without further ado. Campbell concluded to it after satisfying himself that behaviors and choices

MacIntosh, Duncan, ***Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem*** , Southern Journal of Philosophy, 26:4 (1988:Winter) p.499

issuing from a stable causally determining disposition freely and rationally acquired involved as much freedom as was worth wanting. He could then discount further pleas for contra-causal freedom from the Libertarian.[25] But that a disposition is one it is rational to choose or acquire does not entail that every choice resulting from it is rational, nor that every behavior issuing from it is an action and a rational one. For it is problematic whether genuine choices or actions *can* issue from dispositions which *prevent* immediate utility maximizing action. Further, the conditions making selection or possession of a disposition rational can change (e.g., when one chooses second in a sequential PD), rendering both it and any choices or actions issuing from it irrational in the new situation.[26]

I. An Argument for the Irrationality of Morality

Genuinely moral action is widely acknowledged to have three ingredients. First, the action must be free in a variety of senses: It must be physically unforced. It must not have occurred *only* because caused by such extra-moral inducements as would make even an evil person behave morally. The action must be psychologically non-pathological (i.e., free of non-rational internal compulsion). The agent must be informed about his options and the consequences of his choices. It is hotly debated whether the action must also be metaphysically undetermined and unnecessitated; also whether it must be politically and culturally uninfluenced in certain ways. Second, moral behavior must be *intentional* behavior; not mere behavior, but *action*. Action is self-aware behavior proximately caused by something like a complex of beliefs and desires (to take the standard theory of agency as offered by Donald Davidson), or at any rate, by something which has both cognitive and conative components. A paradigm of morally evaluable action is behavior proximately caused by *reasons* which rationalize or justify the behavior in their own terms. Further, the reasons must be rationally acquired, typically by reflection on antecedent desires plus present beliefs. Moreover, the beliefs and desires must be rationally sustained; beliefs and desires which were rational in one situation do not necessarily count as rational in an entirely different one simply because they once were rational. Rationality is a form of response to perceived circumstances, and to continue to be rational, one must continue to be responsive to believed-to-be-relevant perceived changes in circumstances.[27] Thirdly, in addition to being a free and intentional—ideally, a reasoned or reason*able*—action, a truly moral action must, obviously, be morally correct. Call the first set of requirements the freedom requirements, the second, the agency requirements, and the third, the rectitude requirements.

Moral rationalists think the requirements of morality are a subset of those of instrumental rationality, i.e., that it is instrumentally rational to be moral; also, that to be moral is to be instrumentally rational. In

MacIntosh, Duncan, ***Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem*** , Southern Journal of Philosophy, 26:4 (1988:Winter) p.499

contractarian moral rationalism, the correct moral rules are just those (1) by which individual utility maximizers would agree to be governed, and (2) with which the parties to the agreement would be led to comply by their instrumental rationality alone.[28] Put another way, anyone who counts as an agent by virtue of his instrumental rationality, will, if he is free to act, and after rational reflection, act in compliance with the obligations of morality.

The big internal problem for (2) is nicely modeled in the one-shot PD under opacity.[29] Imagine two rational bank robbers have been captured, and are separately offered the following deal. "If you confess and implicate your colleague, you get partial immunity, he takes the fall. You get one year in jail, he, nine. If you do not confess but he does, he gets one year and you, nine. If you both confess, you both get five years; if neither of you confesses, you both get three." The prisoners are allowed to confer, and are then separated again to make their choices in secret. By co-operating with each other and not confessing, they can reduce their jail time to three years each. But if they do not trust each other and both defect from any agreement they may make to co-operate, they will each get five years. Assume: Rational individuals choose actions which maximize individual expected utility on an act-to-act basis. They choose actions which maximize satisfaction of the coherently ordered, all-things-considered preferences which they have at the time of choice. (Hereafter, I shall call coherently ordered, all-things-considered preferences "sum-preferences" for short, and preferences the agent has at the time of choosing, "concurrent" preferences. Finally, I shall call individual expected utility maximizers, "maximizers.") The POS for their choices gives the highest reward to a defector who meets a co-operator, the second highest to a co-operator who meets a co-operator, the third highest to a defector who meets a defector, and the fourth highest to a co-operator who meets a defector. Neither's actual choice entails or causally influences the other's. Suppose one of these people is you. If you defect and the other co-operates, you get the highest payoff, while if you had co-operated, only the second highest. If you defect and the other defects, you get the third highest, while if you had co-operated, only the worst payoff, the fourth highest. Thus you always do better by defecting, no matter how the other chooses, and so it is always rational to defect. This is the so-called dominance argument.

This is worrisome if the PD is a microcosm for the predicament of moral agents. Morality presumably requires agents to co-operate with each other (e.g., to keep the agreements described in (1) above). Yet individual rational agents apparently have no motivation to co-operate, at least not one guaranteed *simply* by their being rational. Yet if everyone would only co-operate, they would all do even better than if they all defected—mutual co-operation would get them each the second rather than the third best payoff. Thus modeled, we have the problem of

MacIntosh, Duncan, ***Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem*** , Southern Journal of Philosophy, 26:4 (1988:Winter) p.499

compliance: There are principles everyone grants universal compliance with which would make everyone as well or better off than universal compliance with any other principles. But since any given individual can do better by violating those principles, how can we motivate individual rational agents to co-operate, to act morally, i.e., in accordance with those principles?[30]

II. The Gauthier Solution

What is wanting to make free co-operation rational? Is it just that you can not guarantee that others will co-operate? No. If you could, it is even more rational for you to defect, for that nets you the highest payoff. Is it just that others have no guarantee you will co-operate? No. That would give them all the more reason to defect against you, again, being assured of the highest payoff for themselves. What then?

David Gauthier thinks one is wanting a guarantee that others are *disposed* to co-operate with you *iff* you are *disposed* to co-operate with them. We need a reason to trust each other, but not to think each other fools. When would we have such a reason? According to Gauthier, under the following conditions[31]: Assuming the standard PD POS, people must be able, with some reliability—the higher the better—to tell what kind of people they themselves and others are; would they tend to co-operate with co-operators, or would they try to cheat? One only wants to co-operate with those who are likely to reciprocate, and they are only likely to co-operate with you if they think you are likely to reciprocate. Further, assuming they will perceive you relatively accurately, you need to know whether you are the kind of person disposed to co-operate with co-operators and so likely to be co-operated with.[32] Given this mutual psychological transparency, it is rational to co-operate iff you are disposed to co-operate just with the similarly disposed (otherwise, defect), think you recognize someone of that kind in the dilemma with you, and think that person recognizes that you are a similar kind of person. (This is the imperfectly explained step in Gauthier, on which I will accost him shortly.)

If you lack that disposition, it is rational, according to Gauthier, to acquire it.[33] You lose nothing by having it, because it does not result in you co-operating with defectors.[34] Yet you gain the benefits of co-operation denied to people disposed always to defect, since possessing it induces others with it to co-operate.[35] As Gauthier puts it, one gains opportunities for increases in one's utilities denied those without the disposition.[36] If one can choose dispositions to follow strategies, then, one rationally must choose conditional co-operation, sometimes called narrow compliance, constrained maximization, or CM for short. It beats total non-compliance, or straightforward maximization (SM), since it guarantees you your second-best outcome whenever you meet a CM, something the ever-defecting SM never achieves (except if he can pass for a CM; but for simplicity, I shall assume agents are transparent

MacIntosh, Duncan, ***Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem*** , Southern Journal of Philosophy, 26:4 (1988:Winter) p.499

here, and since they cannot be misidentified, deception will not be a consideration).[37] And it beats broad compliance, always co-operating, which will have you victimized by SMs.[38] Whatever non-tuistic preferences (preferences defined independently of the welfare or ilfare of others) he begins with, then, a person able to choose his dispositions will rationally choose the CM disposition.

Gauthier claims, further, that after the acquisition of the CM disposition we may speak of at least the *derivative* rationality, and, indeed, even of the rationality *simpliciter,* of the co-operative actions it induces. Co-operation with people disposed to co-operate would be rational because expressive of an actually held disposition it is rational to have. As he puts it,

> The rationale for disposing oneself to constraint does not appeal to any weakness or imperfection in the reasoning of the actor; indeed, the rationale is most evident for perfect reasoners who cannot be deceived. The disposition to constrained maximization overcomes externalities; it is directed to the core problem arising from the structure of interaction. And the entire point of disposing oneself to constraint is to adhere to it in the face of one's knowledge that one is not choosing the maximizing action.
>
> Imperfect actors find it rational to dispose themselves to make less than rational choices. No lesson can be drawn from this about the dispositions and choices of the perfect actor. If her dispositions to choose are rational, then surely her choices are also rational.[39]
>
> The dispositions of a fully rational actor issue in rational choices. Our argument identifies practical rationality with utility maximization at the level of dispositions to choose and carries through the implications of that identification in assessing the rationality of particular choices.[40]

If people can influence each other's choices by their selection and revelation of dispositions, they should choose to have dispositions to co-operate with just those similarly disposed, and should learn to reveal them to and detect them in others.[41]

III. Campbell's Addenda to Gauthier's Solution

Campbell, while generally sympathetic to Gauthier's position, thinks certain conditions must be added before co-operation becomes rational.[42] He doubts whether it is really rational for an agent to act on his disposition to co-operate with those with a similar disposition. That is, he doubts whether what offers agents guarantees of increased probabilities of being met with co-operation is simply,

> (3) the mutually recognized possession of instrumental dispositions to co-operate with those similarly disposed.

(By an instrumental disposition, I mean one had only because and only insofar as having it is instrumentally needed to increase the expected utility of a choice by altering the probability of another's action. One does not value co-operation or fair outcomes for their own sakes, but has only disposed oneself to behaviors which contribute to them because, and for just as long as, this helps maximize satisfaction of some other preference.)

MacIntosh, Duncan, ***Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem*** , Southern Journal of Philosophy, 26:4 (1988:Winter) p.499

His doubts are best appreciated where one person makes his choice after that of another so that the issue of trust is paramount (though the question also arises for simultaneous secret choosing). Suppose, as per Libertarianism, that people have the freedom to do anything (that they are not physically restrained from doing) at any moment of choice; they are not psychologically bound by any prior disposition. Suppose two parties, First and Second, meet, size each other up, discover each other to be rational and Libertarian-free (hereafter, L-free), and to have CM dispositions.[43] First then chooses in secret. Second must now choose, knowing First has already chosen. What should he do? In a PD, the highest utility accrues to Second from unilateral defection, the second highest from mutual co-operation, etc. Dominance reasoning tells Second he does best no matter how First chose if Second defects, so rationally he must defect if he can. L-free agents always *can* defect. Thus, even if he began with the CM disposition (perhaps hoping to induce First to co-operate), he should quit it, acquire the disposition to defect, and defect. An L-free agent always *can* acquire the defector's disposition, since nothing deterministically prevents him from choosing dispositions on the basis of instrumental rationality. (Gauthier can hardly object to this. If instrumental thinking originally motivated acquisition of the CM disposition, why can it not motivate its abandonment and the acquisition of the SM disposition when faced with a suitable opportunity?) What should First have done? Knowing how Second will find himself after First chooses, First expects Second to defect. He should choose whichever way has the highest expected utility given what Second will choose. Dominance reasoning tells him that no matter how Second chooses, First does better by defecting, *a fortiori* so if Second is expected to defect. So First should acquire the SM disposition, act on it, and defect. Result: mutual defection.[44]

Thus, far from having good reason to think such people will co-operate, we may generally expect them to defect.[45] For nothing in Gauthier's argument makes co-operation on an act-to-act basis suddenly rational. *If* they *could* do otherwise than co-operate in the moment of choice, rationality (as individual utility maximization) requires them to defect. According to Campbell, only the disposition to co-operate could prevent people from doing just this in Gauthier's account—its role is to serve as a *constraint* on individual utility maximization, after all—and that preventative would be, *ex hypothesi*, ineffective in L-agents.[46]

Note: What makes defection rational here is not just the absence of soft-determinism. (I do not think Campbell recognizes this.) Indeed, suppose soft-determinism, i.e., compatibilism, were true: every rational action is also a determined action without threat to our responsibility. Since it is rational for Second to reacquire the SM disposition after First makes his choice, nothing in soft-determinism *per se* would prevent him from doing so. And since it is rational to defect if one has

that disposition, it is rational for Second to defect. (Gauthier, of course, thinks the CM disposition the rational one to have; so from *that* vantage defection is irrational. Nonetheless, he thinks that if you actually have the SM disposition, SM rationality dictates that you defect. I am claiming that while Gauthier may have shown that it is rational to have the CM disposition when being sized up prior to choosing, he has not shown that it continues to be rational to have it. I think instrumental rationality should, indeed, induce its abandonment after one has passed scrutiny. Having abandoned it and reverted to the SM disposition, *defection* is thereafter rationally mandated.) It would appear, then, that it is not reasonable to take as a guarantee of a high probability of co-operation being reciprocated that there is, initially,

(4) the mutually recognized possession of instrumental CM dispositions, plus the assumption of compatibilist determinism.

At any rate, to escape Campbell's original worry, in addition to Gauthier's condition (3), Campbell offers the following conditions. These are needed both to give sense to the idea that agents *actually have* a CM disposition, and to strongly enough guarantee that co-operation will be met with co-operation, defection with defection:

(5) People must know that CM dispositions are strong enough to override any other current dispositions they may have, specifically, the temptation or disposition to cheat.[47]

(6) We must be able to tell (with some accuracy) that people cannot suddenly change their CM dispositions to take advantage of the possibility of cheating; that they are not prone to the sudden acquisition of the SM disposition.[48]

(7) We must know that people do not have the power, in spite of the foregoing, to unpredictably and spontaneously just brutely and freely choose to cheat; that they do not have the contra-causal ability to step outside their dispositions altogether.[49]

Gauthier is somewhat unclear about the nature of the dispositions he has in mind, but Campbell thinks that the guarantees given in conditions (5)-(7) either amount to, or will only be satisfied if, we each have a stable *mechanical disposition* to co-operate with others similarly *mechanically disposed.* To have sufficient guarantee that co-operation will be reciprocated as to make it rational for myself, I must know that you are as a matter of irrevocable psychological necessity disposed to co-operate with me if you know that as a matter of irrevocable psychological necessity I am disposed to co-operate with you.[50] The guarantee need not always be perfect; the rationality of co-operation depends on the POS of choices under risk: as long as the probability of mutual co-operation on the condition of the presence of shared irrevocable CM dispositions times the payoff for mutual co-operation is sufficiently high, rational co-operation may ensue. Thus people's natures may be merely translucent rather than transparent to each other. Their dispositions may be to co-operate only most of the time.

MacIntosh, Duncan, ***Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem*** , Southern Journal of Philosophy, 26:4 (1988:Winter) p.499

There may be some degree of objective indeterminacy in their behavior, so that, while they may be predicted to co-operate some percentage of the time, co-operation is not certainly predictable on any particular occasion.[51] But on balance, Campbell, and perhaps Gauthier, think(s) that with the standard PD POS, co-operation is rational only so far as agents have relatively irrevocable psychological dispositions which (in significant degree) are known (with some certainty) to make them (to some degree) unfree. These dispositions must *determine* them *irresistibly* (to some extent) to co-operate with people similarly disposed. Apparently L-free people cannot rationally choose to co-operate with each other knowing they could freely choose to co-operate *or not*. Thus, what offers guarantees of mutual co-operation must be:

(8) (i) the mutually recognized possession of instrumental CM dispositions, plus (ii) the assumption that the dispositions it was rational to acquire prior to being sized up will endure, where (iii) that assumption is licensed by the mutual belief that the dispositions will endure because of rationality-*in*dependent, causal forces (for as we have seen, it is not instrumentally rational to retain such dispositions).

Now, recall that what was originally thought to be needed to make co-operation rational was a guarantee that co-operation would or would likely be met with co-operation, defection with defection. The mere mutual possession of the CM disposition, even under the weak determinism of compatibilism, does not provide that guarantee. For the sequential PD shows the *ir*rationality of keeping such dispositions once one has them, and thereafter, of acting on them. (At least it does assuming constancy of POS. The CM disposition is then adopted only instrumentally to inducing co-operation in the other. Since a decrease in jail term is the ultimately desired consequence, it remains more desirable to have succeeded in unilateral defection than in mutual co-operation). So far, it would seem that the only way to get *any* such guarantee, is if people have deterministically *irrevocable and efficacious* CM dispositions. However rational it may be to change a CM disposition when given the opportunity to cheat, something must prevent one from doing that.

IV. The Libertarian Objection to the Gauthier/Campbell Solution

Libertarians think one acts in morally responsible, morally evaluable ways only if one's behaviors are finally undetermined (except by one's preferences and choices), so that one is free to do anything whatever that one might prefer or choose to do, physical limitations apart.[52] If people only do something because of causally stabilized deterministic dispositions, they have not chosen to co-operate, and so do not act rationally in co-operating.

Campbell's defense is that, though his argument is not intended to prove determinism or compatibilism, it implies that it is worth wanting same to be true for the resulting benefits of rational co-operation.[53] It

would be irrational to want people to be free to revoke, change, or escape their dispositions at the moment of actual choice, or to lack dispositions altogether. For were they known to have such freedom, wanting a chance at their best outcome, and fearing that each other will try for their own best outcomes, rationally they must defect. They would be stuck again with the lower payoff from mutual defection. Conditional co-operation then, is something rational people would conditionally deterministically dispose themselves towards if they could.[54]

Campbell notes that while the CM disposition makes one co-operate with other CMers, one remains free in several senses. One can do otherwise if circumstances are different, namely, if one meets a SM.[55] Also, as noted above, there may be some objective indeterminacy, randomness, and unpredictablity in one's behavior.[56] Finally, one is free to acquire the disposition or not; it is acquired without coercion and the behaviors it induces are caused by something itself freely chosen with foreknown effects.[57]

But I think the libertarian, speaking as a philosopher of action, has a reply. To see it we must reflect on the necessary properties of (rational) actions.[58]

Actions must be caused by rationalizing reasons, rational actions by rational reasons. What is a reason for action? Standardly, it is a complex of beliefs and desires—a belief that a certain behavior would have a certain property, and a desire to do something with that property. Thus a behavior is an action if caused by and concordant with, a reason. An action is rational iff concordant with a reason the desire component of which reflects the agent's sum-desires and the belief component of which reflects what the agent believes. Thus, a rational action is the reason-caused behavior most appropriate on the occasion of action for furthering the agent's total self-perceived interests, for causing what he sum-prefers to have brought about at the time given his beliefs. Put this way, it is odd to think of actions as able to issue from mere dispositions, things not obviously reasons.[59] But we can count such behaviors actions if they issue from dispositions under certain conditions, namely, if the dispositions are had for reasons—i.e., are caused and rationalized by a complex of beliefs and desires—and if they induce behaviors concordant with the preference component of that complex.[60] Moreover, such behaviors may count as rational actions iff they are caused by dispositions in turn rationalized, caused, and sustained by reasons the preference components of which are rational to have given one's beliefs.

Is it not possible to act rationally without having a full-fledged reason? Perhaps. But the arguments I am about to give will go through even granting this, so long as it is impossible to act rationally in behaving a certain way if one knows of a conclusive rational reason *not* to behave that way. And I think we have just seen that one has such a

MacIntosh, Duncan, ***Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem*** , Southern Journal of Philosophy, 26:4 (1988:Winter) p.499

reason not to behave co-operatively where Gauthier and Campbell think co-operative behavior would comprise rational action.

We are considering the proposal that a disposition is,

(9) a rationally acquired, informed, irrevocable causal tendency to behave co-operatively with those with a similar tendency.

Here dispositions are like irrevocable 'reflexes' (so-called to distinguish them from belief-desire complexes), perhaps little devices (natural, artificial, whatever) in people's heads, which are *causally sensitive* to each other. When agents recognize conditionally co-operative reflexes in each other, these reflexes cause each other to 'go off,' directly causing their carriers to behave co-operatively.[61]

Now, people with such reflexes *will behave* co-operatively. But the behaviors they cause will not be rational. As we have just noted, rationality is a property of actions, not of mere behaviors. Actions are proximally caused by complexes of beliefs and desires, usually intentions, but perhaps even by dispositions or reflexes, so long as *these in turn* express concurrent sum-desires given beliefs (or at least do not go against them). Ones which do not so express are not actions, or at least not rational, directly morally evaluable ones. Nothing rational *can* prevent free agents from acquiring the SM disposition. For no beings governed solely by rationality, and so by no rationality-extraneous form of causation, would fail, it seems, to reacquire the SM disposition. Thus it must be something non-rational—say a cause that is not a reason—that prevents them from reacquiring it. But that means the continuing CM disposition is not rational at the time it is supposed to take effect.[62] And a behavior caused by a disposition it is not rational to have is not a rational action. Thus if they have (9) in mind, neither Gauthier nor Campbell has proved the rationality of co-operative behavior *qua action.*

The difficulty with (9) then, is that no right reason (*qua* concurrent complex of beliefs and desires figuring as the immediate antecedent of the action) of the agent's proximally causes and rationalizes his behavior. He is, instead, directly induced to co-operate indifferently to, and indeed, against, his own desires or preferences at the time of co-operation. To be sure, he knows he will not be cheated against; the other's disposition prevents this happening given his own disposition. But that is not sufficient to make co-operation rational. The motivational component of a rational complex of beliefs and desires that would issue in a freely chosen action is missing, however much the cognitive component may be satisfied by the belief that the co-operator is safe from defection.

Note that this objection to (9) is *not* that one can not *act* rationally *and* be *determined* to act rationally; one can. What is problematic is being determined to act *ir*rationally, to *behave non*-rationally, to behave *without* an appropriate reason as the proximal and rationalizing

cause of the behavior supposed to count as an action. Compatibilism works fine so long as what is determined are rational actions *via* rational dispositions. Act-rationality needs no other kind of freedom. It is enough that one's behaviors are directly caused by a concurrently sum-preferable disposition, and would not have occurred otherwise.[63] But choices and actions are not *bona fide* and *free* (not even in the compatibilist sense) if they do not express concurrent preferences, but rather merely unshakeable reflexes. Such dispositions do not express extant preferences, so behaviors issuing from them satisfy neither the freedom nor the agency conditions, outlined above, and so are not moral actions.

Rational actions must issue from dispositions it is rational to have. That is, they must issue directly from reasons (preferences given beliefs) as they would *if* it were straightforwardly rational for the L-agent to co-operate, since, *ex hypothesi,* the *only* thing that makes a rational L-agent do *anything*, is a *rational* reason. It is apparently irrational for an L-agent to co-operate with another L-agent, because he cannot have the guarantee that if he does, so, likely, will the other person, and that if he does not, neither, likely, will the other. It seems the only way he could get that guarantee is if he and his 'opponent' in the game were to acquire causally efficacious and irrevocable CM dispositions. But if *these* cause co-operative behaviors, the behaviors are not *rational* actions. Campbell, and perhaps Gauthier, wanted a hybrid system where what makes it epistemically rational to *expect* reciprocal co-operation is perception of irrevocable mechanical determining CM dispositions in each other, these expectations, in turn, supposedly justifying co-operative *actions*. Yet the behaviors involved are precisely things prevented from being rational actions because they are brought about, in contravention of extent preferences, by rationality-*frustrating* mechanisms. One co-operates not because it is immediately rational to do so, but because directly caused to do so by one's CM disposition.

V. Possible Replies to the Libertarian Objection

i) It might be replied that it *is* rational for the agent to *choose* to co-operate given the strategic belief that his CM disposition increases the probability of him meeting with co-operation over that of him meeting defection, which it does.[64] But this is not enough. For the agent to be rational in co-operating, he must do so because its *expected utility* is higher than that of defection.[65] But while the disposition increases the probability of reciprocal co-operation, it leaves one with a sum-preference for minimization of one's jail sentence. Since one's utility from successful unilateral defection would be higher than from mutual co-operation, the expected utility of defection is made even higher than that of co-operation by the guarantee that the other agent will co-operate. So rationality requires one to defect.[66] Thus, though the agents will co-operate, they do so *not* because that is rational. Rather, they are

merely *directly* caused to co-operate by their disposing devices. The dispositions do not induce co-operation by affording a reason to co-operate, but by a causal process itself independent of reason (at the time of choice). The disposition, at the time it causes co-operative behavior, is not rationally had (however rational it was to have acquired it) and so cannot issue in a rational action.

ii) It might be objected that people with an enduring CM disposition are not being irrational, or non-agents, when behaving from that disposition, even if it would have been more utility maximizing for them to have changed or escaped from it and defected. Rather, they chose the irrevocable CM disposition. They knew it would cause them to co-operate even where, had they not chosen it (but say, only seemed to have in order to trick another agent into co-operating), they would not have co-operated. And when they co-operate they *are* directly acting on a rational disposition, one rational to form on Gauthier's rationale. Since it *is* rational to *form* that disposition and since *it* is the direct cause of co-operative behavior, such behavior *is* rational action on a disposition it is rational to have. Thus co-operation under these conditions is rational action in the fullest possible sense.[67]

Unfortunately, the preferences which cause and rationalize acquisition of the disposition formed under the Gauthian rationale do not do the same for the co-operative behavior it causes. The former are preferences to have a reflex tendency to behave according to the general strategy of narrow compliance, not to undertake the specific behavior of co-operating *now* (i.e., at the time Second finally chooses in the sequential PD). They are preferences for dispositions to co-operate, not for specific co-operative behaviors. A disposition must be one it is sum-preferable to have to be sufficiently rationalized to serve as the cause of a rational action. But the preferences that rationalized the CM disposition only did so at the time of its selection, not of its activation. This is evident from the identity conditions for preferences.[68] X and Y are the same preferences only if preferences for the same thing. They are preferences for the same thing just if whatever events would consist in their satisfaction necessarily co-occupy the same spatio-temporal regions. They do that only if they or their counterparts so co-occupy in all possible worlds in which either exists.[69] But the preference for adoption of a co-operative strategy is realized by different events than would realize such specific preferences for co-operative behavior as might be had in a specific instance, and as would be needed to rationalize behavior from a disposition on a given occasion. The former is realized by a psychological event—by acquisition of a disposition to behave co-operatively—the latter by a specific behavior—by co-operating on some occasion. As they are preferences for different events, they are different preferences. Thus, the preference to acquire the CM disposition does not rationalize the specific co-operative behavior in question. Therefore, the latter is not made rational behavior

MacIntosh, Duncan, ***Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem*** , Southern Journal of Philosophy, 26:4 (1988:Winter) p.499

simply because caused, in part, by the preference to form the CM disposition. Indeed, in the Gauthier/Campbell model, the co-operative behavior appears to have *no* preference as *its* proximate *rationalizing* cause. It runs *against* the agent's concurrent sum-preferences.

iii) It might be replied that this misdescribes the preference for a co-operative disposition, making it unable to rationalize specific co-operative behaviors, and creating a spurious need for additional preferences for specific behaviors not found in Gauthier's account. Suppose we eliminate this need by construing the preference to acquire the CM disposition as simply a standing preference to perform any behavior dictated by the strategy of narrow compliance in specific situations. Suppose the specific behavior of co-operation on a given occasion is a behavior dictated by the strategy in light of concurrent information about that situation. Then the behavior is caused directly by the preference to perform any behavior dictated by the strategy, and so is rational in the fullest possible sense: it is directly caused and immediately rationalized by the aforementioned preference, a preference the details of which are completed by the acquisition of specific beliefs about specific occasions from occasion to occasion.[70]

This would be terrific but for one thing. The agent knows the utility of having the disposition *qua* preference for following a certain strategy has changed between the time of the formation of the preference to conform to the strategy (or the time at which his psychology will be inspected by his competitor in the game) and the time at which a specific co-operative behavior is to be performed (in secret, or at a later time after the other has chosen). It is known to be more utile *not* to have the CM disposition *at the later time*. Rationality thus dictates its abandonment at that later time. Only a non-rational force could prevent that. The CM disposition *qua* standing preference for conditional co-operation is thus, at the time of the behavior in question, one that it is irrational to have. It is thus, *then*, an irrational disposition, issuing in irrational behavior.[71]

Think of this example: The preference to go to town by following the strategy of taking the bridge is one it becomes irrational to have and act on upon one's learning that the bridge is out. If one nonethelss behaves concordantly with the original preference because of the disposition it causes in one, or because the preference is somehow unrevocable, one is a paradigm of irrationality. Likewise, if one formed the preference to follow a strategy of co-operation just in order to induce another to co-operate so as to minimize one's jail sentence, and he is, just before the time of one's prospective action, believed to have co-operated, it is no longer rational to retain that preference, nor to act on whatever disposition it induced. For even greater minimization of one's jail sentence would result from abandonment of the disposition and adoption of one to defect.[72] Thus retention of, and action upon, the disposition or preference to adhere to the co-operative strategy is irrational, and can only result in an irrational action.

MacIntosh, Duncan, ***Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem*** , Southern Journal of Philosophy, 26:4 (1988:Winter) p.499

iv) Perhaps it will be replied that the bridge and co-operation cases are crucially different. Whether the bridge will be intact is not influenced by one's disposition to use it to get to town. By contrast, adoption of the CM disposition increases the probability of reciprocal co-operation. Moreover, unless the disposition is known by both parties to be enduring, it will not have that effect. So one acquires it with the shared knowledge that it will endure. Now Second comes along and co-operates from a disposition to perform any co-operative actions dictated by the strategy of narrow compliance, intending to be unable to abandon it given the opportunity to cheat, and having put himself in this condition for a perfectly rational reason, namely, that it increases the probability of a shorter jail sentence. Surely now Second's co-operation is a rational action from a disposition it is rational to have, in every possible sense, even if he cannot escape acting on it.[73]

I answer with two observations. First, that it was rational to acquire a CM disposition understood to be irrevocable, does not prove it is irrational to revoke it in light of a change of circumstances *if one can*. It is presumptively rational for Second to do that after First has chosen, given the standard PD POS. This is because the standard PD POS implies that both parties would sum-prefer to have succeeded at unilateral defection rather than to have participated in mutual co-operation. Thus the only thing that could possibly prevent Second from revoking it is something non-rational, something which is not itself a *reason* not to do it. Thus, continuing to have and act on the disposition at the time of choice is not rational. Therefore, actions resulting from the disposition are irrational. Of course Second will co-operate, but not because doing so is rational for him.

Second observation: Suppose my disposition to go to town by the bridge probabilifies the intactness of the bridge because I tell my foreman I prefer to get to town by following the strategy of crossing the bridge. He only attempts to maintain the bridge because he thinks my commitment to the strategy will endure, making it likely I will try to use the bridge. I only commit to *that* strategy on the understanding that it will be irrevocable because I know that doing so will induce him to attempt to maintain the bridge. Alas, try as he might, the bridge still washes out. Along I come, I see that the bridge is out. Surely my attempted crossing of the bridge through the influence of the disposition to cross it would be irrational. If only I could revoke the disposition I ought to. Only some non-rational, causal force could keep me on my doomed track. Attempting to use the bridge then is, perhaps, an action, and one flowing from a disposition it was rational to adopt, but not from one it is now rational to have. So it is not a rational action.[74]

v) One final attempt to support my objector's suggestion: Surely since what prevents revocation of the co-operative disposition is not any present intention, desire, or action of the agent's, the influence of that preventative on the agent's behaviors does not render them irrational.[75]

MacIntosh, Duncan, ***Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem*** , Southern Journal of Philosophy, 26:4 (1988:Winter) p.499

After all, if something like this *makes* my finger pull the trigger, that does not mean that *I* have *done* something immoral. (I may have adopted an immoral disposition in disposing myself to murder, but if the actual killing is done, say, in a trance over which at the time I have no voluntary control, the killing is a non-moral behavior. I still get the chair because of my originally preferring to kill someone, and for causing myself to go into a killing trance, but not for an immoral action in a trance. It was not an action, and so not an immoral one either.) Why say something similar implies that *I did* something *irrational*?

This is, perhaps, a reasonable point, but conceding it does not defeat my thesis. Suppose I grant that the influences of such forces make one's behaviors non-moral and non-rational rather than *immoral* and *irrational* respectively. Then, *a fortiori*, they are not fully rational and moral actions. So it still has not been demonstrated that co-operation under such conditions is *rational action*. Think of it like this. One co-operates not because of a direct, presently rational preference to do so, but because of a *consequence* of a past rational preference, namely, the preference to acquire the irrevocable tendency to behave co-operatively. The thing has 'taken on a will of its own' so that you now co-operate willy-nilly *whether it is rational to do so now or not.* Behaviors of yours that occur whether it would at present be rational for them to do so or not are precisely behaviors which do not count as actions, or at least not as fully rational ones. (Or, if they are actions, ones initiated earlier in the choice of a disposition which would cause them, certainly they are not presently rationally chosen actions, as Gauthier thinks they are.) And the same is true of behaviors caused by dispositions you would have whether it was immediately rational to have them or not.

The only rationally and morally evaluable thing in the vicinity in (9) is that of the *acquisition* of the reflexes or devices. For the agents, we are assuming, have rationally and wittingly acquired them, foreseeing the morally and practically significant consequences of doing so. And where the agent foresees what such a device or reflex would force him to do, perhaps we can morally and rationally evaluate his *behaviors.* But I think it preferable to evaluate him for *putting* himself in a condition with more or less desirable behavioral *consequences.* We do not, after all, speak of the rationality or morality of falling, only of jumping. However responsible agents may be for acquiring co-operative dispositions when foreseeing their effects, that does not, I think, by itself make behaviors issuing from those dispositions into *actions*, i.e., into things consequent upon individual *choices* for each *occasion* of behavior. The dispositions analyzed as in (9) produce no action, then, because the resulting behavior does not express concurrent sum-preferences given beliefs. It occurs instead as a mere reflex.

MacIntosh, Duncan, ***Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem*** , Southern Journal of Philosophy, 26:4 (1988:Winter) p.499

VI. Other Construals of the Nature of CM Dispositions Face Similar Problems

We still need, then, some other way of giving guarantees that co-operation will or will likely be met with co-operation, defection with defection; some way other than contra-rationally, deterministically guaranteed, enduring dispositions; some way that will make the resulting behaviors count as actions.

Some candidates immediately come to mind. We have just seen a conclusive difficulty with the proposal according to which one's own and the other's psychological natures make CM dispositions, once (instrumentally rationally) acquired, a-rationally caused to endure through all temptations which would rationalize the reacquisition of the SM disposition. This sluggishness of disposition, this non-responsiveness to new circumstances, precisely consists in being contra-rationally determined not to acquire a disposition it is rational to acquire. Alas, most other proposals I know of have the same structure. Consider one I learned from Geoffrey Sayre-McCord, one which may be very close to Gauthier's intent:[76]

(10) The CM disposition is an Aristotelian character trait acquired first by brute choice, then strengthened by the enhabituating effect of repeated practice, by choosing the same thing again and again, each choice making it harder to choose otherwise the next time.

There are two conclusive difficulties with this. First, if it is not rational to co-operate, then it is not rational to do so the first time either. This is especially true here, where, because by hypothesis it *is* the first time, it is not even rational in the derivative (and contentious) sense of being rational given an enduring disposition, itself rationally acquired. The disposition is not yet in place. Secondly, if one does something merely out of habit or through acquired character, even were one rational in and morally responsible for the acquisition of the character, one is not necessarily responsible for actions done once it is acquired in situations where it would be rational not to act from it if one could refrain from so acting. One must be allowed a reasonable time to alter it and must have the capacity to alter it. The foreseeability of such situations does not make behaviors from that character into rational actions either. Of course, one may be accountable on other grounds for the behaviors that character causes because accountable for putting oneself in a condition from which they would issue, i.e., for acquiring that character—see the original objection to (9), above.

Now consider a science fiction proposal (which I stole from a *Star Trek* episode):

(11) The CM disposition is sustained by the implantation of a device (or through the inculcation by hypnosis of a psychological condition) which is correctly anticipated by the agent to cause acute pain (physical pain, moral anguish, whatever) upon defection against someone known to have a similar device (or condition).

It might be thought that here, the CM disposition is sustained by threat of pain rather than by an a-rational causal force. Thus, one's rationality is still rationally responsive to circumstances (one of which happens to involve one's own psychology). Thus the CM disposition remains one it is rational to have and act upon. Therefore, it issues in fully rational actions.

But there are two problems with this. First it alters the POS. This device or condition effectively makes it sum-preferable to have participated in mutual co-operation rather than in successful unilateral defection, to which there now attaches great disutility. Second, if this is somehow false, if one continues to prefer to have successfully unilaterally defected, but feels coerced into co-operation by this internal psychological trait, then we just have something like an external coercion method, which Gauthier dismisses as uninteresting.[77] The only difference is that the coercive force is provided by a psychological mechanism. Co-operation is, of course, rational even for contra-causally free persons where there is an external system of enforcement that detects and penalizes defectors. But as just noted, this alters the POS for choices. It also violates the freedom conditions (on some interpretations), stated above, something already corroded with the apparent need for some degree of determinism in rational co-operation.

The point is, it involves something one wishes was not operating whenever put in a position where one could otherwise cheat. This makes behaviors issuing from it ones that do not issue from an impulse it is rationally desirable to have *at the time* it issues in co-operation. To be sure, it may be rational to acquire these things, and then rational to act so as to avoid the punishment they mete out. But the same can be said of an authoritarian state which detects and punishes defectors. And in the moment of a potential cheat, if one could throw them off, rationality would oblige one to do so.[78] Instead, we want to know whether free persons in a PD could ever voluntarily rationally co-operate, absent threat, and absent rationality-resistantly-stable, determining dispositions. These, as we have seen, are little more than internal coercions.

VII. The Basic Problem with the Gauthier/Campbell Solution

We have noted a problem with the Gauthier/Campbell Solution: no theory of dispositions in which they are independent of concurrent sum-preferences will allow all its claims be true together.

Both writers think rational agents would prefer to have a CM disposition. Now such preferences might take one of two forms. First, agents could prefer one only so far as having it alters the *probability* of reciprocal co-operation, it remaining sum-preferable to have succeeded in unilateral defection. Call this an instrumental preference. Second, agents could prefer one for its value in securing co-operative behavior for its own sake, this increasing the *utility* attaching to mutual co-operation, it becoming sum-preferable to have participated in it. Call this a non-instrumental preference.

MacIntosh, Duncan, ***Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem*** , Southern Journal of Philosophy, 26:4 (1988:Winter) p.499

Gauthier wants to keep the PD POS constant, and merely alter the *expected* utilities by offering a change in the probability that co-operation and defection will each be reciprocated by introducing the CM disposition. The desire to effect that change justifies acquiring the disposition. But if it is merely instrumentally preferred, the justification for keeping it vanishes in the sequential PD. If having it makes others co-operate, it becomes rational, once they have made their choice, to quit the disposition and defect. Instrumental reasoning justifies cheating since agents have no commitment to co-operation for its own sake. Thus, players have no reason to think it rational for each other to co-operate just because each now has a disposition to do so. Instead, each rationally should change his disposition when actually choosing, if he can. This is why L-free rational agents with only an instrumental preference, and with the metaphysical ability to defect in the course of maximizing, may be predicted to defect.

Campbell's proposal is supposed to guarantee reciprocal co-operation in spite of (what I would call) its *act*-irrationality. For him, dispositions appear to be stable reflex mechanisms which, once rationally acquired, are caused non-rationally to endure. They operate, again, not as a changed set of total preferences, but as concurrent-preference-independent reflex tendencies. They simply persist in causing co-operative behavior, indifferently or contrarily to the immediate dictates of act-rationality. But then, as we have seen, the behaviors they cause will not be rational actions. Rational actions must express concurrent sum-preferences given beliefs. When one sum-prefers unilateral defection to minimize one's jail term given an opportunity for it, sustaining and acting on any disposition or isolated preference to the contrary—e.g., the CM disposition or the instrumental preference to conform to the CM strategy, respectively—is irrational action on a disposition (or instrumental preference) it is now irrational to have. This holds even assuming actions can issue directly from stable reflex tendencies out-dating the preferences which gave them legitimacy. The CM disposition is flat-out incompatible with the preferences determining one's utility functions at the time of choice of action. Thus actions from it would not express one's rationally ordered preference set. If, all this not withstanding, one sustains and acts on an instrumental preference for conditional co-operation given an opportunity to cheat, we may have to say that one's preferences are not coherently ordered: the preference to co-operate is incompatible with the preference to minimize individual jail time.

To insist in the face of this that co-operative choices and actions *would* express one's concurrent sum-preferences as reflected in the standard PD POS, or would in some *other* sense be rational actions, would require rewriting action theory. One might, for instance, try rehabilitating the defenses of (9) I have rejected. But this is unpromising. There is the strong intuition that one acts (perhaps

rationally) in jumping, but (certainly) not in falling. Similarly, one may act in choosing the stable CM disposition, but not, I claim, in exhibiting behaviors compliant with it. Another alternative, one which I take up elsewhere,[79] would be to conceive rationality in a different way. (There is some textual evidence to think that Gauthier in fact does this, though as much to think that he does not.) Perhaps we should conceive an action as rational just if it expresses a stable rationally chosen disposition, rather than if it maximizes satisfaction of present preferences. That route too, though, presents difficulties, since it challenges the core concept of practical rationality expressed in voluntary behavior, *viz.*, that a rational person does what he at present prefers to do all things considered, given his beliefs. Moreover, as we have seen here, if the standard by which one evaluates a disposition is whether having the disposition is maximizing, co-operation is still not rational. For while it is maximizing to adopt the CM disposition, it is not maximizing to retain it. This required that it be made irrevocable. But then one is faced with a conflict: It is maximizing to adopt an irrevocable CM disposition, maximizing to revoke it though it cannot be revoked. If, to be rational, one must always have a maximizing disposition, it is impossible to be continuingly rational in this situation. One would have to adopt and revoke a disposition by hypothesis irrevocable. Thus, since the theory that to be rational is always to have a maximizing disposition gives incoherent advice to the rational agent, it cannot be a correct theory of rationality.

Since perfectly free and rational persons will defect even if they begin with Gauthier's CM disposition, they will not comply with the obligations of morality. Unfree persons with a stable reflex to co-operate discordant with their concurrent sum preferences will not act rationally in so co-operating, however morally compliant their behavior will be. Thus, neither writer has a solution to the compliance problem in which moral behavior proves to be rational action.

Their proposals attempt to resolve the PD without requiring people to come to the game with a non-instrumental preference for co-operation with those with similar preferences. Each assumes the standard PD POS and then tries to generate a co-operative solution while trying to leave it intact. This scruple is usually thought essential to the objectives of contractarianism. But in both of these proposals, given the standard POS, it is always rational to defect. Any other choice, no matter how caused, will not be the most rational action. Thus, these proposals cannot consist in rational-action solutions to the PD.

Is it impossible then for rational agents who originally lack the non-instrumental preference for co-operation with those with a similar preference to ever come to freely act co-operatively in a PD? Is morally compliant behavior a possibility for them with the standard POS only so far as they have so causally configured themselves as to behave irrationally? Must we accept the paradox that one's preferences are best

MacIntosh, Duncan, ***Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem*** , Southern Journal of Philosophy, 26:4 (1988:Winter) p.499

satisfied by conduct that defies them? I think not, though there is not space to explain in detail here why. But in a companion paper,[80] I develop the idea that Gauthier-like arguments are able to rationalize revising one's preference set, so that one comes, all things considered, to non-instrumentally prefer co-operating with (just) those with similar sum-preferences. These arguments do not just rationalize acquiring a preference-resistant constraint on the expression of extant preferences in choices. Once having acquired a non-instrumental preference for co-operative behavior with similar agents, agents, in the course of individually maximizing their utility, would find it rational to co-operate with those who, because of possession of a similar non-instrumental preference, would reciprocate. The agents would thus act in ways that would mutually minimize jail time, rather than just individually minimize it. Co-operation would then be rational action in PDs.

NOTES

[1] This paper is, in effect, a reply to Richmond Campbell's (1988a), first presented at Dalhousie University in 1986. I am grateful to Campbell for pointing me at the relevant literature, for his searching criticism, and for his editorial comments. My thanks are also due to Neera Badhwar, Robert Bright, Douglas Butler, Robert Martin, Geoffrey Sayre-McCord, Terrance Tomkow, Kadri Vihvelin, and Sheldon Wein, and to an anonymous referee who requested some important clarifications in the exposition of Gauthier's program. I am especially indebted to Julia Colterjohn, whose role as foil is evident throughout, and who read earlier versions with undaunted incredulity. My thanks to the Killam Trust of Dalhousie University, whose post-doctoral fellowship support I enjoyed during the spring and summer of 1986.

[2] See Campbell (1988a), pp. 192-195, where the doctrine is referred to as the 'Reduction Thesis,' morality reducing to a sub-portion of rationality, and Gauthier (1986).

[3] Gauthier (1986), p. 15, and Ch. VI.

[4] *Ibid.*, pp. 164-5, and more generally, pp. 170-189.

[5] *Ibid.*

[6] Campbell (1988a), pp. 200-208. I also borrow some of his more explicit phrasings from Cambpell (1988b) and (1986).

[7] Gauthier (1986), p. 177.

[8] *Ibid.*, p. 2.

[9] *Ibid.*, p. 2.

[10] *Ibid.*, p. 3.

[11] *Ibid.*, p. 79.

[12] *Ibid.*, p. 158.

[13] *Ibid.*, p. 168.

[14] *Ibid.*, p. 169.

[15] *Ibid.*, p. 169.

[16] *Ibid.*, p. 189.

[17] *Ibid.*, pp. 164-165, my emphasis.

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

[19] *Ibid.*, p. 37.

[20] *Ibid.*, p. 343.

[21] *Ibid.*, p. 187.

[22] Campbell argues for this in his (1988a), pp. 200-205.

[23] Gauthier (1986), pp. 185-186, and p. 165, where he says, in another connection, that "it is rational to comply with an agreement if it is rational to make it." See also, Campbell (1988a), pp. 200-205.

[24] See the passages cited in the preceding note.

[25] Campbell (1988a), pp. 200-208.

[26] Campbell argues in his (1988a), pp. 200-208, that co-operation is irrational if one can escape the CM disposition; it is rational only if that disposition continues to thoroughly condition one's choices. I doubt if it is rational even then.

[27] The foregoing conditions on moral action are widely discussed in several literatures. Campbell discusses the problem of the origin of preferences for rationalist and reductionist programs in his (1988a), Section IV. On the matter of rationality requiring responsiveness to new situations, Gauthier himself, in his (1986), remarks of rational action that "Utility is . . . the measure of present preference[s] . . . of the self at a particular time; practical rationality is the maximization of utility and so the maximization of the satisfaction of present [or, as I shall say, "concurrent"—concurrent with the time of choice or action—D.M.] preferences" (p. 343).

[28] Gauthier (1986), Chapter I. E.g., "We shall defend the traditional conception of morality as a rational constraint on the pursuit of individual interest" (p. 2). Contrasting his position with that of other philosophers, he writes, "Neither Rawls nor Harsanyi treats moral principles as a subset of rational principles for choice" (p. 4). See also Campbell (1988a), pp. 192-194.

[29] This is a variation on standard expositions of the Prisoner's Dilemma. See, e.g., Campbell (1985), pp. 3-8, and Gauthier (1986), pp. 79-80.

[30] Gauthier states the problem in Hobbesian terms in his (1986), pp. 158-166.

[31] Gauthier (1986), Chs. V and VI. His general treatment is given in Gauthier (1986), Chs. I, VI, X, and XI.

[32] Gauthier (1986), pp. 168-169, pp. 173-183.

[33] *Ibid.*

[34] *Ibid.*, pp. 178-180.

[35] *Ibid.*, p. 173.

[36] *Ibid.*, pp. 15, 170, 173.

[37] *Ibid.*, p. 173.

[38] *Ibid.*, p. 178-9.

[39] *Ibid.*, p. 186.

[40] *Ibid.*, p. 187.

[41] *Ibid.*, Ch. VI, Sect. 2.5.

[42] For references, see note 6, above.

[43] Who knows what having a CM disposition may amount to for a Libertarian free agent? Perhaps it consists in a tendency to feel guilt when exploiting others, or in a propensity, but not an irresistable one, to co-operate. Of course, there is some sense in which an L-free agent with a CM disposition thus construed is practically indistinguishable from one with an SM disposition, which is just what Campbell's argument is designed to reveal, and what it exploits in criticism of Gauthier. For vividness, I shall continue to speak as if it made some sense to ascribe a CM disposition to an L-free agent, acknowledging its evident hopelessness as a way of guaranteeing that he will co-operate.

[44] Campbell (1988a), pp. 196-198. Note that this worry holds even assuming Gauthier could explain how dispositions can rationally issue in behaviors commensurate with them. Even if we were to grant that it is rational to act on any disposition one has, we would need another argument to show why we should not revise our dispositions before acting on them.

[45] Campbell (1988a), pp. 196-198. Being totally free, they do not *have* to defect; they may yet behave irrationally. But defection will be rational for them, and we are assuming they will perform rationally.

[46] Campbell (1988a), pp. 196-198, and pp. 199-200. Gauthier in his (1986), p. 177, says that, "[a] constrained maximizer is conditionally disposed to co-operate in ways that, followed by all, would yield nearly optimal and fair outcomes, and *does* co-operate in such ways when she may actually expect to benefit." He does not say exactly *why* she *does* co-operate (at least not in that context); surely she would benefit more, in the maximizing sense, if she did not co-operate. Campbell thinks she would only do so if she *had* to give

her disposition; otherwise she would defect. It seems to me then, that on this interpretation, the disposition has to *determine* co-operation, for it cannot rationalize it. At the end of his book, Gauthier sometimes implies that the reason she will co-operate is that she will rationally come to *prefer* to do so. I take up this alternative interpretation of the force of Gauthier's arguments in my (1988a), arguing that that interpretation affords the only hope for a solution to the compliance problem that yields co-operation as free, rational, voluntary action. In my (1988b), I deploy a similar argument to defend Gauthier's claim that nuclear retaliation is rational upon provocation if it follows formation of a rational intention to retaliate prior to provocation.

[47] Campbell (1988a), pp. 203-204.

[48] *Ibid.*, pp. 200-202.

[49] *Ibid.*, p. 203.

[50] *Ibid.*, p. 203, and Campbell (1988b), (1986).

[51] Campbell (1988a), pp. 201-202, and Campbell (1988b), (1986).

[52] Campbell (1988a), p. 203.

[53] *Ibid.*, pp. 192-194, 203-205.

[54] Note: This does not prove co-operative behavior issuing from such a disposition would be rational action. Only that whether those behaviors are actions and rational ones or not, it may be rational to want everyone to have and continue to have, such a disposition, and in wanting to acquire it, or in wanting to continue to have it, to want it to have those issuances, this to stave off mutual defection.

Now, someone with no preferences for justice for its own sake might think it yet more preferable to have a CM disposition *he alone* can escape. He might then use it to trick others who will be bound by their irrevocable dispositions to co-operate and whom he can then exploit. But of course if he is transparent to others (as we are assuming he will be) they would detect the revocability of his disposition and refuse to co-operate with him. Knowing this, then, he has excellent instrumental reason to prefer that his dispositions genuinely bind his behaviors. But it does not follow that behaviors issuing from those dispositions are rational. Only that whether rational or not, they issue from a disposition welcome at the time of its adoption. One is glad to have the option of being subsequently forced to behave irrationally.

Campbell thinks that the rational preferability of the stable CM disposition over any other assures the rationality of co-operative choices, since, he thinks, according to Gauthier, rationality as utility maximization is applied first at the level of dispositions to choose, rather than at the level of first-order choices. Were it not—were it applied first to choices, then to dispositions, one's 'disposition' thereafter being read off one's choices—rationality as individual act-to-act utility maximization would leave one as if with the SM disposition, a disposition which, paradoxically enough, affords one less utility than the CM disposition; hardly utility maximizing. To keep the conception of instrumental rationality consistent and "self-supporting" then (Gauthier's phrase), rationality must attach first to dispositions, then to choices. See Campbell (1988a), pp. 200-205.

However, I think we have the reverse paradox if we go Campbell's route (and Gauthier's, if Campbell's reading of him is correct); we end up with specific choices which are not utility maximizing, though with dispositions which are (at least at the time of their initial selection). Something is wrong. I argue in my (1988a) that the most consistent way out is to acknowledge the rationality of preference revision, and of co-operative choice upon acquisition of a non-instrumental preference for co-operating with those with a similar preference. Here, instrumental rationality will, throughout, be simple maximization of expected utility, except applied first to a second-order choice among non-instrumental preferences for actions, and then to the first-order problem of choice among actions given those new preferences.

[55] Campbell (1986).

[56] Campbell (1988a), pp. 201-202.

[57] *Ibid.*, p. 204.

[58] I here follow the classic early Davidson (1963) conception of practical rationality and of rationalizing explanation. I have also consulted Brand (1984). Generally, I take what

follows to be entirely consistent with (at least the bulk of) what Gauthier says on the subject in his (1986), Ch. II and final chapter. E.g., "[rational] choice maximizes preference fulfillment given belief." (p. 30); "rational choice must be directed to the maximal fulfillment of our present considered preferences" (p. 37). See also p. 343.

[59] They are not obviously intentions either, things many philosophers think must be the proximal causes of behaviors before the latter may be called actions. I here follow Myles Brand, however, who thinks that there are non-intentional actions, and that in any case, the proximal cause of actions must be something more complicated, fine-grained and sub-conscious than intentions or reasons (*qua* complexes of beliefs and desires); they are more like dispositions conceived as complex kinematical programs comprising information and command sets as their cognitive and conative elements, respectively. Good athletes want to win and believe brilliant moves achieve victories, but what directly causes their brilliant moves in competition are rehearsed dispositions, inculcated to the point of being automatic reflexes. Their proximal causes are not clumsy belief-desire complexes, which would be too slow to effectively choreograph the many intricate details of the behavior. Specific co-operative behaviors and choices could conceivably share this dispositional etiology. See Brand (1984), Ch. 6, "Intending and Believing," pp. 147-170. This is all just a nicety of action theory though, and what will prove important is how dispositions must fit with concurrent preferences if they are to cause rational actions.

[60] This allows us to distinguish the inborn dispositions figuring in such things as knee-jerk reflexes, from those rationally acquired, like the complex of dispositions involved in executing a good jump-shot.

[61] See Campbell (1986).

[62] See my discussion of Campbell's views to the contrary in note 54, above.

[63] Thus, for present purposes at least, I embrace standard compatibilism. See Campbell (1988a), p. 203.

[64] Certainly Gauthier thinks that the sharing of CM dispositions is a sufficient condition for rational agents to voluntarily co-operate. But how can it be voluntary if it is not maximizing? For more on this, see my notes 46 and 54, above, and the concluding remarks in the final section of the present work.

[65] See Gauthier (1986), Ch. 2.

[66] This parallels Campbell's argument in his (1988a), pp. 196-197.

[67] My thanks to Julia Colterjohn for helping to make this objection clear to me.

[68] The following is derived from Myles Brand's account of identity conditions for intentions, actions, and events. See his (1984), especially Ch. 3, "Events as Spatiotemporal Particulars."

[69] See Brand (1984), Ch. 3, "Events as Spatiotemporal Particulars," Ch. 5, "Desiring," and Ch. 6, "Intending and Believing."

[70] Again, my thanks to Julia Colterjohn for helping to formulate this objection.

[71] Both Gauthier and Campbell acknowledge the change in the utility of the disposition, Gauthier in his discussion of Economic man empowered with the Ring of Gyges, in his (1986), Ch. X, Campbell in his demonstration of the impossibility of rational co-operation for L-free agents.

This situation differs from that of Ulysses and the Siren. Agents in the PD have a POS such that they most benefit from successful unilateral defection. An agent whose CM disposition has presumptively induced another to choose to co-operate, now does best if he can abandon his CM disposition and defect. That is why he is irrational if he does not do so. Ulysses' POS is one in which he does best if he does not succumb to the Siren. Thus, once bound to the mast, he does not do better if he is released and goes to the Siren. So arguably, if he does not go to the Siren—because he is bound to the mast—he is not failing to act rationally. On the other hand, he is irrational at the time of being seduced by the Siren's voice since he has an irrational preference to go to her. Just the reverse of the PD predicament.

[72] See Campbell (1988a), pp. 196-197, and Gauthier (1986), Ch. X.

[73] My thanks again to Julia Colterjohn.

[74] Gauthier has argued that it is sometimes rational to carry through with threats, the

analogy here being the 'threat' to attempt crossing the bridge whether it is safe or not. See his (1986), pp. 185-6, for instance. In my (1988b) I argue that threat following-through is not rational if the disposition to follow through is merely a permanent mechanism. If I am right there, Gauthier's arguments on the rationality of threat follow-through are no help in the present case.

[75] The useful and tenacious Julia Colterjohn was once again a help in advancing this suggestion.

[76] He suggested this to me in a quick conversation while distracted by some photo-copying he was doing; I may not have his view right.

[77] Gauthier (1986), pp. 163-165.

[78] *Ibid.*, pp. 163-165.

[79] See my (1988a) and (1988b).

[80] See my (1988a).

REFERENCES

Brand, Myles (1984): *Intending and Acting: Toward a Naturalized Action Theory,* A Bradford Book, The MIT Press, Cambridge, Massachusetts.

Campbell, Richmond, and Sowden, Lanning (eds.) (1985): *Paradoxes of Rationality and Co-operation: Prisoner's Dilemma and Newcomb's Problem,* The University of British Columbia Press, Vancouver.

Campbell, Richmond (1985): "Background for the Uninitiated," in Campbell and Sowden (eds.) (1985), pp. 3-41.

Campbell, Richmond (1986): Correspondence with Campbell following his presentation of his paper, (1988a), at Dalhousie University.

Campbell, Richmond (1988a): "Moral Justification and Freedom," *The Journal of Philosophy*, Volume LXXXV, Number 4, pp. 192-213.

Campbell, Richmond (1988b): "Critical Study: Gauthier's Theory of Morals by Agreement," forthcoming in *The Philosophical Quarterly*, April, 1988.

Davidson, Donald (1963): "Actions, Reasons, and Causes," *Journal of Philosophy* 60, pp. 685-700.

Gauthier, David (1986): *Morals By Agreement,* Clarendon Press, Oxford.

MacIntosh, Duncan (1988a): "Two Gauthiers?", forthcoming in *Dialogue.*

MacIntosh, Duncan (1988b): "Retaliation Rationalized," manuscript, presented to the 1988 meetings of the *Canadian Philosophical Association.*